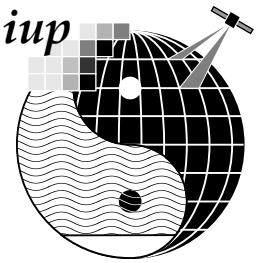



Institute of Environmental Physics
University of Bremen


# Automated Detection of Canola/Rapeseed Cultivation from Space

Application of new Algorithms for the Identification of Agricultural Plants with Multispectral Satellite Data on the Example of Canola Cultivation

Dissertation
Submitted for a degree of
Doctor in Natural Sciences (Dr. rer. nat.)
of the University of Bremen

presented by
Dipl. Phys. Hendrik Oliver Arp Laue

September 2004

Institut für Umweltphysik
Universität Bremen

# Automatische Detektion des Rapsanbaus aus dem Weltraum

Anwendung von neuen Algorithmen zur automatischen Identifikation von landwirtschaftlichen Anbauflächen mit multispektralen Satellitendaten am Beispiel von Raps

Dissertation zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

präsentiert von
Dipl. Phys. Hendrik Oliver Arp Laue

September 2004

Erster Gutachter: Prof. Dr. phil. Klaus Künzi
Zweiter Gutachter: Dr. habil Broder Breckling

Eingereicht am: 23. September 2004
Tag des Promotionskolloquiums: 25. Oktober 2004

# Abstract

The advances in biotechnology allow the use of genetically modified plants in agriculture. Whereas in the EU, this is still limited to experimental sowing, it is already practised commercially in Argentinia, Canada, China and the USA. The dispersal resulting from such cultivation holds risks that are difficult to assess. In the joint research project "*Generische Erfassung und Extrapolation der Rapsausbreitung*" (Generic analysis and extrapolation of oilseed rape dispersal, GenEERA) funded by the *Bundesministerium für Bildung und Forschung* (Federal Ministry of Education and Research, BMBF) the hybridisation and dispersal of canola and its wild relatives is investigated exemplarily. In this context the situation of canola cultivation, described by the mean field size and the mean minimum distance between canola fields, is of particular interest, especially since canola fields are potential sources of the transfer of new genes to non-modified or related plants. The aim of this work, which is part of the GenEERA project, is the identification of canola cultivation areas in northern Germany in the studied period from 1995 to 2002. The sizes of the fields and the investigation area pose requirements on the satellite data best met the LANDSAT Thematic Mapper (TM)/Enhanced Thematic Mapper+ (ETM+) and the Indian Remote Sensing Satellite (IRS) Linear Imaging Self Scanner/3 (LISS/3) sensors which allow to detect the individual fields. Complete coverage of the investigation area requires about 12 TM/ETM+ images. Considering the period of 7 years, and although only 47 images could be obtained for this study due to cloud cover, the amount of data to be processed is very large (14 GB). Therefore one focus of this work is the autonomous processing of the satellite data.

The processing of the data is performed in several steps: The first processing step is the georectification, assigning map positions to the satellite pixels. The georectification is done by a passpoint correlation. An even more accurate assignment is necessary between the pixels of different satellite images in order to allow an automated selection of training data sets from overlapping images. This is accomplished in an additional correction step, based on the correlation of image clips. The next processing step is the identification of clouds and their shadows. Opaque clouds can be identified by their brightness and low top temperature. The cloud shadows are identified by looking for dark patches near clouds that are located in the opposite direction of the sun azimuth angle. Thin clouds are identified based on the haze optimized transform (HOT) method which had to be adapted in order to compensate for the high albedo

of flowering canola. The effect of thin clouds can be compensated to some extent by a histogram-based method. The third processing step, the classification, is performed by the Mahalanobis distance classifier (MDC) because it only requires training data for one single surface type. Since the MDC is not as accurate as the commonly used maximum likelyhood classifier (MLC), its accuracy is enhanced by a segmentation of the MDC result used to identify single wrongly identified pixels and to perform region growing to include pixels missed by the MDC.

The resulting segments are approximated by rectangles of equal orientation and area which allow a vectorised data representation and a simple evaluation of the field distances and several other parameters of interest for the dispersal of canola pollen. Furthermore, the results are used to produce statistics of the complete investigation area allowing to investigate derived parameters on the situation of canola cultivation in northern Germany. The results of the classification are compared to validation data, i.e., edges and positions of known canola fields and agricultural statistics for 1995 and 1999. This validation showed that the total acreage of canola is identified with 70 to 90% accuracy, whereas larger fields are identified more accurately because of the lower ratio of border to inner field pixels. The accuracy also depends on the strength of the flowering of canola, which causes overestimation of the field size due to the high brightness of the flowers.

The methods presented in this work show that an automated classification based on a large number of satellite images with good accuracy is possible. The georectification and the cloud identification can easily be adapted to the classification of other agricultural crops, like e.g, maize or cereals. However, the actual classification of these crops might require to select satellite images from different acquisition times and to adapt the classification algorithm.

# Zusammenfassung

Der biotechnologische Fortschritt hat den Einsatz von gentechnisch modifizierten Pflanzen in der Landwirtschaft ermöglicht. In der EU beschränkt sich der Anbau zur Zeit noch auf experimentelle Aussaaten, in Argentinien, China, Kanada und den USA wird er jedoch bereits kommerziell betrieben. Die damit verbundene Freisetzung von gentechnisch modifizierten Pflanzen birgt allerdings auch schwer abzuschätzende Risiken. Im vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Verbundprojekt „Generische Erfassungs- und Extrapolationsmethoden der Rapsausbreitung“ (GenEERA) wurden die Hybridisierungs- und Ausbreitungsdynamik von Raps und verwandten Wildarten exemplarisch untersucht. Hierzu ist die Anbausituation von Raps, beschrieben durch die Anbaudichte, mittlere Feldgröße und den minimalen Abstand zwischen den Feldern von besonderem Interesse, da Rapsfelder potentielle Quellen für den Transfer von neuen genetischen Eigenschaften zu nicht modifiziertem Raps oder zu verwandten Pflanzen darstellen.

Ziel dieser Arbeit, die im Rahmen eines Teilprojektes von GenEERA angefertigt wurde, ist daher die Identifizierung und Charakterisierung der Rapsanbauflächen in Norddeutschland mit Satellitendaten im Zeitraum von 1995 bis 2002. Die Identifikation von Rapsfeldern stellt Mindestanforderungen an die Auflösung und Abdeckung der Satellitendaten. Diese werden am besten von den Daten der LANDSAT Thematic Mapper (TM)/Enhanced Thematic Mapper+ (ETM+) und der Indian Remote Sensing Satellite (IRS) Linear Imaging Self Scanner/3 (LISS/3) Sensoren erfüllt, da sie eine Identifikation einzelner Felder ermöglichen. Allerdings erfordert die räumliche Abdeckung des Untersuchungsgebietes mindestens 12 TM-Bilder. Zieht man den Untersuchungszeitraum von sieben Jahren in Betracht, so ergibt sich, obwohl die Anzahl der verfügbaren Daten auf 47 Bilder beschränkt war, eine sehr große Datenmenge (14 GB). Ein Schwerpunkt dieser Arbeit ist daher die möglichst automatische Verarbeitung der Satellitendaten.

Der erste Verarbeitungsschritt ist die Georektifizierung, die einzelnen Pixeln im Satellitenbilder mittels einer Passpunktskorrektur Koordinaten auf einer Landkarte zuordnet. Eine genauere Zuordnung ist zwischen Pixeln aus unterschiedlichen Satellitenbildern notwendig, um eine automatische Auswahl von Trainingsdatensätzen aus überlappenden Satellitenbildern zu ermöglichen. Dies wird durch eine Korrektur der Georektifizierung mittels einer Bildkorrelation erreicht. Der nächste Verarbeitungschritt ist die Identifizierung von Wolken und deren Schatten. Undurchsichtige Wolken können durch ihre große

Albedo in den sichtbaren Wellenlängen und die geringe Temperatur an ihrer Obergrenze identifiziert werden. Wolkenschatten werden zunächst über dunkle Flecken in der Nähe von Wolken im Satellitenbild erkannt; als weiteres dient die Tatsache, daß solche dunklen Flecken in der entgegengesetzten Richtung zum Sonnenazimuthwinkel der Wolke liegen. Dünne Wolken werden durch die haze optimized transform (HOT) Methode identifiziert, welche allerdings an die hohe Albedo von blühendem Raps angepasst werden muß. Der Einfluss dünner Wolken konnte durch den Einsatz eines Histogrammvergleiches teilweise korrigiert werden. Der dritte Verarbeitungschritt, die Klassifikation, wird mittels des Mahalanobis distance classifier (MDC) durchgeführt, da dieser nur den Trainingsdatensatz für einen einzigen Oberflächentyp benötigt. Allerdings erreicht der MDC nicht die Genauigkeit des üblicherweise verwendeten maximum likelyhood classifier (MLC) durch die Segmentierung des MDC Ergebnisses kompensiert, indem einzelne falsch klassifizierte Pixel entfernt und Pixel am Rand von Segmenten durch ein Region-Growing Verfahren auf ihre Zugehörigkeit zum Segment bzw. zum Feld überprüft werden.

Die Ergebnisse der Segmentierung werden durch Rechtecke mit gleicher Größe und Orientierung angenähert, was eine einfache Vektordarstellung der Klassifikation erlaubt, es aber auch ermöglicht, Parameter, die für die Ausbreitung von Rapspollen von Bedeutung sind (z.B. die Entfernungen zwischen den Feldern), einfach zu ermitteln. Des Weiteren wurden die Ergebnisse genutzt, um Statistiken für die abgeleiteten Parameter zu erstellen, welche die Anbausituation von Raps in Norddeutschland charakterisieren. Die Ergebnisse der Klassifikation wurden mit verschiedenen Validierungsdatensätzen verglichen, so z.B. mit den Rändern und Positionen von bekannten Rapsfeldern, und mit den Agrarstatistiken von 1995 und 1999. Diese Validierung ergab, dass die Gesamtanbaufläche von Raps mit einer Genauigkeit von 70 bis 90% erfasst werden konnte. Die Genauigkeit hängt dabei von der Größe und Struktur der Rapsfelder ab, da eine höhere Zahl von Randpixeln eine größere Unsicherheit für die Klassifikation bedeutet. Außerdem hängt die Genauigkeit von der Stärke der Rapsblüte ab, da die hohe Albedo der Rapsblüten eine Überschätzung der Anbaufläche zur Folge hat.

Die in dieser Arbeit vorgestellten Methoden zeigen, dass eine überwiegend automatische Klassifikation einer großen Anzahl von Satellitendaten mit guter Genauigkeit möglich ist. Die vorgestellten Verfahren zur Georektifizierung und Wolkenerkennung können problemlos auf die Klassifikation anderer Ackerfrüchte übertragen werden. Die Übertragung der eigentliche Klassifikation könnte die Auswahl anderer Satellitendaten und die Anpassung des Klassifikationsalgorithmus erfordern.

# Contents

# Chapter 1

# Introduction

Satellite-based remote sensing is a valuable tool to obtain various geophysical and biological parameters over large areas of the earth surface. In contrast to ground-based sampling, which is usually based on spot measurements, remote sensing is able to cover large areas rapidely. Moreover, depending on size and location of the investigation area, ground-based methods are frequently too time-consuming, too costly or for other reasons not possible to perform.

A good example for the advantages of satellite based remote sensing is the monitoring of the deforestation in the tropical rainforest in South America (Short, 2003), which is an important issue in the protection of biodiversity and would not be possible without the aid of satellite data. Another example is the measurement of global photosynthetically active radiation (PAR) (Ganapol et al., 1998; Myneni et al., 1997; Asrar, 1989; Alados et al., 1996). The PAR indicates the carbon dioxide assimilation by plants and algae, which is an important parameter for the greenhouse effect and is impossible to obtain from ground-based measurements.

## 1.1 Surface Type Identification

An important application field of remote sensing techniques is the identification of different ground surface types. These can either be different plant communities (e.g., coniferous forests (Hansen et al., 2001), grasslands (Tucker, 1985) or man-made structures (e.g., industrial or agricultural areas). Especially the identification of areas used for the cultivation of agricultural or horticultural plants is of interest. Such techniques may be employed to improve yield predictions (Allen, 1990; Bellow and Ozga, 1991; Genovese, 2004) or to efficiently control agricultural aid spending (Leo, 2004). Besides these commercial and administrative applications of surface type identification, the characteristics of crop cultivation is also used to investigate influences of crops on the surrounding fauna and flora. For instance, Wright (1994) used LANDSAT Thematic Mapper images to investigate poisoning of roe deer by oilseed rape in Scotland by identifying rape seed fields near to forests, i.e., habitats of deer.

## 1.2 GenEERA

The subject of this thesis too, is the influence of oilseed rape/canola[1] (*Brassica napus*) on surrounding ecosystems. Here, the influence of genetically modified (GM) canola on wild vegetation and on non-modified canola is to be investigated. Although remote sensing is a valuable tool to estimate the cultivation characteristics of agricultural crops, it does not allow to investigate the influences directly. Therefore, further information on pollen dispersal, meteorology, botany, agriculture and geography is necessary.

Consequently, this study is part of a interdisciplinary joint research project "*Generische Erfassung und Extrapolation der Rapsausbreitung*" (Generic analysis and extrapolation of oilseed rape dispersal, GenEERA) funded by the *Bundesministerium für Bildung und Forschung* (Federal Ministry of Education and Research, BMBF). GenEERA investigates the potential of GM canola with conventional canola crops, feral canola populations and potential hybridisation partners. The aims of GenEERA are necessary in order to understand the background of the various parameters obtained in this study. Therefore, a short outline of GenEERA is given here.

The type of potential influences and hazards caused by GM plants depends on the added genetic properties. The modification to canola already used for commercial cultivation is a resistance against acetolactate synthase-inhibiting herbicides (Rieger et al., 2002). A herbicide resistance allows the GM plant to survive the application of a specific herbicide. The herbicide kills all or the majority of other plants and lets the modified plant mostly unharmed. Therefore, the GM plant can grow on the field without weedy competitors which increases the yield for this crop. There are two worries concerning the herbicide resistance:

**Contamination:** A large number of consumers prefer food produced from non-modified plants, thus the non-modified crop has an economical advantage. Thus, the breeding of modified and non-modified plants is an economical threat to the farmers growing conventional plants if the new gene moves to their seeds. Moreover, the new gene is usually patented and the farmer can be forced to pay licence fees for their crops. There already has been a precedent-setting in Canada (Simon, 2004; Schimmeck, 2002).

**Hybridisation:** The herbicide resistance may be transferred to the feral relatives (see below) of the plant. This is important for the canola cultivation in Europe since a number of interbreeding partners are native in Europe, e.g., field mustard (*Brassica rapa*), Indian mustard (*Brassica juncea*) and wild radish (*Raphanus raphanistrum*). A more complete list of interbreeding partners can be found in Breckling et al. (2003). The

[1] Canola are special breeds of rapeseed. Originally, Canola is a trademark for edible oil from oilseed rape (CANadian Oil – Low Acid). Canola is the most common rape breed today and since the name is less ambiguous than "rape" it will be used here instead.

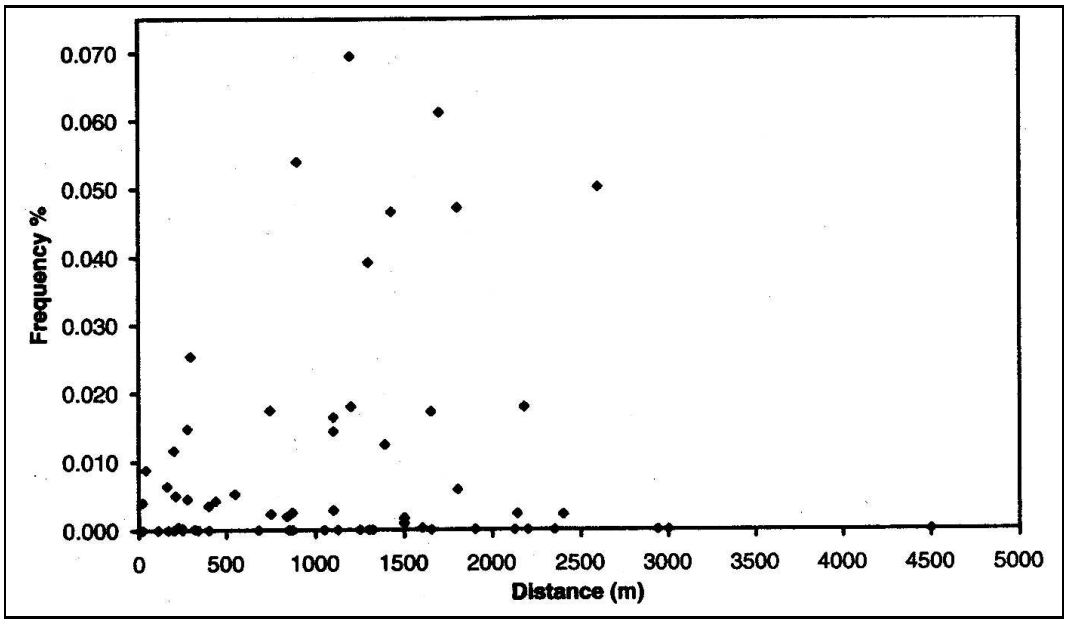



Figure 1.1: Pollen transfer distance investigated with genetically modified canola. Shown is the percentage of herbicide-resistant individuals in seeds from non-resistant canola plotted over the distance to a field with cultivated herbicide-resistant canola (adapted from Rieger et al., 2002). Note the hardly visible point at 4500 m, which is the maximum distance found in this study.

hybridisation[2] is especially important since the majority of relatives of canola are common weeds, which are the original target of the herbicide, i.e., these weeds might acquire a herbicide-resistance as an unwanted side-effect.

Consequently, the main problems can arise from the transfer of the new gene to non-modified canola and to the interbreeding partners of canola. Therefore, an important parameter is the distance by which canola pollen is transported. Recently, Rieger et al. (2002) investigated the distribution of canola pollen. The results displayed in Figure 1.1 show that the pollen is transported to a distance of up to 4.5 km. Further information on the transport range of canola pollen can be found in Breckling et al. (2003). In combination with information on the cultivation characteristics obtained from the satellite data, like e.g., the mean field distances, this information can be used to quantify the gene flow from modified to unmodified canola for different regions.

The estimation of hybridisation can be supported by satellite data. The potential gene flow from canola to turnip rape (*Brassica rapa*) and wild cabbage (*Brassica oleracea*) has been estimated by Davenport et al. (2000) and Wilkinson et al. (2000), using a LANDSAT TM (TM) and a Indian Remote Sensing Satellite (IRS))/ Linear Imaging Self Scanner/3 (LISS/3) image to identified canola fields nearby riverbanks and in coastal regions, potential habitats of

[2]Hybridisation is the process of interbreeding plants of two different species.

Figure 1.2: Volunteer plant in a wheat field located south of Bremen (adapted from Menzel et al., 2003).

these two interbreeding partners. The study of Davenport et al. (2000) only covers two interbreeding partners and one principal type of habitat. But since the interbreeding partners are also growing in other habitats, it is necessary to extend this study to other interbreeding partners and habitats. In GenEERA, the satellite data are used for a comparison with botanical population maps of various interbreeding partners.

Moreover, hybridisation depends on the flowering date of the different plants. Since the flowering period of agriculturally cultivated canola usually does not overlap with that of many potential interbreeding partners, a number of hybridisations are unlikely to occur. Nonetheless, interbreeding is still possible since wild growing canola is flowering outside of the main flowering period (see Schlink, 1994). Therefore, potential locations of wild growing canola, like grass stripes along roads, railroads, sand pits or building lots, also have to be identify (see Menzel et al., 2003). A particular case of "wild" growing canola are volunteer canola plants[3] on agricultural fields which are quite common, since canola seeds can remain in the ground for several years before germinating (see Schlink, 1994). They generally appear in cereal fields (see Figure 1.2). The number of volunteer canola plants on a field depends on the cultivation frequency of canola. Therefore it is important to investigate also the frequency of canola cultivation in the crop rotation cycle.

The appraisal of the different types of dispersal requires a number of pa-

[3]Generally, a volunteer plant is one that grows in unexpected places. In the context of this work, a volunteer plant is a canola plant that is growing within a field of other crops, e.g., cereals.

rameters, e.g., prevailing winds or seed persistence in the ground.

## 1.3 Parameters Retrievable with Satellite Remote Sensing

Satellite remote sensing can provide the area contiguously covered with canola plants. The identification is limited to populations of sufficient extent. The minimum size thereby depends on the sensor's spatial resolution, the typical spatial resolution being metres to hundreds of metres. Thus, it is not possible to identify single or small colonies of plants. This limits the use of remote sensing to the identification of agriculturally cultivated canola, i.e., canola fields. This is sufficient since the majority of canola plants are cultivated plants. Considering the mechanism of dispersal described above, the information about canola fields should be used to derive the following parameters:

**Cultivation density:** The cultivation density is the fraction of area covered by canola fields, i.e., it is a direct measure for the number of plants present in a region. This parameter is obviously linked to the pollen density or to the seed dispersal.

**Mean Minimum distance between canola fields:** The minimum distance is important for estimating the probability of pollen transfer between different fields. This is the most important parameter for the gene transfer from modified to conventional canola plants.

**Mean field size:** The field size gives information on the characteristics of cultivation in the investigation area. The spreading of pollen and seed outside canola fields is more likely in regions with a large number of small fields than in regions with fewer large fields.

**Length of field borders:** The length of field borders describes the contact region between cultivated canola and the surrounding vegetation, i.e., the direct neighbourhood of potential interbreeding partners.

**Frequency of canola cultivation:** Canola is usually cultivated within a crop rotation cycle. The frequency of canola cultivation is important for the appearance of volunteer plants.

## 1.4 Basics of Land Surface Remote Sensing

This section gives a brief overview of the physical properties of radiation and its interaction with the sensor, the surface and the atmosphere, mainly focused on reflection properties of plants and scattering of clouds for visible and infrared radiation.

### 1.4.1 Electromagnetic Radiation

Remote Sensing is the measurement of object properties with a distant sensor. This is true for many physical measuring techniques, but in general, the term remote sensing refers to satellite and aircraft based methods. In most cases, the sensors detect electromagnetic radiation reflected or emitted by the observed object. Therefore the interaction of the surface and the atmosphere with electromagnetic radiation is important to understand the physical principles of remote sensing.

