Investigating the MetOp ASCAT vegetation parameters

Nicael A. Jooste

TU Delft

On the cover

The cover photo is a visualization of the curvature parameter currently derived in the TU Wien soil moisture retrieval approach. More specifically, the plotted lines belong to ASCAT grid points that contain a large percentage of cereal crops. In this study it was found that the curvature parameter exhibits particularly interesting seasonal behavior in such grid points, which is why they are highlighted on the cover page.

Investigating the MetOp ASCAT vegetation parameters

N.A. Jooste

In partial fulfilment of the requirements for the degree of

Master of Science
in Civil Engineering

at Delft University of Technology,
to be defended publicly on Friday March 20, 2020 at 14:00 PM.

Student number:	4156137	
Date:	6 March 2020	
Thesis committee:	Prof. Dr. ir. Susan Steele-Dunne,	TU Delft
	Dr. Stef Lhermitte,	TU Delft
	Dr. ir. Miriam Coenders,	TU Delft

Delft University of Technology
Faculty of Civil Engineering and Geosciences
Department of Water Management
Section Water Resource Management

TU Delft

Preface

Dear reader,

The document that lies before you marks the end of my Master studies in Water Management at Delft University of Technology. My time as a student has not always been easy, and I can say with certainty that this thesis has pushed my limits – it has been the hardest thing I have done to date. I have grown a lot over the past years, both academically and as a person, and I am very proud of what I have accomplished during this time. However, it would not have been possible without the support of the people around me.

First of all, I would like to thank my Susan Steele-Dunne, my daily supervisor; your guidance throughout this project has been truly great. Regardless of the problem I was facing, your door was always open and your advice was always honest. Thank you for your supervision and for giving me the opportunity to work on a very challenging yet interesting subject. Additionally, I want to thank Stef Lhermitte and Miriam Coenders for their support, discussions and valuable feedback.

To all of my dear friends, thank you for all the discussions, sparring sessions and emotional support. Everyone from room 4.84, I will never forget our shared fun and suffering. To my colleagues at Crunch Analytics, thank you for believing in me, being patient with me, and pushing me forward. I greatly appreciate having you as my colleagues and I am very excited about everything the future will bring us!

Finally, I want to thank my family. None of this would have been possible without your unwavering love, patience and support – both emotional as well as financial. You always offered me an ear and a shoulder especially during my lowest points, even though I know this has not been easy for you either. Another great source of inspiration that must be mentioned is my grandfather; throughout my life, he instilled in me the value of hard work and perseverance. He is no longer with us today, and I am deeply saddened that he did not get the opportunity to see me graduate – but I know he would have been extremely proud of my accomplishments.

N.A. (Nicael) Jooste
Delft, January 2020

Abstract

In this study, an unsupervised classification approach is used to investigate and characterize the spatial and temporal variability of MetOp-A ASCAT backscatter ($\sigma°$) data and the TUW SMR vegetation parameters across mainland France between 2007 and 2017. Currently, soil moisture data is retrieved from ASCAT backscatter measurements using the TU Wien Soil Moisture Retrieval (TUW SMR) approach. To correct for the influence of vegetation on soil moisture, two so-called 'vegetation parameters' are also estimated from the backscatter measurements. These vegetation parameters are the slope (σ') and curvature (σ'') of a second-order Taylor polynomial which describes the incidence angle dependence of backscatter.

While the slope was always seen as a measure for vegetation density, little research has been done into the value of the curvature as a source of information. However, a recent study by Steele-Dunne et al. [59] showed that both the slope and curvature contain significant information about vegetation phenology and vegetation water dynamics across the North-American grasslands, which suggests that the TUW SMR vegetation parameters are a potentially valuable source of information on vegetation dynamics. This study further investigates the value of the TUW SMR vegetation parameters as a source of information about vegetation dynamics for a number of land cover types present in mainland France.

The 3492 ASCAT grid points in France were separated into ten groups using agglomerative hierarchical clustering based on the climatology of σ' and subsequently analysed. The results show that clusters based on σ' are generally contiguous and are able to resemble distinct land cover features; areas such as Paris, the Alps, and the Landes forest are clearly visible in cluster maps. Even though it is expected that the clusters differ in terms of σ' – which inherently follows from a clustering based on σ' – the results show that the clusters generally have distinct and unique σ°_{40} and σ'' characteristics as well. This suggests that the clusters represent 'scattering surfaces' that differ in terms of their seasonal scattering characteristics.

In general, it was found that grid points with a heterogeneous land cover footprint tend to have noisy seasonal backscatter signatures, while those with a homogeneous land cover footprint tend to have clear and recognizable seasonal behavior. Additionally, the results suggest that certain backscatter signatures correspond to certain land cover footprints; in particular the predominantly agricultural area around Paris produced very specific σ°_{40}, σ', and σ'' signatures that correspond to specific growth stages of wheat and the rapid land cover change that occurs during the agricultural growth season. In general, the results are consistent with the previous assumptions that σ' is a measure for vegetation density and σ'' is a measure for the relative dominance of ground-bounce and direct scattering from vertical vegetation constituents.

Finally, clustering was performed on ten years of dynamically estimated σ' and a measure for robustness was introduced to quantify the certainty of clustering for each grid point. Very robust grid points are found in areas that have a relatively stable land cover such as the Alps or Paris, suggesting that areas with stable land cover exhibit predictable seasonal backscatter behavior with low interannual variability. On the other hand, poor robustness scores are mainly found in north-west France, where land cover is heterogeneous and seasonal backscatter behavior is highly variable, perhaps due to crop rotation.

This study confirms that the TUW SMR vegetation parameters contain valuable information about vegetation phenology in both homogeneous and mixed land cover footprints. Furthermore, it was shown that unsupervised classification methods based on the vegetation parameters are able to identify areas with similar scattering characteristics, and are able to show how these areas change over time.

Contents

List of Tables

List of Figures

Nomenclature

List of Abbreviations

AMI	Active Microwave Instrument	1
ASCAT	Advanced Scatterometer	1
CH	Calinski-Harabasz index	26
DB	Davies-Bouldin index	26
ERS	European Remote Sensing	1
ESA	European Space Agency	10
ESCAT	European Scatterometer	1
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites	10
MetOp	Meteorological Operational	1
PC	Principal Component	12
PCA	Principal Component Analysis	5, 12
SIL	Silhouette index	26
TU Wien	Vienna University of Technology	1
TUW SMR	TU Wien Soil Moisture Retrieval	1

List of Symbols

σ''	Curvature of the $\sigma(\theta)$ relationship [dB/deg^2]	1
ϵ	Electric permittivity [-]	6
ϵ_0	Permittivity of free space, equal to 8.8542×10^{-12} $C^2N^{-1}m^{-2}$	6
ϵ_r	Relative permittivity, also known as the dielectric constant [-]	6
σ	Scattering cross-section [m^2]	7
σ°	Normalized backscatter coefficient [dB]	1, 7
σ°_{40}	Normalized backscatter coefficient [dB] at a reference angle of 40°	1
σ'	Slope of the $\sigma^\circ(\theta)$ relationship [dB/deg]	1
θ	Incidence angle [deg]	1
θ_r	Reference incidence angle [deg]	1

1 Introduction

1.1 Context

Soil moisture and vegetation are key variables in the water, energy and carbon cycle, as they determine the energy and water fluxes between the soil surface and the atmosphere interface. Consequently, many scientific disciplines rely on soil moisture and vegetation data for developing, evaluating and improving their descriptive and predictive models; examples include climate forecasting [34, 35, 13], hydrological modeling [38, 62, 72, 8], drought monitoring [60, 25, 40], flood forecasting [73], forest monitoring [28], and agricultural drought detection [2]. However, data has not always been as accessible as it is today. In recent years, developments in the field of satellite remote sensing have dramatically improved our ability to monitor the Earth, leading to the wide availability of land surface data we have today.

At the core of this study is the Advanced Scatterometer (ASCAT), which is a C-band active microwave remote sensing instrument carried on board the series of Meteorological Operational (MetOp) satellites. Even though ASCAT was initially designed to measure wind speed and direction over the Earths oceans in support to numerical weather prediction, tropical cyclone analysis, and ocean waves forecasting, research carried out with its predecessor instrument – the Active Microwave Instrument (AMI) on board the ERS-1/2 satellites – confirmed the ability of C-band scatterometry to provide reliable soil moisture observations on a global scale [15, 1, 9]. Furthermore, many studies have shown that C-band scatterometer data correlates with the seasonal behavior of vegetation, indicating that C-band scatterometry may also be a potentially valuable source of information for vegetation monitoring [21, 30, 31, 19].

Several soil moisture products are derived from the ASCAT backscatter observations using the so-called TU Wien Soil Moisture Retrieval (TUW SMR) approach. This algorithm was first developed by the Vienna University of Technology (TU Wien) for ESCAT on-board the ERS-1/2 satellites [70] and was later translated to ASCAT [5, 45]. The TUW SMR algorithm uses a change detection approach to retrieve soil moisture. An important early step in the derivation of soil moisture is the normalization of all backscatter observations to the reference angle $\theta_r = 40°$; using the multiangle measurement capabilities of ASCAT, the TUW SMR algorithm is able to describe the incidence angle (θ) dependence of backscatter ($\sigma°$). The resulting $\sigma°(\theta)$ relationship – which is characterized by a second-order Taylor polynomial in the TUW SMR algorithm – is particularly important, as it contains information about soil moisture, vegetation characteristics and scattering mechanisms and their variations through time [24]. This relationship is used to normalize the backscatter measurements to the reference incidence angle (θ_r) and to correct for the effects of vegetation on the normalized backscatter signal.

Firstly, this step yields the normalized backscatter coefficient ($\sigma°_{40}$), which is used to determine soil moisture relative to the historically wettest and driest observations. Secondly, the vegetation correction is applied during this step; vegetation is characterized by the behavior of the slope (σ') and curvature (σ'') of the $\sigma° - \theta$ relationship, which are the so-called "vegetation parameters". An increase in vegetation cover is assumed to cause only a rotation of the $\sigma° - \theta$ curve (i.e. a change in σ' and σ''), while an increase in soil moisture causes only a vertical translation and no rotation of the $\sigma° - \theta$ curve, see Fig. 1.1. With these assumptions the TUW SMR algorithm is able to correct for the influence of vegetation.

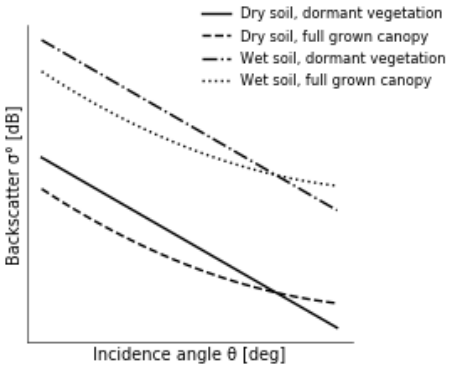


Fig. 1.1: Conceptual $\sigma(\theta)$ relationship and the influence of soil moisture and vegetation

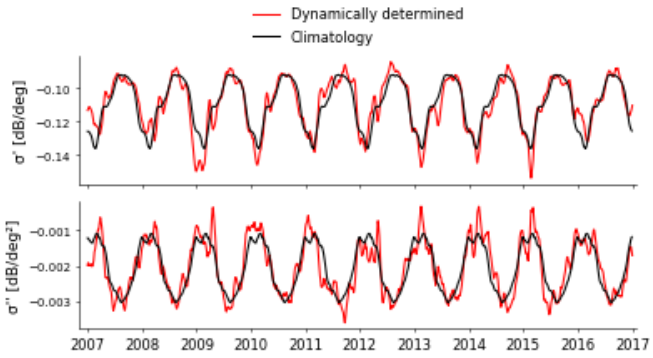


Fig. 1.2: Climatology vs. dynamically determined vegetation parameters

Due to constraints in data density and the significant amount of noise in the individual values, several years of data have to be combined in order to ensure robust estimates of the σ' and σ'' coefficients [70, 24]. As such, all existing ASCAT soil moisture products derived using the TUW SMR approach are based on a seasonal climatology of σ' and σ'' coefficients. This was particularly necessary for the ERS-1/2 scatterometer, which had only one set of three fan-beam antennas and hence, low data density. ASCAT has two sets of three fan-beam antennas – one for each swath – allowing for three independent backscatter measurements at three different azimuth angles and two different incidence angles for every pixel [71]. The increased data density unlocks the possibility to estimate the σ' and σ'' coefficients dynamically.

A recent algorithmic development by Melzer [42] allows the vegetation parameters to be estimated dynamically, making it possible to investigate the interannual variation of the vegetation parameters, see Fig. 1.2. Based on this new method, Steele-Dunne et al. [59] recently examined multiple years of ASCAT backscatter data to characterize spatial and temporal variability in the vegetation parameters across the North American Grasslands. In their study, data was aggregated across four Regions of Interest (ROIs), each defined by uniform Köppen-Geiger climate class (KGCC) and ecoregion. The seasonal climatology and interannual variability of the backscatter variables of each of the ROIs was then investigated. Their results show that the seasonal climatology, spatial patterns, and interannual variability of both σ' and σ'' vary considerably between the ROIs, but there does not seem to be a simple relationship between σ' and σ''. However, it may be preferable to define ROIs based on scattering characteristics instead of KGCC and ecoregion, so that each ROI describes a distinct scattering surface with unique scattering characteristics.

This study builds on the insights obtained by Steele-Dunne et al. [59] and is focused on exploring the possibility of identifying distinct scattering surfaces using unsupervised clustering, as well as investigating seasonal and interannual variation of the vegetation parameters based on these scattering surfaces.

1.2 Current challenges and knowledge gaps

One of the main concerns surrounding ASCAT has been that C-band has a lower sensitivity to soil moisture in the presence of vegetation compared to longer wavelengths [33] which makes it difficult to separate the individual effects of soil moisture and vegetation. Moreover, the intricate relationship between soil moisture and vegetation only adds to the complexity of the task at hand. Even though validation studies have shown that the quality of the ASCAT soil moisture product is generally comparable to (or even better than) currently available soil moisture data sets derived from passive microwave sensors [71] there are still considerable knowledge gaps surrounding the vegetation parameters of the TUW SMR algorithm.

As discussed previously, significant efforts have been undertaken to correct for the influence of vegetation in the retrieval of soil moisture. However, even though vegetation is corrected for in the current implementation of the TUW SMR algorithm, the physical meaning of the vegetation parameters is not

fully understood and is still an area of active study. A general consensus exists on the meaning of σ' as a measure for "vegetation density"; previous research has linked σ' to seasonal dynamics in wet vegetation biomass [69]. However, no direct link between σ' and (wet) biomass or vegetation water content exists.

Moreover, the meaning of σ'' is less well understood, and little research has been done into its potential value as a source of information about vegetation. In a recent study, Steele-Dunne et al. [59] showed that σ'' is clearly influenced by vegetation phenology, vegetation water content, as well as the distribution of water in the vegetation components. Furthermore, their results indicate that σ'' may contain valuable information about the drought response of grassland vegetation. While these results are reassuring, the potential value of the dynamically determined σ' and σ'' as a source of information about vegetation dynamics must be further investigated in other land cover types.

1.3 Research aim

The aim of this research is to investigate the temporal and spatial characteristics of the backscatter coefficient (σ°_{40}) and TUW SMR vegetation parameters (σ' and σ''), in order to gain an improved understanding of their physical meaning and behaviour, as well as to further explore their value as a source of information. In order to achieve this aim, the main research question has been defined as followed:

Can distinct scattering surfaces be identified and used to obtain an improved understanding of the observed σ°_{40}, σ', and σ'' behavior?

In order to answer the main research question, the following sub-questions are defined:

1. *Which data preprocessing and clustering techniques are required and suited to solve this problem?*

   Many techniques exist for the purpose of (unsupervised) classification. However, a 'one size fits all' solution does not exist. Hence, the first question is dedicated to exploring data preprocessing and classification techniques so as to gather a suitable set of tools for solving the problem at hand.

2. *Can distinct and meaningful scattering surfaces be identified using unsupervised classification?*

   A set of clusters is generated using an unsupervised classification approach based on the climatology of σ'. The input data must first be preprocessed to ensure meaningful results. The obtained clusters are investigated in terms of their characteristics and their σ°_{40}, σ', and σ'' signatures in order to investigate whether they are distinct and meaningful. The generated clusters should not be noisy and should ideally bear a resemblance to land cover.

3. *What is the influence of sub-footprint land cover heterogeneity on the σ°_{40}, σ', and σ'' signatures?*

   The observed σ°_{40}, σ', and σ'' signatures are compared to land cover data in order to better understand the relationship between sub-footprint land cover heterogeneity and seasonal scattering behavior, as well as to investigate how sub-footprint heterogeneity relates to the generated clusters.

4. *Are the grid points and generated clusters "robust"?*

   A measure for robustness is proposed and used to determine how well each grid point belongs to its assigned cluster. Robustness scores are compared to land cover footprint to investigate whether a relationship exists between land cover and robustness.

1.4 Research significance

Nowadays, many real-world applications such as numerical weather prediction and hydrological modelling rely on the ASCAT soil moisture products in their day-to-day operation, and other potential applications are increasingly being identified. For example, initial results for the application of ASCAT data products in other areas of expertise such as crop yield monitoring, epidemic risk modelling, and societal risk assessments are positive, but still require additional improvements [71]. As such, it is imperative that the ASCAT soil moisture data products is as accurate and reliable as possible.

As previously discussed, the TUW SMR vegetation parameters are essential components for vegetation correction in the derivation of soil moisture from ASCAT backscatter. Having an improved understanding of the vegetation parameters will help improve the soil moisture derivation from backscatter observations, ultimately leading to increased quality of soil moisture data products. This benefits all existing and (currently unidentified) future applications that make use of these data products.

Furthermore, previous research has indicated the potential value of the TUW SMR vegetation parameters as a source of information about vegetation. A better understanding of the vegetation parameters may lead to new ASCAT derived data products describing seasonal and interannual vegetation dynamics. Such new vegetation data products may provide significant value to a diverse set of current and future operational applications and a wide range of scientific research fields, especially considering the large archive of existing ASCAT data. Moreover, the promising future of the MetOp mission and next generation MetOp-SG mission ensure the availability of backscatter data until at least 2040, which further underlines the potential value of improving and expanding the existing suite of ASCAT data products.

1.5 Thesis outline

Any terms, acronyms, symbols and operators that may be unknown to the reader are explained in the Nomenclature. Chapter 2 covers most of the background theory that support the performed analyses, and will be of interest mainly to readers that are not familiar with ASCAT, the TUW SMR algorithm, and/or (unsupervised) classification techniques. Chapter 3 describes the study area as well as the different data and methods that are used in this study. The most important results are presented and discussed in Chapter 4. Additionally, all results – including those less relevant – can be found in Appendix C. Finally, the main conclusions and answers to all the research questions are presented in Chapter 5, together with any assumptions and limitations of the performed work, recommendations for further research, and the contributions of this study.

2 Theoretical Background

This chapter provides an overview of relevant background theory for this research. A short introduction to remote sensing is given in section 2.1. Several important concepts in microwave scatterometry are treated in section 2.2. Section 2.3 gives a synopsis of the ASCAT instrument, and section 2.4 serves to provide the necessary theory about the TUW SMR algorithm, which is at the core of this study.

The remainder of this chapter is focused on theory of data pre-processing techniques and unsupervised classification. Section 2.5 explains the concept of Principal Component Analysis (PCA) and its value in this research. Finally, section 2.6 provides an introduction to clustering as well as a description and comparison of two popular clustering algorithms: agglomerative hierarchical clustering and k-means clustering.

2.1 Remote sensing

Remote sensing is defined as the process of obtaining information about a target object without making actual physical contact with it. This definition is most often used to describe the acquisition of information about the Earth's surface and atmosphere using different remote sensors, i.e. devices that measure one or multiple types of electromagnetic radiation reflected or emitted by a target object. The different types of electromagnetic radiation together make up the continuous electromagnetic spectrum, and can be divided into the following regions: (long) radio waves, microwaves, far- and near infrared waves, visible light, ultraviolet, X-rays, and gamma-rays, see Fig. 2.1 [17].

Remote sensing platforms can be terrestrial, airborne, or spaceborne and can be classified into passive or active systems. Passive sensors detect natural electromagnetic radiation that is emitted or reflected by the target object. In most cases, reflected sunlight is the radiation source for passive sensors. Examples of passive sensors include photographic cameras, radiometers and spectrometers. On the other hand, active sensors illuminate the target by emitting radiation themselves, and subsequently measure the amount of radiation that is reflected from the observed target. An example of an active sensor is ASCAT, which is a radar scatterometer operating in the microwave region of the electromagnetic spectrum.

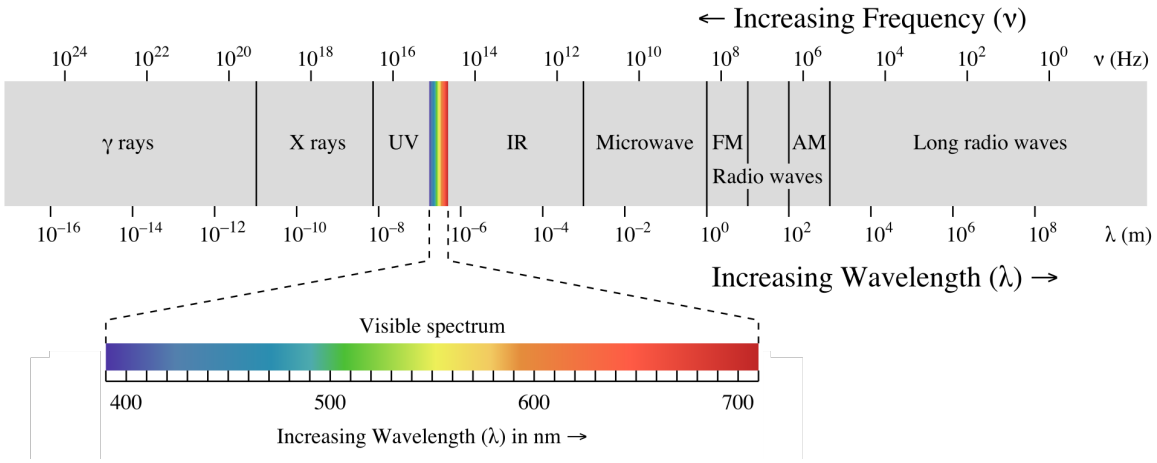


Fig. 2.1: The electromagnetic spectrum

2.2 Microwave remote sensing

The use of microwaves for remote sensing is a relatively recent phenomenon, having been in use only since the early 1960s. There are some disadvantages to using microwaves in remote sensing. As a result of the long wavelengths of microwaves, active microwave instruments require relatively large antennas to achieve acceptable spatial resolutions. Furthermore, active microwave instruments are large and heavy and consume a significant amount of power.

However, microwaves have several characteristics that make them valuable for remote sensing applications. Microwaves are able to penetrate vegetation more deeply than optical waves, with longer wavelengths penetrating better than shorter wavelengths. Microwaves are also relatively unaffected by clouds and rain, and are able to penetrate the top soil layer. Moreover, microwave interactions are generally controlled by other physical parameters compared to different types of electromagnetic radiation, meaning that microwaves contain different information. As such, the use of microwaves provides observation capabilities that assist methods in other spectral regions.

Even though the field of microwave remote sensing is relatively young, a significant amount of research on microwaves and their use in remote sensing has been carried out; clearly, this research is unable to fully capture the existing body of background knowledge. Several important concepts pertaining to microwave scatterometry are treated in this section, and we refer to the literary works of Ulaby [65, 66, 67], Woodhouse [75] and Rees [52] for an excellent and comprehensive overview of the fundamental principles of microwaves and microwave remote sensing.

2.2.1 Microwave propagation

In order to understand how microwaves interact with the real world, several concepts related to microwave propagation must be understood first. Every material reacts differently to electromagnetic radiation: in the case of visible light, it is general knowledge that glass is transparent, wood is opaque, mirrors are reflective, and water is refractive. The same holds true for the microwave region of the electromagnetic spectrum: materials can be transparent, opaque, reflective, refractive, and so forth. These effects are the result of the electromagnetic properties of a material, which are described by: the *electric permittivity* (ϵ); the *magnetic permeability* (μ); and the *electric conductivity* (g).

The electric conductivity describes the (lack of) mobility of electrons in a material. For example, electrons are free to move in metals, which translates to a high electric conductivity. The magnetic permeability describes how well a material is able to maintain the establishment of a magnetic field within itself, i.e. it quantifies the amount of magnetization in a material under an imposed magnetic field. The magnetic permeability of a vacuum is equal to $\mu_0 = 4\pi \times 10^{-7}\ Ns^2C^{-2}$ and $\mu \approx \mu_0$ for nonmagnetic materials (i.e. most practical applications). While it is important to understand these material characteristics, they do not need to be considered in detail for most remote sensing purposes.

On the other hand, the electric permittivity is a very important material characteristic for microwave remote sensing. The electric permittivity is a measure for the capacitance of a material under influence of an electric field – it describes how well the molecules of a medium polarize in an electric field. The higher the electric permittivity of a medium, the better its molecules polarize and the more that medium is able to resist the imposed electric field. Similar to the magnetic permeability, the *vacuum permittivity* (ϵ_0) is equal to $8.8542 \times 10^{-12}\ C^2N^{-1}m^{-2}$. The permittivity of a dielectric medium is then defined as the product of the vacuum permittivity and the *relative permittivity* of a medium (ϵ_r), see Eq. 2.1.

$$\epsilon = \epsilon_r \epsilon_0 \qquad [-] \tag{2.1}$$

Most solid materials on Earth are non-conducting; such a material is also called a *dielectric*. Microwaves lose energy exponentially as they travel through a dielectric material, i.e. the incoming waves are *attenuated*. The dielectric properties of a medium are described by the relative permittivity, which consists of a real and complex part so that $\epsilon_r = \epsilon_r' - i\epsilon_r''$. The real part of the complex electric permittivity (ϵ_r') is also known as the *dielectric constant*, while the imaginary part ($i\epsilon_r''$) describes wave attenuation.

The value of microwaves for remote sensing is in part due to the dielectric properties of water, mainly in the form of soil moisture and vegetation moisture. Water is a good conductor due to the permanent electric dipole of its molecules, and has a relatively high relative permittivity of approximately 80 in the microwave region while most dry materials have a dielectric constant in the range of 3 – 8 [75]. As such, the presence of moisture in both soil or vegetation significantly increases the strength of the return signal, making it possible for ASCAT to observe changes in soil- and vegetation moisture.

2.2.2 Scattering

2.2.2.1 The backscatter coefficient

When an electromagnetic wave reacts with a target object, the incident wave may be redirected into different directions compared to the incidence direction in a process called *scattering*, which is a fundamental concept in many remote sensing applications. The effectiveness of a scattering object is described by the *scattering cross-section* (σ), which quantifies scattering in all directions. However, for active systems such as ASCAT the amount of incident energy ($I_{incident}$) to the target area is known, and such systems measure the amount of energy that is scattered back (*backscatter*) by the target object to the sensor along the incidence direction ($I_{received}$) at a range R. Hence, for active systems it is more useful to define the backscatter using the proportion of returned energy over incident energy. This is called the *radar scattering cross-section*, which is defined as:

$$\sigma = \frac{I_{received}}{I_{incident}} 4\pi R^2 \qquad [m^2] \tag{2.2}$$

Like most active instruments, ASCAT observes the backscatter generated by an extended area (or *footprint*) rather than individual objects. To ensure that ASCAT observations can be compared to those of other instruments, a backscatter measure independent of the instrument footprint is required. One such measure is the *(normalized) backscatter coefficient* (σ°), which is also known as the *normalized radar cross-section* or *sigma nought*. The backscatter coefficient relates the radar scattering cross section defined by Eq. 2.2 to the satellite footprint A and is defined as:

$$\sigma^\circ = \frac{\sigma}{A} \qquad [-] \tag{2.3}$$

The backscatter coefficient is unitless (m^2/m^2) and quantifies the average reflectivity of a target normalized to a unit area on the horizontal ground plane. As such, σ° is independent of footprint size, meaning that it is a feature of the observed target and not of the measuring instrument, which ensures that σ° observed by different instruments can be directly compared.

2.2.2.2 Parameters affecting radar backscatter

Backscatter characteristics depend on a combination of different parameters; both surface parameters and instrument features determine the observed scattering behavior. Important surface parameters are dielectric properties (mainly due to moisture content), roughness, and geometric shape. Instrument features affecting backscatter are frequency, polarization, and incidence angle. However, since ASCAT has a fixed pulse magnitude, frequency and polarization, their influence are not further treated here.

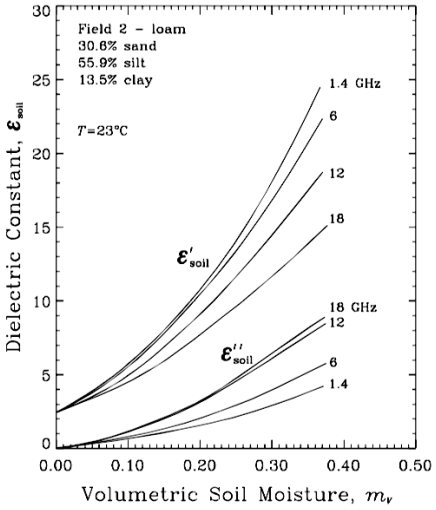


Fig. 2.2: Relationship between soil moisture content and dielectric constant (Ulaby et al. [67])

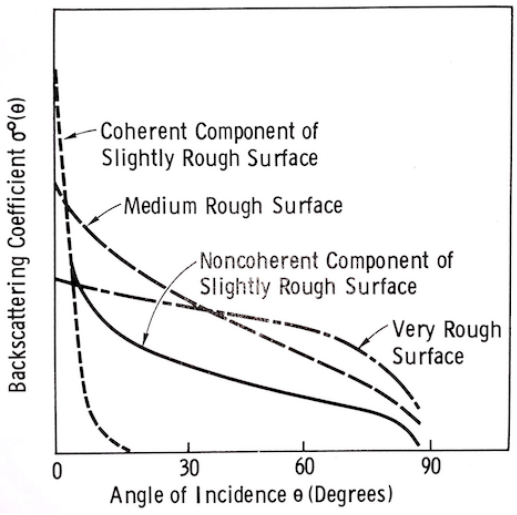


Fig. 2.3: Influence of incidence angle on surface roughness effects (Ulaby et al. [67])

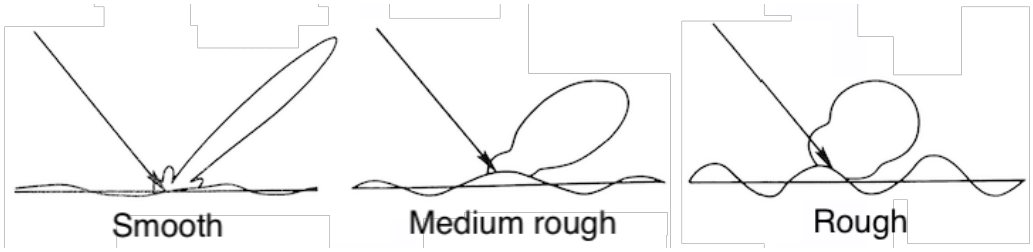


Fig. 2.4: Effect of increasing surface roughness on scattering patterns

Dielectric constant

The dielectric constant depends on many variables, such as frequency, temperature, water content of soil and vegetation, soil texture, and salinity [14]. Additionally, the different components of the dielectric constant discussed in section 2.2.1 (i.e. ϵ' and ϵ'') are sensitive to different variables. However, both the real and complex part of the dielectric constant are highly sensitive to water content.

In the microwave range, the dielectric constant of water is an order of magnitude larger compared to the dielectric constant of dry materials. For soil this means that the dielectric constant is larger for wet soil than for dry soil, which is shown for different soil types in Fig. 2.2 [67, 57]. Since an increase of the dielectric constant is associated with increased direct backscatter, the contribution of the soil to total backscatter will be larger for surfaces with higher soil moisture content.

Similarly, the dielectric constant of vegetation is strongly influenced by vegetation water content; the dielectric constant of dry vegetation is significantly lower compared to the dielectric constant of vegetation with high moisture content [7, 63]. Consequently, vegetation with a high moisture content is associated with larger direct backscatter as well as higher attenuation of the incident wave.

Roughness

Surface roughness is a relative concept which depends on the emitted wavelength of the instrument and the incidence angle. In general, a surface is considered to be rough if the dimensions of its structure are similar in size to the incident wavelength. As shown in Fig. 2.4, a rough surface scatters an incident wave partially in the specular direction and partially in all directions. Specular scattering becomes negligible for increasingly surface roughness and instead, diffuse scattering dominates. Furthermore, this relationship between surface roughness and backscatter changes with the incidence angle as shown in Fig. 2.3.

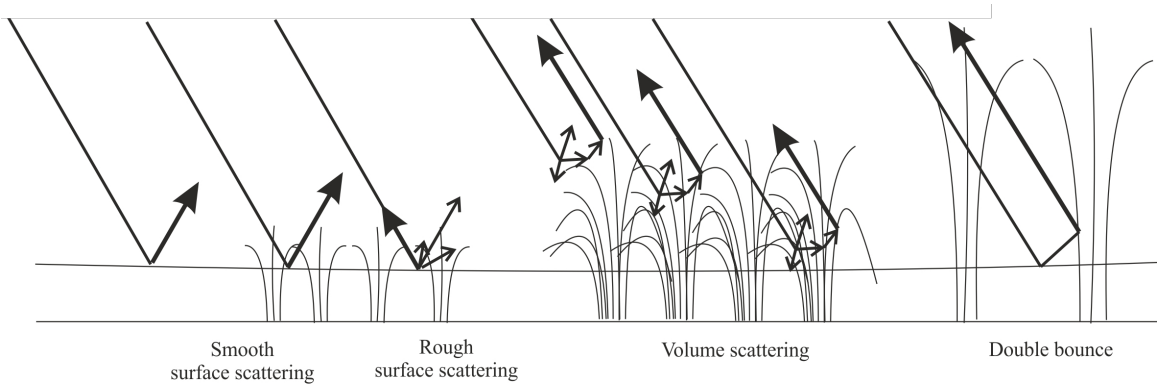


Fig. 2.5: Different scattering mechanisms occur due to the changing geometry of the growing crop.

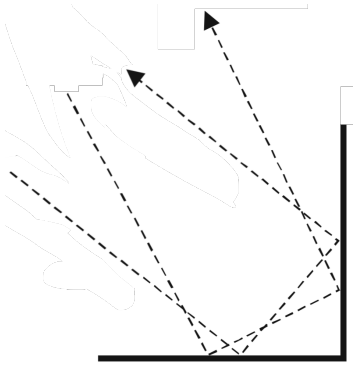


Fig. 2.6: Corner reflector

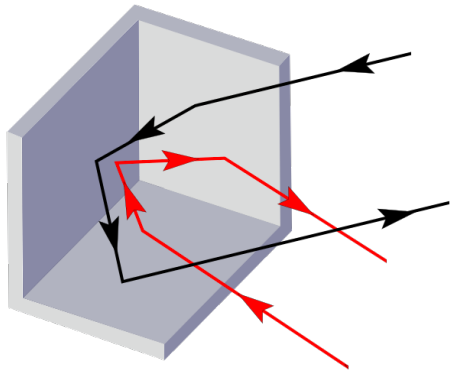


Fig. 2.7: Triple corner reflector

Incidence angle

As discussed previously, the incidence angle strongly influences the amount of surface scattering from rough surfaces. Generally, low incidence angles return the largest backscatter from surface scattering regardless of the surface roughness (see Fig. 2.3). The incidence angle also strongly influences backscatter from and attenuation by vegetation. For low incidence angles, the path through the vegetation layer becomes longer and consequently, attenuation of the incident wave increases.

Geometrical shape

The geometry of vegetation consists of many different structural elements, all differing in size, density and orientation. Such scattering elements include leaves, branches, fruits, flowers, and stems, all of which cause a different type of scattering response. Furthermore, certain types of vegetation (such as agricultural crops) have a rapidly changing geometry and consequently, rapidly changing scattering behavior, see Fig. 2.5. As discussed earlier in this section, the geometry of the ground surface matters too; rough surfaces generally lead to more diffuse scattering.

An interesting example of the effect of geometry are corner reflectors, such as the ones shown in Fig. 2.6 and Fig. 2.7. Corner reflectors are made by joining two or three smooth, reflective surfaces at a 90° angle. As a result of their geometry, corner reflectors are able to generate very high backscatter over a range of incidence angles. This type of geometry is generally seen in man-made structures as well as in urban, residential, and industrial areas. This is why urban areas generally have high backscatter returns.

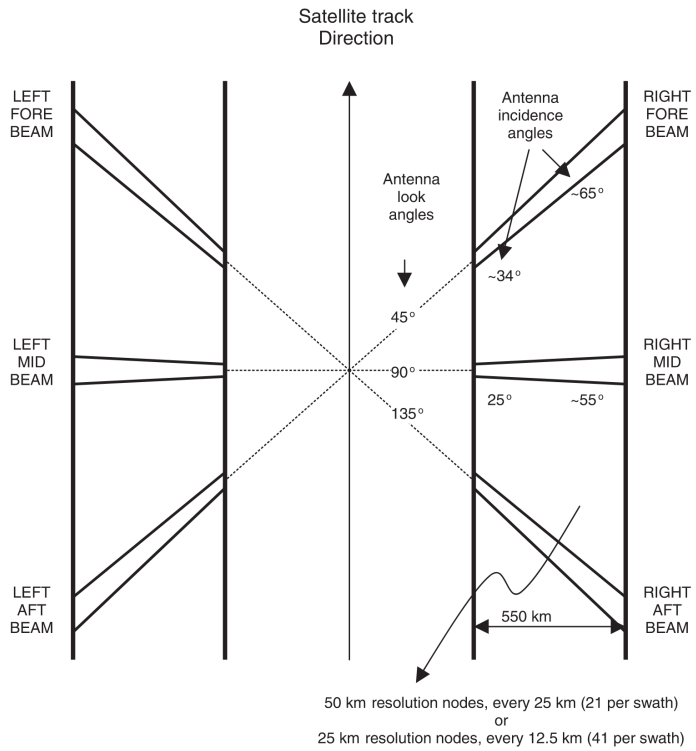


Fig. 2.8: Geometry of the ASCAT swath (Figa-Saldaña et al. [20])

2.3 ASCAT on-board MetOp

The MetOp satellite programme is a joint undertaking by the European Space Agency (ESA) and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) that provides weather data services for climate monitoring and weather forecasting. The mission consists of three satellites – MetOp-A, MetOp-B, and MetOp-C – which were launched in October 2006, September 2012, and November 2018 respectively. Initially, the MetOp satellites were intended to be operated sequentially. However, because the in-orbit performance of MetOp-A and MetOp-B exceeded expectations, all three MetOp satellites are now planned to be operated simultaneously until the de-orbiting of MetOp-A in 2022. The MetOp satellites fly in a sun-synchronous orbit and are equally spaced in orbit around 120° apart. Each satellites completes 14 orbits per day and has a daily global coverage of approximately 82% [71].

On board the MetOp satellite series is ASCAT, which is an active microwave instrument that was initially designed for measuring ocean wind vectors for numerical weather prediction and climate research. Even though ASCAT was initially not supposed to support operational services over land, it has proven its usefulness in a number of other applications. Nowadays, ASCAT backscatter data also supports areas such as the monitoring of land-ice, sea-ice, snow cover, and soil moisture. Moreover, recent research indicates that ASCAT backscatter data may be a valuable source of information for vegetation monitoring [59].

Like its predecessor instrument, ASCAT is a fixed fan-beam scatterometer that operates at 5.255 GHz (C-band) and VV polarization [71]. ASCAT carries two sets of three side-ways looking antennae, each covering a 550 km wide swath to the right and left side of the satellite ground track. Compared to the AMI on-board ERS-1/2, ASCAT has double the spatial coverage as well as an improved spatial resolution of 25 x 25 km^2. The three antennae on each side take backscatter measurements in different directions and are oriented at 45° forward, 90°, and 45° backward. The incidence angle ranges of the antennae are 34–65° for the fore and aft antennae, and 25–55° for the middle antenna [59]. The ASCAT swath geometry and antenna configuration are visualized in Fig. 2.8.

ASCAT consecutively emits short and well-characterised microwave pulses from each of its antennae and records the resulting echoes. If a point on the Earth's surface falls within either swath, it will be seen by the aft-, mid-, and forebeam antennae on that side of the satellite. Hence, three independent backscatter measurements (i.e. a *backscatter triplet* of σ°_a, σ°_m and σ°_f) are made for each location, with each measurement taken at three different azimuth angles and two different incidence angles [20].

The ability of ASCAT to simultaneously obtain a backscatter triplet allows for the calculation of an instantaneous backscatter slope, which is also known as the *local slope*. As will be discussed in section 2.4, computation of the local slope is essential for the retrieval of soil moisture data from ASCAT backscatter observations, as local slope values are used to estimate both σ' and σ'' in the TU Wien soil moisture retrieval algorithm [24].

2.4 TU Wien Soil Moisture Retrieval (TUW SMR) algorithm

The TUW SMR algorithm is responsible for the derivation of several soil moisture products from ASCAT backscatter observations. Initially designed for the AMI on board the ERS-1/2 satellites, the TUW SMR method exploits the multi-incidence angle measurement capabilities of ASCAT to retrieve soil moisture content relative to the historically driest and wettest condition. This section serves to explain the general concept of the TUW SMR algorithm and to describe the derivation of soil moisture and the vegetation parameters (σ' and σ'').

2.4.1 Concept

Most approaches aimed at retrieving vegetation and soil properties from scatterometer observations rely on inversion methods based on physical approximations of the different scattering processes; such inversion methods generally have significant problems related to their parameterization and their physical validity at large scales is questionable [45]. However, the TUW SMR approach refrains from extensive physical modelling. Instead, it is based on an entirely different concept; it has been a *change detection method* since its conception, i.e. it is based on identifying differences between subsequent observations.

The main idea that underlies the TUW SMR approach is to express the surface soil moisture (Θ_s) relative to the historically lowest backscatter measurement (dry reference σ°_d) and highest backscatter measurement (wet reference σ°_w), assuming that the backscatter coefficient (σ°) and surface soil moisture are linearly related [42]. The surface soil moisture for some location at reference angle θ_r and a certain time t is then calculated using Eq. 2.4:

$$\Theta_s(t) = \frac{\sigma^\circ(\theta_r, t) - \sigma^\circ_d(\theta_r, t)}{\sigma^\circ_w(\theta_r, t) - \sigma^\circ_d(\theta_r, t)} \tag{2.4}$$

The backscatter coefficient is influenced by a combination of static and dynamic factors. Static factors such as soil composition, surface roughness and land cover are assumed to be constant in time at the scatterometer measurement scale [59]. In the TUW SMR approach, these static factors are accounted for by subtracting the dry reference σ°_d from the actual backscatter measurement σ° as shown in Eq. 2.4.

