upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Department of Statistics, Computer Science, and Mathematics

PhD Thesis

Software Tools and Statistical Methods for Downloading, Processing, and Analysing Satellite Images

Author:
Unai Pérez-Goya

Supervisors:
Dr. Ana Fernández Militino
Dr. María Dolores Ugarte Martínez

Pamplona, 2019

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Nire gurasoei eta nire ondoan egon diren guztiei, izan zirelako gara

Acknowledgments

I would like to thank all the people that made this thesis possible. First of all, I would like to express my sincere gratitude to my supervisors, Dr. Ana F. Militino and Dr. Lola Ugarte for their continuous support, help, and guidance since my arrival at the Public University of Navarre until the final stage of this dissertation.

I am very grateful to the Department of Statistics, Computer Science, and Mathematics of the Public University of Navarre, and in particular to Gonzalo M. Vicente and Dr. Manuel Montesino for their help and support during these years.

I wish to thank Dr. Peter M. Atkinson for his kindness and hospitality during my research stay at Lancaster Environmental Center at the University of Lancaster.

I express my gratitude to the three institutions: a) the Spanish Ministry of Economy and Competitiveness (project MTM2017-82553-R AEI/FEDER grants, MTM2014-51992-R, and MTM2011-22664), b) the Government of Navarre (projects PI015-2016 and PI043-2017), and c) "la Caixa" Foundation (ID 1000010434), Caja Navarra Foundation and UNED Pamplona, under agreement LCF/PR/PR15/51100007, for the financial support.

Finally, my warmest thanks to all my family, especially to my parents and my sister, and to my friends.

Contents

List of Figures

List of Tables

Introduction

Remote sensing deals with the acquisition of information about an object or phenomenon from distance, usually from the Earth surface through sensors or satellites. Since the 60s, the satellites launched for earth monitoring are abundant. However, in the last few years, the new open-access data policy has produced a tremendous growth in this area. Satellites retrieve information with many types of sensors as spectral, charge-coupled devices, or radiometers. The information provided from these sensors can be used in meteorological, agricultural, hydrological, or environmental applications (Ban, 2016a). The spectral satellites usually capture data as images (see NASA, 2018) but these images are not free of errors. The literature provides different methods to reduce outliers and avoid missing values (see Holben, 1986; Vancutsem et al., 2010). Traditionally, mathematical methods based on harmonic analysis or filtering are the most popular for this aim (Roerink et al., 2000; Jönsson and Eklundh, 2004). These methods usually do not use spatial or temporal dependence for image smoothing. Indeed, the use of stochastic spatio-temporal dependence is still more scarce (Bivand et al., 2013).

The main objective of this thesis is the introduction and development of statistical methods in satellite imagery to improve the processing, smoothing, prediction, and inference of remote sensing data. This objective can be split into the following more concrete goals (each goal is studied in a different chapter).

The first goal is acquiring, managing, and automatizing processes to download remote sensing data from different platforms in a standardised way. Chapter 1, entitled "Satellite imagery: data access and download" explains how to achieve this goal, and describes the satellites that nowadays provide images as open-access data. The specific characteristics of each satellite are explained, emphasizing the differences between raw and composite images. The platforms that provide these data are also introduced, and the pros and cons of each platform are presented. All

the satellite programs split the entire world in tiles, but usually, to compose a region, the combination of a set of tiles or the combination of some pieces from different tiles is needed. Therefore, it is important to give procedures for generating regions from the tiles of the satellite programs. Chapter 1 also explains the development of a set of functions to automatize the downloading, customizing, and processing time series of satellite images from Landsat, MODIS, and Sentinel-2 platforms. All the functions have been developed in R (R Core Team, 2019), and encapsulated in a unique package called 'RGISTools' (see Pérez-Goya et al., 2019).

The second goal is to provide a brief review of the main geostatistics tools used in satellite imagery, emphasizing the importance of considering stochastic spatio-temporal methods. Chapter 2, named "The use of geostatistics in the analysis of satellite imagery" achieves this goal providing an overview of existing methods, and analysing the R packages that can be used for satellite image processing. It also explains the procedures for deriving variables such as the normalize difference vegetation index (NDVI) (see Rouse Jr et al., 1974), the enhanced vegetation index (EVI) (Huete et al., 2002), and the land surface temperature (LST) (Sobrino et al., 2004) from spectral images, among others. The procedure for obtaining these variables are explained taking as a reference the images obtained using the 'RGISTools' package. This chapter introduces images preprocessing for avoiding the fails produced by atmospheric factors, clouds, or distortions produced by mechanical components of the sensor (Sola et al., 2014). Some methods proposed in the literature to reduce the effect of atmospheric factors are also described. We also explain specific methods defined for each variable such as the creation of image compositions using Maximum Value Composition (MVC) method for vegetation indices (see Holben, 1986), or averaging a set of images for Land Surface Temperature (LST).

The third goal consists in exploring some techniques to detect trend changes when analysing the natural evolution of certain indices. In particular Chapter 3, named "Detecting change-points in a predefined surface in Spain", uses NDVI data from satellite imagery providing an example of an statistical study. The objective of the study is detecting trend changes in a time series of NDVI images in Spain in 35 years. As a preliminary work using satellite imagery, and to carry out the study for this long time period, 3rd generation Global Inventory Modelling and Mapping Studies data (GIMMS NDVI3g) are used in a time series from 1981 to 2015 (see Detsch, 2019). The study reveals trend changes in all the considered methods, but the location of the change-points does not coincide from method to method.

The fourth goal is focused on the use of spatial and temporal dependence with satellite images. Chapter 4, entitled "Stochastic spatio-temporal models for analysing the NDVI distribution", presents a work dealing with the use space-time dependence of the pixels in a time series of NDVI images to estimate the distribution of NDVI in Spain. The study uses a stochastic state space model (SSM) (see Cameletti, 2012).

However, the model has been designed to use a small set of locations, and do not support the use of images. This work adapts the SSM to accept data coming from NDVI images extracting all the locations from it. The model treats the fluctuations generated by atmospheric factors, errors in the capture instant, or the changes in the time series of NDVI as random fluctuations. The result of the modelling is a set of smoothed images showing the distribution of the NDVI in time and space. These images reveal a clear seasonality in NDVI that intensifies the effect of spring vegetation, where a higher level of rainfall than average is documented.

The fifth goal is to develop new methods for filling gaps and smoothing errors in satellite images using spatial and temporal dependence. Chapter 5 and Chapter 6, entitled "Interpolation of the mean anomalies for cloud-filling" and "Using ground-truth data for improving satellite imagery" respectively, provide new methods for filling gaps and smooth outliers in time series of satellite images. The new procedure explained in Chapter 5 is called Image Mean Anomaly (IMA). The method is based on creating a temporal window around a target image to select similar images. The temporal window is averaged generating an average image. The target image can be expressed as the sum of the average image and a residual image, which is called anomaly. The methods are compared in a simulation study using already filled and processed images as a reference. Chapter 6 explains an alternative to IMA using covariates. The study shows how the use of covariates can improve the quality of satellite images. To perform the study a new method named TpsWc based on the same premise of IMA is presented. This procedure is compared to the SSM model introduced in Chapter 4. The results reveal better predictions when using covariates, and the new TpsWc gets better predictions than SSM.

Finally, this dissertation ends with some conclusions and comments on further research lines.

All the contents in this dissertation have been published in book chapters or specialized impact factor journals. Some of the new methods have been also implemented in a new R package.

- The contents of Chapter 1 have been published in the Comprehensive R Archive Network (CRAN) as an R package named 'RGISTools':
  Pérez-Goya, U., Militino, A. F., Ugarte, M. D., and Montesino-SanMartin, M. (2019). *RGISTools: Handling Multiplatform Satellite Images*. R package version 0.9.7.

- The contents of Chapter 2 have been published as a chapter of the book *Handbook of Mathematical Geosciences*:
  Militino, A. F., Ugarte, M. D., and Pérez-Goya, U. (2018c). An introduction to the spatio-temporal analysis of satellite remote sensing data for geostatisticians. In Sagar, B. D., Cheng, Q., and Agterberg, F., editors, *Handbook of Mathematical Geosciences*, chapter 13, pages 239–253. Springer.

- The contents of Chapter 3 have been published as a chapter of the book *The Mathematics of the Uncertain*:
  Militino, A. F., Ugarte, M. D., and Pérez-Goya, U. (2018b). Detecting change-points in the time series of surfaces occupied by pre-defined ndvi categories in continental spain from 1981 to 2015. In Gil, E., Gil, E., Gil, J., and Gil, M. Á., editors, *The Mathematics of the Uncertain*, chapter 28, pages 295–307. Springer.

- The contents of Chapter 4 have been published in the journal *Remote Sensing*:
  Militino, A. F., Ugarte, M. D., and Pérez-Goya, U. (2017). Stochastic spatio-temporal models for analysing ndvi distribution of gimms ndvi3g images. *Remote Sensing*, 9(1):76.

- The contents of Chapter 5 have been published in the journal *IEEE Transactions on Geoscience and Remote Sensing*:
  Militino, A. F., Ugarte, M. D., Pérez-Goya, U., and Genton, M. G. (2019). Interpolation of the mean anomalies for cloud filling in land surface temperature and normalized difference vegetation index. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):6068–6078.

- The contents of Chapter 6 have been published in the journal *Remote Sensing*:
  Militino, A., Ugarte, M., and Pérez-Goya, U. (2018a). Improving the quality of satellite imagery based on ground-truth data from rain gauge stations. *Remote Sensing*, 10(3):398.

1 Satellite imagery: data access and download

1.1 Introduction

Satellite imagery is commonly used for monitoring the Earth. It includes different sensors categorized in two main groups: active and passive sensors. Active sensors use on board energy to illuminate the objects they observe, and they capture the reflection from the target. The best known active sensor is RADAR. This sensor emits microwave radiation and captures the reflected energy. In contrast to active sensors, passive sensors detect natural energy reflected by the target objects. The most common source of radiation measured by passive sensors is the sunlight reflection, captured by spectral radiometer. This type of sensors measure the radiance at multiple regions of the electromagnetic spectrum, called bands.

All the satellites have on board at least one sensor, but usually they include more. For example, Terra satellite has on board five different sensors (see NASA, 2004). However, spectral images are a key source for getting information in agricultural applications. Spectral radiometer can capture from tens to hundreds bands depending on the type of sensor. The combination of bands produces different variables, some of them very interesting for vegetation monitoring. The most common example is the Normalized Difference Vegetation Index (NDVI), published for the first time by Tucker (1979).

This chapter analyses the satellites that currently provide open-access spectral images. The specific characteristics of each satellite are explained emphasizing the differences between raw and composite images. Nowadays, many platforms provide satellite images. Here we will be dealing with Landsat, MODIS, and Sentinel. There are many programs publishing open-access satellite images, but most of them are focused on specific areas or regions.

The platform for each satellite program is also introduced, and the pros and cons of each platform are presented. All the analysed platforms provide a graphical

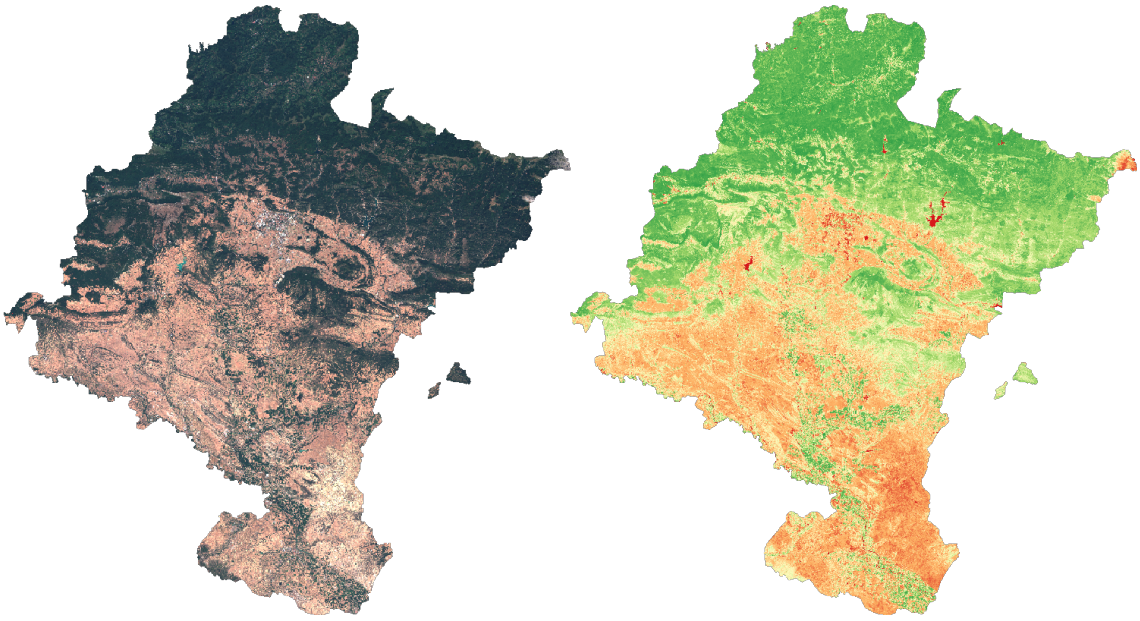

Figure 1.1: On the left RGB image and on the right NDVI image of Navarre in Spain. Both are captured by Sentinel-2A satellite on 23 of August, 2016.

user interface (GUI) as the main downloading method, showing a world map and a search box. This type of interface implements tools for searching images defining a time interval and creating a spatial polygon with the cursor on the map. The result of the search is a list of images that allows the pre-visualization of the images one by one. This search procedure is intuitive for downloading a single image, but the downloading process of a long time series of images is more difficult. However, if the user is interested in doing spatio-temporal analyses, a time series of images is required. Therefore, this chapter presents the development of a set of functions to automatize the downloading process of a time series of images.

The downloading process is only the first step. Each satellite program split the entire world in tiles, but usually, regions are the focus of the analysis. At this point, procedures for generating regions from a group of tiles are as important as the download. In this chapter, jointly with the download procedure, there is a description of the applications developed for merging and cutting the tiles, in order to set a particular region. Both procedures, the downloading and the merging have been developed as R functions and have been encapsulated in the 'RGISTools' package. The function development requires to fulfil some dependencies in R.

The rest of this chapter is organized as follows. In Section 1.2, named "Satellite images access platforms", the most known satellites are described. The section also presents the platforms used by each organization to publish the data and describes some possibilities to develop an automatic download procedure. In Section 1.3,

entitled "Using R as GIS environment" the use of R programming language for downloading and customizing the satellite images is presented, explaining the most important packages needed for this purpose. In Section 1.4, named "Development of the 'RGISTools' package", the development of an automatic procedure to get time series of satellite images with the same image format is described. The chapter finishes with some conclusions.

1.2 Satellite images access platforms

We focus on three satellite programs publishing spectral open access satellite images: MODIS, Landsat, and Sentinel. They were launched in different times, and with different objectives. Each of the three projects has two satellites in orbit at the moment. MODIS have Terra (1999) and Aqua (2001), Landsat have Landsat-7 (1999) and Landsat-8 (2013), and Sentinel have Sentinel-2A (2015) and Sentinel-2B (2017). The general information of each satellite project can be seen in Table 1.1.

MODIS satellite has been designed to provide a coarse view of the entire Earth every day. It gives one image every day and captures a great variety of bands (up to 36). As a negative point, the spatial resolution of its captures is too low if we compare it with satellites from other programs. However, Landsat and Sentinel-2 programs provide fine images. In fact, the main objective of these satellites is to provide high resolution images to allow data analysis in smaller regions. Landsat and Sentinel-2 programs have been designed to have two satellite in orbit at the same time. Each program can reduce the temporal resolution resolution by 50% combining both satellite images.

Table 1.1: Main characteristics of open access multi spectral images satellites: MODIS, Landsat, and Sentinel

Satellite	MODIS		Landsat		Sentinel	
	Terra	Aqua	Landsat-7	Landsat-8	Sentinel-2A	Sentinel-2B
Temporal resolution	daily	daily	16 days	16 days	10 days	10 days
Spatial resolution	250m	250m	30-60m	15-30m	10-60m	10-60m
Number of bands	36	36	8	11	12	12
Image size approx.	70mb	70mb	200mb	200mb	500mb	500mb
Image format	HDF-EOS	HDF-EOS	TIF	TIF	JP2	JP2
Launch date (mm/dd/yyyy)	12/18/1999	05/04/2002	04/15/1999	02/11/2013	06/23/2015	03/07/2017

Sentinel program can provide one fine image every five days using Sentinel-2A and Sentinel-2B images, only since April of 2017, when Sentinel-2B was launched. Even having the images of both satellites, the main problem of the high resolution images are the clouds. If a cloud appears in the image, the temporal gap between

the images increases to ten days. With two or more images including clouds, the time series may become unusable for the analysis.

The organizations publishing MODIS, Landsat, and Sentinel images are different, and each project has its own download platform. Landsat has a web platform called Earth Explorer (USGS, 2012), MODIS has an application programming interface (API) named Nasa inventory (USGS, 2012), and Sentinel has an SciHub platform (USGS, 2012).

1.2.1 Earth explorer

Earth Explorer is a satellite image platform designed for centralizing different open-access products. The platform is managed by United Stage Geological Survey (USGS) that publishes images from different satellites as Sentinel-2, even if they are not part of the same organization. The platform needs a user account that can be created for free. It is an easy platform to download a couple of images because it provides a Graphic User Interface (GUI), see Figure 1.2. The GUI allows definition of a spatial location and a date range to perform the search. The main problem of this platform is that it was designed for downloading images one by one and it is not easy to download time series of images. If the user wants to download time series of satellite images, all the images need to be downloaded separately. Downloading Landsat-8 images one-by-one can be a tedious work, but it may be possible because Landsat-8 only provides two images per month.

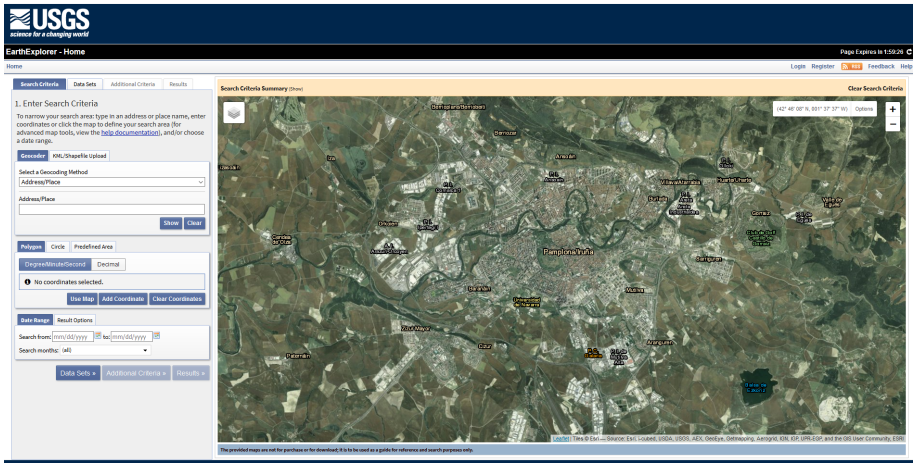

Figure 1.2: City of Pamplona (Spain) using the Earth Explorer platform search screen.

1.2.2 NASA inventory

Downloading MODIS images has an alternative called Nasa Inventory (NASA, 2016). The platform integrates an Application Programming Interface (API) instead of a Graphic User Interface (GUI). The API implements a standard protocol for communication defining the requests and the responses. It defines methods of communication among National Aeronautics and Space Administration (NASA) Inventory Service and an application in the user computer. This publishing platform provides an easy way to search and download a time-series of MODIS satellite images.

The communication with this API is done using a usual web protocol HTTPS. The platform is not optimized for manual searches, because all the attributes for the search need to be written in the url. Some of the search attributes of the platform are the product name, the spatial location or the date range.

The search is done defining the search arguments as "*product=MOD09GA*" to define MODIS product, *latitude/longitude* to define the spatial location and *date=2019-01-01/2019-12-31* to define the date ranges. This is an example of searching the product "*MOD09GA*", in the city of Pamplona in the whole 2019 using Nasa Invetory:

> https://lpdaacsvc.cr.usgs.gov/services/inventory?product=MOD09GA&version=6&latitude=42.81&longitude=-1.65&date=2019-01-01/2019-12-31

The response to the query given by the platform is an XML file with a list of images, that matches with the search. This list provides directly all the url for downloading the images. It is a simple platform but simplifies the search and download process, and it accepts implementations for automatic downloads in any programming language that implements web communication technology.

1.2.3 Earth data

Since May 2014, and with the objective of centralizing all the information coming from satellites, NASA starts developing a new remote sensing download platform. This platform is called Earth Data (see NASA, 2014). The platform not only manages satellite images, but it has data sets as rainfall data or soil sampler. Nowadays, the download of any image coming from United States satellite requires an Earth Data user account, even if the search is done with other platform. Platforms as Nasa Inventory permit the search of images but for downloading, an Earth Data account is required. Earth Data is a new and modern platform, and it integrates both, a GUI and an API. In the GUI you can perform searches specifying a spatial location and a product. The downloads are done selecting a set of images and sending a request. After some processing time, the download is done from a user panel. In the API, as with the Nasa Inventory service, the downloads can be automatized developing an

application to communicate with Earth Data. The main disadvantage of Earth Data is that the API communication has the same constraint as the GUI. The automatic application firstly needs to search and selects the images, and after the selection, some extra time is needed for images downloading. In other platforms such as Earth Explorer or Nasa Inventory service there is no need to wait.

1.2.4 Scihub

The Sentinel program uses another search and download platform called SciHub (see ESA, 2016). It is a platform that is still on development, because with each new satellite the support for the new images need to be implemented. Sentinel is an European Space Agency (ESA) program, and for downloading images from this platform, a user account is needed. This platform is similar, in terms of technology, to Earth Data. The platforms have two parts, one for a GUI and the second one for an API. The GUI, is a simple map-based GUI where you can select with the mouse some spatial locations. In addition to spatial locations, the search allows to define a period range and the satellite. The API implements two communication standards: OpenSearch API and OpenData API, giving support for many programming languages to create automatics searches and downloads. However, as with Earth Data, its main searching tool is an interactive map.

Platform comparison

Table 1.2 shows the comparison of different download platforms. Earth Data, Nasa Inventory, and SciHub. The three platforms implement a GUI, but in all the cases the download of images using GUI is a hard work if the user needs to download images in a wide temporal range. The searches support time ranges and spatial location, but the downloading needs to be done by the supervision of a person. No GUI is prepared for time series downloading. That is why it is necessary to implement an automatic application for downloading time-series of satellite images. The best way to automatize the download procedure is using some of the defined API, one for each satellite program. The main difficulty is the download of Landsat images because these images only are published in Earth Explorer and Earth Data. Earth Explorer is an only-GUI platform, and Earth Data has an API but it has the constraint of the waiting time after the time-series selection. To automatize the downloading for Landsat it is mandatory to use Earth Explorer or Earth Data platforms.

The case of MODIS and Sentinel are different. MODIS has Nasa Inventory, a service dedicated for time-series searching and downloading. Sentinel has SciHub, a platform dedicated for downloading entire time series of satellite images with an API.

Table 1.2: Characteristics of publishing platforms for satellite images

	Earth Explorer	Earth Data	Nasa Inventory	SciHub
GUI	yes	yes	no	yes
API	no	yes	yes	yes
Direct Download	man	man	auto	auto
Search response	HTML	XML	XML	JSON
Download Account	Earth Data	Earth Data	Earth Data	SciHub
MODIS Images	yes	yes	yes	no
Landsat Images	yes	yes	no	no
Sentinel-2 Images	yes	no	no	yes

1.3 Using R as GIS environment

There are many geographic information system (GIS) tools designed for visualizing or processing spatial data. For example, ArcGIS, QGIS, or GRASS are GIS widely used for local computing, while Google Earth Engine is most used for cloud computing. There are advantages and disadvantages using some of these tools, but all platforms are limited to download series of images. The availability of statistical procedures is another limitation in these type of platforms. ArGIS or QGIS provide tools for sending data to R, but there are only designed for using a single image band or layer, and they can not used time series of images.

R, in addition to the core package, has native access to CRAN repository, and to its more than 15.000 packages. Since 2000s, the R developers community has built a wide variety of packages to use R with geographic data. The main advantage of the package developed in this dissertation is its ability to manage time series of images and to work with the whole data.

1.3.1 Libraries for downloading images

The automatic procedure for downloading satellite images can be implemented in any programming language that supports web communications. R supports natively networks communications with multiple package implemented in CRAN. The 'curl' package has been used to develop this work (Ooms, 2019). Curl or client URL is a library for downloading files performing requests on the Web. It simulates a web browser and it supports all the new web technologies. Using this package some functions have been developed for downloading images. The search response for each publishing platform is different (see Table 1.2). Earth Explorer provides HTML file format, Earth Data (ED) and Nasa Inventory (NI), XML data format and SciHub, JSON data format. To perform automatic download, R needs to read correctly these

file formats. CRAN provides libraries to read all these formats. 'RGISTools' uses 'rjson' and 'XML' packages for reading the API responses. In addition, 'urltools' has been used for managing API URLs. Here there are some details of the main packages used in 'RGISTools' downloading methods:

'curl' (Ooms, 2019): implements functions to use the last version of *libcurl*, that is a system library to perform web requests. It has full support for secure connections (HTTPS) and, redirection, and its able to have an active user session in connections.

'rjson' (Couture-Beil, 2018): JSON is an advanced and structured data format standard designed for web communications. It is defined to sent data in web requests and to receive and read the data of the response. As it is an structured data format, a complex set of data can be sent using JSON. In R, there are more than one package designed for reading and creating JSON files, but 'rjson' has been used because it is very complete..

'XML' (Lang and the CRAN Team, 2019) : XML is an structured data format mainly used in web communications. Many APIs or web pages used XML to publish data. In R, the most common packages to manage XML are *XML* and *XML2*. *XML* used for downloading because of its simplicity for reading simple data, similar to the used platform.

'urltools' (Keyes et al., 2019): the implementation of an automatic downloading procedure needs some tools to check and manage different requested urls. This package provides some functions to simplify these url operations. For example, to get the root of an url or to add attributes for creating a new url.

1.3.2 Libraries for loading GIS images

'raster'

'raster' (Hijmans, 2019) is the R package that loads images using the coordinate system of the image. The package has been optimized to allow hard drive processing. This means that R can process images directly from the hard drive, without loading them to the Random Access Memory (RAM). The process of loading an image into the R environment can be done with only three functions: *raster*, *stack*, and *brick*.

raster: it loads one image with only one layer. By default, the image is referenced in the hard drive and the processing may run in the hard drive.

stack: it loads a group of images with one or more layers with the same spatial extent and resolution. By default the image is referenced in the hard drive .

brick: it is similar to stack, but it loads additional data into the RAM. It is faster when processing data but may produce out of memory problems using high resolution images or large amount of images.

Satellite images usually have huge amounts of data. The hard drive processing is an important characteristic, otherwise it may be very difficult to load the whole images into the RAM. The 'raster' package is able to run in the hard drive some basic GIS functions like *crop*, *merge*, or *projectRaster*.

projectRaster: it is an important function because operations between rasters can only be run if the images have the same coordinate systems. *projectRaster* is able to change the coordinate system of an image and to transform the data to a new projection using this function.

crop: this function is able to cut an image. It can be used for selecting the region of interest of an image.

merge/mosaic: there are two functions used for merging a set of images: *merge* only accepts images that do not overlap, while *mosaic* may merge overlapping images. Any set of images you want to merge needs to have the same coordinate system and image resolution.

These operations are really necessary because R only accepts operations with images that have the same number of rows and columns. Before any analysis of time series of satellite images it is necessary to standardise the time series to the same coordinate reference and matrix dimension.

'rgdal' and 'gdalUtils' packages

The package 'raster' is a powerful tool, but is not able to load all satellite imagery data formats. For example, HDF format of MODIS cannot be loaded using the package 'raster'. In such cases, the library that includes almost all image file format is GDAL. (Geospatial Data Abstraction Library). GDAL is an open source library for raster and vector geospatial data formats. Almost all modern GIS environments use GDAL library in any part of the environment and it has been ported to lots of programming languages including R. GDAL Supports 155 raster formats and 95 vectorial formats. There are two packages to use GDAL in R, 'rgdal', that is a native implementation of GDAL and 'gdalUtils', that is an interface between R and the package in its C programming language version.

'rgdal' (Bivand et al., 2019): *rgdal* is a native implementation of GDAL for loading and saving satellite images. The package is focused on loading satellite images and creating raster or vectorial object in R. It does not support as many

formats as GDAL native C implementation, for example, HDF-EOS format used by MODIS satellites is not supported by 'rgdal'. Another example is jp2 that was not supported until February 2018.

'gdalUtils' (Greenberg and Mattiuzzi, 2018): this package works as an interface between R and the C implementation of GDAL. To use the library, the GDAL binary needs to be installed in the computer. The function in R calls GDAL system instruction. When a function of 'gdalUtils' is ran, the data in the image is not loaded to the R environment, but the function can cut, merge, or change the format directly in the hard drive.

1.4 Development of the 'RGISTools' package

The automatic procedure for downloading satellite images has been done developing some functions in R. This Section presents the downloading and customizing process for Landsat-7, Landsat-8, Sentinel-2, and MODIS satellite images. In Section 1.2 the platform used by each satellite program is presented, defining the three platforms for downloading satellite images. Multiple platforms need the development of multiple function, that is, each satellite program needs its specific download function. These functions compose the core of the package 'RGISTools', defined as a set of functions for searching, downloading, merging tiles, and cropping a region of interest. Due to the wide variety of procedures and sources of information being handled in 'RGISTools', the functions are divided into seven categories, which are identified by the first three characters of the function names;

1. **ls7** identifies Landsat 7 functions.
2. **ls8** identifies Landsat 8 functions.
3. **ls** identifies both Landsat 7 and 8 functions.
4. **mod** identifies MODIS Terra and Aqua satellite functions.
5. **sen** identifies Sentinel functions.
6. **gen** identifies function for being used in any of the three platforms.
7. **var** identifies function for deriving variables in any of the three platforms.

1.4.1 The downloading process

The downloading process needs a previous search step. The search step locates the images matching within a pre-defined time interval and a spatial location. Both processes save the images in the hard drive for future processing. Each of the satellite program implements both actions as functions.

Application for downloading Landsat images

Landsat images are published into two platforms: Earth Explorer (EE) and Earth Data (ED). Earth Explorer is not prepared for automatic downloading because it does not have an API for this purpose. Earth Data, the second alternative, has an API, but the user needs to perform a request and wait for a while before downloading the images. For automatic downloads the Earth Data platform is preferred.

The first step for downloading Landsat images is to get the names of the images to download. EE has not an API for the searching procedure. However, the Landsat program publishes the information in a csv file where all captures done by its satellites are registered with additional meta data. The file is renewed every day with the entries of the new images. Landsat publishes one csv for Landsat-7 and another one for Landsat-8.

The way to automatize the downloading of Landsat images is to download the csv file with the meta data of all captures, and to create a search function using the data from the csv file. Once the search function gets a list with the names of the images for downloading, it is possible to perform all the downloading automatically from the Earth Explorer platform. The procedure has been implemented with three different actions:

Downloading meta data file: for downloading the meta data file 'RGISTools' provides two functions *ls8LoadMetadata* for Landsat-8 and *ls7LoadMetadata* for Landsat-7. These functions download the meta data file to a directory defined in the *AppRoot* argument. If the file does not exists or is obsolete, the functions automatically download the new meta data file and transform the csv to an R data frame. The function is implemented in the search function to make the procedure easier.

Searching in meta data: the spatial search in a time range for Landsat images implements two functions *ls7Search* and *ls8Search*. The functions read the Landsat meta data file and identify images that matched with the temporal range and the spatial location. For the temporal location only a data filter is needed because one of the meta data columns contains the capturing date. For the spatial search in the meta data there are four columns that contain the boundaries position in latitude/longitude. Using the boundaries a spatial

search can be done. The function returns the filtered data frame with matching images and all its meta data.

Downloading time series of images: the download function called *lsDownload* for Landsat images is designed to take the resulting data frame from *ls7Search* and *ls8Search* functions and select the SceneID column. For each SceneID and with the static web address, a download url is created. This url will download the images from the Earth Explorer platform, so a USGS account is needed. The download procedure is done starting a user session using the 'curl' package, and performing all the url for downloading the images.