The whole spectrum of electromagnetic radiation is divided into the subranges x-rays, ultraviolet, visible light, infrared, microwaves and radiowaves. The parts of the electromagnetic spectrum mainly used in this study are visible (VIS, 0.4–0.7 $\mu$m), near infrared (NIR, 0.7–1.1 $\mu$m) and middle infrared (MIR, 1.1–5.0 $\mu$m). The source of this radiation is the sun. The radiation emitted by the sun can be described by *Planck's blackbody equation* for a blackbody with a temperature of about 5,900 K. The wavelength with the strongest emission can be calculated from *Wien's displacement law* and is 0.49 $\mu$m, corresponding to the colour green (Schowengerdt, 1997; Elachi, 1987).

Another part of the spectrum used in this study is the thermal infrared (TIR, 8–14 $\mu$m ). The earth surface, which has a typical temperature of 300 K emits mainly at these wavelengths, with a maximum at 9.66 $\mu$m. Additionally, since the majority of earth surface types have a very low albedo in the TIR, they can be considered black bodies. Thus, the measured radiation intensity can be directly converted into temperatures. This is useful to identify clouds that have a lower temperature than the surface in mid-latitudes in the non-winter seasons.

The satellite sensor measures radiation reflected or emitted by the earth surface. The sensor usually has a number of channels, where each channel integrates the energy of the radiation over a defined range of wavelength called band and a defined viewing angle range called instantaneous field of view (IFOV). The area corresponding to this viewing angle and IFOV on the earth surface is called a pixel and all pixels constitute a satellite image like the one in Figure 3.3 (p. 46).

#### Reflectance Properties of the Earth Surface

The reflectance properties of earth surfaces are described by the bidirectional scattering coefficient which depends on the illumination direction, viewing angle and wavelength (Asrar, 1989, Chapter 4). In satellite remote sensing the

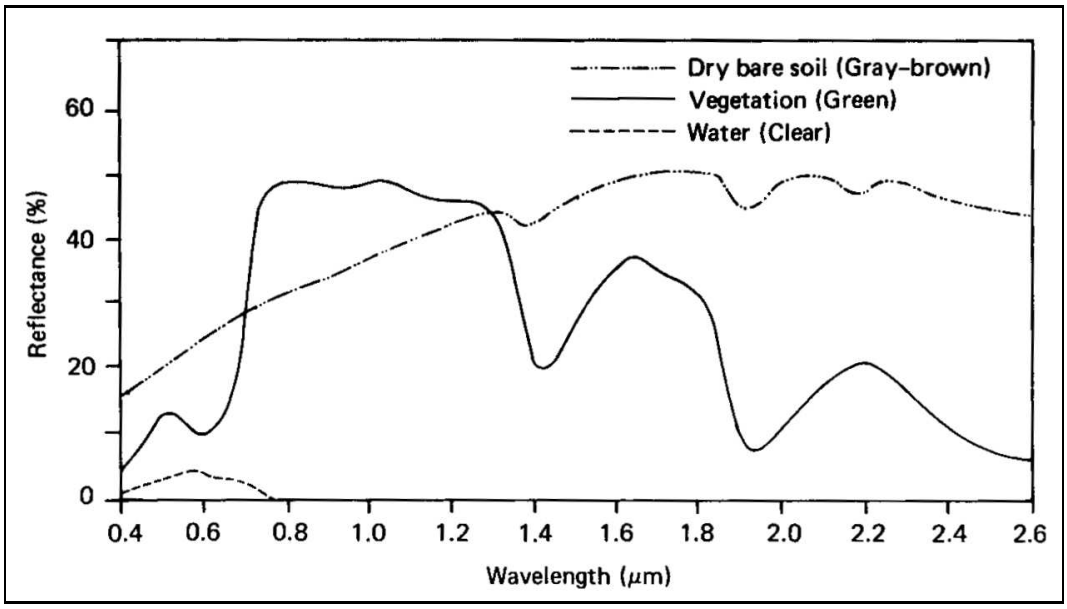



Figure 1.3: Typical reflectances in dependence of wavelength for vegetation, soil and water. These distinct spectra allow an easy discrimination of these surface types (from Lillesand, 2000).

bidirectional scattering coefficient is replaced by a reflectivity $\sigma_i$ or albedo for each individual channel $i$. The angular dependence of the reflectivity can be neglected for a satellite image since the viewing angles of the multispectral sensors only cover a small range ($\pm 7.2\,^\circ$ for LANDSAT TM, Kramer, 1996) and the sun direction is constant for a single image because it is acquired in a very short period ($< 2\,\mathrm{min}$).

The satellite measurements are only useful to distinguish different surface types if these have different reflectances in some channels. For instance, the surface types water, bare soil, sand and vegetation, can easily be distinguished since their reflectances are very unique for the various wavelengths in the visible (VIS), near infrared (NIR) and middle infrared (MIR) (see Figure 1.3).

The reflectance of plants depend on a large number of parameters, e.g., growth stage, water content and health state. Especially, photosynthesis is a complicated process (Hall and Rao, 1999) influenced by the plant species, temperature, time of day or even past illumination conditions (Hall and Rao, 1999; Gates et al., 1965).

Therefore, it is not possible to obtain a "spectral signature" of plants as it is possible for minerals (Vincent, 1997; Asrar, 1989; Price, 1994). Identification of plant types with remote sensing is thus generally based on the reflectance of plants that grew under comparable conditions, i.e., sample spectra generally have to be obtained from the satellite data itself (Lillesand, 2000).

Nonetheless, it still has to be discussed if the different crops are distinguishable in the satellite data.

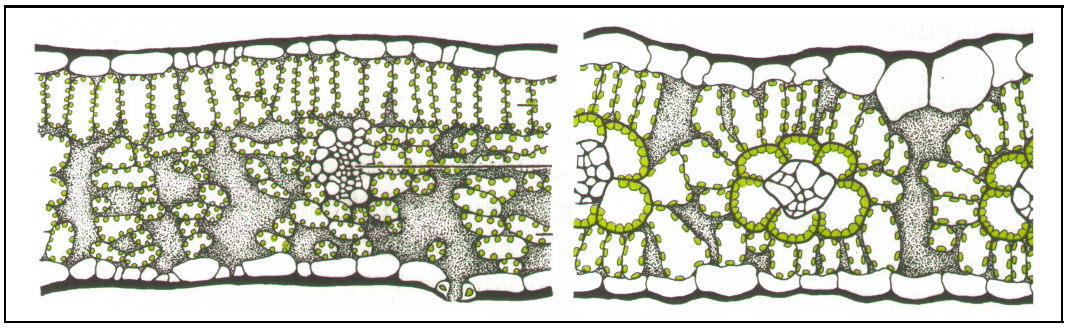


Figure 1.4: Leaf structures for two of the three basic photosynthesis apparatuses. Left: Cross section of a hellebore leaf displayed exemplarily for $C_3$ plants. Right: Similar cross section for maize as an example for $C_4$ plants.(adapted from Schopfer and Brennicke, 1999).

### Plant Leaf Reflectance

There are three different mechanisms that are mainly responsible for the reflectance properties of plant leaves:

**Absorption by leaf pigments:** The absorption of light by plants in the VIS results mostly from photosynthesis. There are a number of pigments involved in photosynthesis, mostly chlorophyll-a, chlorophyll-b and carotenoids (Zwiggelaar, 1998; Blackburn, 1998), which are responsible for converting water ($H_2O$) and carbondioxide ($CO_2$) to carbonhydrates with the aid of a photon with the energy $h\nu$. Simplified, photosynthesis can be described by:

$$6CO_2 + 6H_2O \xrightarrow[\text{photosynthesis}]{h\nu} C_6H_{12}O_6 + 3H_2O + 6O_2$$

Since only photons in the VIS have enough energy for these processes, these pigments mostly absorb in this part of the spectrum (Hall and Rao, 1999). In the VIS, the leaf pigments responsible for the photosynthesis are the main source of variation in the reflectance. The leaf pigmentation can be different depending on the type of photosynthesis and the plant species (Gates et al., 1965; Blackburn, 1998, 1999). Nonetheless, the leaf pigmentation is not sufficient to distinguish different plants from each other since they are too similar (Gitelson and Merzlyak, 1997).

**Scattering by leaf structure:** The incoming radiation is scattered by the leaf structures, e.g., cell walls, air-water interfaces, chloroplasts and mitochondria, since they have sizes in the dimension of the wavelength (Zwiggelaar, 1998; Lillesand, 2000; Sims and Gamon, 2002; Gates et al., 1965). Because of the high absorption in the VIS and MIR this is most important for the NIR radiation. Thus, plants have a high reflectance in the NIR which depends on the leaf internal structure. An example for

two different structures is given in Figure 1.4. Since the cell structure can be different for the various plants, these wavelengths pose a possibility to distinguish different species.

**Water absorption:** Plant leaves predominantly consist of water and thus the absoption of water is important. Additionally, proteins, cellulose and lignin show increased absorption for wavelength longer than $2\,\mu$m (Asrar, 1989).

The above discussion showed that the reflectance properties of leaves potentially allow a discrimination of plants. Nonetheless, field crops are usually quite similar plants concerning the cell structure and photosynthesis, e.g., canola and wheat both are plants with a $C_3$ type of photosynthesis (Hall and Rao, 1999; Schopfer and Brennicke, 1999) and are likely to have similar pigmentation and VIS reflectances. This is probably also true for the cell structures and the water content and would result in similar NIR and MIR reflectances.

Therefore, the spectral properties of leaves are not distinct enough do separate the different plant. This has been investigated and confirmed by Zwiggelaar (1998), who tried to distinguish weeds from crops by their leaf reflectance.

### Plant Cover Reflection

The above discussion only takes the releflectance properties of plant leaves into account. Since light is scattered within the plant cover depending on the leaf density, leaf form or leaf angular distribution (Asrar, 1989), the plant cover reflectance properties therefore can give further indication on the crop grown on the field.

**NIR Radiation:** The wavelength range with the strongest influence of scattering by the plant cover density and structure is the NIR. The internal leaf structure scatters this radiation (see above). Therefore, the reflectance propeties of the plant cover depend on the leaf density. In dense and high vegetatation covers VIS is frequently scattered by the leaf and other plant structures. This results in a diffuse reflection of the complete plant cover, which therefore appears bright.

In thin vegetation, the scattering is much weaker and due to the high transmitans of leafes for this radiation, a larger part of radiation is then absorbed by the underlying surface, which has a low reflectance for these wavelengths (see Figure 1.3).

Therefore, thicker vegetation cover appears brighter in the NIR as thin ones, this feature is especially valuable to identify canola, since it is a quick and high growing crop.

**VIS Radiation:** The absorption of plant covers for the VIS reflectances is very high. Therefore, scatterig and consequently, the density of vegetation cover is not important for this spectral range. Nonetheless, for crops only

covering parts the underlying soil, the reflectance will allow to distinguish these crops from acreage with a closed vegetation cover because of the high VIS soil reflectance.

**MIR Radiation:** Similarly, the MIR is also absorbed strongly, predominently by water, but also by lignin and cellulose. This absorption depends on the amount of water present in vegetation and therefore reflects the amount of vegetation present.

#### Canola flowering

The above discussion is limited to the reflectance properties of green leaves. Especially for canola, another part of the plant becomes important for the reflectance: The petals of the flowering plant. The flowering is responsible for a change of reflectance for the plant cover and the reflectance for green and red is increased. This can be seen on the titlepage of this work. Obviously, these changes in reflectance ease the identification of canola fields.

Nonetheless, the flowering is only observalble in a short period from end of April to beginning of May and it is not guaranteed to have satellite data available from that period. Moreover, the flowering of canola is not homogeneous and can change from field to field and also within fields. Thus, the flowering is not a reliable parameter for the identification of canola fields, especially when performed over a large area in which canola fields with different phenological stages (i.e., strong flowering, weak flowering and non-flowering) are present.

#### Conclusion

The measurement of reflectances in the VIS, NIR and MIR probably allows to distinguish different agricultural crops. Nonetheless, this has to be tested with real satellite data since the spectral properties are unknown for the crops and also depends on a large number of parameters. Probably, all three types of radiation are necessary to identify canola, where VIS allows to distinguish vegetated from non-vegetated areas, the NIR allows to identify different types of vegetation and MIR allows to distinguish plant cover from water bodies.

Flowering of canola makes the reflectance properties of canola more unique but also increases the variations of spectral properties for all canola fields.

### 1.4.2 Atmospheric Influences

Satellite-based remote sensing measure radiation transmitted by the atmosphere. Therefore, this section will give a short overview over the physical processes influencing VIS, NIR and MIR radiation on its way from the ground to the satellite.

### Gaseous absorption and emission

The atmosphere consists mainly of oxygen, nitrogen, water vapor and carbon dioxide. These gases absorb radiation in dependence of their molecular structure. In the optical wavelength range, absorption is predominantly caused by electronical and vibrational transitions of the molecules quantum states which have discrete energy levels (Asrar, 1989). The width and strength of the absorption lines depend on temperature, pressure and molecule density profiles in the atmosphere. Oxygen, nitrogen and carbon dioxide are horizontally homogeneously distributed and contribute a constant factor to the radiation received by the sensor. This is not the case for water vapor which is highly variable in the atmosphere.

### Scattering

Two different types of scattering do occur in the atmosphere: Rayleigh and Mie scattering (Schowengerdt, 1997).

**Rayleigh Scattering** occurs with particles much smaller than the radiation wavelength. In the optical and infrared part of the spectrum, these are the air molecules. The intensity of Rayleigh scattering varies as $\lambda^{-4}$. Therefore, radiation of shorter wavelengths are scattered more strongly than longer wavelengths. Since the atmospheric composition is mostly constant, the Rayleigh scattering does not show much variation for different weather situations.

**Mie-Scattering** occurs with particles in the dimension of the wavelength. In the atmosphere such particels are ice crystals, small water droplets or dust(Asrar, 1989). Usually these particles are called aerosols, although this term is sometimes used only for non-water particles. The Mie scattering is constant over a longer range of wavelengths, which is the reason why clouds appear white. Nonetheless, Mie scattering is stronger for shorter wavelength, i.e., it decreases from VIS over NIR to MIR. The high concentration of aerosols can be found in clouds and thus clouds have also the greatest influence on the measurements of the satellite sensor.

### Conclusion

The most important scattering process is Mie scattering since it has a stronger influence on the radiation than the Rayleigh scattering and depends on the highly variable aerosol density. Considering the absorption of gases, the most important gas is water vapor since it is distributed inhomogeneously and strongly depends on the weather situation.

## 1.5 Objectives and Outline of this Thesis

### 1.5.1 Objectives and Requirements

The aim of this work is the identification of agricultural fields used for canola cultivation in northern Germany for the period of 1995 to 2002. The results of this investigation give an overview of the cultivation characteristics in different types of landscape, which is necessaty to estimate the potential dispersal of GM canola seeds or pollen.

This aim poses a major difficulty since the fields to be identified are small compared to the investigation area. The investigation area is too large to be covered by a single image from a sensor that allows to identify individual fields. Therefore, it is necessary to process a large number of satellite images. However, the preprocessing and classification of satellite data can be labour intensive: The atmospheric conditions are different for different images and the spectral reflectances are usually also different (see Cihlar et al., 2000). An automatic scheme needs to adapt to the different conditions.

Therefore, the main focus of this work is to automate the processing of data to the greatest possible extent.

### 1.5.2 Outline

This section gives a short overview on the structure of the thesis and briefly describes the content of the different chapters. The focus of this work lies on Chapters 3 and 4 since they describe general methods for the georectification, cloud-/cloud-shadow/haze detection and classification.

**Chapter 2 – Data selection and Description:** This chapter gives an overview of satellite data usable to identify agricultural crops and the most appropriate sensors to identify canola fields in northern Germany. Additionally, an overview of the sensor and the validation data is given.

**Chapter 3 – Preprocessing:** This chapter describes the preprocessing applied to the satellite images to allow a classification and mapping of the classification result: The satellite data is to be projected to a map (Georectification) and clouds are identified and corrected, if possible. The methods have been selected, optimised and automated to allow a mostly autonomous preprocessing.

**Chapter 4 – Classification:** In this chapter, the agricultural crops cultivated in northern Germany will be examined for their separability by multispectral sensors for an exemplary region. Based on this examination, an appropriate classification algorithm is selected and used to classify the remaining regions. Also described are the methods applied to obtain field information from the pixel-based classification.

**Chapter 5 – Results:** The various results of the classification are presented, with a main focus on the parameters important for the dispersal of GM canola. The presentation is followed by a validation with known canola fields and a comparison with agricultural statistics available for 1995 and 1999. The chapter is closed with a discussion of error sources for the classification.

**Chapter 6 – Summary and Outlook:** This chapter gives a final appraisal of the results of this study and on the methods applied for preprocessing and classification. Moreover, possible approaches for improvements of these methods are suggested and the applicability of these methods for other plants is discussed.

# Chapter 2

# Data Selection and Description

In this chapter, a description of the selection and technical properties of the appropriate satellite data will be presented. There are a number of space-borne remote sensing sensors available that are capable of distinguishing different crops and it is necessary to identify those most suitable for the canola acreage detection in Northern Germany. Subsequently, the most suitable sensors is described in detail.

Besides the satellite data, ground gathered data are also needed in order to identify training data sets for the classification and to validate the classification results. The available ground gathered data is described at the end of this chapter.

## 2.1 Requirements for Satellite Data

The objective of this project is to obtain statistics for the cultivation of canola in Northern Germany as a base for gene flow estimation between different canola fields and wild growing canola plants or akin plants like wild cabbage (*Brassica oleracea*), wild mustard (*Sinapis arvensis*) and turnip rape (*Brassica rapa*).

Turnip rape and mustard are also cultivated as forage or green manure. To estimate the gene flow from one field to another or from field to wild growing plants, some cultivation characteristics are required (see Section 1.2 (p. 3)).

The number of sensors can be restricted by these requirements to sensors with a spatial resolution below 1 km, since this resolution is far too coarse for the detection of typical fields in the investigation area. Table 2.1 lists the available sensors potentially suitable for agricultural crop detection. Two additional sensors, the *Haute Résolution Visible* (HRV) and High Resolution Visible - Infrared (HRVIR) on *Système Pour l'Observation de la Terre* (SPOT) 3 and 4 are not listed, because their characteristics are similar to those of the Advanced Spaceborne Thermal Emisssion and Radiation Radiometer (ASTER) which will be discussed instead. Moreover, there is only one synthetic aperture radar (SAR) sensor listed, the European Remote Sensing Satellite-SAR (ERS-SAR), since the other SAR-Systems, e.g., the Japanese Earth Resource

Table 2.1: List of sensors usable for crop identification. "Resolution" always applies to the spatial resolution of the multispectral data (in case there is also a panchromatic band present like for IKONOS and ETM+). "Coverage" refers to the number of necessary frames to cover the entire study area. The price per frame is the average price for a complete frame in July 2001. The information has been compiled from Kramer (1996); Space Imaging Eurasia (2003); Eurimage (2001); Euromap (2001).

| Sensor | MODIS | TM/ETM | LISS/3 | ERS SAR | ASTER | IKONOS |
|---|---|---|---|---|---|---|
| Satellite | AQUA/ TERRA | LAND-SAT 5/7 | IRS 1C/1D | ERS 1/2 | AQUA/ TERRA | IKONOS |
| resolution [m×m] | 500 × 500 | 30 × 30 | 23.5× 23.5 | 12.5× 12.5 | 15 × 15 | 4 × 4 |
| frame width [km] | 2330 | 183 | 142 | 102 | 70 | 12 |
| frames needed for coverage | 1 | 10 | 16 | 24 | 46 | 160 |
| repeat cycle [days] | 16 | 16 | 24-25 | 35 | 4-16 | 3-25 |
| available since year | 1999 | 1972 | 1995 | 1991 | 1999 | 2000 |
| price per frame € | 0 | 1000 | 2700 | 800 | 55 | 1,728 |
| price for coverage € | 0 | 10,000 | 43,200 | 30,000 | 2,530 | 276,480 |

Satellite-SAR (JERS-SAR) are similar in resolution, frequency and coverage to the ERS-SAR. Also not listed are panchromatic sensors, which usually have a better spatial resolution, but only one channel; however this type of sensor will be discussed later.

For this study, the sensors have to comply with the following requirements:

1. The spatial resolution has to be fine enough to detect most of the fields in the investigated area, since too coarse a resolution will lead to errors at the field boundaries and therefore to inaccuracies in the determined field sizes.

2. The spatial coverage must be large enough to cover most of the study area.

3. The sensor has to be appropriate to discriminate the spectral properties of canola plant covers from those of other plants.

4. Historic data must be available for an evaluation of canola cultivation over at least the past 5 years.

5. Since there is only a limited period of the year when canola can be discriminated from other field crops, the revisit time has to be short and the frame width wide enough to ensure several coverages of the area during this period. This is important for optical sensors because the investigation area is frequently covered by clouds.

6. As satellite data can be quite expensive, especially high resolution data, the price has also be taken in account, too.

The following sections will discuss the important sensor characteristics and conclude which sensors are most suitable.

### 2.1.1 Spatial Resolution and Coverage

High resolution sensors like IKONOS and LANDSAT usually have small frame size and longer revisit times, and sensors with lower resolution like the Moderate Resolution Imaging Spectroradiometer (MODIS) have better coverage. The spatial resolution is necessary to detect small fields ($\approx 2$ ha)[1] on the ground and a good coverage is important to get data for the complete area. Therefore it is necessary to find the best tradeoff between the minimum field size detectable and the maximum coverable area (see Table 2.1).

#### Minimum resolution

The common field size in different regions varies from less than 2 ha in central Europe to more than 350 ha in the Midwest of North America (Colwell, 1983, Chapter 21). The study area is the northern part of Germany. Here, the field size ranges from less than two hectare in the West to large fields with 200 ha in the East.

The sensor has to be capable to detect smaller fields also, in Northern Germany down to a minimum size of 2 ha. There have been some studies on sub-pixel methods that derive the acreage for a certain crop by using the spectral properties of different surface types for the calculation of the crop fraction within one pixel (Gross and Schott, 1998; Kerkes and Baum, 2002). This would allow to use sensors with a coarser resolution. However, since the size of individual fields is of special interest in this study and sub-pixel methods only deliver the total acreage of the canola within one pixel, these methods are not applicable for this study. Therefore, the spatial resolution of the sensor has to be fine enough to separate different fields. The available sensors have spatial resolutions ranging from 1 m (IKONOS) to 500 m (MODIS).