Soil moisture and vegetation are dynamic factors influencing the backscatter coefficient. In order to obtain a reliable soil moisture estimate, the variation of vegetation must be accounted for. Vegetation correction is performed when the ASCAT backscatter measurements are normalized to the reference angle θ_r using the relationship between backscatter coefficient and incidence angle. In the TUW SMR, the incidence angle dependence of backscatter is described by the following second order Taylor polynomial:

$$\sigma^\circ(\theta) = \sigma^\circ(\theta_r) + \sigma'(\theta_r) \cdot (\theta - \theta_r) + \frac{1}{2}\sigma''(\theta_r) \cdot (\theta - \theta_r)^2 \tag{2.5}$$

Eq. 2.5 can be rearranged to:

$$\sigma^\circ(\theta_r) = \sigma^\circ(\theta) - \sigma'(\theta_r)\cdot(\theta-\theta_r) + \frac{1}{2}\sigma''(\theta_r)\cdot(\theta-\theta_r)^2 \tag{2.6}$$

Once the slope ($\sigma'(\theta_r)$) and curvature ($\sigma''(\theta_r)$) are known, backscatter measurements can be extrapolated from any incidence angle to the reference incidence angle using Eq. 2.6. As discussed in section 1, this step yields the normalized backscatter (σ°) which is used in Eq. 2.4 to obtain the surface soil moisture. Moreover, using the knowledge that a change in vegetation state causes a rotation of the $\sigma^\circ(\theta)$ curve – i.e. a change in σ' and/or σ'' – the influence of vegetation on σ° is corrected for, as σ' and σ'' mediate the effect of vegetation on the $\sigma^\circ(\theta)$ relationship [42].

As previously mentioned, the local slopes are used to determine σ' and σ''. The local slope is an estimate of the first derivative of the $\sigma^\circ(\theta)$ relationship and is determined using the backscatter triplets (σ°_a, σ°_m and σ°_f) using Eq. 2.7:

$$\sigma'\left(\frac{\theta_m-\theta_{a/f}}{2}\right) = \frac{\sigma^\circ_m(\theta_m)-\sigma^\circ_{a/f}(\theta_{a/f})}{\theta_m-\theta_{a/f}} \tag{2.7}$$

Until recently, a large number of local slope values were combined in order to account for the noise in individual local slope values. However, Melzer [42] recently introduced a new method that allows for the dynamic estimation of σ' and σ''. This approach uses a kernel smoother with a fixed time window of 42 days ($\lambda = 21$ days) to compute a weighted linear fit of σ' and σ''. In other words, σ' and σ'' are estimated using all local slope values within the prescribed time window of 42 days, where local slopes are linearly weighted by their distance in time, with larger weights for those closer in time.

2.5 Principal component analysis

2.5.1 Concept

Principal Component Analysis (PCA) is an old, well-known and widely applied technique introduced by Pearson [46] in 1901 and later independently developed by Hotelling [27] in 1933. The main goal of PCA is dimensionality reduction of a data set containing many interrelated variables, while preserving as much variation present in the original data set as possible [32]. This is done by transforming the original data to a new set of uncorrelated variables called the Principal Components (PCs). The PCs are ordered so that the first *few* PCs contain most of the variation present in the *entire* original data set.

Besides being a dimensionality reduction method, PCA is also a *feature extraction* method, which is an important preparatory step in the application of machine learning algorithms (see section 2.6). This can be illustrated using the following example. Assume we have a set of ten variables. PCA allows us to create a set of ten "new" uncorrelated variables – the PCs – where each PC is a combination of each of the ten original variables. The number of PCs to retain is chosen by the user, and since the PCs are ordered from most important (containing a lot of the original variation) to least important (containing very little of the original variation) the most important PCs can be retained while the least important ones can be dropped. Hence, not only does PCA reduce dimensionality (which improves computation time), it also separates essential information from non-essential information.

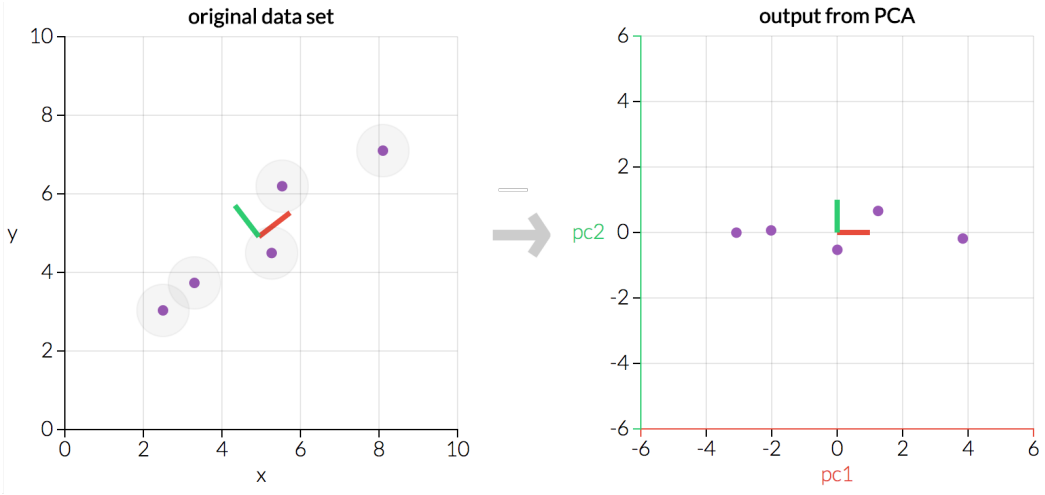


Fig. 2.9: An example of the application of principal component analysis on five data points. The left graph visualizes the original data $\mathbf{X}$ while the right graph visualizes the transformed data $\mathbf{Z}^*$ (setosa.io).

2.5.2 Derivation of principal components

Assume a data set in the form of a matrix $\mathbf{X}$. First, each column of $\mathbf{X}$ is standardized to ensure that the importance of the features is independent of their variance. Standardization of $\mathbf{X}$ is done using Eq. 2.8:

$$\mathbf{Z}(m,n) = \frac{\mathbf{X}(m,n) - \mu_m}{s_m} \tag{2.8}$$

where $\mathbf{Z}$ is the resulting standardized matrix and the mean (μ_m) and standard deviation (s_m) of each column m are calculated using Eq. 2.9 and Eq. 2.10, respectively:

$$\mu_m = \frac{1}{N}\sum_{n=1}^{N}\mathbf{X}(m,n) \tag{2.9}$$

$$s_m^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{X}(m,n) - \mu_n)^2 \tag{2.10}$$

Next, the covariance matrix of $\mathbf{Z}$ is determined by calculating $\mathbf{Z}^T\mathbf{Z}$. An eigendecomposition is then performed on the covariance matrix $\mathbf{Z}^T\mathbf{Z}$ which yields $\mathbf{PDP}^{-1}$, where $\mathbf{P}$ is the matrix containing the *eigenvectors* of $\mathbf{Z}$ and $\mathbf{D}$ is a diagonal matrix containing the *eigenvalues* of $\mathbf{Z}$ on its diagonal. Each element on the diagonal of $\mathbf{D}$ is associated to a corresponding column in $\mathbf{P}$, i.e. the first element of $\mathbf{D}$ is the eigenvalue λ_1 and its corresponding eigenvector is the first column of $\mathbf{P}$. The "importance" of each eigenvector is (roughly) described by its corresponding eigenvalue, as each eigenvalue measures the amount of variance in the original data in the direction of its corresponding eigenvector. As such, the eigenvector that corresponds with the largest eigenvalue is the first and most important principal component.

The eigenvalues $\lambda_1 \ldots \lambda_n$ are then sorted from largest to smallest and the eigenvectors are sorted accordingly, yielding the sorted matrix of eigenvectors $\mathbf{P}^*$. Finally, the matrix $\mathbf{Z}^* = \mathbf{ZP}^*$ is calculated, with $\mathbf{Z}^*$ containing the PCA projections (also called *PCA scores*) of the standardized version of the original matrix $\mathbf{X}$. In other words, the matrix $\mathbf{Z}^*$ consists of weights, each of which belonging to a different eigenvector. The original data can also be reconstructed by multiplying each weight in a column with its corresponding eigenvector and adding μ_m. To illustrate, the difference between the matrices $\mathbf{X}$ and $\mathbf{Z}^*$ are visualized in Fig. 2.9.

2.5.3 Determining the number of principal components

It is important to know how many PCs can be dropped before significant information loss occurs. Many rules exist for determining how many PCs should be retained in order to adequately account for the variation present in the original data (**X**) or in the standardized data (**Z**). One of the most straightforward rules for choosing the number of PCs – which is also the method used in this research – is to determine a desired minimum threshold for the cumulative percentage of total variation that the PCs should account for. For example, if a threshold of 90% is chosen, then the number of PCs that should be retained is the smallest number for which at least 90% of the total variance present in the original data is explained.

Practical threshold values for the amount of explained variance are generally in the range of 70–90% [32]. However, the threshold may be lower or higher depending on the original data set; for example, when the first and second PCs represent overly dominant and obvious sources of information and if the smaller and less obvious variations are of particular interest, setting a threshold larger than 90% may be more appropriate. However, setting the threshold too high may result in too many PCs being retained, which complicates the interpretation of additional analyses and increases computation time. On the other hand, setting a too low threshold (e.g. $< 60\%$) may result in significant loss of information as too few PCs are retained. It should be noted that many more methods and rules of thumb exist for determining how many PCs to retain, some of which have been developed from a strong statistical point of view.

2.5.4 Benefits and drawbacks of PCA

Because PCA is a straightforward and non-parametric method for separating relevant information from irrelevant information, it has become a staple data analysis technique across a wide range of disciplines. Even though PCA clearly provides a number of benefits as indicated by its widespread use, every method also has (potential) disadvantages that one should be aware of. This section serves to discuss some of the benefits and drawbacks of PCA.

Since PCA makes no special assumptions on the input data it can be applied to nearly all numerical datasets, which is one of the reasons for why PCA is nowadays applied in many fields. PCA can be applied to small datasets, but becomes particularly useful when applied to large datasets (both in terms of objects and variables) for a number of reasons; firstly, by extracting the most important information and dropping redundant information and noise, PCA can reduce overfitting as a result of having too many variables. The resulting dimensionality reduction has the added benefit of improving algorithm performance, since a smaller data set directly translates to lower computation time. This is especially valuable for computationally heavy algorithms such as agglomerative hierarchical clustering (see section 2.6), which has a time complexity of $\mathcal{O}(n^3)$ and requires a memory of $\mathcal{O}(n^2)$.

However, because PCA transforms a data set into a combination of independent weights, eigenvectors and eigenvalues, the data produced by PCA can be difficult to interpret. Moreover, the meaning of the original variables may be harder to assess based on the obtained PCs. Another drawback of PCA is directly related to dimensionality reduction; even though PCA aims to explain as much as possible the variance present in the original features, a reduction of dimensionality inherently comes with a loss of information. This is particularly true when the PCs are not selected with care, which may result in the user unknowingly omitting important information present in the original data.

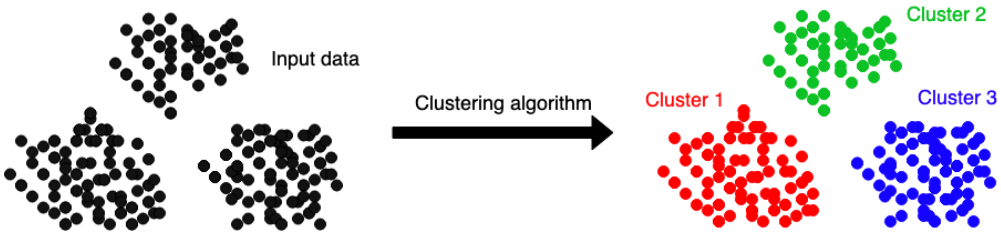


Fig. 2.10: Example of clustering, where the input data is separated into three clusters.

2.6 Clustering

Clustering (also called *unsupervised classification*, *cluster analysis*, or *segmentation analysis*) is a popular data analysis technique for classifying a group of observations into several distinct groups called *clusters*. In general, the act of data clustering separates a data set into several clusters in such a way that observations belonging to the same cluster have similar characteristics, while observations in other clusters have different (or *dissimilar*) characteristics, see Fig. 2.10.

Obtaining a set of clusters is rarely the end goal of a cluster analysis. Rather, the generated clusters give insights into the structure of the data, which can aid the user in developing a better understanding of the data. Hence, clustering should instead be seen as a knowledge-building tool for suggesting hypotheses [76]. In the context of this research, clustering is used to define groups of grid points which may or may not have distinct scattering characteristics. By investigating and comparing the obtained clusters, a better understanding of the vegetation parameters may then be obtained.

Many clustering algorithms have been developed, each with their own assumptions on what exactly constitutes a cluster and how to efficiently determine them. There does not exist one "best clustering algorithm" that consistently outperforms the others, as different methods may be appropriate depending on the type of problem and the type of data under investigation. Two popular clustering methods are examined in the remainder of this section; *hierarchical clustering* and *k-means clustering*.

2.6.1 Hierarchical clustering

Hierarchical clustering methods seek to group data in a hierarchical structure based on a certain similarity metric, yielding a sequence of nested clusters. In general, hierarchical clustering algorithms can be classified into *agglomerative* methods or *divisive* methods, see Fig. 2.11. Agglomerative hierarchical clustering algorithms are bottom-up algorithms that start with singleton clusters (i.e. clusters consisting of a single element) and end with a single cluster containing all elements. During each step of the algorithm, the two clusters that are most similar are merged into a single cluster. Conversely, divisive algorithms are top-down approaches that work the other way around, starting with one cluster containing all elements and splitting clusters along the way until each cluster consists of only one element.

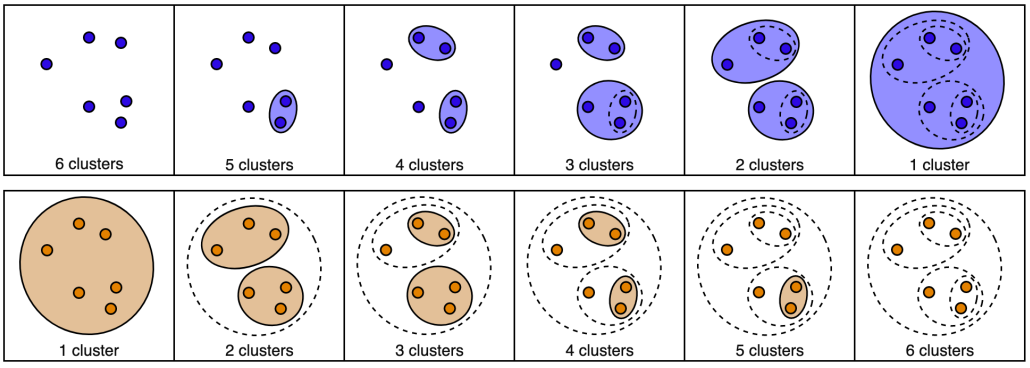


Fig. 2.11: Top: agglomerative clustering, which starts with singleton clusters and merges clusters until one cluster remains. Bottom: divisive clustering, which starts with one cluster and splits clusters until singleton clusters remain.

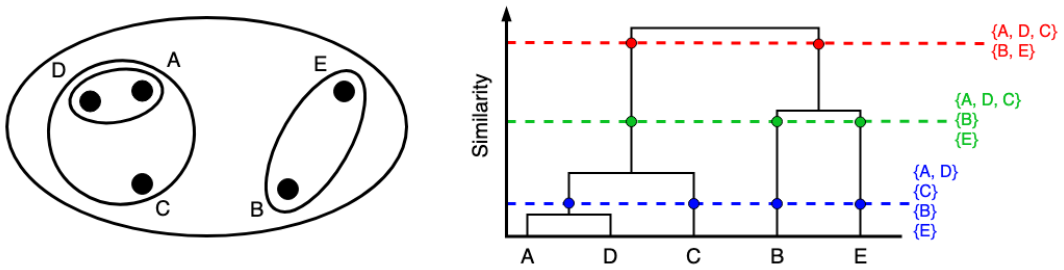


Fig. 2.12: Example of the nested clusters obtained using hierarchical clustering (left), visualized by a corresponding dendrogram (right). The data points ($A \dots E$) are clustered based on the distances between them. The height of each sub-tree represents the similarity between each of the clusters along the hierarchy, and a set of clusters is obtained by cutting across the dendrogram.

Even though both approaches are able to determine the hierarchical structure present in a data set, it should be noted that divisive algorithms are not commonly used due to their computational inefficiency [48]. On the other hand, agglomerative algorithms are much more simple and efficient [22].

After the hierarchical clustering has been performed, the obtained hierarchy of clusters can be depicted in a *dendrogram*, which is a tree-like graph that describes how easily certain (nested) clusters can be merged based on how similar they are (Fig. 2.12). Moreover, a dendrogram visualises the order of cluster merges or splits and how all nested clusters relate to each other in the hierarchy. After the dendrogram has been determined, a specific set of clusters can be obtained by cutting across the dendrogram horizontally at a certain height. As shown by the dashed lines in the dendrogram depicted in Fig. 2.12, different sets of clusters are obtained when placing horizontal cuts in a dendrogram at different levels; the blue line returns four clusters, the green line returns three clusters, and the red line returns two clusters. Hence, hierarchical clustering can be performed without knowing how many clusters to generate.

2.6.1.1 Distances and similarities

As previously stated, clusters are defined as groups containing similar elements while the elements of different clusters are dissimilar. In order to determine which clusters to merge (for agglomerative algorithms) or split (for divisive algorithms), a measure for the similarity or dissimilarity between the different data objects is needed. To quantify how similar two data points or two clusters are, similarity metrics are used: the greater the similarity metric, the more similar are the two data points or two clusters. Conversely, dissimilarity metrics quantify the differences between data points, with a greater dissimilarity metric implying a larger difference between two data points or two clusters.

The concepts of similarity and dissimilarity are generally synonymous with the concept of distance, as small distances between objects imply similarity, while large distances imply dissimilarity. In hierarchical clustering, similarity is described by a *distance metric* combined with a *linkage criterion*. The distance metric determines how the distance between individual objects should be calculated, while the linkage criterion determines how the distance between different clusters should be determined, as a function of the distances between their individual objects [76].

Distance metrics

Many functions exist for calculating the distance between observations in the case of continuous data sets, with each distance function describing a different geometrical view of the data. Examples of distance functions include Manhattan distance, Minkowski distance, Tchebyschev distance, Canberra distance, and cosine similarity [22]. However, perhaps the most well-known distance function is the Euclidean distance, which is also the distance that is usually implied when we speak of "distance" in the English language. Since there are too many distance functions to treat each of them individually, three examples are illustrated in this section: Euclidean distance, Manhattan distance, and Minkowski distance. The differences between these metrics are best illustrated when considering two data points, see Fig. 2.13.

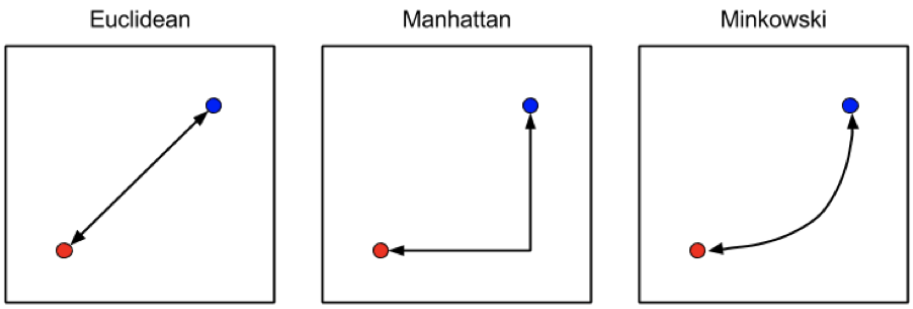


Fig. 2.13: Euclidean distance, Manhattan distance, and Minkowski distance visualised.

The Euclidean distance is the distance of the shortest straight line that connects two points in n-dimensional space. Assume we have two n-dimensional data points $\mathbf{p} = (p_1, p_2, \dots p_n)$ and $\mathbf{q} = (q_1, q_2, \dots q_n)$. The Euclidean distance $d_{euc}(\mathbf{p}, \mathbf{q})$ between points $\mathbf{p}$ and $\mathbf{q}$ can then be calculated using Eq. 2.11:

$$d_{euc}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{2.11}$$

The Manhattan distance (also called the *taxicab distance*) is obtained by taking the sum of the horizontal and vertical distances between two points on a grid. It is named after the shortest path a car would be able to take on the island of Manhattan, where the streets are aligned in a grid layout. The Manhattan distance $d_{man}(\mathbf{p}, \mathbf{q})$ between points $\mathbf{p}$ and $\mathbf{q}$ is calculated using Eq. 2.12:

$$d_{man}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i| \tag{2.12}$$

The Minkowski distance is a generalization of the Euclidean distance ($a = 2$), the Manhattan distance ($a = 1$), and the Chebyshev distance ($a \to \infty$). The Minkowski distance $d_{min}(\mathbf{p}, \mathbf{q})$ between points $\mathbf{p}$ and $\mathbf{q}$ can be calculated using Eq. 2.13:

$$d_{min}(\mathbf{p}, \mathbf{q}) = \sqrt[a]{\sum_{i=1}^{n} |p_i - q_i|^a} \tag{2.13}$$

It is important to select an appropriate distance metric when performing hierarchical clustering, as the way in which pairwise distances are calculated can strongly influence the shape and size of the obtained clusters. Two data points may have a small distance between them for one distance metric, but may be farther apart when using a different distance metric. For example, consider the two points $\mathbf{p}$ at $(0, 0)$ and $\mathbf{q}$ at $(1, 1)$. The Euclidean distance between $\mathbf{p}$ and $\mathbf{q}$ is equal to $d_{euc}(\mathbf{p}, \mathbf{q}) = \sqrt{2} = 1.41$, the Manhattan distance is equal to $d_{man}(\mathbf{p}, \mathbf{q}) = 2$, and the Minkowski distance ($a = 5$) is equal to $d_{min}(\mathbf{p}, \mathbf{q}) = \sqrt[5]{2} = 1.15$. As such, significantly different clusters may be obtained for two different distance metrics.

Linkage criteria

Besides defining how to determine the distance between individual points, the linkage method for determining the distance between sets of points must also be defined. The clustering algorithm will then merge pairs of clusters so that the linkage criterion is minimized. While many of such linkage criteria exist, this section limits itself to four well-known linkage criteria: single-, complete-, average-, and Ward linkage.

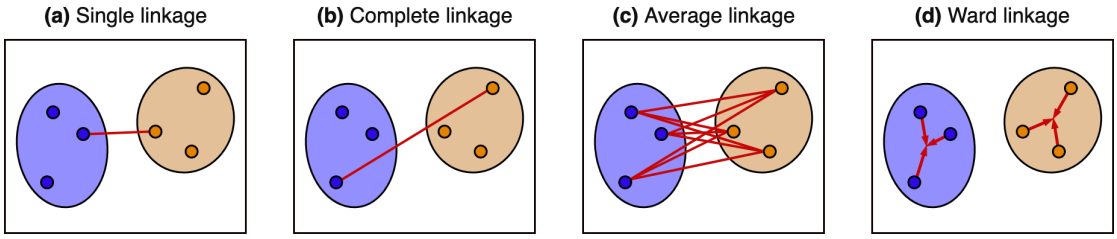


Fig. 2.14: Examples of linkage criteria for agglomerative hierarchical clustering: single, average, complete, and Ward linkage.

Single linkage

In the single linkage method (Fig 2.14a) the distance between a pair of clusters is defined by the distance between the two closest objects of each cluster [41, 58]. Hence, single linkage is also known as the *nearest neighbor* method. Single linkage suffers from so-called *chaining effects*; only one pair of points need to be close in order to merge two clusters, irrespective of all other points. This can result in elongated and spread out clusters which may not be compact enough. Consequently, two clusters with clearly different characteristics could be merged if noise is present in the data set [56]. Another potential problem with single linkage is that it exacerbates the "greedy" or "rich get richer" behavior of agglomerative hierarchical clustering, meaning that larger clusters will tend to merge faster than smaller clusters. However, even though single linkage is not robust to noise, it works well when clusters are far apart. Moreover, single linkage is very efficient, making it the only practical linkage criterion when dealing with very big data sets.

Complete linkage

In contrast to single linkage, complete linkage (Fig 2.14b) defines the inter-cluster distance as the farthest distance between a pair of points [56]. Complete linkage does not suffer from the chaining effects present in the single linkage method, but instead has issues with *crowding effects*; because the inter-cluster distance assumes the worst-case dissimilarity between pairs, a point can be closer to points belonging to other clusters than to points in the same cluster. Even though complete linkage is able to find small and compact clusters, the obtained clusters may not be far enough apart.

Average linkage

The average linkage method (Fig 2.14c), also called *unweighted average linkage*, defines inter-cluster distance as the average of all pairwise distances between the points of two clusters [22]. Average linkage tries to strike a balance between single linkage and complete linkage, so clusters are neither too compact nor too far apart. The average linkage method exists in several similar forms: weighted average linkage, in which the distances between a new cluster and the other clusters are weighted by the number of observations in every cluster; and centroid linkage method, which defines the inter-cluster distances between clusters as the distances between their centroids.

Ward linkage

Ward's linkage method (Fig 2.14d), also known as the *minimum variance* method, aims to minimize the increase of the within-cluster sum of squared errors [74, 56]. In other words, Ward's method minimizes the variance of the clusters being merged so that minimum information loss occurs with each merging. Compared to single-, complete- and average linkage, Ward's method yields the most regular cluster sizes and is less sensitive to noise.It must be noted that Ward linkage should only be applied in conjunction with the Euclidean distance metric, as Ward's method computes centroids in Euclidean space. Since Ward's method is a variance-minimizing approach (i.e. it minimizes the sum of squared differences within all clusters), it is somewhat comparable to the least-squares objective function of the k-means clustering algorithm, but applied with an agglomerative hierarchical approach [76].

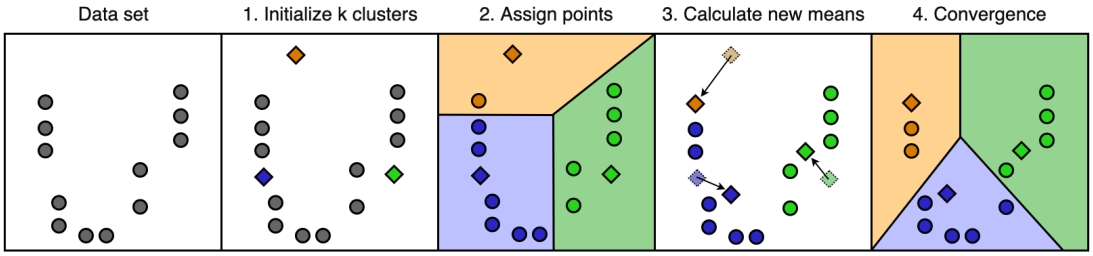


Fig. 2.15: Visual example of the K-means algorithm, with randomly initialized means

2.6.2 K-means clustering

The k-means algorithm was introduced by Macqueen [39] and has grown to be one the most well known and popular clustering algorithms due to its ease of implementation [16]. In contrast to hierarchical clustering – which builds a nested cluster hierarchy – the k-means algorithm is a so-called *partitioning method* which seeks to divide a given data set into a predefined number of k clusters based on an initial partitioning. Even though k-means does not generate a cluster hierarchy, it is similar to hierarchical clustering in that both are *hard* clustering methods, which means that each observation is assigned to only one cluster. Moreover, k-means is also aimed at maximizing the similarity of points within one cluster while minimizing the similarity of points belonging to different clusters. The algorithm uses an iterative refinement method to find the best partition of k clusters by minimizing a sum-of-squared-error criterion. The most basic form of k-means (see Fig. 2.15) is outlined by the following computation procedure:

1. Initialize k random cluster means (or *centroids*);
2. Assign every point to its closest centroid, i.e. the centroid with the least squared Euclidean distance;
3. Recalculate the new centroid of every cluster;
4. Repeat steps 2 and 3 until point assignments stop changing for every cluster (convergence).

Several initialization methods exist, such as the Forgy method and the Random Partition method: the Forgy method randomly selects k points from the data set as starting points, while the Random Partition method first assigns each point to a random cluster and then uses the cluster centroids as starting points.

K-means has several desirable properties that have contributed to its popularity. Firstly, K-means can be easily implemented and functions well for a wide range of problems, especially when the clusters are dense and spherical. Moreover, for an input data set of size n and a number of clusters k, the time complexity of k-means is defined as $\mathcal{O}(nk)$, which is approximately linear as the value of k is often much smaller than n. In other words, the computational complexity of k-means is approximately linearly proportional to the size of the input data set, which makes it relatively efficient at clustering big data sets [76].

However, the k-means algorithm also has significant disadvantages. One big drawback of k-means is caused by the initialization step: since the end result depends on randomly initialized cluster centroids, it is not certain that the global optimum will be found by the k-means algorithm – instead, k-means often terminates at a local optimum. To make matters worse, k-means may not converge at all if any other distance metric than Euclidean distance is used. The random initialization may also result in different clusters for each run. Sadly, no efficient and universal method exists for centroid initialization [22].

Another issue with k-means is that it assumes that the value k is known *a priori*, even though this is not the case in practice. Choosing a proper value for k is important, as poor results could be obtained when choosing an inappropriate choice of k. Similar to the problem of centroid initialization, no efficient and universal method exists for determining an appropriate value of k [22].

Finally, k-means is quite sensitive to noise, as the algorithm also considers outliers in calculating the cluster means. Points that located far from any cluster centroid are forced into one of the clusters, which shifts its cluster mean far away from its "true" mean and distorts the shapes of all clusters.

2.6.3 Comparing clustering algorithms

Agglomerative hierarchical clustering and k-means clustering both have their strengths and weaknesses. The choice for which algorithm and corresponding settings to use depends on the purpose of the clustering and on the type and size of the data set. This section serves to compare the characteristics, advantages and disadvantages of both clustering methods.

Efficiency

The time complexity of k-means increases linearly with the amount of observations n, i.e. $\mathcal{O}(n)$, while the time complexity of most agglomerative hierarchical clustering algorithms is qubic, i.e. $\mathcal{O}(n^3)$. This makes agglomerative hierarchical clustering impractical when dealing with big data sets, whereas k-means is more scalable. However, as discussed in section 2.5, the application of PCA may significantly lower the size of the data set. Hence, even though k-means will still be more efficient than hierarchical clustering when PCA is applied, hierarchical clustering may still be a practical choice. Put simply, if hierarchical clustering takes one second to complete and k-means clustering takes 0.01 second to complete after applying PCA, the fact that k-means is 100 times faster than hierarchical clustering is practically irrelevant.

Consistency

Hierarchical clustering is deterministic, meaning that the same results are obtained for different runs if the settings are the same. On the other hand, the initialization step of the k-means algorithm introduces randomness, which may lead to k-means generating different results for different runs even if the settings are identical. As such, results obtained by k-means may not always be reproducible. Moreover, hierarchical clustering produces a hierarchy of nested clusters that is informative and internally consistent, while k-means simply partitions the input data without providing any additional information of its internal structure.

Similarity

The basic k-means algorithm specifically requires the use of Euclidean distance and may not be able to converge if any other distance metric is used. On the other hand, many distance metrics and linkage criteria can be used to perform hierarchical clustering, with the choice of similarity metrics often depending on the application and goals of the clustering. While the ability to choose from different distance metrics and linkage criteria in hierarchical clustering can be interpreted as additional complexity, it also makes hierarchical clustering more flexible and provides many more possibilities than k-means.

Number of clusters

K-means requires prior knowledge about the number of clusters k, which is often unavailable in practice. On the other hand, hierarchical clustering does not require the user to know k, as k can be chosen afterwards based on knowledge obtained from the resulting hierarchy and dendrogram. However, when a certain set of clusters and their characteristics are further investigated – as is done in this research – an appropriate value for k must be chosen at some point. Therefore, several methods that help determine an appropriate value of k are discussed in section 3.5.2.

2.6.4 Choice of algorithm

Even though hierarchical clustering is less computationally efficient than k-means clustering, iterative testing showed that a combination of hierarchical clustering and PCA for dimensionality reduction resulted in acceptable computation times that are comparable to k-means. Since this solves the main drawback of hierarchical clustering, the aforementioned combination was chosen as the preferred clustering approach.

3 Data and methods

This chapter describes the data products used in this research, as well as all methods regarding data preprocessing and clustering. Firstly, a description of the study area is given in section 3.1. The ASCAT data and land cover data are discussed in section 3.2 and section 3.3, respectively. The implementation of the principal component analysis is treated in section 3.4, and the methods related to clustering are explained in section 3.5. Finally, section 3.6 introduces the concept of *cluster robustness*, as well as a method for determining cluster robustness.

3.1 Study area

The study region consists of the entirety of France, extending from 42°N to 51°N and 4°W to 8°E and covering over 640.000 km^2. Most of France has a temperate oceanic climate [36]. However, four distinct climatic zones exist within France: oceanic climate (Köppen-Geiger climate type: Cfb), continental climate (Dfb), Mediterranean climate (Csa, Csb), and mountain climate (Dfc, ET) [49].

The oceanic climate can mainly be found in the western parts of France (e.g. Brittany, Normandy) and is characterized by cool summers and mild winters, resulting in a relatively narrow annual temperature range compared to other areas at comparable latitudes. Moreover, temperate oceanic climates generally lack a dry season, with average precipitation spread relatively evenly throughout the year. The continental climate found in the eastern and central areas of France (e.g. Champagne, Burgundy and Alsace) is characterized by higher precipitation, warmer summers and colder winters compared to areas with an oceanic climate. The Mediterranean climate is found in southern France (e.g. Provence, Côte d'Azur) and is characterized by little precipitation, mild winters, hot summers, and a distinct dry season. Finally, the mountain climate is located in areas with relatively high altitude such as the Alps, Pyrenées, and Central Massif, and are characterized by cold winters, cool summers and abundant precipitation.

Due to its overall temperate climate and the aforementioned climate zones, vegetation cover in France is relatively heterogeneous. The main vegetation cover types are rainfed cropland (both summer and winter crops), grasslands, deciduous and coniferous forests. Over 50% of the total land area of France consists of arable and pastoral land [44]. Arable land is mainly used for the production of cereal crops, of which wheat and corn are the most dominant crops, while barley and oats are significantly less popular. Even though the majority of cereal production occurs in south-west France and in the Paris Basin, there are few areas in France where cereal crops are entirely absent [51]. Deciduous and coniferous forests cover an area of approximately 180.000 km^2, or around 28% of the total surface area of France. Finally, grasslands make up approximately one-third of the agricultural area [26]. The heterogeneous land cover of France provides the opportunity to study the behavior of σ°_{40}, σ', and σ'' over a diverse set of land cover classes and vegetation types.

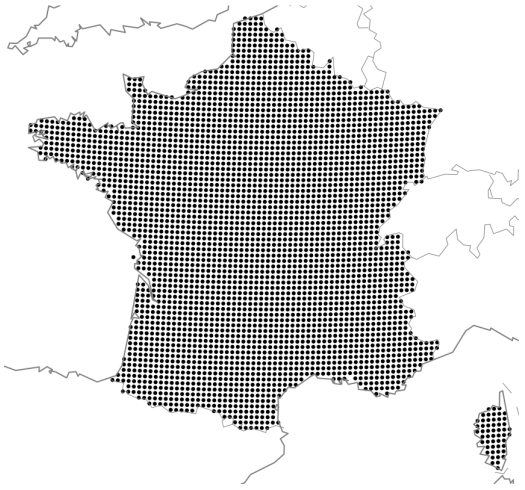


Fig. 3.1: Map depicting all 3492 grid points in the study region of France. Each of the 3492 grid points represents an area of 25 x 25 km^2.

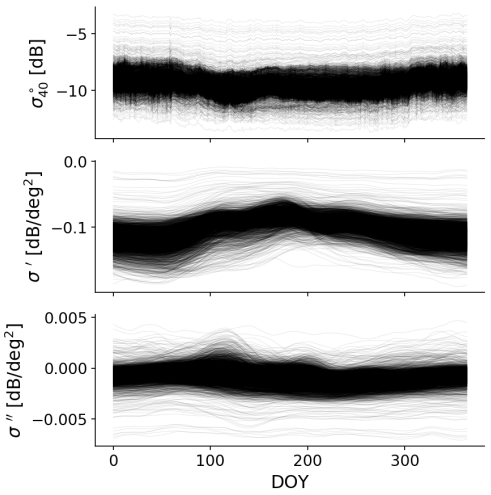


Fig. 3.2: Seasonal climatology of σ°_{40}, σ' and σ'' for each of the grid points shown in Fig. 3.1, derived from the 10-year time series shown in Fig. 3.3.

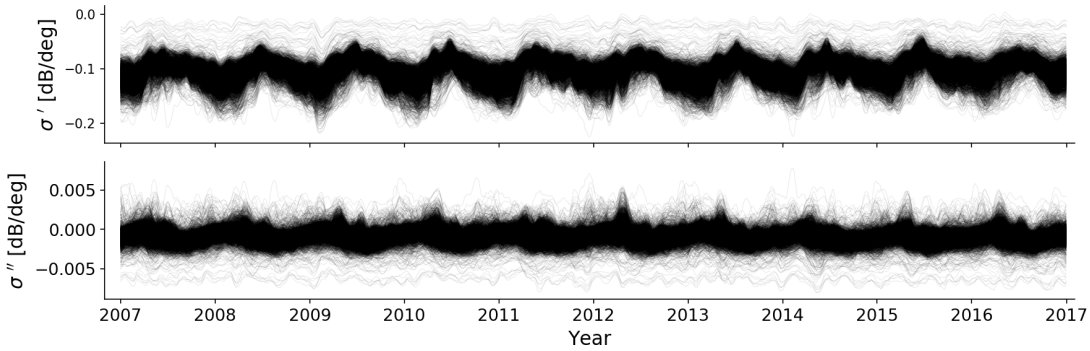


Fig. 3.3: The 10-year time series of dynamically estimated slope (σ') and curvature (σ'') for each grid point shown in Fig. 3.1.

3.2 ASCAT data

Ten years of MetOp-A ASCAT SZR Level 1b Fundamental Climate Data Record backscatter data were obtained, to which three preprocessing steps were applied: the backscatter observations were resampled to a fixed Earth grid [45]; an intra- and interbeam calibration was performed [53]; and azimuthal effects were accounted for [4].

For every grid point in the study region, the 10-year time series of the backscatter coefficient (σ°_{40}, normalized to the reference incidence angle $\theta_r = 40°$), slope (σ'), and curvature (σ'') were extracted. The slope and curvature were dynamically estimated with the new method proposed by Melzer [42] and a kernel width of $\lambda = 21$ days. Moreover, the seasonal climatology was determined for σ°_{40}, σ', and σ''. For σ' and σ'', the seasonal climatology was determined by averaging the daily values across the obtained 10 years of dynamically estimated daily σ' and σ'' data. However, in order to determine the seasonal climatology of σ°_{40}, the data was first aggregated into 10 day intervals before averaging across the 10 years. This was necessary because only a limited number of values are available for any given day of the year due to the revisit time of MetOp-A, which takes two days to entirely cover the Earth's surface.

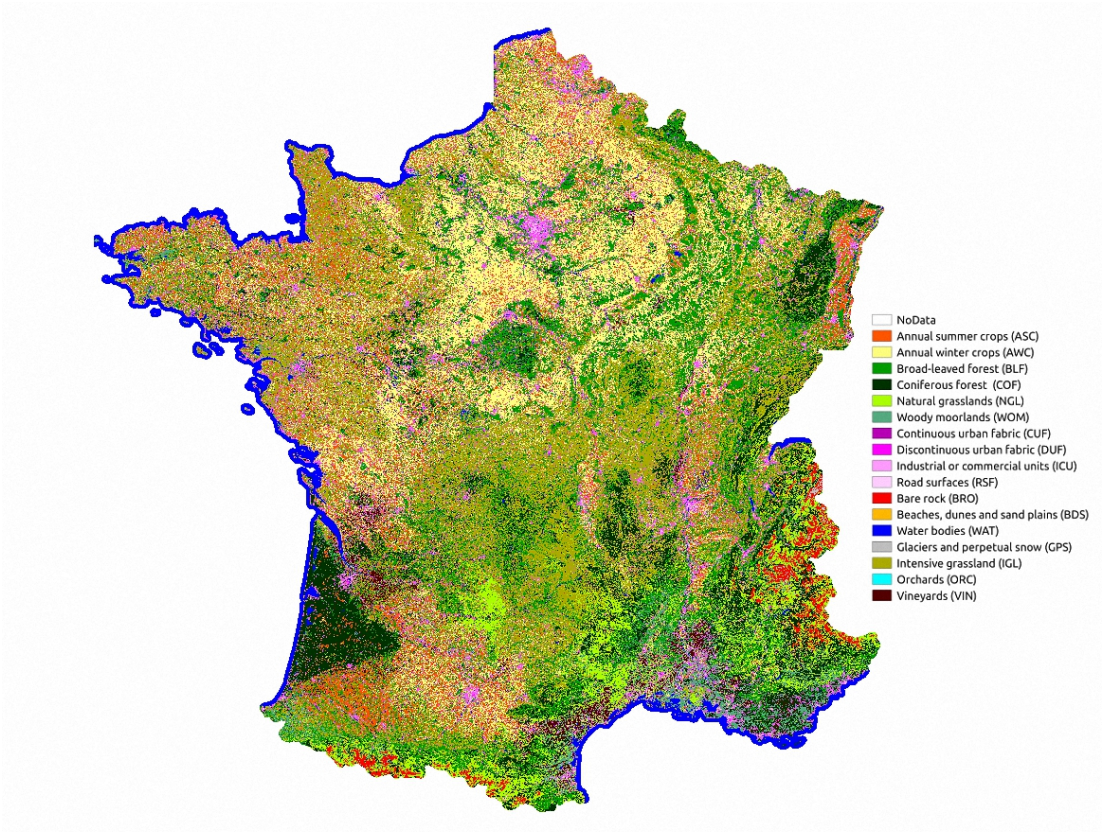


Fig. 3.4: Land cover data set for France for the year 2016, generated using the approach proposed by Inglada et al. [29]. This land cover data set has a resolution of 10 m and is available at http://osr-cesbio.ups-tlse.fr/

3.3 Land cover data

In order to improve our understanding of the relationship between land cover and the vegetation parameters, an accurate land cover data set of France is needed. The land cover data set used in this research was obtained from the Theia Data and Services Center (https://www.theia-land.fr/en), which is a France-based initiative aimed at developing and distributing high-quality data products related to the observation of continental surfaces, primarily based on satellite data. This section describes the Theia land cover data set, and how this data set was rescaled to the ASCAT grid.

3.3.1 Theia land cover data set

This research makes use of the 2016 Theia land cover data set of metropolitan France (Fig. 3.4). This data set was generated using a classical supervised classification procedure in which existing data bases were used as reference data for training and validation [29]. The procedure described is able to handle large volumes of data and can be applied over large territories to produce land cover maps automatically in very short production times; a detailed description of the procedure is given by Inglada et al. [29].

The 2016 Theia land cover data set is mainly based on Sentinel-2 data acquired between the end of 2015 to the end of 2016. Landsat-8 data was not directly used for land cover classification, but the use of Landsat-8 data from 2014 until the end of 2016 was necessary to transfer classifier training to 2016 Sentinel-2 data. The data set has a resolution of 10 m, with each pixel assigned one of 17 possible land cover classes. In order to relate the σ°_{40}, σ', and σ'' of a grid point to its land cover footprint, the 10 m resolution Theia data set was first rescaled to the 25 km resolution ASCAT grid.