The package for downloading Landsat images provides these functions: *ls7LoadMetadata*, *ls7Search*, *ls7DownSearch*, *ls8LoadMetadata*, *ls8Search*, *ls8DownSearch*. The package also provides an additional function implementing the three actions needed for downloading a set of images, making the downloading procedure more automatic. The function is *lsDownload*, giving a total of seven functions only for downloading Landsat images.

Application for downloading MODIS images

The downloading of MODIS images has been developed using Nasa Inventory. The platform implements a simple API where all MODIS lands products can be searched. The response of the search is a list of web addresses where the images can be downloaded directly. There is no need for downloading any meta data data file for MODIS images, and the downloading only needs two functions, one for image searching, and the other for image downloading. Nasa inventory provides some arguments for image searching. These are the more important arguments for performing the search query in Nasa Inventory:

product: All MODIS published images are products. The products can be raw images or pre-processed images. In the search the product needs to be specified (i.e. MOD09A1, MOD13Q1, MYD11A2, ...).

version: MODIS pre-processes all images before publishing, and the algorithm for pre-processing changes once in a while. The last algorithm is version 006, but previous versions can be downloaded as well.

bbox: the spatial location is defined with a bounding box with the extension of the spatial region in latitude/longitude coordinates.

date: the temporal range can be defined using this attribute. The user needs to specify a temporal range with two date (starting and ending date).

The response of the Nasa inventory is an XML format file with the list of the url for downloading its image. The search function reads all the data and it creates an R vector with the response. The function for MODIS image search in 'RGISTools' is *modSearch*.

Nasa Inventory does not need any user account to perform the search but it requires an Earth Data user account for image downloading. As it happens with Landsat functions, MODIS needs another function for downloading. In this case its name is *modDownSearch*. The name of the function is the same of Landsat (*ls8DownSearch*), but changing the first three characters. An additional function to run the search and the downloading actions has been developed, called *modDownload*. The use of the same name for the same actions to all satellites gives homogeneity to the package standardizing the procedures for different sources.

MODIS provides the images in HDF format. This format is not supported by 'raster' package. The format compresses all the layers and additional data from the capture instant in a unique file. Therefore, MODIS images need an addition function to extract the data from the images. This function is called *modExtractHDF*. The extraction function, is also included in the MODIS automatic download function named *modDownload*. In total, the package includes four functions for downloading MODIS images.

Application for downloading Sentinel images

Sentinel-2 is the newest satellite program publishing open-access images from the entire Earth surface. The platform is implemented in the newest API technology supporting OpenSearch and OpenData query types for searching in the Sentinel data base. The functions created for searches use OpenSearch standard for image searching. OpenSearch is similar to Nasa Inventory, but using new open data standard for APIs. SciHub API implements more attributes than Nasa Inventory to perform the searches because it is an API designed for all Sentinel images (Sentinel-1, Sentinel-3, ...). Here there are the most important attributes implemented in Scihub:

platformname: Sentinel is a huge project with five satellite programs. This attribute refers to the Sentinel program. It can be Sentinel-1, Sentinel-2, etc.

product: Sentinel-2 provides different levels of processing or different products. The options for Sentinel-2 are SLC, OLCI, SLSTR, etc.

beginposition: temporal range specifying both, the starting and the ending dates.

intersects: the spatial location is defined by an spatial polygon defining the target area. It does not have to be a bounding box, it can be a polygon or a region.

The function developed in R for searching the images is similar to MODIS functions. Only two actions are needed for image downloading, the search function (*senSearch*) and the download function (*senDownSearch*). Unlike Nasa Inventory, SciHub needs a user account for both the search and the download processes.

senSearch uses 'curl' package, and appends SciHub user account information to perform the search query. The response of the SciHub platform is parsed using 'rjson' package. This function transforms json data format to R data frame from where the list of images and its meta data can be extracted.

senDownSearch takes the response of the *senSearch* and downloads all the images in the response. The set of Sentinel functions also implements a function for searching and downloading in a unique function, the function is *senDownload*, giving a total of three functions for downloading Sentinel images.

1.4.2 Customizing satellite imagery data

The customizing procedure standardizes the downloaded data to get a time series of images with the same image name, the same image format, the same dimensions, and same date format. One customizing function has been developed for each satellite program: *lsMosaic* for Landsat, *modMosaic* for MODIS, and *senMosaic* for Sentinel-2.

The first difficulty for merging the images is the image file format. Each satellite platform uses its own file format (See Table 1.1). MODIS satellites use HDF-EOS, Landsat uses TIF, and Sentinel-2 uses JP2 or JPEG2000. Depending on the need of each satellite, the program selects the format that fulfil its requirements. For example, not all Sentinel satellites use the same image format, Sentinel-2 images are huge and the images are published in JPEG2000 because it allows compression to manage large amount of data. For example Sentinel-1 satellites use TIF format. The common characteristics of all formats is that they are Geographic Information System (GIS) formats. When an image is loaded in R using any of the R packages for this purpose both, data in the image and the coordinate system are loaded.

The variety of image formats makes it difficult merging the data in the analysis. To avoid this problem the customizing functions load the satellite images in its native format and generate a TIF image as output.

Application for standardizing data automatically

The load of time series of images in R needs the same format, dimension, and projection in all the images in the time series. The definition of a target region may imply that more than one tile is needed. The region can be a entire country or a region in a country. Each satellite program defines its own tiles dividing the regions or countries in more than one tile. If this happens, it is necessary to merge all the

tiles for each period in the time series. Another possibility is that the target region appears only in a part of the image, and in this case it is recommended to extract the region to save computer resources.

The procedure to standardize the data may be intensive in terms of computing, and it needs lots of time to process all the data. For each image format, specific standardization procedures have been developed. All of the procedures have the same output format but they need to be adjusted to correctly define each satellite projection. The standardization automatic procedure has been developed in three main actions: assignation of the coordinate reference system, crop the region from images, and merge chunks if necessary. In the following we describe the three actions:

1. **Change the coordinate system and image format**: when an image is downloaded the first action of the automatic procedure is extracting all the layers or bands from the tile. Then, the procedure changes the projection to one defined by the user, usually to UTM to get the image without deformation. Finally, the transformed image is saved in a user defined format. This procedure is different for each satellite because MODIS or Sentinel images cannot be loaded with 'rgdal' package and need to be transformed with 'gdalUtils' package.

2. **Crop the region**: the automatic download procedure only download images of the target region. This means that at least a fragment of the region will appear in all downloaded images.

3. **Merge chunks**: the standardized procedure automatically reads the date of all images and merges only the images with the same date, creating an image of a region for each period. If there is only a single image in a day period the merging process is omitted.

The package provides a single function per satellite program for customizing the images. It looks simple, but these functions make easier the management of large amount of data. The functions: *lsMosaic*, *modMosaic*, and *senMosaic*, automatically reads all the downloaded tiles from a folder, get the dates of all the captures, and group the tiles by day. Finally, these functions merge the tiles corresponding to the same day in the region of interest. This procedure makes easier the management of satellite images to obtain time series of images.

1.5 Deriving variables from spectral images

The satellites capturing spectral images can provide more than one hundred spectral bands. One spectral band only explains a fraction of the surface, but the combinations of them can derive interesting variables, called derived variables.

The 'RGISTools' package enables the calculation of the most common variables derived from spectral bands. All the functions for calculating variables start with the prefix 'var'. The package provides 9 variables functions, to measure the vegetation, the water presence or burned areas. The input of *var* functions are the spectral bands in *raster* class. It returns the computed variable in a new *raster* class.

In addition to the variable functions, the package provides one function for each satellite to automatically select the bands of each layer and compute an entire time series of variables. The most common derived variables are the Normalized Difference Vegetation Index (NDVI), the Normalized Difference Water Index (NDWI), the Normalized Burn Ratio (NBR), the Modified Soil-adjusted Vegetation Index (MSAVI), the Normalized Difference Moisture Index (NDMI) or Enhanced Vegetation Index (EVI).

1.5.1 Normalized difference vegetation index (NDVI)

From all the derived variables, one of the more common one is the the Normalized Difference Vegetation Index (NDVI). NDVI is an important index that reflects vegetation growth and it is closely related to the amount of photosynthetically absorbed active radiation as indicated by Slayback et al. (2003) and Tucker et al. (2005). Called for the first time as Transformed Vegetation Index by Rouse Jr et al. (1974), this variable can be calculated by a simple combination of two bands, the band that corresponds the red color (Red), and the band that corresponds to near infra-red (NIR). Using the the red (Red) and the near infra-red (NIR) bands the NDVI is obtained calculating the following formula:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1.1}$$

As mentioned in Sobrino and Julien (2011), this parameter is sensitive to the blueness of the observed area, which is closely related to the presence of vegetation. Although numerical limits of NDVI can vary for the vegetation classification, it is widely accepted that negative NDVI values correspond to water or snow. NDVI values close to zero could correspond to bare soils, yet these soils can show a high variability. Values between 0.2 and 0.5 (approximately) to sparse vegetation, and values between 0.6 and 1.0 conform to dense vegetation such as that found in temperate and tropical forests or crops at their peak growth stage. Therefore, NDVI provides a very valuable instrument for monitoring crops, vegetation, and forestry, and it is directly calculated in specific images by the aforementioned satellites missions. On the left of Fig. 1.3 a Sentinel NDVI satellite image of Funes, a village of Navarra (Spain) is shown, and on the right of the same Figure, the NDVI for the whole region of Navarra is given.

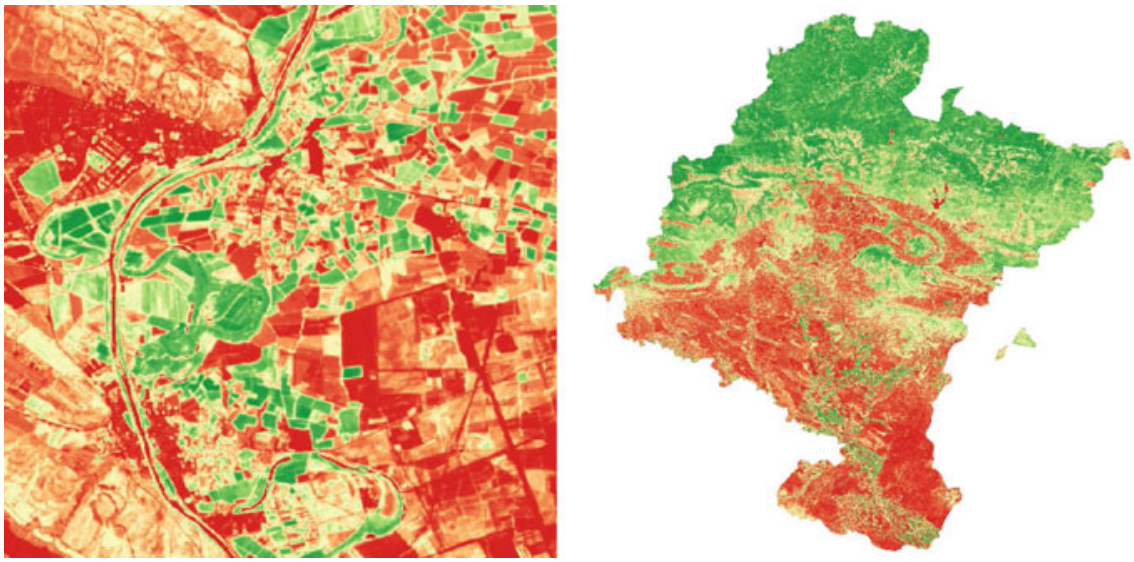

Figure 1.3: NDVI image captured by Sentinel-2A. (Left) Zoom of Funes village in Navarra, and (Right) the NDVI for the whole Navarra (Spain)

1.5.2 Enhanced vegetation index (EVI)

The Enhanced Vegetation Index (EVI) is a vegetation indicator that improves sensitivity towards high biomass densities compared to NDVI. EVI accomplishes a higher sensitivity to the green biomass through a de-coupling of the background signal. EVI uses the Near Infra-Red (NIR), Red and Blue atmospheric corrected surface reflectance, plus a background correction factor (L). The blue band allows correcting for aerosol disturbances in the red band.

$$EVI = G \times \frac{(NIR - RED)}{(NIR + C1 \times RED - C2 \times Blue + L)} \tag{1.2}$$

1.5.3 Normalized difference water index (NDWI)

The Normalized Difference Water Index (NDWI) is a ratio between bands of the spectrum that was developed to detect open water areas minimizing the influence of the soil and vegetation variations. The ratio involves the Green and the Near Infra-Red (NIR) reflectances. Values of water bodies are generally larger than 0.5.

$$NDWI = G \times \frac{(Green - NIR)}{(Green + NIR)} \tag{1.3}$$

1.5.4 Normalized burn ratio (NBR)

The Normalized Burn Ratio (NBR) is an index to identify burned areas by comparing its value before and after the fire event ($\triangle NRB$). NBR uses the reflectance of Near Infra-Red (NIR) and the Shortwave Infra-Red (SWIR) parts of the spectrum. The $\triangle NRB$ of unburned areas oscillate between -0.1 and +0.1. Values of $\triangle NRB$ higher than 0.66 indicates a high level severity of burn. These interpretations of the $\triangle NRB$ should be taken with care. Results should be evaluated against ground truth data.

$$NBR = \frac{(NIR - SWIR)}{(NIR + SWIR)} \tag{1.4}$$

$$\triangle NRB = NBR_{prefire} - NBR_{postfire} \tag{1.5}$$

1.5.5 Normalized burn ratio 2 (NBR2)

The Normalized Burn Ratio 2 (NRB2) is an index to identify burned areas. In contrast to NBR, the NBR2 replaces the Near Infra-Red band by the Shortwave Infra-Red at 1566-1651 nanometres (SWIR1) to highlight the sensitivity to water in vegetation. The NRB also uses the Shortwave Infra-Red at 20107-2294 nanometres (SWIR2)

$$NBR2 = \frac{(SWIR1 - SWIR2)}{(SWIR1 + SWIR2)} \tag{1.6}$$

1.5.6 Modified soil-adjusted vegetation index (MSAVI2)

The Modified Soil Adjusted Vegetation Index 2 (MSAVI2) is a vegetation indicator that removes the effect of background variations. It is an improvement of Modified Soil-adjusted Vegetation Index (MSAVI) that represents vegetation greenness with values ranging from -1 to +1. MSAVI2 prevents from explicitly specifying the canopy background correction factor (L). The MSAVI2 uses the reflectances of the Near Infrared (NIR) and Red parts of the spectrum.

$$MSAVI2 = \frac{2NIR + 1 - \sqrt{(2NIR + 1)^2 - 8(NIR - Red))}}{2} \tag{1.7}$$

1.5.7 Normalized difference moisture index (NDMI)

The Normalized Difference Moisture Index (NDMI) is an index that represents the water stress levels of the canopy. NDMI uses the Near Infra-Red (NIR) and Shortwave Infra-Red (SWIR) reflectances of the spectrum. NDMI oscillates between -1 and +1:

Values closer to -1 represent bare soil, values around 0 represent canopy under water stress, and values closer to 1 represents vegetation with no water stress.

$$NDMI = \frac{NIR - SWIR}{NIR + SWIR} \tag{1.8}$$

1.5.8 Land surface temperature (LST)

Another important variable derived from satellite images is the land surface temperature (LST). The precision of the wavelength from LST derives changes from satellite to satellite. This is why LST variable can be obtain with different algorithmic procedures. As an example, Sobrino et al. (2004) compare three methods to retrieve the LST from thermal infrared data supplied by band 6 of the Thematic Mapper (TM) sensor onboard the Landsat 5 satellite. The first is based on the radiative transfer equation using in situ radiosounding data. The others are the mono-window algorithm developed by Qin et al. (2001) and the single-channel algorithm developed by Jiménez-Muñoz and Sobrino (2003). Many satellites platforms provide specific images of LST all over the Earth, because it is also a very outstanding variable for many environmental process. Figure 1.4 shows the daily land surface temperature in Navarra (Spain) the 13th of July 2015 from TERRA satellite

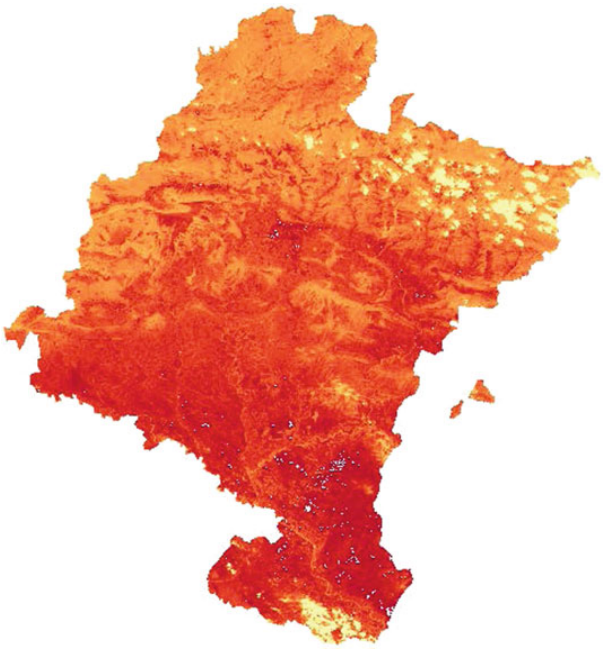

Figure 1.4: Land Surface Temperature of Navarra the 13th of July 2015. Image captured by Landsat-8 satellite.

1.6 Additional functions in 'RGISTools'

The package, in addition to downloading and customizing functions, provides additional functions to facilitate working with satellite images in R. Functions for previewing the images before the download, creating cloud masks, deriving variables, and smoothing/gap filling are also provided. Here a small description of these functions is presented:

- **Preview**: the user can preview the images using *senPreview*, *modPreview* or *lsPreview* functions previous to downloading them.
- **Deriving variables**: the package provides functions for automatizing the procedure of deriving variables from spectral images. The function reads all the layers in a folder, and computes the derived variables. Some derived variables are explained in Chapter 2.
- **Cloud masking**: as it is difficult to found a satellite image without presence of clouds, the package provides functions for creating masks that removes the pixels where a cloud is located.
- **Smoothing data**: functions for smoothing and gap-filling data are also provided. In particular, the two procedures explained in Chapters 5 and 6. The functions are *genSmoothingIMA* and *genSmoothingCovIMA*, respectively.
- **Getting meta data**: functions for getting the date from the name convention used by each satellite are also provided (*lsGetDates*, *senGetDates*, and *modGetDates*). In addition, the relation between the satellite band ("B01", "B02", "B03", ...) and the name of the band ("red", "blue", "green", ...) is given in the package making easier band identification.
- **GIS plot**: the images in raster format can be plotted using multiple functions. However, the package provides a preprogrammed function for plotting the images with a proper GIS format. The function is *genPlotGIS*.

1.7 Conclusions

The spatio temporal analysis of satellite images needs to download a huge among of data. The procedure for downloading complete time series of satellite images can be automatic, but as far as we know, no such procedure exits yet. Depending on the requirements of each satellite, the image format, projection o size may change. For handling images from multiple satellites, a standardization of all images is needed. Nowadays, there are applications for downloading and standardising a single image, but the process for creating time series is still time consuming.

In this chapter we analyse and provide solutions to both, the downloading procedure and the standardization. The variety of platforms and formats requires an automatic process for downloading the entire time series of images. We use the R programming language to develop automatic downloading and standardizing procedures dealing with images from open access multi-spectral satellites. All the procedures have been developed as functions and encapsulated in a R package called 'RGISTools'[1]. The package organizes the functions and the procedures to be similar between the different satellites. These procedures provide tools for downloading, standardizing, and managing all the downloads, reducing the time needed to generate ready-to-use time series of satellite images. In particular, the package provides 60 functions including the downloading, customizing, variable extraction, cloud removal, and tools for data managing in a 1.5MB package. All the function are explained in a manual with 81 pages, that it is available in the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/RGISTools/RGISTools.pdf).

1 Pérez-Goya, U., Militino, A. F., Ugarte, M. D., and Montesino-SanMartin, M. (2019). *RGISTools: Handling Multiplatform Satellite Images*. R package version 0.9.7.

2 The use of geostatistics in the analysis of satellite imagery

2.1 Introduction

The satellite programs publish variables in different images processing levels, closely related with different smoothing levels. The first processing levels are almost raw images, and the higher levels are smoothed images. However, all the levels may have gaps or distorted values at some point. There exist many procedures and techniques to fill and correct the images by combining several bands of the satellite images. The aim of this chapter is to describe the use of geostatistical tools in the analysis of satellite imagery. The chapter provides an overview of the most well known methods in the literature, and analyses R packages that can be used for satellite image processing. The procedure for obtaining these variables is explained taking as a reference the images obtained by the applications developed in Chapter 1. This chapter describes the solutions proposed in the literature to reduce the effect of atmospheric factors, for example the presence of clouds or distortions made by electronic components. Some alternatives for smoothing these errors are the creation of image compositions using Maximum Value Composition (MVC) method for vegetation indexes, or averaging a set of images for Land Surface Temperature (LST).

The chapter also provides a review of some applications (Sagar and Serra, 2010). In particular, we explain pre-processing, analysing, interpolating, smoothing, and modelling these data. The chapter encloses five additional sections where a short explanation of the state of the art in the analysis of remote sensing data using free statistical software is given. Particular attention is devoted to the use of geostatistical tools in this subject. The rest of the chapter, is organized as follows Section 1.5, entitled "Deriving variables from spectral images", explains how different variables

can be derived by the combination of bands. Section 2.2, named "Preprocessing variables", explains the procedures needed for correcting some errors and the most common methods of pre-processing and smoothing the satellite imagery data before its use in statistical analysis. Section 2.3, entitled "Spatial interpolation", introduces the importance of the spatial interpolation in remote sensing data for removing errors and gaps. The section finishes reviewing the most popular interpolation methods. Section 2.4, named "Spatio-temporal interpolation", presents the advantage of using both, the spatial and the temporal correlation for improving the interpolation procedure. In addition, we comment the use of R and some R packages for applying spatial and spatio-temporal geostatistics techniques with satellite images. The chapter ends up with some conclusions in Section 2.6.

2.2 Preprocessing variables

The effect of the atmosphere can distort, blur or degrade the images. Even in a sunny day the atmosphere and the sun radiation affect the images. The position of the sun in the sky at the capture instant produce variations in radiation levels affecting the images. The satellites record the position of the sun for correcting the differences in the radiation. The atmospheric correction is a method that try to remove influence of a portion of light reflected by the atmosphere on the image in order to preserve the part reflected off the surface below. Using specific procedures for each satellite, these effects can be corrected getting bottom of atmosphere (BOA) or surface radiance. The correction of atmospheric radiance is always necessary but when other type of atmosphere distortion such as clouds, fog, flash appear other strategies need to be used. The usual correction for smoothing these higher distortion consists in composing several images into a new single one. Different algorithms have been used in the literature according to the derived variable. For vegetation variables such as NDVI, the maximum value composite (MVC) procedure (Holben, 1986) is used. MVC assigns the maximum value of the time-series of pixels across the composite period. The procedure is based on the assumption that lower values in the time-series correspond to an error produced by atmospheric factors. The main problem with MVC is that in regions with sharp changes as farming fields the procedure does not work properly. Alternative techniques include using a bidirectional reflectance distribution function (BRDF-C) to select observations and the constraint view angle maximum value composite (CV-MVC) (NASA, 2018). These two algorithms analyse all pixels in the time-series to select a better value than the maximum but if they are not, the maximum value is selected. For LST day/night there is no sense to get the maximum value across a time-series because the maximum temperature may be overestimated. It is more common to average the cloud-free pixels over the compositing period (Vancutsem et al., 2010). The presence of a clouds in a LST

image can drop the temperature to very low values, becoming mandatory cloud removal. In the preprocessing time of all open access satellite images a cloud analysis is done. The result of this analysis is band with the position of the cloud. Using these bands the clouds can be removed from the images to enable the averaging process. Nowadays, many composite images can be directly downloaded with different spatial and temporal resolutions. For example, already cloud-free daily LST images from all over the world of Aqua or Terra satellites can be downloaded, but the use of these images is always scarce because the effect of clouds use to be high. So, usually composite images of at least weekly are used but also bi-weekly temporal resolution.

Spatial and temporal resolutions are also different from the same or different satellites as it can be seen in Table 1.1 (Chapter 1). Two big satellite groups can be distinguished, the satellites that have high temporal resolution and the satellite with low temporal resolution. For example, Terra and Aqua are high temporal resolution satellites. They take an image of the entire Earth every day. These type of data can be useful when tracking seasonal changes in vegetation on continental and global scales. But the main disadvantage of these type of images is the analysis of small regions, where you can get only a small set of pixels. In these type of analysis higher resolution images are needed, but they use to have lower temporal resolution, getting only one image per week or less. The small number of high resolution images joining with the clouds can complicate the analysis of small regions. Daily time-series enable the use of composition images. But when only having one image every two weeks the time distance between the images may involve great season differences making it the composition impossible. At this step, numerical, physical or mechanical analyses and procedures may reconstruct the image. Sometimes, the highest presence of clouds determine the drop out of some images in the time-series, but if they are only partially clouded, different approaches for removing these effects can be used. Noise reduction in image time series is neither simple nor straightforward. Many spatial, temporal or spatio-temporal alternatives have been provided. For example R.HANTS macro of GRASS, SPIRITS, BISE, TIMESAT, Gapfill, or the CACAO methods are very well spread. R.HANTS performs an harmonic analysis of time series in order to estimate missing values and identify outliers (Roerink et al., 2000). SPIRITS is a software that processes time series of images (Eerens et al., 2014). It was developed by PROBA-V data provider and gives four smoothing options, including MEAN (Interpolate missing values & apply Running Mean Filter RMF) and BISE (Best Index Slope Extraction), (Viovy et al., 1992). TIMESAT uses numerical procedures based on Fourier analysis, Gauss, double logistic or Savitzky-Golay filters (Jönsson and Eklundh, 2004). Gapfill uses quantile regression to produce smoothed images where the effect of the clouds has been reduced. Usually, every software has different requirements with regard to the number of images necessary for smoothing (Atkinson et al., 2012). Finally, CACAO software (Verger et al., 2013) provides smoothing, gap

filling, and characterizing seasonal anomalies in satellite time series.

All these procedures give composite images that are smoothed versions of the raw images, but very often they are not completely free of noise. Many of the attributes that can be extracted from the combination of satellite image bands are still vulnerable to many atmospheric or electronic accidents. For example, highly reflective surfaces, including snow and clouds, and sun-glint over water bodies may saturate the reflective wavelength bands, with saturation varying spectrally and with the illumination geometry (Roy et al., 2016). Land surface temperature or normalized vegetation index are examples of attributes where these type of errors can be present. Therefore, after pre-processing is done, interpolation and smoothing methods can be very useful for drawing or detecting trend changes, clustering, or many other processes on remote sensing data.

2.3 Spatial interpolation

Likely, interpolation and classification are among the most used tools with remote sensing data. Classification of satellite images in supervised or unsupervised versions are important research areas not only with satellite images, but also with big data and data mining, where there are a great number of algorithmic procedures (see for example Benz et al., 2004). Here, we are more interested in interpolation as it is more closely related to geostatistics.

Interpolation has been widely used in environmental sciences. Li and Heap (2011) revise more than 50 different spatial interpolation methods that can be summarized in three categories: non-geostatistical methods, geostatistical methods, and combined methods. All of them can be represented as weighted averages of sampled data. Among the non-geostatistical methods the authors find: nearest neighbours, inverse distance weighting, regression models, trend surface analysis, splines and local trend surfaces, thin plate splines, classification, and regression trees. The different versions of simple, ordinary, disjunctive or model-based kriging are among the geostatistical methods. The combined methods include: trend surface analysis combined with kriging, linear mixed models, regression trees combined with kriging or regression kriging.

Recently, Li and Heap (2014) present an excellent review of spatial interpolation methods in environmental sciences introducing 10 methods from the machine learning field. These methods include support vector machines (SVM), random forests (RF), neural networks, neuro-fuzzy networks, boosted decision trees (BDT), the combination of SVM with inverse distance weighting (IDW) or ordinary kriging (OK), the combination of RF with IDW or OK (RFIDW, RFOK), general regression neural network (GRNN), the combination of GRNN with IDW or OK, and the combination of BDT with IDW or OK. Although all these methods were not developed specifically

for remote sensing data, nowadays the majority of them have been implemented in different packages of the free statistical software R, and can be used with satellite images. Many of these methods are ready to use and interpret, but the family of kriging methods as the core of geostatistics, are preferred and widely used.

2.4 Spatio-temporal interpolation

Since the publication of the seminal book on Spatial Autocorrelation (Cliff Andrew David, 1973), and at latter date Spatial Statistics (Ripley, 1981), Statistics for Spatial Data (Cressie and Wikle 2015), and Multivarate Geostatistics (Swan 1996) books, there has been a rapid growth of spatial geostatistical methods, as they are essential tools for interpolating meteorological, physical, agricultural or environmental variables in locations where these variables are not observed.

The use of spatial geostatistics with remote sensing data is also very well widespread, and its procedures are present in many specific softwares of satellite image analysis (Skidmore et al., 1999). Geostatistics techniques can help to explore and describe the spatial variability, to design optimum sampling schemes, and to increase the accuracy estimation of the variables of interest. These models can be enriched with auxiliary information coming from classified land cover or historical information (Curran and Atkinson, 1998). Kriging is the most popular geostatistical method with several versions such as block kriging, universal kriging, ordinary kriging, regression kriging or indicator kriging. It provides the spatial interpolation of different spatial variables through the use of spatial stochastic models, and it is the best linear unbiased predictor under normality assumptions when using spatially dependent data.

However, the extension to the spatio-temporal methods is more complicated. Time series models typically assume a regularly sampling over time, but the temporal lag operator cannot be easily generalized to the spatial domain, where data are likely irregularly sampled (Kyriakidis and Journel, 1999). Scales of time and space are different, therefore defining joint spatio-temporal covariance functions is not a trivial task (De Iaco et al., 2002). Recently, Cressie and Wikle (2015) show the state of the art in this area and explain the difficulties of inverting covariance matrices in spatio-temporal kriging, because it becomes problematic without some form of separable models or dimension reduction. Modelling the spatio-temporal dependence is frequently case-specific. Therefore, yet the presence of the spatio-temporal keyword is abundant in many satellite imagery papers, the use of spatio-temporal stochastic models is scarce. Very often, space-time refers only to descriptive analyses of time series of satellite images where every image is analysed as a set of separate pixels, i.e., when estimating trends, or trend changes, statistical methods of univariate time series are used for every pixel. For example, when completing, reconstructing or

predicting the spatial and temporal dynamics of the future NDVI distribution many papers use a time series of images (Forkel et al. 2013; Tüshaus et al. 2014; Klisch and Atzberger 2016; Wang et al. 2016; Liu et al. 2015; Maselli et al. 2014). These studies include temporal correlation of individual pixels at different resolutions but ignoring spatial dependence among them.

Spatio-temporal stochastic models use the spatial or temporal dependence to estimate optimally local values from sampled data. In satellite images, sampled data can be a huge amount of spatially and temporally dependent pixels, if a sequence of images is involved. In what follows a briefly review of some stochastic spatio-temporal models that can be used when analysing remote sensing data.

1. Spatio-temporal kriging (Gasch et al., 2015). This paper uses spatio-temporal R packages for fitting some of the following spatio-temporal covariance functions: separable, product-sum,metric and sum-metric classes in a spatio-temporal kriging model, and a random forest algorithm for modelling dynamic soil properties in 3-dimensions.

2. State-space models (Cameletti et al., 2011). The authors apply a family of state-space models with different hierarchical structure and different spatio-temporal covariance function for modelling particular matter in Piemonte (Italy).

3. Hierarchical spatio-temporal model (Cameletti et al., 2013). The paper introduces a hierarchical spatio-temporal model for particulate matter (PM) concentration in the North-Italian region Piemonte. The authors use state-space models involving a Gaussian Field (GF), affected by a measurement error, and a state process characterized by a first order autoregressive dynamic model and spatially correlated innovations. The estimation is based on Bayesian methods and consists of representing a GF with Matérn covariance function as a Gaussian Markov Random Field (GMRF) through the Stochastic Partial Differential Equations (SPDE) approach. Then, the Integrated Nested Laplace Approximation (INLA) algorithm is proposed as an alternative to MCMC methods, giving rise to additional computational advantages (Rue et al. 2009).

4. Spatio-temporal data-fusion (STDF) methodology (Nguyen et al., 2014). This method is based on reduced-dimensional Kalman smoothing. The STDF is able to combine the complementary GOSAT and AIRS datasets to optimally estimate lower-atmospheric CO2 mole fraction over the whole globe.