As an example for the effects of spatial resolution on the identification of field sizes, a comparison of an aerial photograph with TM and simulated MODIS images is shown in Figure 2.1. In the region displayed, the fields are

[1] The field sizes in this thesis are not stated in the SI-Unit $m^2$ but rather in ha since this is the common Unit used in agriculture in Germany. The relations between square metres, square kilometres and hectare is: 1 ha=10,000 $m^2$=0.01 $km^2$.

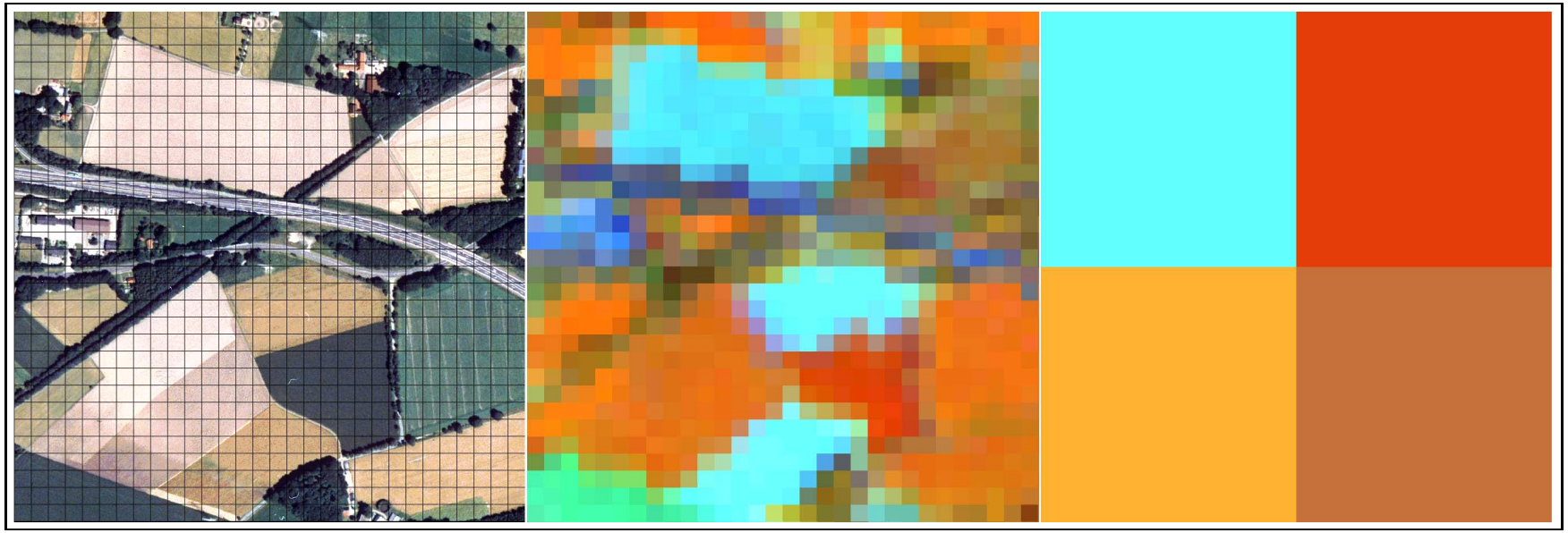


Figure 2.1: Comparison between an aerial photograph (left), TM (middle) and simulated MODIS data (right) west of Delmenhorst, near Bremen. The MODIS image was simulated by using TM data. These images demonstrate the impact of different sensor resolutions on the discrimination of agricultural fields. The field boundaries can easily be distinguished in the aerial photograph and also in the TM data, but not in the simulated MODIS data. The satellite images are displayed in false colour representation with near infrared, shortwave infrared and red as red, green and blue. This channel combination allows to distinguish between different crop types. The spatial resolution is one meter for the photograph, 30 m for the TM and 500 m for MODIS. The black grid lines on the aerial photograph indicates the spatial resolution of the TM data. Note that the images were acquired in different years and the field crops in the satellite images are not related with the field crops in the aerial photograph. Original data: LANDSAT ETM+ ©ESA, 2001. Distributed by Eurimage.

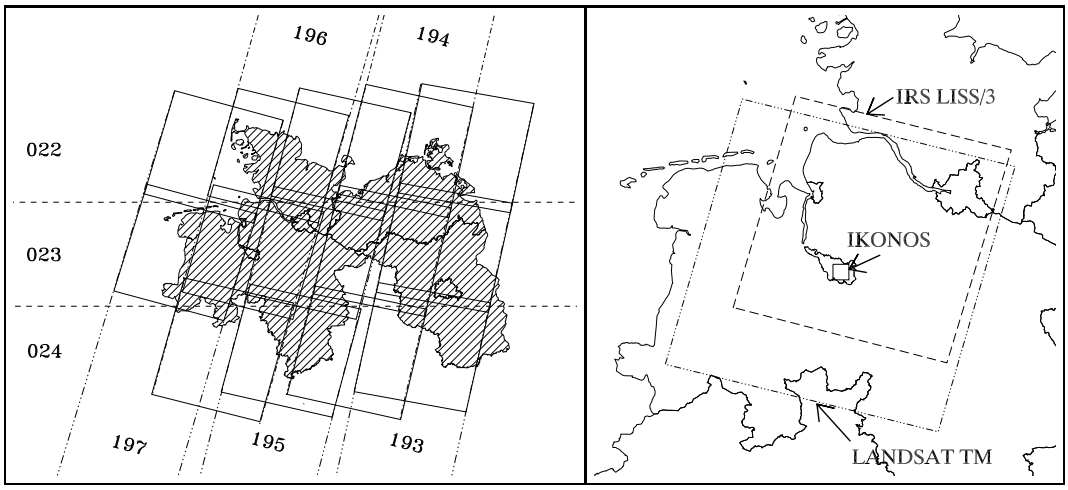



Figure 2.2: TM coverage for the study area (left). The numbers on the left represent the rows and the numbers on top and bottom the path numbers. The map shows the contours of the investigated federal states (*Bundesländer*): Niedersachsen, Bremen, Schleswig-Holstein, Mecklenburg-Vorpommern, Brandenburg and Berlin. The right figure shows an example for the TM, LISS/3 and IKONOS frame sizes.

small compared to the typical field size of other regions in the study area. The aerial photograph has a resolution of one meter. For better comparison, the resolution of the TM sensor is indicated by the black grid lines in the aerial photograph. In the TM data, the different fields can be separated from each other. The only problematic field is the small one directly above the highway in the right part of the image, but as the small field just below the farm can be easily identified, this is a problem of the grown crop in that year and not due to the sensor resolution. Nonetheless, fields of similar size can be identified and the resolution of TM data is therefore sufficient to detect the majority of fields. This is also true for the LISS/3 with comparable 23.5 m resolution, but not for MODIS: the right image in Figure 2.1 shows that the resolution is not sufficient to detect single fields. The MODIS resolution of 500 m was simulated by averaging the TM pixels with a resolution of 30 m. MODIS has two higher resolved channels with 250 m resolution, but this resolution is also too coarse to detect smaller fields, which can be seen in Figure 2.1.

The aerial photograph gives an idea about the resolution of IKONOS, which is 4 m for the multispectral bands, such that with classification, IKONOS would lead to the best results. The resolution of ASTER is 15 m and would also yield good results.

The conclusion is that with the exception of MODIS, all sensors listed in Table 2.1 are suitable to detect individual fields in the investigation area. Obviously, the finer the resolution, the more accurate are the results (Colwell, 1983, Chapter 21).

#### Frame coverage

The investigated area has a size of about 600 km by 500 km. A large frame size covering this area completely is desirable since only one frame needs to be processed. This is even more important for optical sensors because the investigation area is frequently hampered by clouds.

In Table 2.1 the sensors frame width and the frames needed for a complete coverage are listed. The widest frames, 2330 km, are provided by MODIS. Consequently, a single frame would be sufficient to cover the whole area. Higher resolved data have smaller frame sizes. The widest here is TM with 183 km, thus it is capable of covering the whole area with 10 frames (see Figure 2.2). In the same figure the frame sizes of LISS/3 and IKONOS are also displayed. Since the LISS/3 frame is only 143 km wide, 16 frames are necessary for one complete coverage. IKONOS has the smallest frame with a width of 12 km and would thus require about 160 frames.

Consequently, the best coverage is provided by MODIS, but TM and LISS/3 also have good coverage. ASTER and ERS-SAR frames are smaller and would need a greater number of frames. The frame size of IKONOS is far too small for the postulated coverage.

### 2.1.2 Spectral Information

The wavelength range used by the sensor is important because canola and the other surface types must be distinguishable in the satellite data. Sensors with spatial resolutions smaller than 1 km can be divided into three types:

- Multispectral sensors that have several bands in the visible and infrared spectral range.
- Panchromatic sensors that use the complete range of the visible and near infrared in order to achieve better spatial resolution.
- SAR that actively emits microwaves (i.e., wavelengths in the cm range) and records the backscattered signal.

In the following sections the spectral interactions for the different sensors will be discussed.

#### Multispectral sensors

As the name suggests, multispectral sensors have several bands in the spectral range of the VIS, NIR and MIR. For multispectral sensors, the range of wavelengths for all bands is usually from 0.4 to 2.2 $\mu$m. As discussed in Section 1.4.1 (p. 8) the reflectivity of plants in these spectral ranges depend on the pigmentation, cell and canopy structure of the plants. These parameters depend on the crop type cultivated on a field

All these parameters differ usually for different crop types and influence the reflectivity of the multispectral channels differently. Agricultural plants

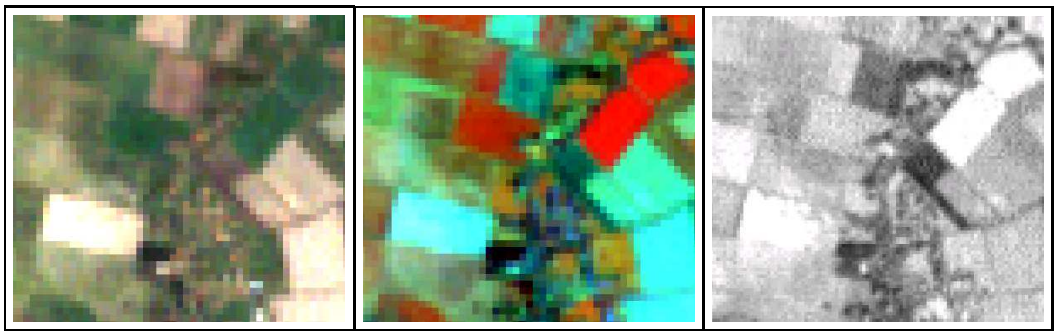


Figure 2.3: ETM+ panchromatic and multispectral data. Left: True colour image of channels 3 (red), 2 (green) and 1 (blue) with 30 m resolution. Middle: False colour image channels 4 (NIR), 5 (MIR) and 3 (red). Right: Panchromatic data image with a spatial resolution of 15 m. The displayed region is in the south-east of Bremen. The image was acquired on April 3, 2002. Different crops can be easily discriminated in the false-colour image: orange-red is canola and turquoise bare soil, respectively. This is not possible for the panchromatic image. Original data: LANDSAT ETM+ ©ESA, 2002. Distributed by Eurimage.

are generally cultivated in mono-cultures, therefore the leaf structure and the plant cover structure are mostly homogeneous over the area of a field and are useful parameters to distinguishing agricultural crops.

Adding up these properties, multispectral sensors have a good potential to discriminate different crop types. This has been confirmed in studies by Wright (1985, 1994); Davenport et al. (2000) who successfully detected canola by using the red, NIR and MIR channels of TM and LISS/3. All listed multispectral sensors have these channels and are therefore suitable for the detection of canola.

### Panchromatic sensors

The spectral range of panchromatic sensors is similar to those of the multispectral sensors, but these sensors integrate over a broader spectral range. This allows to improve the spatial resolution. In Figure 2.3 a comparison of multispectral and panchromatic data from ETM+ is shown. Here it is demonstrated that different crops can be distinguished in the multispectral images. This is only partly true for the panchromatic image, e.g., in the multispectral image the canola fields can easily be recognised by their orange-red colour in the false-colour image, whereas in the panchromatic image both fields appear in light-grey, which is also the case for other fields, e.g., the turquoise field (bare soil) on the left side of the false colour image. Therefore, it is difficult to distinguish canola from other surface types without the use of additional multispectral information. However, panchromatic data can be used to improve the spatial accuracy of multispectral classifications, which has been shown by Müschen et al. (2001).

#### Synthetic Aperture Radar

SAR can acquire data at day and night and also through clouds, which is a great advantage for the temporal coverage. The interaction of microwaves with plant cover can be found in Lillesand (2000). Because of the noise in SAR data, a pixel based classification is not possible. This problem can be circumvented by using the field boundaries from ground surveys and maps as shown by Michelson et al. (2000), who used ERS-SAR data to fill cloud-related gaps in TM data and for crop type classification, which worked well for different agricultural crops, also for canola. Such field boundary data are usually not available, especially for larger areas. This difficulty was overcome by Lopès et al. (1999) by overlaying four SAR frames from the same year to reduce the speckle noise.

Concluding on the spectral information, multispectral sensors provide the best ability for the distinguishing different field crops with one single image. SAR sensors may be an alternative but require accurate maps or multiple images, which would increase costs for data and algorithm development. Panchromatic data can be used to improve the spatial resolution of multispectral data (Müschen et al., 2001), but is not suitable for the discrimination of crops without additional information from multispectral sensors.

### 2.1.3 Temporal Requirements

In addition to the spatial coverage, the temporal coverage is also of importance. The sensor must have acquired data in the relevant period and the data have to be available from a data archive. Moreover, the frequency of data acquisition, which depends on the revisit time and the frame width (see Section 2.1, p. 16), must be taken into account. The revisit time can be shortened by using overlapping frames or by multiple sensors, e.g., there are two IRS satellites available since the end of 1997, and LANDSAT 5 and 7 were both operational from end of 1999 to early 2002.

#### Temporal coverage

Investigation of canola cultivation over the past years is one of the main objectives of this project because the crop on a field is usually changed every year. This is called crop rotation and is necessary since each crop needs different types of nutrients. The order of planted crops depends on the soil and on the field crops the farmer wishes to cultivate. Gene flow from genetically modified to the unmodified canola is possible by seed from previous years, since canola seed can persist for up to ten years in the ground and still be germinable (Cramer, 1990). Hence, to estimate the probability for a genetically modified canola plant to grow in a field of non-modified plants, it is necessary to get information on the repeat times for canola cultivation in the crop rotation cycle for different regions. Table 2.2 shows a list of common crop rotation cycles in northern Germany. The longest repeat time for canola is

four years. Thus, the observing time span for the satellite data should be at least four years. Table 2.1 shows the dates since when different satellite data are available. The sensors available for the required time span from 1997 to 2001 are TM, ERS-SAR and LISS/3. Newer sensors like MODIS, ASTER and IKONOS, launched in 1999, could be useful to fill gaps in later periods.

### Temporal resolution

In addition to the spectral separability that is described above, the development stages of the different agricultural plants are equally important to identify different crops . An overview of agricultural crops and their acreage is shown in Table 2.3. The main crops grown in Northern Germany are winter wheat, rye and winter sown canola. Obviously, the plant cover of a field has to be dense enough to be detectable by the sensor and the spectral signature of canola has to be distinguishable from other field crops by the sensor at that time.

There are two types of canola cultivation, winter-sown and spring-sown canola. Actually, winter-sown canola is sown in late summer. The plants grow vegetatively and flower the next spring. Since winter-sown canola brings higher yields for northern Germany, only 2.8 % of canola cultivation is spring-sown canola (Cramer, 1990). Mostly it is used when the winter-sown canola has been damaged by frost or drought. Hence, this work concentrates on winter-sown canola.

Figure 2.4 displays the period between sowing and harvesting of different types of agricultural crops in Belgium. Since Belgium is a neighbouring country with a comparable climate, these periods are also valid for Northern Germany. Figure 2.4 shows that most crops are sown in early May. This gives a good chance to distinguish these plants from winter-sown canola with satellite data acquired before or soon after sowing, since the plant cover in that time is sparse or non-existent. The remaining plants, wheat, barley and winter sown canola, have to be distinguished by their spectral properties since they have a similar cultivation period and it is necessary to consider the development of winter-sown canola in detail. The development of the plants is depending on the local climate and will be discussed for Northern Germany only.

Table 2.2: Common crop rotation cycles used in canola cultivation in Germany listed in order of frequency, (Cramer, 1990).

| | | | |
|---|---|---|---|
| **canola** | → wheat | → rye | |
| **canola** | → wheat | | |
| **canola** | → wheat | → peas | |
| **canola** | → wheat | → oat | → barley |

Table 2.3: Acreage of main agricultural crops for the federal states Niedersachsen (NS), Schleswig-Holstein (SH), Mecklenburg-Vorpommern (MV) and Brandenburg (BB) for 2001. The acreage of summer canola also includes the turnip rape acreage, Source: Saaten-Union GmbH, Isernhagen HB.

| Crop Acreage [1000ha] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Federal State | winter wheat | rye | winter canola | sugar beet | oat | summer wheat | peas | spring barley | summer canola |
| NS | 384.5 | 155.6 | 73.8 | 115.0 | 23.3 | 50.9 | 7.0 | 97.4 | 4.5 |
| SH | 193.0 | 33.5 | 89.0 | 13.3 | 9.0 | 1.7 | 1.9 | 12.8 | 0.3 |
| MV | 294.7 | 111.0 | 203.8 | 27.9 | 12.0 | 2.5 | 13.7 | 12.8 | 4.3 |
| BB | 128.4 | 253.2 | 95.3 | 11.3 | 15.8 | 3.1 | 24.0 | 9.7 | 4.3 |
| Total acreage | 1000.6 | 553.3 | 461.9 | 167.5 | 60.1 | 58.2 | 46.6 | 135.3 | 13.4 |

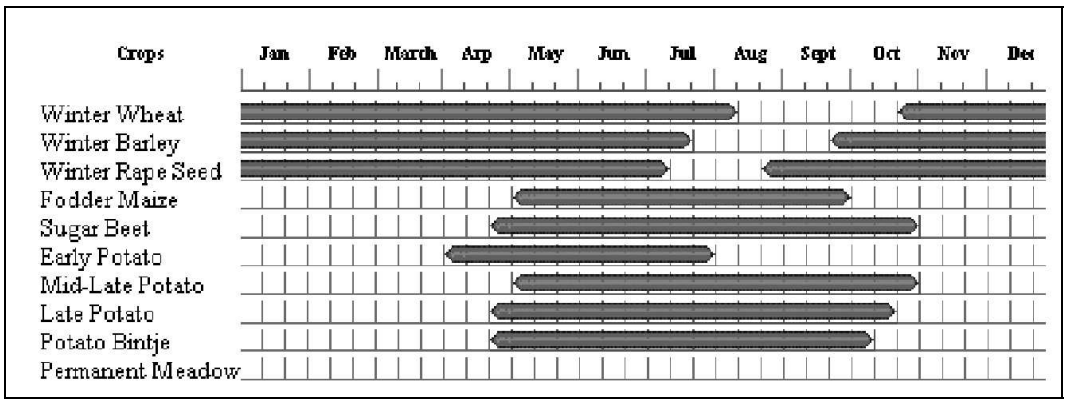



Figure 2.4: Period from sowing to harvesting from different field crops in Belgium, adapted from Blaes et al. (2001).

Figure 2.5: Difference between blooming (left clip) and not flowering (right clip) canola fields as seen by the satellite sensor; Both images display clips from the frame 196/023 in true colour representation of the TM/ETM+ channels 3, 2 and 1; the left image was acquired on May 10, 2000 and the right on April 3, 2002. The image clip displays the region South of Bremen. Original data: LANDSAT TM ©ESA, 2000 and LANDSAT ETM+ ©ESA, 2002. Distributed by Eurimage.

**The development stages of winter-sown canola.** Winter-sown canola is sown in late august and starts germinating soon afterwards. It grows until the end of the vegetation period in November, and continues growing at the beginning of the next vegetation period in March. The blooming period starts in late April or May and lasts for two to three weeks. At the end of July winter-sown canola starts to mature and is harvested at the beginning of August. The dates stated above depend on the weather conditions and can shift by as much as 10 days in either direction (Cramer, 1990).

**Best times for detection.** At the end of the vegetation period in November, the canola coverage of the soil is dense enough to be detected by satellite sensors. Problematic in that period is turnip rape, a close relative to canola that is sown as green manure and as a fodder plant. Turnip rape might be indistinguishable from canola, in fact, even at the ground they are difficult to distinguish. The discrimination of these plants is possible after the beginning of the vegetation period because canola grows higher than turnip rape. The best time for the detection of canola is the blooming phase when the fields appear in bright yellow in the true colour image (see Figure 2.5).

Taking all this into account, the best time for detecting winter-sown canola from satellite is in the period from early April to June, especially because of the blooming in May. Non blooming canola fields are detectable in that period due to the fast growth causing a higher plant cover in comparison to other agricultural crops. Data taken at the end of the vegetation period and during

wintertime might lead to confusion with turnip rape.

### 2.1.4 Costs of Satellite Data

The costs depend on the sensor, preprocessing applied by the data provider and years of the acquisition. The prices for images of satellite data are difficult to compare because of various discount options (primarily for TM-data).

#### Prices for Data of Different Sensors

In Table 2.1 (p. 16), the prices for data from different sensors are displayed. The price for TM data is an estimated average, since there are various discounting regulations (Eurimage, 2001).

IKONOS data are not ordered by frame, but rather by covered square kilometres. In July 2001 the price per square kilometre was €12, with a minimum order of €3000 for one image (Space Imaging Eurasia, 2003).

The data for the remaining sensors do not have complicated discount options and the prices in Table 2.1 are the exact prices. Comparing the prices necessary for a complete coverage, the most cost-effective sensor would be MODIS, because its data are free. Also quite inexpensive are ASTER and TM. The most costly data would be IKONOS data with more than a quarter million of Euro for a complete coverage of the investigation area.

Satellite data providers offer the satellite data on different processing levels. For a surcharge, the providers offer some further processing like passpoint or terrain correction or an atmospheric correction (see Section 3.1.3 (p. 39)). Further processing would increase the costs per frame, for example the surcharge for a pass point and terrain correction for TM data would be €450 per frame at Eurimage (Eurimage, 2001).

Ordering of passpoints or atmospheric corrected data would increase costs by about 50 % for Landsat data and therefore diminish the number of affordable frames by one third.