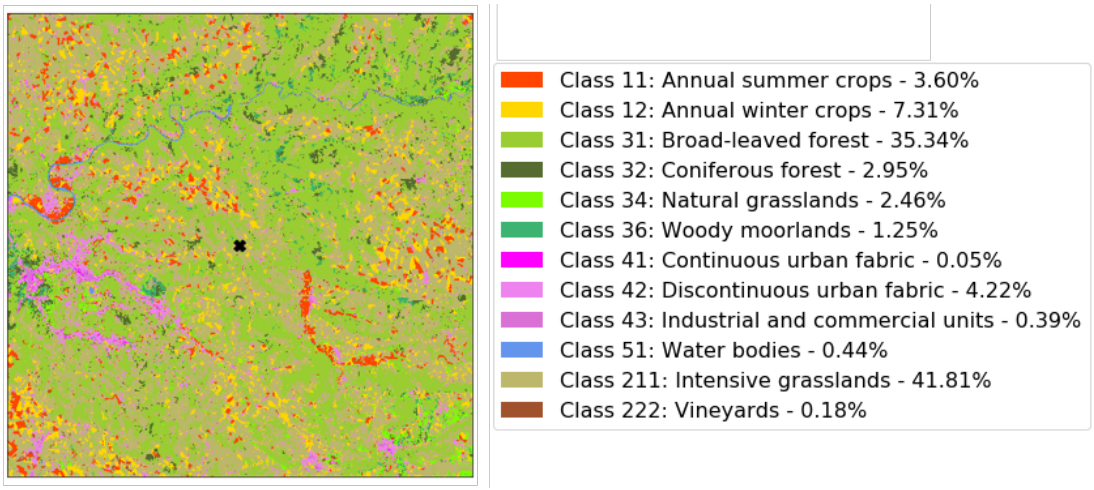


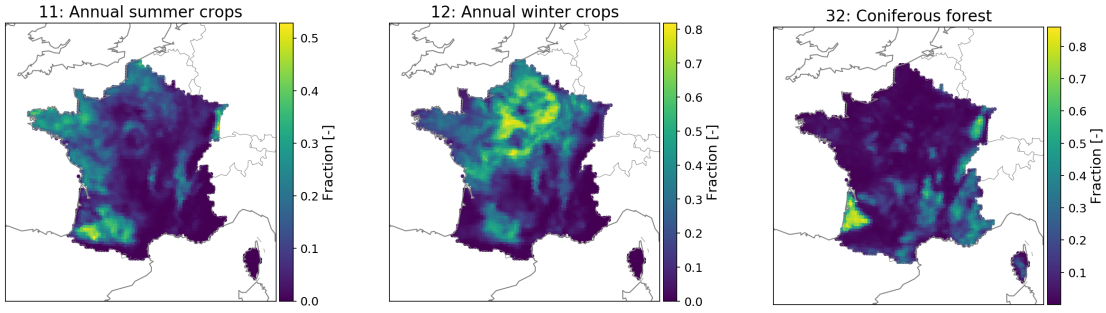
Fig. 3.5: Rescaling the Theia land cover data set to the ASCAT grid. The exact location of an example ASCAT grid point (GPI = 2283605) with coordinates (44.573316°, 2.360628°) is denoted by a black cross.

3.3.2 Rescaling to ASCAT grid

This section will describe the processing steps that were performed to rescale the Theia land cover data set to the ASCAT grid. The resulting land cover data set contains the land cover composition of each ASCAT grid point, i.e. the average fractions of each of the 17 land cover classes for each grid point.

Since each ASCAT grid point covers an area of 25 x 25 km^2 and each Theia grid point covers an area of 10 x 10 m, each ASCAT grid point will consist of 2500 x 2500 Theia grid points. Hence, for each ASCAT grid point an array of 2500 x 2500 land cover grid points must be extracted from the Theia land cover data set. It is assumed that each ASCAT grid point is located at the center of its corresponding 2500 x 2500 land cover array. For example, Fig. 3.5 shows an ASCAT grid point (GPI = 2283605, marked with a black cross) with coordinates (44.573316°, 2.360628°) sitting at the center of a 2500 x 2500 array of grid points, extracted from the Theia land cover data set.

After extracting the appropriate Theia grid points, the number of occurrences of every land cover class are counted. Finally, the fraction of every land cover class is calculated relative to the total number of grid points (i.e. $2500 \times 2500 = 6.25 \cdot 10^6$ grid points). For example, if a certain ASCAT grid point contains $2.25 \cdot 10^5$ grid points labeled as class 11 (annual summer crops), then that grid point has a class 11 fraction of $2.25 \cdot 10^5 / 6.25 \cdot 10^6 = 0.036$, or 3.6% (see Fig. 3.5). This procedure is repeated for all land cover classes and all ASCAT grid points, yielding a data set describing the land cover composition of each grid points in terms of 17 land cover classes. Using the resulting data set, the land cover classes can be mapped individually; three land cover classes are mapped in Fig. 3.6, and all land cover maps can be found in Appendix A.3. Finally, it must be noted that recent land cover data sets generated by Theia distinguish between even more land cover classes, which may be interesting for future research.



(a) Class 11: Annual summer crops **(b)** Class 12: Annual winter crops **(c)** Class 32: Coniferous forest

Fig. 3.6: Rescaled Theia land cover data set. Each land cover class is mapped separately and colored by their fractions.

3.4 Principal component analysis

3.4.1 Feature scaling

Since PCA is a variance maximizing exercise, the input data must be standardized before carrying out the PCA to ensure that the proper PCs are found. The `StandardScaler` module of `scikit-learn` [47] was used to perform standardization, taking as input the (3492×365) σ' climatology and yielding the (3492×365) standardized σ' climatology.

3.4.2 Explained variance score

The `explained_variance_score` module of `scikit-learn` is used to calculate the ratio of variance that is explained by the PCA approximation relative to the total variance present in the original data, which is denoted by η^2. The best possible score of $\eta^2 = 1$ indicates that all variance in the original data is accounted for. It is possible for η^2 to be negative – this indicates that the mean of the 'correct' values is a better predictor than the estimated target values. Given an estimated target output $\hat{y}$ and a corresponding correct target output y, the explained variance score η^2 can be calculated using Eq. 3.1:

$$\eta^2(y, \hat{y}) = 1 - \frac{\mathbf{Var}\{y - \hat{y}\}}{\mathbf{Var}\{y\}} \tag{3.1}$$

where $\mathbf{Var}\{y - \hat{y}\}$ and $\mathbf{Var}\{y\}$ are given by Eq. 3.2 and Eq. 3.3, respectively.

$$\mathbf{Var}\{y - \hat{y}\} = \frac{\sum[y_i - \hat{y}_i - \mathbf{E}(y - \hat{y})]^2}{n - 1} \tag{3.2}$$

$$\mathbf{Var}\{y\} = \frac{\sum(y_i - \bar{y})^2}{n - 1} \tag{3.3}$$

3.4.3 Choosing m, the number of principal components

As described in section 2.5.3, the number of PCs to retain can be determined by defining the minimum percentage of total variance present in the input data set that should be accounted for by the retained PCs. In this study, a threshold of $\eta^2 \geq 0.99$ is chosen so that at least 99% of the variance present in the standardized σ' climatology is explained by the first m PCs.

The PCA was performed using the `PCA` module of `scikit-learn`, taking as input the (3492×365) standardized σ' climatology and yielding a (3492×365) array of PCA projections, i.e. 365 weights for each grid point, with each weight corresponding to one of the 365 PCs. The value of m is then equal to the number of PCs for which $\eta^2 \geq 0.99$ is satisfied. Finally, the (3492×365) array of PCA projections is truncated to a size of $(3492 \times m)$, which is the reduced data set on which the hierarchical clustering is performed. The full PCA implementation is given in Listing 3.1.

Listing 3.1: Implementation of the described PCA.