5. Hierarchical statistical model (Kang et al., 2010). This model includes a spatiotemporal random effects (STRE) model as a dynamical component, and a temporally independent spatial component for the fine-scale variation. This article demonstrates that spatio-temporal statistical models can be made

operational and provide a way to estimate level-3 values over the whole grid and attach to each value a measure of its uncertainty. Specifically, a hierarchical statistical model is presented, including a spatio-temporal random effects (STRE) model as a dynamical component and a temporally independent spatial component for the fine-scale variation. Optimal spatio-temporal predictions and their mean squared prediction errors are derived in terms of a fixed-dimensional Kalman filter.

6. Three-stage spatio-temporal hierarchical model (Fassò and Cameletti, 2009). This work gives a three-stage spatio-temporal hierarchical model including spatio-temporal covariates. It is estimated through an EM algorithm and bootstrap techniques. This approach has been used by Militino et al. (2015) for interpolating daily rainfall data, and for estimating spatio-temporal trend changes in NDVI with satellite images of Spain from 2011-2013 (Militino et al., 2017).

7. Space-varying regression model (Bolin et al., 2009). In this space-varying regression model the regression coefficients for the spatial locations are dependent. A second order intrinsic Gaussian Markov Random Field prior is used to specify the spatial covariance structure. Model parameters are estimated using the Expectation Maximisation (EM) algorithm, which allows for feasible computation times for relatively large data sets. Results are illustrated with simulated data sets and real vegetation data from the Sahel area in northern Africa.

2.5 Geostatistical R packages

In this section we briefly describe some of the most useful R packages for geostatistical analysis, including spatial and spatio-temporal interpolation in satellite imagery.

1. 'FRK' (Cressie and Johannesson, 2008) means fixed rank kriging and it is a tool for spatial/spatio-temporal modelling and prediction with large datasets.

2. 'geoR' (Ribeiro Jr et al., 2001) offers classical geostatistics techniques for analysing spatial data. The extension to generalized linear models was made in geoRglm package (Christensen and Ribeiro Jr, 2002).

3. 'georob' (Papritz, 2018) fits linear models with spatially correlated errors to geostatistical data that are possibly contaminated by outliers.

4. 'geospt' (Melo et al., 2012) estimates the variogram through trimmed mean and does summary statistics from cross-validation, pocket plot, and design

of optimal sampling networks through sequential and simultaneous points methods.

5. 'geostatsp' (Brown et al., 2015) provides geostatistical modelling facilities using raster. Non-Gaussian models are fitted using INLA, and Gaussian geostatistical models use maximum likelihood estimation.

6. 'gstat' (Pebesma, 2004) does spatio-temporal kriging, sequential Gaussian or indicator (co)simulation, variogram and variogram map plotting utility functions.

7. 'RandomFields' (Schlather et al., 2015) provides methods for the inference on and the simulation of Gaussian fields.

8. 'spacetime' (Pebesma et al., 2012) gives methods for representations of spatiotemporal sensor data, and results from predicting (spatial and/or temporal interpolation or smoothing), aggregating, or sub-setting them, and to represent trajectories.

9. 'spatial' (Venables and Ripley, 2002) provides functions for kriging and point pattern analysis.

10. 'spatialEco' (Evans, 2016) does spatial smoothing, multivariate separability, point process model for creating pseudo- absences and sub-sampling, polygon and point-distance landscape metrics, auto-logistic model, sampling models, cluster optimization and statistical exploratory tools. It works with raster data.

11. 'SpatialTools' (French, 2018) contains tools for spatial data analysis with emphasis on kriging. It provides functions for prediction and simulation.

12. 'spBayes' (Finley et al., 2015) fits univariate and multivariate spatio-temporal random effects models for point-referenced data using Markov chain Monte Carlo (MCMC).

2.6 Conclusions

This chapter analyses the state of the art of the geostatistical techniques applied to remote sensing data. The number of publications of remote sensing data with geostatistical techniques has rapidly increased. But unfortunately, not all published papers deriving, analysing or monitoring spatio-temporal evolutions, spatio-temporal trends or spatio-temporal changes are necessarily geostatistical papers, because they do not really use spatio-temporal stochastic models. These models are still scarce

in remote sensing data because many of these models are computationally very intensive, and they need huge among of data to run properly. Many of methods are not so broadly applicable as the spatial models are, and largely depends on seasonal dependency. The solutions found in the literature are very well fitted to specific problems, but we cannot always plug-in to other applications. The use of time series analysis in remote sensing opens a great window of opportunities for monitoring, smoothing, and detecting changes in large series of satellite images, but there are still many remote sensing papers ignoring the spatial dependence when analysing time series of images (Ban, 2016b). Instead, a huge discretization of the problem is presented where time-series of pixels are treated as spatially independent.

Nowadays, the upcoming opportunities for geostatisticians in remote sensing data are not based on the use of spatial models and time series separately, but on the use of spatial, temporal, or spatio-temporal stochastic models embedding both types of dependencies when necessary. Moreover, a single free statistical software like R is a powerful tool for downloading, importing, accessing, exploring, analysing and running advanced statistical modelling with remote sensing data in a row.

The contents of this chapter have been published as a chapter of the book *Handbook of Mathematical Geosciences*:
Militino, A. F., Ugarte, M. D., and Pérez-Goya, U. (2018c). An introduction to the spatio-temporal analysis of satellite remote sensing data for geostatisticians. In Sagar, B. D., Cheng, Q., and Agterberg, F., editors, *Handbook of Mathematical Geosciences*, chapter 13, pages 239–253. Springer.

3

Detecting change-points in a predefined surface in Spain

3.1 Introduction

The study of trends and trend changes is of crucial interest in many environmental studies. The objective of this chapter is to analyse the evolution of the NDVI in Spain for over 35 years and to detect trend changes.

Since Tucker (1979), numerous research projects and studies have been carried out. Mann-Kendall non-parametric test is one of the most broadly used methods for the analysis of parametric changes in time series of NDVI pixels. For example, de Jong et al. (2011), Li et al. (2013), or Sobrino et al. (2011). When plotting significant changes, a patchy map can be obtained because every pixel is analysed separately. Indeed, Mann-Kendall test only assumes time dependence within the same pixel across years, but it does not encompass the spatial dependence among neighbour pixels. Therefore, close locations can present different trend changes, something that could be questionable in some real situations. For example, Neeti and Eastman (2011) introduce the contextual Mann-Kendall approach that removes serial correlation through a pre-whitening process.

To detect spatio-temporal change-points in the NDVI trend is not a trivial task because of the different scales and dependencies between space and time. Therefore, we propose to aggregate pixels between pre-defined thresholds of NDVI values for estimating the occupied land cover area. Specifically, a total of four categories are obtained. For everyone of these categories we have a time series of areas where different change-point methods will be applied to detect breakdown points in means and variances.

In order to carry out the study presented in this chapter and to get a long time series, 3rd generation Global Inventory Modeling and Mapping Studies data

(GIMMS NDVI3g) are used.The NDVI 3g data have been modelled and smoothed from different AVHRR sensors since 1981. GIMMS NDVI3g data have been widely used during the last decades for studying large scale changing trends along years, mainly over continental or semi-continental regions. Its actual resolution of 8 Km at the Equator or 1/12 degrees is an attractive feature for monitoring changes of vegetation at any scale. These images are not raw images, but bi-weekly composite images by means of the Maximum Value Compositing (MVC) procedure explained in Chapter 2. This technique suppresses clouds, atmospheric and radiometric effects, and reduces the directional reflectance and off-nadir viewing effects. The result is a smaller number of output images with regard to the original ones, but with better quality and where the spatial and temporal stochastic dependence is still present.

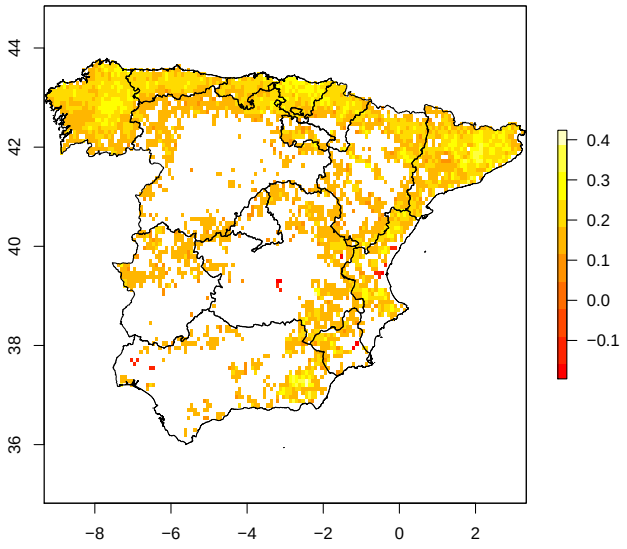

Figure 3.1: Mann-Kendall test applied to NDVI3g data. Coloured pixels correspond to significant changes in trends obtained from July 1981 to December 2015

This chapter contains the following four sections. Section 3.2 presents the GIMMS3g data used in the study. The section explains the importance of this data source focusing on its best characteristics, that is, the temporal availability of the time series. Section 3.3 shows R packages used to detect trend changes. The packages are: 'changepoint', 'structchange', 'bfast' and 'ecp'. Each of these four packages implements a single trend change detection method used in the study. Together with the packages, this section briefly describes the methods on which each of the

packages is based. Section 3.4 presents the real data analysis. The chapter finishes with Section 3.5 presenting the conclusions of the work.

3.2 Data

Remote sensing data were captured from the GIMMS NDVI3g images during the period July 1981-December 2015. More details of GIMMS NDVI3g can be found in Pinzon and Tucker (2014). It has been largely used along recent years, for example in Ahmed et al. (2017), Li et al. (2017) or Yuan et al. (2015). The data have flags accounting for additional pixel-by-pixel information about its quality. These flags can vary between 1 and 7, where 1 or 2 indicates good quality, numbers between 3 and 6 indicate different kinds of processing, and 7 indicates missing data. GIMMS NDVI3g data are bi-weekly composite NDVI data set and it has shown to be more accurate than the GIMMS NDVI predecessors for monitoring vegetation activity and phenological change [36]. GIMMS NDVI3g data can be downloaded from `http://ecocast.arc.nasa.gov/data/pub/gimms/3g.v1/`. For this study, we have downloaded 828 bi-weekly images, but to preserve space Figs. 2 and 3 provide the monthly averages of NDVI3g in continental Spain for the first and second semesters respectively from 2011 to 2015.

These images have been cropped, projected and plotted in the free statistical software R (R Core Team, 2019). In particular, library gimms (Detsch, 2019) has been used for downloading the images and importing in R, yet it can also be done with raster library (Hijmans, 2019). In green colors NDVI3g values closer to 1 are depicted and in brown colors the values closer to 0. In these maps, the pattern of high vegetation in the North and the Central West part of the country is predominant while middle vegetation is concentrated mainly in the watercourse of two important rivers: Guadiana and Guadalquivir and some mountain ranges. In the second semester low vegetation is predominant in the central part of Spain. This seasonality must be removed before applying change-point detection techniques.

3.3 Change-point methods

Change-points methods refer to the inference of a change in distribution for a set of observations. An excellent reference for these procedure is given in Chen and Gupta (2011). They arose in the 1950's from the process of quality control, and yet there were developed for independent and identically distributed random variables (Csörgö and Horváth, 1997), the expansion to the time-ordered observations (Antoch et al., 1997) was immediate. However, the application of these methods to remote sensing data is still very rare. In this work we compare four specific R packages for solving

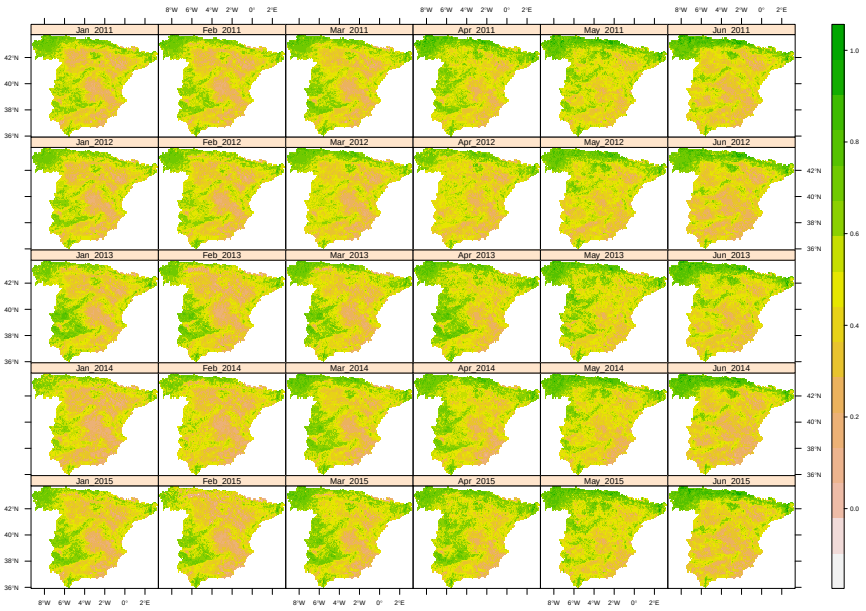

Figure 3.2: GIMMS NDVI3g monthly averaged data from Spain corresponding to the six first months from 2011 to 2015

the change-point detection problem in time series of land cover areas in Spain from 1981 to 2015.

3.3.1 Change-point package: segmented neighbourhood, binary segmentation and PELT

The 'changepoint' package (Killick et al., 2016) contains three methods for multiple change-point detection in addition to a variety of test statistics. The change can be either in mean and/or variance settings with a similar argument structure. The implemented methods are: Segmented neighbourhood, binary segmentation and PELT. See Killick and Eckley (2014) for details. Binary segmentation Edwards and Cavalli-Sforza (1965), Scott and Knott (1974), Sen and Srivastava (1975) first applies a single change-point test statistic to the entire data. If a change-point is identified, the data is split into two at the change-point location. The single change-point procedure is repeated on the two new data sets, before and after the change. If change-points are identified in either of the new data sets, they are split further. This process continues until no change-points are found in any parts of the data. The splitting is based on likelihood ratio-tests similar to those used in cluster analysis. The segment neighbourhood algorithm was proposed by Auger and Lawrence (1989), Bai and Perron (2003). The algorithm minimizes a penalized expression of cost

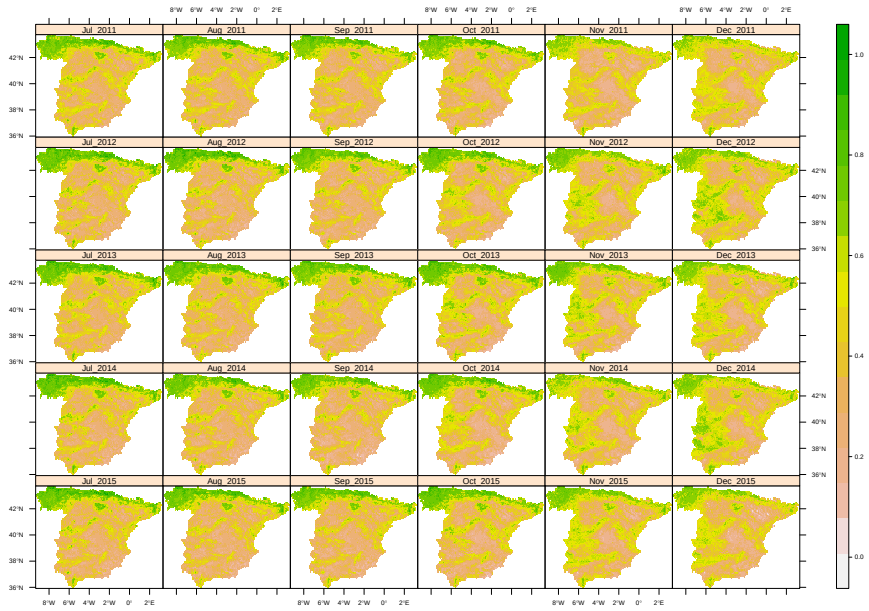

Figure 3.3: GIMMS NDVI3g monthly averaged data from Spain corresponding to the six last months from 2011 to 2015

using a dynamic programming technique to obtain the optimal segmentation for $m+1$ change-points reusing the information that was calculated for m change-points. The PELT algorithm Killick et al. (2012) is similar to the segment neighbourhood algorithm in that it provides an exact segmentation. It is computationally more efficient, due to its use of dynamic programming and pruning. The test statistics are likelihood ratio tests that can be applied to different families of distributions (Chen and Gupta, 2011).

3.3.2 'ecp' package: divisive and agglomerative algorithms

The 'ecp' package (James and Matteson, 2014) contains two algorithms: divisive and agglomerative. These algorithms come from hierarchical cluster analysis and detect changes within the marginal distributions. They do not make any assumption regarding the nature of the change in distribution or any distribution assumptions beyond the existence of the αth absolute moment, for some $\alpha \in (0, 2)$. The agglomerative algorithm estimates change point locations through an optimal segmentation. Both approaches are able to detect any type of distributional change within the data. The divisive method provides consistent estimates of both the number and location of change points, under standard regularity assumptions. These methods also deal with the nonparametric multiple change point analysis of multivariate data. Regardless of

the dimension, the nonparametric estimation can be done for both the number of change points and the positions at which they occur. These procedures have been widely used in financial modeling (Talih and Hengartner, 2005), and bioinformatics (Matteson and James, 2014) to identify genes that are associated with specific cancers and other diseases or to detect credit card fraud (Bolton and Hand, 2002).

3.3.3 'bfast' package: breaks for additive seasonal and trend

The more specific R program to manage with change-point detection in time series of satellite images is 'BFAST' Verbesselt et al. (2010a), Verbesselt et al. (2010b). BFAST is the acronym of "Breaks For Additive Seasonal and Trend" that integrates the decomposition of time series into trend, seasonal, and remainder components with methods for detecting change within time series. It iteratively estimates the time and number of changes characterizing the change by its magnitude and direction and using harmonic seasonal model requiring few observations.

3.3.4 'strucchange' package: generalized fluctuation and F tests

The 'strucchange' package Zeileis et al. (2002) contains Generalized fluctuation and F test for structural change in linear regression models. Here, the null hypothesis of "no structural change" is tested against the alternative that the coefficient vector varies over time for certain patterns of deviation from the null hypothesis. Significance can be also assessed through various tests. (see Zeileis, 2006; Zeileis et al., 2003, for details).

3.4 Results

For everyone of the 828 images, NDVI3g values are assigned to 4 categories: bare soils (ndvi1) for values between 0 and 0.2, sparse vegetation (ndvi2) for values greater than 0.2 and less or equal than 0.5, middle vegetation (ndvi3) for values greater than 0.5 and less or equal than 0.7 and dense vegetation (ndvi4) for values greater than 0.7. Table 3.1 shows the average occupied area (in %) for the four NDVI3g classifications estimated in the continental Spain and in 15 regions. The classified areas have been calculated summing the number of pixels in these categories and multiplying by 65.95 km^2, the mean surface by pixel. In the whole territory the average percentage of bare soils is estimated in 3%, the average percentage of sparse vegetation is 64%, the average percentage of middle vegetation is 26% and 6% is the average percentage of dense vegetation. The 100% corresponds to the 504.537 km^2 of continental Spain.

Table 3.1: Average percentage of the area occupied by the 4 pre-defined NDVI3g categories estimated in Spain and in 15 regions from 1981 to 2015

	num	ndvi1	ndvi2	ndvi3	ndvi4
Andalucia (An)	1	4.87	76.27	18.04	0.82
Aragon (Ar)	2	3.70	78.78	17.20	0.32
Cantabria (Ca)	3	0.19	7.90	47.68	44.23
Castilla-La Mancha (Cm)	4	5.66	81.62	12.59	0.13
Castilla y Leon (Cl)	5	1.60	64.57	32.00	1.83
Cataluna (Ct)	6	1.41	48.80	44.13	5.65
Comunidad de Madrid (Ma)	7	2.78	75.95	21.21	0.06
Comunidad Foral de Navarra (Na)	8	1.32	49.53	32.57	16.58
Comunidad Valenciana (Va)	9	1.76	83.04	15.20	0.00
Extremadura (Ex)	10	1.16	61.99	35.02	1.82
Galicia (Ga)	11	0.01	3.37	55.06	41.57
La Rioja (Ri)	12	0.28	62.04	33.18	4.49
Pais Vasco (Pv)	13	0.05	13.95	46.11	39.89
Principado de Asturias (As)	14	0.29	5.76	44.52	49.42
Region de Murcia (Mu)	15	9.67	90.15	0.18	0.00
Spain	16	3.01	64.34	26.29	6.36

Table 3.2 gives the two last figures of the change-point years detected by the four methods in Spain and by regions. Columns `cp1`, `cp2`, `cp3` and `cp4` show the year of the first detected change-point by 'changepoint' package in ndvi1, ndvi2, ndvi3 and ndvi4 categories respectively. Columns `str1`, `str2`, `str3` and `str4` show the year of the first detected change-point by *strucchange* method in the same four categories. Similarly, columns `bf1`, `bf2`, `bf3` and `bf4` for `bfast` package. Columns `prun1`, `prun2`, `prun3` and `prun4` do the same with 'ecp' package and finally, `prun` shows the year of the detected change-point in the overall NDVI3g. Empty places correspond to an absence of change-point.

In Spain, `cp` and `str` methods exactly coincide detecting the year of changepoint in ndvi1 (year 1996) and ndvi3 (year 2000) categories. They do not coincide in ndvi2 (years 1993 and 1986) and they roughly coincide in ndvi4 (year 1989-1990), the upper category. Likely, both methods are the best candidates to explain the performance of the Spanish land cover change between July 1981 and December 2015. Figures 4 and 5 plot the detected change-points over the seasonally adjusted trends in the four pre-defined categories in Spain from 1981 to 2015 with the 4 methods. The proximity between *changepoint* and *strucchange* methods is clear

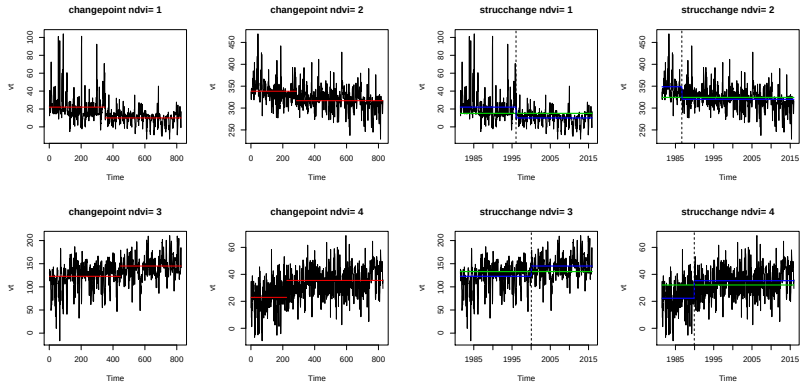

Figure 3.4: Change-points in seasonally adjusted trends of GIMMS NDVI3g data obtained with *changepoint* and *strcchange* packages in the 4 pre-defined categories from 1981 to 2015

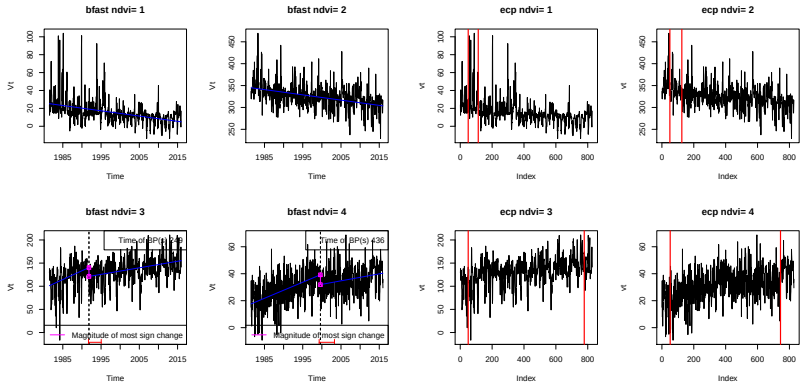

Figure 3.5: Change-points in seasonally adjusted trends of GIMMS NDVI3g data obtained with *bfast* and *ecp* packages in the 4 pre-defined categories from 1981 to 2015

Table 3.2: Years of change-points detected in the overall NDVI3g, and in the four pre-defined categories with *changepoint* (cp), *structchange* (str), *bfast* (bf) and *ecp* (prun) methods calculated in continental Spain and in 15 regions

regions		ndvi1				ndvi2				ndvi3				ndvi4				ndvi
		cp1	str1	bf1	prun1	cp2	str2	bf2	prun2	cp3	str3	bf3	prun3	cp4	str4	bf4	prun4	prun
An	1	96	96	91	86	12	10	92	13	0	0	91	13	4	1		13	86
Ar	2	97	86	91	86	7	7	2	13	96	96	5	10	6	10	2	13	86
Ca	3	87	87		13	88	88	99	86	90	90		13	90	90	98	86	13
Cm	4	96	96		86	10	10		13	0	96		85	4	3		13	86
Cl	5	94	96		86	93	86		86	93	86		13	96	97		13	86
Ct	6	89	88	93	86	89	89	98	86	89	89	92	86	88	88	98	12	13
Ma	7	97	96		86	86	86		13	86	86	92	13	87			13	86
Na	8	89	96	96	90	91	89	99	13	86	87		13	91	94		13	90
Va	9	0	0	91	1	7	7	91	12	6	6	91	12	91			13	1
Ex	10	95	95		13	86	86	4	13	86	96		13	96	1		12	13
Ga	11	99	86		13	89	89	94	86	90	90	99	85	90	90	99	13	13
Ri	12	97	87		13	89	88		13	88	88		13	97	97		13	13
Pv	13	87	87	92	86	89	89	91	13	89	94		13	89	89	97	86	86
As	14	91	86		13	89	88	91	86	89	89	94	13	89	89	97	13	13
Mu	15	86	86	91	86	86	86	91	86	7	7	91	13	15			13	86
Spain	16	96	96		86	93	86		86	0	0	91	13	90	89	99	12	86

in all categories except in ndvi2 where *changepoint* method is more exigent and conservative for detecting change-points. Both methods show a decreasing trend in the lowest categories (ndvi1, ndvi2), but an increased trend in the upper categories (ndvi3 and ndvi4). Package 'bfast' does not detect any change-point neither in ndvi1 nor in ndvi2 but it detects changes in ndvi3 and ndvi4. The *prun* method estimates change-points in the beginning or in the last years, so it seems to be very sensible to small changes. Unfortunately, in the majority of regions, the year of the detected change-point do not coincide neither in methods nor in the categories, although the bigger the regions, the better the approximation.

Figure 6 shows the evolution of the yearly average area in hectares, corresponding to the 4 categories in Spain from the same studied period. Clearly, a decreasing trend in bare and semi-arid soil is observed, corresponding to ndvi1 and ndvi2 categories, an important increase of middle vegetated soil corresponding to ndvi3 category, and a small increase of trend in ndvi4, the dense vegetation category.

3.5 Conclusions

Nowadays, satellite remote sensing is a common instrument for detecting changes in land cover surfaces over time. GIMMS NDVI3g provides a world wide long time series very useful for analysing temporal trends, however and yet the quality of the series have been improved with regard to the old NDVI series, it is known that there is no concordance with other NDVI images coming from alternative sources, as MODIS TERRA or MODIS AQUA. See Atzberger et al. (2013), Kern et al. (2016) where a

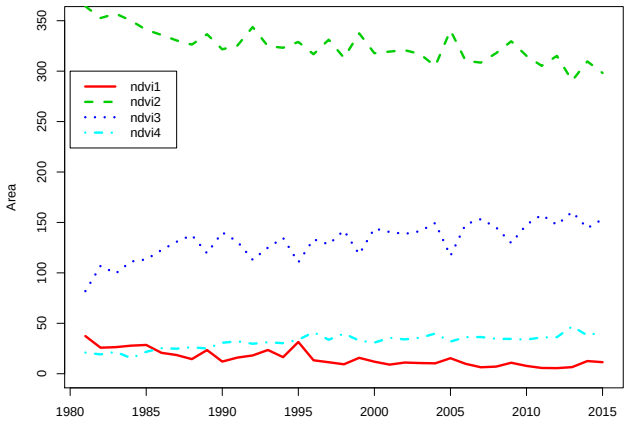

Figure 3.6: Yearly averages of areas corresponding to the 4 pre-defined categories (ndvi1, ndvi2, ndvi3 and ndvi4) in Spain from 1981 to 2015

detailed comparison has been made in Central Europe from 2000 to 2013. Moreover, as long as we down-scaling the spatial resolution, more inaccurate estimations we obtain. From this perspective, we can say that the change-points trends found in Spanish regions are only approximate for small regions such as Navarra, Asturias, Murcia, Pais Vasco o Aragon. The main advantage of using GIMMS NDVI3g is that is the longest NDVI series of images with 34 years, already pre-processed, easily accessible, and from all over the world. As long as we can retrieve longest series of high spatial resolution satellite images, these results could change.

The variety of methods found in the literature for detecting change-points in ordered observations is large, and they do not necessarily provide the same points. At this regard, we consider that matching results at different categories tip the balance in the *changepoint* and *strucchange* favour. Unfortunately, this methodology cannot determine the locations where these changes have been produced, because we loose the spatial location as long as we aggregate different pixels between the same thresholds. For this aim we need not only to develop a specific spatio-temporal methodology but also a larger spatial resolution of time series of images. For example, Sentinel-2A can provide the required spatial resolution, however, the history of these images is still too short for being reliable in the time series analysis.

There is an inherent difficulty in checking the performance of this result, because there are not previous studies similar to this one. Perhaps, this step can only be done looking for vegetation changes previously documented. The most relevant is Julien

et al. (2011) where the authors investigate the NDVI changes in trends happened in Iberian peninsula between 1981 and 2001, using GIMMS NDVI3g data with a pixel by pixel approach. The interpretation of global trends in the peninsula is limited, although the results show a slight desertification in Iberian Mountains, but is dated more than 15 years ago

Based on the results given in *changepoint* and *strucchange* methods we can finally conclude that the detected change-points in Spain show a decrease of bare soils and semi-bare soils starting in the middle nineties or a bit before, and a slight increase of middle-vegetation and high-vegetation soils starting in 1990 and 2000 respectively. Further research is needed to confirm the results found in this pilot study.

The contents of this chapter have been published as a chapter of the book *The Mathematics of the Uncertain*:
Militino, A. F., Ugarte, M. D., and Pérez-Goya, U. (2018b). Detecting change-points in the time series of surfaces occupied by pre-defined ndvi categories in continental spain from 1981 to 2015. In Gil, E., Gil, E., Gil, J., and Gil, M. Á., editors, *The Mathematics of the Uncertain*, chapter 28, pages 295–307. Springer.

4 Stochastic spatio-temporal models for analysing the NDVI distribution

4.1 Introduction

The normalized difference vegetation index (NDVI) is an important indicator for evaluating vegetation change, monitoring land surface fluxes or predicting crop models. Due to the great availability of images provided by different satellites in recent years, much attention has been devoted to testing trend changes with a time series of NDVI individual pixels. However, the spatial dependence inherent in these data is usually lost unless global scales are analyzed. In this chapter, we propose incorporating both the spatial and the temporal dependence among pixels using a stochastic spatio-temporal model for estimating the NDVI distribution thoroughly. The stochastic model is a state-space model that uses meteorological data of the Climatic Research Unit (CRU TS3.10) as auxiliary information. The model will be estimated with the Expectation-Maximization (EM) algorithm. The illustration is carried out with GIMMS NDVI3g images of continental Spain with a temporal resolution of fifteen days from January 2011 to December 2013, yet the model can be applied for many other variables, countries or regions with different resolutions.

In remote sensing data, atmospheric conditions or cloud presence alter the correct estimation of NDVI. A large number of papers have been devoted to completing, reconstructing and predicting the spatial and temporal dynamics of the future NDVI distribution using a time series of images (see, for example, Forkel et al., 2013; Tüshaus et al., 2014; Klisch and Atzberger, 2016; Wang et al., 2016; Liu et al., 2015; Maselli et al., 2014). These studies are mainly based on including temporal correlation of individual pixels at different resolutions but ignoring spatial dependence among them. Perhaps the most broadly used method for analysing NDVI temporal changes is the non-parametric Mann-Kendall test (see, for example, Li et al., 2013;

de Jong et al., 2011; Sobrino et al., 2011). When plotting significant changes, a discrete pixel by pixel map of the NDVI trend changes is obtained. Figure 4.1 shows the coloured pixels where significant trend NDVI changes have been detected in continental Spain from October 2011 to December 2013. This discretization comes because the Mann-Kendall test only assumes a time dependence within the same pixel across years, but it does not encompass the spatial dependence among neighbour pixels. Therefore, unless random disturbances occur because of fire events, land-use/cover changes, crop rotation, land degradation or many other causes, we expect that close locations present similar trend changes. Some improvements of this test have been also provided.