### 2.1.5 Conclusions

Summing up the qualities of sensors listed above, the following conclusion can be drawn:

- The best temporal and spatial coverage is provided by MODIS. A major drawback is the coarse resolution of 0.5 to 1.1 km that makes it impossible to detect smaller fields in the study area. There have been some studies on sub-pixel classification, but these cannot provide information on the size distribution of small fields.

- IKONOS has the best spatial resolution currently available for spaceborne sensors, but the number of needed frames and the resulting costs are too large. This is also true for airborne sensors, which have not been discussed here in detail.

- Most suitable from the technical point of view are TM and LISS/3 data. Taking into account the lower number of frames necessary and the lower cost, TM is preferable over LISS/3, although the latter is useful to fill gaps.

- ERS-SAR images have the advantage of being independent of cloud cover. The disadvantages are that a pixel based classification is not possible mainly resulting from the strong noise in the data (Michelson et al., 2000). With additional maps for the field boundaries a per field classification would be possible but this information was not available for this study. Another possibility is the use of multiple frames for the same region (Lopès et al., 1999), which would however increase the costs.

The selection of satellite data was done with search tools of satellite data providers like Eurimage and Euromap that provide quicklooks of the data, because the majority of the frames were cloud-covered during the period in question. Among the frames with little cloud cover over the study area, the ones with the date closest to the canola blooming where purchased. In Figure 2.6 the selected frames are shown. A total of 46 frames was purchased. An additional LISS/3 frame was bought for the region around Bremen in 1999. For validation, it was necessary to acquire two additional frames: one LISS/3 frame from 2002 for the comparison of the two sensors and another TM frame to evaluate images from early April.

## 2.2 Data Description

The appropriate sensors from the above discussion have been identified as TM/ETM+ and LISS/3. The characteristics of these sensors not listed in Table Section 2.1 (p. 16) are discussed here in detail. Another sensor, IKONOS is not used for the classification, but since its high spatial resolution is useful to investigate the influence of the resolution on the classification result, it will be discussed briefly.

Additionally the various ground survey data available for this study are presented. These data are essential for the selection of training data sets and the validation of the classification results. Described are also the cultivation statistics that are used for further validation of the classification results.

### 2.2.1 Satellite Data

All sensors discussed here are multispectral sensors that measure the radiance in the VIS, NIR and MIR part of the electromagnetic spectrum. All sensors orbit the earth on satellites with a near polar and sun synchronous orbit in order to obtain constant illumination conditions.

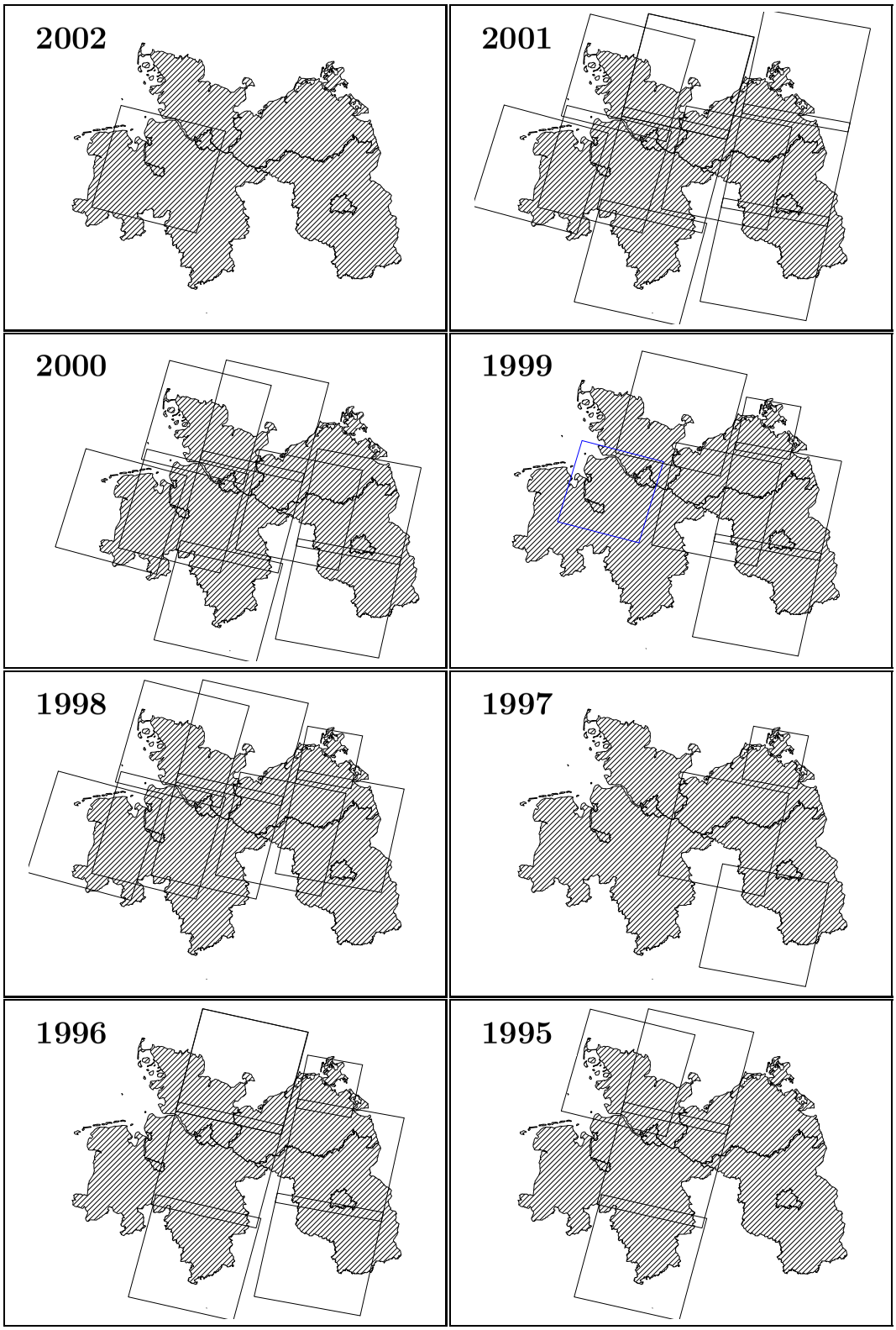



Figure 2.6: TM/ETM+ and LISS/3 coverage for the period from 1995 to 2002. A larger number of frames are available in 2000 and 2001, since LANDSAT 5 and 7 were both operational. The blue frame in 1999 is the one of the LISS/3 image used in this study.

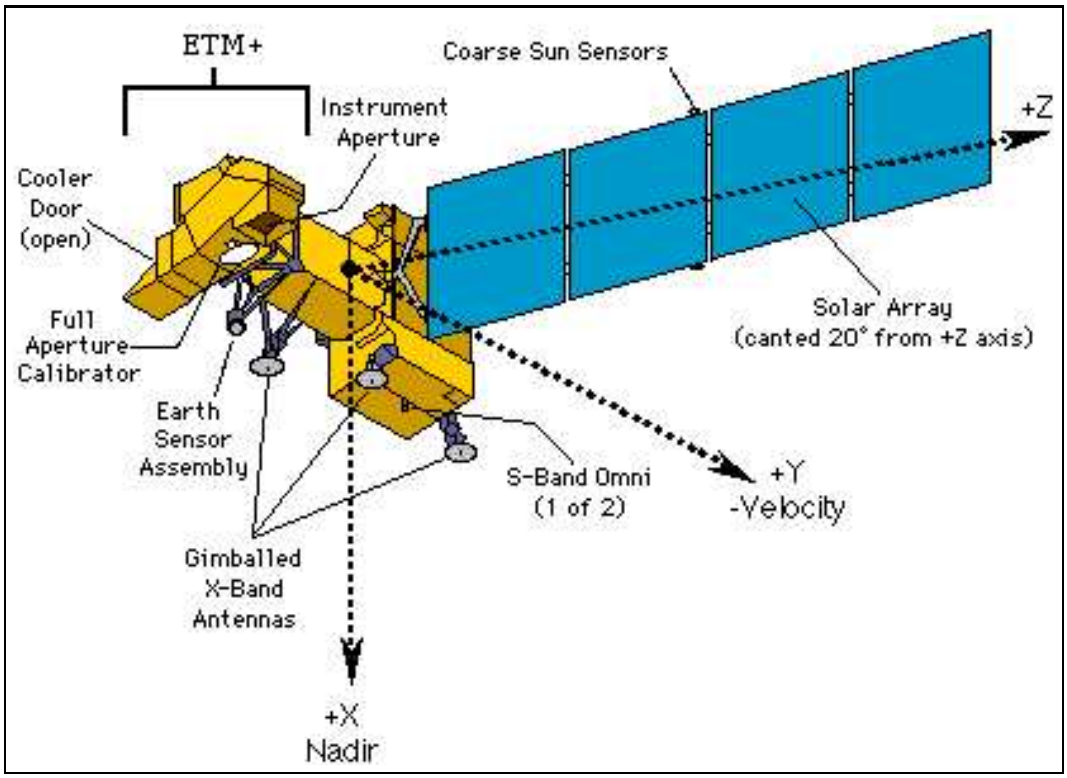



Figure 2.7: Sketch of the LANDSAT 7 satellite (from Irish, 2000).

### TM and ETM+

TM and ETM+ are mounted on the LANDSAT satellites. The first TM sensor was launched with the LANDSAT 4 satellite in March 1972. The successor on LANDSAT 5 was launched in March 1984. Since December 1999 an improved version, the ETM+ is available on LANDSAT 7. Currently, LANDSAT 5 and LANDSAT 7 are operational[2]. Both satellites orbit the earth at a height of 705 km. The orbit inclination is 98.1° and the period is 99 Min. The current equator crossing time is 10:45 local time (Kramer, 1996).

TM has seven channels ranging from VIS to thermal infrared (TIR) (see Table 2.4 and Figure 2.7). The spatial resolution is 30 m for all channels except channel six which has a resolution of 60 m (see Table 2.4).

ETM+ has the same six channels as TM, plus an additional panchromatic channel with a spatial resolution of 15 m. Figure 2.7 shows a sketch of the LANDSAT 7 satellite. The positions of the TM wavelength range for different channels overlaid over a generalised reflectance spectrum of plant cover is shown in Figure 2.8. From the discussion in Section 1.4.1 (p. 8) and this figure it can be seen that TM covers most of the useful wavelength to gather information on plants and other land surface types.

[2]There have been some problems with ETM+ in March 2003 and it is to be seen if this can be fixed (see U.S. Geological Survey, 2003a).

Table 2.4: Spectral range and spatial resolution for the discussed sensors. ETM+ and IKONOS have an additional panchromatic channel with a spatial resolution of 15 m for ETM+ and 1 m for IKONOS, respectively (from Kramer, 1996).

| satellite | spectral response | blue | green | red | NIR | MIR | MIR | TIR |
|---|---|---|---|---|---|---|---|---|
| LS 5 LS 7 | TM/ETM+ channel no. | 1 | 2 | 3 | 4 | 5 | 7 | 6 |
| | wavelength range [$\mu$m] | 0.45-0.52 | 0.52-0.60 | 0.63-0.69 | 0.76-0.90 | 1.55-1.75 | 2.08-2.35 | 10.40-12.50 |
| | spatial resolution [m] | 30 | 30 | 30 | 30 | 30 | 30 | 60 |
| IRS 1C IRS 1D | LISS/III channel no. | | 2 | 3 | 4 | 5 | | |
| | wavelength range [$\mu$m] | | 0.52-0.59 | 0.62-0.68 | 0.77-0.86 | 1.55-1.70 | | |
| | spatial resolution [m] | | 23.5 | 23.5 | 23.5 | 70.8 | | |
| IKONOS | IKONOS channel no. | 1 | 2 | 3 | 4 | | | |
| | wavelength range $\mu$m | 0.44-0.52 | 0.51-0.59 | 0.63-0.68 | 0.75-0.85 | | | |
| | spatial resolution [m] | 4 | 4 | 4 | 4 | | | |

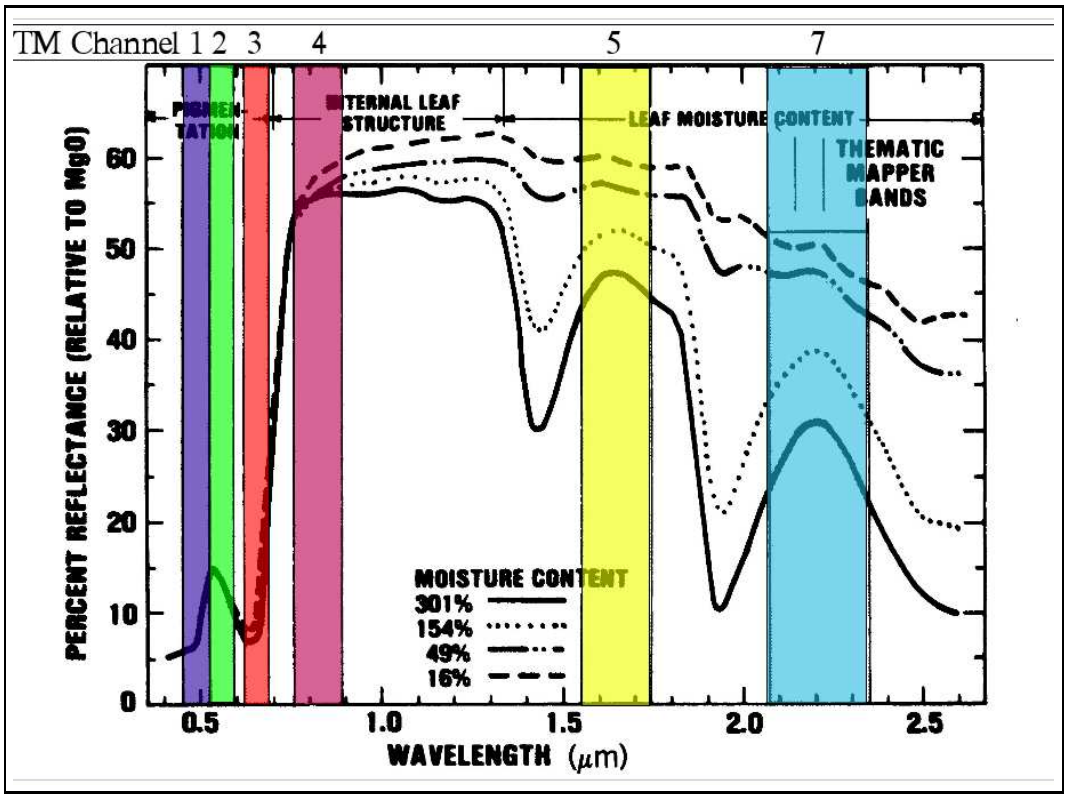



Figure 2.8: Spectral range of the TM channels 1 to 5 and 7, adapted from (Elachi, 1987, Chapter 3).

### LISS/3

LISS/3 is mounted on the IRS satellites 1C and 1D. IRS 1C was launched in December 1995 and 1D in September 1997. The inclination of the orbit is 97.5° and the orbit period is 101 min for both satellites. The orbit height is 817 km for IRS 1C. IRS 1D is on an elliptical orbit resulting from unexpected behaviour during the satellite launch and has a height of 736 km in the perigee and a height of 825 km in the apogee.

The LISS/3 sensor has two VIS channels, one NIR channel and one MIR channel. The first three channels have a resolution of 23.5 m and the MIR has a lower resolution of 70.8 m (see Table 2.4). The LISS/3 channels are comparable with the TM channels 2 to 5 and could therefore be used as a replacement for TM if no TM data are available.

### Sensor Calibration

Satellite sensors convert the incoming radiance for a single pixel into a digital number (digital number (DN)) $q_{cal}$. In order to compare the measurements from different sensors, especially TM and ETM+, it is necessary to convert these values into physical values like radiance or reflectance (see Section 1.4.1 (p. 7)).

Depending on the binary digits used, the detected radiance is quantised into a number of radiance levels. TM and ETM+ use 8 bits per pixel and channel and thus 255 radiance levels[3]. LISS/3 only uses 7 bits per pixel. Therefore, only 126 radiance levels are available.

The original radiance $L_i$ measured by the sensor for channel $i$ can be determined using the gain $g_i$ and the bias $b_i$ which are stored in the header file of the satellite data or are available through the Internet (Irish, 2000, Chapter Calibration). Note that TM is not calibrated for the radiance at the sensor but rather for the surface radiance assuming the American Standard Atmosphere (Irish, 2000, Chapter Calibration).

$$L_i = g_i q_{cal} + b_i \tag{2.1}$$

Since the sensor is calibrated for surface radiances in clear sky conditions, i.e., the atmospheric effect for a clear sky has been taken into account. Assuming a clear sky, the radiance can therefore be converted into the surface reflectance $\rho_s$ by applying:

$$\rho_s = \frac{\pi L_i d^2}{L_{sun} \cos\theta_{sun}} \tag{2.2}$$

where $d$ is the sun-earth distance in astronomical units, $\theta_{sun}$ the sun zenith angle during acquisition time and $L_{sun}$ the mean solar irradiance. These parameters are available from (Irish, 2000, Chapter Calibration). For LISS/3 the radiances and reflectances can be determined similarly (Kalyanaraman et al., 1995).

The above procedure is only applicable for a dry, cloud free atmosphere. Therefore an atmospheric correction is usually necessary (Song et al., 2001; Moran et al., 1992). This will be discussed in Section 3.2 (p. 60).

Channel 6 of the sensors TM and ETM+ measures the TIR radiation of the earth and can be used to determine the surface temperature. This is useful to distinguish the colder clouds from the warmer earth surface (see Section 3.2.1 (p. 62)). The surface temperature $T_s$ can be determined using the following empirical relation (Irish, 2000, Chapter Calibration):

$$T_s = \frac{k_2}{\ln(\frac{k_1}{L_6} + 1)} \tag{2.3}$$

with

$$\begin{aligned} k_1 &= 666.9 \text{ W}/(\text{m}^2\text{sr}\mu\text{m}) \\ k_2 &= 1282.71 \text{ K} \end{aligned}$$

for ETM+ and

$$\begin{aligned} k_1 &= 607.76 \text{ W}/(\text{m}^2\text{sr}\mu\text{m}) \\ k_2 &= 1260.56\text{K} \end{aligned}$$

[3]The first level in the data is reserved for "no data".

for TM. This equation is not valid for all surface types since it assumes the surface emissivity for this wavelength to be approximately 1 (see Section 1.4.1 (p. 7)). This is not true, e.g., for water which has an emissivity of about 0.95 for this wavelength. Therefore the temperature is underestimated, if the above formula is employed (see Section 3.2.1 (p. 62)).

### 2.2.2 Ground Surface Gathered Data

Since the spectral signatures (see Section 1.4.1 (p. 7)) of canola fields and other surface types are unknown, it is important to have information on the location of canola cultivation from ground based sources. In this study two types of ground data are available:

- Ground survey data, where the position of single fields were recorded.
- Agricultural statistics, which give information on the total acreage in a certain area.

#### Ground Survey Data

There are five different sets of ground survey data available in this study:

- The agricultural mapping for the region of the Quillow River (Northern Brandenburg).
- The ground survey results for the mapping of canola interbreeding partners in the surrounding area of Bremen.
- The results from interrogations of a seed producing company on the fields used for canola seed production South of Bremen.
- The archived information on agricultural experiments with different agricultural crops on an experimental farm near Braunschweig.
- The Global Positioning System (GPS) based mapping of canola fields in 2001 in the surrounding area of Bremen.

The type of ground survey data necessary in this study is the location of canola fields in different regions of the study area. In principle it is sufficient to know the position of one point inside or near the borders of a field and the agricultural plants grown there. Accurate information on the field edges, e.g., corner coordinates or complete edges, are of advantage, since it permits to compare the size of the field with the classification result.

### Agricultural Crop Mapping in Quillow

A detailed data set is provided by the *Leibniz-Zentrum für Agrarlandschafts- und Landnutzungsforschung* (Leibniz-Center for Agricultural Landscape and Land Use Research, ZALF) in Müncheberg. The data set is available for the years 1999, 2000 and 2001 and covers an area of 24,200 ha in the North of Brandenburg, 945 fields were mapped for this survey. Figure 2.9 shows these fields for the year 2001 with the type of cultivated plants highlighted in different colours. The field boundaries were determined by the use of a GPS receiver in 1999 and the cultivated plants were updated each year by interviewing the owner of the field in question. Some of the field boundaries changed with the years. This results in inaccuracies for 2000 and 2001 but most of the boundaries remained unchanged.

### Mapping of Canola Interbreeding Partners

Another ground survey was performed by the ecology division of the *Zentrum für Umweltforschung und Umwelttechnologie* (Centre for Environmental Research and Environmental Technology, UFT) for the years 2001 and 2002. The main purpose of the survey was the mapping of wild growing canola and its interbreeding partners. Besides the species and the location of the plant, the type of habitat was recorded. The types of habitat also included agricultural fields and the crop grown on that field. In 2001 this data set contains 273 different fields of which 132 were canola. In 2002 there were 552 different fields of which 289 were canola fields.

### Mapping of Canola Fields by GPS

Additionally canola fields were visited during blooming and the corner coordinates for 5 fields were mapped with a GPS receiver. Additional fields were mapped by two corner points of a border line and the field orientation in respect to North. 12 further fields were mapped with this method. The orientation was determined with a magnetic compass.

After the first classification test 21 classified fields were visited in order to verify the classification.

### Experimental Cultivation

The *Biologische Bundesanstalt für Land- und Forstwirtschaft* (Federal Biological Research Centre for Agriculture and Forestry, BBA) operates an experimental farm in Sickte near Braunschweig. This farm performs agricultural experiments on its fields. The experiments and the crop grown were archived and provided information of canola cultivation on eight fields, including the corner coordinates for the yeas 1995 to 2001.