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

data_zscore = StandardScaler().fit_transform(slo.T) # standardize the slope climatology
min_exp_var = 99 # desired minimum % of explained variance
pca_fit = PCA().fit(data_zscore) # fit PCA model
exp_var_ratio = np.cumsum(pca_fit.explained_variance_ratio_) * 100 # cumulative exp_var ratio [%]
n_PCs = np.where(exp_var_ratio >= min_exp_var)[0][0] # find n_PCs so that exp_var_ratio >= 99%
data_PCA = PCA(n_components=n_PCs).fit_transform(data_zscore) # perform PCA, retain n PCs
```

3.5 Clustering

3.5.1 Algorithm and settings

Taking into account the advantages and disadvantages of k-means clustering and (agglomerative) hierarchical clustering described in section 2.6.3, agglomerative hierarchical clustering with Ward's linkage criterion was chosen as clustering method. As discussed in section 2.6.1.1, there exist many methods to compute the distance between elements of a set. However, Euclidean distance is the only option when applying Ward's method as linkage criterion. Agglomerative hierarchical clustering is available in the `AgglomerativeClustering` module of `scikit-learn`, which also contains support for Ward's method. At the time of writing, visualisation of dendrograms is not implemented in `scikit-learn`; instead, the `linkage` and `dendrogram` modules of `scipy` [68] were used to calculate the linkage and distance matrices and plotting the dendrogram. Clustering is performed on the reduced $(3492 \times m)$ data set obtained from the PCA, see section 3.4.3.

3.5.2 Choosing k, the number of clusters

In order to form a set of k clusters, the dendrogram must be cut at a certain level after hierarchical clustering has been performed. It is important to choose a correct value of k, as either over-estimating or under-estimating k will affect the quality of the obtained clusters. A too high value of k results in too many clusters, which makes it difficult to interpret the results. Conversely, a too low value of k causes significant information loss which can lead to misleading interpretations of the results.

The best value for k can be determined by performing a number of clustering iterations using different values of k and evaluating the performed clustering during each iteration using a performance index; the best value of k is then the value that corresponds with the highest performance index. In general, performance indices assess the quality of the obtained clusters based on within-cluster similarity, between-cluster dissimilarity, or a combination of the two. In their study, Milligan and Cooper [43] compared and ranked 30 performance indices in terms of their performance on several artificial data sets. While the Calinski-Harabasz index (CH) [11] was found to be the best performing index, the Davies-Bouldin index (DB) [12] and the Silhouette index (SIL) [54] are also applied in this study. Additionally, the more recently proposed 'L-method' by Salvador and Chan [55] for determining k was implemented in this study.

The CH, DB, and SIL indices are so-called *internal validation measures*, which are measures that validate clustering based on the compactness of clusters and the separation between clusters. The compactness of a cluster describes how closely related its objects are based either on within-cluster variance or within-cluster distance, while the separation between clusters describes how distinct a given cluster is compared to all other clusters, generally based on between-cluster distances.

3.5.2.1 Calinski-Harabasz index

The CH index introduced by Calinski and Harabasz [11] is implemented in the `calinski_harabasz_score` module of `scikit-learn`. Higher CH scores indicate better clustering performance and hence, the best value of k is found by maximizing CH. To calculate CH for a data set of N elements partitioned into K clusters $C_1 \dots C_K$, first the within- and between-cluster dispersion must be defined. For each cluster k, the within-cluster dispersion W_k is the sum of squared distances between the observations M_k^i and the cluster centroid G_k:

$$W_k = \sum_{i \in C_k} \|M_k^i - G_k\|^2 \tag{3.4}$$

The total within-cluster dispersion W is then the sum of all within-cluster dispersions:

$$W = \sum_{k=1}^{K} W_k \tag{3.5}$$

The between-cluster dispersion B is the weighted sum of squared distances between G_k and the centroid of the entire data set G, weighted by the number of elements n_k in each cluster:

$$B = \sum_{k=1}^{K} n_k \|G_k - G\|^2 \tag{3.6}$$

Ultimately, the CH index is calculated using Eq. 3.7:

$$CH = \frac{B/(K-1)}{W/(N-K)} = \frac{N-K}{K-1}\frac{B}{W} \tag{3.7}$$

3.5.2.2 Davies-Bouldin index

The DB index introduced by Davies and Bouldin [12] attempts to maximize the within-cluster similarity while minimizing the between-cluster similarity. Lower DB values indicate better clustering performance. Hence, the best value of k is found by minimizing DB. The DB index is implemented in the `davies_bouldin_score` module of `sklearn` and can be calculated as follows. First, the mean distance δ_k of the points of cluster C_k to the cluster centroid G_k (i.e. the mean within-cluster similarity) is defined as:

$$\delta_k = \frac{1}{n_k} \sum_{i \in C_k} \|M_k^i - G_k\|^2 \tag{3.8}$$

Furthermore, the distance $\Delta_{kk'}$ between the centroids G_k and $G_{k'}$ of clusters C_k and $C_{k'}$ (i.e. the between-cluster similarity of clusters k and k') is defined as:

$$\Delta_{kk'} = \|G_{k'} - G_k\|^2 \tag{3.9}$$

For each cluster k the maximum M_k of the quotients $\Delta_{kk'}^{-1}(\delta_k + \delta_{k'})$ is calculated for all indices $k' \neq k$. The DB index is then the mean of all M_k values:

$$DB = \frac{1}{K}\sum_{k=1}^{K} M_k = \frac{1}{K}\sum_{k=1}^{K} \max_{k' \neq k}\left(\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}}\right) \tag{3.10}$$

3.5.2.3 Silhouette index

The SIL index introduced by Rousseeuw [54] is implemented in the `silhouette_score` module of `sklearn`. The optimal value of k is found by maximizing SIL. First, the mean within-cluster distance a_i is defined as the mean distance of point M_i to all other points in the same cluster C_k:

$$a_i = \frac{1}{n_k - 1} \sum_{i' \in C_k} d(M_i, M_{i'}) \tag{3.11}$$

Furthermore, the mean distance $\delta(M_i, C_{k'})$ of M_i to the points of other clusters $C_{k'}$ is defined as:

$$\delta(M_i, C_{k'}) = \frac{1}{n_{k'}} \sum_{i' \in C_{k'}} d(M_i, M_{i'}) \tag{3.12}$$

The smallest mean distance b_i of $\delta(M_i, C_{k'})$ is then defined as:

$$b_i = \min_{k' \neq k} \delta(M_i, C_{k'}) \tag{3.13}$$

The silhouette width s_i can then be calculated for a point M_i using:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{3.14}$$

Finally, the SIL index is calculated using Eq. 3.15:

$$SIL = \frac{1}{K} \sum_{k=1}^{K} \mathfrak{s}_k \quad where \quad \mathfrak{s}_k = \frac{1}{n_k} \sum_{i \in C_k} s_i \tag{3.15}$$

3.5.2.4 L-method

After performing a hierarchical clustering, the resulting dendrogram contains the hierarchy of merges as well as the similarity at every merge (see section 2.6.1). From this dendrogram a 'number of clusters vs. similarity' evaluation graph can be constructed (Fig. 3.7). The L-method proposed by Salvador and Chan [55] is used to find the knee of a 'k vs. similarity' graph resulting from hierarchical clustering, where the location of the knee indicates the best value of k. No Python implementation of the L-method exists; instead, the L-method was implemented based on the procedure described by Salvador and Chan [55].

Consider the evaluation graph shown in Fig. 3.7. The x-axis ranges between $2 \ldots b$, so the graph consists of $b-1$ elements. The graph is partitioned at $x = c$ into a left sequence L_c $(2 \leq x \leq c)$ and a right sequence R_c $(c+1 \leq x \leq b)$. If $RMSE(L_c)$ and $RMSE(R_c)$ are the root mean squared error of the best-fit lines for L_c and R_c, then the total root mean squared error $RMSE_c$ is defined by Eq. 3.16, where $RMSE(L_c)$ and $RMSE(R_c)$ are weighted proportional to their size. After partitioning the data at all $3 \leq c \leq b-2$ and calculating $RMSE_c$ for each c, the best value of k is found where $RMSE_c$ is minimized.

$$RMSE_c = \frac{c-1}{b-1} RMSE(L_c) + \frac{b-c}{b-1} RMSE(R_c) \tag{3.16}$$

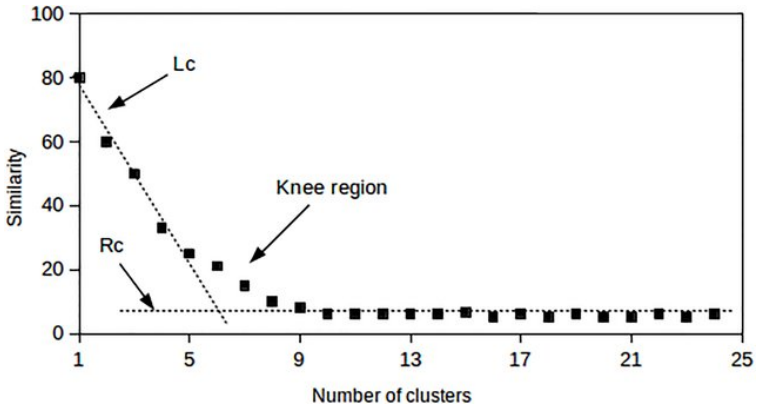


Fig. 3.7: Example of a 'k vs. similarity' to which the L-method was applied. The data is partitioned at $x = c$, yielding L_c, R_c and the knee region, which gives an indication of the best value of k [55].

3.6 Robustness

Currently, the seasonal climatology of σ' and σ'' coefficients are obtained using multiple years of local σ' observations. As a result, a clustering performed on the seasonal climatology of σ' does not provide insight into if and how the clusters change over time. However, if hierarchical clustering is performed on each of the 10 available years of σ' data it becomes possible to investigate the clusters on an annual basis.

For example, consider a certain grid point p for which 10 cluster labels have been determined. If p is assigned the same cluster label for every year of σ' data – e.g. {3, 3, 3, 3, 3, 3, 3, 3, 3, 3} – then the behavior of σ' in point p is relatively similar for all years and point p clearly belongs to cluster 3. In other words, the clustering of point p is *robust*. Conversely, consider another grid point q for which the assigned cluster labels are not the same for every year. For example, if q is assigned the labels {1, 1, 0, 0, 0, 2, 2, 2, 2, 2} then point q does not belong to only one cluster as a result of the larger interannual variability of σ' in point q. In this case, the clustering of point q would be *less robust* than point p.

3.6.1 Annual clustering

In order to perform hierarchical clustering on each of the 10 years of σ' observations separately, some preprocessing steps are required. As shown in Fig. 3.3, the 10-year σ' data set consists of 3492 grid points, with 3653 daily values (7 years of 365 days and 3 leap-years of 366 days) for each grid point. Firstly, the 10-year data set is split by year into 10 separate data sets; each set consists of 3492 grid points, with 365 daily values per grid point (Fig. 3.8). Finally, the 10 sets are combined again into one (34920 × 365) data set, on which a PCA is performed as described in section 3.4 in order to reduce the dimensionality and improve computation time of clustering. Finally, hierarchical clustering is performed on the reduced data set obtained from the PCA. It should be noted that leap-years are disregarded in this analysis, as many clustering algorithms including hierarchical clustering require all objects that are to be clustered to have an equal number of features. Hence, all leap-years are truncated to a length of 365 days.

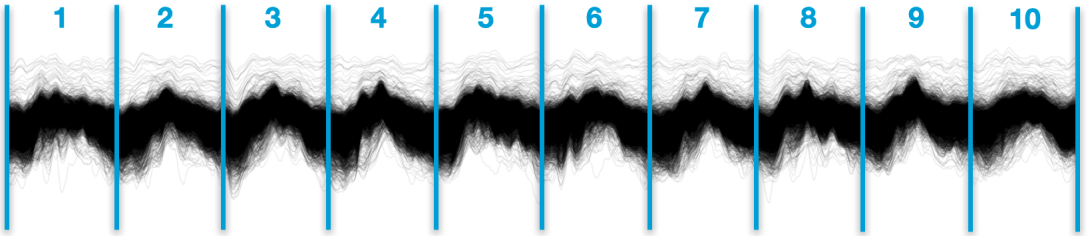


Fig. 3.8: 10 years of σ' data, split into 10 separate (3492 × 365) data sets.

3.6.2 Robustness score

As previously discussed, the hierarchical clustering yields one cluster label for every year for every grid point, i.e. a set of 10 labels are obtained for every grid point. In order to quantify the clustering robustness for every grid point, two factors are taken into account: (1) the number of unique cluster labels assigned a grid point $|N|$; and (2) the number of times that the cluster label changed $\hat{N}$. The *robustness score* R is then defined by $R = |N|\hat{N}$, with lower values indicating better robustness.

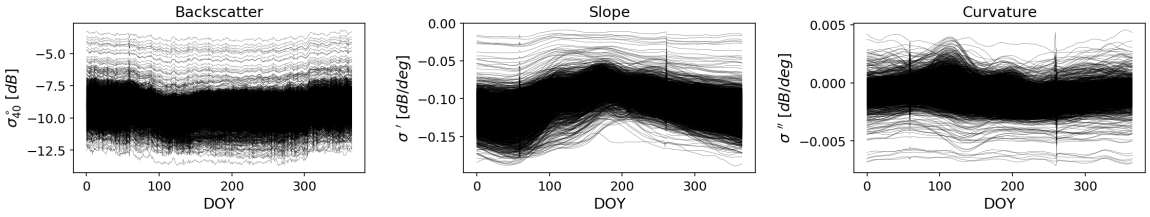
Using the previous example, point p was assigned the cluster labels $N_p = \{3,3,3,3,3,3,3,3,3,3\}$, which gives $|N_p| = 1$ and $\hat{N}_p = 0$ and results in a good robustness score of $R_p = 1 \cdot 0 = 0$. On the other hand, point q was assigned the cluster labels $N_q = \{1,1,0,0,0,2,2,2,2,2\}$, which gives $|N_q| = 3$ and $\hat{N}_q = 2$ and results in a worse robustness score of $R_p = 3 \cdot 2 = 6$.

4 Results and discussion

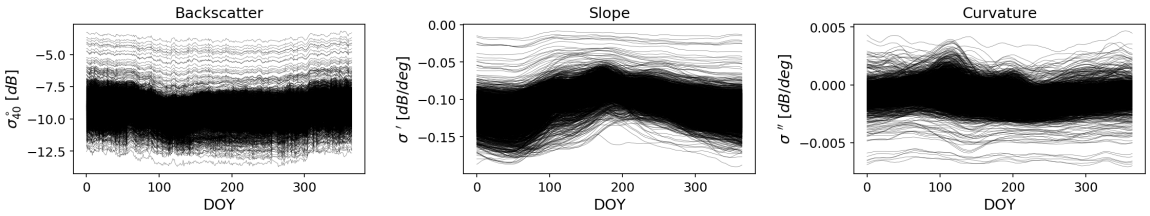
In this chapter the results of the performed analyses are discussed. First, the original data is screened for perturbations. Second, the required number principal components is determined, a PCA is carried out, and the principal components are investigated. Third, the optimal number of clusters is determined and the defining characteristics of the resulting clusters are investigated. Fourth, the potential relationship between σ°_{40}, σ', and σ'' and sub-footprint land cover heterogeneity is explored. Fifth and finally, robustness scores of grid points and clusters are determined and analysed in relation to their land cover footprint.

4.1 Data screening

Perturbations can be observed in the original data set, especially in the seasonal climatology of σ' and σ'' around day 60 and 260, see Fig. 4.1a. These perturbations may be caused by missing/spurious σ°_{40} data during one or several periods, either due to calibration activities or instrument error(s). Perturbations in the input data should be corrected to ensure that the right principal components are found, and hence, to ensure proper clustering. To remove these perturbations, the average σ°_{40}, $\sigma'$and σ'' signatures were recalculated for each grid point using the 10 years of available σ°_{40}, σ', and σ'' data, see Fig. 4.1b.



(a) Perturbations in the original data set around day 60 and day 260



(b) Data after removing perturbations

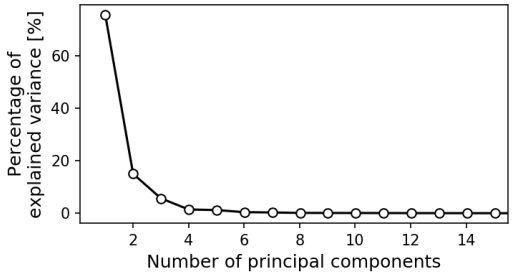
Fig. 4.1: Input data before and after correction of perturbations around day 60 and 260

4.2 Principal component analysis

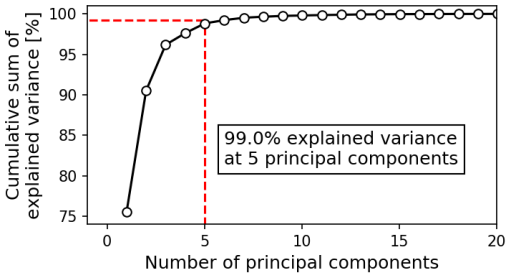
4.2.1 Determining the number of principal components

As explained in section 3.4, the number of PCs is chosen so that the percentage of explained variance is at least 99%. In Fig. 4.2a, it can be seen that the first PC explains the largest amount of the variance present in the data set (approximately 75%), and the percentage of explained variance decreases for subsequent PCs. The cumulative explained variance is plotted against the number of PCs in Fig. 4.2b, which shows that 99% of the total variance present in the input data set is explained when at least five PCs are retained.

Moreover, it should be noted that approximately 91% of the total variance is explained by the first two PCs and approximately 96% is explained by the first three PCs, suggesting there is a possibility that three PCs are sufficient for some grid points. However, this is likely not true for all grid points; it may be that other grid points require significantly more than five PCs before their data is well represented. Hence, it should be investigated whether differences exist between areas in terms of how many PCs are required.



(a) Percentage of total variance explained by each of the PCs.

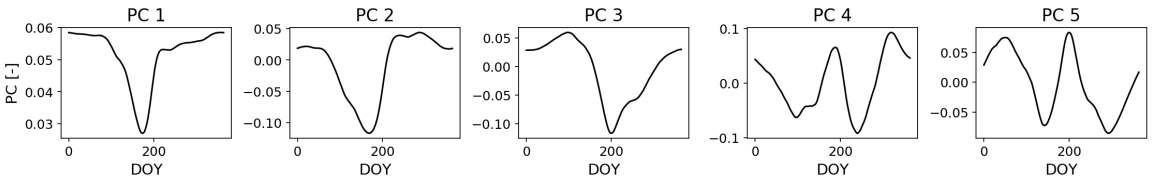


(b) Cumulative percentage of explained variance

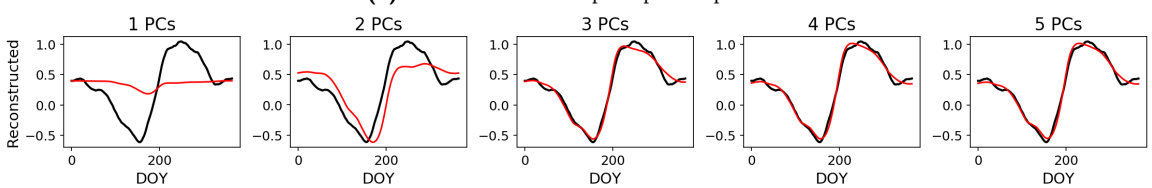
Fig. 4.2: Variance explained by the principal components

4.2.2 Investigating the principal components

In order to explain at least 99% of the variation present in the original data, at least five PCs should be retained; these first five PCs are plotted in Fig. 4.3a. The standardized σ' observations can be reconstructed by combining a set of weights with each of their corresponding PCs (as described in section 2.5). This process is visualized for a single grid point in Fig. 4.3b, which shows that the reconstructed signal of this particular grid point does not significantly improve when more than three PCs are retained. This also suggests that the number of PCs required to accurately estimate the original data differs between grid points, since only three PCs may have been required even though five PCs were retained. Several more examples are provided in appendix B Fig. B.1, which visualizes the differences between a number of grid points in terms of how many PCs are required before their $\eta^2 \geq 99\%$. The number of PCs that are required to describe the original signal depends on the shape of the standardized σ' signal. For example, grid points with a standardized σ' signal that is very similar to the shape of the first PC are likely to be described mostly by the first PC, i.e. the weight corresponding with the first PC is largest. On the other hand, for grid points where the standardized σ' signal is very complex or noisy and is not similar to any specific PC, significantly more PCs are required to properly estimate the original signal. By applying PCA and using the obtained weights as input for clustering, the size of the original data set is reduced from (3492 x 365) to (3492 x 5) – a decrease of 98.6% – while the essence of the original data set is retained.

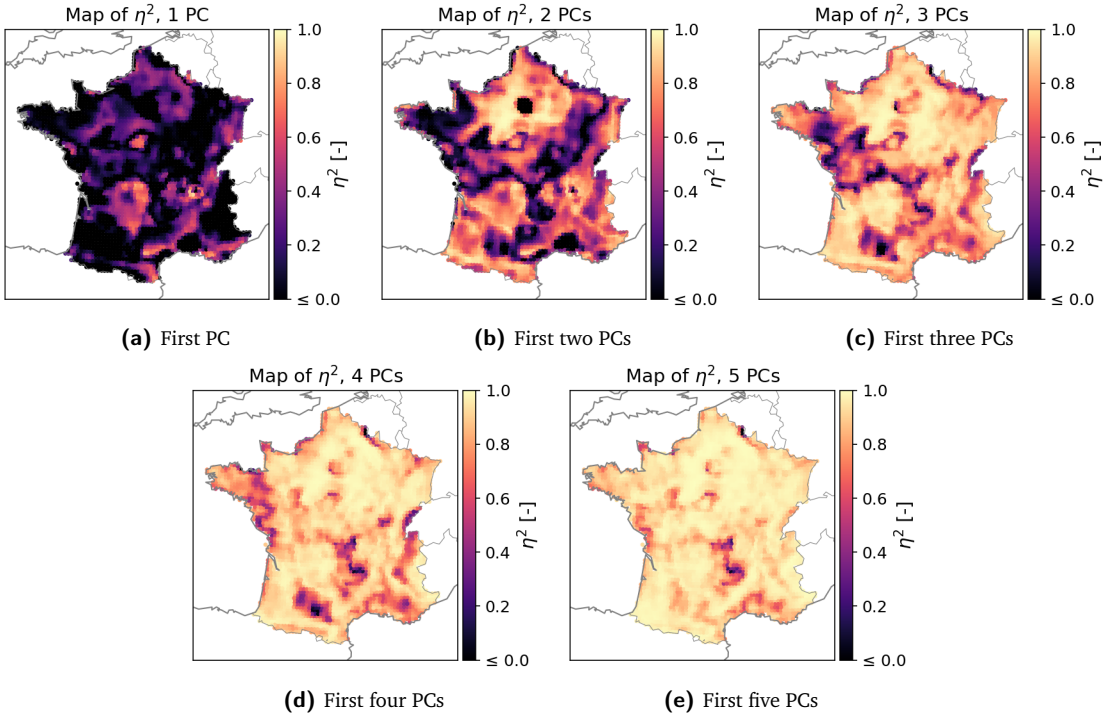


(a) Plots of the retained principal components



(b) Reconstructing the original data of a grid point using the first five PCs.

Fig. 4.3: Principal Components



(a) First PC **(b)** First two PCs **(c)** First three PCs

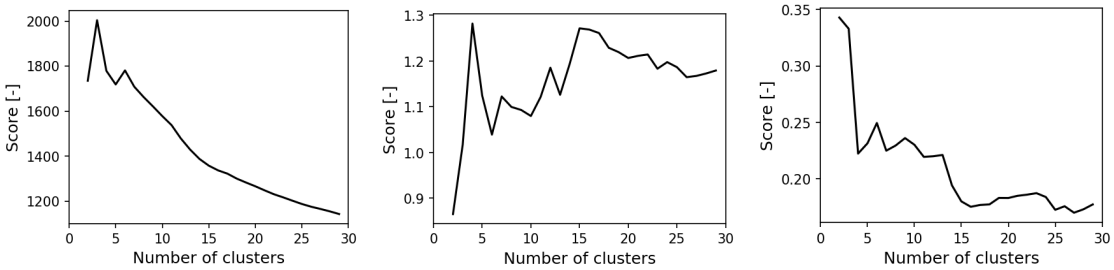
(d) First four PCs **(e)** First five PCs

Fig. 4.4: η^2 mapped for cumulative combinations of PCs. η^2 is set to zero for grid points where $\eta^2 < 0$.

4.2.3 Spatial characteristics of the principal components

In order to investigate whether there is some sort of spatial rationale underlying the PCs, the explained variance score η^2 (see section 3.4.2) is calculated for all grid points and all cumulative combinations of PCs using Eq. 3.1; these values are subsequently mapped in Fig. 4.4. Fig. 4.4a shows that especially Paris (and other smaller, highly urbanized areas) as well as coastal grid points and (to a lesser degree) the Alps and Pyrenees, perform significantly worse in terms of η^2 than the rest of France when only the first PC is used. For two PCs (4.4b), η^2 significantly improves for the area surrounding Paris, the south-west of France and the Alps. For Paris and other urban areas η^2 improved but still performs poorly compared to the rest of France. When using three PCs these area significantly improve, see Fig. 4.4c. The central-north-west area of France has relatively poor η^2 for three PCs compared to the rest of France, but η^2 in this area improves when using four PCs. The maps showing the effect of including the fourth (Fig. 4.4d) and fifth (Fig. 4.4e) are quite similar, with η^2 improving similarly in nearly all areas. Additionally, the performance of cumulative combinations of the first twelve PCs is provided in appendix B.2, which shows that improvements of η^2 are marginal in nearly all grid points when retaining seven or more PCs. However, differences between areas are visible even when twelve PCs are retained, suggesting that some areas are simply more difficult to represent than others using PCA.

Fig. 4.4 indicates that spatial information is present in the retained PCs. This can be explained by the fact that the most important PCs are found in the directions of the largest variations. Since the PCA is based on σ' ("vegetation density") the first few PCs will contain information about the main seasonal cycle of vegetation density. As shown in Fig. 4.3a, the first three PCs suggest that the lowest variation between the grid points occur during summer (when all vegetation is dense) and the largest variation occurs during autumn and winter (when only some vegetation is dense). This may be why areas containing relatively little vegetation (e.g. urban areas such as Paris and Lille) perform poorly in terms of η^2 when only one or two PCs are used. If the first couple of PCs mainly capture information about vegetation, it follows that data from areas with little vegetation (and hence, weak/noisy seasonal σ' behavior) cannot be summarized by the first few PCs but will require more PCs before they can properly be described.



(a) Calinski-Harabasz index **(b)** Davies-Bouldin index **(c)** Silhouette index

Fig. 4.5: Performance metrics for determining the number of clusters

4.3 Determining the number of clusters

In this section, an appropriate number of clusters k is determined. k is chosen based on the methods described in section 3.5.2. For the Calinski-Harabasz index, the Davies-Bouldin index and the Silhouette index, clustering is performed for $k = 2 \ldots 30$. For each value of k, the Calinski-Harabasz score, Davies-Bouldin score, and the Silhouette score are determined as described in section 3.5.2.

4.3.1 Calinski-Harabasz index

The CH index is plotted for different values of k in Fig. 4.5a, where higher scores indicate clusters of higher quality. The highest and second-highest scores are reached for 3 and 6 clusters, respectively. CH decreases for higher values of k. The general optimum is obtained for k between 3–10 clusters.

4.3.2 Davies-Bouldin index

For the DB index, lower values indicate clusters of higher quality. The lowest, second lowest, and third lowest scores are reached for 6, 10 and 13 clusters respectively, see Fig. 4.5b. Even though the choice for two clusters results in a good DB index, it is assumed that two clusters would be too general for further analysis. DB shows a general optimum for k between 6–13 clusters.

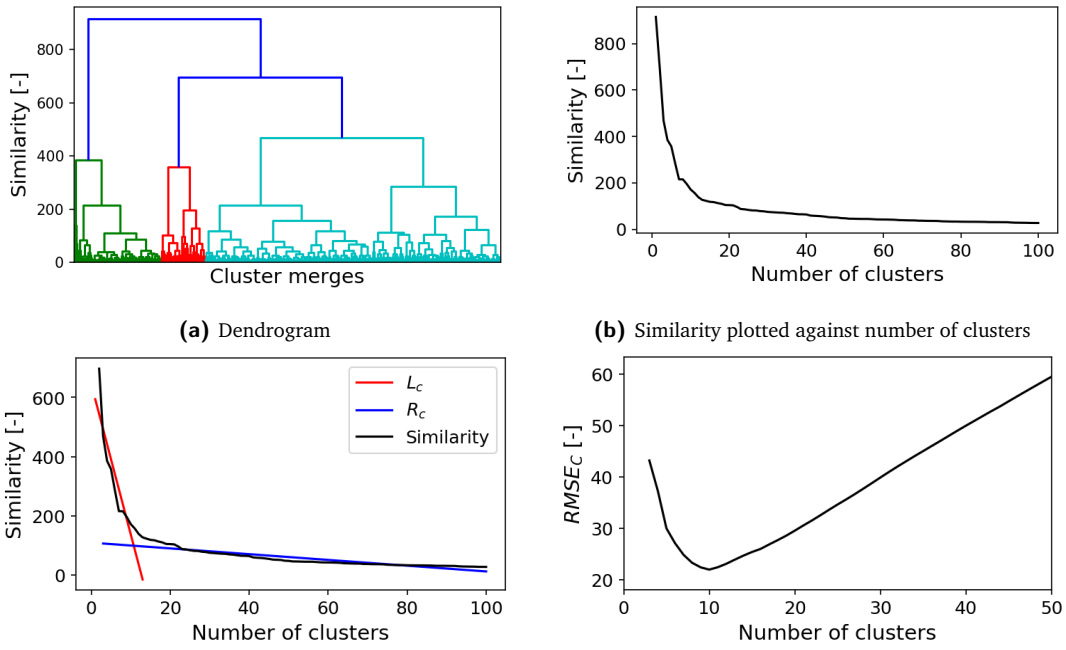
4.3.3 Silhouette index

For the SIL index, higher values indicate clusters of higher quality. Similar to the DB index, it is assumed that two or three clusters would be too general for further analysis, even though Fig. 4.5c shows that these values of k result in relatively good SIL scores. The best scores are found for 6, 9 and 13 clusters. SIL shows a general optimum for k between 6–13 clusters.

4.3.4 L-method

Agglomerative hierarchical clustering is applied in this study, which initially treats each individual observation as a separate cluster. This results in an evaluation graph with a size equal to the input data set (in this case therefore $1 < k < 3492$). However, the merges that occur at a very fine level (for a very large number of clusters) are irrelevant. These relatively irrelevant data points – which are all located on the right side of the evaluation graph – will start to dominate the solution of the L-method. Due to the large number of points located to the right of the "actual" optimum value of k, the relatively few points left of the "actual" optimum become statistically irrelevant, leading to the knee being significantly overestimated. As described by Salvador and Chan [55], in order for the L-method to perform best it is recommended that the sizes of the two best-fit lines (i.e. L_c and R_c, see section 3.5.2.4) are comparable in size. As such, the evaluation graph was truncated after $x = 100$ (i.e. $k_{max} = 100$).

Dendrograms – also known as "tree diagrams" – are used to visualize the internal structure of an agglomerative hierarchical clustering (i.e. the performed subsequent merges from bottom to top), with cluster merges on the x-axis and the similarity between clusters (e.g. distance) on the y-axis. The dendrogram of the performed clustering is plotted in Fig. 4.6a. In order to apply the L-method, the cluster merges originating from the dendrogram are recalculated to a vector containing a range of k-values, which are plotted against their respective similarities in Fig. 4.6b; this is the evaluation graph on which the L-method will be performed. For different numbers of clusters, $RMSE_c$ is calculated by finding the best-fit lines for L_c and R_c as described in 3.5.2.4, see Fig. 4.6c. Finally, when plotting $RMSE_c$ against the number of clusters a clear optimum is found for $k = 10$ clusters, see Fig. 4.6d, with the knee region ranging between approximately 8–13 clusters.



(a) Dendrogram

(b) Similarity plotted against number of clusters

(c) Knee with best-fit lines for L_c and R_c

(d) $RSME_C$ with optimum around 10 clusters

Fig. 4.6: Determining the best number of clusters using the L-method by Salvador and Chan [55]

4.3.5 Choice for number of clusters

Four different methods were applied to determine a good value for k. Even though the Calinski-Harabasz score, the Davies-Bouldin score, and the Silhouette score do not agree on a specific optimal value for k, there is a general agreement between the different metrics. The Calinski-Harabasz score, the Davies-Bouldin score, and the Silhouette score perform well for very low values of k (e.g. 2 or 3 clusters). However, a k-value of just two or three clusters may be insufficient for further analysis, as it would lead to clusters that are very general. In other words, choosing $k = 2$ or $k = 3$ may lead to the resulting clusters containing many different types of signatures, which would have limited value for further analysis.

In general, all aforementioned metrics indicate that a good value of k can be found in the range of approximately 5–15 clusters. The method by Salvador and Chan [55] shows a clear optimum for $k = 10$ clusters. It should be noted that these metrics almost never return a definitive optimal k-value, and that the best value for k may differ for each metric. Instead, these metrics are used as general indicators and as an aid to visual inspection for choosing a good value for k. Since the general optimum for k was found to be around 10 clusters for all scores, and because the L-method by Salvador and Chan [55] returned a clear optimum at $k = 10$, it was chosen to perform the clustering for $k = 10$ clusters.

4.4 Clustering

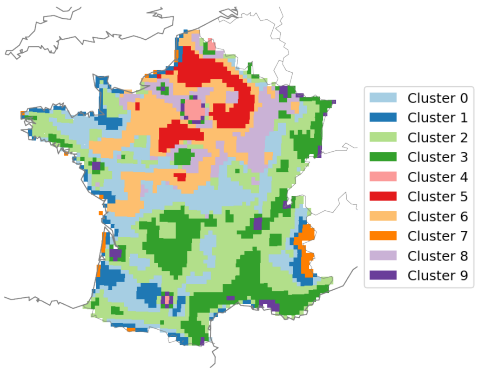
An agglomerative hierarchical clustering was performed using Ward's method, with as input five principal component weights based on the 10-year average σ' data for each of the 3492 grid points, generating a total of 10 clusters. In this section, the characteristics of the resulting clusters will be investigated. The characteristics of all generated clusters can be found in Appendix C.

4.4.1 Generated clusters

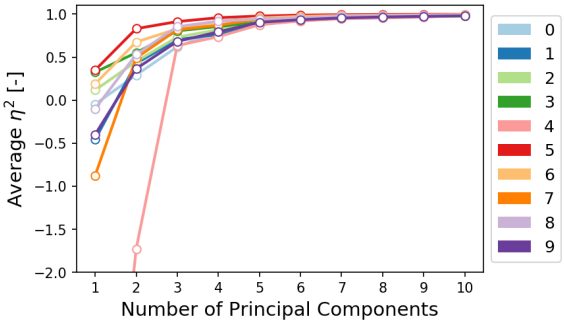
After performing the hierarchical clustering, each grid point is assigned a cluster label between $0 \ldots 9$. The grid points are mapped in Fig. 4.7a, where each grid point is colored based on its respective cluster label. Clear patterns emerge when mapping the individual grid points; instead of producing a noisy field, the generated clusters are generally contiguous and have clearly defined shapes. Spatial consistencies can be identified between the clusters and land cover maps (Appendix A); examples include Paris, the agricultural area surrounding Paris, the Alps, and the Landes coniferous forest south of Bordeaux. These consistencies indicate that the generated clusters may be related to land cover composition.

For each of the generated clusters, the PCA performance is determined by reconstructing the data for all grid points belonging to a certain cluster and subsequently calculating the average η^2 using 1–10 PCs. The resulting PCA performance per cluster is plotted in Fig. 4.7b. On average, the PCs perform differently for every cluster, which may be due to differences between clusters in terms of land cover footprint. This underlines the importance of determining the correct amount of principal components to retain.

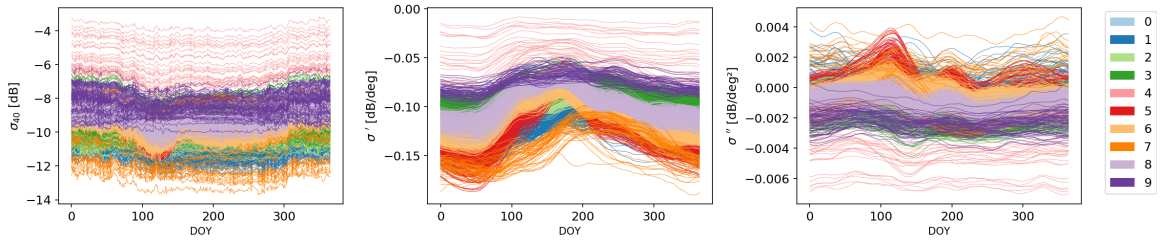
Since clustering is based on σ', is can be expected that the clusters differ in terms of σ'. However, Fig. 4.7c indicates that the clusters also have unique σ°_{40} and σ'' signatures, which suggests that the clusters represent "scattering surfaces" which differ in terms of their scattering characteristics.



(a) Generated clusters



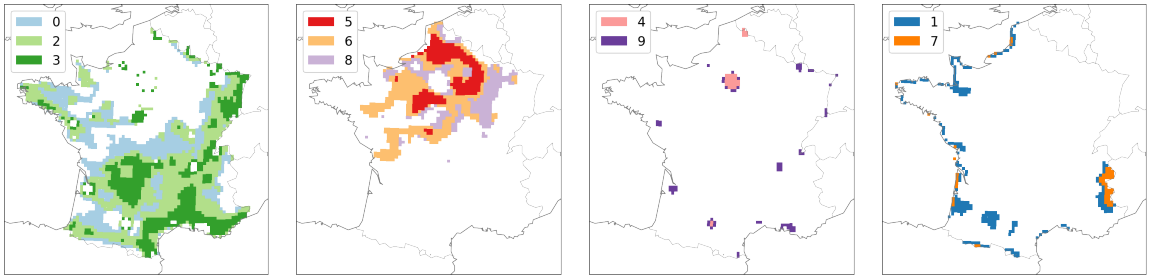
(b) Performance of cumulative PCs for all clusters



(c) Seasonal signatures of σ°_{40}, σ', σ'' for all clusters

Fig. 4.7: Generated clusters

Four categories (Fig. 4.8) have been identified based on the land cover footprints and seasonal signatures of the generated clusters: mixed clusters, agricultural clusters, urban clusters, and miscellaneous clusters. In the following sections, the aforementioned categories and their corresponding clusters will be discussed.



(a) Mixed clusters **(b)** Agricultural clusters **(c)** Urban clusters **(d)** Miscellaneous clusters

Fig. 4.8: Four cluster categories: mixed, agricultural, urban, and miscellaneous.

4.4.2 Mixed clusters

Mixed clusters are defined as clusters that have a relatively heterogeneous land cover footprint as well as noisy/unclear backscatter signatures. Clusters 0, 2 and 3 are similar in terms of their land cover footprint and backscatter signatures, and have been identified as being mixed clusters.

4.4.2.1 Cluster 0: Grassy croplands

Grid points

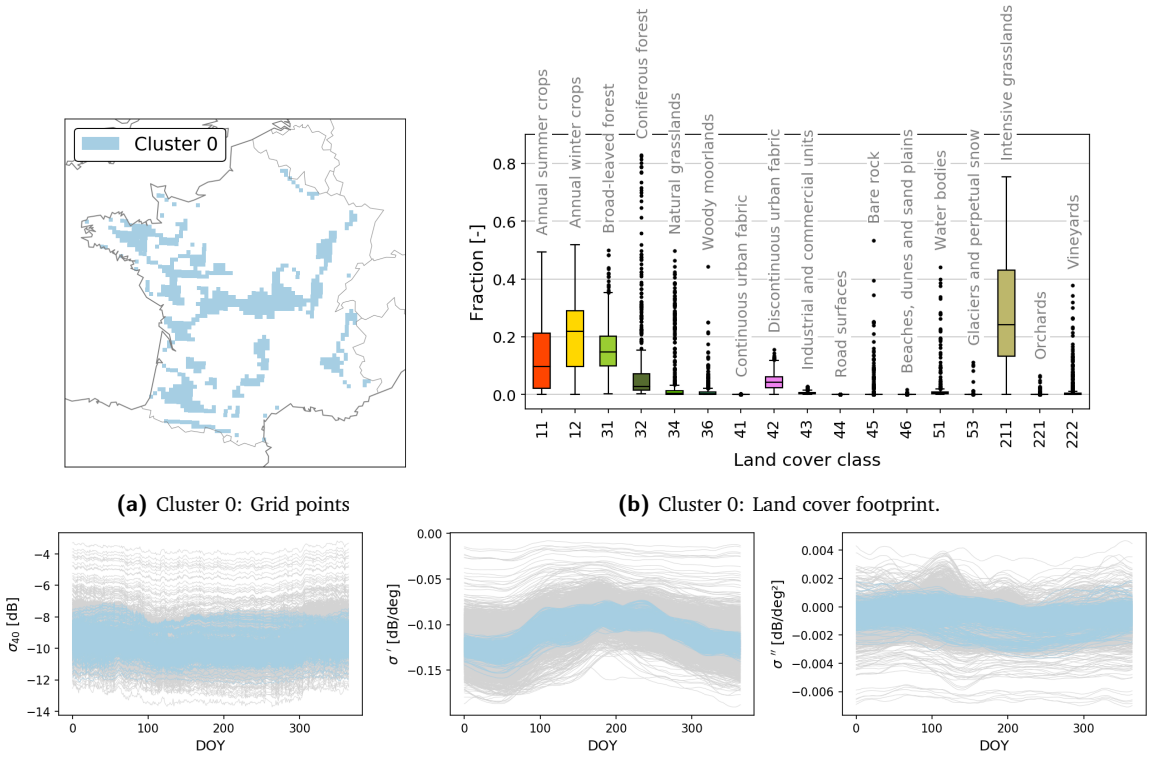
Cluster 0 is spread out throughout France, with grid points located mainly in the north-west, central, and south-west areas of France (Fig. 4.9a). Cluster 0 mainly consists of slim, elongated areas and small isolated areas instead of one single contiguous area.

Backscatter signatures