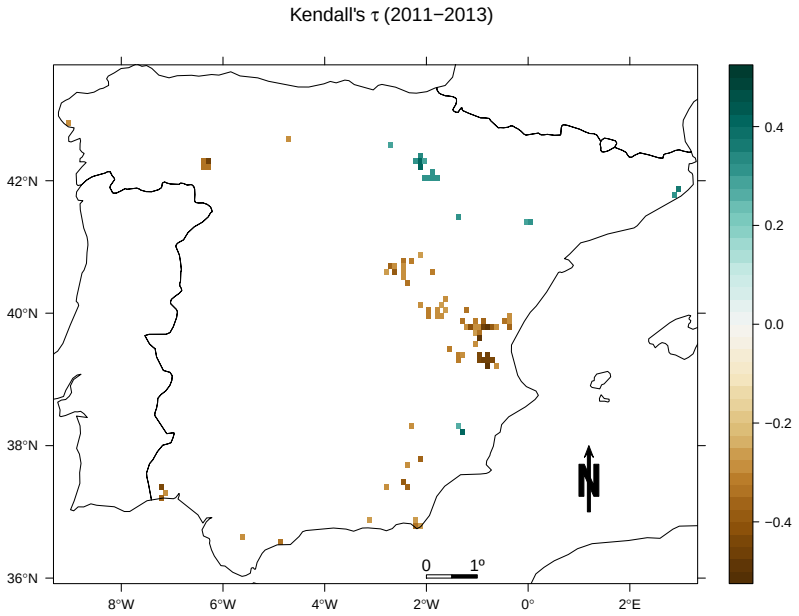

Figure 4.1: Coloured pixels correspond to significant trend changes of NDVI with a Mann-Kendall test in Spain from October 2011 to December 2013.

Jönsson and Eklundh (2004) provides the TIMESAT free program designed primarily for analyzing the time series of satellite data. It uses an adaptive Savitzky-Golay filtering and methods based on upper envelope weighted asymmetric Gaussian and double logistic model functions. This program Eklundh and Jönsson (2012) has been used in this paper for comparison purposes. The use of stochastic spatio-temporal models (Cressie and Wikle, 2015) is scarce with satellite data. Hengl et al. (2012) use a spatio-temporal regression kriging for smoothing land surface temperature data of MODIS MOD11A2.The difficulty of this method lies in fitting the variogram necessary for modelling the spatio-temporal dependence that increases depending on the number of periods and stations. As an alternative, we propose a

stochastic state-space model that simultaneously exploits dependencies across space and time. Figure 4.2 shows the graphical summary followed in the paper.

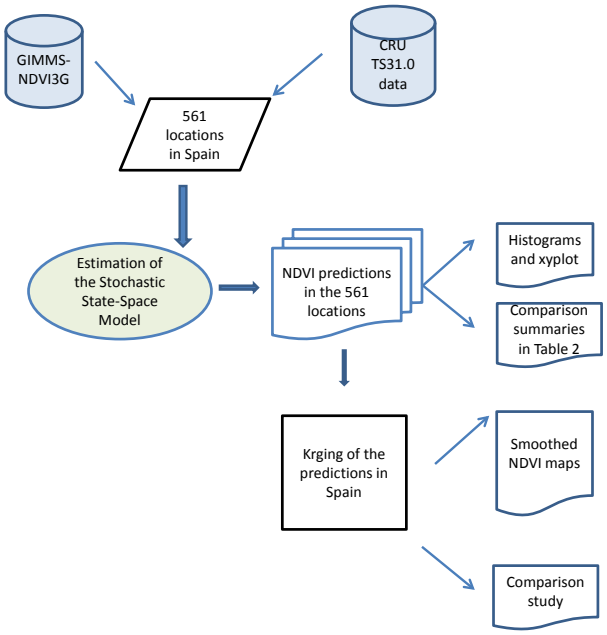

Figure 4.2: Graphical summary of the computational processes in this paper.

This chapter contains the following four sections, Section 4.2 introduces the GIMMS NDVI3g data. The covariates used are also introduced in this section, explaining the rainfall data obtained from the Climate Research Unit (CRU) database. Section 4.3 explains the stochastic state space model (SSM), Section 4.4 presents the results of the modelling, where as expected, the flaws and errors of the capture, represented as high values in the NDVI image, are smoothed. In addition, the SSM is compared with the TIMESAT methods that only use the temporal dependency. Finally Section 4.5 gives the conclusions of the work.

4.2 Data

Global Inventory Modelling and Mapping Studies Normalized Difference Vegetation Index (GIMMS NDVI3g) between 2011 and 2013 are used for analysing the spatio-temporal NDVI distribution in continental Spain. GIMMS NDVI3g data are bi-weekly composite NDVI data. The composite images are obtained by the Maximum Value Compositing (MVC). It has been shown to be more accurate than the GIMMS NDVI data for monitoring vegetation activity and phenological change (Wang et al., 2014). More details on GIMMS NDVI3g can be found in Pinzon and Tucker (2014).

The GIMMS NDVI3g time series is an improved normalized difference vegetation index (NDVI) data set produced from Advanced Very High Resolution Radiometer (AVHRR) instruments that extends from 1981 to the present onboard NOAA satellite. It has been largely used along recent years, for example in Zhang et al. (2016); Yuan et al. (2015). The data have flags accounting for additional information about the pixel quality. These flags can vary between 1 and 7, where 1 or 2 indicates good quality, numbers between 3 and 6 indicate different kinds of processing, and 7 indicates missing data.

The spatial resolution of these data is 8 km at the equator, but it has been corrected for calibration, view geometry, volcanic aerosols, and other effects not related to vegetation changes, providing sometimes unrealistic values of the NDVI when downscaling the NDVI index to smaller regions (Erasmi et al., 2014). Figure 4.3 shows the 72 scenes of original GIMMS NDVI3g data cropped to continental Spain from January 2011 to December 2013 and plotted in the free statistical software R (R Core Team, 2019). In particular, 'gimms' package (Detsch, 2019) has been used for reading the images in R, yet it can also be done with 'raster' package (Hijmans, 2019). According to this figure, western and northern Spanish regions have the maximum limit of NDVI, even in the summer, which is usually the driest season, which is an unlikely case in this country, particularly in the central western regions.

To calibrate the stochastic space-time model, climate data from the Climatic Research Unit (CRU) are used. This is a gridded climate data set of monthly observations taken at meteorological stations across the world land areas and referred to as CRU TS3.10. Station anomalies were interpolated into 0.5 degrees latitude/longitude grid cells covering the global land surface (excluding Antarctica) and combined with an existing climatology database to obtain absolute monthly values. Detailed information can be found in Harris et al. (2014). Figure 4.4 part (a) shows the grid locations of CRU TS3.10 data where auxiliary meteorological information is drawn. This database contains the following auxiliary variables:

```
cld  cloud cover  percentage (%) x 10
dtr  diurnal temperature range  degrees Celsius x 10
frs  frost day frequency  days x 100
pet  potential evapotranspiration  millimetres per day x 10
pre  precipitation  millimetres per month x 10
tmp  daily mean temperature  degrees Celsius x 10
tmn  monthly average daily minimum temperature  degrees Celsius x 10
tmx  monthly average daily maximum temperature  degrees Celsius x 10
vap  vapour pressure  hectopascals (hPa) x 10
wet  wet day frequency  (rain days per month) days x 100
```

In this list, only cld, frs, pre, tmx, vap and wet variables are used because dtr, pet, tmp and tmn can be derived from the rest, and the stochastic spatio-temporal

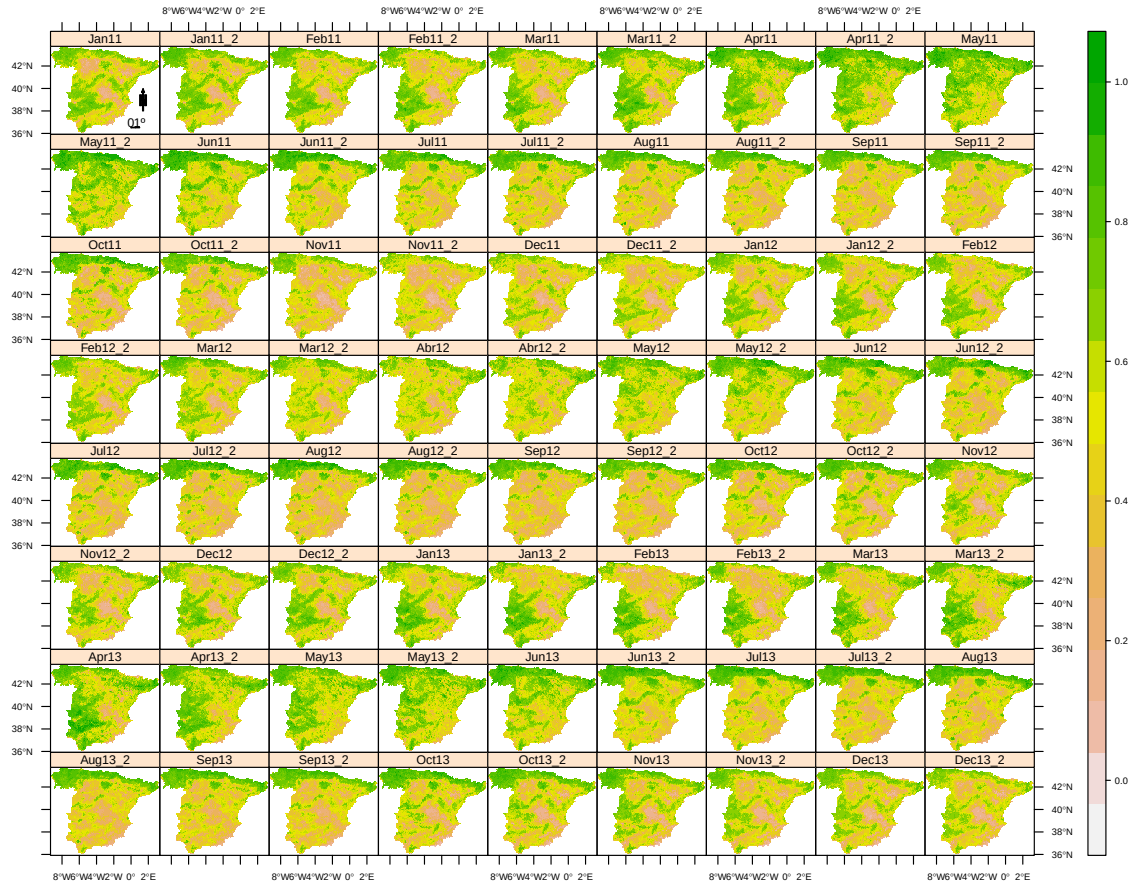

Figure 4.3: GIMMS NDVI3g images in continental Spain from January 2011 to December 2013.

models require independent auxiliary variables for avoiding multicollinearity (Ugarte et al., 2015). The six chosen variables will be called covariates hereafter.

From the GIMMS 3g data, we randomly choose $n = 561$ locations among those with good flag attributes (indicating high quality). These locations are plotted on the right of Figure 4.4, (part b). In these locations, we extract the meteorological information of the six covariates. As the temporal resolution of CRU data differs from GIMMS NDVI3g data (monthly versus bi-monthly data), we decided to transform CRU monthly data in bi-monthly data. In particular, cld, frs, $tmax$ and vap remain invariant in the corresponding fifteen days, but pre, frs and wet are divided by two. Next, the CRU covariates and the altitude of the sampled locations are organized in a 561×433 matrix. The first column corresponds to the height values of the n sampled observations, and the rest are blocks of 72 periods by six covariates. The number of sampled locations have been chosen after checking different sizes between 300 and

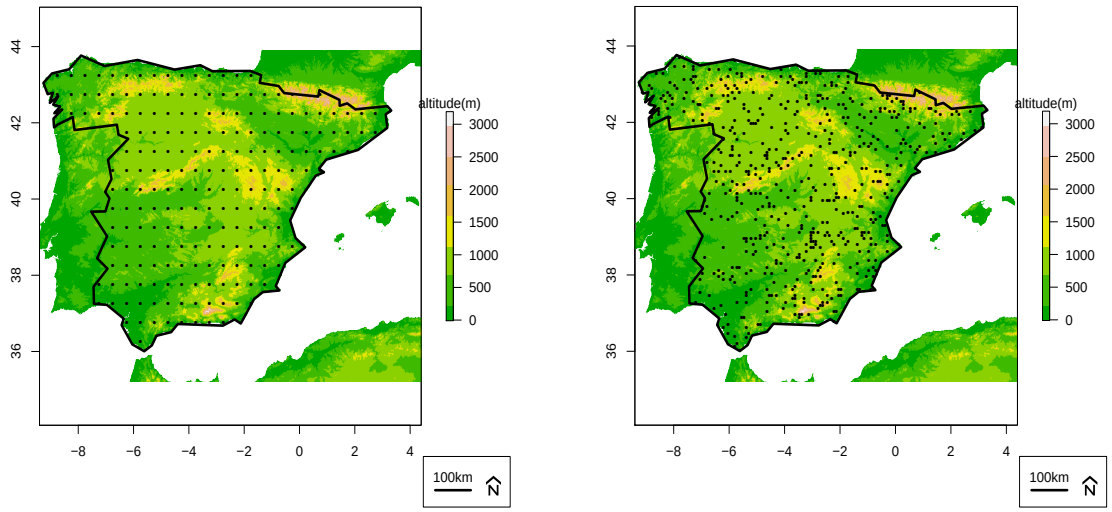

Figure 4.4: (a) grid locations of CRU TS3.10 meteorological data where auxiliary information is drawn for calibrating satellite data and (b) sampled locations used for estimating the state-space model.

1000 locations. From 300 locations, similar results have been obtained. This number is closely related to the meteorological data resolution because meteorological data must be drawn at these sampled locations, yet only a limited number of 211 pixels of CRU TS3.10 data are inside continental Spain. It means that only 211 different sets of covariates are available for being used in the model, and, then, negligible differences in model coefficient estimates are found when increasing the number of sampled locations.

4.3 The state-space model (SSM)

The state-space model is a very well-known mathematical tool used in dynamical systems. It became very used in econometrics since the publication of Durbin and Koopman (2012), and, more recently, Fassò and Cameletti (2009) have developed it in a spatio-temporal context. The state-space model is a spatio-temporal linear model that simultaneously accounts for spatial and temporal dependence. It is a hierarchical model in two steps defined by two equations: the transition Equation (4.1) and the state Equation (4.2). Here, the first equation explains a linear regression between NDVI and the covariates. In this example, the covariates are the meteorological variables and the altitude. The second equation expresses the temporal dependence. More precisely, let us denote z_{st} as the sampled NDVI value at location s and time t. The stochastic process at n locations $s_1, \ldots, s_n$ and T time points t_j, from $j = 1, \ldots, T$, is represented by $\mathbf{z}_{st} = (z(s_1, t_1), z(s_1, t_2), \ldots, z(s_1, t_T), \ldots, z(s_n, t_1), \ldots, z(s_n, t_T))'$.

In this case, $T = 72$ and $n = 561$ locations randomly chosen inside continental Spain. The model is given by

$$\mathbf{z}_{st} = \beta_0 + \beta_1 \mathbf{x}_{1s} + \beta_2 \mathbf{x}_{2st} + \cdots + \beta_7 \mathbf{x}_{7st} + \mathbf{v}_t + \boldsymbol{\epsilon}_{st}, \quad \boldsymbol{\epsilon}_{st} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(d)), \tag{4.1}$$

$$\mathbf{v}_t = \mathbf{G}\mathbf{v}_{t-1} + \boldsymbol{\eta}_t, \quad \mathbf{v}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta), \tag{4.2}$$

where β_i, $i = 0, \ldots, 7$ are the coefficients to be estimated. The first covariate $\mathbf{x}_{1s}$ is time invariant and corresponds to the altitude of the n sampled locations. The rest of the covariates $\mathbf{x}_{ist}, \quad i = 2, \ldots, 7$ are the spatio-temporal meteorological covariates: maximum temperature, frost day frequency, precipitation, wet day frequency, cloudy cover percentage, and vapor pressure respectively, depending upon the location s and time t. The unobservable latent temporal process, $\mathbf{v}_t$, takes account of the temporal dynamics of data through an autoregressive process. It means that the current state $\mathbf{v}_t$ depends on the previous state $\mathbf{v}_{t-1}$ in the state equation through a transition matrix $\mathbf{G}$. The initial $T \times 1$ state vector, $\mathbf{v}_0$ is assumed to be normally distributed with mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$.

The spatial dependence is accounted for in the covariance structure of the model error, $\boldsymbol{\epsilon}_{st}$, given by $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(d)$. It can be estimated using well-known covariance functions such as the Matérn, the exponential or the spherical covariance functions (Militino and Ugarte, 2001). The covariance function depends on the Euclidean distance $d = ||\mathbf{s}_i - \mathbf{s}_j||$, and, therefore, the sampled locations must be Universal Transverse Mercator (UTM) projected. It is invariant to translations, so $\mathbf{z}_{st}$ is assumed to be a second-order stationary process. The additive $T \times 1$ state-estimation errors, $\boldsymbol{\eta}_t$, and the $s \times 1$ measurement errors $\boldsymbol{\epsilon}_{st}$ are uncorrelated Gaussian white noises with zero mean and covariance matrices $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(d)$, respectively. Finally, $\boldsymbol{\eta}_t$ quantifies the uncertainty of the state estimate given the n observations. The transition equation incorporates the spatial dependence and the state equation takes into account the temporal dependence. Therefore, this state-space model can be interpreted as a spatio-temporal kriging model with a separable spatio-temporal covariance function.

4.3.1 Running the state-space model

The state-space model is implemented in the R statistical software package 'Stem' (Cameletti, 2012). It was originally applied to predict air concentrations and to deal with error measurements in instruments (Fassò and Cameletti, 2009). Similar state-space models were used by Amisigo and Van De Giesen (2005) to estimate model parameters and missing data of river basin runoff values, and by Militino et al. (2015) to interpolate daily rainfall in Navarre (Spain). This package uses the function *Stem.Estimation* to carry out the iterations of the EM algorithm Dempster et al. (1977) until convergence. Each iteration calls the function *kalman* to perform both the E-step and the M-step. The exponential covariance function is also assumed for

$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(d)$. The maximization process of the likelihood is done with the Kalman filter. We recommend using the coefficient estimates of the multiple linear regression model without assuming spatial dependence as initial values of the coefficients in the EM algorithm. When the state-space model is fitted, the $\boldsymbol{\beta}$ coefficients are obtained and tested. Additional programming is necessary for calculating $NDVI$ predictions for the whole Spain.

Before running the state-space model, the temporal and the spatial dependence need to be explored. Unfortunately, there are no statistical tools for checking jointly the spatio-temporal dependence. Therefore, it can be checked only marginally. Figure 4.5 shows the autocorrelation function of the first 6 NDVI pixels, although similar results are obtained in the rest. Excluding the first vertical bar that is always equal to one because it is the autocorrelation of a pixel with itself, all the locations have at least one vertical bar above the blue dotted horizontal line, showing that the temporal dependence is significant in at least one lag. As expected, a slight seasonality is also present in these data, yet it can change from one pixel to another. To check marginally the spatial dependence, the Moran test (Moran, 1950) is used. This test has been computed for every one of the 72 GIMMS NDVI-3g Spanish scenes. It varies between 0.72 and 0.84, indicating a strong spatial autocorrelation. The *acf* function from 'base' package (R Core Team, 2019) and *Moran* function from 'raster' package (Hijmans, 2019) have been used in this step.

4.4 Results

Classical statistical tools are used for checking the statistical significance of the model coefficients. Table 4.1 shows the estimates, the standard errors, the t-values, and the confidence intervals of the state-space model coefficients. Standard errors are obtained by bootstrapping 10 replicates, but similar results are derived when increasing the number of replicates. All the coefficients are statistically significant because no one of the confidence intervals contain the zero value, except for the *wet* variable.

Different random sets of 561 sampled locations have been essayed with similar results. In some cases, the *wet* covariate is statistically significant but with a very small estimate. Therefore, this covariate has been kept in the model, yet we know that it has a negligible impact in the predictions. Interpretation of sign estimates allows to conclude that NDVI is positively correlated with altitude, precipitation, and number of cloud days. However, NDVI decreases when maximum temperature or vapour pressure increase as expected. Meteorological covariates have been divided by 100 and altitude by 1000 because scaling covariates help to avoid singularities in the process of inverting matrices. Maximum temperature could be substituted by the average or minimum temperature without altering significantly the model

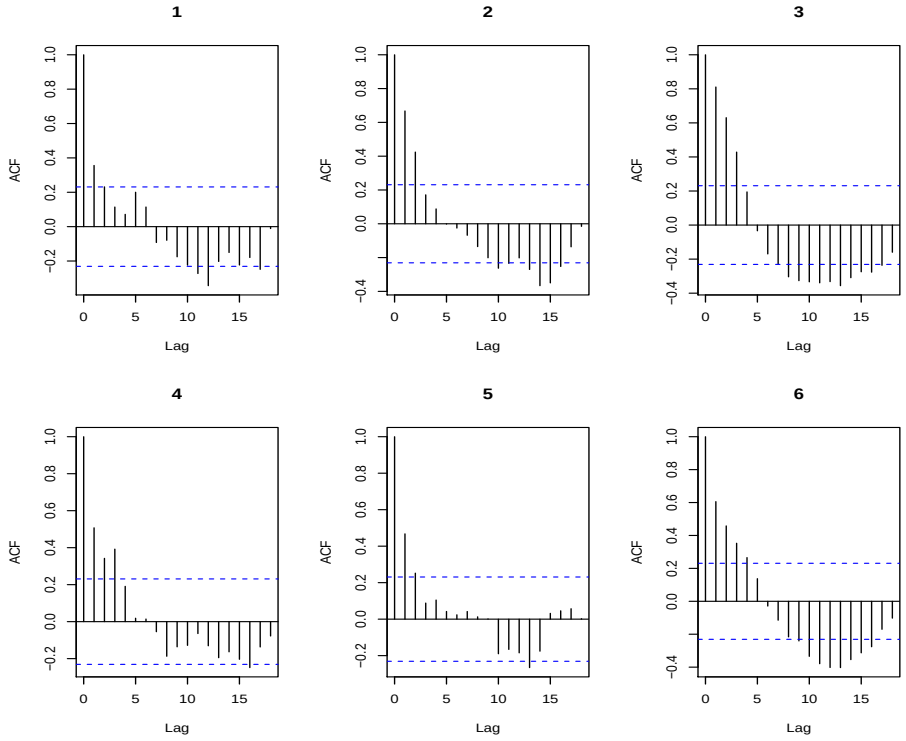

Figure 4.5: Autocorrelation function in six sampled locations of raw GIMMS NDVI3g data.

estimation and the predictions. The model has been statistically validated testing the normality of the residuals.

For checking the model, we firstly compare sampled versus predicted data both in a unique period for all of the 561 locations, and, separately, in every one of the 72 bi-monthly periods. The overall summary of the sampled and predicted values of NDVI from 2011 to 2013 are shown in Table 4.2, where we can observe that the model does not only provide the same average for sampled and predicted values, but also similar quantile values. The smoothing process crosses over the most extreme values as expected.

The state-space model predictions not only follow the pattern of GIMMS NDVI3g data in the overall period (2011–2013), but also in everyone of the 72 bi-monthly periods.Figure 4.6 plots sampled versus predicted NDVI data in the 72 periods, exhibiting also a close proximity between them. Therefore, the good performance of the model in sampled data is not only shown in summary statistics but also in all of the sampled locations. Later, an ordinary kriging was applied in every one of the 72 bi-monthly periods to get an overall image of the whole continental Spain. 'geostatsp' package (Brown, 2015) has been used in this step. Figure 4.7 shows the

Table 4.1: Estimates, standard error, t-values and 95% confidence intervals of the state-space model coefficients.

	Estimate	**SE**	**T-Stat.**	**CI_low**	**CI_upp**
(intercept) β_0	1.1343	0.0086	131.9563	1.1176	1.1435
(height) β_1	0.0471	0.0027	17.2019	0.0425	0.0501
(tmax) β_2	-0.1235	0.0039	−32.0501	−0.1313	−0.1216
(frs) β_3	−0.0153	0.0011	−14.3707	−0.0163	−0.0135
(wet) β_4	−0.0007	0.0008	−0.8950	−0.0010	0.0011
(prec) β_5	0.0190	0.0011	17.7825	0.0176	0.0209
(cld) β_6	0.0142	0.0010	13.7122	0.0116	0.0146
(vap) β_7	−0.0176	0.0070	−2.5172	−0.0208	−0.0013

Table 4.2: Minimum, first quantile, median, mean, third quantile, and maximum of the sampled and state-space smoothed NDVI data.

Summary	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
sampled NDVI	0.0140	0.4130	0.5410	0.5421	0.6740	1.0000
state-space smoothed NDVI	0.0867	0.4211	0.5392	0.5421	0.6604	0.9580

monthly predictions obtained by averaging the bi-monthly predictions. To complete the validation process, we compare these results to the documented information retrieved from the the Spanish National Agency of Meteorology (AEMET) (AEMET, 2019) and the Spanish CRU TS3.10 meteorological data.

Spain is the fifth largest country in Europe with an extension of 505,000 km^2 and an average altitude of 650 m, the third highest country in Europe. It has three climatological regions. The Mediterranean region with dry and warm summers and cool to mild, wet winters. The oceanic region located in the North of Spain and characterised by relatively mild winters and warm summers, and the semiarid region located in the southeastern part of the country. In contrast to the Mediterranean region, the dry season continues beyond the end of the summer. This climatology affects the country vegetation, where differences can be appreciated among and within seasons.

AEMET reveals that the year 2011 was extremely hot with higher temperatures than the historical average (1971–2000). It was also very dry with 25% less rainfall in the North of Spain; however, spring was more humid than normal, particularly in March. The autumn rainfall was 10% lower than usual. The meteorological information drawn from the CRU TS3.10 data is summarized in Figure 4.8. On the left panel, monthly average temperatures are shown, and, on the right panel, the

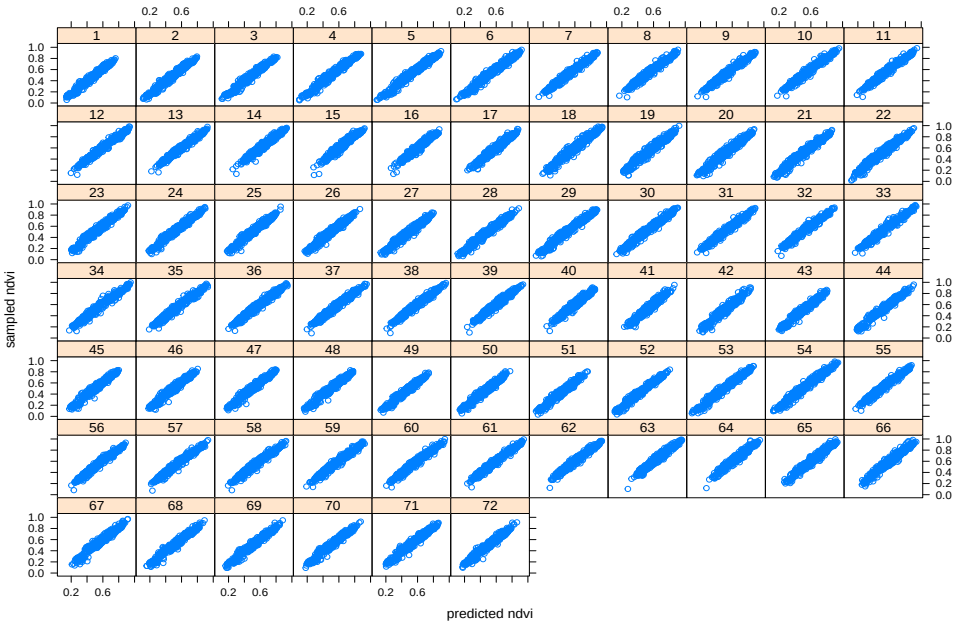

Figure 4.6: Sampled NDVI vs. predicted NDVI data of the 561 locations in the 72 periods from October 2011 to December 2013.

corresponding monthly average rainfall is given. Different colors are used for the different years and the historical mean is plotted in black in both panels. In the spring of 2011, high temperatures and abundant rainfall were also reported, yet the autumn was also very dry. Figure 4.7 shows the NDVI monthly Spanish predictions obtained by averaging the bi-monthly predictions given by the state-space model. In 2011, low values of NDVI are estimated in autumn but have very high values in spring, in agreement with AEMET and CRU TS3.10 data. The year 2012 was also very hot, especially in summer, and rainfall was 15% less than usual, except for autumn, and the region of Galicia, located in the northwest of Spain, which was extremely humid. These features are also observed in Figure 4.7, where a blue color is observed in December 2012 in Galicia, a brown color predominates in the main plateau of Spain, and northern regions show high values of NDVI, particularly in spring.

The year 2013 was hot, but not as hot as 2011 and 2012. January and February were 30% more humid than normal, and March was extremely humid, with more than 340% more rain than the normal average. However, December was very dry. In Figure 4.8, CRU TS3.10 data also show a big pick of rainfall in winter that correspond to high values of smoothed NDVI in spring.

In summary, smoothed NDVI reveals a clear seasonality that intensifies the effect of spring vegetation in 2011, 2012 and 2013, where a higher level of rainfall

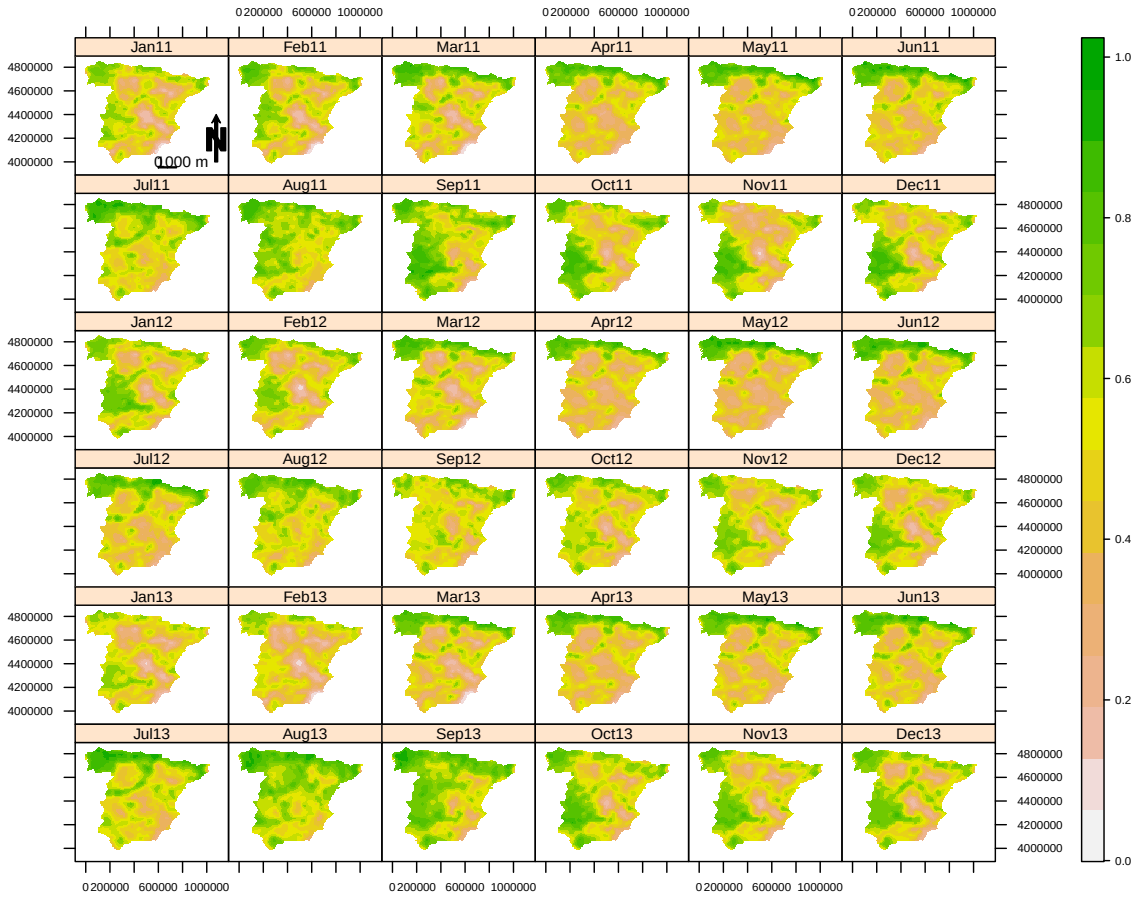

Figure 4.7: Smoothed NDVI in continental Spain from January 2011 to December 2013.

than average is documented. The images preserve the pattern of the original ones but reduce the larger values of NDVI. As expected, the northern regions of Spain maintain higher values around 0.8 and 0.9, mainly in spring and early summer when temperatures and rainfall are more intense. Mountainous regions are also prone to the highest values, and the main plateau reaches values between 0.3 and 0.5, indicating the presence of bare soils or sparse vegetation. Therefore, the smoothed NDVI obtained through the state-space model is close to the climatological real scenario given in Spain between 2011 and 2013. Overall, smoothed images are more sensitive to seasonal and specific meteorological changes than the original ones.