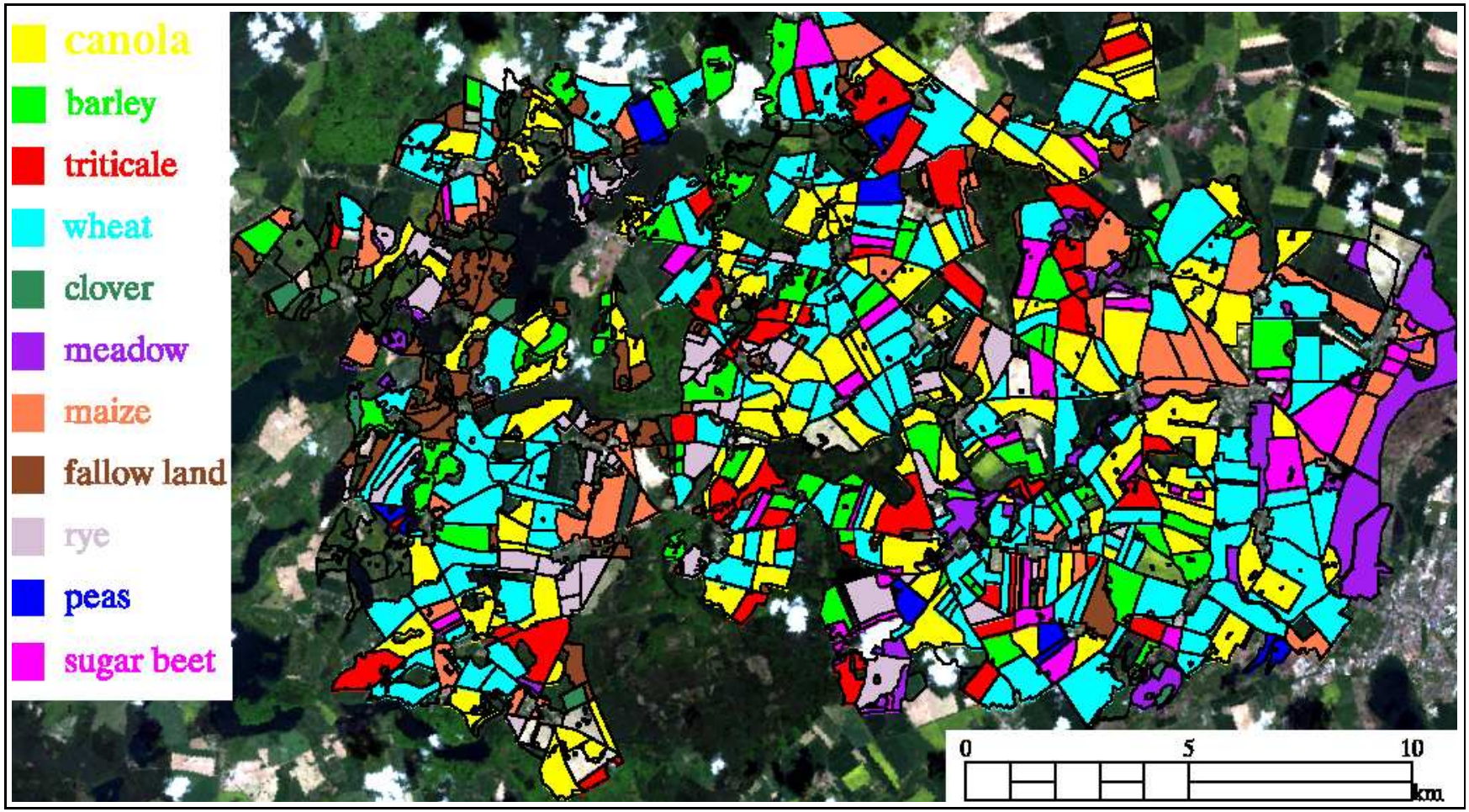



Figure 2.9: Agricultural crop mapping in the region of Quillow performed by the ZALF in Müncheberg. The different colours indicate the different crops grown in that region in 2001. The background is a true colour image from ETM+ acquired May 13, 2001. Original data: LANDSAT ETM+ ©ESA, 2001. Distributed by Eurimage..

**Interview with a Seed Producer**

In order to obtain historical data, a seed producer, the *Deutsche Saatveredelung* (German Seed Refinement, DSV), Lippstadt-Bremen, was interviewed for the fields used for seed production. The interrogation yielded the coordinates for 16 canola fields, 1 in 1999 and 2 in 2000, five in 2001 and eight in 2002.

### 2.2.3 Agricultural Statistics

The agricultural statistics for Germany is collected every year (see (Statistisches Bundesamt Deutschland, 2002)). It is not suitable for the validation of the results, since the study area only covers a part of Germany. Every 8 years a more detailed statistics is published: the *Kreisstatistik* (County Statistics) which is based on 10.000 sample interrogations of farmers from different counties. This statistic gives extrapolated information on the total acreage and yield for every county in Germany for various agricultural crops (Statistisches Bundesamt Deutschland, 2002). It was last collected in 1999 and can be used to validate the classification of 1999.

A more accurate and more frequently collected statistics is the *Gemeindestatistik* (Township Statistics) of Schleswig-Holstein. It is collected every four years, last in 1999 and 1995, and can be used for validation. Like the county statistic, it is based on interrogation of farmers and contains the information on the statistically estimated total acreage of the different crops for each township in Schleswig-Holstein.

# Chapter 3

# Preprocessing

Satellite data have to be prepared in order to extract information on the surface type. The mapping of the pixels available from the satellite data themself is not accurate enough to produce maps of the identified canola pixels that met the requirements of this study. Therefore, the mapping of has to be improved. This is especially important to allow a comparison of these identified fields with validation data from the ground.

Moreover, optical satellite data are influenced by the atmosphere, clouds in particularly. Therefore the atmospheric influences on the satellite data are discussed, and procedures to identify and, if possible, to compensate the effects of cloud aerosols in the images are presented.

## 3.1 Geocoding and Image Registration

Image registration, also named geocoding, is the assignment of satellite data to the corresponding regions (pixels) on the earth surface. High resolution satellite images like TM and LISS/3 data are usually geocoded by the satellite provider using information on the orbit and the sensor attitude. However, this type of geocoding with a typical accuracy of 300 to 1 km (see Section 3.1.2, p. 39) is not accurate enough to identify individual fields in different satellite images, which is necessary for the selection of training data sets (see Section 4.4.1, p. 126).

Therefore an additional correction using passpoints is necessary. The passpoint correction gives sufficient accuracy to generate maps containing the classification results, but it is less accurate than the sensors resolution.

Since classification and cloud masking algorithms require to identify corresponding pixels in different satellite images (see Section 3.2, p. 60 and Section 4.1.2, p. 89), which is not available through the passpoint correction, additional processing step has to be applied. These algorithms do not need an accurate map projection but rather an image-to-image registration. This aim is achieved by combining the passpoint correction and a correlation method between images of the same or overlapping frames.

### 3.1.1 Sources of Image Distortions

High resolution spaceborne sensors acquire surface data over a large distance on a spherical surface. The sensor continuously measures with a scanning mirror that reflects the light coming from the earth surface to the detector. The sensor scans the field of view (FOV), i.e., the angular range scanned by the sensor. One single data point (pixel) represents the integrated radiation over the instantaneous field of view (IFOV). For most high resolution sensors, the scanning speed of the mirror is adapted to achieve a constant pixel size on the surface (Schowengerdt, 1997, Chapter 3). With the knowledge on the viewing direction and position of the satellite, it is possible to determine the location of the pixels on the surface. Inaccuracies of these parameters will result in errors of the pixel localization.

The following influences have to be taken into account (for a more detailed list and description see Colwell (1983, Chapter 21)):

- Changes of the satellite altitude resulting from variations of the earth gravity field lead to scale distortions. Furthermore, the changing attitude influences the IFOV location on the ground.
- The earth rotation beneath the sensor shifts scanlines gradually westward during acquisition.
- The variation of the scanning rate due to imperfections in the scanning mechanism leads to pixel distortion within one scanline.
- Variations in surface elevation cause a misplacement of pixels unless accurate elevation information is available, i.e., a digital elevation model (DEM).

These influences can be compensated if accurate information on the sensor state and satellite position during acquisition are available. This is true for most high resolution sensors and this information is more accurate for recent sensors like ETM+. With additional information of an earth ellipsoid, the satellite data can be projected on a map. The georectification with this sensor and satellite information is called geometric correction. E.g., the accuracy achieved for ETM+ data is specified by the National Aeronautics and Space Administration (NASA) to be lower than 250 m (U.S. Geological Survey, 2003b) for regions at sea level. This is valid for most parts of Northern Germany.

Another possibility for image rectification is the passpoint or ground control point (GCP) correction. It is based on the identification of corresponding points in the satellite data and on a reference map. These pairs of points are used to find a mathematical transformation between their coordinates in the image and on the map. With the passpoint correction earth rotation and earth curvature can be compensated. The varying orbital altitude and attitude can be compensated if these parameters are not changing rapidly with time.

However, inaccuracies on a small scale, caused by variations in scan speed cannot be corrected by this method. For instance, the IFOV of TM is 47 $\mu$rad and an error of 1 mrad would lead to an error of 800 m on the ground, or 26 pixels, respectively[1]. Therefore, a passpoint correction is only useful if these small scale inaccuracies have been corrected by a geometrical correction.

### 3.1.2 Geometrically Corrected Data

In this study, path-oriented satellite data is used, which implies that the stored data are still organized in scan lines. This was necessary to minimize errors caused by resampling. The information necessary for the georectification of the path-oriented data is included in the header file of the satellite data. The required parameters are the centre coordinates $(x_c, y_c)$ and the rotation angle $\alpha$, between the direction of scanlines and the east direction. With the following equation the image-based coordinates $(x, y)$ can be transformed into map-based coordinates $(x', y')$:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} (x - x_c)s_p \\ (y - y_c)s_p \end{pmatrix} + \begin{pmatrix} x'_c \\ y'_c \end{pmatrix} \tag{3.1}$$

$s_p$ is the sampling distance or pixel size of the sensor, $(x'_c, y'_c)$ the centre coordinates in the satellite image in map-based coordinates and $(x_c, y_c)$ the centre coordinate for the image-based coordinates. The data can be resampled to the desired map projection. The international map projection universal transverse Mercator (UTM) is used in this study (UTM Zone 32 and 33 North)[2]. The resampling algorithm was the nearest neighbour method (Lillesand, 2000). Figure 3.1 shows the results of this transformation for of an ETM+ and a TM Image. For a comparison, the satellite data is overlaid with a geographical information system (GIS) road map of Schleswig-Holstein. It visualises that the accuracy of the TM data with an offset of 20 pixels is not accurate enough. The mapping of the ETM+ data has a much better accuracy, but still shows deviations of 6 pixels or 180 m. Therefore, a correction with passpoints is necessary for both sensors.

### 3.1.3 Passpoint Correction

Since the geometric correction is not accurate enough, it has to be improved by using additional information for geolocating the satellite data. The most common method is the use of passpoints, also called GCPs, to find a mathematical

[1] The pixel size is not exactly the size resulting from the IFOV since the pixels overlap (Richards, 1986, Chapter 1).

[2] The official map coordinate system in Germany is Gauss-Krüger but since it is divided into five stripes and UTM only in two it is not suitable for the mapping of the complete area.

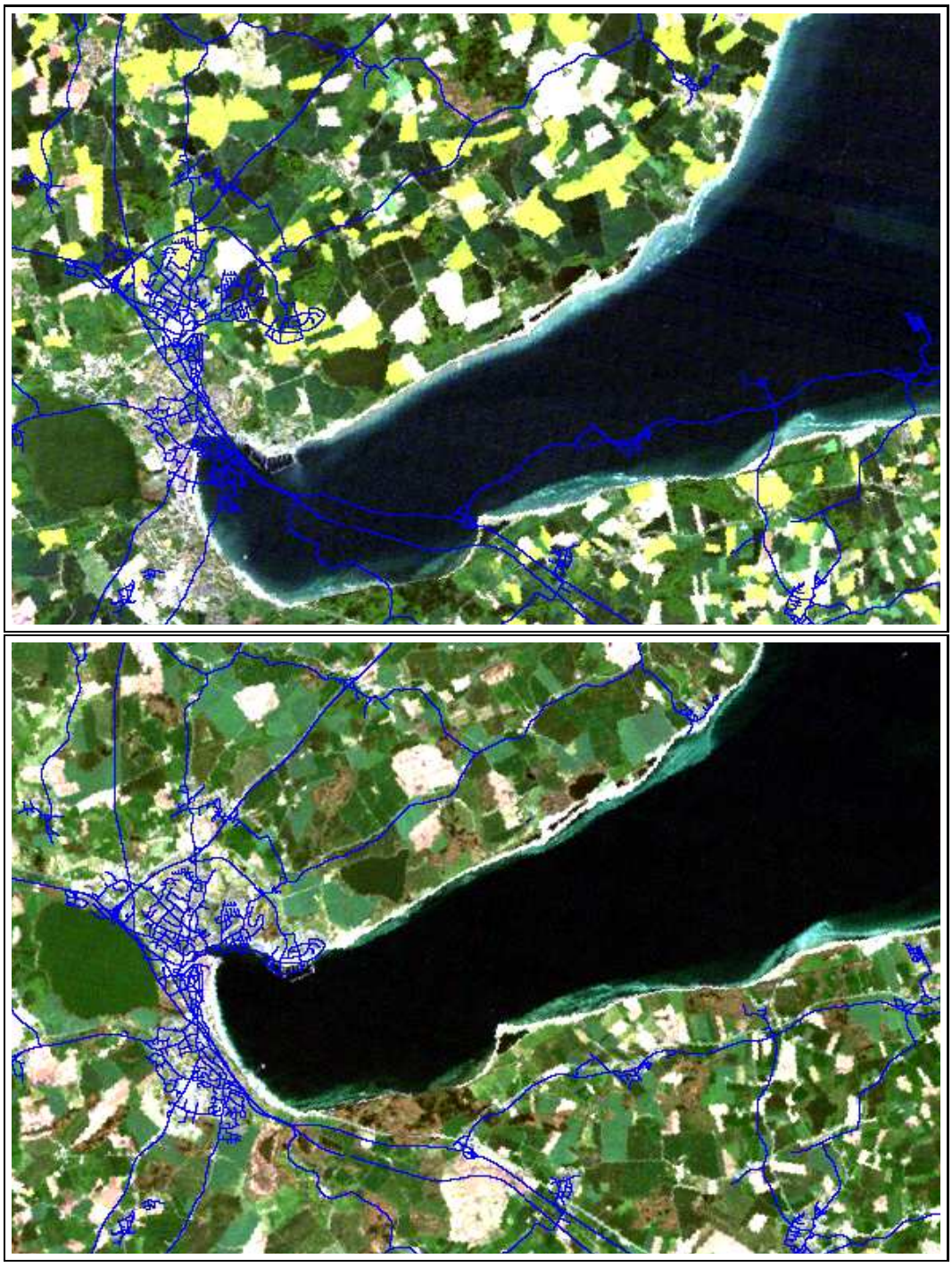


Figure 3.1: Example for the accuracy of geometric correction by the satellite provider. The clips show the bay of Eckernförde in Schleswig-Holstein. Top: Geometrically corrected TM image clip as provided by Eurimage, overlaid with a road map with 30 m resolution, generated from the ATKIS data set by the Ecology Centre at the University of Kiel. Bottom: Corresponding image clip for geometrically corrected ETM+ data. Original data: LANDSAT TM ©ESA, 2000 and LANDSAT ETM+ ©ESA, 2001. Distributed by Eurimage.

transformation $f$ to assign the map coordinates $(x', y')$ to the corresponding image coordinates $(x, y)$:

$$(x', y') = f(x, y) \tag{3.2}$$

To identify corresponding pixels in different satellite images, the map coordinates have to be transformed back and therefore, the reverse transformation $g$ is also required:

$$(x, y) = g(x', y'). \tag{3.3}$$

The accuracy of this transformation depends on the map and sensor information available. The best transformation, usually called orthorectification, can be obtained using a DEM, accurate information on the scanning geometry and GCPs. A DEM was not available and since Northern Germany is mostly flat, an empirical approach using a polynomial approximation is chosen.

The polynomial transformation is based on fitting a polynomial of $n$th order on the base of the chosen GCPs. Since this transformation is not directly related to the imaging system or the shape of the earth surface, it is necessary to validate the results of such a transformation for the sensor in question. A general polynomial transformation $f$ from image-based coordinates to map-based coordinates has the following form Schowengerdt (1997, Chapter 7):

$$x' = \sum_{i=0}^{n} \sum_{j=1}^{n-i} a_{ij} x^i y^j \tag{3.4}$$

$$y' = \sum_{i=0}^{n} \sum_{j=1}^{n-i} b_{ij} x^i y^j. \tag{3.5}$$

The corresponding reverse transformation $g$ can be obtained by exchanging the coordinates $(x', y')$ and $(x, y)$.

Higher order transformations need a larger number of GCPs since the number of required coefficients is higher. Moreover, resulting from the non-linear terms, the transformation might lead to errors in regions only sparsely covered with GCPs or at the border of the image. Transformations with $n > 2$ are therefore rarely used for high resolution data with small frame sizes, assuming that the relatively small distortion in these data can be approximated by a quadratic transformation[3]. This can be justified by the relatively small area covered and the scanning geometry (Schowengerdt, 1997, Chapter 8).

First order or affine transformations can be applied to linear distortions like rotation, shear, scale and translation (see Figure 3.2). A second order transformation can also correct quadratic distortions in the data, which can result from the earth curvature and the scanning geometry. Since the data are already geometrically corrected for these effects, it has to be investigated which types of geometric distortions remain in these images, and the transformations have to be tested to find the most suitable one for TM and LISS/3 images.

[3]Transformations with $n > 2$ are generally used for lower resolution sensors with a larger coverage and a stronger dependency on the earth curvature.

Therefore, the following paragraph will introduce quadratic and affine transformations and the mathematical background to obtain their coefficients. Additionally, another transformation is presented performing only scaling, translation and rotation but not a shearing. This is justified if the geometric correction of the provider accounts for the remaining distortions. Besides the image distortion, the identification of GCPs is a time consuming task and the number of GCPs can be limited due to lack of suitable surface features.

### Quadratic Transformation

The quadratic transformation can be obtained from equations (3.4) and (3.5) as a matrix equation

$$\mathbf{x}' = \mathbf{W}\mathbf{a} \tag{3.6}$$
$$\mathbf{y}' = \mathbf{W}\mathbf{b} \tag{3.7}$$

where $\mathbf{W}$ is a matrix

$$\mathbf{W} = \begin{pmatrix} 1 & x_1 & y_1 & x_1 y_1 & {x_1}^2 & {y_1}^2 \\ 1 & x_2 & y_2 & x_1 y_1 & x_1^2 & y_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_M & y_M & x_M y_M & {x_M}^2 & {y_M}^2 \end{pmatrix} \tag{3.8}$$

$\mathbf{a}$, $\mathbf{b}$ are the coefficient vectors and $\mathbf{x}'$, $\mathbf{y}'$ the coordinate vectors of the GCPs

$$\mathbf{a} = \begin{pmatrix} a_{00} \\ a_{10} \\ a_{01} \\ a_{11} \\ a_{20} \\ a_{02} \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_{00} \\ b_{10} \\ b_{01} \\ b_{11} \\ b_{20} \\ b_{02} \end{pmatrix} \quad \mathbf{x}' = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_M \end{pmatrix} \quad \mathbf{y}' = \begin{pmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_M \end{pmatrix}.$$

The coefficients $a_{ij}$ and $b_{ij}$ are obtained by solving these equations with the corresponding coordinate pairs of the chosen GCPs.

The number of GCPs should be higher than the number of coefficients since some coordinates might be inaccurate and $\mathbf{W}$ might not be regular if some rows are linearly dependent, e.g., if all GCPs are placed on one line. Thus $\mathbf{W}$ is not a square matrix and does not have an exact solution for the vectors $\mathbf{a}$ and $\mathbf{b}$. an approximated solution is necessary. This is achieved by minimizing the errors of predicted and actual positions:

$$\epsilon_x = \mathbf{W}\mathbf{a} - \mathbf{x}' \tag{3.9}$$
$$\epsilon_y = \mathbf{W}\mathbf{b} - \mathbf{y}' \tag{3.10}$$

$\epsilon_x$ and $\epsilon_y$ are the vectors of the differences of the map coordinates. Thus the minimum of the norm of these vectors has to be calculated

$$\min[|\epsilon_x|] = \min[\sqrt{(\mathbf{Wa}-\mathbf{x}')(\mathbf{Wa}-\mathbf{x}')}] \quad (3.11)$$

$$\min[|\epsilon_y|] = \min[\sqrt{(\mathbf{Wa}-\mathbf{y}')(\mathbf{Wa}-\mathbf{y}')}] \quad (3.12)$$

which is the least square solution for the equations (3.6) and (3.7). This least square solution is calculated using the singular value decomposition (SVD) which is most suitable to solve these types of equation (Press et al., 1992, Chapter 15). A detailed description of the SVD can be found in Press et al. (1992, Chapter 3).

The quality of a transformation depends on the number, distribution and accuracy of the selected GCPs. This is more important for higher order polynomials, but also has effects on the quadratic transformation. Therefore, a less sensitive transformation would be advantageous.

### Affine Transformation

An affine transformation is a a polynomial transformation with $n = 1$. The affine equation can be noted in matrix form, analogously to the equations (3.6) and (3.7) by modifying the matrix $\mathbf{W}$ and corresponding coefficient vectors $\mathbf{a}$, $\mathbf{b}$:

$$\mathbf{W} = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_M & y_M \end{pmatrix} \quad \mathbf{a} = \begin{pmatrix} a_{00} \\ a_{10} \\ a_{01} \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_{00} \\ b_{10} \\ b_{01} \end{pmatrix}$$

The solution of this equation is also obtained by using the SVD. To get a better understanding of the affine transformation, it can be written as follows:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a_{00} \\ b_{00} \end{pmatrix} + \begin{pmatrix} a_{10} & a_{01} \\ b_{10} & b_{01} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (3.13)$$

From this equation, the influence of the different coefficients $a_{ij}$ could be identified as different image transformations: The coefficients $a_{00}$ and $b_{00}$ describe a simple translation, and the remaining coefficients are responsible for a shifting, rotation, scaling and shearing of the image. Three of these transformations are sketched in Figure 3.2. Since only these transformations are applied, the affine transform is less sensitive to the distribution and the errors of the GCPs. Thus, the number of GCPs can be limited and an error of a single GCPs is less relevant.