The backscatter signatures of cluster 0 are relatively noisy and contain several temporal patterns (Fig. 4.9c). On average, σ°_{40} decreases between day 50 – 100 and increases between day 100 – 150, and stays relatively constant afterward. Maximum σ' values are reached during summer and minimum values occur during winter, which corresponds with the vegetation growth cycle. Several patterns can be observed in σ', which may be because the different vegetation types occurring in cluster 0 reach maximum biomass at different moments. Maximum σ'' values are observed during (late) winter and minimum values occur during summer. While σ'' is generally close to zero, positive values are observed around day 50 – 100, indicating slightly larger backscatter at larger incidence angles during the winter period. The increased backscatter at larger incidence angles may be caused by exposed trunks and branches generating larger backscatter after seasonal leaf drop. This is corroborated by the larger σ°_{40} values during this period.

Land cover

Cluster 0 has a heterogeneous land cover footprint in which none of the land cover classes are clearly dominant (Fig. 4.9b). Several vegetation types are found in cluster 0: intensive grasslands ($\bar{p} = 28.6\%$), annual winter crops (20.2%), broad-leaved forest (16.0%), annual summer crops (12.7%), and some coniferous forest (8.1%). Furthermore, a significant number of outliers are present in several land cover classes. The heterogeneous land cover footprint may explain the relatively noisy backscatter signatures, as it is likely that different land cover footprints will produce backscatter signatures with different patterns/timings. However, despite the relatively variable land cover footprints of the grid points of cluster 0, their σ' signatures are statistically similar (compared to the other clusters).

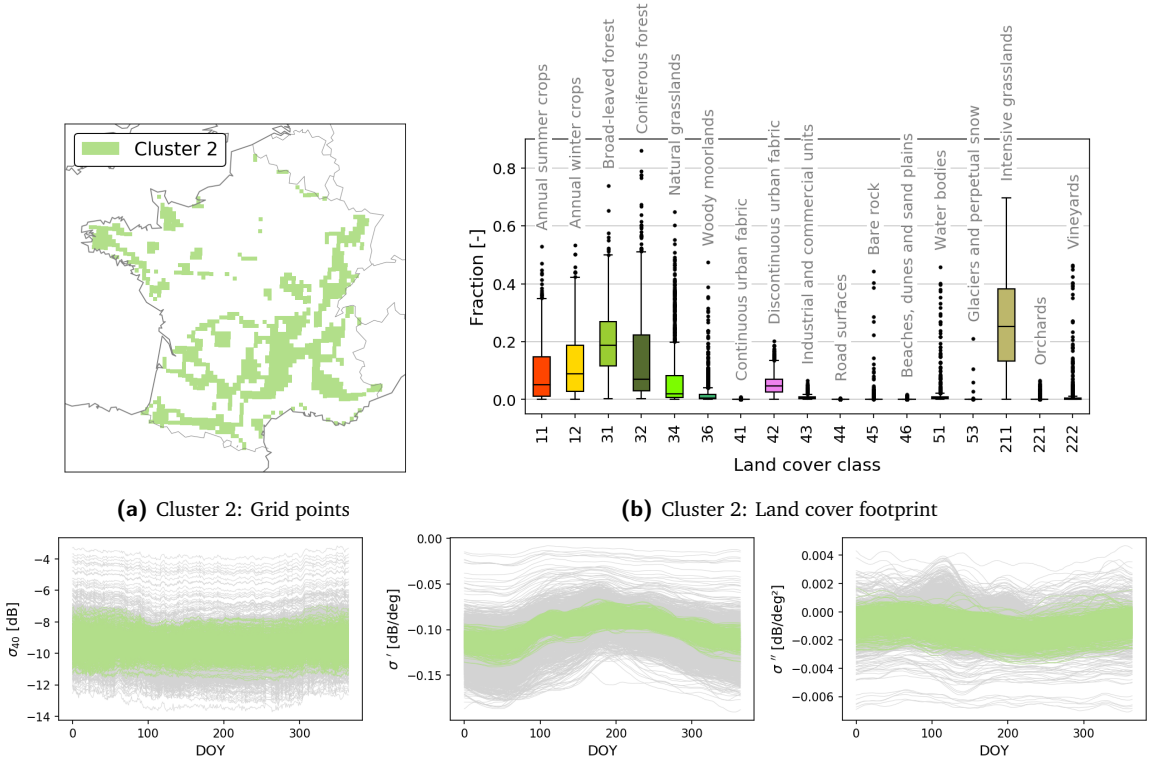


(a) Cluster 0: Grid points

(b) Cluster 0: Land cover footprint.

(c) Cluster 0: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.9: Characteristics of cluster 0 (grassy croplands). All grid points of cluster 0 are mapped in Fig. 4.9a, the land cover footprints of these grid points are visualised in Fig. 4.9b, and the σ°_{40}, σ', and σ'' signatures of cluster 0 are plotted in Fig. 4.9c. For descriptions of each of the land cover classes depicted in Fig. 4.9b see appendix A.2 [29]. The characteristics of all clusters can be found in Appendix C.



(a) Cluster 2: Grid points

(b) Cluster 2: Land cover footprint

(c) Cluster 2: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.10: Characteristics of cluster 2 (wooded grasslands and crops)

4.4.2.2 Cluster 2: Wooded grasslands and crops

Grid points

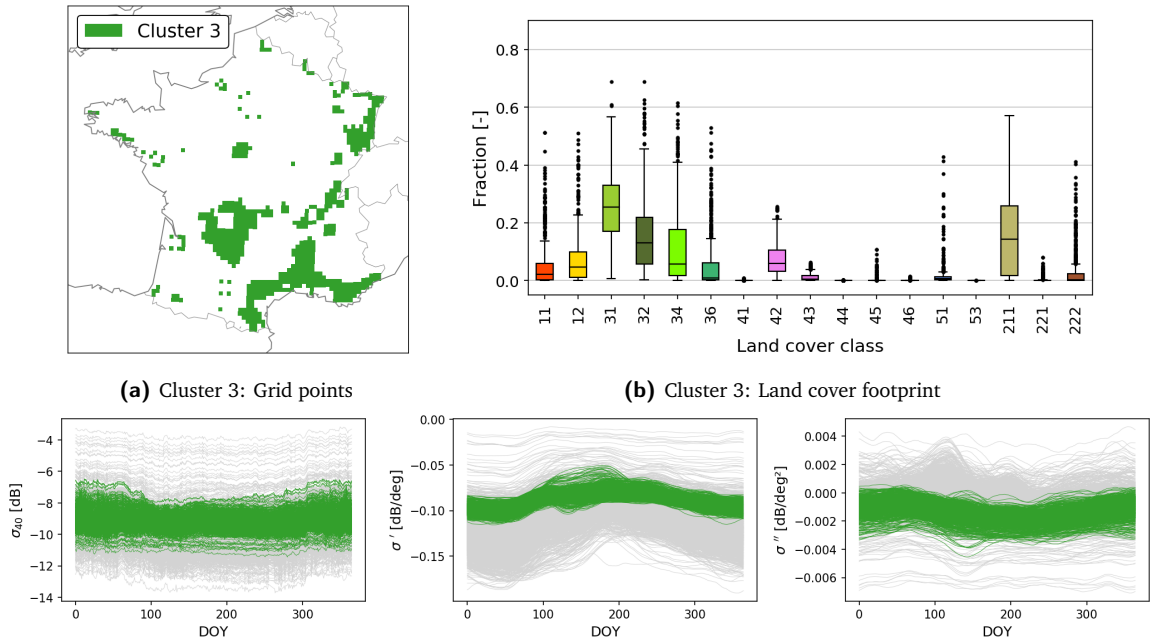
Cluster 2 is located throughout France, ranging from the south-west to the north-east of France, with some grid points in north-west France (Fig. 4.10a). Cluster 2 consists of a few larger, contiguous areas as well as slim, elongated areas, small isolated areas, and some stray grid points. Fig. 4.7a shows that cluster 2 often lies close to or between clusters 0 and 3, which may indicate that it is a transitional area between clusters 0 and 3 and has a land cover footprint and backscatter behavior similar to these areas.

Backscatter signatures

As shown in Fig. 4.10c, the backscatter signatures of cluster 2 are indeed similar to those of clusters 0; σ°_{40} is largest during winter, decreases between day 50 – 100 and slowly increases during the remainder of the year. σ' is lowest during winter and reaches maximum values during summer, again corresponding with vegetation growth cycle (i.e. maximum biomass during summer, minimum biomass during winter). Similar to cluster 0, σ'' is generally negative throughout the year for most grid points, but positive values are observed for some grid points during winter and early spring (day 0 – 100).

Land cover

Cluster 2 has a relatively heterogeneous land cover footprint, which mainly consists of intensive grasslands ($\bar{p} = 26.3\%$), broad-leaved forest (19.8%), coniferous forest (13.8%), annual winter crops (11.7%) and annual summer crops (9.0%), see Fig. 4.10b. Overall, the land cover footprint of cluster 2 is very similar to that of cluster 0; the main occurring land cover classes are the same for clusters 0 and 2, but their fractions differ slightly. For example, cluster 2 has lower fractions of annual summer crops (9.0 % vs 12.7%) and annual winter crops (11.7% vs 20.2%), but has slightly higher fractions of coniferous forest (13.8% vs 8.1%), natural grasslands (7.4% vs 4.0%), and broad-leaved forest (19.8% vs 16%). This may explain why cluster 2 is often located next to cluster 0 and why the seasonal signatures of cluster 2 are similar to those of cluster 0, in terms of both scale and seasonal patterns. The heterogeneous land cover footprint – consisting mainly of different vegetation types – may explain why different seasonal patterns can be observed in σ', while still exhibiting a general seasonal cycle.



(a) Cluster 3: Grid points

(b) Cluster 3: Land cover footprint

(c) Cluster 3: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.11: Characteristics of cluster 3 (grassy forests)

4.4.2.3 Cluster 3: Grassy forests

Grid points

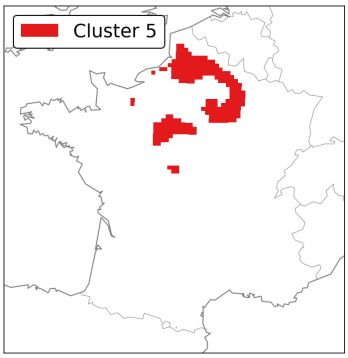
Cluster 3 consists of a few large, clearly defined contiguous areas in central and north-east France, and along the southern coast (Fig. 4.11a). Smaller, less defined areas and stray grid points near urban areas are also visible. Cluster 3 generally borders cluster 2, indicating that they may have similar characteristics.

Backscatter signatures

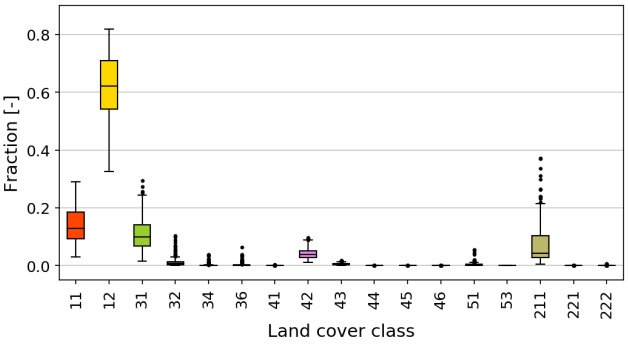
Indeed, Fig. 4.11c shows that the σ°_{40}, σ', and σ'' signatures of cluster 3 are similar to those of clusters 2 (and to a lesser extent, cluster 0). In general, σ°_{40} is at maximum values between day 300 – 80, decreases between day 80 – 100, and slowly recovers between day 100 – 300. Compared to other clusters, σ' is relatively close to zero indicating a stable vegetation cover, part of which may persist year-round. For a number of grid points, σ' exhibits different behavior between day 100 – 200, perhaps due to the disproportionate presence of a vegetation class that is characterized by large changes in σ' during this period. Such locally divergent behavior can be further investigated by performing a within-cluster analysis. Similar to cluster 2, σ'' has a generally negative, gentle seasonal cycle that reaches maximum values during winter and minimum values during summer.

Land cover

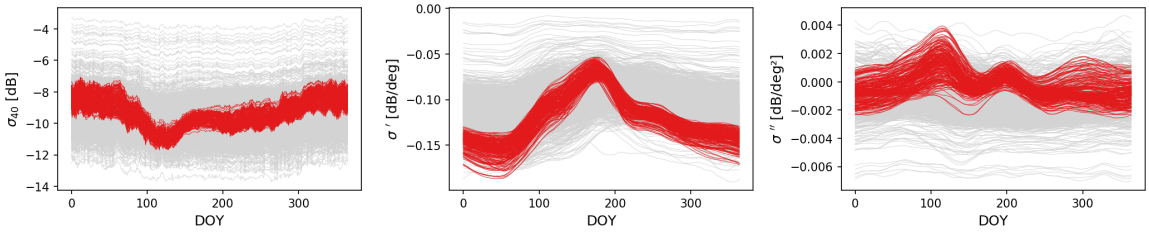
The land cover footprint of cluster 3 is relatively heterogeneous and consists mainly of broad-leaved forest ($\bar{p} = 25.4\%$), coniferous forest (15.9%), intensive grasslands (15.8%), natural grasslands (11.3%), as well as some annual winter and summer crops (7.4% and 5.3%, respectively), see Fig. 4.11b. The land cover footprint of cluster 3 is similar to those of clusters 0 and 2; cluster 3 contains the same main land cover classes in slightly different fractions. Compared to cluster 2, cluster 3 has higher fractions of broad-leaved forest (25.4% vs 19.8%), coniferous forest (15.9% vs 13.8%), and natural grasslands (11.3% vs 7.4%), and has lower fractions of annual summer crops (5.3% vs. 9.0%), annual winter crops (7.4% vs 11.7%), and intensive grasslands (15.8% vs 26.3%). This is consistent with the idea that clusters with similar backscatter characteristics are generally located near each other, and have land cover footprints consisting of the same main land cover classes (usually with slightly different fractions).



(a) Cluster 5: Grid points



(b) Cluster 5: Land cover footprint



(c) Cluster 5: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.12: Characteristics of cluster 5 (intense agriculture)

4.4.3 Agricultural clusters

Several clusters are characterized by a land cover footprint that is dominated by agriculture (i.e. annual summer crops and annual winter crops). Interestingly, these agricultural clusters are also similar in terms of their backscatter behavior, which is relatively interesting and distinct compared to the backscatter characteristics of mixed clusters. Clusters 5, 6 and 8 have been identified as agricultural clusters.

4.4.3.1 Cluster 5: Intense agriculture

Grid points

Cluster 5 consists of two large contiguous areas near Paris, as well as some smaller areas in the general vicinity (Fig. 4.12a). As discussed in section 3.1, this area is well known for its intensive agriculture, and is clearly recognizable in land cover maps (Fig. A.1 and A.2b in appendix A).

Backscatter signatures

Due to the intensive agriculture present in this area, cluster 5 may have the most distinct and interesting σ°_{40}, σ', and σ'' signatures out of all clusters, see Fig. 4.12c. A comprehensive investigation of effect of the growth cycle of winter crops on the σ°_{40}, σ', and σ'' signatures can be found in appendix D.1.2.

σ°_{40} is relatively constant and largest during winter (day 300 – 80). The decrease in σ°_{40} during (early) spring can be explained by the rapid development of the vertical crop structure due to elongation of the main crop stem; the vertical plant components cause increased attenuation of the incident wave, resulting in lower total backscatter (see appendix D.1.2). Direct backscatter rapidly increases between day 130 – 160 due to development of the vegetation canopy, which corresponds to the observed increase in σ°_{40}. During the second half of the year σ°_{40} slowly increases until reaching maximum values during winter.

Minimum values for σ' occur during late fall and winter (day 300 – 50) and maximum values occur during summer (day 150 – 200), which corresponds to the seasonal growth cycle of winter crops. σ' decreases in late summer and early fall (day 180 – 300), which indicates a decrease in (wet) biomass; this is explained by crop maturation (i.e. loss of water through evaporation, grain ripening and leaf/stem senescence) and subsequent crop harvesting.

A relatively distinct and interesting pattern can be observed in σ'', which has two maximum (positive) values around day 120 and 200 and two minimum (negative) values around day 160 and 250. As described in appendix D.1.2, the oscillations between positive and negative σ'' values seem to correspond to changes in the physical structure and water distribution of crops, caused by processes such as rapid elongation of the main stem, transport of water stored in the stem to the developing grains, and loss of water due to grain ripening. These results correspond with the current understanding of σ'' as a measure of the relative dominance of ground-bounce scattering over direct scattering from vertical vegetation constituents.

Compared to other clusters, σ°_{40}, σ', and σ'' signatures exhibit very distinct behavior without much noise, which may indicate that cluster 5 has a relatively homogeneous land cover footprint.

Land cover

Fig. 4.12b shows that cluster 5 has a homogeneous land cover footprint compared to other clusters, consisting mainly of annual winter crops (61.1%), annual summer crops (13.7%), broad-leaved forest (10.9%), and some intensive grasslands (7.4%). Furthermore, the land cover footprint of cluster 5 contains relatively few outliers, so most grid points have a roughly similar land cover footprint. This indicates that the unique backscatter signatures observed in this area are likely caused mainly by distinct crop growth processes and the intensive agricultural cycle. However, even though agriculture is clearly dominant in cluster 5, it should be noted that other land cover classes are always present. Therefore, the observed backscatter signatures should not be seen as characteristic for agricultural vegetation only.

4.4.3.2 Cluster 6: Grassy agriculture

Grid points

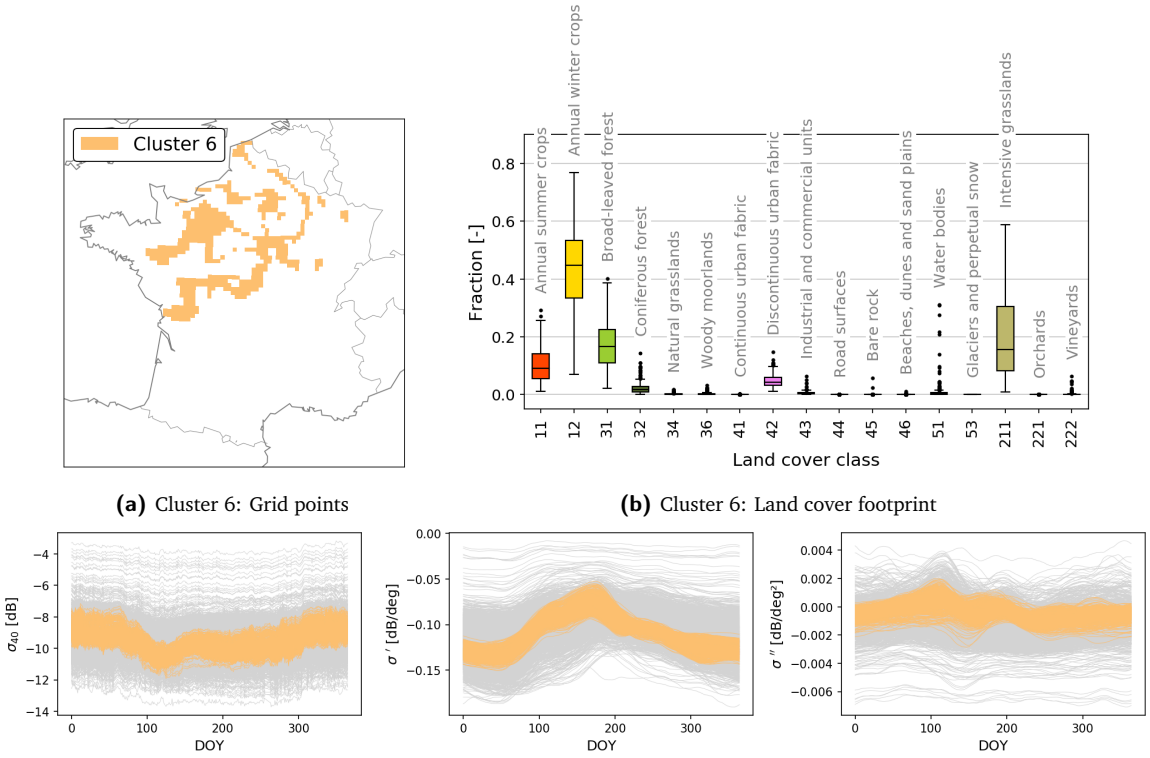
Similar to cluster 5, cluster 6 is located in the central-north of France, an area known for its intensive agriculture. Cluster 6 consists of a few larger contiguous areas, as well as some stray grid points and some long, stretched areas, see Fig. 4.13a. Cluster 6 mainly borders clusters 5 and 8, indicating that their characteristics are likely to be similar.

Backscatter signatures

As can be seen in Fig. 4.13c, the σ°_{40}, σ', and σ'' signatures of cluster 6 are indeed similar to those of cluster 5, but slightly less defined; σ°_{40} is largest between day 300 – 80, decreases sharply between day 80 – 120, increases between day 120 – 150, and slowly increases until reaching maximum values in winter. Additionally, the seasonal cycle of σ' is similar to that of cluster 5, with minimum values in winter and maximum values in summer. Finally, σ'' exhibits the same 'double-peak' behavior that was observed in cluster 5, suggesting a significant presence of agricultural land cover. However, σ'' oscillates in a narrower band compared to cluster 5, which could indicate that changes in dominant scattering mechanism are less significant, possibly due to a lower fraction of agricultural land cover.

Land cover

Fig. 4.13b shows that the land cover footprint of cluster 6 is similar to that of cluster 5, consisting mainly of annual winter crops, intensive grasslands, broad-leaved forest and annual summer crops. However, while clusters 5 and 6 contain the same dominant land cover classes, the fractions in which these classes occur is different. Compared to cluster 5, cluster 6 contains less annual winter crops (44.2% vs 61.1%), less annual summer crops (10.3% vs 13.7%), more intensive grasslands (19.3% vs 7.4%) and more broad-leaved forest (17.3% vs 10.9%). The differences in land cover footprint between cluster 5 and cluster 6 correspond to the observed differences between them in terms of spatial distribution and σ°_{40}, σ', and σ'' signatures.

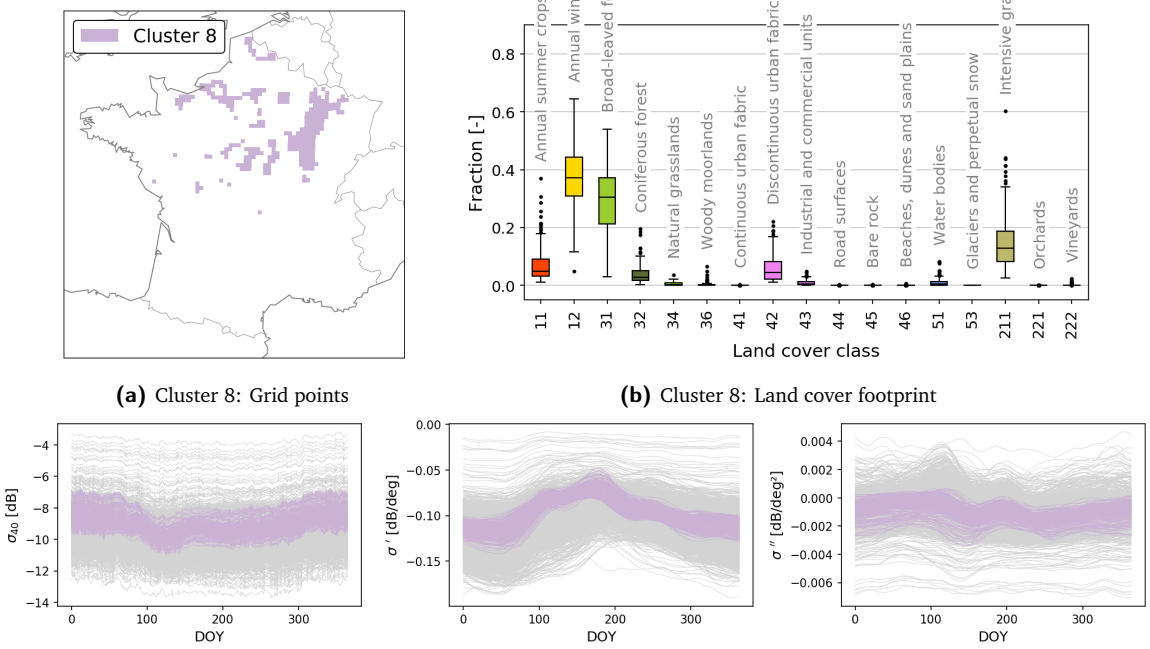


(a) Cluster 6: Grid points

(b) Cluster 6: Land cover footprint

(c) Cluster 6: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.13: Characteristics of cluster 6 (grassy agriculture)



(a) Cluster 8: Grid points

(b) Cluster 8: Land cover footprint

(c) Cluster 8: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.14: Characteristics of cluster 8 (wooded agriculture)

4.4.3.3 Cluster 8: Wooded agriculture

Grid points

Cluster 8 is located near clusters 5 and 6 in the north and north-west of France and consists of one large contiguous area, some small contiguous areas, and several stray grid points (Fig. 4.14a). As discussed in section 4.4.3.2, cluster 8 mainly borders clusters 0 and 6, which suggests that their backscatter and land cover characteristics may be similar.

Backscatter signatures

While the σ°_{40}, σ', and σ'' signatures of cluster 8 are markedly less clear compared to clusters 5 and 6, similar behavior can still be discerned (Fig. 4.14c). σ°_{40} is maximum during winter, decreases during (early) spring, and recovers during the second half of the year. σ' is at a minimum during winter (sparse vegetation), and at a maximum during summer (dense vegetation). Even though the previously observed 'double-peak' behavior of σ'' is less defined than in cluster 5, similar oscillations in σ'' are still visible in cluster 8 – this suggests that scattering mechanism dominance changes throughout the year, likely due to the presence of agriculture.

Land cover

The land cover footprint of cluster 8 is similar to those of cluster 5 and 6, but the fractions in which the dominant land cover classes occur are different (Fig. 4.14b). Compared to cluster 6, cluster 8 generally has less annual winter crops (37.9% vs 44.2%), less intensive grasslands (14.1% vs 19.3%), less summer agriculture (7.0% vs 10.3%) and more broad-leaved forest (28.9% vs 17.3%). Compared to cluster 5, cluster 8 has significantly less winter agriculture (37.9% vs 61.1%), less summer agriculture (7.0% vs 13.7%), more intensive grasslands (14.1% vs 7.4%) and more broad-leaved forest (28.9% vs 10.9%). In this case, the similarities between the land cover footprints of clusters 5, 6 and 8 explain why these clusters all exhibit similar σ°_{40}, σ', and σ'' patterns, while the differences in land cover footprint (i.e. different fractions of dominant land cover classes) explain why these grid points have been split into three clusters.

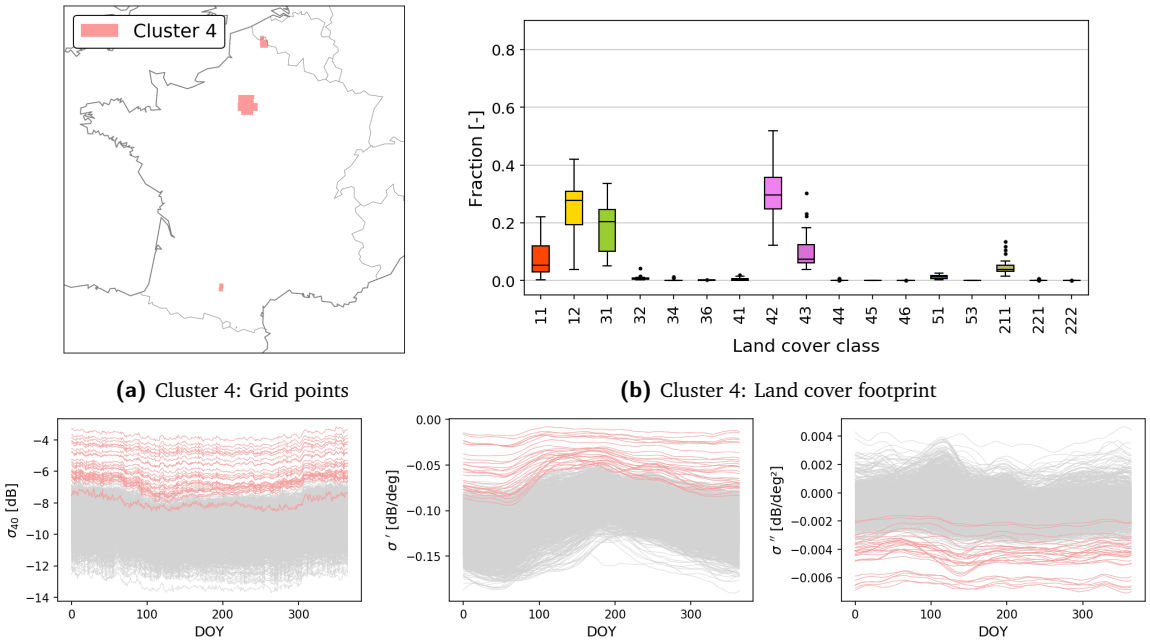
4.4.4 Urban clusters

Urban clusters are defined as clusters containing relatively high fractions of urban land cover classes, mainly class 42 (continuous urban fabric) and class 43 (discontinuous urban fabric). Due to their unique geometry – flat surfaces joined at 90° angles – urban areas act as so-called *corner reflectors* (see appendix D.2.1), generating a strong backscatter signal for (nearly) all incidence angles. Cluster 4 and cluster 9 are both urban clusters.

4.4.4.1 Cluster 4: City centers

Grid points

Cluster 4 is a relatively small cluster located in the highly urbanized areas of Paris, Toulouse and Lille (Fig. 4.15a). Cluster 4 is clearly recognizable in Fig. A.1, Fig. A.2g, Fig. A.2h and Fig. A.2i in appendix A. However, not all urban areas visible in these land cover maps are assigned to cluster 4; this may be because these areas do not contain enough urban area (i.e. contain significantly more vegetation) to be clustered together with highly urbanized areas such as Paris.



(a) Cluster 4: Grid points

(b) Cluster 4: Land cover footprint

(c) Cluster 4: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.15: Characteristics of cluster 4 (city centers)

Backscatter signatures

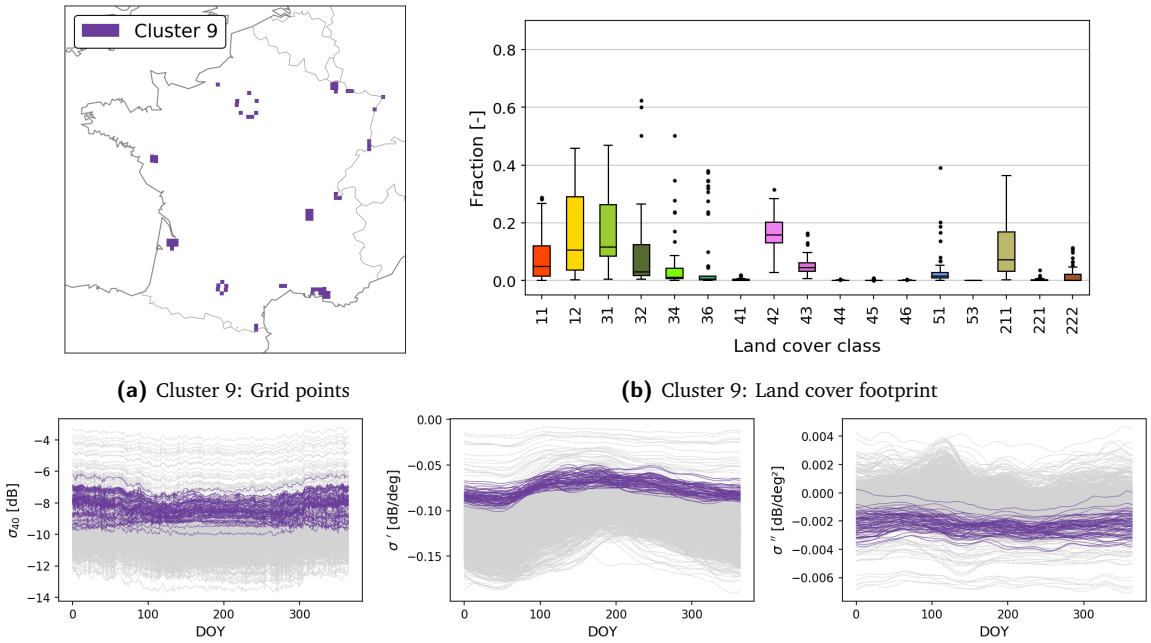
Even though it is expected that there will be little seasonal behavior for highly urban areas, Fig. 4.15c shows that cluster 4 has clear backscatter characteristics; out of all clusters, cluster 4 contains the largest σ°_{40}, the largest (i.e. shallowest) σ', and the most negative σ'', all relatively constant throughout the year. This is explained by the characteristics of urban areas and the (comparatively) sparse vegetation.

Highly urbanized areas act as retroreflectors due to their unique structural characteristics, leading to large σ° values over most incidence angles. Moreover, σ°_{40} is relatively constant in time because the structural characteristics of urban areas do not change significantly throughout the year. In turn, this means that the $\sigma^\circ - \theta$ relationship of highly urbanized areas is relatively shallow and constant, i.e. σ' will be close to zero throughout the year. The observed negative σ'' can be explained by the fact that corner reflectors generate the highest backscatter for $40° < \theta < 50°$, with decreasing σ° for $\theta < 40°$ or $\theta > 50°$ [37]; this results in a concave $\sigma^\circ - \theta$ relationship (i.e. $\sigma'' < 0$).

However, even though the seasonal signatures of cluster 4 are constant compared to other clusters, slight seasonal behavior can be seen in Fig. 4.15c. This is likely because cluster 4 does not consist entirely out of urban area, but its land cover footprint also contains vegetation in smaller fractions. Finally, it should be noted that σ' values closer to zero generally indicate higher vegetation densities compared to lower σ' values, but this is not necessarily true for urban grid points. Even though the σ' values of cluster 4 are closest to zero compared to all other clusters, this is more due to the unique backscatter effects of urban areas rather than due to larger vegetation density.

Land cover

The land cover footprint of cluster 4 mainly consists of discontinuous urban fabric (30.6%) and industrial and commercial units (10.3%), but also has significant fractions of annual winter crops (25.1%), broad-leaved forest (18.4%), annual summer crops (7.8%) and some intensive grasslands (4.9%), see Fig. 4.15b. This is consistent with the observations made earlier in this section. Furthermore, the fact that vegetation is indeed present in cluster 4 explains why variations in σ°_{40}, σ', and σ'' can be identified.



(a) Cluster 9: Grid points

(b) Cluster 9: Land cover footprint

(c) Cluster 9: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.16: Characteristics of cluster 9 (green suburbs)

4.4.4.2 Cluster 9: Green suburbs

Grid points

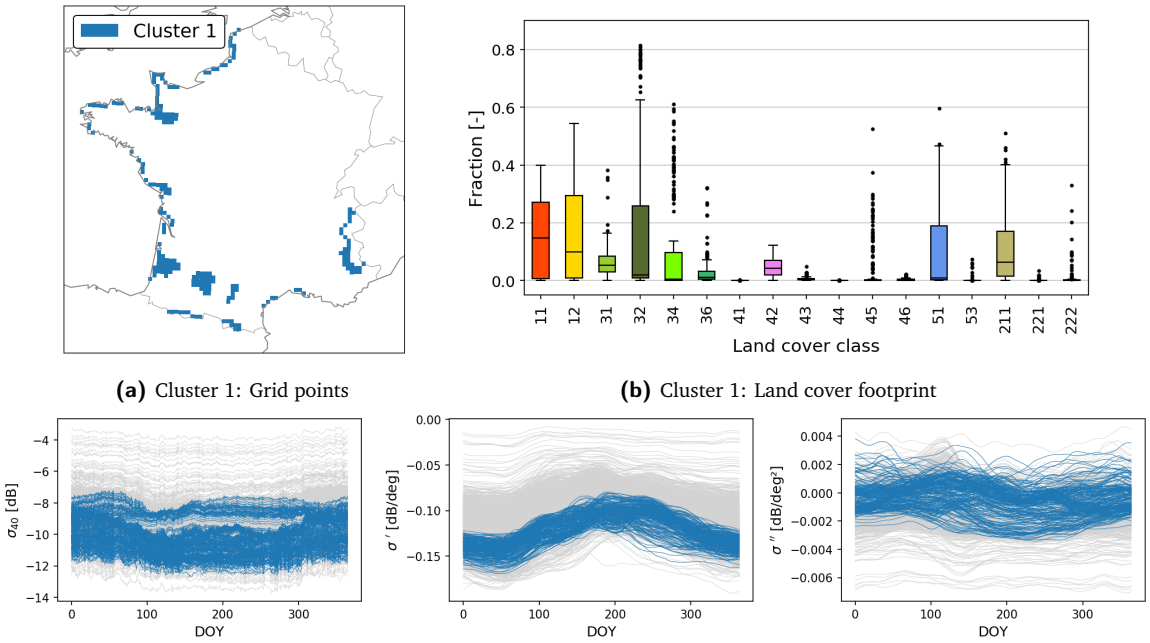
Cluster 9 consists of small areas located near urban areas such as Paris, Bordeaux, Montpelier, Lyon, and Toulouse (Fig. 4.16a). Since these areas are not clustered together with cluster 4 but are always located near urban areas, it is likely that the land cover footprint of cluster 9 has lower fractions of urban area and higher fractions of vegetation classes compared to cluster 4.

Backscatter signatures

The backscatter signatures of cluster 9 are quite constant and noisy, with relatively large σ°_{40}, shallow σ', and negative σ'' compared to the other clusters (see Fig. 4.16c). However, σ°_{40}, σ', and σ'' of cluster 9 show slightly more distinct seasonal behavior than cluster 4. These seasonal signatures suggest that cluster 9 contains a significant amount of urban area, which is also indicated by the spatial distribution of the grid points. However, cluster 9 likely has denser vegetation than cluster 4 as indicated by the slightly more distinct seasonal behavior of σ°_{40}, σ', and σ''.

Land cover

The land cover footprint of cluster 9 is relatively heterogeneous, see Fig. 4.16b. However, cluster 9 indeed contains a relatively large fraction of urban area (16.7% discontinuous urban fabric, 5.2% industrial and commercial units). Other land cover types include annual winter crops (16.7%), broad-leaved forest (16.3%), intensive grasslands (10.2%), coniferous forest (9.6%), annual summer crops (8.6%), woody moorlands (5.9%) and natural grasslands (4.8%). Clearly, the grid points of cluster 9 are spread throughout France, so it can be expected that they do not share the same land cover footprint. However, the grid points have in common a relatively large degree of urban area. Consequently, the seasonal signals originating from the different vegetation classes are disproportionately affected by urban areas, resulting in large σ°_{40} values, shallow σ' values, and negative σ'' values, with noticeable (but mixed) seasonal behavior. This explains why the grid points of cluster 9 have been clustered together: even though their land cover footprints may be very different in terms of vegetation, their scattering characteristics are similar due to the strong scaling effect of urban areas.



(a) Cluster 1: Grid points

(b) Cluster 1: Land cover footprint

(c) Cluster 1: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.17: Characteristics of cluster 1 (mixed coastal and mountain vegetation)

4.4.5 Miscellaneous clusters

4.4.5.1 Cluster 1: Mixed coastal and mountain vegetation

Grid points

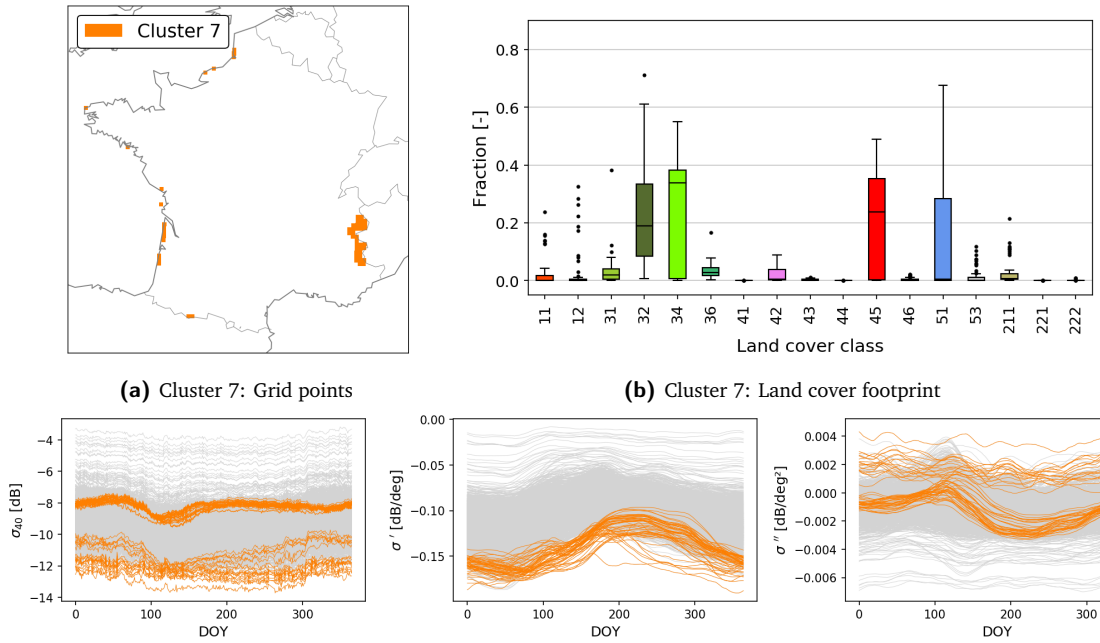
Cluster 1 is located along the French coast, south of Bordeaux, and near the Alps and Pyrenees (Fig. 4.17a). Some smaller, well defined contiguous areas can be identified near Bordeaux and Toulouse, while most stray grid points are found along the coast.

Backscatter signatures

The σ°_{40}, σ', and σ''signatures of cluster 1 are shown in Fig. 4.17c. σ°_{40} shows two distinctly different patterns; larger σ°_{40} values are observed for a subset of grid points, ranging between approximately -9 dB and -7 dB. For the other grid points, σ°_{40} ranges between approximately -12 dB and -8 dB. Despite the differences in σ°_{40}, the main seasonal behavior of σ' is relatively similar for all grid points, with minimum values during winter and maximum values during summer. σ'' is relatively noisy and close to zero.

Land cover

The land cover composition of cluster 1 is relatively heterogeneous and consists mainly of coniferous forest (17.0%), annual winter crops (16.1%), annual summer crops (15.7%), intensive grasslands (11.7%), natural grasslands (9.9%), and water bodies (9.7%), see Fig. 4.17b. The occurrence of several vegetation classes explain why σ' has a general seasonal cycle but lacks a distinct pattern, while the outliers suggest that different land cover footprints exist in cluster 1. However, since the land cover footprint is aggregated across the entire cluster, it is difficult to obtain insights into the several contiguous within-cluster areas that cluster 1 contains. For example, the area south of Bordeaux (clearly visible in Fig. A.1 and Fig. A.2d) has a homogeneous land cover footprint dominated by coniferous forest. It may be possible that such within-cluster areas have homogeneous land cover footprints (and distinct σ°_{40}, σ', and σ'') when viewed separately, while the land cover footprint of the entire cluster seems very heterogeneous.



(a) Cluster 7: Grid points

(b) Cluster 7: Land cover footprint

(c) Cluster 7: σ°_{40}, σ', σ'' seasonal signatures

Fig. 4.18: Characteristics of cluster 7 (sparse coastal and mountain vegetation)

4.4.5.2 Cluster 7: Sparse coastal and mountain vegetation

Grid points

Cluster 7 consists of one contiguous area that coincides with the Alps (which is clearly visible in Fig. A.2k) as well as individual grid points dotted along the east coast of France.

Backscatter signatures

While σ' is relatively similar for all grid points in cluster 7, distinctly different patterns are visible in σ°_{40} and σ'', see Fig. 4.18c. This is likely because cluster 7 is located in two distinctly different areas. Two bands are visible in σ°_{40}; the first band has a clear pattern ranging between -9 dB and -7.5 dB; the second band is less defined and ranges between -13.5 dB and -10 dB. σ' is noisy but similar for all grid points, with minimum values during winter and maximum values during summer. σ'' also consists of two bands; the first band is relatively constant and noisy, ranging between 0.001 – 0.004 dB/deg^2. The second band does have a clear seasonal cycle ranging between -0.003 – 0.001 dB/deg^2, with maximum values around day 100 and minimum values around day 200. This behavior of σ'' suggests a large fraction of coniferous forest (see appendix D.1.4). It is interesting that these areas – while different in terms of σ°_{40} and σ'' – are similar in terms of σ'. This may be due to a similar composition of vegetation cover leading to similar σ'. It may also be that these areas have different land cover footprints resulting in similar σ' signals.

Land cover

The land cover footprint of cluster 7 contains a few dominant land cover classes, see Fig. 4.18b. Notable are the high fractions of bare rock (21.3%) and water bodies (12.1%); clearly, the high fraction of bare rock corresponds to the Alps/Pyrenees, while the high fraction of water bodies belongs to the eastern coast of France. Cluster 7 also has high fractions of natural grasslands (25.6%) and coniferous forest (21.6%), as well as seven other land cover types with fractions < 5%. High fractions of coniferous forest and natural grasslands mainly occur in the Alps/Pyrenees and not near the coast, see Fig. A.2d and A.2b. The coastal and mountainous areas differ in terms of land cover – which explains the differences in σ°_{40} and σ'' – but are similar in terms of σ'. This suggests that grid points with similar σ' signatures do not necessarily have similar land cover; but similar land cover footprints do generally yield similar σ' signatures.

4.4.6 Summary of clustering results

Spatial distribution

Some clusters contain clearly defined shapes that are also recognizable in land cover maps (Appendix A.3). Cluster 1 contains a contiguous area directly south of Bordeaux that is also visible in the map of coniferous forest fractions (Fig. A.2d). Cluster 4 has grid points located in Paris, Lille and Toulouse, all of which are highlighted in the maps visualizing urban land cover types (Fig. A.2g, Fig. A.2h, Fig. A.2i). The two contiguous areas of cluster 5 are also relatively clear in the annual winter crops map (Fig. A.2b). Cluster 7 contains one contiguous area that coincides with the Alps, which can be identified in the map visualizing bare rock fractions (Fig. A.2k). Finally, similar to cluster 4, the grid points of cluster 9 are located in/near urban areas and can be identified in Fig. A.2g, Fig. A.2h, and Fig. A.2i. Other clusters (e.g. clusters 0, 2, 6, and 8) have a less defined shape and are not clearly recognized in land cover maps.

Some clusters are relatively localized, existing only in certain parts of France (e.g. clusters 5, 6, and 8), while others are more dispersed (e.g. clusters 0, 1, 2, 3, and 7). It was found that localized clusters generally have more defined σ°_{40}, σ', and σ'' signatures, while dispersed clusters have noisier signatures. On the other hand, a more localized cluster such as cluster 5 has more distinct σ°_{40} and σ'' signatures with smaller differences between grid points throughout the year.

For clusters that contain several contiguous areas (e.g. clusters 1, 3, and 7) it may be interesting to investigate these areas separately in terms of their land cover footprint and σ°_{40}, σ', and σ'' signatures; the results suggest that separate contiguous areas within a cluster have distinct characteristics (i.e. land cover composition, seasonal backscatter behavior) when investigated separately, and that these areas only appear to have mixed/noisy characteristics when investigated together as one cluster.

Seasonal signatures