Checking the performance of smoothed NDVI with real data is a difficult task because NDVI is only estimated through satellite images. In this regard, comparisons of the mean estimated surfaces in four categories of NDVI are presented in Table 4.3: ndvi1 for data less than or equal to **.2**, ndvi2 for data greater than **.2** and less than

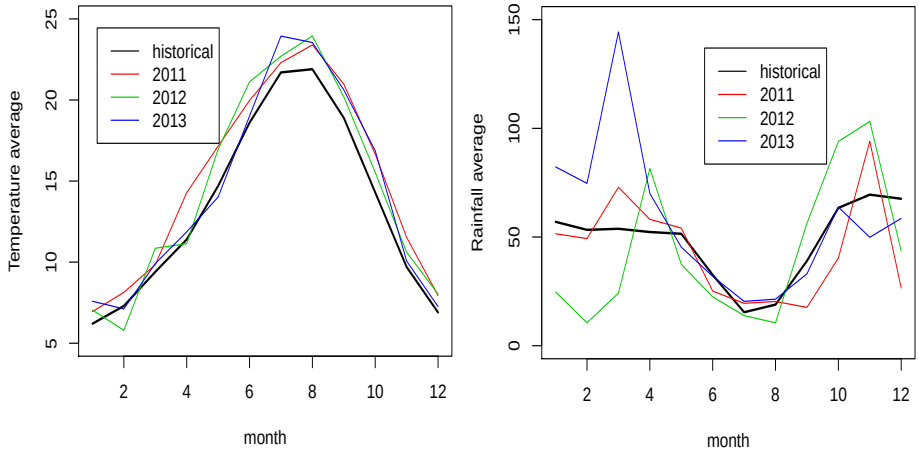

Figure 4.8: Left: average temperatures in °C and Right: average Rainfall in mm on the sampled locations jointly with the historical data.

or equal to **.5**, ndvi3 for data greater than .5 and less then or equal to .7, and ndvi4 for data greater than .7. The mean total surfaces have been calculated with the raw GIMMS NDVI3g images, the state-space smoothed NDVI values, and three versions of the TIMESAT smoothed NDVI values from 2011–2013. The smoothing effect of the state-space model is mainly shown in both ndv1 and ndvi4 categories where smoothed NDVI mean total surfaces are lower than raw averages. These reductions have been added to the ndvi2 and ndvi3 categories. The three TIMESAT versions behave likewise providing close values to those obtained with the original images in both ndvi2 and ndvi3 categories, but important differences are found in the rest of the categories. Figure 4.9 shows the monthly mean surfaces of the raw GIMMS NDVI3g data, and the three smoothing versions: the state-space and two versions of TIMSESAT, the Savitzky–Golay filtering and the Gaussian filtering data by years. The double logistic smoothing version of TIMESAT has been omitted because it is equal to the Gaussian version. The state-space approach follows the same pattern as the original data, but we can see how the ndvi1 category is smoothed mainly in winter and the ndvi4 category in spring and winter. The TIMESAT smoothing versions do not preserve well the pattern of the raw data in the smallest category, and bigger differences than with the state-space procedure can be found, particularly in the first category. In summary, the state-space approach preserves the monthly pattern of raw data by years and smooths mainly the lowest and upper categories. Additionally, this approach incorporates external information coming from CRU TS3.10 meteorological data and agrees with the information provided by the Spanish National Agency of Meteorology.

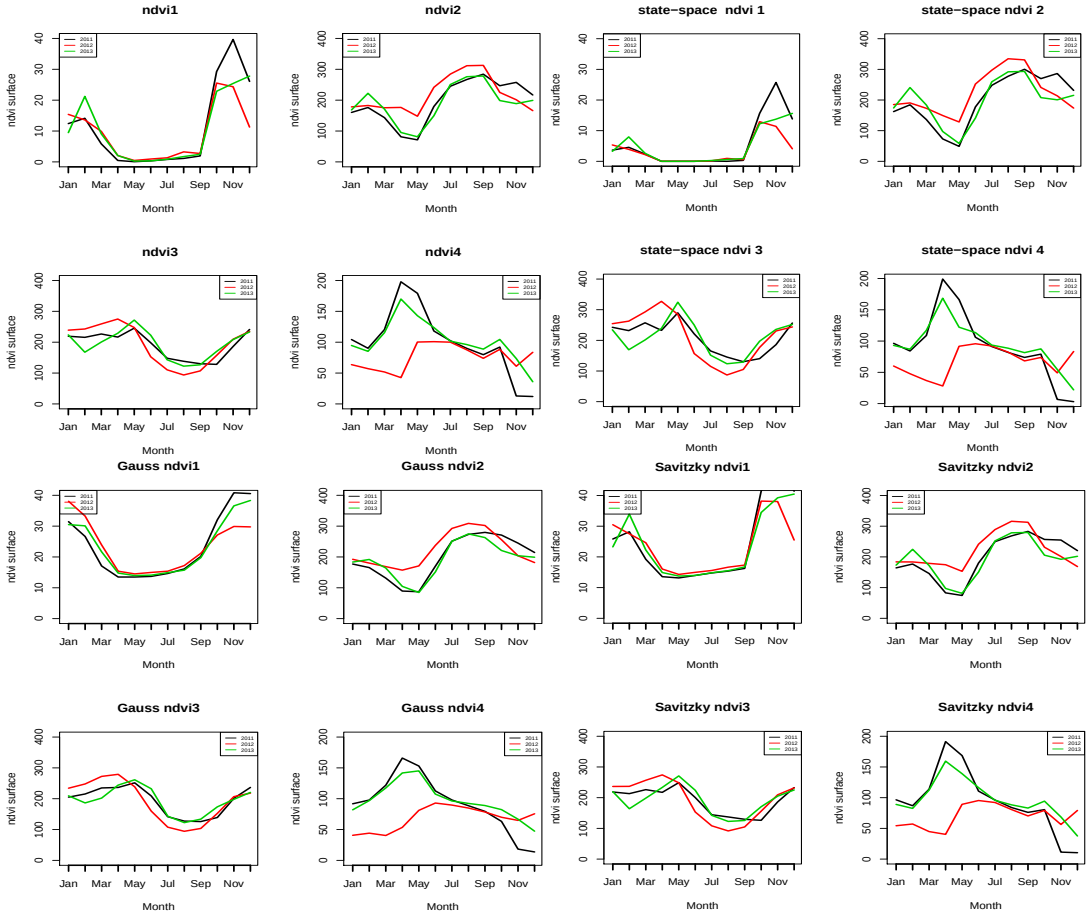

Figure 4.9: Monthly mean surfaces in the four NDVI categories with raw GIMMS NDVI3g data on the upper left, state-space smoothing on the upper right, Gaussian TIMESAT smoothing on the bottom left, and Savitzky–Golay smoothing on the bottom right.

4.5 Conclusions

This chapter is focused on showing that a stochastic spatio-temporal model is a useful tool for overcoming random or fluctuations often present in satellite images, and these fluctuations could interfere with the detection of the NDVI trend changes. The main aim of this work is to show the importance of considering both the spatial and temporal dependence for analysing and smoothing NDVI data. The stochastic spatio-temporal model used here is a useful tool to capture space and time variability for simultaneously smoothing images. Smoothed images have been compared with TIMESAT that only uses temporal dependence. The state-space method outperforms this alternative, as it is able to reduce the most extreme values preserving the original

Table 4.3: Mean total surfaces of four NDVI categories in Spain between 2011 and 2013 in thousands of square kilometers.

	ndvi1	ndvi2	ndvi3	ndvi4
Raw GIMMS NDVI3g	10.88	203.08	195.42	95.42
State-space smoothed NDVI	4.58	206.36	209.57	84.29
TIMESAT Savitzky smoothed NDVI	23.83	202.71	191.32	86.94
TIMESAT Gaussian smoothed NDVI	23.30	202.57	193.31	85.62
TIMESAT double smoothed NDVI	23.30	202.57	193.31	85.62

pattern of raw data. The state-space model also provides the contribution of every covariate to predict NDVI. In this regard, it agrees with other studies such as Wang et al. (2001); Ichii et al. (2002); Schultz and Halpert (1993); Potter and Brooks (1998), where it is shown that, among climatic factors, precipitation and temperature influence both temporal and spatial patterns of NDVI. We have shown that smoothed images are more sensitive to seasonal and specific meteorological changes than the original ones, yet they follow a similar pattern. Moreover, the smoothing method used in this paper provides a calibration method of satellite images with real data. A higher spatial and temporal resolution jointly with auxiliary data at the same resolution level could improve the model performance, but the computational cost will also increase. We are currently working on reducing the computational cost while obtaining accurate predictions in a sensible time.

The actual resolution of 8 km at the equator is an attractive feature for monitoring changes of vegetation at any scale. Unfortunately, this resolution is not enough to warrant high precision images at smaller scales because images have been pre-processed, and, likely, there is also an important ocean border effect, as in the case of Spain. The Maximum Value Compositing (MVC) algorithm used to suppress atmospheric effects also minimizes significant problems associated with short-wave passive remote sensing of the Earth's surface, but the MVC technique itself has generated a second level of problems that must be addressed for proper interpretation of the NDVI MVC images. These are radiometric effects, which are relevant to the stratification assumption, and engineering effects, which are relevant to the MVC technique Holben (1986). Similar situations can also be found with other NDVI global scenes coming from Terra MODIS or SPOT VGT (see, for example, Fensholt et al. (2009), where evaluation of long trends vegetation coming from these satellites is made, revealing differences among them). High bias can also be found when using MVC in mountain regions (see Fontana et al., 2009). Therefore, when down-scaling global scenes to country levels, as in the case of GIMMS NDVI3g in Spain, an adequate smoothing of NDVI data is needed for a proper interpretation of the

spatio-temporal NDVI distribution. Similar situations can be found with other image processing techniques that may require smoothing procedures to analyse the data properly.

The contents of this chapter have been published in the journal *Remote Sensing*: Militino, A. F., Ugarte, M. D., and Pérez-Goya, U. (2017). Stochastic spatio-temporal models for analysing ndvi distribution of gimms ndvi3g images. *Remote Sensing*, 9(1):76.

5

Interpolation of the mean anomalies for cloud-filling satellite imagery

5.1 Introduction

Removing clouds from satellite imagery is an important and crucial task for reconstructing the history and evolution of many remote sensing data. While very cloudy images must be dropped from time series, missing or distorted data in images that are only partially clouded can be filled using series of multi-temporal images. Several procedures have been recently introduced Meng et al. (2009), Zhu et al. (2012), Hermosilla et al. (2015), Roy et al. (2008), Chen et al. (2017), but some of the most popular, e.g., Timesat Eklundh and Jönsson (2012) based on filtering, Hants Verhoef et al. (1996); Roerink et al. (2000) based on harmonic analysis of time series, and Gapfill Gerber et al. (2018, 2016) based on a specific ordering of images and quantile regression, provide free access to users and are easy to run. However, these gap-filling techniques are neither simple nor straightforward.

This chapter proposes a new method for filling gaps or lost data in satellite images, the method is called Image Mean Anomaly (IMA). The method is based on creating a temporal window or neighbourhood around a target image to select similar images. The neighbourhood frame is defined as the set of previous and subsequent images in time periods and years, accommodating the temporal dependence between near images, in the same or different years. The similar images are averaged generating an average image. In this way, the target image can be expressed as the sum of the average image and a residual image, which we call anomaly. the method interpolates the averaged anomalies in the original resolution over the study region with thin-plate splines (Tps). These interpolated anomalies are added to the mean image to fill in the gaps. The flowchart in Figure 5.1 summarizes the IMA process for one image.

Several studies have shown the benefits of using geostatistical methods, over

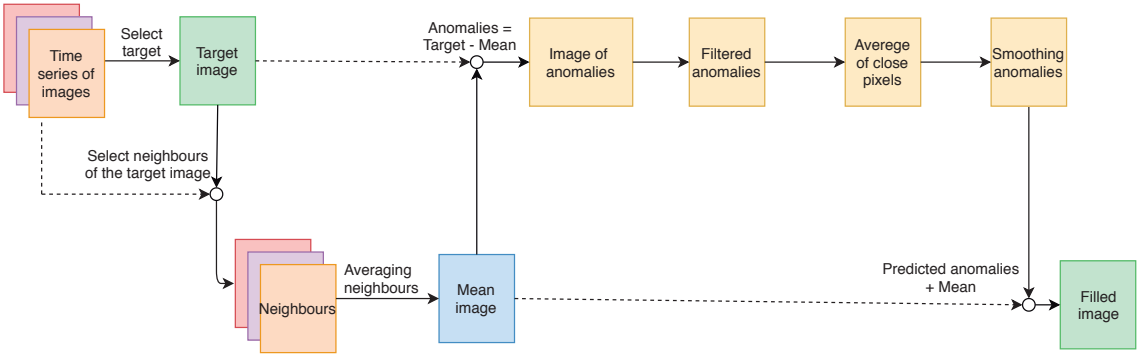

Figure 5.1: Flowchart of IMA for processing one image.

mathematical ones, for filling gaps in satellite imagery. Geostatistical methods explicitly assume that the stochastic spatial dependence inherent in spatial data decreases with distance. For example, Addink (1999) made a comparison of the conventional mathematical and the geostatistical methods to replace clouded pixels in NOAA-AVHRR images. The conventional method is maximum value composition (MVC) and the geostatistical methods are kriging and co-kriging. The geostatistical methods performed better than those ignoring stochastic spatial dependence, though co-kriging methods have computing restrictions when applied to large datasets. Some years later, another study compared noise-reduction NDVI model-based methods Hird and McDermid (2009). In this case, the authors showed that the double logistic and asymmetric Gaussian function-fitting methods of Timesat outperformed the other four alternatives: 4253H Velleman (1980), twice filter Velleman (1980), mean-value iteration filter Ma and Veroustraete (2006) and ARMD3-ARMA5 filter Filipova-Racheva and Hall-Beyer (2000).

The Chapter is organized as follows. Section 5.2 presents the data used in this work and coming from Modis. These data have been downloaded with the applications presented in Chapter 1. The section explains the selected variables, that are, NDVI, day LST and night LST. Section 5.3 provides the explanation of the new IMA method, and a summary of some popular free access alternatives: Hants, TIMESAT, and Gapfill. It includes several subsections for describing its main features. Section 5.4 presents the simulation study uses to check the new method that smooths almost 2500 images for each compared method, giving this quantity of images robustness to the simulation study. Finally Section 5.5 ends with the conclusions.

5.2 Data

Frequently, many raw satellite images are almost unusable, because atmospheric disturbances and electronic radiation from the satellites can distort, blur, or degrade the information. The Moderate Resolution Imaging Spectroradiometer (MODIS) (NASA, 2018) provides time series of pre-processed series of images, where these effects have been mitigated. Additionally, image transformations are made to provide very popular remote sensing data, such as NDVI or LST to the users. These time series of images are already enhanced by composing pre-processed images every 8 or 16 days, and they are available for free. We retrieved remote sensing data from MODIS instrument in TERRA and AQUA satellites. NDVI images are of 16-day temporal resolution, then only 23 images are available each year in TERRA, and the same number in AQUA. To adjust the balance with the same number of images every month, we get 23 images from Version 5-MYD13A2 (AQUA), and we retrieve an additional image from Version-5 MOD13A2 (TERRA) in November, yet this is not a prerequisite for running IMA.

The NDVI reflects vegetation vigour and it is closely related to the amount of photosynthetically absorbed active radiation as indicated in Slayback et al. (2003) and Tucker et al. (2005). It is calculated through the radiometric information obtained for the red (R) and near-infrared (NIR) wavelengths of the electromagnetic spectrum. Then, the index is defined as $NDVI = ((NIR) - R)/((NIR) + R)$ Rouse Jr et al. (1974), and takes values between 0 and 1 with high variability van Wijk and Williams (2005).

Table 5.1: Remote sensing data (Data), climatological season (CS), coefficient of variation (CV), minimum, quartiles and maxima of the day and night LST, and NDVI by climatological seasons in the Navarre tile (Spain), during 2011-2013.

Data	CS	CV	Min	1st Q.	Med	3rdQ.	Max
LST Day	DJF	0.015	252.2	279	281.3	283.5	297.5
	MAM	0.020	254.3	288.8	292.9	296.4	314.9
	JJA	0.019	265.9	298.0	302.0	307.0	320.4
	SON	0.028	251.2	285.5	290.7	297.5	314.0
LST Night	DJF	0.012	247.2	272.4	274.3	275.9	288.5
	MAM	0.015	249.2	277.3	280.3	283	293.3
	JJA	0.012	264.0	285.9	288.4	290.6	297.7
	SON	0.018	256.9	278.6	282.1	285.5	294.6
NDVI	DJF	0.33	0	0.36	0.48	0.60	0.95
	MAM	0.29	0	0.48	0.60	0.70	0.95
	JJA	0.40	0	0.36	0.58	0.78	0.98
	SON	0.39	0	0.34	0.52	0.70	0.98

The LST images of MODIS are derived from the two thermal infrared (TIR) band channels, 31 (10.78-11.28 μm) and 32 (11.77-12.27 μm) Benali et al. (2012). The atmospheric effects are corrected with the split-window algorithm Wan and Dozier (1996), Wan et al. (2002). The algorithm also uses the MODIS Land Cover product (MOD12C1) for correcting the emissivity effects. Composite LST every eight days are downloaded from Version-5 MOD11A2 and they correspond to the eight days average LSTs of the Version-5 MOD11A1 product. In all variables, we cropped the $H17$-$V4$ MODIS tile containing Navarre, to fit the study region. This region consists of a 156×145 (22620 pixels) rectangular array, where each pixel corresponds to 1 km^2 (see Figure 5.2).

Table 5.1 shows the study variables, the climatological seasons (winter (DJF), spring (MAM), summer (JJA) and fall (SON)), the coefficient of variation, the minimum, the first, second, and third quartiles, and the maximum of the daytime LST, nighttime LST and NDVI variables in the study region from 2011-2013. LST day and night are given in Kelvin degrees, and NDVI has no units with a restricted range between 0 and 0.98. NDVI has greater variability than LST in the four climatological seasons.

5.3 Cloud-filling methods

5.3.1 Interpolation of the mean anomalies method (IMA)

We assume that the target image is an LST image, named LST_day_2011.073, which corresponds to the 8-day composite image of March 13, 2011, over Navarre, Spain (see Figure 5.2), although any other time period or variable can also be chosen.

This image is represented by the vector $\mathbf{z}_{\mathbf{s}t_0} = \{z_{s_i t_0} | i = 1, \ldots, m\}$, where $z_{s_i t_0}$ is the remote sensing data observed at location s_i, $s_i \in \mathbf{s} = (s_1, \ldots, s_m)$; $m = 22,620$ is the total number of pixels in the image, and t_0 is the target time period. Note that m can include pixels of missing data.

The IMA method consists of the next seven steps.

1) Define the neighbourhood of the target image. The neighbourhood $\{\mathbf{z}_{\mathbf{s}t_k} | t_k = 1, \ldots, T_0\}$ of the target image $\mathbf{z}_{\mathbf{s}t_0}$, consists of the target image, and the preceding and following images in that year, as well as images from those dates during the previous and subsequent years. If the image belongs to the first or last year of the study period, we can choose more subsequent or preceding years, as it is done in this example. Here, the time series are made of eight-day composite images of daytime LST data collected from 2011 to 2013. Therefore, the neighbourhood of the target image, including this image has $T_0 = 9$ images, which are identified by the year and Julian day on which they were pre-processed: $I_1 = 2011.065, I_2 = 2011.073, I_3 = 2011.081, I_4 = 2012.065, I_5 = 2012.073, I_6 = 2012.081, I_7 = 2013.065, I_8 = 2013.073,$

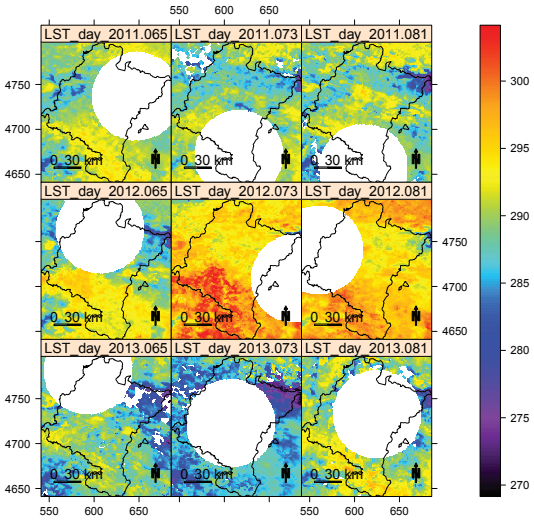

Figure 5.2: Example of the neighbourhood of the target image LST_day 2011_073 (color bar units in Kelvin degrees) used in the IMA and Gapfill methods, where random gaps of size G have been introduced into every image of the neighbourhood. The target image corresponds to the 13th of March 2011.

and $I_9 = 2013.081$. See Figure 5.2 for details. We use the same neighbourhood whether the target image is $I_2 = 2011.073, I_5 = 2012.073$ or $I_8 = 2013.073$ since this study only includes data from 2011-2013.

The size and dimension of this neighbour can be enlarged according to the availability, the quality of satellites images, and the repetitive cloudiness.

2) Compute the mean target image of the neighbourhood. We assign to each pixel, the mean of the non empty pixels at the same location of the neighbour images. Thus, the mean target image is given by $\bar{\mathbf{z}}_{st_0} = \{\bar{z}_{s_i t_0} | i = 1, \ldots, m\}$, where

$$\bar{z}_{s_i t_0} = \frac{\sum_{t_k=1}^{T_0} z_{s_i t_k}}{T_0}, \tag{5.1}$$

$z_{s_i t_k}$ is the ith observed pixel of the t_k period in the neighbourhood of the target image.

3) Estimate the anomalies of the target image by subtracting the mean image from the target image. In other words, $\mathbf{w}_{st_0} = \mathbf{z}_{st_0} - \bar{\mathbf{z}}_{st_0}$ is the target image of the anomalies, and $\mathbf{w}_{st_j} = \{w_{s_i t_j} | i = 1, \ldots, m\}$, where

$$w_{s_i t_j} = z_{s_i t_j} - \bar{z}_{s_i t_0}, \quad \text{for} \quad i = 1, \ldots, m. \tag{5.2}$$

4) Filter the anomalies. The target anomalies are filtered out by removing the upper and lower 5% of the extreme values, i.e., percentiles $(p_{0.95})$ and $(p_{0.05})$ respectively. Removing extreme anomalies prevents from distorted data still present in some images. This step follows the expression

$$w'_{s_i t_0} = \begin{cases} w_{s_i t_0}, & \text{if} \quad p_{0.05}(w_{s_i t_0}) < (w_{s_i t_0}) < p_{0.95}(w_{s_i t_0}) \\ \text{non value}, & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, m_1$. Filtering the target anomalies reduces the maximum total number of pixels in the tile from $m = 22,620$ to $m_1 = 20,358$. Alternative threshold values of filtering may be used depending on the quality of the input images, yet these percentiles are recommended.

5) Average the anomalies over a predefined window. The anomalies are averaged over their neighbouring pixels in a window of 5×5 pixels, when the spatial resolution is 1 km^2. We assign the mean of the non-empty anomalies in the same window to all the pixels of the window. Unless all the anomalies in the same window have missing data we will reduce the number of missing pixels. There are three benefits gained by performing this step: we avoid sudden changes among close pixels, we reduce the number of empty anomalies, and we reduce the number of the equations to be solved when using Tps. Therefore, a good trade-off between computing time and prediction error is achieved. Removing sudden changes among close pixels is especially important in NDVI images, because these remote sensing data are very sensitive to small changes in vegetation, sensor calibration, and atmospheric correction. Alternative shrinking factors can also be used, depending on the image resolution and computing capacities. Here, we have chosen a factor equal to 5 because the spatial dependence in these variables is about 5 km^2, yet it can be changed. Then, the averaged anomalies are defined as

$$w''_{s_i t_0} = \frac{\sum_{s_i \in T_1} w'_{s_i t_0}}{25} \tag{5.3}$$

where $i = 1, \ldots, n$, T_1 is a 5×5 pixel window around s_i. This step further reduces the m_1 maximum number of pixels in the target image to $n = m_1/25 \approx 928$ pixels.

6) Interpolate the averaged anomalies. We choose bivariate thin-plate splines to interpolate the target image of the averaged anomalies because of its well known properties Wood (2003), yet other alternatives can be used Li and Heap (2011).

The thin-plate spline model provides a flexible relationship between the anomalies $\mathbf{w}''_{st_0} = (w''_{s_1t_0}, \ldots, w''_{s_nt_0})$, and the the planar coordinates $(\mathbf{x_s}, \mathbf{y_s})$. Predictions are given by $\hat{\mathbf{w}}''_{st_0} = \sum_{j=1}^{3} a_j p_j(s_i) + \sum_{i=1}^{n} b_i \phi(d)$, where $\phi(d) = d^2 \log(d)$ is a basis function, d is the Euclidean distance between the prediction location s_0, and each data location s_i, and $p_1(s_i) = 1, p_2(s_i) = x, p_3(s_i) = y$. The weights $\{a_j | j = 1, 2, 3\}$ and $\{b_i | i = 1, \ldots, n\}$ are estimated by solving a linear system of order n Luo et al. (2008); Duchon (1977); Boer et al. (2001). The predictions are obtained over the m pixels of the target image. This process is nowadays programmed in mathematical, remote sensing, and statistical software in a very efficient way. Here, we use the R package fields D. Nychka et al. (2015), where uncertainty measures can also be derived.

7) Add the interpolated anomalies to the mean image. Thus, the final predicted image is

$$\hat{\mathbf{z}}_{st_0} = \bar{\mathbf{z}}_{st_0} + \hat{\mathbf{w}}''_{st_0}. \tag{5.4}$$

Finally, we programmed IMA in R; and the code for running IMA is available from the authors.

Table 5.2: Number of filled images for the three remote sensing data (LST day, LST night and NDVI), time periods, years, cloud sizes, and methods (Hants, 3 Timesat, Gapfill, and IMA) used in the simulation study.

	LST Day	LST Night	NDVI
Time periods by year	46	46	24
Years	3	3	3
Cloud Sizes	7	7	7
Number of methods	6	6	6
Total	5,796	5,796	3,024
	Total of 14,616 filled images		

5.3.2 Hants

The Harmonic Analysis of Time Series (Hants) was a procedure originally developed for processing time series of noisy remote sensing data Verhoef et al. (1996), and a few years later the Hants algorithm was published Roerink et al. (2000). The performance of this algorithm has been studied with applications to leaf area index (LAI), land surface temperature (LST), and the polarization difference brightness temperature (PDBT) Zhou et al. (2015). The application was released as plug-in for the geographical information system (GIS) platform called the Geographic

Resources Analysis Support System (GRASS)Neteler and Mitasova (2013). The Hants algorithm uses an iterative procedure to fit a curve based on pixel-wise time series separately, but ignores stochastic spatial dependence. The process follows the steps:

a) checking the time series and flag samples outside the valid range of data,

b) fitting the remaining valid samples of the series by several prescribed harmonic components,

c) if the maximum signed bias between the fitted series and the valid samples is larger than a user defined threshold, and the number of the remaining samples exceeds the minimum number of samples necessary for the reconstruction process, then it rejects the samples with bias larger than half of the maximum bias and return to step b). Otherwise, stop the processing.

5.3.3 Timesat

Timesat (Eklundh and Jönsson, 2012) is a software released in 2002 and coded in Matlab and Fortran, yet those are not necessary for running it. It implements three processing methods based on least-squares fits for satellite time series processing: Savitzky-Golay (SG) filtering, double logistic (DL) and asymmetric Gaussian (AG).

The software was designed to analyze satellite time series data from satellites by extracting seasonal parameters from smoothed versions of the data, but it works with pixel-by-pixel time series. Timesat needs a specific image binary format that can be obtained using the 'raster' package of the R software (R Core Team, 2019). First, the user must specify the number of rows and columns to be processed (156×145), the type of data (16 bit integer), the range of the variables ([0-10000] for NDVI and [0-999.9] for LST day/night), the lag period length (8 or 15), and the time series length (three years). These configuration input parameters must be provided using a graphical interface. When exporting the smoothed image, a matrix is produced, that must be completed in order to be exported to R; the 'raster' package will help again to convert an *hdr* format into a $Tiff$ format.

5.3.4 Gapfill

Gapfill is the specific function of the 'gapfill' R package (Gerber et al., 2018) that fills missing values of satellite data with the Gapfill method. The method ranks the images preceding and following the target image, and predicts the gap using quantile regression. This regression is an extension of the classical estimation of the conditional mean model to conditional quantile functions, which can be explained as follows.

Let $\mathbf{p} = \{p_{ij}\}$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ be the jth pixel in the ith image and let $\mathbf{r} = \{r_1, \ldots, r_I\}$ be the ranks or ordinal sequence of these pixels. Then,

instead of estimating the conditional expectation of the response variable $\mathbf{p}$ given the explanatory variable $\mathbf{r}$, i.e., $E[\mathbf{p}|\mathbf{r}]$, as it is done in classical linear regression, the quantile regression estimates the conditional θ quantile given $\mathbf{r}$, i.e., $Q_\theta(\mathbf{p}|\mathbf{r})$. Therefore, the θ quantile regression model is defined as

$$Q_\theta(\mathbf{p}|\mathbf{r}) = \beta_0(\theta) + \beta_1(\theta)\mathbf{r}, \tag{5.5}$$

where $\beta_0(\theta)$ and $\beta_1(\theta)$ are the regression coefficients with a similar interpretation of the classical regression model (Davino et al., 2013). This algorithm was explained and compared with Timesat for NDVI in a very recent paper (Gerber et al., 2018). To run Gapfill in our simulation study, we tune the programming to tackle the challenge of filling big gaps. First, in the *Predict* function we enlarge to 20 the minimum number of non-empty pixels in the target image, originally written as $nTargetImage = 5$. Second, in the *rank* function we add the argument $ties.method = first$ for avoiding ties in assigned ranks. Third, we include the code $if(sum(!is.na(r)) < 2); return(NA)$ to guarantee at least 2 different ranks for running the quantile regression. When many missing pixels are in the same locations of different images in the same neighbourhood, it is not possible to assign ranks.

Gapfill is easily accesible because of its free distribution, and it provides measures of uncertainty through prediction intervals.

Table 5.3: Cloud size, radius, total surface and mean surface percentage of the distorted images with the artificial clouds used in the simulation study for LST day.

Cloud size	Radius (km)	Surface (km^2)	Mean Surface (%)
A	15	706.8	5.5
B	17.5	962.1	6.6
C	20	1256.6	7.6
D	30	2827.4	13.0
E	50	7852.9	28.4
F	60	11309.7	37.1
G	70	15393.8	44.4

5.4 Analysis and results

The performance of IMA is checked with regard to five alternatives: Hants, Gapfill, and three version of Timesat in both a real example, and a simulation study. The

methods are compared in a simulation study using already filled and processed images as a reference. Figure 5.3 shows the flowchart of the simulation study. In the first step we download the time series of the three variables: 72 composite images of normalized difference vegetation index (NDVI), 138 composite images of daytime land surface temperature (LST), and 138 nighttime LST captured from MODIS (NASA, 2018) in the Spanish region of Navarre between 2011 and 2013. Second, we define the seven sizes (A, B, C, D, E, F, G) of the artificial clouds randomly introduced to each image in the time series. Third, we run the five aforementioned alternatives of cloud-filling methods and IMA. Fourth, we calculate the root mean squared prediction error (RMSE) for each gap, and finally, we average the RMSE by year and method in each remote sensing data, and we explain the conclusions using a collection of plots and tables. The study with almost 15,000 processed images shows that on average IMA obtains the best predictions in all the simulations. In addition IMA improves the runtime to the method with the second best predictions.