### The scale-rotate-translate (SRT)-Transformation

Assuming the shear has been compensated by the geometric correction, it is possible to neglect it in the transfromation. By omitting the shear, the

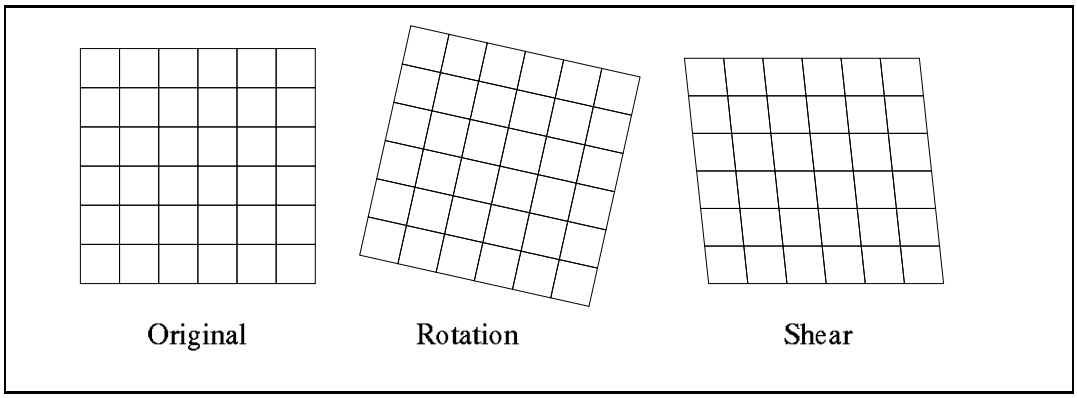



Figure 3.2: Example for the results of an affine transform.

transformation equation is equal to the original transformation of the satellite data in equation (3.1):

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} (x - x_c)s_p \\ (y - y_c)s_p \end{pmatrix} + \begin{pmatrix} x'_c \\ y'_c \end{pmatrix}$$

A comparison of the parameters with the affine transformation in equation (3.13) allows to assign the affine coefficients to the rotation angle, translation and scale:

$$\begin{aligned} c_1 &= a_{10} = b_{01} = s_p \cos(\alpha) && (3.14) \\ c_2 &= -b_{10} = a_{01} = s_p \sin(\alpha) && (3.15) \\ a_{00} &= x_c - x'_c s_p \cos(\alpha) - y'_c s_p \sin(\alpha) && (3.16) \\ b_{00} &= y_c + x'_c s_p \sin(\alpha) - y'_c s_p \cos(\alpha) && (3.17) \end{aligned}$$

With the new parameters $c_1$ and $c_2$, it is possible to obtain a new transformation that only scales, rotates and translates the image. This SRT transformation has the following form:

$$\begin{aligned} x' &= a_{00} + c_1 x + c_2 y && (3.18) \\ y' &= b_{00} - c_2 x + c_1 y && (3.19) \end{aligned}$$

Since equations (3.18) and (3.19) are not independent, a new set of linear equations is necessary. This can be obtained by combining equation (3.6) and (3.7), which leads to a new matrix equation:

$$\mathbf{z}' = \mathbf{W}\mathbf{c} \tag{3.20}$$

with the new matrix $\mathbf{W}$ and vector $z$:

$$\mathbf{W} = \begin{pmatrix} 1 & x_1 & y_1 & 0 \\ 1 & x_2 & y_2 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_M & y_M & 0 \\ 0 & y_1 & -x_1 & 1 \\ 0 & y_2 & -x_2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & y_M & -x_M & 1 \end{pmatrix} \quad \mathbf{z}' = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_M \\ y'_1 \\ y'_2 \\ \vdots \\ y'_M \end{pmatrix} \quad \mathbf{c} = \begin{pmatrix} a_0 \\ c_1 \\ c_2 \\ b_0 \end{pmatrix}$$

This equation can also be solved using the SVD. The advantage of this transformation is that it only corrects for translational, rotational and scale effects and therefore is less sensitive to errors than the usual affine transformation. Since the remaining distortion in the image is unknown, the three transformations presented above have to be evaluated by using example satellite images and testing sets of GCPs. This evaluation will be shown after the description of the reference maps and the methods to determine GCPs are presented.

### Reference maps

The selection of passpoints requires accurate reference maps. Since the investigation area is quite large, these maps have to be available in digital form to permit an accurate localization of surface features and still be manageable. The most accurate map for the investigation area, the *Amtliches Topographisch-Kartographisches Informationssystem* (Authoritative Topographic Cartographic Information System) (ATKIS) map providing an accuracy of ±3 m which is sufficient for the selection of GCPs. However, the price of €15 per km$^2$ is too high to obtain this dataset for the complete investigation area.

Nonetheless, the ATKIS data set for Schleswig-Holstein is available at the Ecology Centre at the University of Kiel and can be used to review the geolocation in that area. The licensing rules prohibited the direct use of this data set, so the ecological institute provided a derived raster map with the spatial resolution of TM (30 m) and LISS/3 (25 m).

For the rest of the investigation area, an inexpensive, but less accurate alternative are the *Top Karten* that are available at the land survey office of each federal state in Germany. The price for a map amount to €40 per state and were obtained for all states in question. The accuracy for these maps is indicated with ±5 m, but tests with a GPS-Receiver, which has an accuracy of 6 m, showed differences of about 40 m. Since only the *Top Karten* were available for the complete investigation area, these were used to select the GCPs.

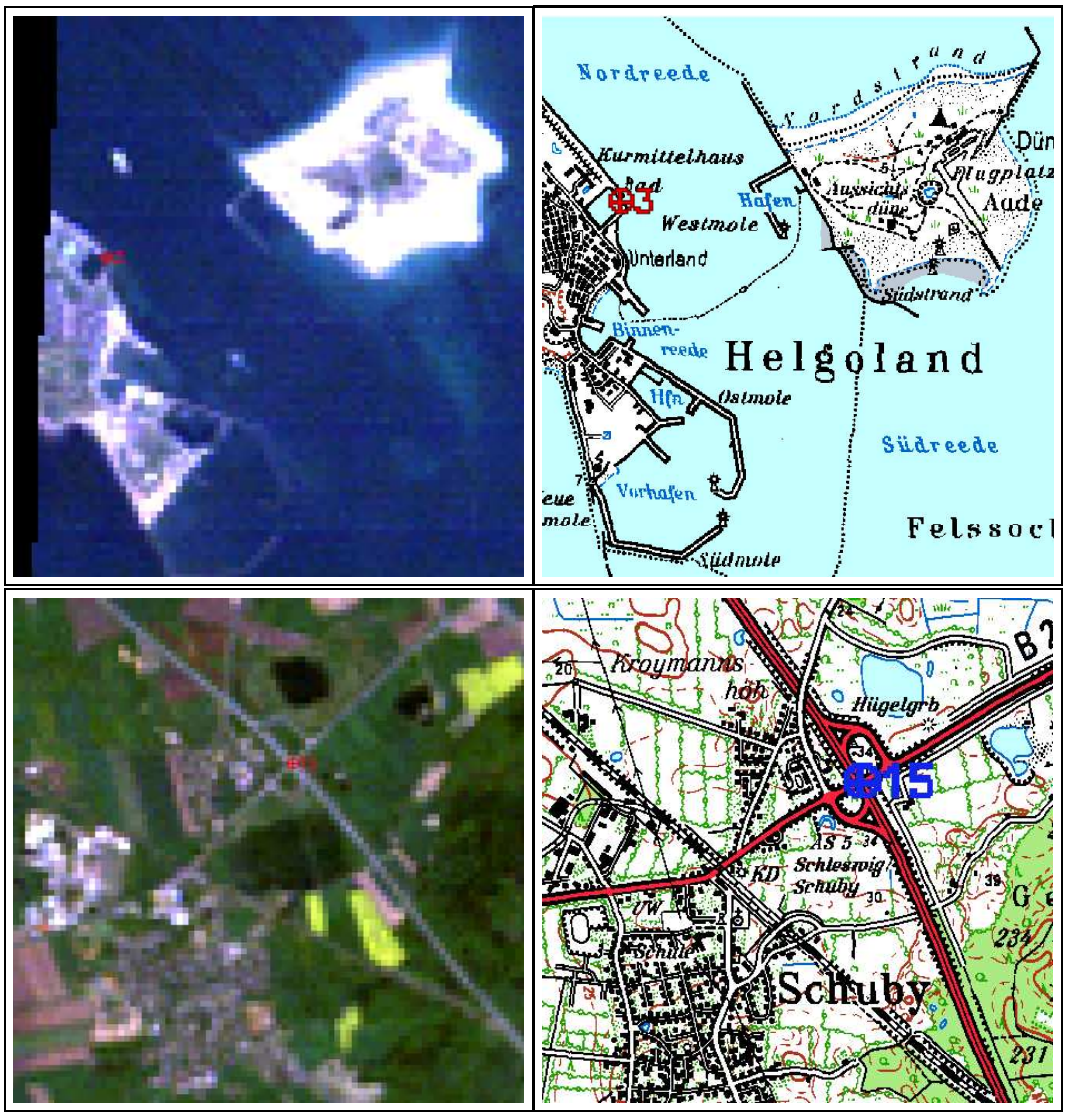



Figure 3.3: GCP selection, top left: clip of Helgoland from a TM image (196/022; May 9, 2000) with the chosen GCP at the old harbour, top right: corresponding clip and GCP from the Top Karten Schleswig-Holstein, bottom: a GCP selected in the vicinity of Schleswig. Original data: LANDSAT TM ©ESA, 2000. Distributed by Eurimage.

### Selection of passpoints

GCPs must be identifiable with surface features in the satellite data. These are generally jetties, highway intersections, railroad crossings or bridges. In the densely populated region of Northern Germany, these are quite numerous, although the sensor resolution limits the usable roads to those broader than *Bundesstraßen* (interstate routes). The coordinates of these characteristics were extracted from the Top Karten. Figure 3.3 shows a jetty and a highway intersection that were selected as GCPs. The whole image has to be georectified, so the GCPs had to be distributed uniformly over the image. Figure 3.3 shows an example of GCP selections for Frame 196/022 (Schleswig-Holstein).

GCP selection is usually performed recursively. After the transformation is calculated, the transformation is used to determine new map coordinates from the satellite coordinates of GCPs. If the discrepancy between original and calculated GCP map coordinates exceeds a maximum distance, these points are either removed or verified with the map. This new set of GCPs is used to calculate a new transformation until all GCPs are within a postulated range.

This procedure is time-consuming and only applicable with few images. Since a minimum of 14 GCPs are selected for every of the fifty available images, the total number of GCPs amounts to more than 700. Therefore the recursive method is not practicable, since it is too time consuming, and the correction of GCPs with the map is omitted. The GCPs that do not match the postulated distances are removed. First, the GCP with the maximum root mean square (RMS) error for the quadratic transformation is removed. Afterwards, the transformation is recalculated and the GCP with the maximum RMS error for the remaining GCPsis removed. This is repeated until all remaining GCP have a RMS error lower than 75 m. If the number of remaining GCP is lower than 14, new GCPs have to be chosen manually. This method is performed using the quadratic transformation only, to obtain the same set of GCPs for all transformations.

### Evaluating of the Transformations

With these restrictions for the selection of GCPs, it is necessary to identify the most suitable of the presented transformations. Because of the large number of images, a representative area is chosen, which has to supply most of the properties that can cause errors in the georectification by GCPs, e.g., the changing elevation or lack of usable surface features. The federal state of *Schleswig-Holstein* is chosen as testing area. For the following reasons, the state of *Schleswig-Holstein* is the best region for this purpose:

1. The most accurate map, the ATKIS road map is only available for Schleswig-Holstein and can be used to evaluate the accuracy of the transformation derived from the *Top Karten* based GCPs.

2. Schleswig-Holstein is surrounded by the North and the Baltic Sea, and is therefore lacking surface features at the border regions of the satellite image.

3. The terrain in Schleswig-Holstein covers the full range of height variations of the complete investigation area in a relatively small region. The regions in the West are very close to sea level, whereas the regions in the South-East have elevations up to 168 m.

4. Images for both LANDSAT Sensors, TM and ETM+ are available for this region and since the georectification has been improved for ETM+, this allows to investigate the differences of these sensors.

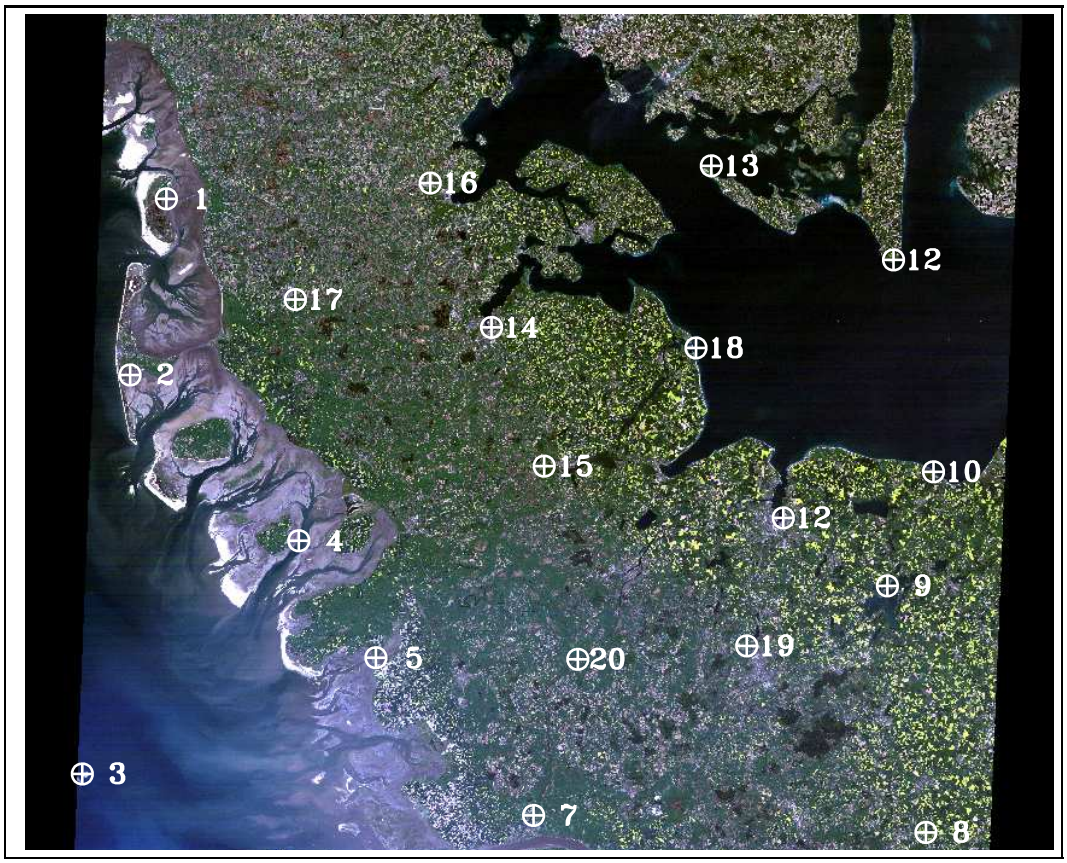



Figure 3.4: Selection of GCPs for one complete image registration, TM image, frame 196/022 acquired on May 9, 2000. The white marks represent the chosen GCPs for this image. Original data: LANDSAT TM ©ESA, 2000. Distributed by Eurimage.

Two different frames are available for this region, frame 196/022[4] and frame 195/022 (see figure 2.6). Since frame 196/022 covers the larger area of Schleswig-Holstein, it will be used. The selected GCPs for this frame are displayed in Figure 3.4. They are evenly distributed over the whole land area and GCPs were used for all images of this frame, where the satellite coordinates had to be adapted for every image. Due to cloud cover, not all of these GCPs have been usable for all images. To evaluate the three transformations, they have been applied to all images and GCP sets from that frame.

Three different methods have been applied to compare the transformations:

- Comparison of the determined transformation coefficients.
- Comparison of predicted and actual position of the GCPs.
- Visual comparison of the resampled image with the digital road map from the ATKIS data.

[4]Frames are always noted as path/row and images as path/row;date.

**Comparison of the Coefficients.** The distortions of the images can be estimated by comparing the coefficients for the different transformations, e.g., small values for the quadratic coefficient $a_{20}$ would indicate that there is only a weak quadratic distortion in $x$-direction. Moreover if all quadratic and bilinear terms are close to zero, the quadratic transformation would give similar results to the affine transformation and the latter would be sufficient to describe the geometric distortions in the image.

In order to determine the influence of the coefficients, it is necessary to perform a sensitivity estimation. The impact of variations on the coefficients $\Delta a_{ij}$ and $\Delta b_{ij}$, on the variation $\Delta\epsilon_x$ and $\Delta\epsilon_y$ of the calculated map coordinates can be estimated by using the equations (3.9) and (3.10).

$$\Delta\epsilon_x = \underbrace{\Delta a_{00}}_{\Delta\epsilon_{a_{00}}} + \underbrace{\Delta a_{10}x}_{\Delta\epsilon_{a_{10}}} + \underbrace{\Delta a_{01}y}_{\Delta\epsilon_{a_{00}}} + \underbrace{\Delta a_{11}xy}_{\Delta\epsilon_{a_{11}}} + \underbrace{\Delta a_{20}x^2}_{\Delta\epsilon_{a_{20}}} + \underbrace{\Delta a_{02}y^2}_{\Delta\epsilon_{a_{02}}} \quad (3.21)$$

$$\Delta\epsilon_y = \underbrace{\Delta b_{00}}_{\Delta\epsilon_{b_{00}}} + \underbrace{\Delta b_{10}x}_{\Delta\epsilon_{b_{10}}} + \underbrace{\Delta b_{01}y}_{\Delta\epsilon_{b_{01}}} + \underbrace{\Delta b_{11}xy}_{\Delta\epsilon_{b_{11}}} + \underbrace{\Delta b_{20}x^2}_{\Delta\epsilon_{b_{20}}} + \underbrace{\Delta b_{02}y^2}_{\Delta\epsilon_{b_{02}}} \quad (3.22)$$

Assuming the error is only resulting from one of the coefficients, $\Delta\epsilon_{a_{ij}}$ and $\Delta\epsilon_{a_{ij}}$ can be estimated as:

$$\Delta a_{00} = \Delta\epsilon_x \qquad \Delta a_{10} = \tfrac{\Delta\epsilon_x}{x_m} \qquad \Delta a_{11} = \frac{\Delta\epsilon_x}{x_m y_m} \quad \Delta a_{20} = \frac{\Delta\epsilon_x}{x_m^2} \quad \ldots (3.23)$$

$$\Delta b_{00} = \Delta\epsilon_y \qquad \Delta b_{10} = \tfrac{\Delta\epsilon_y}{x_m} \qquad \Delta b_{11} = \frac{\Delta\epsilon_y}{x_m y_m} \quad \Delta a_{20} = \frac{\Delta\epsilon_y}{x_m^2} \quad \ldots (3.24)$$

$x_m$ and $y_m$ are the maximum values for the image coordinates. Since the transformation can be assumed to be most accurate at the centre of the image, a realistic estimation is given by using the halved values of the image width: $x_m = 3460$ pixel and the halved height $y_m = 2980$ pixel for a TM/ETM+ image.

A significant deviation for the geocoding of a satellite image is the pixel size $s_p$, at least for a GCP correction, since surface features can only be identified if they are larger than the spatial resolution. Therefore, the minimum expected errors are $\Delta\epsilon_x = \Delta\epsilon_y = 30\,\mathrm{m}$.

These assumptions are used to calculate the resulting deviations necessary to result in a deviation of one pixel. The resulting values for a TM image are listed in Table 3.1. It can be seen that the higher order terms are more sensitive to errors in the coefficients since their values are much smaller.

This sensitivity estimation can be compared to the calculated coefficients of the transformations for the frame 196/022. Table 3.2 shows the results of the different transformations divided by the corresponding deviation from Table 3.1. These values indicate the resulting deviation in pixel size. It can be seen that:

- The quadratic and bilinear coefficients for all images are very small. Only three values exceed the ratio of coefficient and sensitivity by more than one.

Table 3.1: Estimation of the coefficient variations under the assumptions of a resulting transformation variation of one pixel (30 m for TM) and the maximum values for the satellite coordinates, i.e. the halved width ($x_m = 3460$ pixel) and the halved height ($y_m = 2980$ pixel) of a TM/ETM+ image.

| $\Delta a_{00}$[m] | $\Delta a_{10}$[m] | $\Delta a_{01}$[m] | $\Delta a_{11}$ [m] | $\Delta a_{20}$ [m] | $\Delta a_{02}$ [m] |
|---|---|---|---|---|---|
| 30 | 0.009 | 0.010 | 2.9E-6 | 2.5E-6 | 3.4E-6 |
| $\Delta b_{00}$[m] | $\Delta b_{10}$ [m] | $\Delta a_{01}$[m] | $\Delta b_{11}$ [m] | $\Delta b_{20}$ [m] | $\Delta b_{02}$ [m] |
| 30 | 0.010 | 0.009 | 2.9E-6 | 3.4E-6 | 2.5E-6 |

- The situation of the linear terms is similar to the two polynomial transformations. The difference between quadratic and affine transformation is less than two pixels. The difference between SRT and quadratic transformation is slightly larger, which is caused by a remaining shear in the TM data. A possible cause for this shear is the variation of elevation in Schleswig-Holstein. This could be verified by using additional frames from other regions, but since elevation is also present in other frames, the affine transformation is more suitable than the SRT transformation.

- The constant coefficients are mostly within the sensors resolution. There is an exception for the SRT-transformation in 1998. But since the centre of rotation can be chosen freely, this does not lead to errors in the coordinate transformation, which was verified by comparing the registered images.

Summing up these results, affine and quadratic transformation leads to very similar results with just one to two pixels difference, even for the border regions of the image.

**Comparison of predicted and original GCP-positions** The above discussion only compares the results of the different transformations. Thus it is necessary to evaluate the transformation accuracy by using the original GCPs. This is done by comparing the original GCP map-coordinates with the coordinates predicted by the transformations. Table 3.3 shows the obtained deviations. The deviations are noted as RMS-error for $\epsilon_i = \sqrt{\epsilon_{x,i}^2 + \epsilon_{y,i}^2}$ by using the corresponding components of $\epsilon_{\mathbf{x}}$ and $\epsilon_{\mathbf{y}}$ from equations (3.9) and (3.10).

The mean $\bar{\epsilon}$ of this deviation has the size of about one pixel size for the affine and quadratic transformations. The mean for the SRT-transformation is larger, but still in the range of the doubled pixel size. More important for the quality of the transformation is the maximum error $\epsilon_{\max}$, which is lowest for the quadratic transformation, but only slightly larger for the affine transformation. The largest maximum errors are resulting from the SRT-transformation which exceeds 130 m in two cases.