A general seasonal pattern for σ' is identified in most clusters: typically, σ' is at minimum values during winter, increases during spring, reaches maximum values during summer, and decreases during autumn. This corresponds with the seasonal growth cycle of deciduous vegetation and is in line with the interpretation of σ' as a measure for 'vegetation density'. This general seasonal cycle is explained by the fact that deciduous vegetation types are present throughout France (see Fig. A.2c and Fig. A.2o). However, the timing and range of σ' differ between clusters. Some clusters have larger σ' values with lower seasonal variation, indicating that vegetation density is relatively large and part of the vegetation persists year-round (e.g. coniferous forests). Note that this is not true for clusters 4 and 9, where σ' is shallow throughout the year due to the characteristics of urban areas (appendix D.2.1). Other clusters have relatively low σ' values, indicating lower vegetation densities; cluster 7 contains large fractions of bare rock and water bodies and hence, sparser vegetation. In agricultural areas, σ' is characterized by large seasonal variations which corresponds with the growth cycle of agricultural crops (appendix D.1.2).

Two main types of σ'' behavior are identified. The first is characterized by maximum values during winter and minimum values during summer. While positive σ'' values are reached around day 50 – 100 in some grid points, σ'' is generally negative throughout the year. The second type of σ'' behavior is characterized by distinct "double peak" behavior with maxima around day 100 and day 200 and minima around day 150 and 250. Here, σ'' is generally positive or close to zero around the maxima, negative around the minima, and negative or close to zero for the rest of the year. This behavior mainly occurs in highly productive agricultural areas, where vegetation phenology changes significantly during the year.

Even though grid points with similar σ' behavior generally also have similar σ°_{40} and σ'' behavior, this is not always the case. In some cases, grid points with (relatively) similar σ' behavior have markedly different σ°_{40} and σ'' behavior (e.g. clusters 0, 1, and 7). This indicates that areas with similar vegetation density do not necessarily have the same dominant scattering mechanisms.

Land cover composition

Previously discussed results generally indicate that specific land cover footprints correspond with specific combinations of σ°_{40}, σ', and σ'' signatures; clusters containing large fractions of (summer and/or winter) agriculture have distinctly different scattering characteristics compared to clusters located in urban areas or clusters with very heterogeneous land cover footprints. For clusters with relatively homogeneous land cover footprints it may be possible to explain the observed backscatter signatures based on the occurring land cover classes. Based on the seasonal signatures of cluster 4 and cluster 9, it was found that grid points containing a significant fraction of urban area are characterized by large σ°_{40}, shallow σ', and very negative σ'' – all of which relatively stable in time – and that this behavior is explained the unique scattering characteristics of urban fabric. Furthermore, it was found that summer and winter agriculture are characterized by large seasonal variations of σ°_{40}, σ', and σ'', which may be explained by the productive and seasonal nature of agricultural lands.

However, since the clustering is based on σ' each cluster contains grid points that are similar in terms of σ' and not necessarily in terms of land cover footprint. Some clusters have a land cover footprint consisting of multiple land cover classes that is similar for all of its grid points (e.g. cluster 5), while other clusters contain very mixed land cover footprints that differ significantly between its grid points (e.g. clusters 0, 1, 2, and 3). As a result, separating the characteristic σ°_{40}, σ', and σ'' signatures of each land cover class based on the generated clusters is very complex, if not impossible.

Clusters as scattering surfaces

The main research question posed in this study focuses on whether it is possible to find distinct and meaningful scattering surfaces. The aforementioned results indicate that clusters obtained by means of hierarchical clustering based on σ' differ significantly in terms of their scattering characteristics, i.e. σ°_{40}, σ', and σ''. Moreover, differences in scattering characteristics between clusters generally correspond to (and seem to be partially explained by) differences in their land cover footprints. Hence, it can be said that hierarchical clustering based on σ' is indeed able to yield distinct and meaningful clusters, and that each of the obtained clusters represents a scattering surface with certain scattering characteristics. As such, this suggests that the obtained clusters can be seen as a scattering classification or segmentation, which describes how scattering behavior is distributed in space. As will be discussed in the next section, the fact that the vegetation parameters can now be estimated dynamically offers an additional opportunity to investigate and describe how the aforementioned scattering surfaces change over time.

4.5 Robustness

As previously discussed, currently multiple years of measurements are required to estimate the climatology of σ' and σ'' due to the amount of noise present in the observed backscatter signal. However, the recently developed method by Melzer [42] allows for the estimation of σ' and σ'' on a daily basis. In this section, 10 years of dynamically estimated σ' values are used to investigate the temporal stability (i.e. robustness) of the clusters discussed in the previous section, which were based on the climatology of σ'.

4.5.1 Clustering the 10-year data set

After restructuring the 10-year σ' data set from a size of (3492 x 3650) to (34920 x 365) as described in section 3.6, the data set was standardized and a PCA was performed as described in section 3.4. The number of PCs to retain was determined so that $\bar{\eta}^2 \geq 99\%$, resulting in 13 retained PCs. As such, the size of the 10-year data set was reduced by a factor of 28 from (34920 x 365) to (34920 x 13), which significantly improved the computation times of clustering. Hierarchical clustering using Ward's method was performed on this reduced data set, yielding 10 cluster labels per grid point – one for each year.

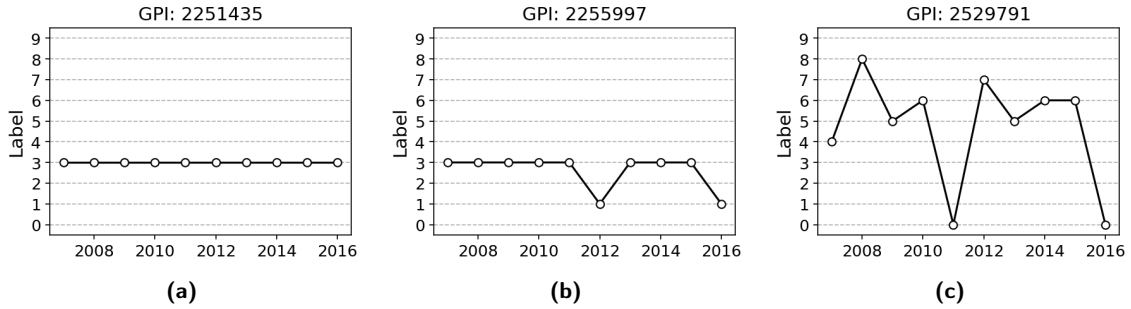


Fig. 4.19: Three examples of grid points and their assigned labels between 2007 and 2017. Fig. 4.19a shows a very stable or *robust* grid point that is assigned to the same cluster throughout the 10 year observation period. Fig. 4.19b shows a neutral grid point that is generally assigned to one cluster for most of the observation period, but may be assigned to different clusters in some years. Finally, Fig. 4.19c shows an unstable grid point that is assigned to many different clusters throughout the years.

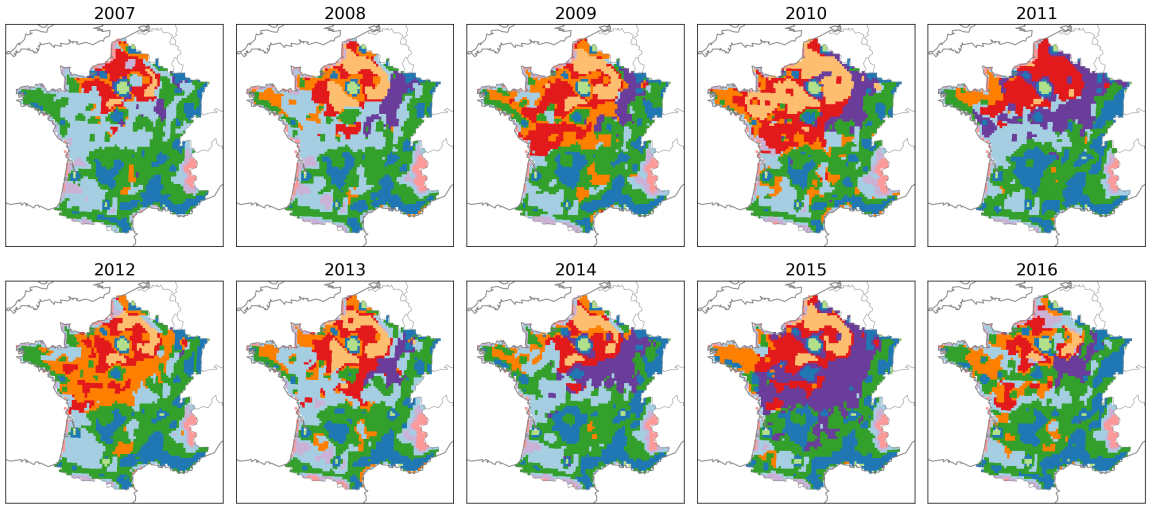


Fig. 4.20: Obtained clusters (i.e. scattering surfaces) for each of the individual years of σ' data. Some areas (e.g. Paris, the Alps) are assigned to the same cluster over the entire observation period, indicating that the scattering behavior in these areas is relatively stable and predictable. On the other hand, some areas are assigned to different clusters during the observation period, indicating that the scattering behavior in these areas can change significantly over time.

Depending on the interannual variability of the observed σ' signal a certain grid point can be very robust or very unstable. If a grid point has σ' observations with negligible interannual variability, it will likely be assigned the same cluster label every year. For such stable grid points, there is a high certainty that they belong to the cluster they have been assigned to – i.e. clustering is regarded as robust. An example of a robust grid point is shown in Fig. 4.19a, which has a constant label between 2007 and 2017. On the other hand, if a grid point has σ' observations with very high interannual variability, the grid point may be assigned to different a different cluster every year. An example of a very unstable grid point is shown in Fig. 4.19c, which shows how its cluster label wavers between six different values between 2007 and 2017. It is impossible to classify this grid point in terms of its scattering characteristics with a high degree of certainty, as its scattering behavior could be different every year. Between very stable grid points and very unstable grid points are grid points that are generally assigned to one cluster, but may be assigned to another cluster occasionally. Fig 4.19b shows an example of such a grid point, which generally belongs to cluster 3 but is assigned to cluster 1 in 2012 and 2016.

The obtained cluster labels are mapped in Fig. 4.20, which visualizes how the clusters change over time. Several shapes that were discussed in section 4.4 can be recognized throughout the years – such as Paris, the agricultural area near Paris, and the Alps – suggesting that some areas have a relatively distinct σ' signal and are relatively consistent over time in terms of their assigned cluster labels. Hence, clustering on an annual basis may help with identifying areas that have stable scattering characteristics.

Another interesting area is the Landes evergreen coniferous forest south of Bordeaux. This area was located directly in the storm trajectory of Cyclone Klaus (Fig. 4.21) which made landfall near Bordeaux on 24 January 2009 with wind speeds over 200 km/h. The before and after photos shown in Fig. 4.22 give an idea of the destruction that occurred – in total, over 220 000 ha of forest was either flattened or badly damaged. Clearly, land cover was significantly altered as a result of the storm. This is also seems to be reflected by the annual cluster labels and σ' signatures, see Fig. 4.23; grid points in the Landes area are assigned to a different cluster in 2009 and 2010 compared to other years, which could be explained by the destruction caused by the storm and the subsequent reforestation efforts and (partial) recovery of the coniferous forest. As a consequence of these events, this area is difficult to classify as coniferous forest even though this area should generally contain a significant amount of trees (see Fig. D.1.3). On the other hand, the results correspond to the events that occurred, which suggests that clustering on an annual basis based on σ' (and perhaps σ°_{40} and σ'') may be useful for identifying disturbances and trends such as storm damage, urban expansion, and conversion of forest/grassland to agriculture.

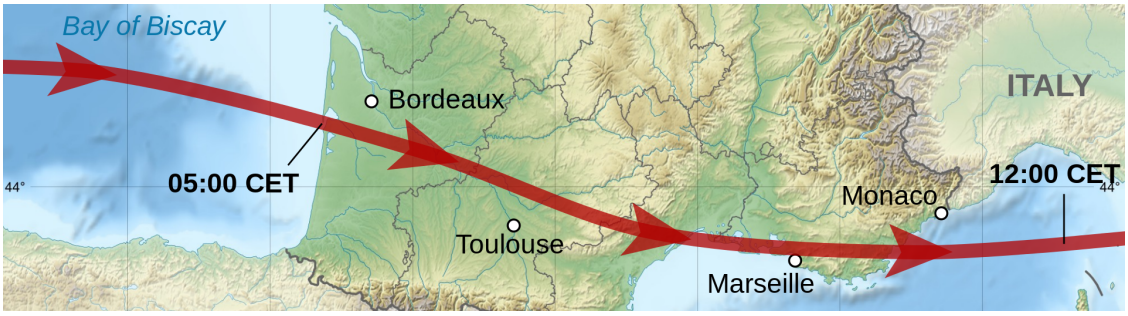


Fig. 4.21: Storm trajectory of cyclone Klaus. The storm made landfall at the Landes forest with wind speeds over 200 km/h.

(a) Before **(b)** After

Fig. 4.22: Photos of the Landes forest before and after cyclone Klaus. The photos show the degree to which the land cover in some areas was altered due to the storm. Even though not the entire area was completely flattened, such radical alterations to land cover would clearly result in different backscattering behavior and altered seasonal behavior of the vegetation parameters.

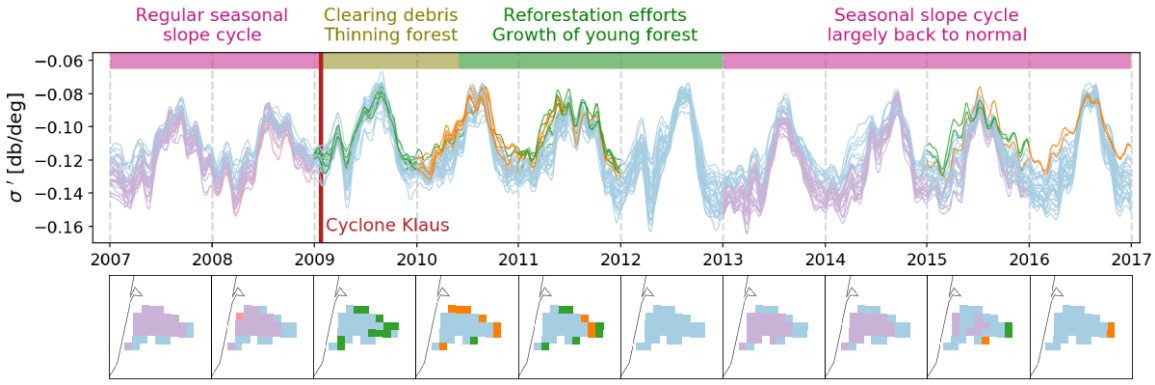


Fig. 4.23: Annual cluster labels and σ' observations of the Landes coniferous forest. A disturbance corresponding to cyclone Klaus is visible in early 2009, after which markedly different cluster labels and seasonal σ' behavior are observed. From 2013 onward, the cluster labels and seasonal σ' cycle seem to have mostly returned to pre-Klaus values, indicating a recovery of the Landes forest.

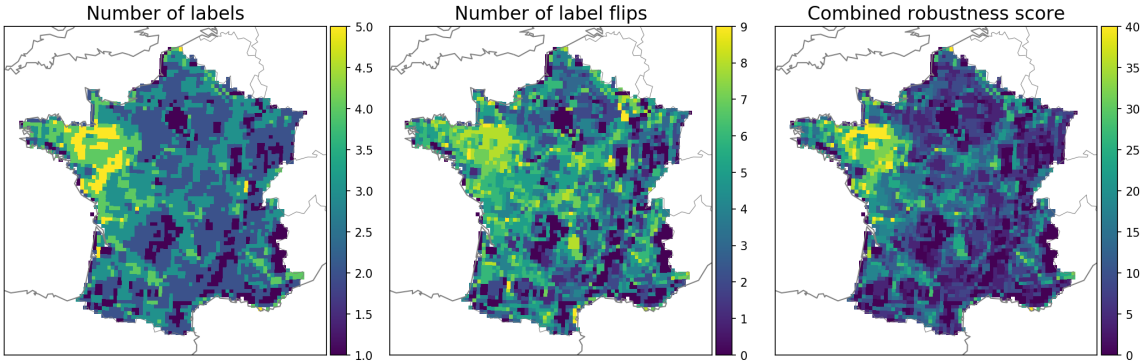


Fig. 4.24: Maps of number of unique labels, number of label flips, and robustness scores per grid point.

4.5.2 Calculating robustness scores

As described in section 3.6.2, the robustness score of each grid point is calculated by multiplying the number of unique labels with the number of times the label of a grid point changed in the subsequent year. The number of unique labels, the number of label flips and the robustness score are mapped in Fig. 4.24.

Number of unique labels

The map describing the number of unique labels clearly shows that cluster label variations are not randomly distributed in space – instead, clumped patterns are visible and several distinct areas can be identified. For example, the labels of Paris, the Alps, the Côte d'Azur, and parts of the Massif Central are constant between 2007 and 2017, which indicates that their σ' signals are relatively similar for all years in the observation period. On the other hand, grid points located in north-west and central France are generally assigned between three and six different labels between 2007 and 2017, indicating that the observed σ' signals of these areas differ significantly during the observation period.

Number of label flips

Even though the map describing the number of label flips is less defined, spatial patterns are visible here as well. Paris, the Alps, the Côte d'Azur, and parts of the Massif Central can identified easily – this is to be expected, as a completely constant cluster label inherently means that no label changes occur, and having many different labels inherently means that many label changes must occur. However, this map mainly contains additional information about grid points with between two and four labels, i.e. it provides a better distinction between grid points that could have the same number of unique labels, but are different in how often they flip between these labels.

Robustness score

Finally, the number of unique labels and the number of label flips are combined to produce a map of robustness scores. The worst robustness scores are obtained in grid points that alternate often between many cluster labels; these grid points likely have very variable σ' observations, which makes it difficult to classify them consistently into one cluster. Interestingly, the south of France is generally robust and easy to classify, while the least robust grid points are concentrated in central and north-west France.

One possible explanation for the poor robustness in central and north-west France is that rotational agriculture is practiced in this region; as can be seen in Fig. A.1, the land cover footprint in the central region of France generally contains significant amounts of annual summer crops, annual winter crops, and intensive grasslands. However, this particular land cover data set is only fully representative for one year. If this region indeed rotates between summer crops, winter crops, intensive grasslands and fallow fields, the backscatter observations – and hence, σ' as well as clusters based on σ' – can change significantly between years. This explains why this region is relatively hard to classify using clustering based on σ'.

As can be seen in Fig. 4.25, the same general region of poor robustness scores is obtained when isolating all grid points that have relatively large fractions of annual summer crops, annual winter crops and intensive grasslands (all $p \geq 10\%$). The area with the poorest robustness scores has a mosaic land cover that is clearly dominated by annual summer crops, annual winter crops, and intensive grasslands, see Fig. 4.26. The potential link between this specific land cover footprint and poor robustness is corroborated by Fig. 4.27, which shows that grid points generally become less robust (i.e. harder to accurately classify) when the combined percentage of annual summer crops, annual winter crops and intensive grasslands increases. This corresponds with the idea that poor robustness may be found in areas where land cover changes significantly between years, such as in agricultural areas where a crop rotation system is applied. This highlights the potential usefulness of clustering based on σ' in combination with the introduced robustness metric for the purpose of identifying areas with significant interannual land use change.

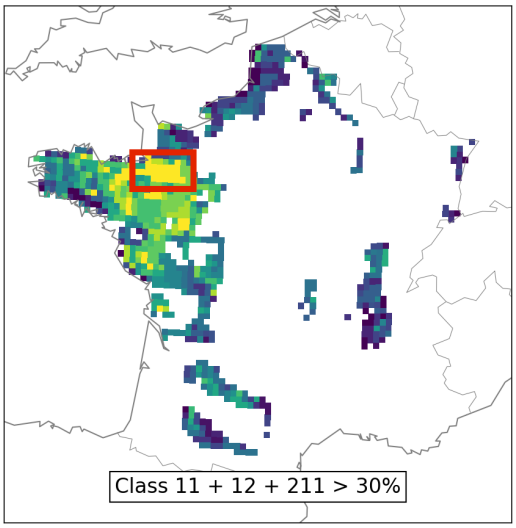


Fig. 4.25: Grid points where each of the fractions of annual summer crops, annual winter crops and intensive grasslands are at least 10%. A sample of the original 2016 Theia land cover data set is taken from the area within the red rectangle and visualized in Fig. 4.26.

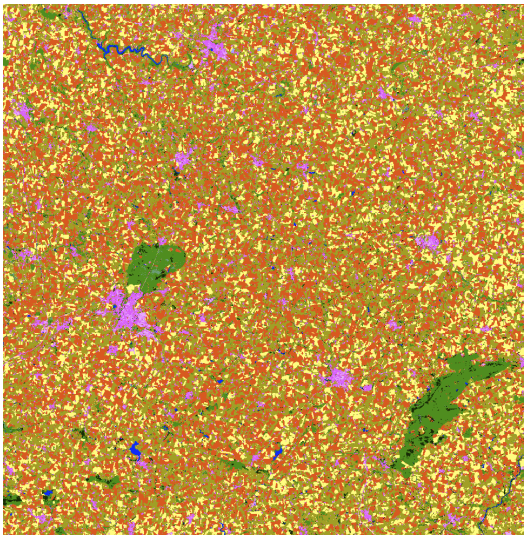


Fig. 4.26: Mapped sample of the original 2016 Theia data set showing the mosaic landscape of north-west France in the original 10 m resolution. This area is dominated by annual summer crops, annual winter crops, and intensive grasslands, which could explain the poor robustness scores.

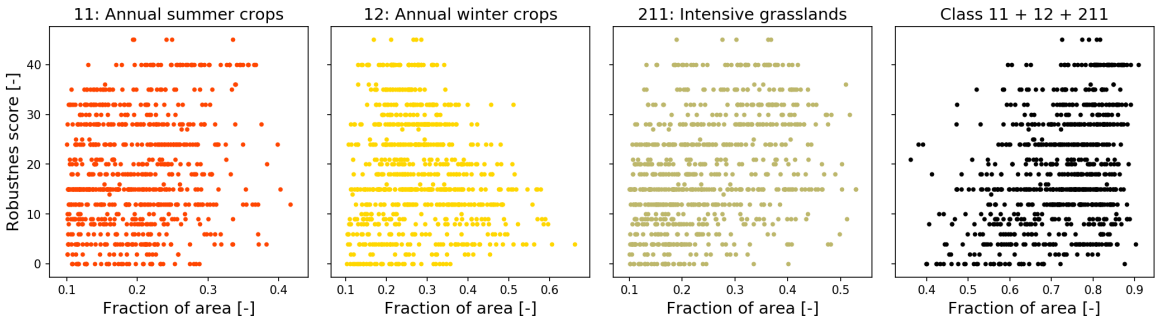


Fig. 4.27: From left to right: Scatter plots of (1) robustness score vs fraction of annual summer crops, (2) robustness score vs fraction of annual winter crops, (3) robustness score vs fraction of intensive grasslands, and (4) robustness score vs the combined fraction of annual summer crops, annual winter crops and intensive grasslands.

5 Conclusions and Recommendations

In this study, the temporal and spatial characteristics of the backscatter coefficient (σ°) and the TUW SMR vegetation parameters (σ' and σ'') were investigated using an unsupervised classification approach in order to gain an improved understanding of their physical meaning and behavior, as well as to further explore their potential value as a source of information. This chapter serves to summarize the main conclusions based on the obtained results, to answer the research questions defined in chapter 1, to describe any assumptions and limitations of this study and to provide recommendations for future research. The conclusions are structured according to the sub-questions; each of the sub-questions are answered first, after which the following main research question will be answered:

Can distinct scattering surfaces be identified and used to obtain an improved understanding of the observed σ°_{40}, σ', and σ'' behavior?

5.1 Conclusions

Question 1: *Which data preprocessing and clustering techniques are required and suited to solve this problem?*

Two clustering algorithms were compared in this research: k-means clustering and hierarchical clustering. The output obtained by hierarchical clustering is significantly more descriptive compared to k-means: hierarchical clustering produces a nested cluster hierarchy which provides information about how clusters merge for different values of k, while k-means simply returns a set of k clusters. Furthermore, assuming the input data is the same, hierarchical clustering will always return the same result while some randomness is inherently present in k-means clustering, which may produce different results for different runs.

In terms of computation time, k-means is faster than hierarchical clustering. However, the use of principal component analysis allows for significant size reduction of the input data set – down to about 2% of the size of the original data set with limited loss of information – thereby improving the computation time of hierarchical clustering to about the same order of magnitude as k-means clustering.

The term 'garbage in, garbage out' is particularly true for machine learning algorithms; in order to obtain meaningful results, the input data must be properly preprocessed. Any missing values in the input data set should be removed or interpolated, as most clustering algorithms cannot deal with this. Then, the data should be standardized before the principal component analysis is performed to ensure that the correct principal components are found. Care should also be taken when determining how many components to retain, as keeping too few components will result in significant loss of information and a possibly incorrect classification, while keeping too many components will lead to longer computation times. Luckily, many methods and guidelines exist for determining how many components to retain.

In order to solve the main research problem, a combination of PCA and agglomerative hierarchical clustering was chosen as clustering approach due to its informative output, internal consistency, deterministic nature, and acceptable computation time.

Question 2: *Can distinct and meaningful scattering surfaces be identified using unsupervised classification?*

A set of 10 clusters is generated based on the climatology of σ' using agglomerative hierarchical clustering and Ward linkage. Instead of a noisy field, clear spatial patterns are visible when mapping the generated clusters. Additionally, land cover features can be identified in the mapped clusters; areas such as Paris, the Alps, the Landes forest, and the intense agricultural around Paris – which are all clearly visible in land cover maps – are well represented by the cluster maps. Areas that are expected to differ significantly in terms of scattering characteristics, such as Paris and the Landes forest, are generally not clustered together.

Since clustering was performed on σ', it is unsurprising that the generated clusters differ in terms of σ' and that grid points within a cluster are relatively similar in terms of σ'. However, the cluster analysis shows that σ°_{40} and σ'' also seem to separate, i.e. grid points that are similar in terms of σ' are generally (but not necessarily!) also similar in terms of σ°_{40} and σ''. Moreover, grid points with similar scattering characteristics are generally near each other, forming contiguous areas that resemble land cover features.

In general, each of the clusters is contiguous, confined to a specific region of France, and has a characteristic set of seasonal σ°_{40}, σ', and σ'' signatures. As such, the obtained clusters represent distinct and meaningful scattering surfaces, each of which having different scattering characteristics.

Question 3: *What is the influence of sub-footprint land cover heterogeneity on the σ°_{40}, σ', and σ'' signatures?*

The results show that grid points that have very similar seasonal σ°_{40}, σ', and σ'' behavior are generally located close to each other and also tend to have very similar land cover footprints. However, the inverse is not necessarily true; slight differences in land cover footprint can lead to significantly different backscatter characteristics. For example, due to the large reflectivity of urban areas a relatively small presence of urban land cover within a grid point will result in significantly larger backscatter compared to a grid point without urban land cover, even if their land cover footprints are nearly identical.

Each land cover class has specific backscatter characteristics and generates specific seasonal σ°_{40}, σ', and σ'' patterns. The backscatter signatures observed in a 25 km^2 ASCAT grid point are therefore generated by the combined effect of all present land cover classes, and thus depend on the relative presence of the land cover classes in each 25 km^2 grid point. The cluster analysis showed that seasonal backscatter behavior becomes less clear and patterns become harder to characterize as the land cover footprint becomes more heterogeneous. Conversely, areas with a land cover footprint that is dominated by one or a few land cover classes show clear backscatter patterns.

In areas that have a relatively 'pure' land cover footprint, the σ°_{40}, σ', and σ'' observations can be (partially) explained by the physical seasonal processes occurring in the land cover types that are present. Overall, σ' corresponds with the seasonal cycle of deciduous vegetation, with minimum vegetation density during winter and maximum vegetation density during summer. On the other hand, σ'' is generally highest during winter and lowest during summer, which may be due to a change in dominant scattering mechanisms following structural seasonal processes such as of leaf drop (in autumn/winter) and canopy growth (in spring/summer). In areas with significant agriculture, σ'' has a 'double-bounce' pattern, which was found to correspond with the different distinct growth stages of cereal crops.

The relationship between land cover footprint and seasonal backscatter behavior is very complex, but the following conclusions can be made: (1) each land cover type has different characteristic seasonal backscatter signatures, (2) a homogeneous land cover footprint results in clear backscatter signatures and clearly identifiable behavior, (3) a heterogeneous land cover footprint results in relatively noisy and unclear signatures that mainly show a 'general' seasonal cycle, and (4) the results are consistent with the interpretation of σ' as a measure for vegetation density and the interpretation of σ'' as a measure for the relative dominance of ground-bounce scattering over direct scattering from vegetation constituents.

Question 4: *Are the grid points and generated clusters "robust"?*

A measure for robustness was introduced in order to quantify the temporal stability of multi-year clusters based on 10 years of σ' for each grid point. The robustness score takes into account the number of unique labels assigned to each grid point as well as how often the label of each grid point changed. A high score indicates an unstable grid point which is assigned to many different clusters over the years, likely due to large interannual variability in the σ' observations. On the other hand, a low score indicates a stable grid point that is always assigned to the same cluster, due to a relatively stable seasonal σ' signal. As such, the robustness score gives an indication of the certainty that a grid point belongs to a specific cluster.

The results show that the robustness score is not randomly distributed; instead, spatial patterns become visible when mapping the robustness score. Areas that can be expected not to change significantly between years in terms of land cover (e.g. Paris, the Alps) are generally robust and always assigned to the same cluster. On the other hand, the poor robustness in west France may be caused by agriculture with a crop rotation schedule, which would lead to large interannual variability of σ' and hence, many different cluster labels over the year. Clearly, it is difficult to decisively classify this area, since the backscatter characteristics seem to be different every year. The results show that multi-year clustering combined with a robustness score can be useful in identifying areas with stable and unstable backscatter characteristics.

Finally, an unexpected but interesting observation in the multi-year clustering is the effect of cyclone Klaus in early 2009. The destruction and subsequent recovery of the Landes forest south of Bordeaux is represented in the multi-year clusters, suggesting that multi-year clustering based on σ' can potentially be a valuable tool for identifying land cover disturbances (e.g. storm damage) or land use change (e.g. urban expansion, conversion of forest to agriculture).

Can distinct scattering surfaces be identified and used to obtain an improved understanding of the observed σ°_{40}, σ', and σ'' behavior?

Based on the sub-questions, the main conclusion of this research is that it is indeed possible to identify distinct and meaningful scattering surfaces based on the climatology of σ', which in turn provide interesting insights into the seasonal and interannual behavior of the backscatter coefficient (σ°) and the TUW SMR vegetation parameters (σ', σ''). These scattering surfaces provide insight into how scattering characteristics vary across the heterogeneous land cover of France. The results presented in this study are consistent with the current understanding of σ' and σ'' as measures for vegetation density and the relative dominance of ground-bounce scattering over direct scattering from vertical vegetation constituents. Additionally, the results show that in some areas – such as agricultural areas with crop rotation – significant interannual variation is present in the dynamically estimated vegetation parameters. This indicates that the vegetation correction of the TUW SMR can be improved significantly by using the dynamically estimated vegetation parameters instead of their multi-year climatology, resulting in a better ASCAT soil moisture data product. Moreover, estimation of the vegetation parameters on a daily basis allows for investigating how scattering surfaces change over time, for example due to land use change or storm damage. This highlights the potential value of the TUW SMR vegetation parameters as a new source of information, especially considering the fact that the MetOp satellites measure backscatter globally on a daily basis and considering that the MetOp mission has a large archive of global backscatter data as well as a promising future with the upcoming MetOp-SG satellite series.

5.2 Recommendations

This study had some assumptions and limitations, which yielded a number of recommendations for future research. The assumptions, limitations, and recommendations are discussed in this section.

Dynamically estimated vegetation parameters

As discussed by Melzer [42], several settings affect the behavior of the dynamically estimated vegetation parameters, e.g. choice of algorithm and kernel window size. It is likely that different settings would change the vegetation parameters, and hence, the clusters presented in this research. In this research it is assumed that the algorithm and settings chosen to dynamically estimate the vegetation parameters are correct, even though different settings may yield better results for some grid points.