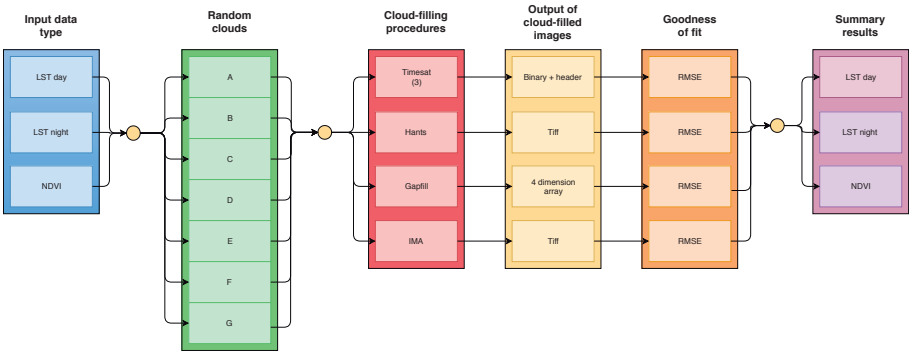

Figure 5.3: Flowchart for the simulation study.

The performance of the proposed IMA method for filling clouds in satellite imagery is evaluated in both a real example and a simulation study using 348 images captured over Navarre, Spain, between 2011 and 2013. In the simulation study, IMA is compared with Hants, the three versions of Timesat, and Gapfill. Time series of daytime LST (LST Day), nighttime LST (LST Night), and NDVI are analyzed. The LST time series each contains 138 composite images of 8-day time periods; the NDVI time series contains 72 composite images of 16-day time periods. In the images of these time series, we introduced artificial clouds of seven different sizes (A, B, C, D, E, F, G) to generate missing data inside a randomly located circle. Figure 5.2 shows an example of randomly introduced size-G clouds. For each of the images of the time series and variables, we run the six cloud-filling methods mentioned above: Hants, three versions of Timesat, Gapfill, and IMA. Table 5.2 shows the distribution of the 14,616 images used in the simulation study according to the derived variables, the cloud sizes, and the methods. Table 5.3 shows the

cloud size, the radius, the total surface (calculated as $\pi \times r^2$), and the mean surface percentage of the artificial cloud coverage. The mean surface percentage varies from 6% to 44% depending on the size of the cloud, and how much of the cloud is located inside the tile.

We evaluate the performance of the methods for each size of artificial cloud, each derived variable, and each model. Pixel-by-pixel square differences between the observed and filled data are averaged for calculating the square root of the mean squared prediction error (RMSE). The expression is given by

$$
\begin{aligned}
RMSE(k,l,p) = & \sqrt{\frac{\sum_{s_i,t_j}(z_{s_i t_j klp} - \hat{z}_{s_i t_j klp})^2}{IT}}, \\
& s_i = 1,\ldots,I \\
& t_j = 1,\ldots,T \\
& k = A,B,C,D,E,F,G \\
& l = \text{LST day, LST night, NDVI, and} \\
& p = \{\text{Gapfill, Hants, Timesat AG,} \\
& \text{Timesat DL, Timesat SG and IMA}\},
\end{aligned}
\tag{5.6}
$$

where $z_{s_i t_j}$ and $\hat{z}_{s_i t_j}$ are respectively the original and predicted values of the remote sensing data, I is the number of pixels inside the cloud gap, T is the number of images, k is the type of cloud, l is the derived variable, and p is the smoothing procedure.

The $RMSE(k,l,p)$ for the six cloud-filling methods, the seven sizes of artificial clouds, and the three derived variables is shown in Figure 5.5. The top-left plot in Figure 5.5 exhibits the LST day RMSE. Different lines correspond to different cloud-filling methods. Pink and black colours are for Gapfill and IMA respectively, while the others colours correspond to Timesat and Hants. In this Figure, Hants and the three Timesat versions show similar RMSE values for the small and moderate cloud sizes, though Hants gives the highest RMSE values for the big clouds. Both Hants and Timesat consistently produce higher RMSE values than Gapfill or IMA. IMA clearly outperforms Gapfill regardless of cloud sizes. Similar conclusions are drawn for LST Night (on the top-right of Figure 5.5), where Hants exhibits the highest RMSE values, though not just for big clouds, and IMA always shows the lowest RMSE values for all cloud sizes. Figure 5.5 shows the RMSE for NDVI. We observe a similar pattern, but we also note a positive correlation between RMSE and cloud sizes. Hants provides the highest RMSE values for almost all the cloud sizes, and IMA the lowest for all the cloud sizes. RMSE estimates are lower for NDVI than for LST because NDVI is constrained to take values between 0 and 1.

Clearly, Gapfill and IMA outperform the others methods, and therefore, both

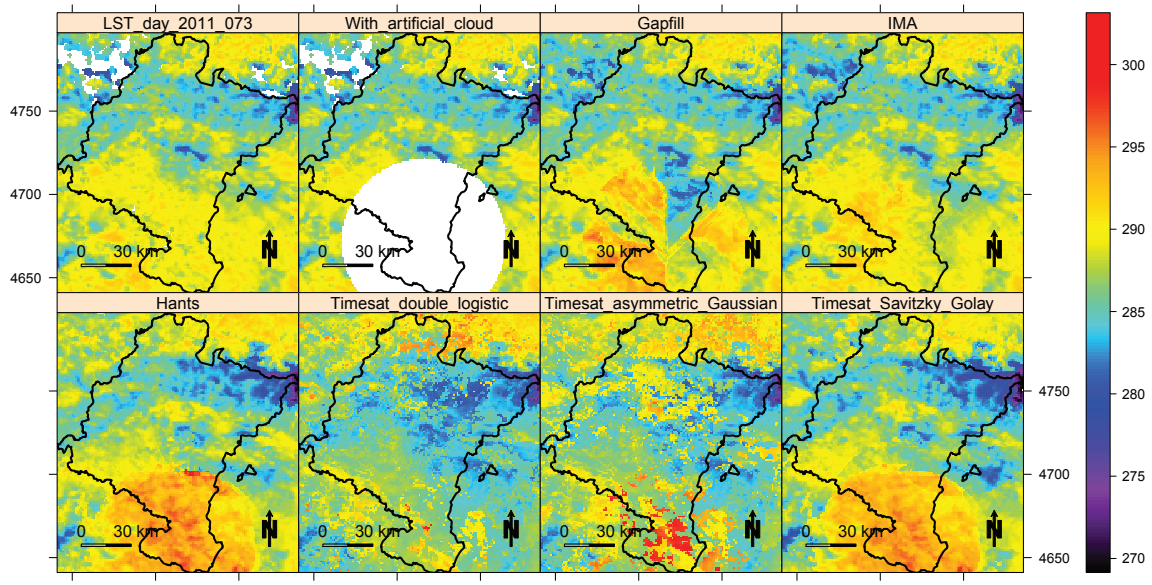

Figure 5.4: LST_2011_073 daytime target image, the target image with artificial cloud, and the reconstructed images with Gapfill, IMA, Hants, Timesat double logistic, Timesat asymmetric Gaussian and Timesat Savitzky-Golay in Kelvin degrees.

Tables 5.4 and 5.5 focus on the RMSE estimates and the percent of reduction. Shrinking percentages are always in favor of IMA, however we see in Table 5.4 that these percentages decrease as cloud size increases, because both methods become less efficient as long as the clouds get bigger. In Table 5.5, we see no such inverse correlation between cloud size and reduction percentage, mainly because NDVI is fairly variable and its values are limited between 0 and 1. Extensive simulation studies not shown here for preserving space reveal that inside a cloud, the pixels closer to the center of the cloud have higher RMSE values than those near the boundary, but the fidelity of the prediction also depends on the similarity of the missing pixels to those from which they borrow information.

Overall, IMA outperforms the three versions of Timesat, Hants, and Gapfill regardless of the cloud size for all three variables, LST Day, LST Night, and NDVI. Gapfill is the closest competitor. Figure 5.4 illustrates the filling processes of the six methods when filling the target image LST_2011_073. Coordinates are given in UTM scaled to km. In this example, IMA provides the best filling.

5.4.1 Running IMA and Gapfill procedures with real data

Additionally, we also illustrate the IMA procedure in real data by filling clouds in the LST day and NDVI daily target images of the 16th (2012198), 17th (2012199),

Table 5.4: Root Mean Squared Prediction Error of Gapfill (GF) and IMA, and Reduction Percentage obtained from the simulation studies of LST day and LST night.

	RMSE LST Day			RMSE LST Night		
	GF	IMA	Reduction (%)	GF	IMA	Reduction (%)
A	1.80	1.41	21.7	1.46	1.22	16.0
B	1.89	1.45	23.2	1.41	1.21	14.6
C	1.93	1.52	21.1	1.57	1.31	16.7
D	2.07	1.64	20.8	1.69	1.38	17.9
E	2.43	2.06	15.5	1.96	1.70	13.4
F	2.59	2.24	13.3	2.04	1.86	8.8
G	2.81	2.48	11.9	2.20	1.99	9.5

Table 5.5: Root Mean Squared Prediction Error of Gapfill (GF) and IMA, and Reduction Percentage obtained in the simulation study of NDVI.

	RMSE NDVI		
	GF	IMA	Reduction (%)
A	0.051	0.046	8.5
B	0.051	0.048	7.0
C	0.050	0.048	4.4
D	0.063	0.055	12.4
E	0.067	0.060	11.3
F	0.075	0.066	11.7
G	0.082	0.073	11.0

and 18th (2012200) of July 2012 in Navarre, Spain. Those days were in principle free of clouds, but we mimic cloudy days adding a real cloud mask to the target images.

Daily images are more variable than composite images, because they are only slightly pre-processed. Therefore, for a robust estimation of the anomalies we need to increase the neighbourhood size of the target images. Now, each target image has a neighbourhood of 3×7 images corresponding to the 3 previous and 3 subsequent images of the same year, and the corresponding images from those dates during the previous and subsequent years. The 3 target images are consecutive, then the neighbourhood is made up of 27 images, and in the end, all the images are filled. The LST day images are daily images retrieved from MOD11A1 (TERRA) Version 5 with a spatial resolution of 1 km^2. The NDVI images are defined using the red and near infrared wavelengths from MOD09GA (TERRA) Version 5 in the same days, because MODIS does not provide NDVI daily images. These images are reprojected for homogeneity reasons to the same resolution of LST day images, because the

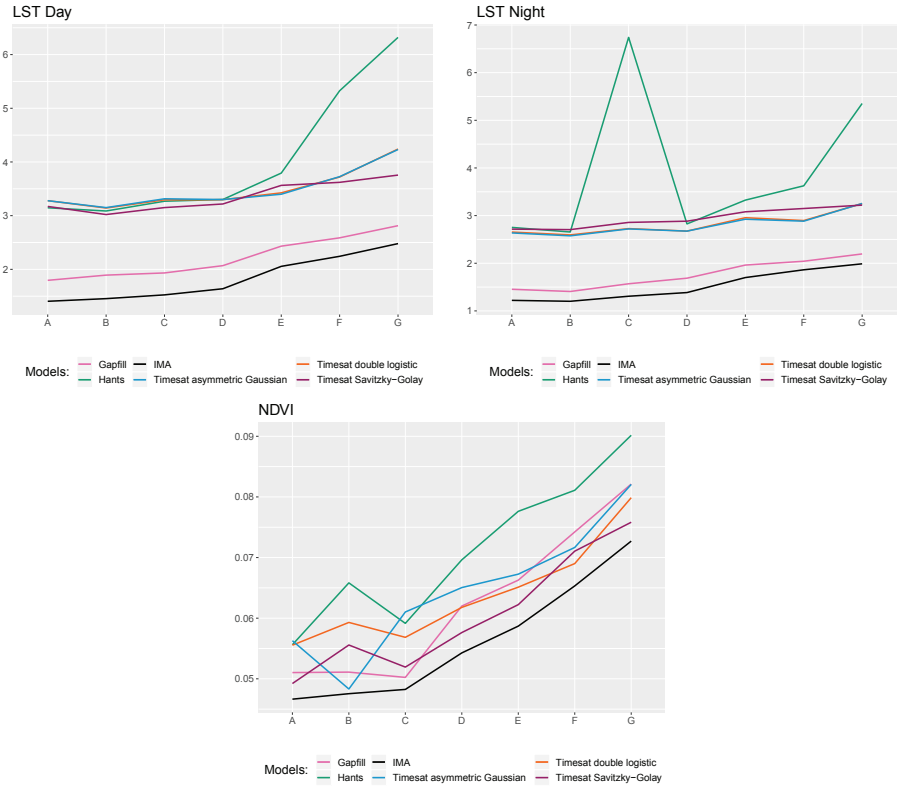

Figure 5.5: Root Mean Squared Prediction Error (RMSE) versus artificial cloud size for the six models compared in the simulation study with LST Day (on the left), LST Night (on the right) and NDVI (on the bottom) images of Navarre, Spain, 2011-2013.

original resolution is .5 km^2, yet this step is not required in the IMA procedure.

The first row of Figure 5.6 shows the observed LST day target images of the 16th (2012198), 17th (2012199) and 18th (2012200) of July 2012. The second row shows the same images masked with real clouds from the 3th, 14th and 21th of July 2011. The third row shows the high fidelity of the IMA predicted images to the original target ones. Table 5.6 summarizes the root mean squared prediction error obtained with IMA and Gapfill in the clouds of the LST day and NDVI target images. Separately and jointly, IMA reduces the root mean squared prediction error estimated by Gapfill in LST day and it is equally competitive than Gapfill in NDVI images, matching the results of the simulation study. The computing time for processing the raster of the three LST day target images in a PC with an Intel Core i7-4790, and 16GB of RAM takes about 4 seconds with IMA, while Gapfill takes 1h30 for filling only the gaps of those images.

For checking the performance of IMA with regard to the aforementioned gap-filling methods, an extensive simulation study involving the filling process of 14,616

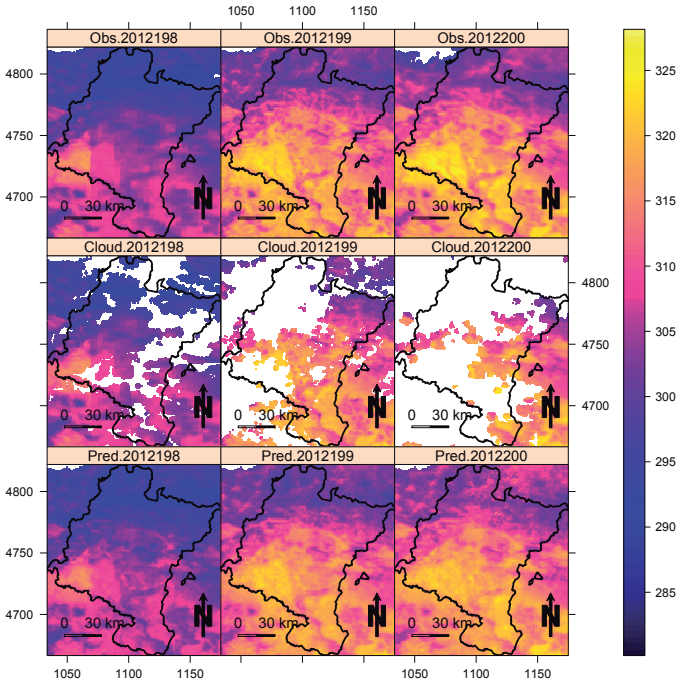

Figure 5.6: The first row shows the observed LST day target images (in Kelvin degrees) of the 16th (2012198), 17th (2012199), and 18th (2012200) of July 2012 in Navarre. The second row shows the same images masked with real clouds borrowed from the 3th, 14th and 21th of July 2011 respectively, and the third row shows the IMA predicted images.

LST day and NDVI images in different scenarios was conducted. Every gap-filling method is based on different models and assumptions, becoming difficult a theoretical comparison. Therefore, the evaluation of the RMSE is proposed. The RMSE is an indicator of the gap-filling fidelity because it has the same measurement units than the original variable, and it is very intuitive. Nevertheless, it can be calculated only when we know the ground-truth data, because then we can subtract the ground-truth data from the prediction. Simulation procedures make it possible, because we know real data. Furthermore, a data set of 3 target LST day and NDVI images using a neighbourhood of 27 images was created for illustrating the performance of IMA with regard to Gapfill algorithm. Overall, both the simulation study and the real data proved the outperformance of the IMA method.

As for computing time, Table 5.7 shows the running times required for processing 5,796 images. They correspond to 138 different LST Day images processed with all methods, and different sizes of artificial clouds on a Windows PC with an Intel

Table 5.6: Root Mean Squared Prediction Error in the cloud set of the LST day and NDVI images of Navarre of the 16th (2011198), 17th (2011199) and 18th (2011199) of July 2012 obtained with Gapfill (GF) and IMA.

	Filling cloud	RMSE LST day		RMSE NDVI	
Julian Day	Surface km^2	GF	IMA	GF	IMA
2011198	7332.53	1.58	1.28	0.07	0.07
2011199	12081.15	4.78	1.99	0.08	0.08
2011200	16877.81	4.87	2.03	0.10	0.11
sample mean	12097.16	3.74	1.77	0.08	0.09

Core i7-4790, and 16GB of RAM. Hants and Timesat are really fast, but Gapfill is slowed-down by its ranking process. The IMA method maintains a constant running time for both small and large gaps in the data, and is faster than Gapfill when processing moderate or large gaps. Though it is not as fast as Timesat or Hants, IMA processes the target images in less than 14 seconds on average, regardless of the gap size.

Table 5.7: Running times in minutes (m) and hours (h) when processing 138 LST Day time series of 1 km^2 resolution images in Navarre with Hants, the tree versions of Timesat, Gapfill and IMA.

		Hants	Timesat			Gapfill	IMA
			DL	AG	SG		
	Pre-process	1m	2m	2m	2m	1m	0m
	Configuration	5m	10m	10m	10m	0m	0m
Running time	Type A	2m	2m	2m	2m	27m	35m
	Type B	2m	2m	2m	2m	59m	35m
	Type C	2m	2m	2m	2m	1h 3m	34m
	Type D	2m	2m	2m	2m	2h 10m	31m
	Type E	2m	2m	2m	2m	11h 20m	29m
	Type F	2m	2m	2m	2m	20h 2m	26m
	Type G	2m	2m	2m	2m	1d 6h 13m	21m
	Export to TIFF	2m	6m	6m	6m	1m	0m
	Total time	21m	31m	31m	31m	2d 17h 16m	3h 31m

5.5 Conclusions

This Chapter proposes a new gap filling procedure to increase the quality of satellite images called Image Mean Anomaly (IMA). The IMA method assumes that remote

sensing data can be expressed as the sum of a trend plus a random error. The trend is assumed to be constant in the neighbourhood of the target image, and it is estimated with the mean of this neighbourhood, and the residuals or anomalies are the estimates of the random error. Using repeated measurements from the same and contiguous time periods across several years provides a more robust estimation of the mean. Anomalies need to be filtered because images are not always free of altered or distorted data. Shrinking spatial resolution of the filtered anomalies is also a necessary step to reduce the dimension of the equations to be solved in the thin-plate splines, and to mitigate the border effect. After interpolating the averaged anomalies, the new predictions are added to the mean image.

The IMA method is competitive for several reasons: a) it shows a strong agreement between the benchmark image and the filled image; b) it preserves the inherent phenology of the remote sensing data by estimating the mean of the same time periods in different years; c) neither the tuning constant nor any other parameter in the input configuration need to be specified in advance; d) its image-processing runtime is consistent regardless of the data-gap size to be filled; e) it exploits the benefits of the spatio-temporal dependence among time series of images, and f) it is easy to use. The simulation study reveals that IMA outperforms Timesat, Hants, and Gapfill, by reducing the RMSE for all three variables that were tested (LST Day, LST Night, and NDVI). The real case study also exhibits a good performance of IMA versus Gapfill, particularly when filling daily LST day images.

Cloud-filling methods have some limitations. For example, clouds that are found in the same locations across neighbouring images in a systematic and periodic way could hinder efficient filling of gaps. Larger clouds cause estimation methods to lose robustness and become unstable. In those cases, wider neighbourhoods are required in IMA, and likely programming improvements based on distributed programming and parallelizing are also necessary when input images are of high resolution. The independent steps required for running IMA can be easily fitted in the map-reduce theory Leskovec et al. (2014), and used in Hadoop cluster Hadoop (2009). We are currently dealing with these issues.

The contents of this chapter have been published in the journal *IEEE Transactions on Geoscience and Remote Sensing*:
Militino, A. F., Ugarte, M. D., Pérez-Goya, U., and Genton, M. G. (2019). Interpolation of the mean anomalies for cloud filling in land surface temperature and normalized difference vegetation index. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):6068–6078.

6 Using ground-truth data for improving satellite imagery

6.1 Introduction

The presence of clouds, atmospheric absorption, weather effects and sensor-introduced noise can be important causes of distorted and missing data in satellite imagery. Therefore, numerous gap-filling and smoothing techniques have been developed in the past few years. However, even after the composition process, residual noise can still arise in images, and therefore, a smoothing procedure using high-spatio-temporal-resolution ground-truth data is proposed in this work.

This chapter proposes a study on how the use of covariates can improve the quality of satellite images. In the study, the SSM model introduced in chapter 4 is used and is compared with a new smoothing method designed in this work. The proposed method is based on the same premise of IMA where a satellite image is expressed as the sum of an average image and an anomaly. This new smoothing method uses covariates to model the anomaly and thus improve the predictions. The name of the new method is thin-plate splines with covariates (TpsWc), because it uses thin plate splines as a generalization of a regression model wherein the linear relationship between the response variable and the coordinates is substituted by a non-parametric function providing a more flexible fitting Wahba (1990). Under certain conditions, the thin-plate splines are formally equivalent to kriging Hutchinson and Gessler (1994), and because the measure of smoothness is invariant under the rotation of the coordinates, thin-plate splines are specially suited for spatial data. The thin-plate splines have an important advantage with regard to universal kriging because they do not need to fit variograms. Moreover, the estimators of the thin-plane splines can be obtained in a closed form by solving a linear system of n equations, where n is the number of data points.

Both methods, the SSM and the TpsWc are used in two different scenarios to conduct the study. In the first scenario, the methods are executed without covariates and in the second scenario, the rainfall covariates are added. To see the response of the models with different variables, the study is carried out on three types of satellite variables, the NDVI, the day LST and the night LST. The results show that the use of covariates improves the predictions of NDVI and the two LST variables analysed in all the proposed scenarios. The results also show better predictions in the new TpsWc method compared to SSM predictions.

Figure 6.1 shows the flowchart of this work. First, meteorological and remote sensing data are taken. Second, daily ground-truth information needs to be averaged over 8-day and 16-day periods to match the corresponding day and night LST and *NDVI* time periods. These data are interpolated in the study region of Navarre (Spain) (see Figure 6.2) through a thin-plate spline model with altitude as the external covariate. This is a preliminary task because these data are a small discrete set, and greater spatial resolution is recommended for predicting. For the simulation study, remote sensing data are altered with different sizes of distortions, i.e., 5%, 10%, 15% and 20%, and different magnitudes are used according to the derived variables. Third, we fit different models for every outlier outbreak: a state-space model with covariates (SSMWc), a thin-plate spline model with covariates (TpsWc), a state-space model without covariates (SSMWoc) and a thin-plate spline model without covariates (TpsWoc). All of these methods provide multi-temporal smoothed series of the three remote sensing datasets. Finally, to compare model performances, we obtain the square root of the mean squared prediction error, which will be summarized in graphs and tables.

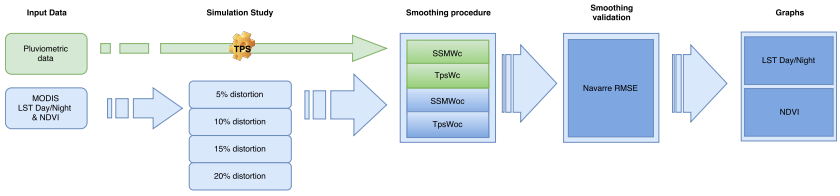

Figure 6.1: Flowchart of the process for evaluating the performance of the smoothing methods: state-space model with covariates (SSMWc), Tps with covariates (TpsWc), state-space model without covariates (SSMWoc) and Tps without covariates (TpsWoc).

The chapter contains the following sections, Section 6.2 explains the Modis satellite images used in the simulation study. The section introduces three time series corresponding to the three different variable types: NDVI, day LST and night LST. The section also describes the data obtained from rainfall stations in Navarra. However, the rainfall data is a set of stations that must be adapted to be compatible with the images. That is why this section presents the processing done on rainfall

data. The section ends with a descriptive analysis of the data. Section 6.3 describes in detail the spatio-temporal procedures used to analyze the usefulness of covariates in the smoothing of satellite images. Special mention is made of the TpsWc method, because it is a new proposal for smoothing satellite images, explaining the necessary adaptations of the data and the model, Section 6.4 explains the proposed simulation study, and presents its results. The results show better predictions in all proposed the scenarios for both models in the version using the covariates compared to the same model without covariates. Finally Section 6.4 presents the conclusions.

6.2 Data

Navarre is a region of approximately 10,000 km^2 located in the north of Spain (see Figure 6.2). Elevations vary between 200 and 2500 meters in the highest zone of the Pyrenees, located in Northeastern Navarre. Valleys and mountains are ubiquitous in the north, and small hills are common in the central part of the province. The northwest of Navarre is humid, but not highly so, and the northeast is a mountainous region with elevations between 1459 m and 2438. The central area is characterized by a temperate Mediterranean climate, with a tendency towards a continental climate. The south is mainly flat; the climate is Mediterranean and continental with dry summers; temperatures exhibit large annual variations; there is minimal and irregular rainfall; and northerly winds are frequent. Clearly, the climate varies across the province, and large weather differences can be experienced on the same day.

In this study, we have drawn classical variables, such as the maximum temperature and humidity, from 48 rain gauge stations. The data are of daily temporal resolution, from which weekly and biweekly mean data are derived. The average distance among rain gauge stations is approximately 15 km. The Moderate Resolution Imaging Spectroradiometer (MODIS) provided the day and night LST composite images from Version-5 MOD11A2 (Terra) and the *NDVI* composite images from both Version-5 MOD13A2 (Terra) and Version-5 MYD13A2 (Aqua); see the URL NASA (2018) for details. We have focused on composite images of the Navarre region from the 2011 to 2015 time period. Each of the day LST and night LST remote sensing datasets require 230 tiles for enclosing Navarre. These datasets correspond to 46 scenes with a temporal resolution of 8 days every year for 5 years. Terra and Aqua have different starting dates each year because Terra starts on the first day of January and Aqua starts on the ninth day of January. Therefore, to achieve two composite *NDVI* images per month every year, we have retrieved 23 images from Aqua and one image from Terra, fitted to November. In total, 120 scenes with a 16-day temporal resolution have been captured across 5 years of study. The spatial resolution of each tile is equal to 166×154 (25,564) pixels of approximately 1 km^2. Inside the Navarre borders, 11,691 pixels are enclosed. All images have been downloaded from USGS

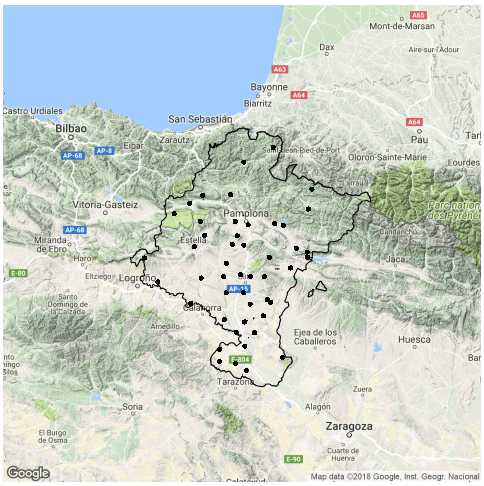

Figure 6.2: Map of Navarre region, located in the north of Spain and with a common border to the south of France. Black dots correspond to the rain gauge stations used in this study.

(2012) in Hierarchical Data Format (HDF). This format helps catalog geo-referenced images, but makes data processing more difficult. Then, all images were transformed into TIFF format to be processed by the R software (R Core Team, 2019).

The MODIS LST product is derived from two thermal infrared (TIR) channels: 31 (10.78 to 11.28 µm) and 32 (11.77 to 12.27 µm). More information can be found in Benali et al. (2012). Assuming that the signal difference in the two TIR bands is caused by differential absorption of radiation in the atmosphere, the correction of atmospheric effects is performed with the split-window algorithm (Wan and Dozier, 1996). Using prior knowledge of the land cover type classification based on the MODIS land cover product (MOD12C1), this algorithm corrects for emissivity effects. Uncertainty in LST estimates increases when significant variations in temperature occur at the 5-km^2 scale Wan and Li (1997). Errors in LST retrieval may be larger in bare soil and highly heterogeneous areas due to large uncertainties in surface emissivities and when the column water vapor content is high (Wan and Li, 1997). An eight-day composite period was chosen because twice this period is the exact ground track repeat period of the Terra platform. LST over eight days is the averaged LSTs of the MOD11A2 product for eight days. Processing these images will preserve the original temperature in Kelvin.

Both derived variables come from composite images, meaning that they have been pre-processed to reduce noise coming from atmospheric and electronic effects and that their accuracy has also been checked via ground-truth data at a coarse spatial resolution. However, when downscaling these data to small regions and comparing them with ground-truth data, we can observe, in some seasons, typically in autumn

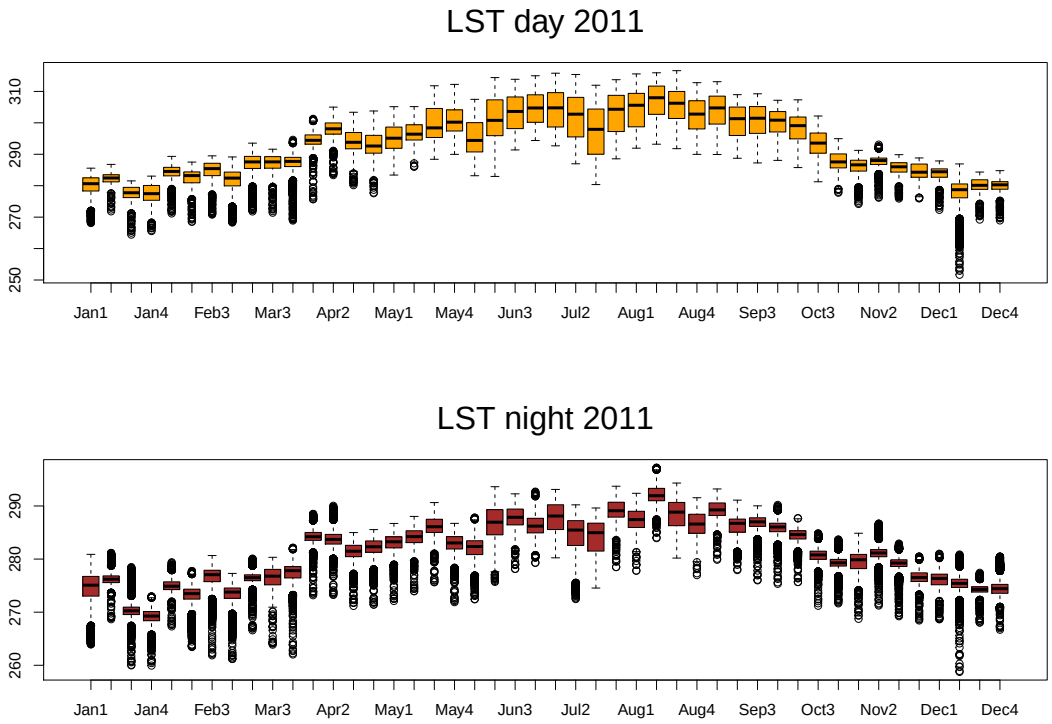

Figure 6.3: From the top to bottom, boxplots of day LST (in Kelvin), night LST (in Kelvin) for the 46 time periods of 2011.

and winter, some abnormal results. Figures 6.3 and 6.4 shows from the top to bottom the boxplots of the day LST and night LST remote sensing data and those of the mean maximum temperature (T_{max}) and the mean humidity (H_{mean}) in Navarre during 2011 for 46 time periods. Each boxplot depicts the median in the horizontal line and the dispersion of every image in a particular stage. Within each box, data between the first and third quantiles are plotted. Dots outside vertical extensions might be, but are not necessarily outliers. The meteorological variables T_{max} and H_{mean} have been chosen because they are the most highly correlated with day and night LST and *NDVI* during the time periods. Figure 6.5 shows the boxplots of *NDVI* in the 24 time periods of 2011. In Figures 6.3, 6.4, and 6.5, all the variables show a clear pattern of seasonality, yet in H_{mean}, larger variability is observed. Day and night LST and T_{max} show a parallel concave shape and therefore a positive correlation across time periods; H_{mean} is more convex, showing a negative correlation with day and night LST. *NDVI* presents a different seasonality pattern, which is negatively correlated with T_{max} and positively correlated with H_{mean} across the time periods. In summer, all the variables are more stable and show less variability. Similar phenology patterns are found across the study period of 2011- to 2015.