Table 3.2: Resulting coefficients for the different transformations divided by the estimated coefficient variation for a deviation of one pixel (see Table 3.1). The quadratic and the bilinear coefficients are mostly smaller than one pixel and the linear terms for the affine and quadratic transformation have deviations of about two pixels.

| year | trans-formation | ratio of coefficients | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\frac{a_{00}}{\Delta a_{00}}$ | $\frac{a_{10}}{\Delta a_{10}}$ | $\frac{a_{01}}{\Delta a_{01}}$ | $\frac{a_{11}}{\Delta a_{11}}$ | $\frac{a_{20}}{\Delta a_{20}}$ | $\frac{a_{02}}{\Delta a_{02}}$ |
| 2001 | n=2 | 728 | 3236 | -713 | 0.4 | 0.3 | -0.8 |
| | n=1 | 728 | 3237 | -715 | | | |
| | SRT | 728 | 3238 | -715 | | | |
| 2000 | n=2 | 870 | 3236 | -717 | 0.4 | 0.0 | -0.8 |
| | n=1 | 871 | 3236 | -719 | | | |
| | SRT | 873 | 3237 | -720 | | | |
| 1998 | n=2 | 92 | 3236 | -715 | 0.1 | 0.1 | 0.6 |
| | n=1 | 91 | 3236 | -713 | | | |
| | SRT | -637 | 3236 | -716 | | | |
| 1995 | n=2 | 621 | 3243 | -711 | -0.1 | 2.3 | 0.2 |
| | n=1 | 623 | 3239 | -711 | | | |
| | SRT | 623 | 3239 | -711 | | | |
| | | $\frac{b_{00}}{\Delta b_{00}}$ | $\frac{b_{10}}{\Delta b_{10}}$ | $\frac{b_{02}}{\Delta b_{02}}$ | $\frac{b_{11}}{\Delta b_{11}}$ | $\frac{b_{20}}{\Delta b_{20}}$ | $\frac{b_{02}}{\Delta b_{02}}$ |
| 2001 | n=2 | 6990 | -716 | -3236 | 0.4 | 0.4 | -1.4 |
| | n=1 | 6991 | -715 | -3238 | | | |
| | SRT | 6989 | -715 | -3238 | | | |
| 2000 | n=2 | 6721 | -725 | -3234 | 0.6 | 1.0 | 0.3 |
| | n=1 | 6718 | -722 | -3237 | | | |
| | SRT | 6715 | -720 | -3237 | | | |
| 1998 | n=2 | 6659 | -717 | -3238 | -1.6 | 1.0 | 0.4 |
| | n=1 | 6659 | -717 | -3239 | | | |
| | SRT | 5002 | -716 | -3237 | | | |
| 1995 | n=2 | 7025 | -712 | -3239 | 1.0 | 0.2 | 0.3 |
| | n=1 | 7026 | -712 | -3239 | | | |
| | SRT | 7023 | -711 | -3239 | | | |

The error of the transformation can also be estimated by the quantity of GCPs with a large deviation of predicted and GCP map position. As maximum deviation 75 m, i.e, or two and a half TM pixel, have been chosen. Except for the year 2000, this quantity of GCPs is the same for the affine and quadratic transformation. For both transformations, there is mostly only one GCP which

Table 3.3: Statistic, showing the RMS-errors $\epsilon_i$ between predicted and actual map position comparing the different transformations. $\bar{\epsilon}$ is the mean and $\delta\epsilon$ the standard deviation of all coordinate RMS-errors. Also displayed is the number $n$ of available GCPs, the number $n_{\epsilon>75\,m}$ of $\epsilon$ larger than 75 m and the maximum deviation $\epsilon_{\max}$.

| sensor | year | trans-formation | $\bar{\epsilon}$ [m] | $\delta\epsilon$ [m] | $\epsilon_{\max}$ [m] | $n$ | $n_{\epsilon>75\,m}$ |
|---|---|---|---|---|---|---|---|
| ETM+ | 2001 | n=2 | 31.1 | 13 | 67.0 | 20 | 0 |
| | | n=1 | 31.1 | 12 | 72.1 | 20 | 0 |
| | | SRT | 33.6 | 12 | 89.0 | 20 | 1 |
| TM | 2000 | n=2 | 31.5 | 15 | 72.4 | 19 | 0 |
| | | n=1 | 33.3 | 15 | 97.5 | 19 | 1 |
| | | SRT | 50.9 | 17 | 139.2 | 19 | 1 |
| TM | 1998 | n=2 | 26.3 | 13 | 55.7 | 16 | 0 |
| | | n=1 | 28.0 | 14 | 59.4 | 16 | 0 |
| | | SRT | 52.1 | 15 | 87.7 | 16 | 2 |
| TM | 1995 | n=2 | 33.0 | 16 | 100 | 19 | 1 |
| | | n=1 | 33.6 | 12 | 112 | 19 | 1 |
| | | SRT | 41.5 | 15 | 137.1 | 19 | 1 |

exceeds 75 m. For the SRT-transformation, there are one or two GCPs with this characteristics. The comparison of the statistics also shows that the affine and the quadratic transformation yields very similar results, which confirm the results of the coefficient comparison.

**Validation with the ATKIS road-map** Apart from the direct comparison of the transformation with the selected GCPs, it is necessary to compare the results of the image-to-map transformation with an additional digital map since the transformations are adjusted to the GCPs and not to the map itself, i.e., the transformation might be optimized for the coordinates of the chosen GCPs, but not necessarily give a correct transformation for the map itself.

Since the ATKIS road map was transformed to a geocoded raster image, the satellite image has to be converted to this format. Therefore, the satellite image was resampled using the nearest neighbour method (Schowengerdt, 1997). An example of a resampled map-based image is shown in figure 3.5 for the image 196/022; May 7, 2000. The image was geocoded to UTM coordinates in zone 32 for the northern hemisphere. A detailed description of this coordinate system can be found in Snyder (1984, Chapter 8).

The image is too large to recognize different small scale features in the shown representation Figure 3.6 shows four image clips from different regions of this image. The clips are displayed in false colour representation using channels

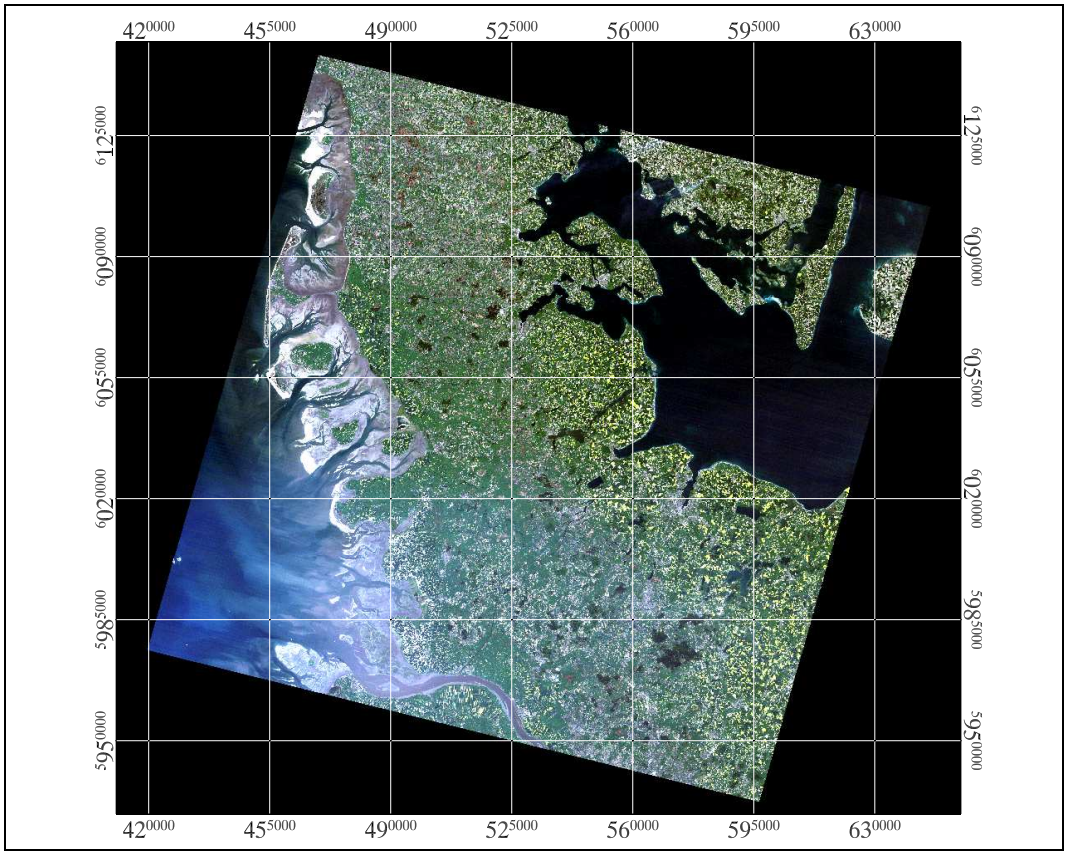



Figure 3.5: Example for a geocoded TM image. The image was acquired on May 9, 2000. The map system is UTM Zone 32 North and the left axis shows the northing and the lower axis the easting values. The easting and northing values are noted in Meters. Original data: LANDSAT TM ©ESA, 2000. Distributed by Eurimage.

3, 4 and 5 of the TM sensor. This representation allows to distinguish different field crops and therefore allows to identify roads with a width smaller than the size of a pixel since the field borders also indicate these roads. Roads broader than the pixel size appear in blue or green in Figure 3.6, which can be seen in the original clips on the left. The roads from the ATKIS road map are overlaid as white lines in the other clips. These comparisons indicate a good agreement with the clips 1 and 4 for all transformations. The results for the clips 2 and 3 show a deviation of about one to two pixel for the polynomial transformation, whereas the result for the SRT-transformation shows larger deviations with about the size of four pixels. These results correspond well to the previously discussed deviation obtained by evaluating the GCPs. Moreover, the difference between affine and quadratic transformations are minimal.

**Conclusion** The TM images can be geocoded by all of the three transformations with an accuracy acceptable for this study. The affine and quadratic transformations are produce mostly identical results, thus it is not necessary to use the quadratic transformation to correct the remaining distortions in

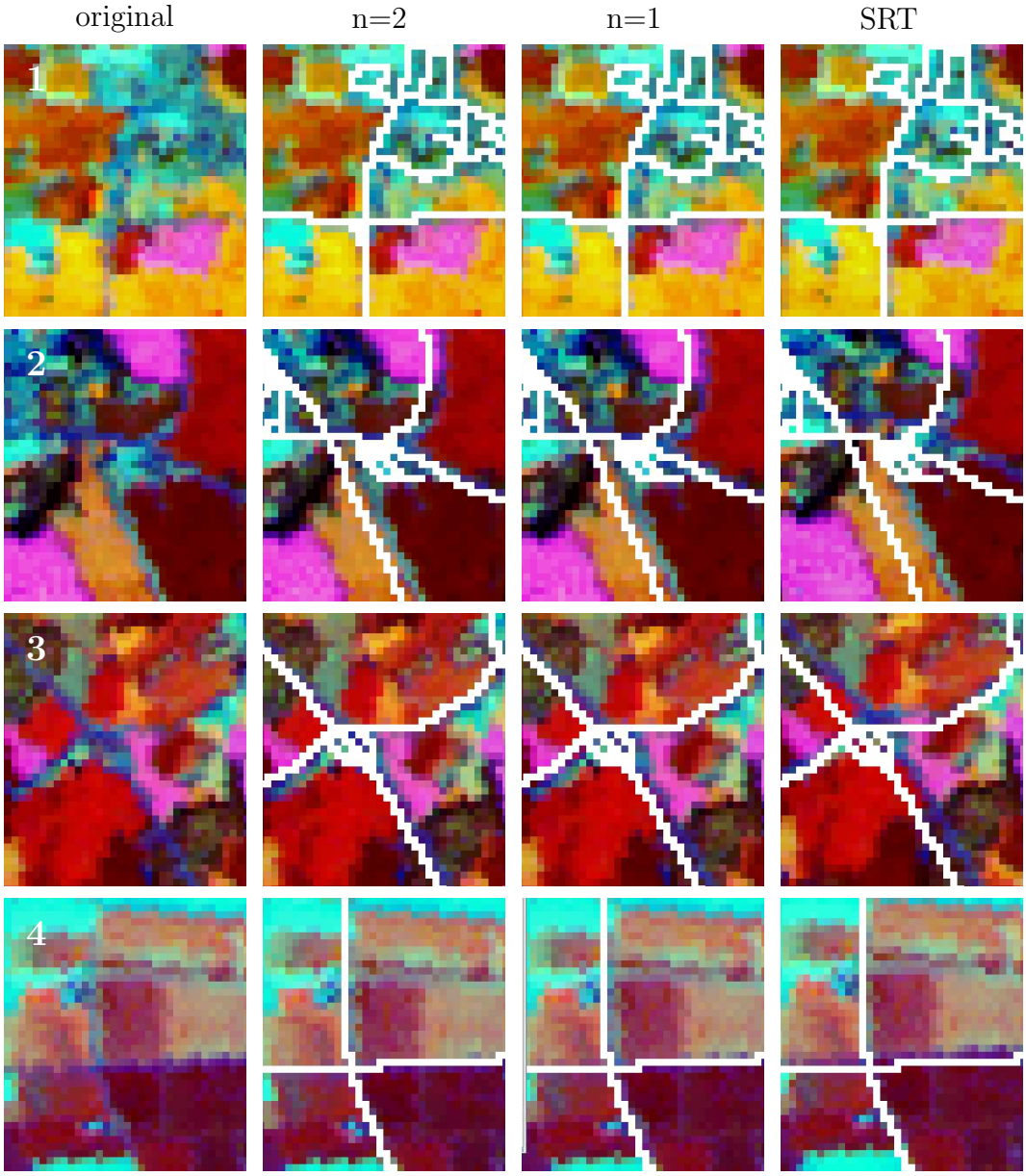




Figure 3.6: Validation of the polynomial transformation with the ATKIS street map. The clips are taken from different parts of a TM image (196/022; May 7, 2000) to show the quality of the different transformations. In the left column, the original images are displayed as a false colour image with the channels 4, 5 and 3. The second, third and right-most columns show the result of the GCP correction using 2nd order polynomials, 1st, and SRT transformations, respectively. The different positions of the clips are shown on the right map. The correction using the polynomials for $n = 1$ and $n = 2$ yields very similar results, whereas the SRT-Method shows larger differences for the eastern side of the satellite image. Original data: LANDSAT TM ©ESA, 2000. Distributed by Eurimage.

the satellite image. Furthermore, the affine transformation is less sensitive to small errors in the coordinates of the selected GCPs.

The SRT-transformation shows a higher deviation for the obtained map coordinates. This indicates that there is still a shear in the geometrically corrected data, which might be induced by the elevation in the eastern regions of Schleswig-Holstein. This could be evaluated by using an additional DEM which was not available for this study. Therefore, the most suitable transformation is the affine transformation which is therefore used to obtain the coordinate transformation for the remaining data.

### 3.1.4 Image-to-Image Registration

The classification algorithm and the atmospheric correction require the identification of corresponding pixels in different images. These images originate either from equal or overlapping frames. If the image-to-map transformations $f_1$ and $f_2$ for both images are at least as accurate as the pixel size, corresponding pixels can be identified by using these transformations. The map coordinates $(x'_1, y'_1)$ and $(x'_2, y'_2)$ of two corresponding pixels should be in the range of the sensor's spatial resolution, and thus corresponding pixels can be identified by using the image-to-map transformations and the image coordinates $(x_1, y_1)$ and $(x_2, y_2)$:

$$(x'_1, y'_1) = f_1(x_1, y_1) \approx f_2(x_2, y_2) = (x'_2, y'_2) \tag{3.25}$$

Unfortunately, the image-to-map transformations are not accurate enough for this relation. Hence, an additional method has to be applied to adjust the transformation.

As discussed in section 3.1.3 the remaining location errors in the images are quite small compared to the size of the image. Therefore, most of these deviations can be compensated for smaller regions by a simple translation. Assuming $(s_o, t_o)$ to be the optimal translation vector, the registration of the second image to the map coordinates of the first can be expressed by

$$(x'_1, y'_1) = f_2(x_2, y_2) + (s_o, t_o) = (x'_2 + s_o, y'_2 + t_o) \tag{3.26}$$

Therefore a method is necessary to calculate the translations for different parts of the image.

As stated above, the accuracy for the GCP-based image-to-map transformations is about 75 m, which equals 2.5 TM pixels. An error in applying both transformations can result in an overall deviation of about 150 m or five pixel. Although this is not sufficient for the identification of corresponding pixels, the transformations are accurate enough to identify corresponding regions or image clips in both images. With the above assumption that the local deviation of the image clips is merely a shift, the vector $(s_o, t_o)$ can have values ranging from -5 to 5 pixels in $x$ and $y$ direction.

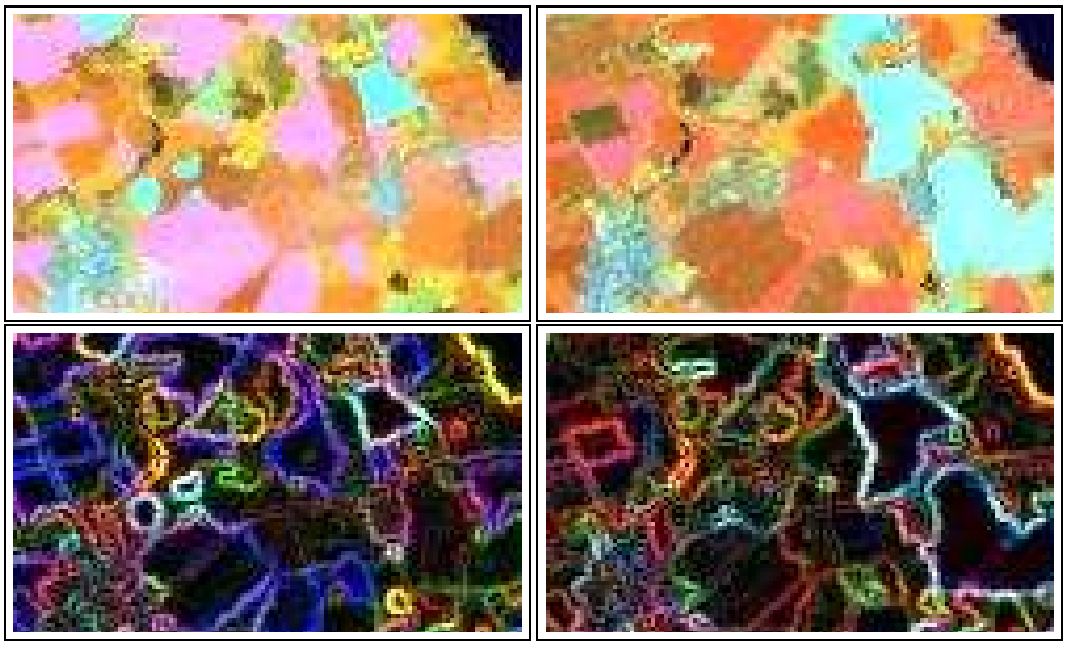


Figure 3.7: Comparison of original (top) and Sobel/Gradient (bottom) filtered clips of the images 196/022; May 9, 2000 (left) and 196/022; May 2, 2001 (right). Top: False colour images of channels 4, 5 and 3. Nearly all agricultural crops have changed from 2000 to 2001, which is indicated by the different colours. Bottom: False colour representation of the Sobel filtered radiance for the same image clips. The field boundaries are clearly visible. Some field boundaries have changed in between the years, but the majority remains unchanged. The displayed region is located south of the city of Eckernförde. Original data: LANDSAT TM ©ESA, 2000 and LANDSAT ETM+ ©ESA, 2001. Distributed by Eurimage.

### Gradient/Sobel Filtered Images

In order to identify the solution for $(s_o, t_o)$, it is necessary to quantify how well two images match. This requires the images to be comparable.

The satellite images used in this study are acquired in different years and the agricultural crops have changed most likely in between the years due to crop rotation (see Section 2.1.3 (p. 22)). This is demonstrated in Figure 3.7 by different colours in the upper image clips, which indicate the changes of crops cultivated on the field from 2000 to 2001. Therefore, the reflectance of similar fields usually changed in between the years and makes it difficult to compare the original images directly for agricultural regions.

Unlike the surface types of different fields, the field borders are seldom changed. The field borders can be determined by a gradient filter (Castleman, 1996). The lower clips in Figure 3.7 show a gradient image, calculated by applying the Sobel operator[5] to the radiance of each of the three channels. The results obtained for the three channels are displayed using the same false colour

[5]A Sobel operator is only one type of several possible gradient operator, nonetheless, here, the term gradient image is used synonymously with the term sobel-filtered image.

representation as in the original image clips. All field boundaries are clearly visible. The different colours of the edges indicate that all three channels have to be used in order to identify the borders of all fields with different crops.

The comparison of the gradient images clips in 3.7 shows that the majority of edge shapes remained unchanged in between 2000 and 2001. But since the colours of the edges are different for equal edges, it is necessary to use the gradients of all channels. Therefore the mean $G$ of the three gradients for channels 3, 4 and 5 is calculated for each pixel. This provides coordinate dependent gradient images for each image.

### Correlation of Image Clips

Two gradient images from different years, like the ones displayed above, can be compared directly. A commonly used method to judge the quality of image matching automatically is the correlation method (Schowengerdt, 1997). This method uses the correlation coefficient $\gamma$ which describes the degree of linear dependency between two data sets. In image processing, correlation is usually applied to corresponding image values, where higher correlation indicates better matching. In this case the mean gradients $G_1(x', y')$ and $G_2(x' + s, y' + t)$ with the translation $(s, t)$ are used as data sets:

$$\gamma(s,t) = \frac{\sum_{x'}\sum_{y'}[G_1(x',y') - \bar{G_1}][G_2(x'-s,y'-t) - \bar{G_2}]}{\sqrt{\sum_{x'}\sum_{y'}[G_1(x',y') - \bar{G_1}]^2 \sum_{x'}\sum_{y'}[G_2(x',y') - \bar{G_2}]^2}} \tag{3.27}$$

The highest value for the correlation $\gamma(s_o, t_o)$ indicates the best solution for the shifting vector (Gonzalez and Wintz, 1987, Chapter 8).

### Complete Frame Registration

The above described correction by a translation cannot be applied to the complete satellite image since the distortions cannot be described for the complete image. Nonetheless, the correction by a translation can be applied to correct smaller clips of the image. In order to register a complete frame, it is necessary to identify different shifting vectors for different regions of the image. In each of these regions it is necessary to identify image clips in both images, which allow to apply the correlation of the gradient images.