Land cover data

In this study, a 2016 land cover data set of France developed by Theia was used. While this data set provided interesting insights, it should be noted that this data set was generated specifically for the year of 2016, while the σ°_{40}, σ', and σ'' data was observed between 2007 – 2017. Therefore, for the purpose of this study the 2016 land cover data set is assumed to be representative for the 2007 – 2017 period. However, it is recommended that future research uses a multi-year land cover data set to ensure that any potential relation between σ°, σ', and σ'' is correctly investigated.

Use of PCA

To improve computation times, PCA is used in this study for dimensionality reduction; this was particularly useful for the 10-year data set, but not so much for the climatology. It is assumed that sufficient information is retained when selecting the number of principal components so that at least 99% of the variance present in the original data set is explained. However, information is inherently lost when choosing to retain a select number of principal components. If processing power is not a limiting factor, better results may be obtained by performing hierarchical clustering on the original σ' data. Additionally, using the original data is simply more straightforward; understanding the inner workings of PCA and the meaning of its outputs requires a significant amount of time and further complicates the interpretation of the results, while microwave scatterometry is a difficult subject in itself.

Clustering algorithms

While two clustering algorithms are compared in this study, there exist many algorithms for the purpose of unsupervised classification. Moreover, each algorithm typically has its own list of settings that can be varied, e.g. linkage criteria, choosing the correct number of clusters, distance metrics, and so forth. However, this research was not aimed at testing all algorithms and combinations of settings – even though other algorithms and/or settings may be better suited for clustering time series data. It was assumed that hierarchical clustering was sufficiently suitable for the purpose of this research, but it is recommended that future studies explore potentially better clustering methods. For example, search for clustering techniques that are able to handle objects leaving and entering over time, as well as recognizing events like shrinking, growing, splitting, merging, dissolving and forming of clusters.

Robustness

The robustness score proposed in this study was useful, as it shows that some areas are more robust than others. However, it is also a very straightforward metric, that perhaps does not take all relevant factors into account. It would be interesting to define robustness in a more comprehensive way and to see if the results presented in this research hold true.

5.3 Implications of this study

Previous research by Steele-Dunne et al. [59] identified significant contiguous variations in the seasonal cycle of σ' and σ'' over the North-American grasslands, and also found that a clear link exists between σ'' and structural changes in grassy vegetation over the seasonal vegetation growth cycle. This study investigated the TUW SMR vegetation parameters over Metropolitan France in order to further explore the value of σ' and σ'' as a source of information, particularly in land cover types other than grasslands.

The results presented in this study clearly support the idea that there exists a relationship between σ'' and (vegetation) land cover composition. It was found that seasonal variations in σ'' seem to be driven mainly by the vegetation growth cycle and seasonal changes in vegetation structure. As such, the results support the idea that σ'' contains information about the relative dominance of direct scattering over ground-bounce scattering, also in land cover types other than grasslands.

Interestingly, the devastating effects of cyclone Klaus over the Landes forest were clearly visible in the dynamically estimated σ'. While the effects of the storm on σ'' were not investigated here, this does highlight the potential value of the TUW SMR vegetation parameters for vegetation monitoring and land use change monitoring. However, the value of σ' and σ'' for these purposes must be further investigated.

One of the more impactful findings of this study is that the interannual variability of the dynamically estimated vegetation parameters can vary significantly depending on the land cover composition; certain areas such as Paris and the Alps show relatively predictable seasonal behavior of σ' and σ'' that can be easily classified, while the interannual variability in σ' and σ'' is much larger in areas with a more heterogeneous land cover footprint and in areas containing both grassland and cropland. This suggests that the current implementation of the TUW SMR vegetation correction – which is based on multi-year averages of σ' and σ'' – is likely over- and/or underestimating the influence of vegetation on backscatter. This is especially true in areas where the interannual variability of σ' and σ'' is large, which negatively affects the accuracy of the TU Wien soil moisture retrieval approach and results in lower quality of the resulting soil moisture data products in these areas. By correcting for vegetation dynamically (i.e. using the dynamically estimated vegetation parameters) instead of performing the vegetation correction using a climatology of σ' and σ'', the TU Wien soil moisture retrieval approach can be improved significantly. In turn, this will yield higher quality ASCAT-derived soil moisture products.

Bibliography

[1] C. Albergel, P. de Rosnay, C. Gruhier, J. Muñoz-Sabater, S. Hasenauer, L. Isaksen, Y. Kerr, and W. Wagner. Evaluation of remotely sensed and modelled soil moisture products using global ground-based in situ observations. *Remote Sens. Environ.*, 118: 215–226, 2012. doi: 10.1016/j.rse.2011.11.017.

[2] A. Anyamba and C. J. Tucker. Historical perspectives on AVHRR NDVI and vegetation drought monitoring. *Remote Sens. Drought Innov. Monit. Approaches*, pages 23–49, 2012. ISSN 0002-9165. doi: 10.1201/b11863.

[3] S. Baronti, E. Del Frate, P. Ferrazzoli, S. Paloscia, P. Pampaloni, and G. Schiavon. SAR polarimetric features of agricultural areas. *Int. J. Remote Sens.*, 16(14):2639–2656, 1995. ISSN 13665901. doi: 10.1080/01431169508954581.

[4] Z. Bartalis, K. Scipal, and W. Wagner. Azimuthal anisotropy of scatterometer measurements over land. *IEEE Trans. Geosci. Remote Sens.*, 44(8):2083–2092, 2006. ISSN 01962892. doi: 10.1109/TGRS.2006.872084.

[5] Z. Bartalis, W. Wagner, V. Naeimi, S. Hasenauer, K. Scipal, H. Bonekamp, J. Figa, and C. Anderson. Initial soil moisture retrievals from the METOP-A Advanced Scatterometer (ASCAT). *Geophys. Res. Lett.*, 34(20):5–9, 2007. ISSN 00948276. doi: 10.1029/2007GL031088.

[6] A. Bosc. EMILION, a tree functional-structural model: Presentation and first application to the analysis of branch carbon balance. *Ann. For. Sci.*, 57(5-6):555–569, 2000. ISSN 12864560. doi: 10.1051/forest:2000142.

[7] T. W. Brakke, E. T. Kanemasu, J. L. Steiner, F. T. Ulaby, and E. Wilson. Microwave radar response to canopy moisture, leaf-area index, and dry weight of wheat, corn, and sorghum. *Remote Sens. Environ.*, 11(C):207–220, 1981. ISSN 00344257. doi: 10.1016/0034-4257(81)90020-1.

[8] L. Brocca, F. Melone, T. Moramarco, W. Wagner, V. Naeimi, Z. Bartalis, and S. Hasenauer. Improving runoff prediction through the assimilation of the ASCAT soil moisture product. *Hydrol. Earth Syst. Sci.*, 14(10):1881–1893, 2010. doi: 10.5194/hess-14-1881-2010.

[9] L. Brocca, S. Hasenauer, T. Lacava, F. Melone, T. Moramarco, W. Wagner, W. Dorigo, P. Matgen, J. Martínez-Fernández, P. Llorens, J. Latron, C. Martin, and M. Bittelli. Soil moisture estimation through ASCAT and AMSR-E sensors: An intercomparison and validation study across Europe. *Remote Sens. Environ.*, 115(12):3390–3408, 2011. doi: 10.1016/j.rse.2011.08.003.

[10] S. C. Brown, S. Quegan, K. Morrison, J. C. Bennett, and G. Cookmartin. High-resolution measurements of scattering in wheat canopies - Implications for crop parameter retrieval. In *IEEE Trans. Geosci. Remote Sens.*, volume 41, pages 1602–1610, 2003. doi: 10.1109/TGRS.2003.814132.

[11] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. (1974) Communications in Statistics Volume 3, Issue 1, 1974. *Commun. Stat.*, 3(1):1974, 1974. ISSN 00903272.

[12] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-1(2):224–227, 1979. ISSN 01628828. doi: 10.1109/TPAMI.1979.4766909.

[13] P. de Rosnay, G. Balsamo, C. Albergel, J. Muñoz-Sabater, and L. Isaksen. Initialisation of Land Surface Variables for Numerical Weather Prediction. *Surv. Geophys.*, 35(3):607–621, 2012. ISSN 01693298. doi: 10.1007/s10712-012-9207-x.

[14] M. C. Dobson, F. T. Ulaby, M. T. Hallikainen, and M. A. El-Rayes. Microwave Dielectric Behavior of Wet Soil-Part II: Dielectric Mixing Models. *IEEE Trans. Geosci. Remote Sens.*, GE-23(1):35–46, 1985. ISSN 15580644. doi: 10.1109/TGRS.1985.289498.

[15] W. A. Dorigo, A. Gruber, R. A. De Jeu, W. Wagner, T. Stacke, A. Loew, C. Albergel, L. Brocca, D. Chung, R. M. Parinussa, and R. Kidd. Evaluation of the ESA CCI soil moisture product using ground-based observations. *Remote Sens. Environ.*, 162: 380–395, 2015. doi: 10.1016/j.rse.2014.07.023.

[16] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, Inc., New York, 2 edition, 2001. ISBN 0471056693.

[17] C. Elachi and J. van Zyl. *Introduction to the Physics and Techniques of Remote Sensing: Second Edition*. John Wiley & Sons, Inc., 2006. ISBN 9780471475699. doi: 10.1002/0471783390.

[18] Eurostat. Crop production in eu standard humidity, 2019. Data retrieved from Eurostat, https://ec.europa.eu/eurostat/data/database.

[19] C. Fatras, F. Frappart, E. Mougin, P. L. Frison, G. Faye, P. Borderies, and L. Jarlan. Spaceborne altimetry and scatterometry backscattering signatures at C- and Ku-bands over West Africa. *Remote Sens. Environ.*, 159:117–133, 2015. ISSN 00344257. doi: 10.1016/j.rse.2014.12.005. URL http://dx.doi.org/10.1016/j.rse.2014.12.005.

[20] J. Figa-Saldaña, J. J. Wilson, E. Attema, R. Gelsthorpe, M. R. Drinkwater, and A. Stoffelen. The advanced scatterometer (ASCAT) on the meteorological operational (MetOp) platform: A follow on for European wind scatterometers. *Can. J. Remote Sens.*, 28(3):404–412, jan 2002. ISSN 0703-8992. doi: 10.5589/m02-035.

[21] P. L. Frison, E. Mougin, and P. Hiernaux. Observations and Interpretation of Seasonal ERS-1 Wind Scatterometer Data over Northern Sahel (Mali). *Remote Sens. Environ.*, 63(3):233–242, 1998. ISSN 00344257. doi: 10.1016/S0034-4257(97)00137-5.

[22] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM, 2007. doi: 10.1137/1.9780898718348.

[23] N. Gyllenstrand, D. Clapham, T. Källman, and U. Lagercrantz. A Norway spruce FLOWERING LOCUS T homolog is implicated in control of growth rhythm in conifers. *Plant Physiol.*, 144(1):248–257, 2007. ISSN 00320889. doi: 10.1104/pp.107.095802.

[24] S. Hahn, C. Reimer, M. Vreugdenhil, T. Melzer, and W. Wagner. Dynamic Characterization of the Incidence Angle Dependence of Backscatter Using Metop ASCAT. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 10(5):2348–2359, 2017. ISSN 21511535. doi: 10.1109/JSTARS.2016.2628523.

[25] Z. Hao and A. AghaKouchak. A Nonparametric Multivariate Multi-Index Drought Monitoring Framework. *J. Hydrometeorol.*, 15(1):89–101, 2014. doi: 10.1175/JHM-D-12-0160.1.

[26] E. Heidorn, K. Utvik, C. Gengler, K. Alati, D. Collet, V. Attivissimo, and M. Colantonio. *Agriculture, forestry and fishery statistics: 2017 edition*. Publications Office of the European Union, 2017. ISBN 9789279757655. doi: 10.2785/570022.

[27] H. Hotelling. Analysis of a complex of statistical variables into principal components (continued from September issue). *J. Educ. Psychol. Oct.*, 1933.

[28] A. R. Huete. Vegetation Indices, Remote Sensing and Forest Monitoring. *Geogr. Compass*, 6(9):513–532, 2012. doi: 10.1111/j.1749-8198.2012.00507.x.

[29] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes. Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sens.*, 9(1):95, 2017. doi: 10.3390/rs9010095.

[30] L. Jarlan, E. Mougin, P. L. Frison, P. Mazzega, and P. Hiernaux. Analysis of ERS wind scatterometer time series over Sahel (Mali). *Remote Sens. Environ.*, 81(2-3):404–415, aug 2002. ISSN 00344257. doi: 10.1016/S0034-4257(02)00015-9. URL http://linkinghub.elsevier.com/retrieve/pii/S0034425702000159.

[31] L. Jarlan, P. Mazzega, E. Mougin, F. Lavenu, G. Marty, P. L. Frison, and P. Hiernaux. Mapping of Sahelian vegetation parameters from ERS scatterometer data with an evolution strategies algorithm. *Remote Sens. Environ.*, 87(1):72–84, 2003. ISSN 00344257. doi: 10.1016/S0034-4257(03)00164-0.

[32] I. T. Jolliffe. *Principal Component Analysis. Second Edition*. Springer, 2002. ISBN 0-387-95442-2. doi: 10.1007/b98835.

[33] Y. H. Kerr. Soil moisture from space: Where are we? *Hydrogeol. J.*, 2007. ISSN 14312174. doi: 10.1007/s10040-006-0095-3.

[34] R. D. Koster, S. P. P. Mahanama, T. J. Yamada, G. Balsamo, A. A. Berg, M. Boisserie, P. A. Dirmeyer, F. J. Doblas-Reyes, G. Drewitt, C. T. Gordon, Z. Guo, J. H. Jeong, D. M. Lawrence, W. S. Lee, Z. Li, L. Luo, S. Malyshev, W. J. Merryfield, S. I. Seneviratne, T. Stanelle, B. J. Van Den Hurk, F. Vitart, and E. F. Wood. Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophys. Res. Lett.*, 37(2):1–6, 2010. ISSN 19448007. doi: 10.1029/2009GL041677.

[35] R. D. Koster, G. K. Walker, S. P. P. Mahanama, and R. H. Reichle. Soil Moisture Initialization Error and Subgrid Variability of Precipitation in Seasonal Streamflow Forecasting. *J. Hydrometeorol.*, 15(1):69–88, 2014. doi: 10.1175/JHM-D-13-050.1.

[36] M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel. World Map of the Köppen-Geiger climate classification updated. *Meteorol. Zeitschrift*, 15(3):259–263, jul 2006. doi: 10.1127/0941-2948/2006/0130.

[37] C. Li, J. Yin, J. Zhao, G. Zhang, and X. Shan. The selection of artificial corner reflectors based on RCS analysis. *Acta Geophys.*, 60(1):43–58, 2012. ISSN 18956572. doi: 10.2478/s11600-011-0060-y.

[38] Q. Liu, R. H. Reichle, R. Bindlish, M. H. Cosh, W. T. Crow, R. de Jeu, G. J. M. De Lannoy, G. J. Huffman, and T. J. Jackson. The Contributions of Precipitation and Soil Moisture Observations to the Skill of Soil Moisture Estimates in a Land Data Assimilation System. *J. Hydrometeorol.*, 12(5):750–765, oct 2011. ISSN 1525-755X. doi: 10.1175/JHM-D-10-05000.1.

[39] J. Macqueen. Some methods for classification and analysis. In *Proc. Fifth Berkeley Symp. Math. Stat. Probab. Vol. 1 Stat.*, volume 233, pages 281–297, 1967. URL http://projecteuclid.org/bsmsp.

[40] J. Martínez-Fernández, A. González-Zamora, N. Sánchez, A. Gumuzzio, and C. M. Herrero-Jiménez. Satellite soil moisture for agricultural drought monitoring: Assessment of the SMOS derived Soil Water Deficit Index. *Remote Sens. Environ.*, 177: 277–286, 2016. ISSN 00344257. doi: 10.1016/j.rse.2016.02.064.

[41] L. L. Mcquitty. Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. *Educ. Psychol. Meas.*, 17(2):207–229, 1957. ISSN 15523888. doi: 10.1177/001316445701700204.

[42] T. Melzer. Vegetation Modelling in WARP 6.0. In *Proc. EUMETSAT Meteorol. Satell. Conf.*, pages 1–7, Vienna, Austria, 2013.

[43] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985. ISSN 00333123. doi: 10.1007/BF02294245.

[44] Ministère de l'Agriculture, de la Pêche, et de l'Alimentation. *L'agriculture, la forêt et les industries agroalimentaires: Agreste GraphAgri 2013*. GraphAgri, Paris, 2013. ISBN 2-11-090437-2.

[45] V. Naeimi, K. Scipal, Z. Bartalis, S. Hasenauer, and W. Wagner. An improved soil moisture retrieval algorithm for ERS and METOP scatterometer observations. *IEEE Trans. Geosci. Remote Sens.*, 47(7):1999–2013, 2009. ISSN 01962892. doi: 10.1109/TGRS.2008.2011617.

[46] K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.*, 1901. ISSN 19415982. doi: 10.1080/14786440109462720.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[48] W. Pedrycz. *Knowledge-Based Clustering: From Data to Information Granules*. John Wiley & Sons, Inc., 2005. ISBN 0471469661. doi: 10.1002/0471708607.

[49] M. C. Peel, B. L. Finlayson, and T. A. McMahon. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.*, 11(5):1633–1644, 2007. ISSN 16077938. doi: 10.5194/hess-11-1633-2007.

[50] G. Picard, T. Le Toan, and F. Mattia. Understanding C-band radar backscatter from wheat canopy using a multiple-scattering coherent model. In *IEEE Trans. Geosci. Remote Sens.*, volume 41, pages 1583–1591, 2003. doi: 10.1109/TGRS.2003.813353.

[51] M. Ray and Britannica Educational Publishing. France, 2014.

[52] W. G. Rees. *Physical Principles of Remote Sensing (Topics in Remote Sensing)*. Cambridge University Press, 2001. ISBN 0521660343.

[53] C. Reimer. *Calibration of Space-borne Scatterometers: Towards a Consistent Climate Data Record for Soil Moisture Retrieval*. Msc. thesis, Vienna University of Technology-http://repositum.tuwien.ac.at/obvutwhs/download/pdf/1634734?originalFilename=true, 2014. URL http://repositum.tuwien.ac.at/obvutwhs/download/pdf/1634734.

[54] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20 (C):53–65, 1987. ISSN 03770427. doi: 10.1016/0377-0427(87)90125-7.

[55] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, 2004. ISBN 076952236X. doi: 10.1109/ICTAI.2004.50.

[56] W. S. Sarle, A. K. Jain, and R. C. Dubes. Algorithms for Clustering Data. *Technometrics*, 32(2):227, 1990. ISSN 00401706. doi: 10.2307/1268876.

[57] Y. Shao, H. Guo, Q. Hu, Y. Lu, Q. Dong, and C. Han. Effect of Dielectric Properties of Moist Salinized Soils on Backscattering Coefficients Extracted from RADARSAT Image. In *Int. Geosci. Remote Sens. Symp.*, volume 4, pages 2789–2791, 2003.

[58] P. H. Sneath. The application of computers to taxonomy. *J. Gen. Microbiol.*, 17(1):201–226, 1957. ISSN 00221287. doi: 10.1099/00221287-17-1-201.

[59] S. C. Steele-Dunne, S. Hahn, W. Wagner, and M. Vreugdenhil. Investigating vegetation water dynamics and drought using Metop ASCAT over the North American Grasslands. *Remote Sens. Environ.*, 224:219–235, 2019. doi: 10.1016/j.rse.2019.01.004.

[60] M. Svoboda, D. Le Comte, M. Hayes, R. Heim, K. Gleason, J. Angel, B. Rippey, R. Tinker, M. Palecki, D. Stooksbury, D. Miskus, and S. Stephens. The Drought Monitor. *Bull. Am. Meteorol. Soc.*, 83(8):1181–1190, 2002. doi: 10.1175/1520-0477(2002)083<1181:TDM>2.3.CO;2.

[61] R. Sylvester-Bradley, P. Berry, J. Blake, D. Kindred, J. Spink, I. Bingham, J. McVittie, and J. Foulkes. *Wheat Growth Guide*. AHDB Cereal. Oilseeds (Agriculture Hortic. Dev. Board), 2015.

[62] E. Tebbs, F. Gerard, A. Petrie, and E. De Witte. *Emerging and Potential Future Applications of Satellite-Based Soil Moisture Products*. Elsevier Inc., 2016. ISBN 9780128033890. doi: 10.1016/B978-0-12-803388-3.00019-X.

[63] F. T. Ulaby and R. P. Jedlicka. Microwave Dielectric Properties of Plant Materials. *IEEE Trans. Geosci. Remote Sens.*, GE-22(4):406–415, 1984. ISSN 15580644. doi: 10.1109/TGRS.1984.350644.

[64] F. T. Ulaby and E. A. Wilson. Microwave Attenuation Properties of Vegetation Canopies. *IEEE Trans. Geosci. Remote Sens.*, GE-23:746–753, 1985. ISSN 15580644. doi: 10.1109/TGRS.1985.289393.

[65] F. T. Ulaby, R. K. Moore, and A. K. Fung. *Microwave Remote Sensing - Active and Passive - Volume I - Microwave Remote Sensing Fundamentals and Radiometry (v. 1)*. Artech House, 1981. ISBN 0201107597.

[66] F. T. Ulaby, R. K. Moore, and A. K. Fung. *Microwave Remote Sensing, Active and Passive: Vol II, Radar Remote Sensing and Surface Scattering and Emission Theory*. Artech House, 1982. ISBN 0201107600.

[67] F. T. Ulaby, R. K. Moore, and A. K. Fung. *Microwave Remote Sensing: Active and Passive, from Theory to Applications: 3 (Artech House Remote Sensing Library)*. Artech House, 1986. ISBN 0890061920.

[68] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv e-prints*, art. arXiv:1907.10121, Jul 2019.

[69] W. Wagner, G. Lemoine, M. Borgeaud, and H. Rott. A Study of Vegetation Cover Effects on ERS Scatterometer Data. *IEEE Trans. Geosci. Remote Sens.*, 37(2 II):938–948, mar 1999. doi: 10.1109/36.752212.

[70] W. Wagner, J. Noll, M. Borgeaud, and H. Rott. Monitoring soil moisture over the canadian prairies with the ERS scatterometer. *IEEE Trans. Geosci. Remote Sens.*, 37(1 PART 1):206–216, 1999. ISSN 01962892. doi: 10.1109/36.739155.

[71] W. Wagner, S. Hahn, R. Kidd, T. Melzer, Z. Bartalis, S. Hasenauer, J. Figa-Saldaña, P. De Rosnay, A. Jann, S. Schneider, J. Komma, G. Kubu, K. Brugger, C. Aubrecht, J. Züger, U. Gangkofner, S. Kienberger, L. Brocca, Y. Wang, G. Blöschl, J. Eitzinger, K. Steinnocher, P. Zeil, and F. Rubel. The ASCAT Soil Moisture Product: A Review of its Specifications, Validation Results, and Emerging Applications. *Meteorol. Zeitschrift*, 22(1):5–33, 2013. doi: 10.1127/0941-2948/2013/0399.

[72] N. Wanders, M. F. P. Bierkens, S. M. de Jong, A. de Roo, and D. Karssenberg. The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological models. *Water Resour. Res.*, 50(8):6874–6891, aug 2014. ISSN 00431397. doi: 10.1002/2013WR014639.

[73] N. Wanders, D. Karssenberg, A. De Roo, S. M. De Jong, and M. F. P. Bierkens. The suitability of remotely sensed soil moisture for improving operational flood forecasting. *Hydrol. Earth Syst. Sci.*, 18(6):2343–2357, 2014. doi: 10.5194/hess-18-2343-2014.

[74] J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.*, 1963. ISSN 1537274X. doi: 10.1080/01621459.1963.10500845.

[75] I. Woodhouse. *Introduction to microwave remote sensing*. Taylor & Francis, Boca Raton, 2006. ISBN 978-0415271233.

[76] R. Xu and D. Wunsch. *Clustering (IEEE Press Series on Computational Intelligence)*. John Wiley & Sons, Inc., 2008.

A Land cover

A.1 Theia France 2016 land cover classification

From Inglada et al. [29].

Fig. A.1: Theia land cover classification [29]

A.2 Theia France 2016 land cover classification: nomenclature

As described by Inglada et al. [29], several data sources were combined to build a land cover nomenclature for mainland France for 2016 consisting of different 17 land cover classes. It should be noted that the most recent land cover map of metropolitan France published by Theia for the year 2018 further refines the nomenclature from 17 to 23 land cover classes, mainly providing an improved distinction between different types of agricultural crops. For example, while the 2016 land cover data set only differentiates between annual summer crops, annual winter crops, orchards and vines, the 2018 land cover data set distinguishes between rapeseed, straw cereals, protein crops, soy, sunflower, corn, rice, and tubers/roots, in addition to orchards and vines. It is recommended that the 2018 land cover data set is used in future studies, as having more land cover classes may help provide an even better understanding of the σ°, σ', and σ'' behavior of different types of agricultural land cover. However, at the time of writing the nomenclature consisted of the following 17 land cover classes [29]:

Arable land

- **Class 11** Annual summer crops: annual crops which are seeded from March to mid June and harvested between mid August to mid September; mainly corn and sunflower.
- **Class 12** Annual winter crops: annual crops which are seeded between November and February and harvested between mid June and late July; mainly wheat, barley and rapeseed.
- **Class 12** Intensive grasslands: dense grass cover of floral composition.

Perennial crops

- **Class 221** Orchards: parcels planted with fruit trees or shrubs.
- **Class 222** Vineyards: areas planted with vines.

Forests

- **Class 31** Broad-leaved forest: areas that consist mainly of broad-leaved trees, including shrub and bush undergrowth. Broad-leaved trees represent at least 75% of the area, the minimum tree height is 5 m and crown cover density exceeds 30%.
- **Class 32** Coniferous forest: areas that consist mainly of coniferous trees, including shrub and bush undergrowth. Coniferous trees represent at least 75% of the area, the minimum tree height is 5 m and crown cover density exceeds 30%.

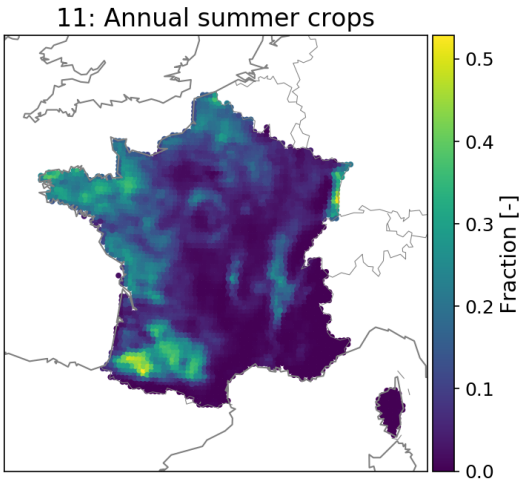
Shrubs and herbaceous vegetation

- **Class 34** Natural grasslands: low productivity grassland, often situated in areas of rough, uneven ground and frequently includes rocky areas, briars and heathland.
- **Class 36** Woody moorlands: spontaneous vegetation dominated by woody plants (heather, briar, broom, etc.) and semi-woody plants (fern, phragmites, etc.) shorter than 5 m.

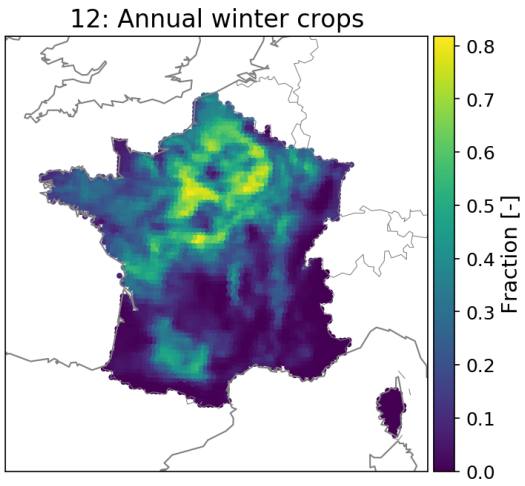
Open spaces with little or no vegetation

- **Class 45** Bare rock: Cliffs, rock outcrops, areas of active erosion, rocks and reef flats situated above the high-water mark.
- **Class 46** Beaches, dunes and sand plains: beaches, dunes and expanses of sand or pebbles in coastal or continental locations, including beds of stream channels with torrential regime.
- **Class 51** Water bodies: all water bodies longer than 20 m and all water courses longer than 7.5 m.
- **Class 53** Glaciers and perpetual snow: land covered by glaciers or permanent snowfields.

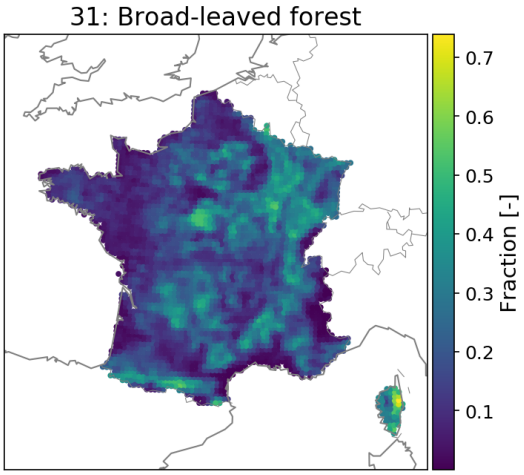
A.3 Land cover fractions, mapped per class



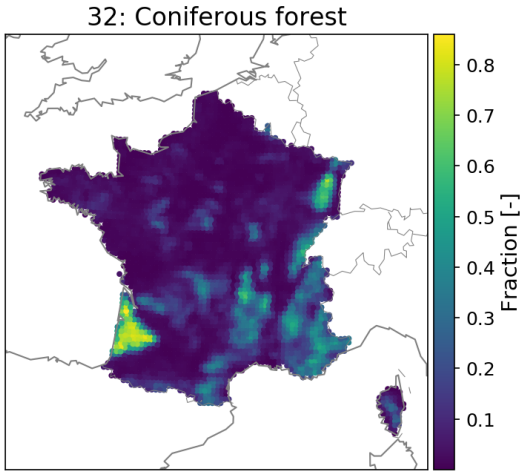
(a) Class 11: Annual summer crops



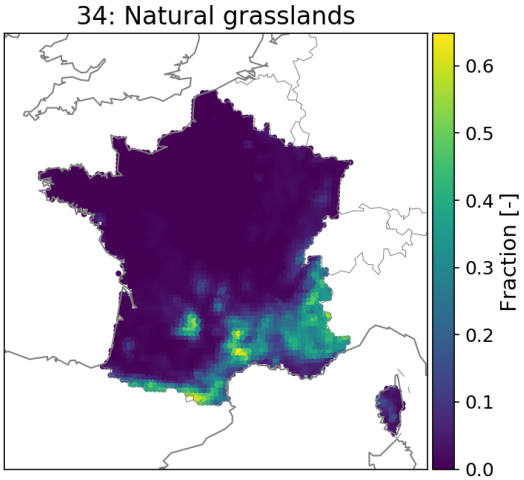
(b) Class 12: Annual winter crops



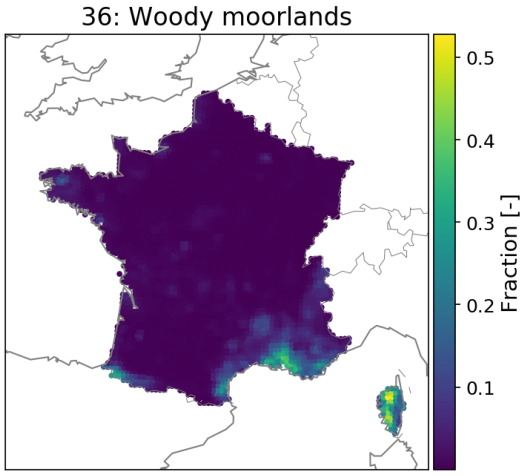
(c) Class 31: Broad-leaved forest



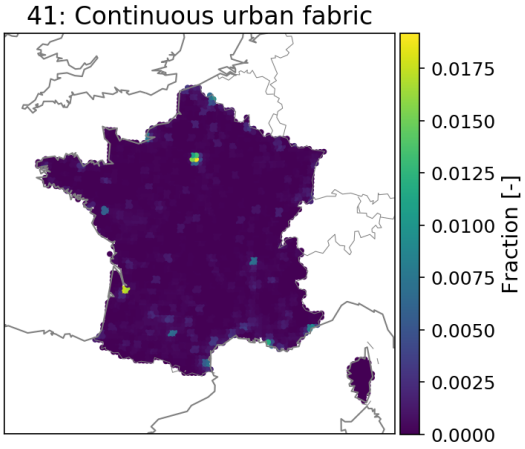
(d) Class 32: Coniferous forest



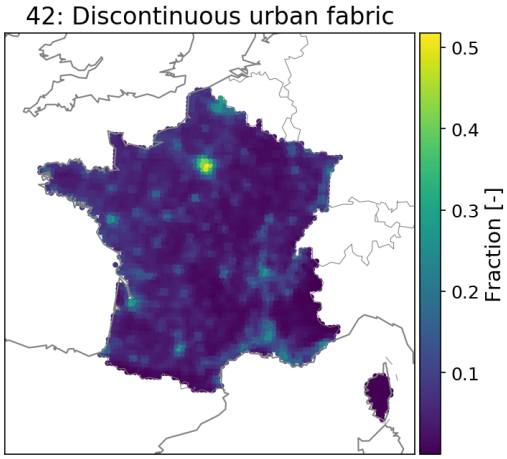
(e) Class 34: Natural grasslands



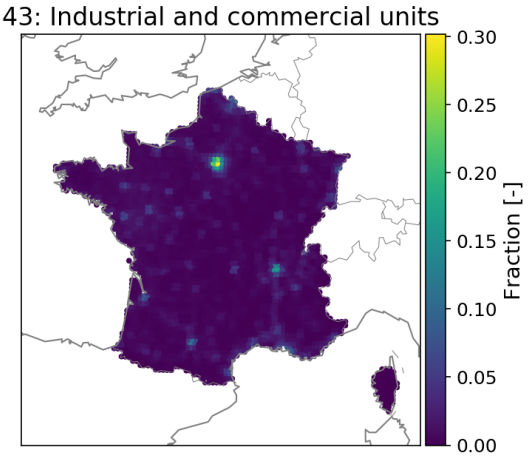
(f) Class 36: Woody moorlands



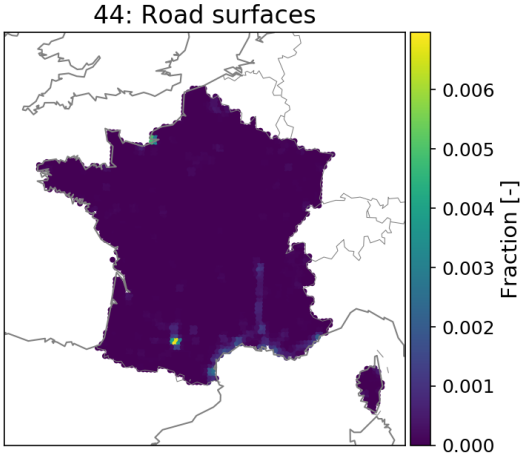
(g) Class 41: Continuous urban fabric



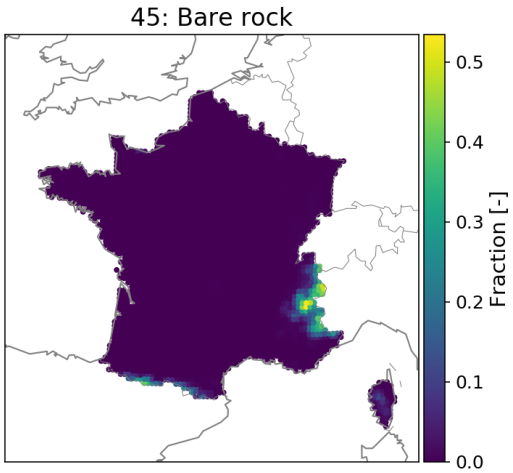
(h) Class 42: Discontinuous urban fabric



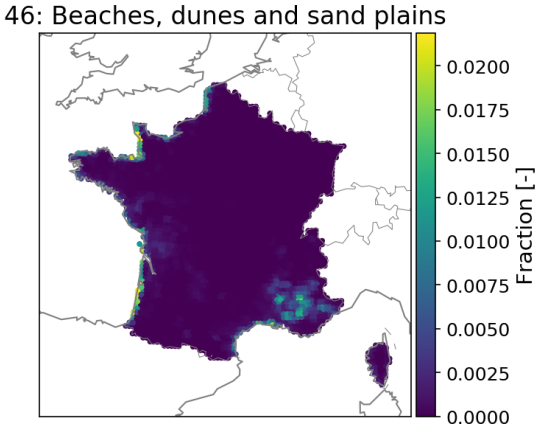
(i) Class 43: Industrial and commercial units



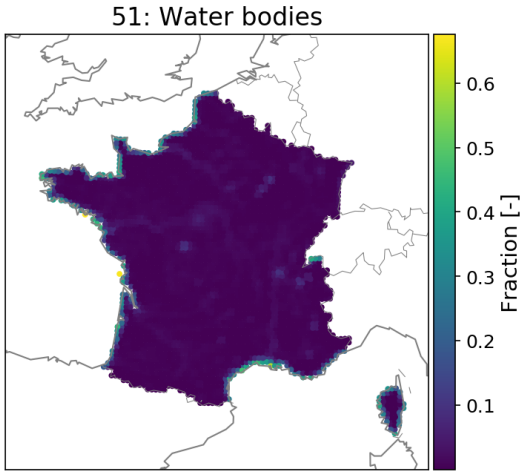
(j) Class 44: Road surfaces



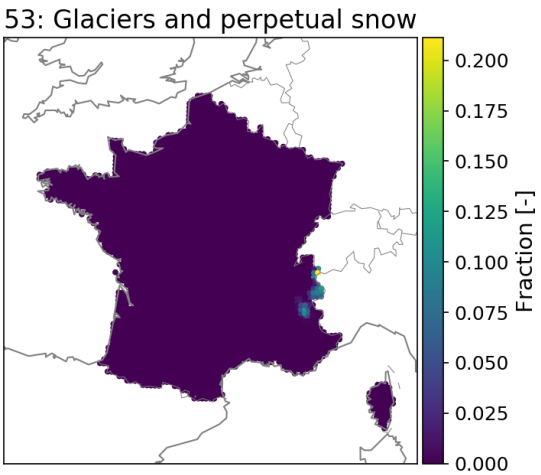
(k) Class 45: Bare rock



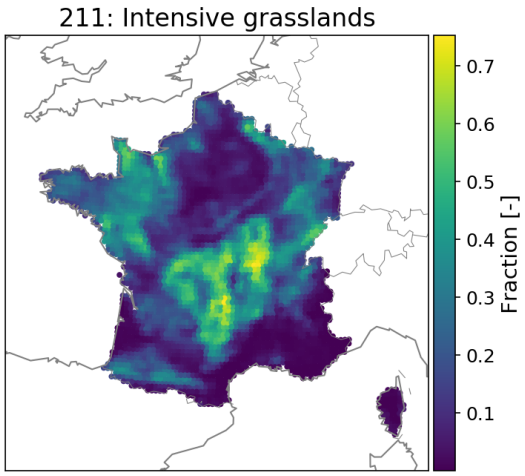
(l) Class 46: Beaches, dunes and sand plains



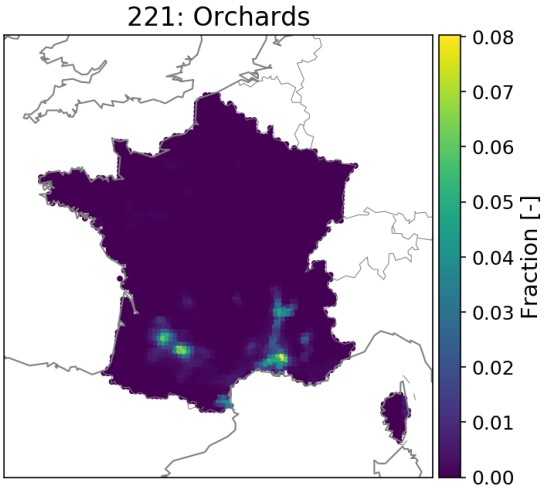
(m) Class 51: Water bodies



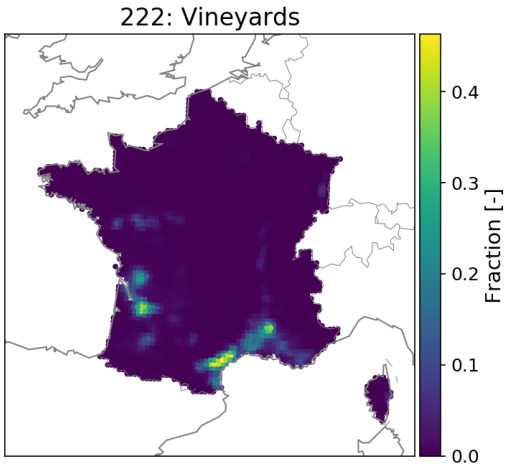
(n) Class 53: Glaciers and perpetual snow



(o) Class 211: Intensive grasslands



(p) Class 221: Orchards



(q) 222: Vineyards

Fig. A.2: Land cover fractions, mapped per land cover class

B Principal Component Analysis

B.1 Reconstructing original data from PCA output

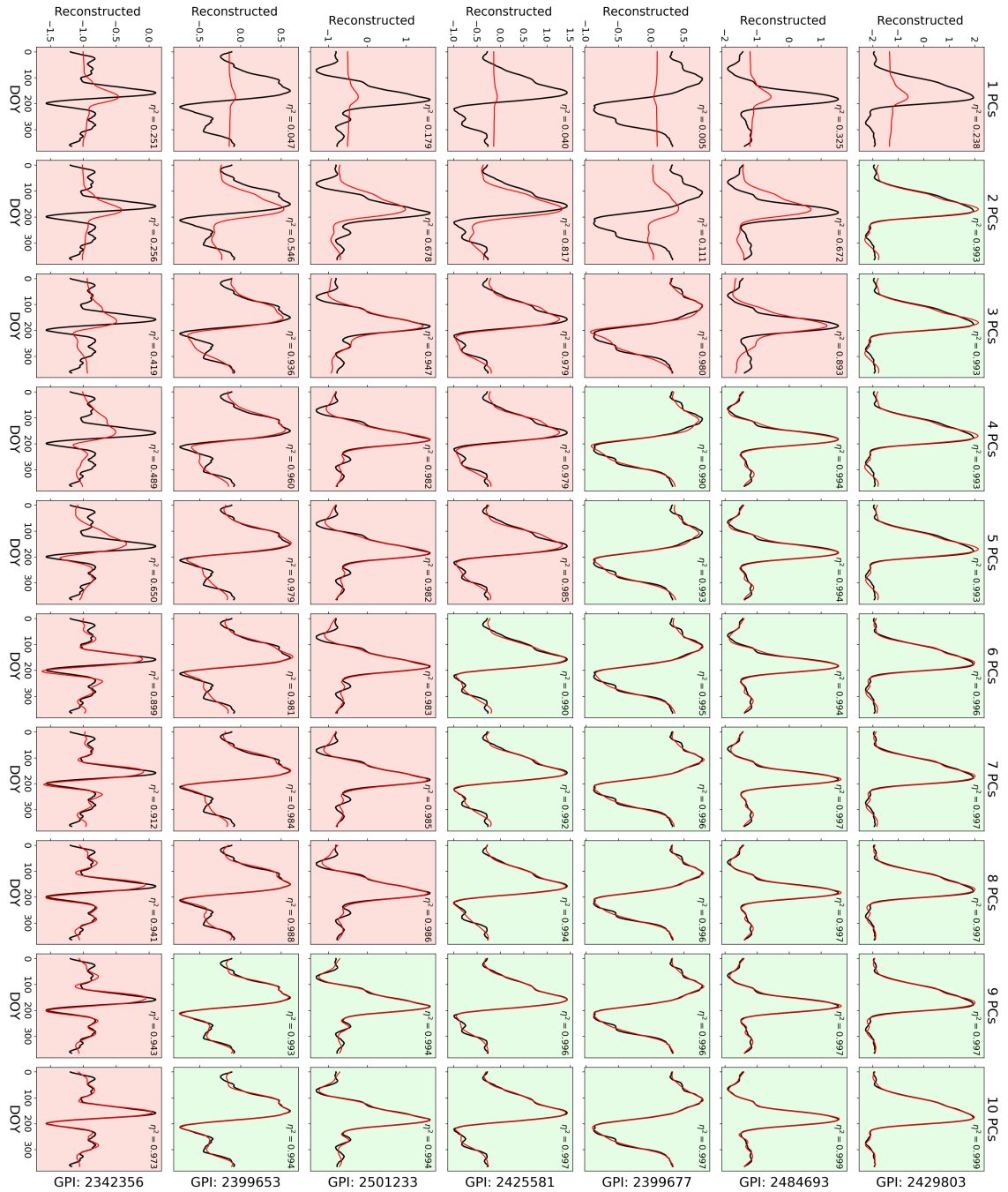


Fig. B.1: Several examples of how original standardized σ' data (dB/deg) is reconstructed from the PCA output (i.e. headings and loadings). A plot with a red background indicates that $\eta^2 < 0.99$ for that grid point and number of retained PCs, while a green background indicates that $\eta^2 \geq 0.99$. It can be seen that the data of some grid points is properly reconstructed for relatively few (i.e. one or two) PCs, while some grid points require over eight PCs before their data is properly reconstructed.

B.2 PCA performance for different number of retained PCs

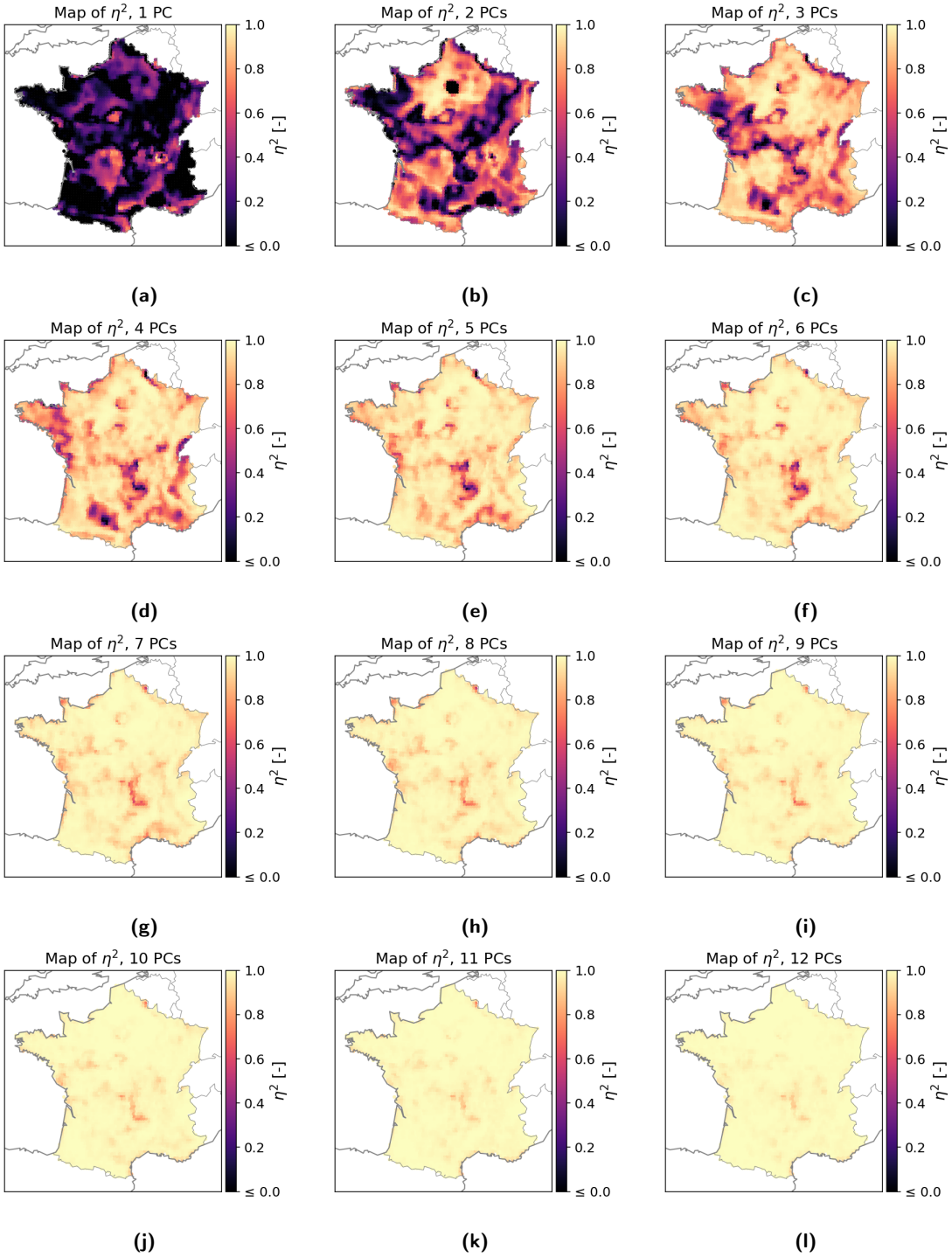


Fig. B.2: Mapped PCA performance for different numbers of PCs. η^2 is the explained variance score (see section 3.4.2).

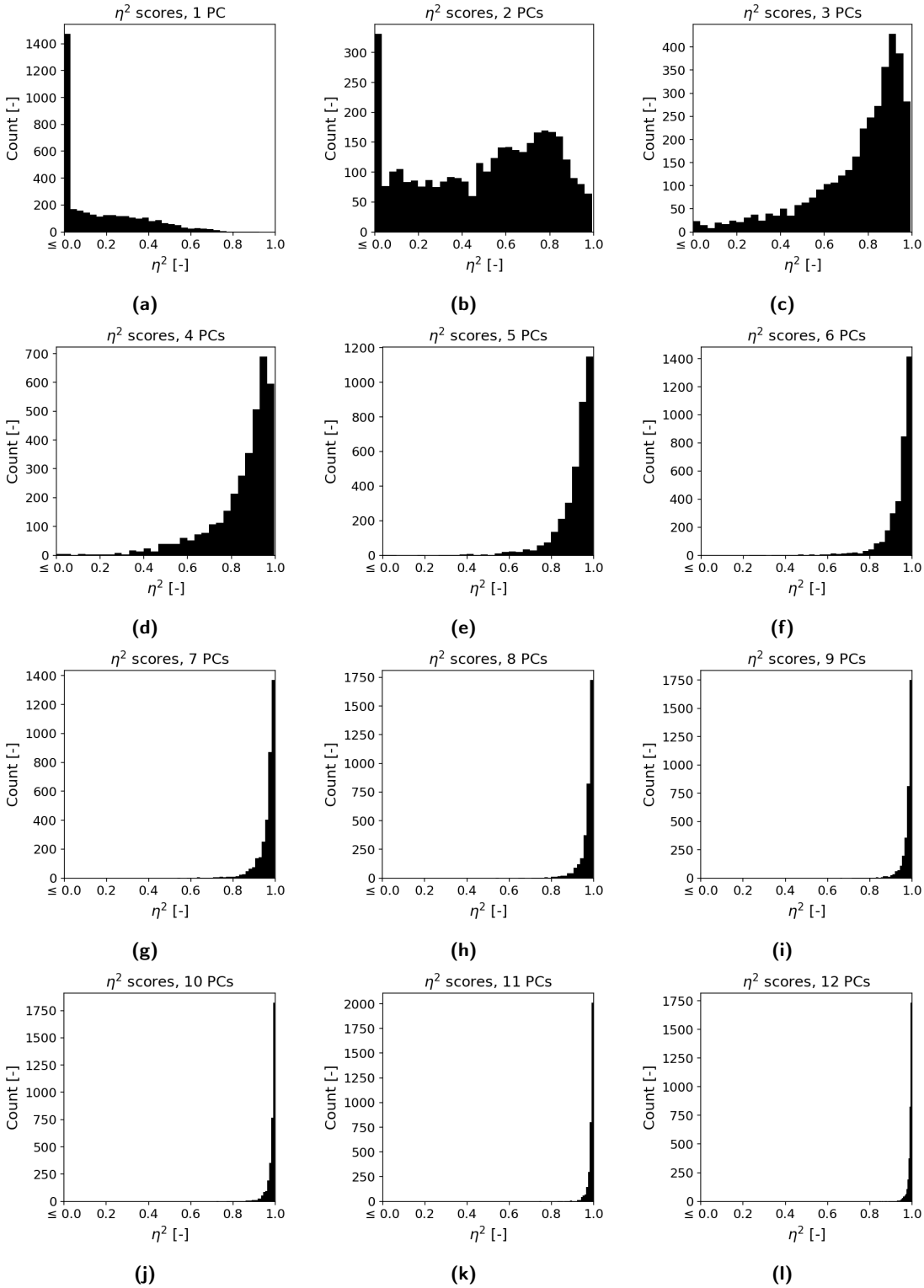
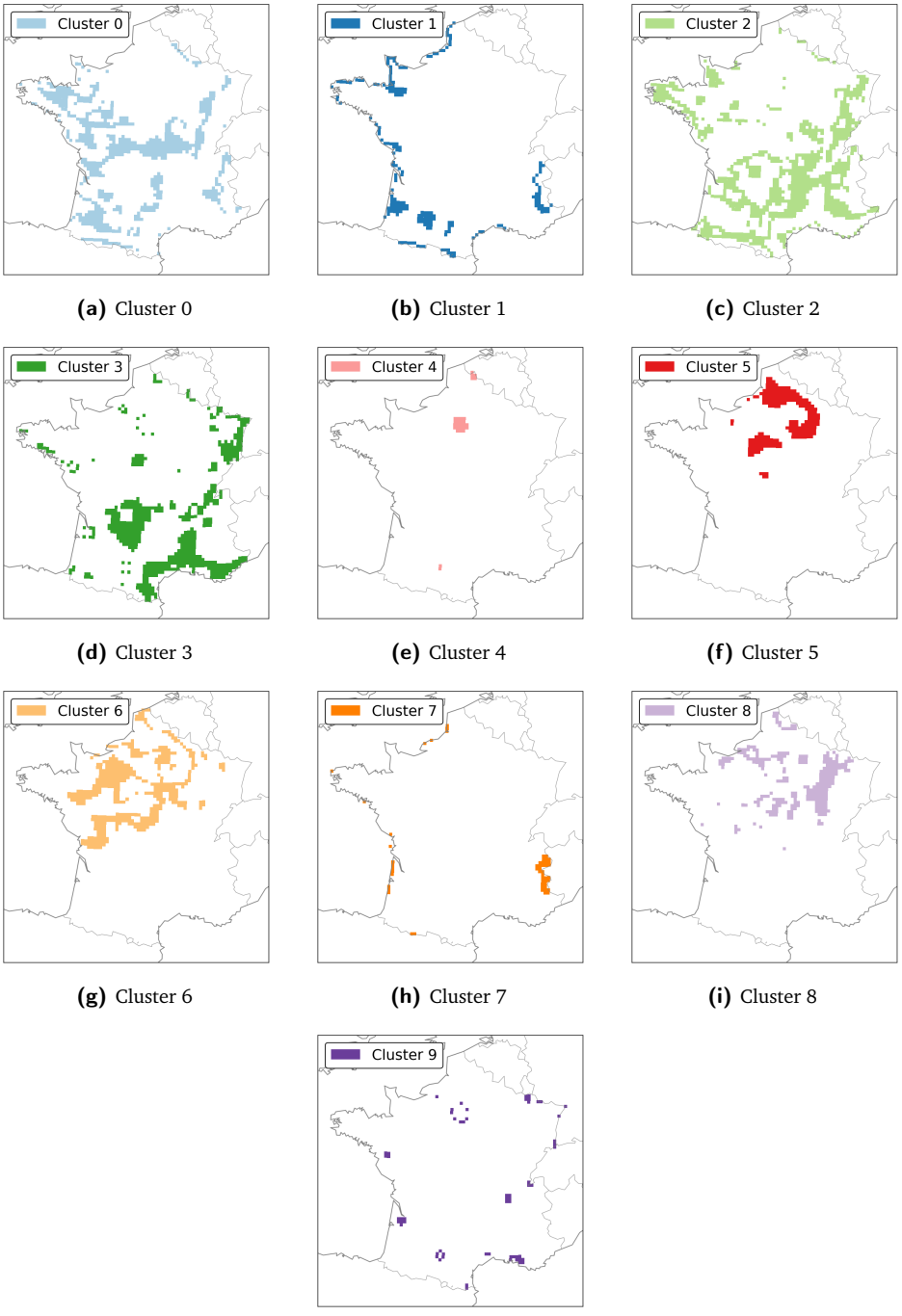


Fig. B.3: PCA performance histograms for different numbers of PCs. η^2 is the explained variance score (see section 3.4.2).

C Characteristics of generated clusters

C.1 Spatial distribution



(a) Cluster 0 **(b)** Cluster 1 **(c)** Cluster 2

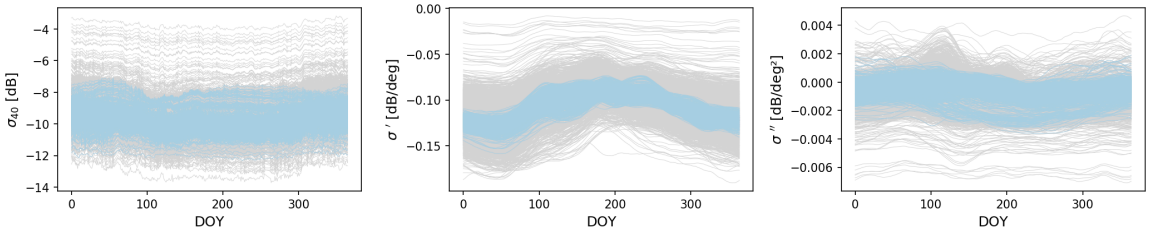
(d) Cluster 3 **(e)** Cluster 4 **(f)** Cluster 5

(g) Cluster 6 **(h)** Cluster 7 **(i)** Cluster 8

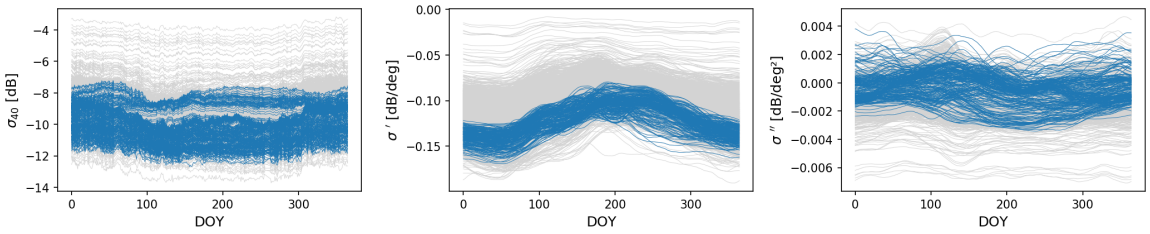
(j) Cluster 9

Fig. C.1: Grid points, mapped per cluster

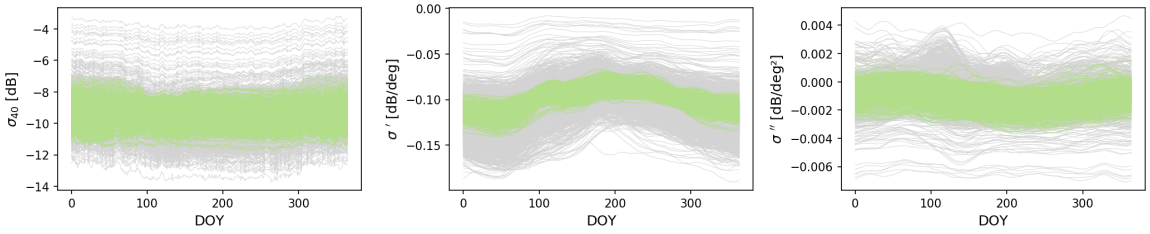
C.2 Relative seasonal signatures



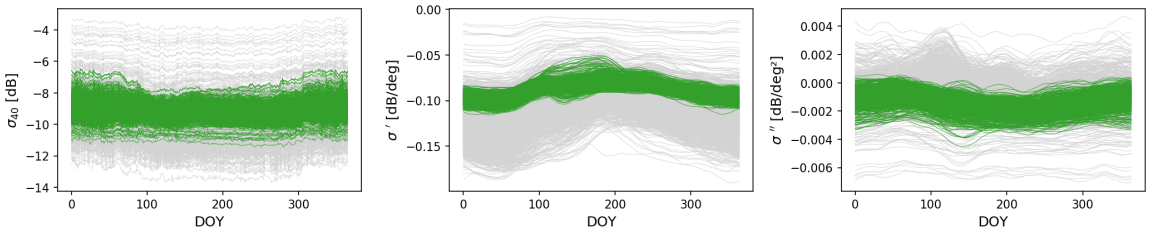
(a) Cluster 0



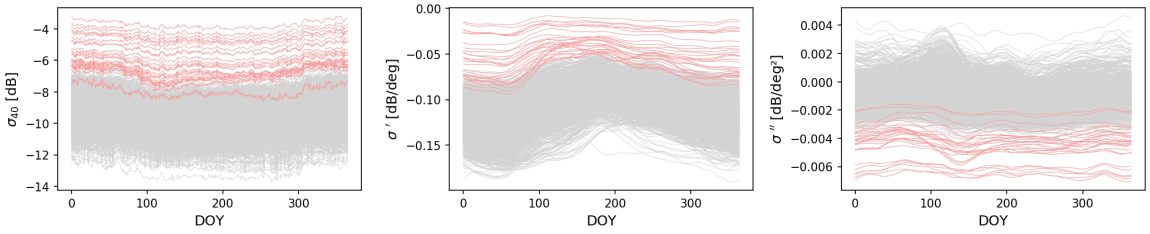
(b) Cluster 1



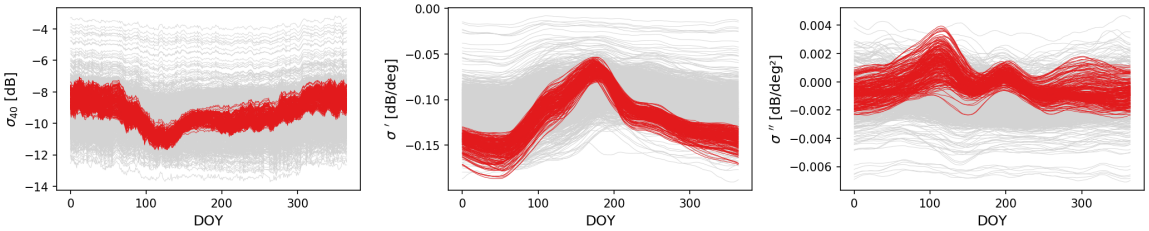
(c) Cluster 2



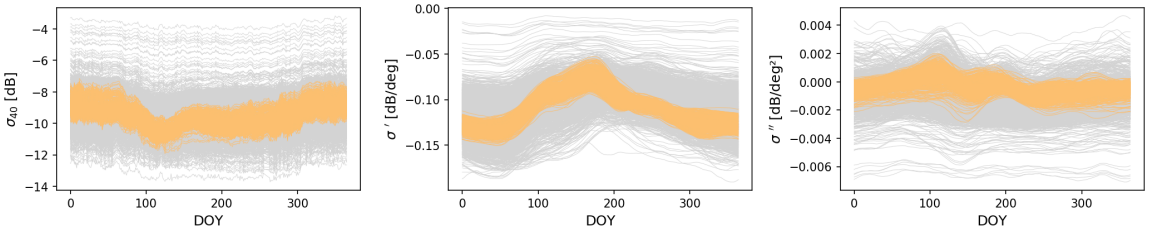
(d) Cluster 3



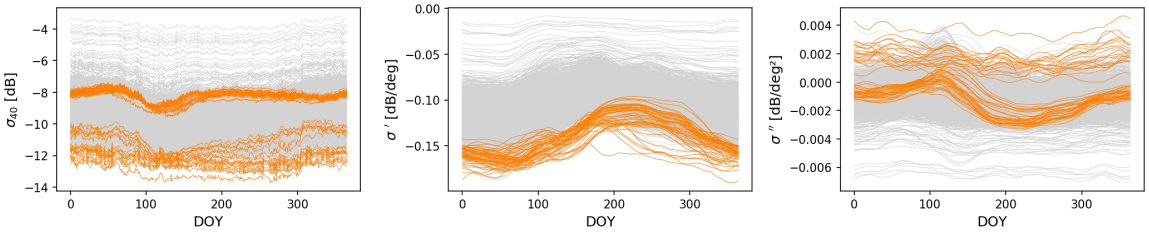
(e) Cluster 4



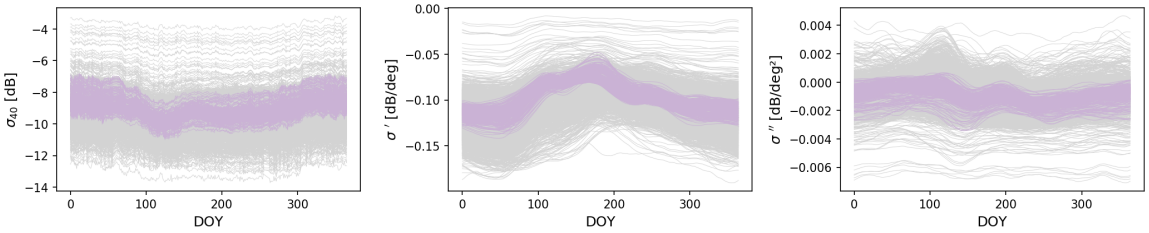
(f) Cluster 5



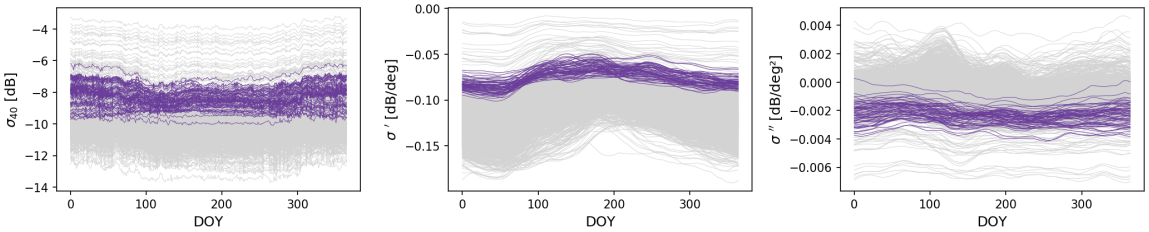
(g) Cluster 6



(h) Cluster 7



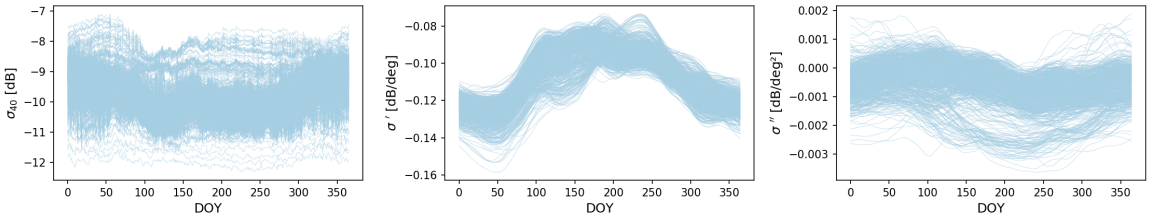
(i) Cluster 8



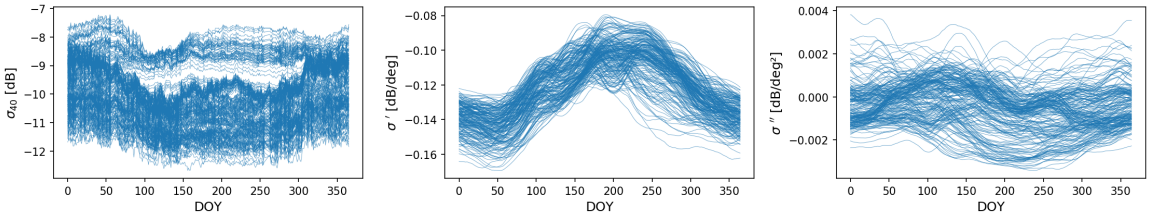
(j) Cluster 9

Fig. C.2: Seasonal climatology of σ°_{40}, σ', σ'' per cluster, plotted relative to those of all other grid points (in grey).

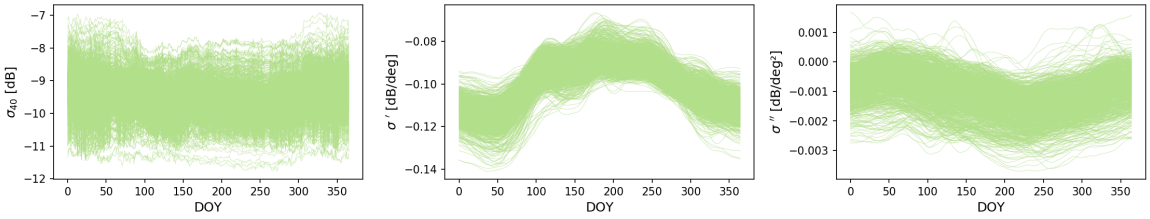
C.3 Scaled seasonal signatures



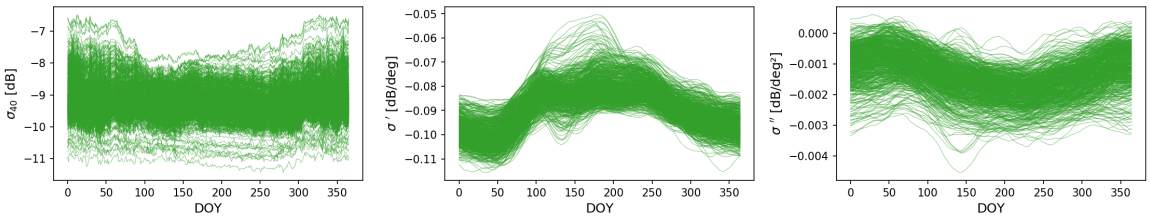
(a) Cluster 0



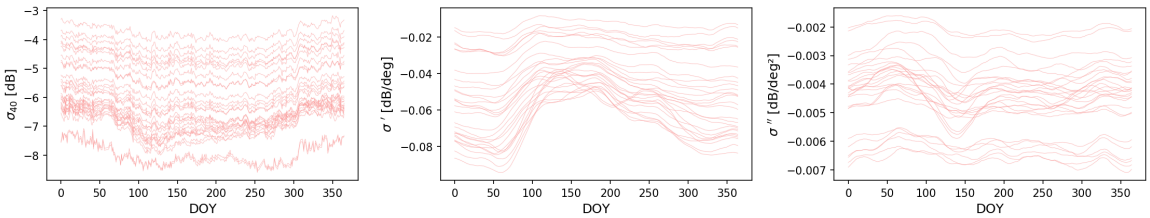
(b) Cluster 1



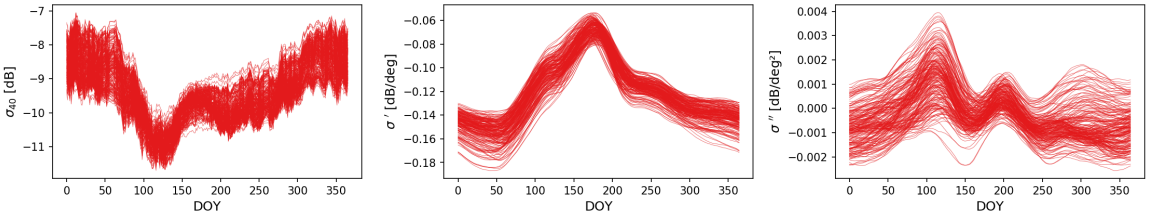
(c) Cluster 2



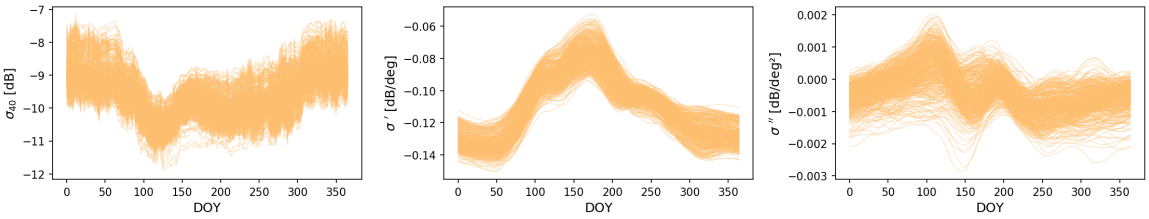
(d) Cluster 3



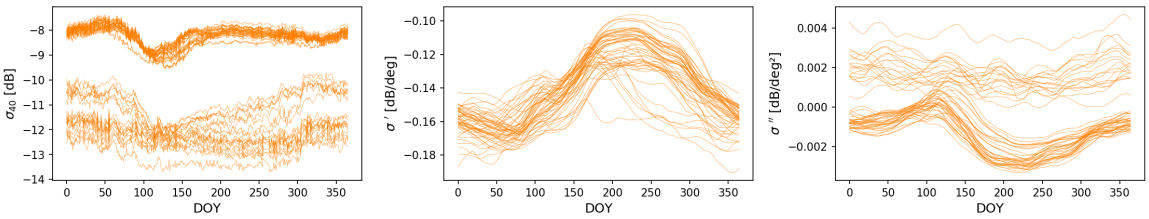
(e) Cluster 4



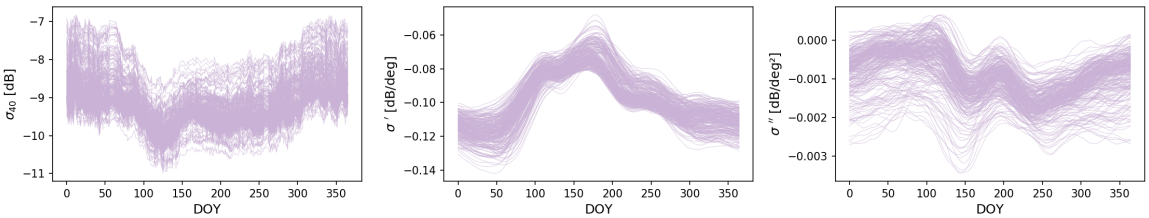
(f) Cluster 5



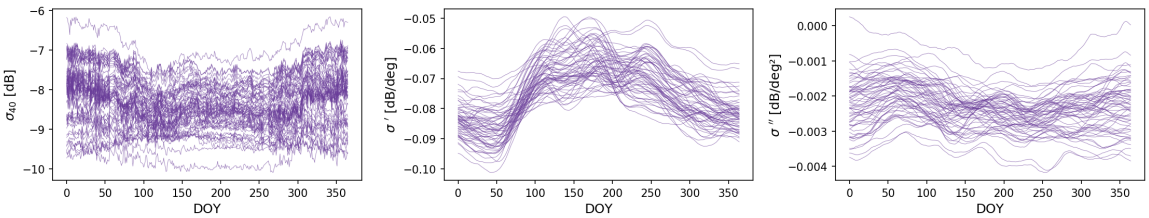
(g) Cluster 6



(h) Cluster 7



(i) Cluster 8



(j) Cluster 9

Fig. C.3: Seasonal climatology of σ°_{40}, σ', σ'' per cluster, plotted relative to all within-cluster grid points.