Figure 6.6 shows Navarre images in the third week of 2011. At the top, day LST

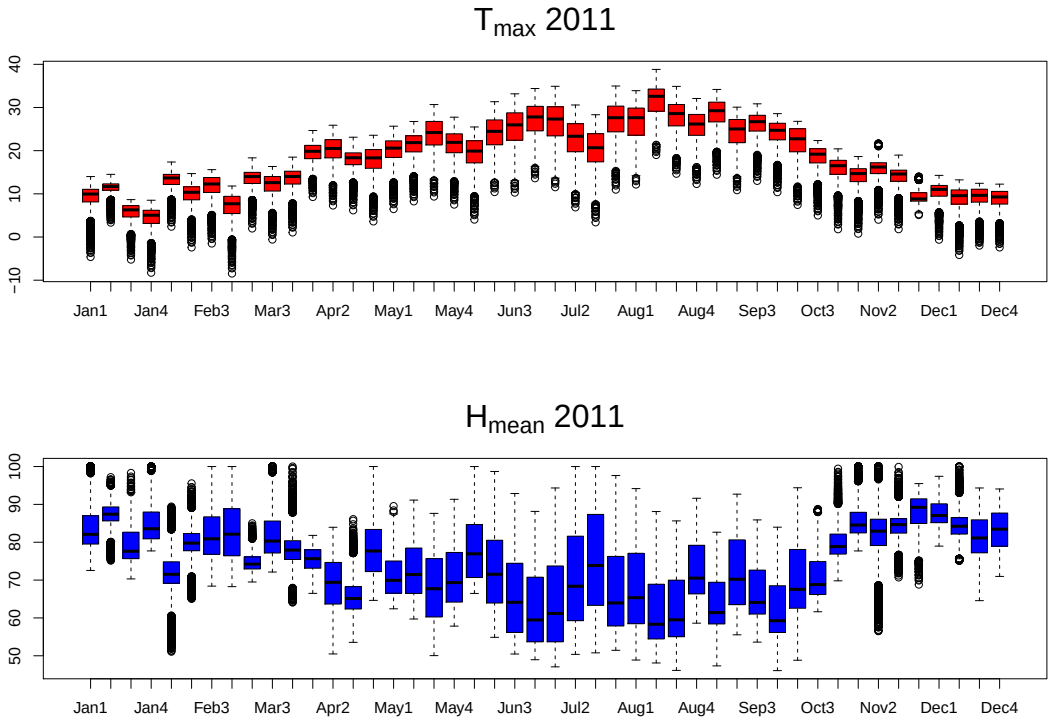

Figure 6.4: From the top to bottom, T_{max} (in Celsius) and H_{mean} (in percentages) for the 46 time periods of 2011.

and T_{max} (in Celsius) are shown, and at the bottom, night LST (in Celsius) and H_{mean} (in percentages) are presented. We can also see the large similarity among these patterns on these dates. The dots on the right images correspond to the rain gauge stations used as ground-truth data. Figure 6.7 shows the altitude of Navarre and the *NDVI* on the second fortnight of February 2014 because *NDVI* includes 24 periods. In this figure, we can also check why similar patterns correspond to a high correlation.

6.3 Methods

Filling gaps and removing abnormal observations in satellite imagery are traditionally performed by mathematical or statistical procedures based on modelling similarities of the same historical series of images. Few methods use external information. The newly-proposed method in this work, named TpsWc, uses additional auxiliary data from external sources, and it is compared with a state-space model that uses the same auxiliary variables or covariates.

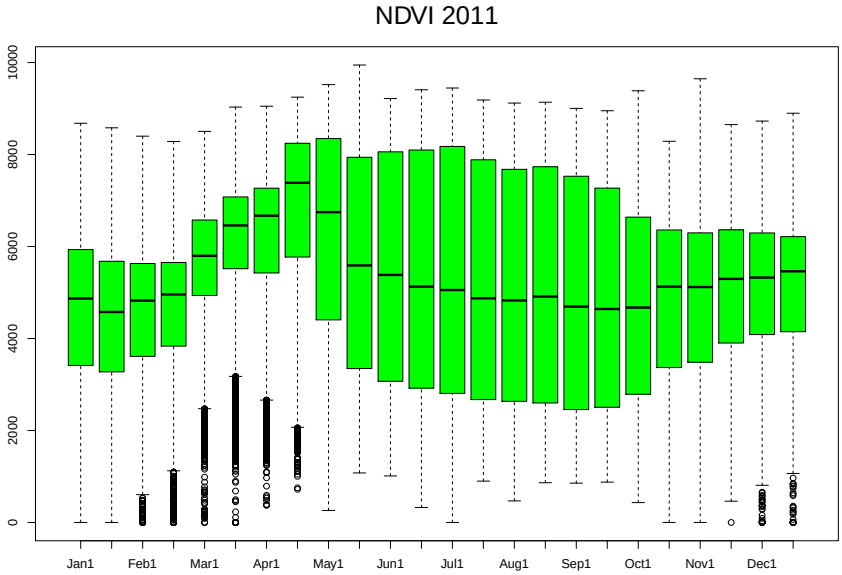

Figure 6.5: Boxplots of *NDVI* (with a zero to 10,000 scale) in the 24 time periods of 2011.

6.3.1 The thin-plate spline model with covariates

First, we fix the target image to be smoothed and define its neighborhood. Let us assume that we have an day LST target image. In this case, $G = 46$ images are available every year from $(r = 2011, \ldots, 2015)$, which should be arranged into a $5 \times G = 5 \times 46$ matrix, where the rows of the matrix correspond to different years. In the *NDVI* case, $G = 24$ images are available each year; they should be arranged into a 5×24 matrix. All the images in the same column correspond to the same time period, but different years. They share a neighbor composed of this column and the previous and subsequent columns of images; therefore, the neighbor of every target image consists of 15 images. In the first time period of 2011 and in the last time period of 2015, previous and subsequent images are also needed, but they do not correspond to the years under study. The second step of this procedure is to compute the median image out of those 15 images and obtain the corresponding anomalies for the target image. The anomalies could come from the mean or median. It is more convenient to calculate them from the mean in the gap-filling process because the mean can be obtained without missing values; calculating from the median should be performed when smoothing altered pixels, because it is more robust, and we are likely not aware of altered pixels.

In greater detail, let us denote $z_{s_i rg}$ as the derived variable in location $s_i = (x_i, y_i)$, $(i = 1, \ldots, n_0)$, where $n_0 = 25{,}564$ is the total number of pixels in the target image for year r. Then, the s_i-th pixel of the median image of time g across the years is

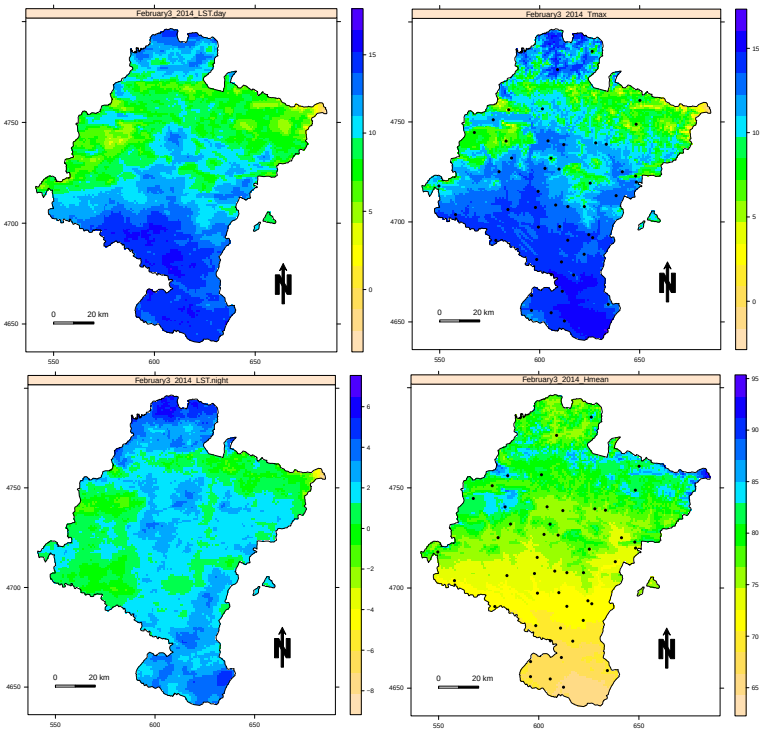

Figure 6.6: Images of Navarre for the third week of February 2014. At the top, day LST and T_{max} images (in Celsius) are presented, and at the bottom, night LST (in Celsius) and H_{mean} (in percentages) are shown. These images show similar patterns. Black dots represent rain gauge stations.

defined over the 15 images involved in its neighborhood, i.e.,

$$z_{s_i 0g} = median\{z_{s_i r g_0}\} \begin{cases} g_0 = (g-1), g, (g+1) \\ r = 2011, \ldots, 2015 \end{cases} \tag{6.1}$$

and the anomalies are obtained as the differences between the original values and the median:

$$w_{s_i rg} = z_{s_i rg} - z_{s_i 0g}, \quad \text{for} \quad i = 1, \ldots, n_0, \quad g = 1, \ldots, G, \quad r = 2011, \ldots, 2015. \tag{6.2}$$

Next, a thin-plate spline model (Tps) is applied to a 5-times lower resolution of the anomalies inside Navarre obtained through a median aggregation. Therefore, the median can be calculated within the tile of 25,564 pixels, but the model will be constrained to the $n_1 = 11{,}691/25 \approx 468$ pixels inside Navarre because that is where we possess ground-truth data. The reduction factor is recommended because it speeds up the running process and facilitates the previous image smoothing, but

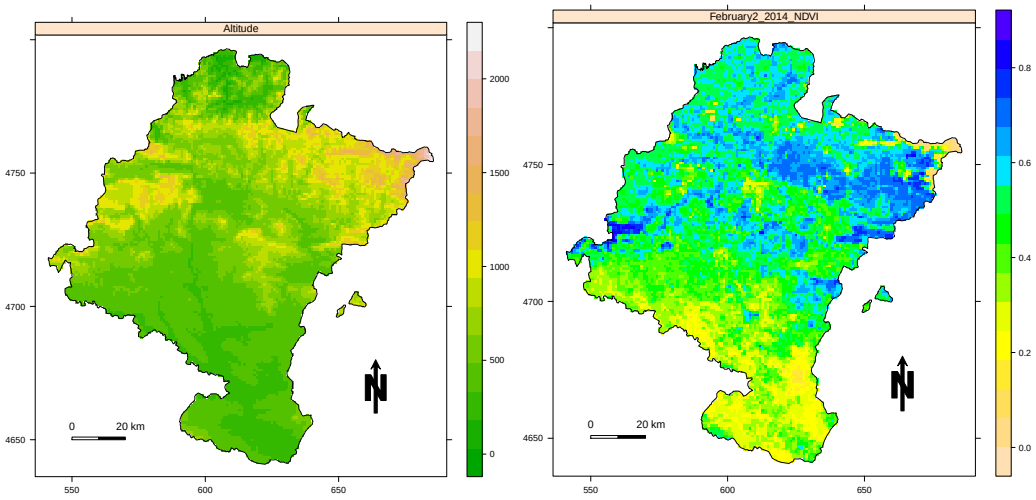

Figure 6.7: On the left is the altitude map, and on the right is the *NDVI* image of Navarre on the second fortnight of February 2014. Both figures show similar patterns because they are highly correlated.

this can vary depending on the computer speed and memory. The thin-plate spline additive model for every image in a fixed time period g and year r with covariates is expressed as a non-parametric function of the coordinates plus the sum of three transformed external covariates: the altitude ($\boldsymbol{u}_{1\mathbf{s}}$), maximum mean temperature ($\boldsymbol{u}_{2\mathbf{s}rg}$) and mean humidity ($\boldsymbol{u}_{3\mathbf{s}rg}$). Then,

$$\mathbf{w}_{\mathbf{s}rg} = f(\mathbf{s}) + \beta_1 \boldsymbol{u}_{1\mathbf{s}} + \beta_2 \boldsymbol{u}_{2\mathbf{s}rg} + \beta_3 \mathbf{u}_{3\mathbf{s}rg} + \boldsymbol{\epsilon}_{\mathbf{s}rg}, \quad g = 1, \ldots, G, \quad r = 2011, \ldots, 2015, \tag{6.3}$$

where $\mathbf{s} = (s_1, \ldots, s_{n_1})'$ is the vector of locations, $f(\mathbf{s})$ is a non-parametric function of the coordinates and β_j are the model coefficients ($j = 1, 2, 3$) associated with the covariates. They need to be estimated from the data because, later, we will be able to calculate the smoothed images. For every year r and stage g, $\mathbf{w}_{\mathbf{s}rg} = (w_{s_1rg}, \ldots, w_{s_{n_1}rg})$ is the vector of remote sensing anomalies, and $\boldsymbol{u}_{j_{\mathbf{s}rg}} = \boldsymbol{u}_{0j_{\mathbf{s}rg}} - \{z_{\mathbf{s}0g}\}$ ($j = 1, 2, 3$) are the transformed covariates of their original covariates: altitude ($\boldsymbol{u}_{01\mathbf{s}}$), maximum temperature ($\boldsymbol{u}_{02\mathbf{s}rg}$) and mean humidity ($\boldsymbol{u}_{03\mathbf{s}rg}$).

This transformation is needed to preserve the correlation between the response variable calculated as median anomalies of the remote sensing data and the transformed covariates of T_{max}, H_{mean} and altitude. This is because day LST, night LST and *NDVI* remote sensing data are correlated with T_{max}, H_{mean} and altitude, respectively, but their anomalies are not necessarily correlated with these. Therefore, a simple transformation that consists of subtracting the remote sensing median from the covariates is needed. The error vector $\boldsymbol{\epsilon}_{\mathbf{s}rg}$ has zero mean and variance σ^2. The thin-plate spline prediction is obtained as a weighted average of the observed data

because the optimal estimate is linear in the observations. Finally, the predictions $\hat{\mathbf{z}}_{\mathbf{s}rg} = \mathbf{z}_{\mathbf{s}rg} + z_{\mathbf{s}0g}$ are computed over the n_0 pixels of the original resolution. Model (6.3) with covariates (TpsWc) and without covariates (TpsWoc) is run for all rg time periods, the three remote sensing datasets and the four sizes of distortions for the simulation study. The R package fields (D. Nychka et al., 2015) estimates second-order thin-plate spline models by fitting a surface to irregularly-spaced data.

6.3.2 The state-space model with covariates

This model is a stochastic spatio-temporal model (SSM) (Durbin and Koopman, 2012) widely used for dynamical systems (see Amisigo and Van De Giesen, 2005; Militino et al., 2015, 2017). The model includes two equations: the transition Equation (6.4) and the state Equation (6.5). The first equation explains a linear regression between the response variable and the covariates. In this example, the response variable is the remote sensing data (day and night LST and *NDVI*), and the covariates are the meteorological variables T_{max}, H_{mean} and altitude. The second equation expresses the temporal dependence. The stochastic process at n_2 locations $s_1, \ldots, s_{n_2}$ and $t = 1, \ldots, T_m$ time points is represented by $\mathbf{z}_{st} = (z(s_1, t_1), z(s_1, t_2), \ldots, z(s_1, t_3), \ldots, z(s_{n_2}, t_{T_m}))'$, where $\mathbf{z}_{st}$ can be day and night LST or *NDVI*, and $T_m = 5$ years $\times G_m$. The value of G_m depends on both the remote sensing data and the m-th climatological season. In winter (January, February and March) and summer (July, August and September), there are $G_m = 11$ day and night LST composite images, and in spring (April, May and June) and fall (October, November and December), there are $G_m = 12$ day and night LST composite images. When using *NDVI*, there are $G_m = 6$ images in each climatological season.

In this application, we consider $n_2 = 208$ locations obtained by defining a 7×7 km^2 grid inside Navarre. The state-space model with covariates (SSMWc) is given by:

$$\mathbf{z}_{st} = \gamma_0 + \gamma_1 \mathbf{u}_{01s} + \gamma_2 \mathbf{u}_{02st} + \gamma_3 \mathbf{u}_{03t} + \mathbf{v}_t + \boldsymbol{\epsilon}_{st}, \quad \boldsymbol{\epsilon}_{st} \sim N_{n_2}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(d)), \tag{6.4}$$

$$\mathbf{v}_t = \mathbf{G}\mathbf{v}_{t-1} + \boldsymbol{\eta}_t, \quad \mathbf{v}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta), \tag{6.5}$$

where now γ_i $(i = 0, 1, 2, 3)$ are the model coefficients to be estimated. The first covariate $\mathbf{u}_{01s}$ is time invariant and corresponds to the altitude of the n_2 gridded locations. The remainder of the covariates, $\boldsymbol{u}_{02st}$ and $\boldsymbol{u}_{03st}$, are the spatio-temporal meteorological covariates: maximum temperature (T_{max}) and mean humidity (H_{mean}) depending on the location s and time t. The unobservable latent temporal process, $\mathbf{v}_t$, considers the temporal dynamics of data through an autoregressive process. This means that the current state $\mathbf{v}_t$ depends on the previous state $\mathbf{v}_{t-1}$ in the state equation through a transition matrix $\mathbf{G}$. The initial state vector $\mathbf{v}_0$ is assumed to be normally distributed with mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$.

This state-space model is fitted in the R statistical software using the Stem package (Cameletti, 2012). In this application, state-space models with covariates (SSMWc) and without covariates (SSMWoc) are estimated for the three remote sensing datasets, the four types of outlier outbreaks and all images between 2011 and 2015.

6.4 Results and discussion

To check the contribution of the ground-truth data for improving the quality of satellite imagery, a simulation study is conducted in Navarra using 120 composite images of *NDVI*, 230 composite images of day LST and 230 composite images of night LST between 2011 and 2015. In this simulation study, we randomly include four different sizes of outlier outbreaks with 5%, 10%, 15% and 20% abnormal observations in each image. The distortion consists of altering 50% of the *NDVI* raw data, 5% of the day LST raw data and 5% of the night LST raw data. The distortion percentages look different, but all represent approximately 50% of the range of the three remote sensing datasets used in this work. The performance of the methods is evaluated with the square root of the mean squared prediction error, defined by:

$$RMSE(j,k,l,m) = \sqrt{\frac{\sum_{i,t}(z_{itjklm} - \hat{z}_{itjklm})^2}{n_4 T_m}}, \tag{6.6}$$
$$i = 1, \ldots, n_4$$
$$t = 1, \ldots, T_m$$
$$j = TpsWc, TpsWoc, SSMWc, SSMWoc$$
$$k = 5\%, 10\%, 15\%, 20\%$$
$$l = dayLST, nightLST, NDVI$$
$$m = winter, spring, summer, fall,$$

where z_{ijkl} and $\hat{z}_{ijkl}$ are the original and predicted derived variables, respectively; $n_4 = 11{,}691$ is the number of pixels inside the Navarre borders; and T_m is the number of images in the m-th climatological season across the five years (see Subsection 6.3.2). The index $j = TpsWc, TpsWoc, SSMWc, SSMWoc$ indicates the type of model; k is the type of distortion; and l is the type of derived variable. $RMSE(j,k,l,m)$ is calculated and plotted in Figures 6.8 and 6.9. Figure 6.8 depicts the RMSE of day LST and night LST for the two versions of Tps, i.e., TpsWc (blue with covariates) and TpsWoc (purple without covariates), and SSM, i.e., SSMWc (red with covariates) and SSMWoc (green without covariates). In all cases, TpsWc outperforms the remaining models, as it achieves the lowest RMSE. Figure 6.9 shows the RMSE obtained after

smoothing *NDVI*. TpsWc is again the best method in all the climatological seasons and again presents a greater difference with regard to the SSM with and without covariates (SSMWc and SSMWoc). Nevertheless, differences between thin-plate splines with and without covariates are apparently less important than in day LST and night LST, but only because the range of the *NDVI* variable is smaller than that of LST. Summarizing, all the figures show lower RMSE when using covariates in both thin-plate splines (Tps) and state-space models (SSM). Table 6.1 shows the RMSE reduction percentage when both models include ground-truth information compared to when not including this information. The maximum percentage reduction is approximately 20% in night LST for TpsWc and in *NDVI* for SSMWc; however, TpsWc clearly provides the lowest values of RMSE with and without covariates.

Figure 6.10 shows the effects of distorting the images and smoothing them in the fourth time period of November 2011. At the top and from left to right, we see the distorted image of LST with an outlier outbreak of 5%, the TpsWc smoothed image and the SSMWc smoothed image with altitude, T_{max} and H_{mean} as the ground-truth covariates. At the bottom, the distorted image with an outlier outbreak of 20% in the same time period and the derived smoothed images are shown. Both models remove distorted data, and SSMWc seemingly better smooths the images, but TpsWc better preserves the original image pattern. The top of Figure 6.11 shows the boxplots of the distorted images, and at the bottom, the boxplots of the smoothed images in the 46 periods of 2011 are shown. We can see how the phenology of the remote sensing data is preserved after smoothing.

Table 6.1: Reduction percentage of the RMSE in SSM and Tps smoothing procedures with and without covariates for day LST, night LST and *NDVI* for different sizes of outlier outbreaks.

		SSM			Tps		
Outlier		**RMSE**			**RMSE**		
Derived Variable	**Out. %**	**Without Covariates**	**With Covariates**	**Reduction %**	**Without Covariates**	**With Covariates**	**Reduction %**
Day LST	5	1.80	1.63	10.19	0.54	0.51	7.75
	10	2.33	2.20	6.21	0.76	0.68	12.09
	15	2.86	2.74	4.47	1.00	0.89	12.56
	20	3.54	3.44	2.89	1.37	1.21	13.31
Night LST	5	1.34	1.29	3.90	0.41	0.37	11.89
	10	1.91	1.85	2.83	0.57	0.48	19.70
	15	2.52	2.47	1.73	0.78	0.65	19.94
	20	3.12	3.08	1.28	1.13	0.94	20.36
NDVI	5	0.11	0.09	19.59	0.04	0.04	1.68
	10	0.11	0.10	15.11	0.05	0.05	1.61
	15	0.12	0.11	12.43	0.05	0.05	1.31
	20	0.12	0.11	10.08	0.05	0.05	1.56

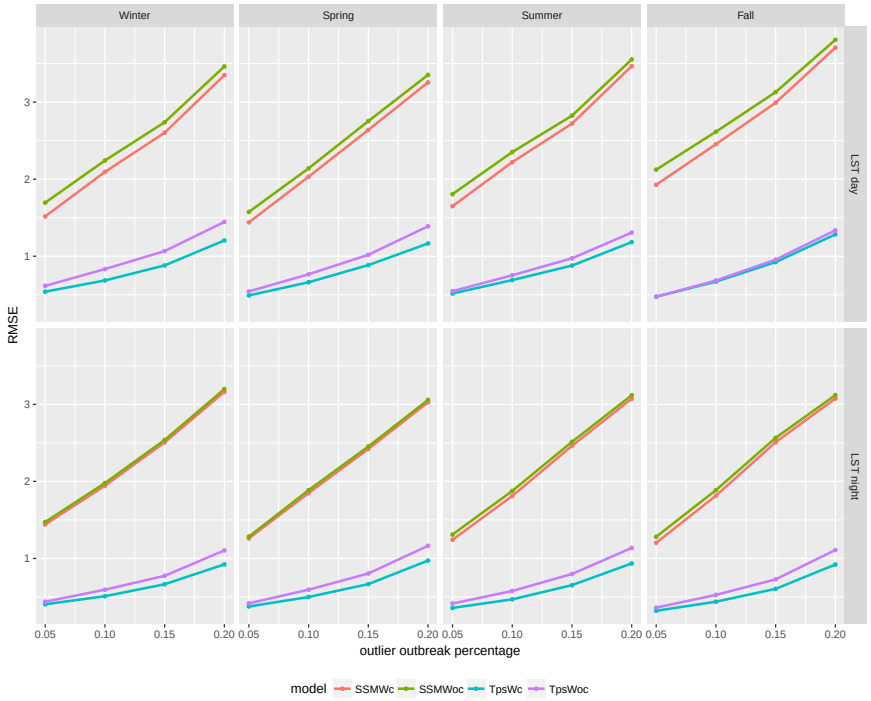

Figure 6.8: Root mean square prediction error versus outlier outbreak percentage obtained for day (on the top) and night (at the bottom). Land surface temperature (LST) by climatological seasons with the four models: space-state model (SSM) with and without covariates (SSWc in red and SSMWoc in green) and Tps with and without covariates (TpsWc in blue and TpsWoc in purple).

The simulation study shows an important decrease in the RMSE calculated with both the TpsWc and SSMWc methods; however, unless we alter the ground-truth data, we cannot evaluate what occurs in terms of RMSE when the covariates exhibit higher or lower correlations with the dependent variable. This is why we have defined a new artificial covariate linearly correlated with the remote sensing data. This correlation takes on values of $0.66, 0.75, 0.83, 0.92$ and 1. After introducing at random 20% abnormal observations in the time series of raw images, we smoothed them using TpsWc and SSMWc with the artificial covariate. When the correlation increases, a greater reduction in the RMSE is observed, although specific values are not shown here to preserve space. On average, each time we increase the linear correlation of the artificial variable with the remote sensing data by one tenth, a 3 to 4% RMSE reduction in the TpsWc and SSMWc models is observed. Therefore, using ground-truth data for increasing the quality of composite images is a recommended option in both models, although TpsWc achieves better results, as it has the lowest RMSE. The mean running time of day and night LST for processing 230 images is

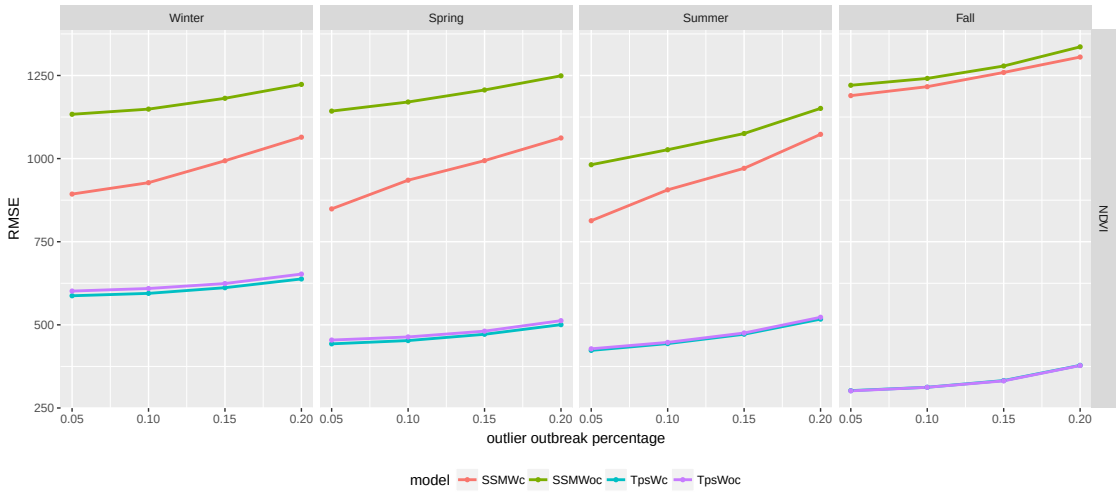

Figure 6.9: Root mean square error versus outlier outbreak percentage obtained for the normalized difference vegetation index ($NDVI$) by climatological season.

approximately 13 min with TpsWc and 11 min with SSWc on a PC with an Intel Core i7-4790 3.60 GHz with 4 cores and 16 GB of RAM; therefore, both models are very fast models.

6.5 Conclusions

In this work, we propose to increase the quality of satellite images through the use of ground-truth data from rain gauge stations in different stochastic models. A preliminary analysis for assessing the quality of day LST, night LST and $NDVI$ remote sensing data in Navarre (Spain) during 2011 to 2015 has revealed that these composite images preserve the temporal and seasonal patterns for each year; however, fall and winter represent the most likely periods whereby these data could be more vulnerable to atmospheric and electronic errors, and some atypical observations can emerge. One way of avoiding abnormal values in remote sensing data is accommodating ground-truth data at high temporal and spatial resolutions when applying statistical models for smoothing. However, models involving both types of data and that are able to manage spatial and temporal stochastic dependences remain scarce. In this study, we propose a new method called TpsWc that is based on smoothing the median anomalies of the remote sensing data with a thin-plate spline model wherein the anomalies of the ground-truth data are the external covariates. The method is compared with a version without covariates (TpsWoc) and a state-space model with (SSMWc) and without covariates (SSMWoc).

The new approach (TpsWc) encompasses the temporal dependence among multi-

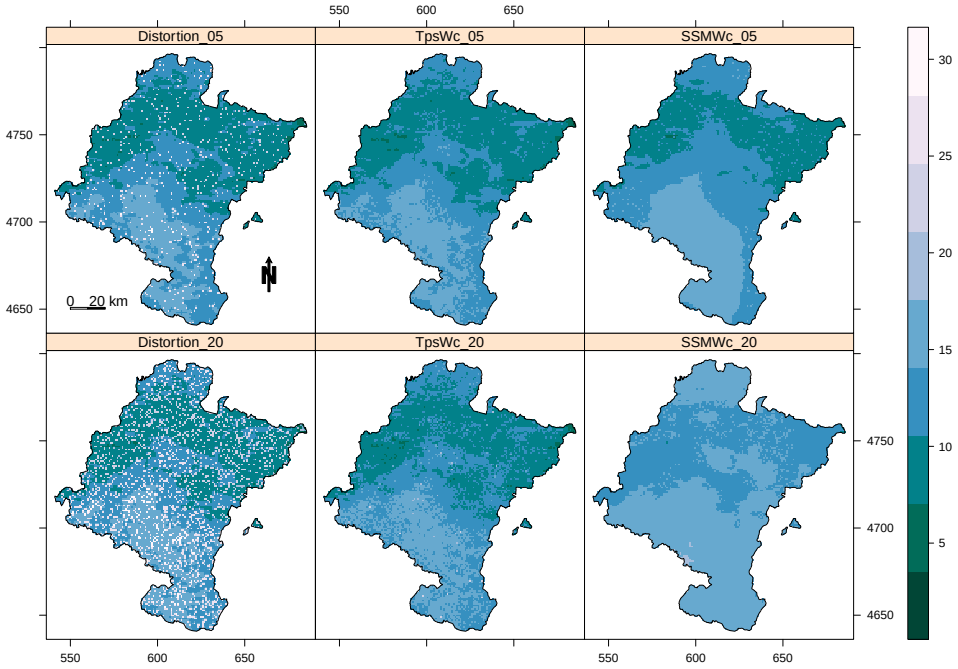

Figure 6.10: LST Navarra image in the fourth week of November 2011. In the upper row and from left to right, the 5% distorted image, the thin-plate spline (TpsWc) and the state-space (SSMWc) smoothed images with covariates. In the lower row and from left to right, the 20% distorted image and their respective TpsWc and SSMWc smoothed images with covariates.

temporal satellite images because it smooths the anomalies of the previous and subsequent images across years, and it accommodates the spatial dependence fitting non-parametric functions of the coordinates with a thin-plate spline. Ground-truth data are included as external covariates in the model after subtracting the median. We have conducted a simulation study wherein different outlier outbreaks have been randomly introduced in the original images. The study finds that TpsWc is the best option for all the variables, although SSWc presents a greater reduction in the RMSE for *NDVI* when using covariates. Moreover, we have found that when the covariates are more strongly correlated with the remote sensing data, a greater reduction in the root mean squared prediction error is achieved. Finally, the phenology of all the variables is preserved, but the potential outliers are removed in all the remote sensing data.