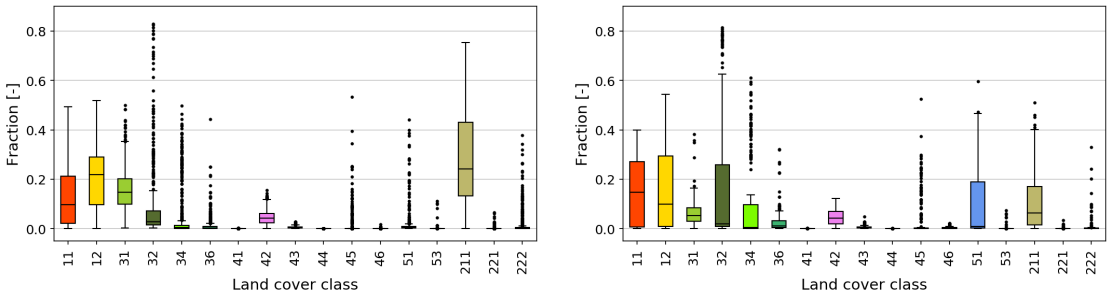
C.4 Land cover composition

Color	Class	Name
	11	Annual summer crops
	12	Annual winter crops
	31	Broad-leaved forest
	32	Coniferous forest
	34	Natural grasslands
	36	Woody moorlands
	41	Continuous urban fabric
	42	Discontinuous urban fabric
	43	Industrial and commercial units
	44	Road surfaces
	45	Bare rock
	46	Beaches, dunes and sand plains
	51	Water bodies
	53	Glaciers and perpetual snow
	211	Intensive grasslands
	221	Orchards
	222	Vineyards

Table C.1: Description of land cover classes.

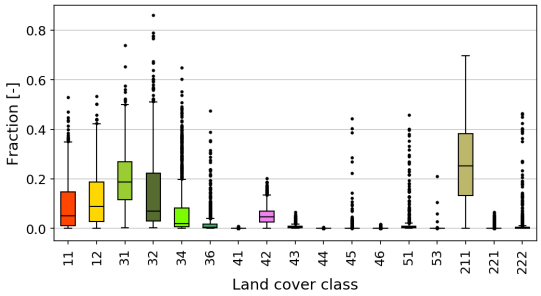
Class	Cl. 0	Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5	Cl. 6	Cl. 7	Cl. 8	Cl. 9
11	12.67	15.66	9.02	5.3	7.84	13.74	10.29	2.47	6.99	8.58
12	20.2	16.07	11.65	7.41	25.07	61.14	44.16	3.3	37.88	16.66
31	15.95	6.76	19.81	25.37	18.4	10.86	17.32	3.21	28.87	16.34
32	8.05	17.03	13.83	15.88	0.8	1.18	2.12	21.6	3.68	9.57
34	4.03	9.92	7.42	11.25	0.1	0.21	0.15	25.63	0.54	4.84
36	1.2	2.86	2.25	5.28	0.13	0.29	0.27	3.4	0.33	5.89
41	0.02	0.02	0.03	0.05	0.46	0.02	0.02	0.01	0.02	0.33
42	4.55	4.55	5.23	7.13	30.62	4.16	4.68	1.97	5.61	16.65
43	0.55	0.63	0.68	1.09	10.29	0.53	0.58	0.25	0.92	5.22
44	0.0	0.0	0.0	0.01	0.06	0.0	0.0	0.0	0.0	0.06
45	0.95	3.6	0.45	0.26	0.0	0.0	0.02	21.28	0.02	0.07
46	0.04	0.33	0.06	0.07	0.0	0.0	0.02	0.38	0.01	0.05
51	1.67	9.72	1.64	1.72	1.28	0.46	0.92	12.08	0.92	3.52
53	0.07	0.16	0.05	0.0	0.0	0.0	0.0	1.55	0.0	0.0
211	28.56	11.74	26.25	15.81	4.9	7.36	19.28	2.82	14.09	10.24
221	0.1	0.1	0.21	0.33	0.04	0.0	0.01	0.0	0.0	0.25
222	1.38	0.86	1.41	3.05	0.01	0.04	0.16	0.06	0.1	1.73

Table C.2: Mean percentage of area per land cover class for each cluster.

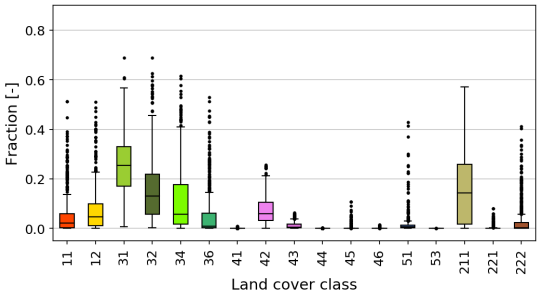


(a) Cluster 0

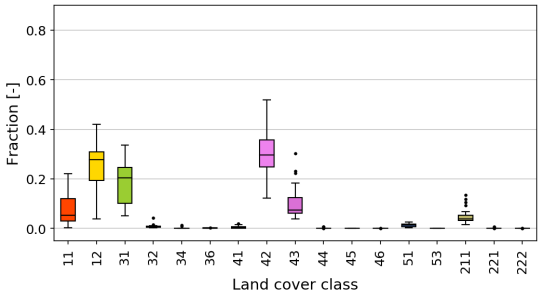
(b) Cluster 1



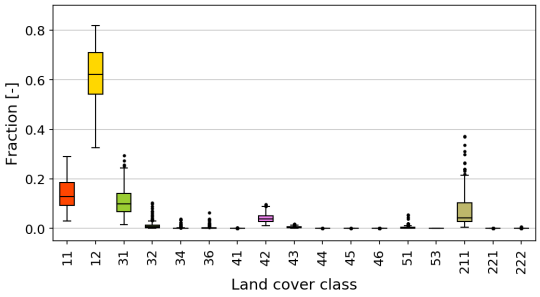
(c) Cluster 2



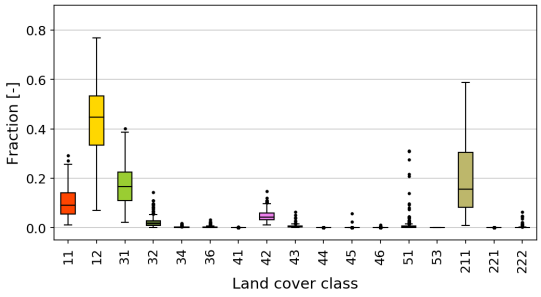
(d) Cluster 3



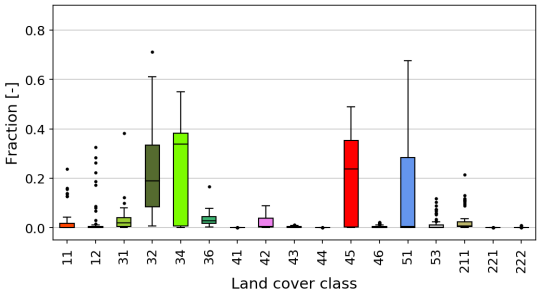
(e) Cluster 4



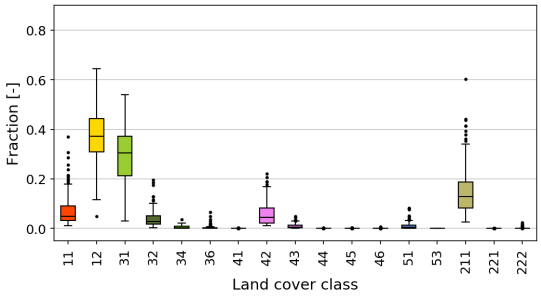
(f) Cluster 5



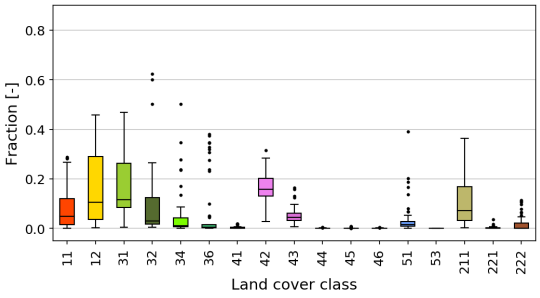
(g) Cluster 6



(h) Cluster 7



(i) Cluster 8



(j) Cluster 9

Fig. C.4: Boxplots of the land cover footprint for each cluster. The box extends from the lower (Q1) to upper quartile (Q3) values of the data, with a line at the median. The lower and upper whiskers extend to $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, respectively, where $IQR = Q3 - Q1$. Data beyond the whiskers are considered outliers and are plotted as individual points.

D

Relating land cover classes to backscatter signatures

The cluster analysis performed in section 4.4.2 indicates that consistencies exist between σ°_{40}, σ', and σ'' signatures, and between land cover footprint and the aforementioned seasonal signatures. However, because the clusters are generated based on σ', the clusters generally have footprints consisting of several dominant land cover classes, or have footprints that are very mixed. Even though it is possible to (partially) describe the backscatter characteristics of some land cover classes (e.g. urban fabric, agriculture), it is impossible to determine the specific σ°_{40}, σ', and σ'' signatures for all classes based on the clusters only.

Clearly, Appendix A.2 shows the land cover footprint of France is relatively heterogeneous compared to, for example, the Sahara desert, the Amazon rainforest, or the North-American grasslands; a total of 17 land cover classes are identified in France. Moreover, since the ASCAT data has a resolution of 25 km, no grid points with a 100% "pure" land cover footprint exist in France, i.e. a land cover footprint consisting of one land cover class. However, as can also be seen in Appendix A.2, grid points where one or two land cover types are clearly dominant do exist. By investigating only the grid points with a relatively pure land cover footprint and/or grid points with a certain known land cover footprint, it may be possible to relate each land cover class to a characteristic set of σ°_{40}, σ', and σ'' signatures. This section serves to further investigate the relationship between the land cover classes and their seasonal scattering behavior, with as goal to characterize – as best as possible – each land cover class in terms of σ°_{40}, σ', and σ''.

In order to achieve this, grid points are first selected based on which specific land cover class or land cover footprint is under investigation. For example, we can investigate the grid points where class 12 (annual winter crops) has a fraction larger than 60% ($p_{12} > 0.6$), or where class 11 (annual summer crops) is larger than 40% *and* class 12 (annual winter crops) is smaller than 10% ($p_{11} > 0.5$ & $p_{12} < 0.10$). After the grid points that satisfy the defined rules have been selected, their σ°_{40}, σ', and σ'' signatures are plotted, as well as their (aggregated) land cover footprint. The histograms of the investigated land cover classes are shown in Fig. D.1 in order to visualize the ranges in which they occur. The following classes have fractions that are assumed to be too low for further investigation: class 41, 43, 44, 46, 53 and 221.

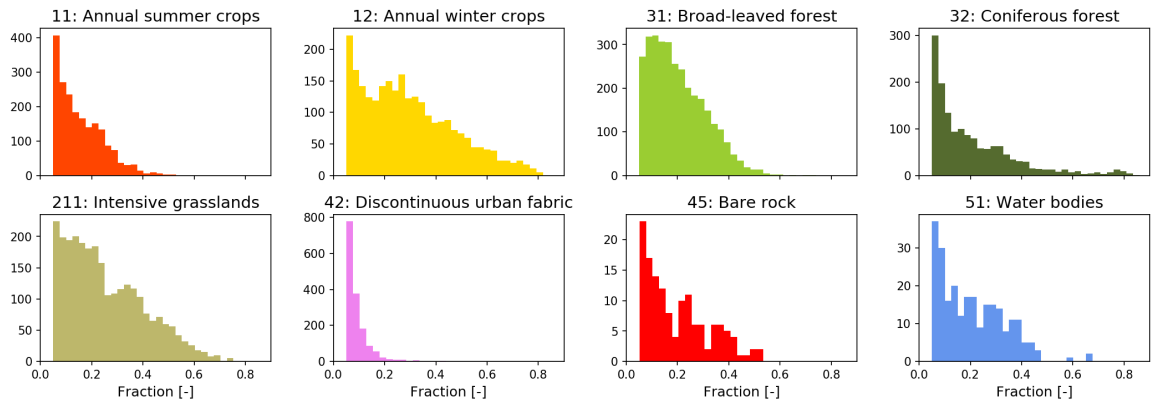
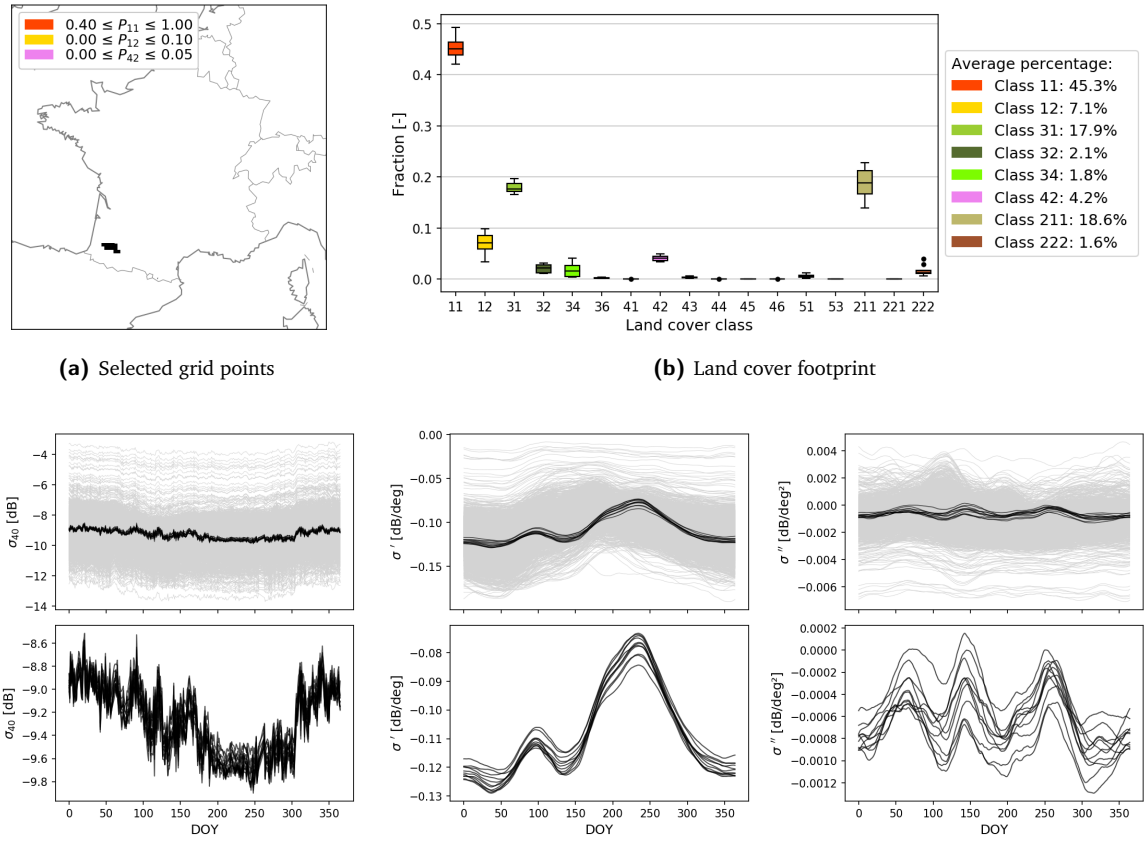


Fig. D.1: Histograms of all land cover classes in France that occur in relatively large fractions ($p_{max} > 0.25$)



(a) Selected grid points

(b) Land cover footprint

(c) σ°_{40}, σ', σ'' seasonal signatures

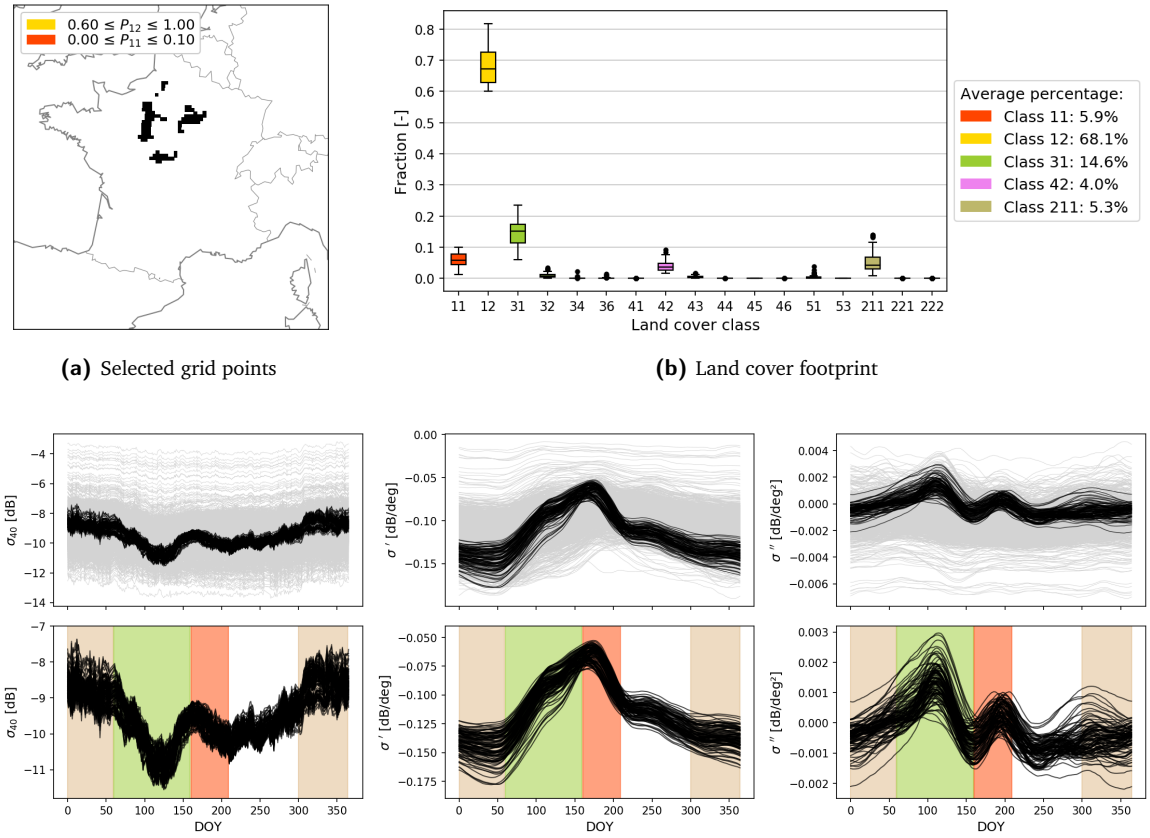
Fig. D.2: Characteristics of grid points with a high fraction of class 11 (annual summer crops)

D.1 Vegetation classes

D.1.1 Class 11: Annual summer crops

In order to find grid points with a homogeneous land cover footprint where annual summer crops are the dominant land cover class, three rules are defined; $p_{11} > 0.4$ to find the grid points with high fractions of annual summer crops; $p_{12} < 0.1$ to minimize the influence of annual winter crops; and $p_{42} < 0.05$ to exclude urban areas and their strong influence on the backscatter signal. The resulting grid points are located in the south of France and are mapped in Fig. D.2a. While annual summer crops are dominant in this area, Fig. D.2b shows that the selected grid points also contain significant fractions of intensive grasslands ($\bar{p}_{211} = 0.186$), broad-leaved forest ($\bar{p}_{31} = 0.179$), and annual winter crops ($\bar{p}_{12} = 0.071$).

The seasonal signatures of the selected grid points are shown in Fig. D.2c. σ°_{40} is relatively stable throughout the year, reaching minimum values around the end of summer and maximum values during winter. σ' has two distinct peaks; one lower peak around day 100 and one higher peak between day 200 – 250. Annual summer crops mainly consist of corn and sunflower (about 15% and 5% of total agricultural area [18]), which are sown between March and mid June (day 60 – 170) and harvested between mid August to mid September (day 220 – 260). As such, the second peak observed in σ' coincides with with the main growing season of annual summer crops. The first, smaller peak may be due to the (smaller) presence of annual winter crops, which start growing earlier in the year as they are sown during autumn and early winter. σ'' is close to zero and slightly noisy, but seems to contain three peaks of roughly equal magnitude around day 50, 150, and 250. These peaks indicate significant structural changes and may be caused by (1) rapid growth of winter crops, (2) rapid growth of summer crops, and (3) crop harvesting.



(a) Selected grid points

(b) Land cover footprint

(c) σ°_{40}, σ', σ'' seasonal signatures, also showing the development phases described in Fig. D.4a

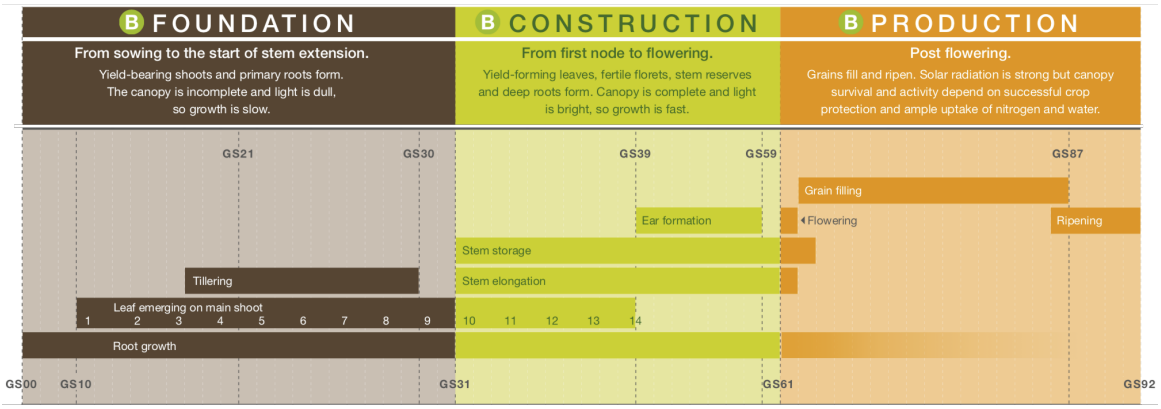
Fig. D.3: Characteristics of grid points with a high fraction of class 12 (annual winter crops)

D.1.2 Class 12: Annual winter crops

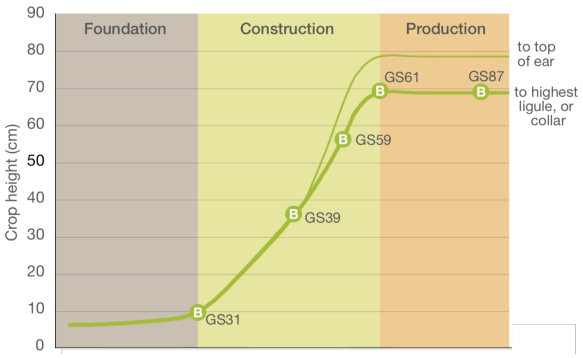
Two rules are defined in order to find the grid points where annual winter crops dominate the land cover footprint: $p_{12} > 0.6$ to maximize the fraction of annual winter crops, and $p_{11} < 0.1$ to minimize the fraction of annual summer crops. The resulting grid points and their land cover footprint are shown in Fig. D.3a and Fig. D.3b. The selected grid points also consist of broad-leaved forest ($\bar{p}_{31} = 0.146$), annual summer crops ($\bar{p}_{11} = 0.059$), intensive grasslands ($\bar{p}_{211} = 0.053$), and some urban area ($\bar{p}_{42} = 0.04$).

Since winter wheat is the dominant winter crop in this area (about 45% of total agricultural area, compared to 10% for barley and 5% for rapeseed [18]), the observed seasonal signatures are likely best described by the growth cycle of wheat. In France, wheat is sown in October and harvested the following year in July and August. As shown in Fig. D.4a, three distinct development phases are identified: foundation (October – February), construction (March – May), and production (June – August).

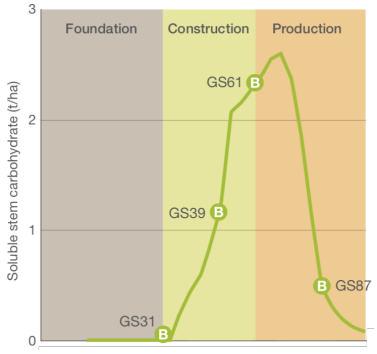
The foundation period (day 300 – 60) is characterized by seedling establishment and the emergence of leaves and tillers. Vegetation is sparse during this period, which corresponds to the low σ' values observed between day 300 – 60. Due to the low vegetation density the backscatter signal is dominated by bare soils, which mainly produce surface scattering [3]. Hence, σ°_{40} decreases significantly for increasing incidence angles. Moreover, soil moisture is highest during the winter months in France, which explains the observed maximum σ°_{40} values between day 300 – 60. Due to the lack of vegetation, ground-bounce scattering and direct scattering from the vertical canopy constituents both do not occur and are therefore equally dominant. This corresponds with the observed σ'' values, which are close to zero during the foundation period. However, σ'' increases toward the start of the construction period indicating that ground-bounce scattering becomes dominant over direct scattering from the canopy. This may be explained by the end of winter dormancy and the start of (mainly) vertical vegetation growth.



(a) Wheat development phases



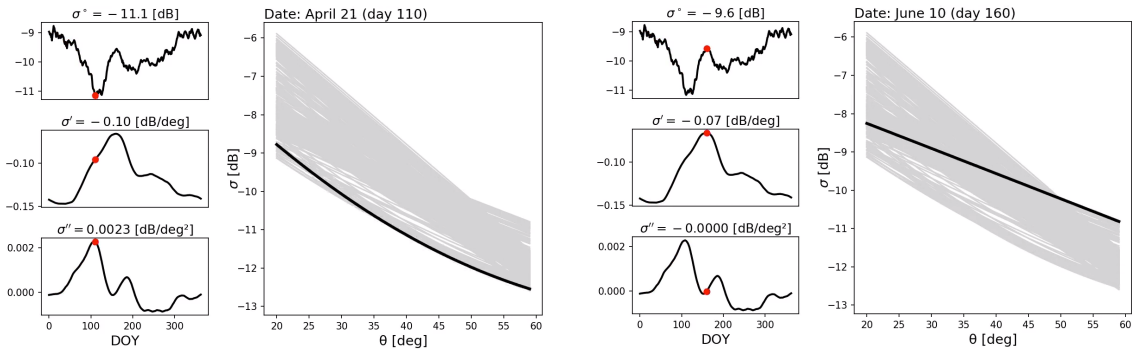
(b) Wheat crop height



(c) Wheat stem storage

Fig. D.4: Characteristics of the annual growth cycle of winter wheat (adapted from Sylvester-Bradley et al. [61])

The construction period (day 60 – 160) is characterized by rapid stem elongation, increasing stem reserves and the formation of grain ears. As can be seen in Fig. D.3c, the backscatter signatures between day 60 – 110 are markedly different than those between day 110 – 160. The first half of the construction period shows sharply decreasing σ°_{40}, increasing σ', and increasing σ''. As indicated by σ°_{40}, total backscatter sharply decreases during the first part of the construction period; this behavior has been attributed to the rapid development of the wheat stem. As shown in Fig. D.4b, the crop height significantly increases during stem elongation and hence, the crop structure changes from a predominantly horizontal structure to a vertical structure dominated by the main stem. The vertically oriented wheat stems couple much more effectively with VV-polarized waves, which results in lower backscatter due to strong attenuation on both the incoming and return paths [64]. Research by Picard et al. [50] showed that attenuation of VV-polarized waves not only increases with stem height but also with stem gravimetric moisture content, which increases throughout the construction period (see Fig. D.4c). The increase in crop height, leaf coverage and soluble stem storage result in increasing vegetation density, which corresponds with the increasing σ' values observed between day 60 – 120. During this period σ'' increases until reaching maximum (positive) values around day 120, meaning that σ°_{40} increases for large incidence angles (i.e. the $\sigma^\circ_{40} - \theta$ relationship curves upwards for large θ [59]), see Fig. D.5a. This can be explained by the stem dominated vertical structure and the incidence angle dependence of the occurring scattering mechanisms. For low θ, the total backscatter is dominated by direct scattering from the soil as well as ground-bounce scattering between the soil and stems. The soil return decreases with increasing θ, as does the ground-bounce contribution; the path through the canopy becomes longer for increasing θ, causing the incoming and outgoing waves to interact with more and more stems which leads to increased attenuation [10]. On the other hand, direct backscatter from the crop canopy increases with θ [66]. The combination of these higher order effects cause σ°_{40} to decrease with increasing θ if $\theta < 40°$, while σ°_{40} increases with increasing θ if $\theta > 40°$ [10]. This corresponds with the observed positive σ'' values.



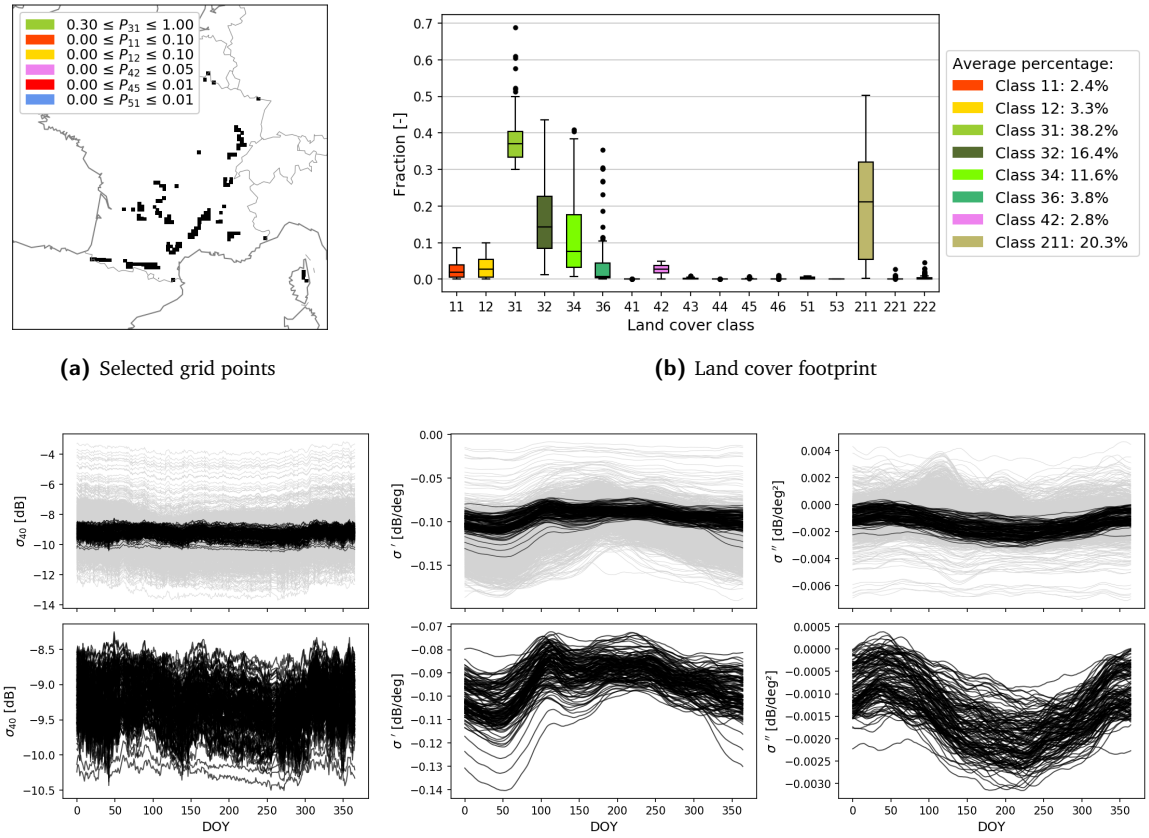
(a) $\sigma^\circ - \theta$ relationship at day 110 **(b)** $\sigma^\circ - \theta$ relationship at day 160

Fig. D.5: $\sigma^\circ - \theta$ relationship at day 110 and day 160

The backscatter signatures change significantly during the second half of the construction period (day 110 – 160); σ°_{40} sharply increases, σ' increases, and σ'' rapidly drops from maximum values to approximately zero. The resulting $\sigma^\circ - \theta$ relationship of day 160 shows that σ° increases for all incidence angles compared to day 110, but the biggest increase is found for large θ (i.e. $\theta > 40°$), see Fig. D.5. Furthermore, since ground-bounce scattering is dominant for small θ and direct scattering from vertical canopy components is dominant for large θ, it follows that the contribution of the vegetation canopy to total backscatter increases significantly more than the contribution of ground-bounce scattering via the wheat stems. This change in scattering behavior could be explained by canopy maturation and the formation of grain-bearing ears during this period, see Fig. D.4a. Direct scattering from the dense canopy becomes dominant over the attenuating effects of the vertical stems, leading to larger σ°_{40}. Furthermore, vegetation density reaches a maximum as the canopy matures and grain-ears form, which corresponds with increasing σ'. Finally, direct scattering from the vertical constituents of the canopy becomes increasingly important relative to ground-bounce scattering, which corresponds with the observed decrease of σ''.

Finally, the production period (day 160 – 210) is characterized by the filling, ripening, and harvesting of grains. Between day 160 – 210, σ° and σ' both decrease, while σ'' initially increases and subsequently decreases again. The strongest decrease in σ° is found for large incidence angles ($\theta > 50°$) while σ° stays relatively constant for small incidence angles ($\theta < 30°$). This suggests that the influence of the canopy decreases significantly during the production period, since backscatter is mainly generated by the ears and leaves at large incidence angles. This can be explained by the loss of (wet) biomass in the canopy during the production period; the leaves (largely) senesce during this period, and the moisture content of the grains drops from approximately 45% to 20% during ripening as moisture is replaced by starch and nutrients [61]. This corresponds with the decrease in σ°_{40} and σ' observed during this period. Furthermore, at the point of harvest almost all soluble stem resources have been transferred to the grains or lost through transpiration, see Fig. D.4c [61]. This may explain the increase and subsequent decrease in σ''; perhaps the water content of the canopy decreases earlier than the water content of the stem, leading to temporary dominance of the stem over the canopy (i.e. σ'' increases). As the water content of the stem continues to decrease until harvest, the dominance of the stem decreases again (i.e. σ'' decreases).

Clearly, the different development phases of wheat are reflected in the $\sigma^\circ_{40} - \theta$ relationship and the σ°_{40}, σ', and σ'' signatures. Throughout the growth cycle of wheat, the observed backscatter signatures can be logically explained by the structural changes occurring in the different vegetation components. It was found that σ°_{40}, σ', and σ'' are not only influenced by the orientation of the different vegetation elements, but also by the distribution of moisture in these elements. The results are consistent with the idea that σ' is a measure of above-ground wet biomass ("vegetation density"), and that σ'' is a measure of the relative dominance of ground-bounce scattering and direct scattering from vertical canopy components.



(a) Selected grid points

(b) Land cover footprint

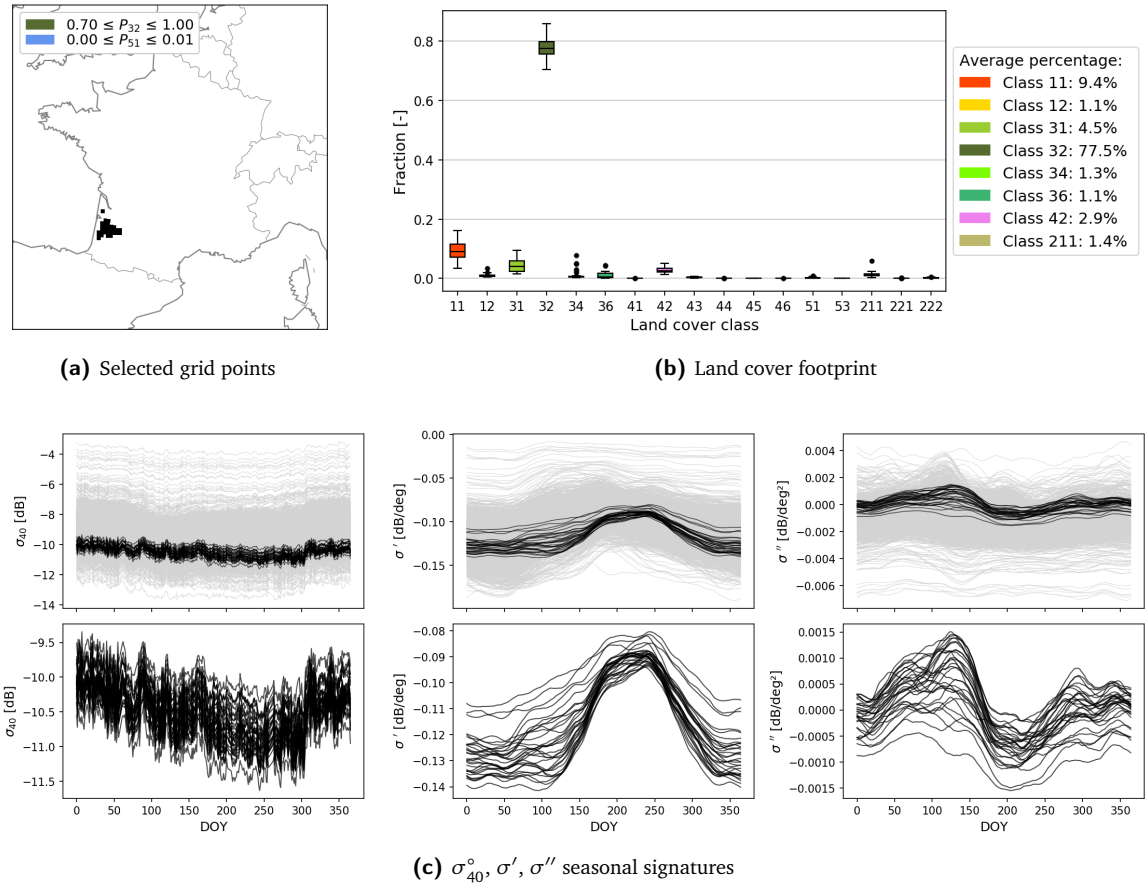
(c) σ°_{40}, σ', σ'' seasonal signatures

Fig. D.6: Characteristics of grid points with a high fraction of class 31 (broad-leaved forest)

D.1.3 Class 31: Broad-leaved forest

As can be seen in Fig. D.6a and Fig. D.7b, grid points with a relatively high fraction of deciduous broad-leaved forest ($p_{31} > 0.3$) are spread throughout France and have mixed land cover footprints. Even for relatively high fractions (e.g. $p_{31} > 0.3$), broad-leaved forest coincides with significant fractions of other land cover classes such as intensive grasslands ($\bar{p}_{211} = 0.203$), coniferous forest ($\bar{p}_{32} = 0.168$), and natural grasslands ($\bar{p}_{34} = 0.116$). As a result, the seasonal signatures of the selected grid points are relative noisy, making it difficult to investigate the backscatter effects of broad-leaved forest separately.

While σ°_{40} is relatively constant throughout the year, σ' and σ'' exhibit discernible seasonal variations. σ' generally reaches minimum values during winter (day 300 – 50) and maximum values between spring and summer (day 100 – 250). Several patterns are visible in σ', which is likely due to the the fact that the land cover footprint of the selected grid points contains significant fractions of several other vegetation types which may reach maximum vegetation density at different times in spring and/or summer. On the other hand, σ'' exhibits a more distinct seasonal cycle, reaching maximum values during winter and early spring (day 350 – 100) and minimum values during summer (day 150 – 250). One possible explanation for this behavior is the growth and fall of leaves; direct scattering from tree canopies is most dominant during summer when deciduous tree canopies reach maximum density, which corresponds with the observed decrease of σ''. As the trees drop their leaves, direct scattering from the canopy decreases and ground-bounce scattering from tree trunks and branches becomes more dominant, which corresponds with the observed increase of σ''.



(a) Selected grid points

(b) Land cover footprint

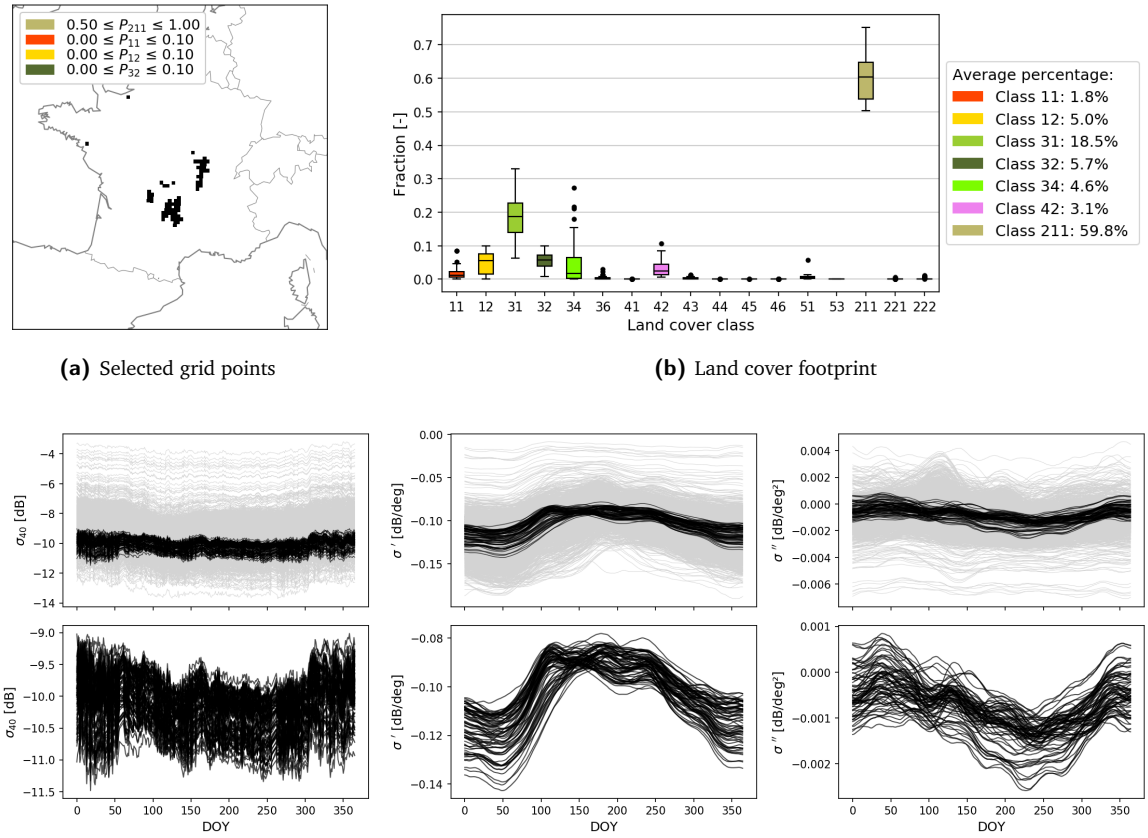
(c) σ°_{40}, σ', σ'' seasonal signatures

Fig. D.7: Characteristics of grid points with a high fraction of class 32 (coniferous forest)

D.1.4 Class 32: Coniferous forest

As shown in Fig. D.7a, an area with high fractions of coniferous forest ($p_{32} > 0.7$) exists in the Landes area south of Bordeaux. This area is the largest man-made woodland in Western Europe and mainly consists of maritime pine. Additionally, this area contains some annual summer crops ($\bar{p}_{11} = 0.094$) and some broad-leaved forest ($\bar{p}_{31} = 0.045$), making the land cover footprint of this area relatively pure.

Clear seasonal behavior can be identified in the backscatter signatures shown in Fig. D.7c. σ°_{40} is maximum during winter and steadily decreases throughout spring and summer. σ' is constant between day 300 – 120 and shows increased activity between day 120 – 300. σ'' oscillates around zero, reaching maximum (positive) values around day 120 and subsequently drops to minimum (negative) values around day 200. This behavior corresponds with the annual growth cycle of the maritime pine and may be explained specifically by the different stages of needle development. Conifers generally have a dormancy period in fall and winter, which is when new buds develop [23]. Growth mainly occurs between in summer months, which is reflected by σ'. Bud burst occurs in spring, after which the new shoots elongate rapidly. The needles stay close to the shoot axis during the first summer, and during the second summer the needles open up until oriented almost perpendicular to the shoot axis [6]. This behavior may explain why σ'' initially increases to positive values and subsequently drops to negative values. Spring is characterized by rapid elongation growth of new shoots, which have a straight structure as the needles are flat against the shoot. On the other hand, summer is characterized by the needles of one year old shoots opening up until perpendicular to the shoot axis. Due to the different timing of these processes, ground-bounce scattering increases in dominance during spring (i.e. σ'' increases), while the dominance of direct/volume scattering from the canopy increases during summer due to larger canopy density (i.e. σ'' decreases).



(a) Selected grid points

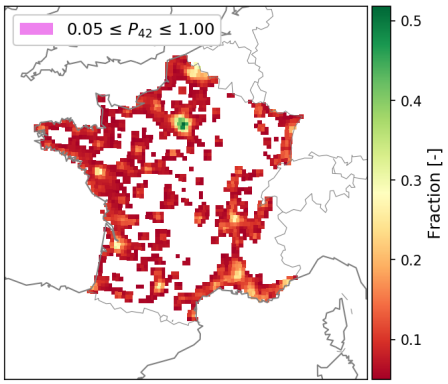
(b) Land cover footprint

(c) σ°_{40}, σ', σ'' seasonal signatures

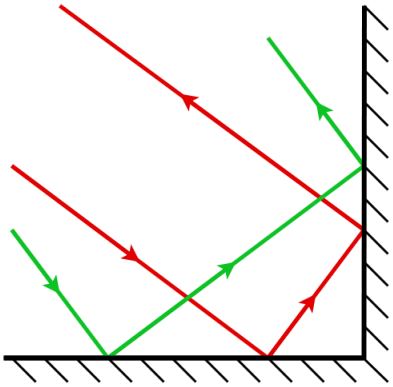
Fig. D.8: Characteristics of grid points with a high fraction of class 211 (intensive grasslands)

D.1.5 Class 211: Intensive grasslands

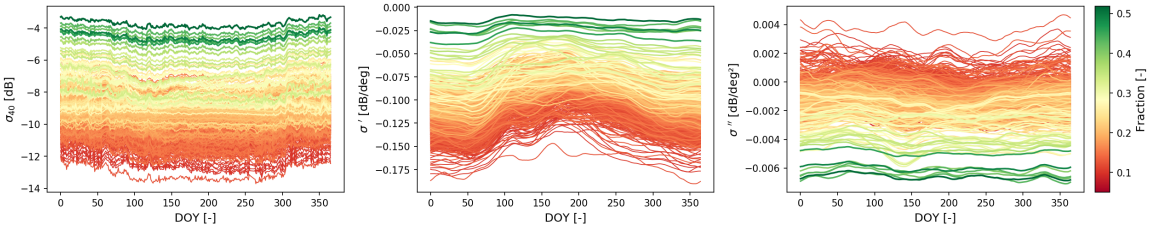
Four rules are defined to determine grid points where intensive grasslands are the dominant land cover class: the fraction of intensive grasslands should be at least 50% and the fractions of annual summer crops, annual winter crops and deciduous forest cannot be higher than 10% to minimize their influence. The resulting grid points are mapped in Fig. D.8a and are mainly located in the Central Massif, an area consisting of mountains and plateaus. Besides intensive grasslands, Fig. D.8b shows that the selected grid points contain a significant amount of deciduous forest ($\bar{p}_{31} = 0.185$). The seasonal signatures plotted in Fig. D.8c show that σ°_{40} is relatively low and has little seasonal variation in this area, which may be (in part) due to relatively low and stable soil moisture levels in highlands and plateaus. On the other hand, both σ' and σ'' show clear seasonality. σ' reaches minimum values during winter (day 350 – 50) and sustains maximum values throughout spring and summer (day 100 – 250), which corresponds to the general seasonal growth cycle of deciduous vegetation. σ'' reaches maximum values during winter (day 350 – 70) and minimum values during summer (day 200 – 300). While σ'' is mostly negative, some grid points reach (near) positive values during winter. In areas where intensive grasslands is the dominant land cover class, σ'' is largest (i.e. close to zero) when σ' is lowest, suggesting an approximately equal dominance of ground-bounce scattering and direct scattering when vegetation is sparse. This could indicate that surface scattering is the main scattering mechanism in highlands and/or plateaus when grass cover is most sparse. Conversely, σ'' is most negative when σ' is largest, suggesting that direct scattering from vertical vegetation constituents is dominant when vegetation is dense.



(a) Selected grid points, colored by class 42 fraction



(b) Corner-reflector effect of urban areas



(c) σ°_{40}, σ', σ'' seasonal signatures, colored by class 42 fraction

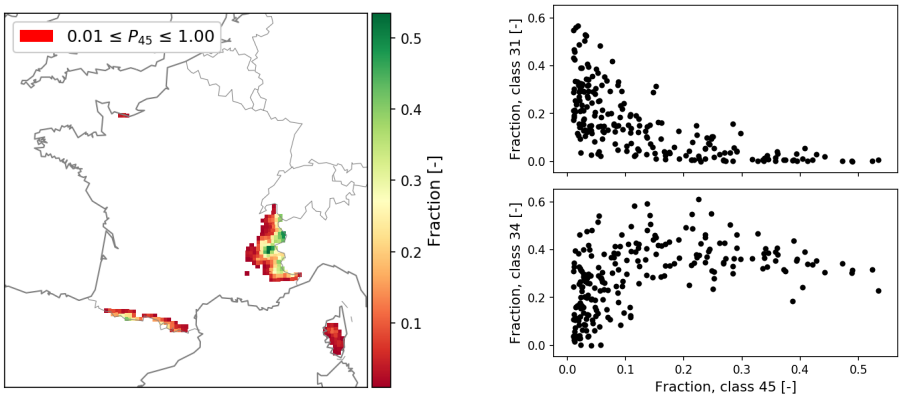
Fig. D.9: Influence of class 42 (discontinuous urban fabric) on backscatter signatures

D.2 Non-vegetation classes

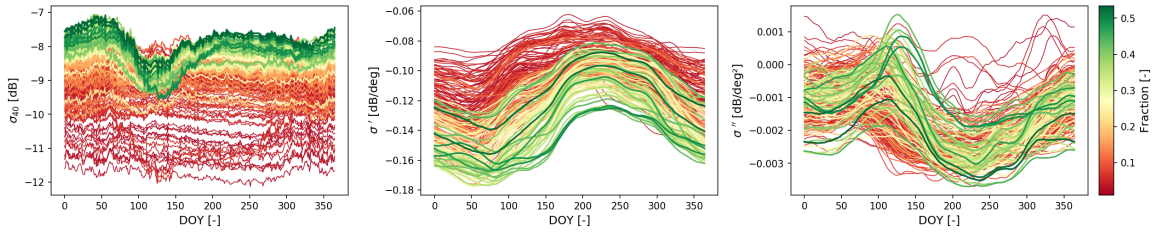
While non-vegetation land cover classes themselves may not exhibit seasonal behavior in terms of σ°_{40}, σ', and σ'', they may affect the backscatter signatures originating from vegetation located in the same grid point. In order to better understand the relationship between land cover and σ°_{40}, σ°_{40}, and σ'', the effect of non-vegetation land cover classes should be investigated. This section serves to investigate and explain the influence of discontinuous urban fabric, bare rock, and water bodies on σ°_{40}, σ', and σ''.

D.2.1 Class 42: Discontinuous urban fabric

All grid points points for which $p_{42} > 0.05$ are plotted in Fig. D.9a. The backscatter signatures of these grid points are plotted in Fig. D.9c, where the signatures of each grid point are colored according to their respective p_{42} value. As previously observed in section 4.4.2, grid points with a land cover footprint containing relatively large fractions of urban classes generally have large σ°_{40} values, shallow σ' values, and negative σ'' values, which are all relatively constant throughout the year. Fig. D.9c clearly shows that larger values of p_{42} indeed result in larger σ°_{40}, shallower σ', and more negative σ''. Moreover, seasonal variations in σ°_{40}, σ', and σ'' are damped as p_{42} increases. This can be explained by the combined effect of increasing urban area and decreasing vegetation. Firstly, urban areas behave as corner-reflectors due to their unique structural characteristics (i.e. flat surfaces joined at 90° angles, see Fig. D.9b), leading to large σ° for a wide range of incidence angles (i.e. shallow σ'). The observed negative σ'' values are explained by the fact that the corner-reflector effect is strongest for approximately $40° < \theta < 50°$ and decreases for lower and higher values of θ [37]. Secondly, vegetation cover decreases as p_{42} increases, resulting in lower seasonal variations. Due to the aforementioned characteristics, urban areas have a strong "scaling" effect on the backscatter response generated by vegetation in the same grid point. In terms of clustering, this scaling effect means that two grid points with similar land cover footprints may be assigned to different clusters if they have slightly different values of p_{42}.



(a) Selected grid points, colored by class 45 fraction **(b)** Class 45 vs class 32 and class 34 fractions



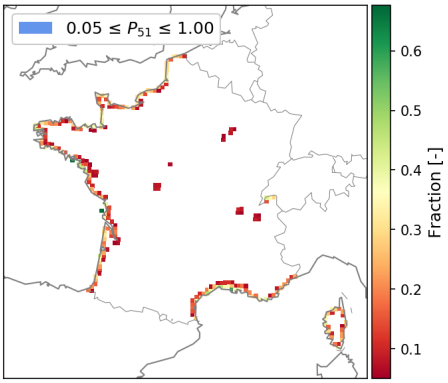
(c) σ°_{40}, σ', σ'' seasonal signatures, colored by class 45 fraction

Fig. D.10: Influence of class 45 (bare rock) on backscatter signatures

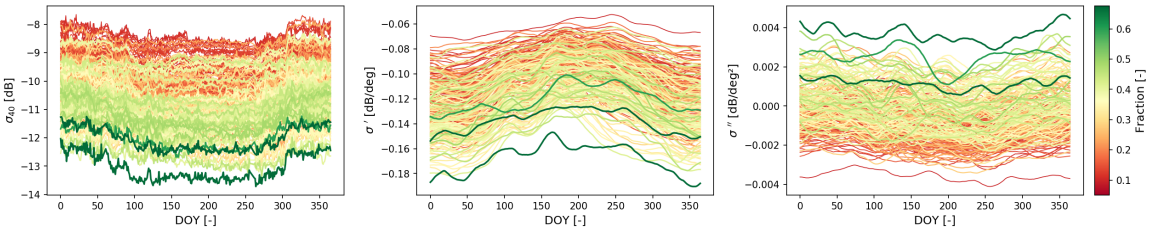
D.2.2 Class 45: Bare rock

The grid points and backscatter signatures for points with a bare rock fraction of $p_{45} > 0.01$ are plotted in Fig. D.10a and Fig. D.10c, respectively, colored by their respective p_{45} value. In general, higher values of p_{45} correspond with higher σ°_{40} and steeper σ', both with increased seasonal variation. There also seems to be a relationship between p_{45} and σ''; the moment of both maximum and minimum σ'' seem to occur later in the year for larger values of p_{45}.

Bare rock characteristically causes larger backscatter than vegetation at $\theta = 40°$, explaining why σ°_{40} increases when p_{45} increases [66]. As can be seen in Fig. A.2k, bare rock exclusively occurs in the Alps, the Pyrenees, and central Corsica, implying that p_{45} is related to elevation. This explains why lower σ' values are observed for increasing p_{45} values; as elevation increases, environmental factors (e.g. temperature, precipitation, humidity) become increasingly unfavorable for vegetation, which results in decreasing vegetation cover (i.e. lower σ'), as vegetation is eventually replaced by bare rock (i.e. higher p_{45}). Moreover, the composition of vegetation changes with elevation; as elevation increases, deciduous forest is gradually replaced by coniferous forest, which in turn is replaced by natural grasslands (see Fig. D.10b), until eventually all vegetation is replaced by bare rock and snow. Different scattering mechanisms occur when land cover composition changes, explaining why the behavior of σ'' changes for increasing fractions of bare rock.



(a) Selected grid points, colored by class 51 fraction



(b) σ°_{40}, σ', σ'' seasonal signatures, colored by class 51 fraction

Fig. D.11: Influence of class 51 (water bodies) on backscatter signatures

D.2.3 Class 51: Water bodies

The grid points and backscatter signatures for points with a fraction of water bodies $p_{51} > 0.05$ are plotted in Fig. D.11a and Fig. D.11b, where each grid point is colored by its respective p_{51} value. Higher values of p_{51} correspond with lower σ°_{40}, lower σ', and higher σ''. Water acts as a smooth surface and results in low backscatter, which explains why σ°_{40} decreases for increasing p_{51}. Furthermore, larger values of p_{51} inherently mean that vegetation cover fractions must decrease, which translates to lower average vegetation density and hence, decreasing σ'. Since water acts as a smooth surface, surface scattering becomes the dominant scattering mechanism as p_{51} increases and vegetation cover fractions decrease; this corresponds with the observed increasing σ'' values. The fact that slight seasonal behavior is present in σ°_{40}, σ', and σ'' for all values of p_{51} is explained by the fact that the maximum observed value of p_{51} is approximately 0.65; there exist no grid points with a land cover footprint consisting entirely of water. Instead, vegetation is always present and hence, seasonal σ°_{40}, σ', and σ'' behavior is observed even for high values of p_{51}.