The thin-plate splines and state-space models used in this work are applied in time-consuming procedures, and the running time increases when we increase the spatial and temporal resolutions of the images. Therefore, improving these procedures from a computational point of view remains a matter of further research, because

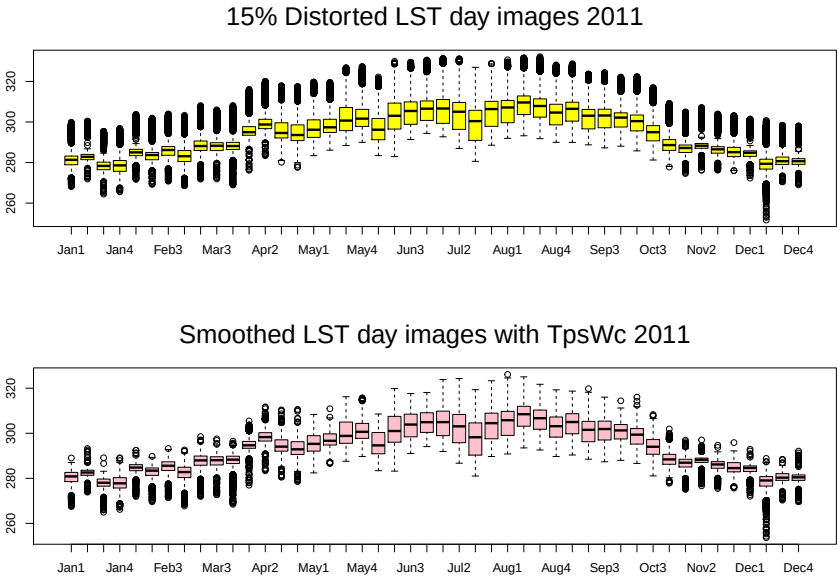

Figure 6.11: At the top, boxplots of the 15% distorted images of day LST in the 46 time periods of 2011 are shown, and the bottom presents the boxplots of the smoothed day LST images by TpsWc in the same time periods.

there are many alternatives for improving the quality of the satellite images, but only a few such methods utilize ground-truth data.

The contents of this chapter have been published in the journal *Remote Sensing*: Militino, A., Ugarte, M., and Pérez-Goya, U. (2018a). Improving the quality of satellite imagery based on ground-truth data from rain gauge stations. *Remote Sensing*, 10(3):398.

Conclusions and further work

Satellite platforms providing open-access satellite imagery have become a key source in many analysis. One of the main contribution of this thesis is the development of the R package 'RGISTools'. This package encapsulates in 1.5MB all the tools for downloading and processing satellite images to get ready-to-use data sets. To run all the procedures there are 60 functions explained in a manual with 81 pages. The package has been published in August 2019 on the main R repository, the Comprehensive R Archive Network (CRAN), giving global diffusion in the greatest statistical programming project. 'RGISTools' has specific functions to automatize the acquisition of remote sensing data.

The first procedure defines new functions for downloading, preprocessing, and loading into R, time series of satellite images from three different repositories. The satellite platforms provide the images in particular formats, tiling systems, and processing levels. The package standardises downloading procedures to be similar for Landsat-7, Landsat-8, MODIS, and Sentinel-2 satellites. 'RGISTools' also considers the process of changing the image format, cropping, and tile mosaicking to create time series of images for a region of interest. Tile mosaicking is required when the region of interest extends over several tiles, so they can be combined into a single image. Cropping returns a tile of the region of interest, making any analysis more computationally and memory efficient.

The second procedure is devoted on deriving variables from the pre-processed spectral images. These functions allow the definition of new indices from spectral images, such as NDVI or EVI, among others. Functions to reduce the effect of atmospheric factors as clouds or measurements errors are also provided. 'RGISTools' facilitates the entire processing from downloading the scratch images, up to get ready-to-use time series of images in R.

This thesis also presents two statistical analysis using time series of satellite

images, one dedicated to search change-points or trend changes, and the other for analysing the NDVI distribution. The first work looks for change-point detection in the size of surfaces occupied by four categories of NDVI in continental Spain. The use of the categories make it easier the detection of the change-points, showing an slight increase of vegetation in Spain.

The second work gives a smoothing procedure incorporating the spatio-temporal dependence of a time series of NDVI remote sensing data. This study uses the state space model (SSM) in a time series of NDVI images in continental Spain. Rainfall data is included as auxiliary variable in order to improve the estimation procedure. The result of the modelling shows the distribution of the NDVI in time and space revealing a clear seasonality that is very correlated with the levels of rainfall. This work proves the improvement of the estimation when using spatio temporal dependence in satellite images.

Finally, two new procedures for filling gaps or smoothing data including spatio temporal dependence in satellite images are presented. The first procedure is named Image Mean Anomaly (IMA). IMA is based on a model that expresses the images as the sum of a trend plus a random error. The trend is explained as a mean or a median computed from a spatio temporal neighbourhood around the target image. The prediction is made on the random error image or residual, a very common technique in statistics but not in satellite imagery. This new procedure proves to be superior in terms of RMSE and runtime over its main competitors.

The second procedure is an extension of IMA using covariates to improve the predictions. The procedure is named TpsWc because uses Thin Plate Splines (Tps) with covariates to interpolate the anomalies. TpsWc is designed to use covariates for smoothing outliers in the images. The procedure is evaluated in a simulation scenario where outliers are introduced and compared with the state space model (SSM). The results show that the use of covariates improves the predictions in all of the considered variables (NDVI, LST day and LST night). The results also show better predictions in the new TpsWc method compared to SSM predictions. Both procedures IMA and TpsWc are also implemented and published in the 'RGISTools' package.

Further work

1. **The border effect**: This thesis shows the suitability of using the stochastic spatio-temporal dependence for improving the processing and analysis of time series of images. However, many procedures seem very affected by the border effect. That is, the predictions, when there are no observations near the border, are usually badly predicted. A solution may be to enlarge the observed region in order to avoid the border effect, but this is not a good option because it

increases the number of observations slowing the runtime of the procedures. The adaptation and improvement of the reconstruction methods to reduce the border effect will be necessary in a near future.

2. **Using LIDAR data to reconstruct multispectral images**: In Chapter 6 a satellite smoothing method using covariates was presented. The covariates were obtained interpolating data from rainfall stations. The satellite programs presented in this dissertation provide additional sources. For example, Sentinel provides LIDAR images. LIDAR images are not as affected by atmospheric factors as multispectral images. We think that the data obtained from LIDAR images could also be used as covariates to reconstruct multispectral images in the presence of clouds or other image distortions. The use of radar images can be an interesting topic to continue the research in gap filling.

3. **Data fusion for downscaling**: The variety of spectral satellites as Landsat, MODIS, and Sentinel-2, with different spatial and temporal resolutions, make the images not compatible with each other. However, we can make fusion of data from different satellites to obtain daily time series of high resolution images. The procedures improving satellite imagery spatial resolution are known as downscaling. The integration of the prediction by means of a residual image for the implementation of downscaling opens up possibilities for research.

4. **Improving the procedures for future data sets**: In this thesis very competitive and efficient procedures have been developed. The procedures have been parallelized to reduce the runtime. However, the size of the images is increasing with the newest satellites. For example, one band from a tile of Landsat-8, launched in 2013, contains more than 50 millions of pixels. The newest Sentinel-2, launched two years later, for the same tile and band has more than 120 millions of pixels. The magnitude of the problem with the data management and processing is increasing with the new satellites. In a near future, the use of GPUs and distributed storage systems will be necessary for analysing time series of satellite images.

References

Addink, E. (1999). A comparison of conventional and geostatistical methods to replace clouded pixels in NOAA-AVHRR images. *International Journal of Remote Sensing*, 20(5):961–977.

AEMET (2019). Spanish national agency of meteorology (aemet). *Retrieved from* *http://www.aemet.es/es/portada*. Accessed: 2019-09-28.

Ahmed, M., Else, B., Eklundh, L., Ardö, J., and Seaquist, J. (2017). Dynamic response of ndvi to soil moisture variations during different hydrological regimes in the sahel region. *International Journal of Remote Sensing*, 38(19):5408–5429.

Amisigo, B. and Van De Giesen, N. (2005). Using a spatio-temporal dynamic state-space model with the em algorithm to patch gaps in daily riverflow series. *Hydrology and Earth System Sciences Discussions*, 9(3):209–224.

Antoch, J., Hu**vs**ková, M., and Prá**vs**ková, Z. (1997). Effect of dependence on statistics for determination of change. *Journal of Statistical Planning and Inference*, 60(2):291–310.

Atkinson, P. M., Jeganathan, C., Dash, J., and Atzberger, C. (2012). Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sensing of Environment*, 123:400–417.

Atzberger, C., Klisch, A., Mattiuzzi, M., and Vuolo, F. (2013). Phenological metrics derived over the european continent from ndvi3g data and modis time series. *Remote Sensing*, 6(1):257–284.

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54.

Bai, J. and Perron, P. (2003). Critical values for multiple structural change tests. *The Econometrics Journal*, 6(1):72–78.

Ban, Y. (2016a). *Multitemporal Remote Sensing. Methods and Applications*, volume 1. Springer Remote Sensing and Digital Image Processing.

Ban, Y. (2016b). *Multitemporal Remote Sensing: Methods and Applications*, volume 20. Springer.

Benali, A., Carvalho, A., Nunes, J., Carvalhais, N., and Santos, A. (2012). Estimating air surface temperature in Portugal using MODIS LST data. *Remote Sensing of Environment*, 124:108–121.

Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I., and Heynen, M. (2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(3-4):239–258.

Bivand, R., Keitt, T., and Rowlingson, B. (2019). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.4-4.

Bivand, R. S., Pebesma, E. J., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R*. Second edition. Springer.

Boer, E. P., de Beurs, K. M., and Hartkamp, A. D. (2001). Kriging and thin plate splines for mapping climate variables. *International Journal of Applied Earth Observation and Geoinformation*, 3(2):146–154.

Bolin, D., Lindström, J., Eklundh, L., and Lindgren, F. (2009). Fast estimation of spatially dependent temporal vegetation trends using gaussian markov random fields. *Computational Statistics & Data Analysis*, 53(8):2885–2896.

Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, pages 235–249.

Brown, P. E. (2015). Model-based geostatistics the easy way. *Journal of Statistical Software*, 63(12):1–24.

Brown, P. E. et al. (2015). Model-based geostatistics the easy way. *Journal of Statistical Software*, 63(12):1–24.

Cameletti, M. (2012). *Stem: Spatio-temporal models in R*. R package version 1.0.

Cameletti, M., Ignaccolo, R., and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in piemonte. *Environmetrics*, 22(8):985–996.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, 97(2):109–131.

Chen, B., Huang, B., Chen, L., and Xu, B. (2017). Spatially and temporally weighted regression: a novel method to produce continuous cloud-free Landsat imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 55(1):27–37.

Chen, J. and Gupta, A. K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media.

Christensen, O. F. and Ribeiro Jr, P. J. (2002). georglm-a package for generalised linear spatial models. *R News*, 2(2):26–28.

Cliff Andrew David, O. J. (1973). *Spatial autocorrelation*. Pion.

Couture-Beil, A. (2018). *rjson: JSON for R*. R package version 0.2.20.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Csörgö, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*, volume 18. John Wiley & Sons Inc.

Curran, P. J. and Atkinson, P. M. (1998). Geostatistics and remote sensing. *Progress in Physical Geography*, 22(1):61–78.

D. Nychka, R. Furrer, J. Paige, and S. Sain (2015). fields: Tools for spatial data. R package version 9.0.

Davino, C., Furno, M., and Vistocco, D. (2013). *Quantile regression: theory and applications*. John Wiley & Sons.

De Iaco, S., Myers, D. E., and Posa, D. (2002). Nonseparable space-time covariance models: some parametric families. *Mathematical Geology*, 34(1):23–42.

de Jong, R., de Bruin, S., de Wit, A., Schaepman, M. E., and Dent, D. L. (2011). Analysis of monotonic greening and browning trends from global ndvi time-series. *Remote Sensing of Environment*, 115(2):692–702.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38.

Detsch, F. (2019). Gimms: download and process gimms ndvi3g data. *R package version 1.1.1.*, 1(0).

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. *Constructive theory of functions of several variables*, pages 85–100.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

Edwards, A. W. and Cavalli-Sforza, L. L. (1965). A method for cluster analysis. *Biometrics*, pages 362–375.

Eerens, H., Haesen, D., Rembold, F., Urbano, F., Tote, C., and Bydekerke, L. (2014). Image time series processing for agriculture monitoring. *Environmental Modelling & Software*, 53:154–162.

Eklundh, L. and Jönsson, P. (2012). Timesat 3.2 with parallel processing-software manual. *Lund University*.

Erasmi, S., Schucknecht, A., Barbosa, M. P., and Matschullat, J. (2014). Vegetation greenness in northeastern brazil and its relation to enso warm events. *Remote Sensing*, 6(4):3041–3058.

ESA (2016). Scihub platform. *Retrieved from* https://scihub.copernicus.eu/. Accessed: 2019-09-28.

Evans, J. (2016). spatialeco: Spatial analysis and modelling. r package version 0.1-5.

Fassò, A. and Cameletti, M. (2009). The em algorithm in a distributed computing environment for modelling environmental space–time data. *Environmental Modelling & Software*, 24(9):1027–1035.

Fensholt, R., Rasmussen, K., Nielsen, T. T., and Mbow, C. (2009). Evaluation of earth observation based long term vegetation trends–intercomparing ndvi time series trend analysis consistency of sahel from avhrr gimms, terra modis and spot vgt data. *Remote Sensing of Environment*, 113(9):1886–1898.

Filipova-Racheva, D. and Hall-Beyer, M. (2000). Smoothing of NDVI time series curves for monitoring of vegetation changes in time. In *Ecological Monitoring and Assessment Network National Science Meeting*, pages 17–22.

Finley, A. O., Banerjee, S., and E.Gelfand, A. (2015). spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63(13):1–28.

Fontana, F. M., Trishchenko, A. P., Khlopenkov, K. V., Luo, Y., and Wunderle, S. (2009). Impact of orthorectification and spatial sampling on maximum ndvi composite data in mountain regions. *Remote Sensing of Environment*, 113(12):2701–2712.

Forkel, M., Carvalhais, N., Verbesselt, J., Mahecha, M. D., Neigh, C. S., and Reichstein, M. (2013). Trend change detection in ndvi time series: Effects of inter-annual variability and methodology. *Remote Sensing*, 5(5):2113–2144.

French, J. (2018). *SpatialTools: Tools for Spatial Data Analysis*. R package version 1.0.4.

Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T. S., and Brown, D. J. (2015). Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3d+ t: The cook agronomy farm data set. *Spatial Statistics*, 14:70–90.

Gerber, F., de Jong, R., Schaepman, M. E., Schaepman-Strub, G., and Furrer, R. (2018). Predicting Missing Values in Spatio-Temporal Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5):2841–2853.

Gerber, F., Furrer, R., Schaepman-Strub, G., de Jong, R., and Schaepman, M. E. (2016). Predicting missing values in spatio-temporal satellite data. *ArXiv e-prints*.

Greenberg, J. A. and Mattiuzzi, M. (2018). *gdalUtils: Wrappers for the Geospatial Data Abstraction Library (GDAL) Utilities*. R package version 2.0.1.14.

Hadoop, A. (2009). The Apache Hadoop. Oopen-source software for reliable, scalable, distributed computing. Accessed: 2019-09-28.

Harris, I., Jones, P., Osborn, T., and Lister, D. (2014). Updated high-resolution grids of monthly climatic observations–the cru ts3. 10 dataset. *International Journal of Climatology*, 34(3):623–642.

Hengl, T., Heuvelink, G. B., Tadić, M. P., and Pebesma, E. J. (2012). Spatio-temporal prediction of daily temperatures using time-series of modis lst images. *Theoretical and Applied Climatology*, 107(1-2):265–277.

Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C., and Hobart, G. W. (2015). An integrated Landsat time series protocol for change detection and generation of annual gap-free surface reflectance composites. *Remote Sensing of Environment*, 158:220–234.

Hijmans, R. J. (2019). *raster: Geographic Data Analysis and Modeling*. R package version 2.9-23.

Hird, J. N. and McDermid, G. J. (2009). Noise reduction of NDVI time series: An empirical comparison of selected techniques. *Remote Sensing of Environment*, 113(1):248–258.

Holben, B. N. (1986). Characteristics of maximum-value composite images from temporal avhrr data. *International Journal of Remote Sensing*, 7(11):1417–1434.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote Sensing of Environment*, 83(1-2):195–213.

Hutchinson, M. and Gessler, P. (1994). Splines–more than just a smooth interpolator. *Geoderma*, 62(1-3):45–67.

Ichii, K., Kawabata, A., and Yamaguchi, Y. (2002). Global correlation analysis for ndvi and climatic variables and ndvi trends: 1982-1990. *International Journal of Remote Sensing*, 23(18):3873–3878.

James, N. A. and Matteson, D. S. (2014). ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62(7):1–25.

Jiménez-Muñoz, J. C. and Sobrino, J. A. (2003). A generalized single-channel method for retrieving land surface temperature from remote sensing data. *Journal of Geophysical Research: Atmospheres*, 108(D22).

Jönsson, P. and Eklundh, L. (2004). Timesat–a program for analyzing time-series of satellite sensor data. *Computers & Geosciences*, 30(8):833–845.

Julien, Y., Sobrino, J. A., Mattar, C., Ruescas, A. B., Jimenez-Munoz, J. C., Soria, G., Hidalgo, V., Atitar, M., Franch, B., and Cuenca, J. (2011). Temporal analysis of normalized difference vegetation index (ndvi) and land surface temperature (lst) parameters to detect changes in the iberian land cover between 1981 and 2001. *International Journal of Remote Sensing*, 32(7):2057–2068.

Kang, E. L., Cressie, N., and Shi, T. (2010). Using temporal variability to improve spatial mapping with application to satellite data. *Canadian Journal of Statistics*, 38(2):271–289.

Kern, A., Marjanović, H., and Barcza, Z. (2016). Evaluation of the quality of ndvi3g dataset against collection 6 modis ndvi in central europe between 2000 and 2013. *Remote Sensing*, 8(11):955.

Keyes, O., Jacobs, J., Schmidt, D., Greenaway, M., Rudis, B., Pinto, A., Khezrzadeh, M., Meilstrup, P., Costello, A. M., Bezanson, J., Meilstrup, P., and Jiang, X. (2019). *urltools: Vectorised Tools for URL Handling and Parsing*. R package version 1.7.3.

Killick, R. and Eckley, I. A. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Killick, R., Haynes, K., and Eckley, I. A. (2016). *changepoint: An R package for changepoint analysis*. R package version 2.2.2.

Klisch, A. and Atzberger, C. (2016). Operational drought monitoring in kenya using modis ndvi time series. *Remote Sensing*, 8(4):267.

Kyriakidis, P. C. and Journel, A. G. (1999). Geostatistical space–time models: a review. *Mathematical geology*, 31(6):651–684.

Lang, D. T. and the CRAN Team (2019). *XML: Tools for Parsing and Generating XML Within R and S-Plus*. R package version 3.98-1.20.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge university press.

Li, H., Wang, C., Zhang, L., Li, X., and Zang, S. (2017). Satellite monitoring of boreal forest phenology and its climatic responses in eurasia. *International Journal of Remote Sensing*, 38(19):5446–5463.

Li, J. and Heap, A. D. (2011). A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecological Informatics*, 6(3-4):228–241.

Li, J. and Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, 53:173–189.

Li, Z., Huffman, T., McConkey, B., and Townley-Smith, L. (2013). Monitoring and modeling spatial and temporal patterns of grassland dynamics using time-series MODIS NDVI with climate and stocking data. *Remote Sensing of Environment*, 138:232–244.

Liu, Y., Li, Y., Li, S., and Motesharrei, S. (2015). Spatial and temporal patterns of global ndvi trends: Correlations with climate and human factors. *Remote Sensing*, 7(10):13233–13250.

Luo, W., Taylor, M., and Parker, S. (2008). A comparison of spatial interpolation methods to estimate continuous wind speed surfaces using irregularly distributed data from England and Wales. *International journal of climatology*, 28(7):947–959.

Ma, M. and Veroustraete, F. (2006). Reconstructing pathfinder AVHRR land NDVI time-series data for the Northwest of China. *Advances in Space Research*, 37(4):835–840.

Maselli, F., Papale, D., Chiesi, M., Matteucci, G., Angeli, L., Raschi, A., and Seufert, G. (2014). Operational monitoring of daily evapotranspiration by the combination of modis ndvi and ground meteorological data: Application and evaluation in central italy. *Remote Sensing of Environment*, 152:279–290.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

Melo, C., Santacruz, A., and Melo, O. (2012). *geospt: An R package for spatial statistics.* R package version 1.0-0.

Meng, Q., Borders, B. E., Cieszewski, C. J., and Madden, M. (2009). Closest spectral fit for removing clouds and cloud shadows. *Photogrammetric Engineering & Remote Sensing*, 75(5):569–576.

Militino, A., Ugarte, M., Goicoa, T., and Genton, M. (2015). Interpolation of daily rainfall using spatiotemporal models and clustering. *International Journal of Climatology*, 35(7):1453–1464.

Militino, A., Ugarte, M., and Pérez-Goya, U. (2018a). Improving the quality of satellite imagery based on ground-truth data from rain gauge stations. *Remote Sensing*, 10(3):398.

Militino, A. F. and Ugarte, M. D. (2001). Assessing the covariance function in geostatistics. *Statistics & Probability Letters*, 52(2):199–206.

Militino, A. F., Ugarte, M. D., and Pérez-Goya, U. (2017). Stochastic spatio-temporal models for analysing ndvi distribution of gimms ndvi3g images. *Remote Sensing*, 9(1):76.

Militino, A. F., Ugarte, M. D., and Pérez-Goya, U. (2018b). Detecting change-points in the time series of surfaces occupied by pre-defined ndvi categories in continental spain from 1981 to 2015. In Gil, E., Gil, E., Gil, J., and Gil, M. Á., editors, *The Mathematics of the Uncertain*, chapter 28, pages 295–307. Springer.

Militino, A. F., Ugarte, M. D., and Pérez-Goya, U. (2018c). An introduction to the spatio-temporal analysis of satellite remote sensing data for geostatisticians. In Sagar, B. D., Cheng, Q., and Agterberg, F., editors, *Handbook of Mathematical Geosciences*, chapter 13, pages 239–253. Springer.

Militino, A. F., Ugarte, M. D., Pérez-Goya, U., and Genton, M. G. (2019). Interpolation of the mean anomalies for cloud filling in land surface temperature and normalized difference vegetation index. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):6068–6078.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

NASA (2004). *Science writers guides to TERRA*. The National Aeronautics and Space Administration (NASA).

NASA (2014). Earthdata platform. *Retrieved from* *https://search.earthdata.nasa.gov/search*. Accessed: 2019-09-28.

NASA (2016). Nasa the lp daac web service for querying lp daac archive. *Retrieved from* *https://lpdaacsvc.cr.usgs.gov/services/*. Accessed: 2019-09-28.

NASA (2018). Modis2018. *Retrieved from* *https://modis.gsfc.nasa.gov/about*. Accessed: 2019-09-28.

Neeti, N. and Eastman, J. R. (2011). A contextual mann-kendall approach for the assessment of trend significance in image time series. *Transactions in GIS*, 15(5):599–611.

Neteler, M. and Mitasova, H. (2013). *Open source GIS: a GRASS GIS approach*, volume 689. Springer Science & Business Media.

Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014). Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics*, 56(2):174–185.

Ooms, J. (2019). *curl: A Modern and Flexible Web Client for R.* R package version 4.0.

Papritz, A. (2018). *georob: Robust Geostatistical Analysis of Spatial Data.* R package version 0.3-7.

Pebesma, E. et al. (2012). spacetime: Spatio-temporal data in r. *Journal of Statistical Software*, 51(7):1–30.

Pebesma, E. J. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, 30(7):683–691.

Pérez-Goya, U., Militino, A. F., Ugarte, M. D., and Montesino-SanMartin, M. (2019). *RGISTools: Handling Multiplatform Satellite Images.* R package version 0.9.7.

Pinzon, J. E. and Tucker, C. J. (2014). A non-stationary 1981–2012 avhrr ndvi3g time series. *Remote Sensing*, 6(8):6929–6960.

Potter, C. and Brooks, V. (1998). Global analysis of empirical relations between annual climate and seasonality of ndvi. *International Journal of Remote Sensing*, 19(15):2921–2948.

Qin, Z., Karnieli, A., and Berliner, P. (2001). A mono-window algorithm for retrieving land surface temperature from landsat tm data and its application to the israel-egypt border region. *International Journal of Remote Sensing*, 22(18):3719–3746.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro Jr, P. J., Diggle, P. J., et al. (2001). geoR: a package for geostatistical analysis. *R News*, 1(2):14–18.

Ripley, B. D. (1981). Spatial statistics. *Wiley Series in Probability and Mathematical Statistics: Applied Probability & Mathematical Section, New York: Wiley, 1981.*

Roerink, G., Menenti, M., and Verhoef, W. (2000). Reconstructing cloudfree ndvi composites using fourier analysis of time series. *International Journal of Remote Sensing*, 21(9):1911–1917.

Rouse Jr, J., Haas, R., Schell, J., and Deering, D. (1974). Monitoring vegetation systems in the great plains with erts. *NASA Special Publication*, 351:309.

Roy, D. P., Ju, J., Lewis, P., Schaaf, C., Gao, F., Hansen, M., and Lindquist, E. (2008). Multi-temporal MODIS–Landsat data fusion for relative radiometric normalization, gap filling, and prediction of Landsat data. *Remote Sensing of Environment*, 112(6):3112–3130.

Roy, D. P., Kovalskyy, V., Zhang, H., Vermote, E. F., Yan, L., Kumar, S., and Egorov, A. (2016). Characterization of landsat-7 to landsat-8 reflective wavelength and normalized difference vegetation index continuity. *Remote Sensing of Environment*, 185:57–70.

Sagar, D. B. and Serra, J. e. (2010). Spacial issue on spatial information retrieval, analysis, reasoning and modelling. *International Journal of Remote Sensing*, 31(22):5747–6032.

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., Strokorb, K., et al. (2015). Analysis, simulation and prediction of multivariate random fields with package randomfields. *Journal of Statistical Software*, 63(8):1–25.

Schultz, P. and Halpert, M. (1993). Global correlation of temperature, ndvi and precipitation. *Advances in Space Research*, 13(5):277–280.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512.

Sen, A. and Srivastava, M. S. (1975). On tests for detecting change in mean. *The Annals of Statistics*, pages 98–108.

Skidmore, A. K., Stein, A., van der Meer, F., and Gorte, B. (1999). Spatial statistics for remote sensing.

Slayback, D. A., Pinzon, J. E., Los, S. O., and Tucker, C. J. (2003). Northern hemisphere photosynthetic trends 1982–99. *Global Change Biology*, 9(1):1–15.

Sobrino, J. and Julien, Y. (2011). Global trends in ndvi-derived parameters obtained from gimms data. *International Journal of Remote Sensing*, 32(15):4267–4279.

Sobrino, J. A., Jimenez-Munoz, J. C., and Paolini, L. (2004). Land surface temperature retrieval from landsat tm 5. *Remote Sensing of Environment*, 90(4):434–440.

Sobrino, J. A., Julien, Y., and Morales, L. (2011). Changes in vegetation spring dates in the second half of the twentieth century. *International Journal of Remote Sensing*, 32(18):5247–5265.

Sola, I., González-Audicana, M., Alvarez-Mozos, J., and Torres, J. (2014). Synthetic images for evaluating topographic correction algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 52:1799–1810.

Swan, A. (1996). Wackernagel h., 1995. multivariate geostatistics. an introduction with applications. xiv+ 256 pp. berlin, heidelberg, new york, barcelona, budapest, hong kong, london, milan, paris, tokyo: Springer-verlag. price dm 74.00, ös 540.20, sfr 71.50 (hard covers). isbn 3 540 60127 9. *Geological Magazine*, 133(5):628–628.

Talih, M. and Hengartner, N. (2005). Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):321–341.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150.

Tucker, C. J., Pinzon, J. E., Brown, M. E., Slayback, D. A., Pak, E. W., Mahoney, R., Vermote, E. F., and El Saleous, N. (2005). An extended avhrr 8-km ndvi dataset compatible with modis and spot vegetation ndvi data. *International Journal of Remote Sensing*, 26(20):4485–4498.

Tüshaus, J., Dubovyk, O., Khamzina, A., and Menz, G. (2014). Comparison of medium spatial resolution envisat-meris and terra-modis time series for vegetation decline analysis: A case study in central asia. *Remote Sensing*, 6(6):5238–5256.

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2015). *Probability and Statistics with R*. CRC Press.

USGS (2012). Earthexplorer platform. *Retrieved from https://earthexplorer.usgs.gov/*. Accessed: 2019-09-28.

van Wijk, M. T. and Williams, M. (2005). Optical instruments for measuring leaf area index in low vegetation: application in arctic ecosystems. *Ecological Applications*, 15(4):1462–1470.

Vancutsem, C., Ceccato, P., Dinku, T., and Connor, S. J. (2010). Evaluation of modis land surface temperature data to estimate air temperature in different ecosystems over africa. *Remote Sensing of Environment*, 114(2):449–465.

Velleman, P. F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association*, 75(371):609–615.

Venables, W. N. and Ripley, B. D. (2002). Tree-based methods. In *Modern Applied Statistics with S*, pages 251–269. Springer.

Verbesselt, J., Hyndman, R., Newnham, G., and Culvenor, D. (2010a). Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, 114(1):106–115.

Verbesselt, J., Hyndman, R., Zeileis, A., and Culvenor, D. (2010b). Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. *Remote Sensing of Environment*, 114(12):2970–2980.

Verger, A., Baret, F., Weiss, M., Kandasamy, S., and Vermote, E. (2013). The cacao method for smoothing, gap filling, and characterizing seasonal anomalies in satellite time series. *IEEE transactions on Geoscience and Remote Sensing*, 51(4):1963–1972.

Verhoef, W., Menenti, M., and Azzali, S. (1996). Cover A colour composite of NOAA-AVHRR-NDVI based on time series analysis (1981-1992). *International Journal of Remote Sensing*, 17(2):231–235.

Viovy, N., Arino, O., and Belward, A. (1992). The Best Index Slope Extraction (BISE): A method for reducing noise in NDVI time-series. *International Journal of Remote Sensing*, 13(8):1585–1590.

Wahba, G. (1990). Spline models for observational data. *CBMS-NSF Regional Conference Series in Applied Mathematics; Society for Industrial and Applied Mathematics (SIAM)*.

Wan, Z. and Dozier, J. (1996). A generalized split-window algorithm for retrieving land-surface temperature from space. *IEEE Transactions on geoscience and remote sensing*, 34(4):892–905.

Wan, Z. and Li, Z.-L. (1997). A physics-based algorithm for retrieving land-surface emissivity and temperature from EOS/MODIS data. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4):980–996.

Wan, Z., Zhang, Y., Zhang, Q., and Li, Z.-l. (2002). Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data. *Remote Sensing of Environment*, 83(1):163–180.

Wang, J., Dong, J., Liu, J., Huang, M., Li, G., Running, S. W., Smith, W. K., Harris, W., Saigusa, N., Kondo, H., et al. (2014). Comparison of gross primary productivity derived from gimms ndvi3g, gimms, and modis in southeast asia. *Remote Sensing*, 6(3):2108–2133.

Wang, J., Price, K., and Rich, P. (2001). Spatial patterns of ndvi in response to precipitation and temperature in the central great plains. *International Journal of Remote Sensing*, 22(18):3827–3844.

Wang, R., Cherkauer, K., and Bowling, L. (2016). Corn response to climate stress detected with satellite-based ndvi time series. *Remote Sensing*, 8(4):269.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Yuan, X., Li, L., Chen, X., and Shi, H. (2015). Effects of precipitation intensity and temperature on ndvi-based grass change over northern china during the period from 1982 to 2011. *Remote Sensing*, 7(8):10164–10183.

Zeileis, A. (2006). Implementing a class of structural change tests: An econometric computing approach. *Computational Statistics & Data Analysis*, 50:2987–3008.

Zeileis, A., Kleiber, C., Krämer, W., and Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44:109–123.

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38.

Zhang, R., Ouyang, Z.-T., Xie, X., Guo, H.-Q., Tan, D.-Y., Xiao, X.-M., Qi, J.-G., and Zhao, B. (2016). Impact of climate change on vegetation growth in arid northwest of china from 1982 to 2011. *Remote Sensing*, 8(5):364.

Zhou, J., Jia, L., and Menenti, M. (2015). Reconstruction of global MODIS NDVI time series: Performance of Harmonic ANalysis of Time Series (HANTS). *Remote Sensing of Environment*, 163:217–228.

Zhu, X., Gao, F., Liu, D., and Chen, J. (2012). A modified neighborhood similar pixel interpolator approach for removing thick clouds in Landsat images. *IEEE Geoscience and Remote Sensing Letters*, 9(3):521–525.