A simple solution is to partition the image according to the shortest distance to a GCP. This divides the image into as many regions as there are GCPs. This method is suitable because:

- The surface features selected as GCPs are the same for all images and allow to identify corresponding regions.
- The region in the vicinity of a GCP is structured, which is also necessary for the selection of GCPs.

- The image can be divided into at least 14 regions with different shifting vectors. This is sufficient if the different shifting vectors from neighbouring regions do not change by more than one pixel.
- The correlation is not possible in cloud-covered image clips, which is also valid for the selection of the GCPs. Therefore, regions with a GCP are usually cloud free or only partly cloud-covered.
- The selected region must be present in both images; This is true if corresponding GCPs appear in both images. This is important for overlapping frames, where only the overlapping region is present in both images.

With this method, the image is divided into a number of regions that depends on the number of selected GCPs. For each GCP, the following steps are performed:

1. GCPs are identified that are present in both images.
2. An image clip of $50 \times 50$ pixels with the corresponding GCP in the centre of the clip is selected in the first image.
3. A second clip with the doubled size is selected in the second image.
4. The mean gradients $G_1$ and $G_2$ are calculated with the Sobel operator for both clips.
5. The region is compared to the cloud cover dataset for both images. If clouds are present, the corresponding values for $G_1$ and $G_2$ are set to 0 for cloudy pixels in both clips.
6. Equation (3.27) is calculated for $s$ and $t$ ranging from -15 to 15 in steps of one for the gradients of the clips. In order to use only steep edges and mask out clouded regions, only pixels with $G > 0.75(\max(G) - \min(G)) + \min(G)$ are used.
7. The combination of $s$ and $t$ with the highest value for $\gamma(s,t)$ is selected as shifting vector.

This is applied to all GCPs of the second image and the results of each image are a list of shifting vectors $(s_i, t_i)$ where $i$ is the number of the corresponding GCP. The identification of corresponding pixels is obtained by adding the calculated shift $(s_i, t_i)$ for the nearest GCP $i$ to the image coordinates $(x_2, y_2)$ in equation (3.25).

Since map coordinates are not necessary for the processing of the images, the back transformation $g_2$ and the results of the shifting vector are used to identify corresponding pixels in two different satellite images with the following equation:

$$(x_2, y_2) = g_2(f_1(x_1, y_1) - (s_i, t_i)) \tag{3.28}$$

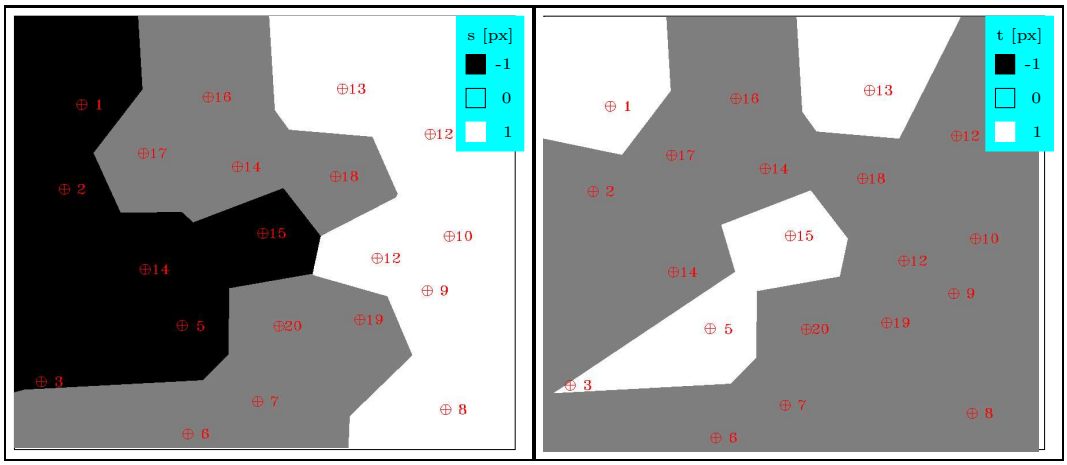



Figure 3.8: Results of the correlation method for the complete image 196/022; July 7, 2000 (see Figure 3.4). Displayed are the components of the shifting vector, $s$ (left) and $t$ (right) and the GCPs with their numbers. The image is registered to the image 196/022; May 11, 2001. GCP 3 is not used since it is outside the image acquired in 2001.

### Results for the Image-to-Image-Registration

The correlation method is applied to all images available in this study. Each image is registered to the image of the same frame of the year 2001. This is necessary to build an invariant map (see 3.2). The results of the correlation is used to assign a shifting vector to every pixel in the image.

The result of this assignment is shown in Figure 3.8 which displays both components of the shifting vectors for the TM image 196/022;2000. It is correlated to the image from 2001 of the same frame. The shifts calculated for the different images are always smaller than two pixels. Additionally, the regions that belong to different shifts are homogeneously distributed. The maximum deviation found by applying the correlation method to the other images of the same frame is 2 pixel. Thus, the correlation method allows to identify corresponding or at least neighbouring pixels in different satellite images originating from the same frame.

An important application of the image-to-image correction is the comparison of training data of overlapping frames (see Section 4.4.1, p. 125). An example of the translation calculated for overlapping images is displayed in Figure 3.9. The corrected image originates from the frame 195/022 which is adjacent to the frame 196/022. Both images are from the year 2001. The frames overlap only in the lower left corner of the image 195/022 and thus only some GCPs are used. Similar to the shifting vector of images from the same frames, the deviation of the shifting vector in the case of overlapping is also small. However, the absolute value is higher than expected though since only a deviation of five pixels was expected.

The results of the image-to-image registration shown in the Figures 3.8 and

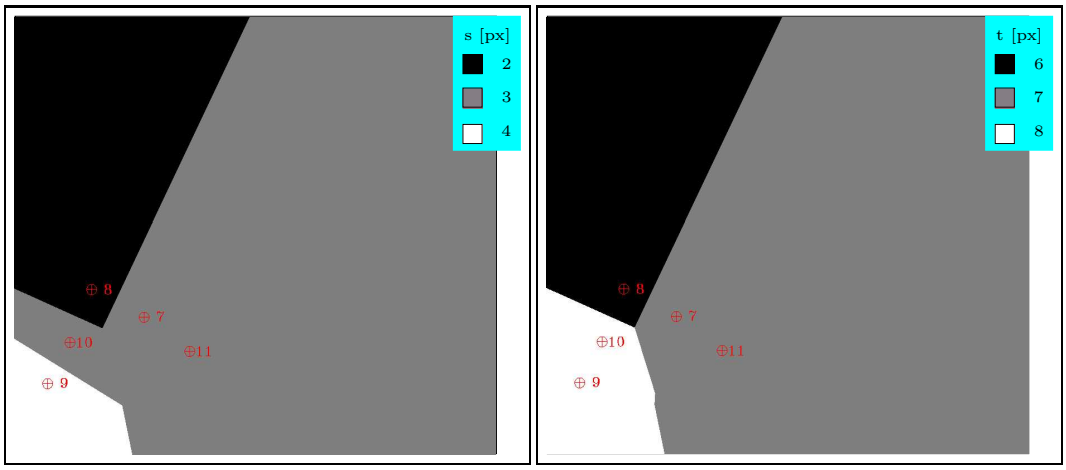



Figure 3.9: Result of the correlation method for neighbouring frames. The shifting vector is displayed for the image 195/022, May 11, 2001). It is registered to the adjacent on the left frame (196/022, May 11, 2001).

3.9 are expedient and demonstrate that the method is working correctly. In order to show the results of this registration directly, figure 3.10 shows a mosaic composed of two different TM images. The squares are alternately displaying the data and gradients from both years where the coordinates have been determined by the image-to-image registration. The year 2001 is represented by blue squares and 1995 by green squares. It can be seen that the images match very well at the transitions edges from one square to the next one for both the original and the gradient image.

## 3.2 Atmospheric Influences

As discussed in Section 1.4.2 (p. 10) the radiation reflected by the surface is modified by scattering and absorption of gases and aerosols in the atmosphere. Moreover, the incoming sunlight is to some extend directly reflected to the sensor by aerosols in the atmosphere which can be seen as clouds and haze in the satellite images. These effects of the atmosphere have to be taken into account for an accurate interpretation of satellite images.

Applications that are based on knowledge of the surface reflectivity require an accurate correction of atmospheric influences (Thome, 2001; Moran et al., 2001, 1992; Zhao et al., 2000; Du et al., 2002; Wrigley et al., 1992; Hall et al., 1991). Although there are various automatic correction algorithms available (Richter, 1997), these corrections require information on the atmospheric composition, e.g, information on temperature, humidity or liquid/ice water content. This information is not available from the image to be classified and has to be obtained from alternative sources like, e.g, radiosonde measurement.

Fortunately, the classification of surface types do not require such a correction, if the training data sets (see Section 4.1.2, p. 89) are taken from the

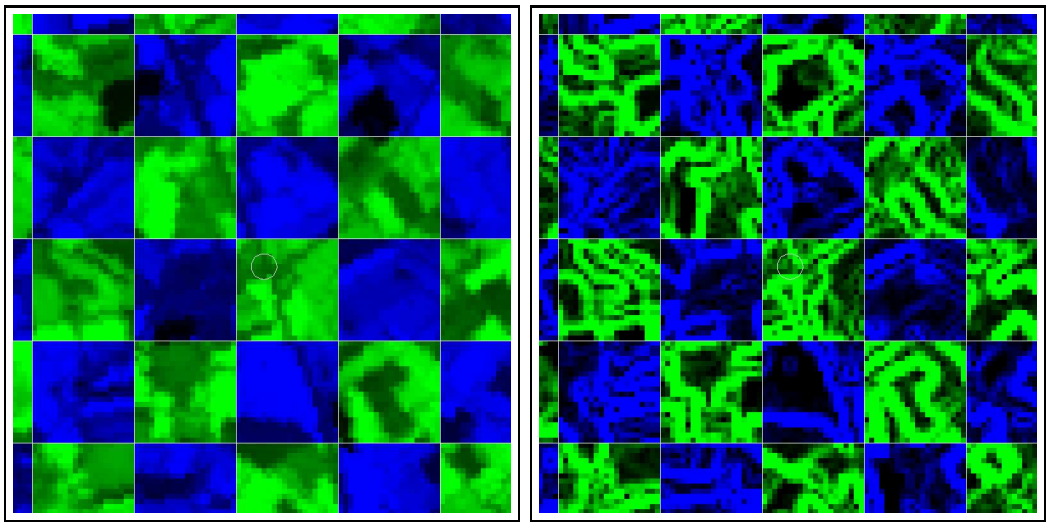


Figure 3.10: Mosaic of channel 4 of TM from the years 2001 (blue) and 1995 (green) for the original (left) and the sobel filtered data (right). The clips are from the frame 196/022 with GCP 19 (see Figure 3.4) at the centre. Both, original and filtered mosaic show a smooth transitions of the structures from the rectangles of the different images. Original data: LANDSAT TM ©ESA, 1995 and LANDSAT ETM+ ©ESA, 2001. Distributed by Eurimage.

image to be classified (Song et al., 2001). This procedure compensates for all atmospheric influences that are constant regarding the complete image, i.e., Rayleigh-scattering and absorption by homogeneously distributed gases like oxygen or carbon dioxide (Liang et al., 2001, 2002). The remaining influences in optical remote sensing are scattered by aerosols and molecular absorption by variable gases like ozone or water vapour.

Ozone mainly absorbs radiation of shorter wavelengths (0.4 to 0.7 $\mu$m) (Asrar, 1989) and thus affects the TM channels 1 to 3. The first two channels are not used for the classification of canola (see Section 4, p. 87) but since the third channel is influenced, the classification algorithm can generate incorrect results if ozone is not homogeneously distributed in the satellite image. Fortunately, the frame sizes of TM and LISS are quite small and ozone only shows small day-to-day variations (Tanré et al., 1992) which implies also a spatial homogeneous distribution.

The third channel is necessary and may be affected by ozone (Asrar, 1989). According to Tanré et al. (1992); Forster (1984) ozone is homogeneously distributed in comparison to the size of a AVHRR image and since this has five times the size of a TM image, this is also valid for TM and LISS/3.

Absorption by water vapour affects the NIR and MIR wavelengths and therefore the channels 4,5 and 7 of TM (Asrar, 1989). Channels 4 and 5 are used for the classification and therefore the classification may be affected. The variations of the water vapour concentration are on a much larger scale than the sensor resolution and most images were acquired during fair weather

conditions in which the water vapour is distributed homogeneously. Therefore the correction or water vapour influences is generally not necessary although it might affect the classification if the training data sets are located far from the region to be classified or in cloudy regions. These special cases will be discussed in Section 4.2.3 (p. 116).

Ozone and water vapour are usually distributed homogeneously compared to the size of a TM or LISS/3 image. Still, it is possible that these gases change within one image and impede a correct classification. Apart from the concentration of the gases, the influence on the classification also depends on the classification method and the surface type to be observed. In this study, no misclassification was found that could be linked to variations of ozone. Nonetheless, it might be necessary to correct the influence of ozone absorption for a different set of satellite data or the classification of other plants. In this case, it is possible to use information on ozone concentration, that are available through additional sensors like Global Ozone Monitoring Experiment (GOME) or Scanning Imaging Absorption Scatterometer for Atmospheric Cartography (SCIAMACHY) or climatology data (Erbertseder et al., 1999; Ouaidrari and Vermote, 1999).

Water vapour definitely has an influence on the classification, but variations on the small scale of the satellite data are linked to the formation of clouds (Richter and Lüdeker, 1999) and can therefore be corrected with the cloud algorithm.

In contrast to the gaseous components of the atmosphere, aerosols can be distributed very inhomogeneously and have a larger effect on the measured radiance than the gaseous components of the atmosphere. Therefore they have to be discussed in detail.

### 3.2.1 Influence of Aerosol Scattering on Satellite Images

Aerosols are suspended droplets or particles that can scatter or absorb incoming sunlight or light reflected by the surface. Clouds and cloud shadows are the most obvious influences of aerosol scattering on optical remote sensing data. The aerosols responsible for clouds are water droplets or ice crystals, which are the most common aerosols in the atmosphere. Another aerosol is dust which is rarely observed with the necessary concentration to affect the classification in central Europe, but has to be taken into account for arid regions.

The simplest way to take clouds into account is to use only cloud free images. This is not possible because the number of available images in the period suitable for canola detection is limited (see Section 2.1.3, p. 22). Besides small or thin clouds often are not visible in the preview image (quicklook) used to select and order the suitable full resolution images from the satellite data provider. Therefore a number of partly cloud-covered images have to be used as well and it is necessary to identify clouds in the images.

This can be dealt with by masking out partly cloud covered regions completely, but more desirable is an accurate cloud identification since it allows to

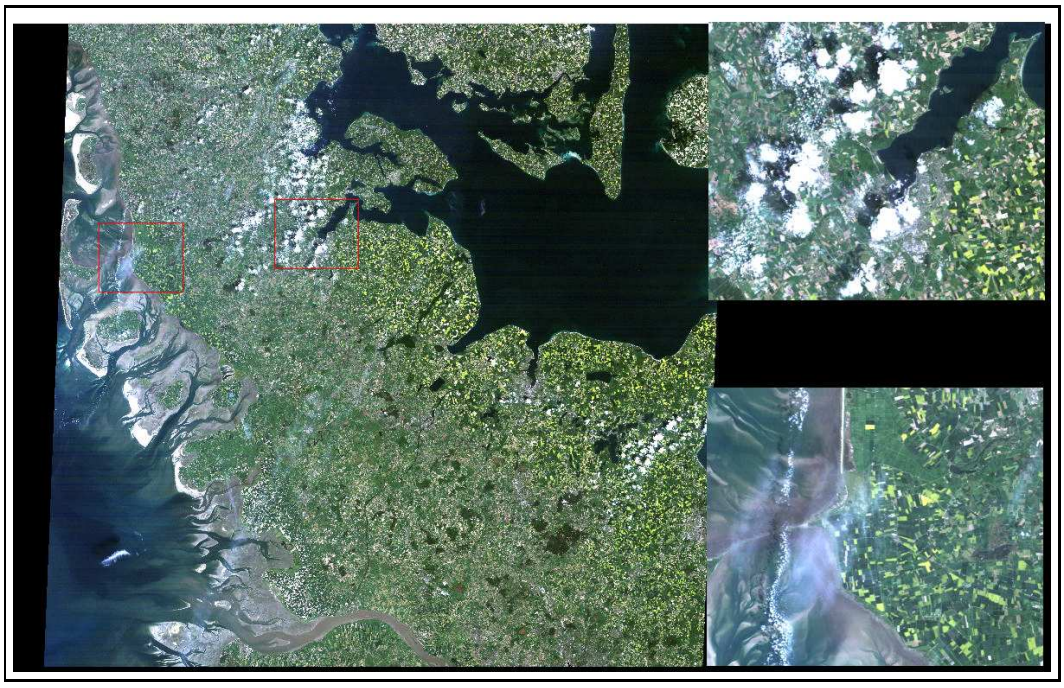


Figure 3.11: Example for a clouded image (frame 196/022; May 4, 1998). The image clips on the right side of the figure show an enlargement of two different region (marked by the red rectangles). The upper clip shows a number of cumulus clouds and their shadows and the lower clip a hazy region. Original data: LANDSAT TM ©ESA, 1998. Distributed by Eurimage.

use the satellite images more efficiently.

Figure 3.11 shows an example of a partly cloud-covered TM image 196/022; May 4, 2000. The two image clips on the right side illustrate the effects of different clouds. Since the influence on the classification depends on the type of cloud influence, it is convenient to distinguish three classes of cloud influences: thick clouds, cloud shadows and haze. These three classes have the following characteristics and effects on a surface type classification:

**Clouds:** Here, clouds are layers of aerosols, that are opaque, thus it is not possible to get information from the earth surface underneath (see the upper right clip in Figure 3.11). This has no effect on the classification itself since the spectral signature of clouds is quite distinct from that of canola and other plant covers. Nonetheless, it is necessary to get obtain the size of the area that is not accessible by the sensor since there might be some canola in these areas.

**Cloud Shadow:** Cloud shadows are directly related to clouds, since clouds scatter a large part of sunlight on its way to the surface (see the upper right clip in Figure 3.11). A classification in shaded areas is still possible, but requires a different set of training data sets or an adaption of the known training data sets. Regarding a whole image, the parts of shaded areas can be neglected and the improvement with such an adaption is

not worth the effort. Therefore cloud shadows will be treated similarly to clouds as regions, where classification is not possible. Along with clouds, their extent still has to be measured.

**Haze:** Haze refers to thin clouds that are mostly transparent (see the lower right clip in Figure 3.11). Usually haze in TM images results from high clouds or mist. The influence of haze depends on the wavelength: the shorter the wavelength, the stronger is the influence on the signal received by the satellite sensor. A simple method to minimize the influence of haze is to use only the bands with longer wavelength. However, even for longer wavelengths, there is still some influence left that can affect the classification results, which will be discussed in detail in Section 3.2.4 (p. 71). Therefore, haze has to be detected and quantified for the detection of canola, even for a single image classification.

Consequently, the image classification requires methods to identify these three classes. The following sections describe the methods used to identify cloud-covered and shaded regions in the images and build masks for cloud-covered and shaded pixels in the satellite data. These masked pixels are not used for a classification. Besides these masked out pixels another mask is created for pixels that are haze-covered. These pixels are still used for the classification but the effect of the haze is quantified and considered for the classification. The method used to quantify this effects is presented here. Its correction is discussed later in Section 4.2.3 (p. 116).

### 3.2.2 Cloud Detection

The properties that allow to identify clouds with multispectral sensors are the high reflectivity in the VIS, NIR and MIR part of the spectrum and the low cloud top temperature (Mölders et al., 1995). These properties allow to build two kinds of masks, a bright mask which is constructed from the bright pixel and a cold mask which is derived from the cold pixels in a satellite image. A combination of these two masks can delineate clouds.

#### Bright Mask

Clouds have a height reflectivity in the VIS, MIR and NIR part of the spectrum, since the cloud particles, i.e., ice crystals and small water drops, strongly scatter light at these wavelengths. Thick clouds scatter the main part of the radiation back into space and appear bright in the satellite image. Since the scattering is mainly Mie-scattering (see Section 1.4.2, p. 11) the amount of scattered light is independent of the wavelength. This is illustrated in the left two image clips in figure 3.12. The white colour of clouds in both clips demonstrates the high reflectivity of the clouds for all channels at these wavelength. Therefore a first criterion for a clouded pixel is the brightness. Channels 3

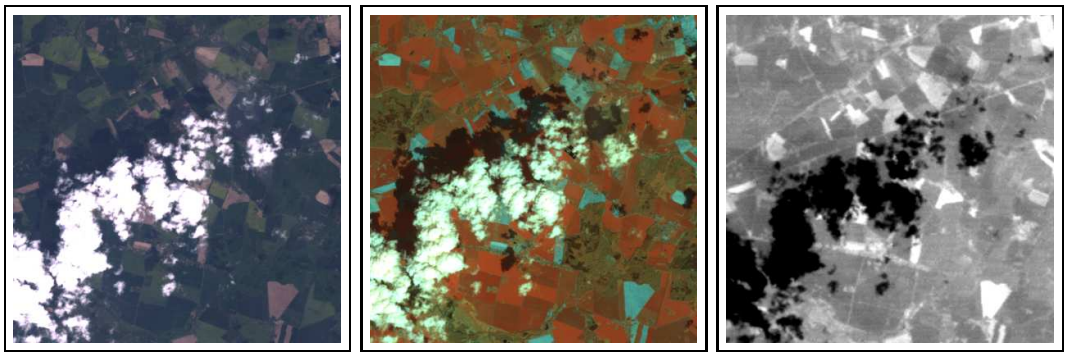


Figure 3.12: Example for the effect of cloud cover in different channels. Left: true colour image clip of the channels 1 (blue), 2 (green) and 3 (red). Middle: false colour image clip of the channels 4 (NIR), 5 (MIR) and 7 (MIR). right: channel 6 (TIR). The two clips on the left illustrate the higher sensitivity to clouds for the shorter wavelength. The right clip demonstrates the possibility to detect cold surfaces by using the thermal infrared channel of the TM sensor. Original data: LANDSAT TM ©ESA, 2001. Distributed by Eurimage.

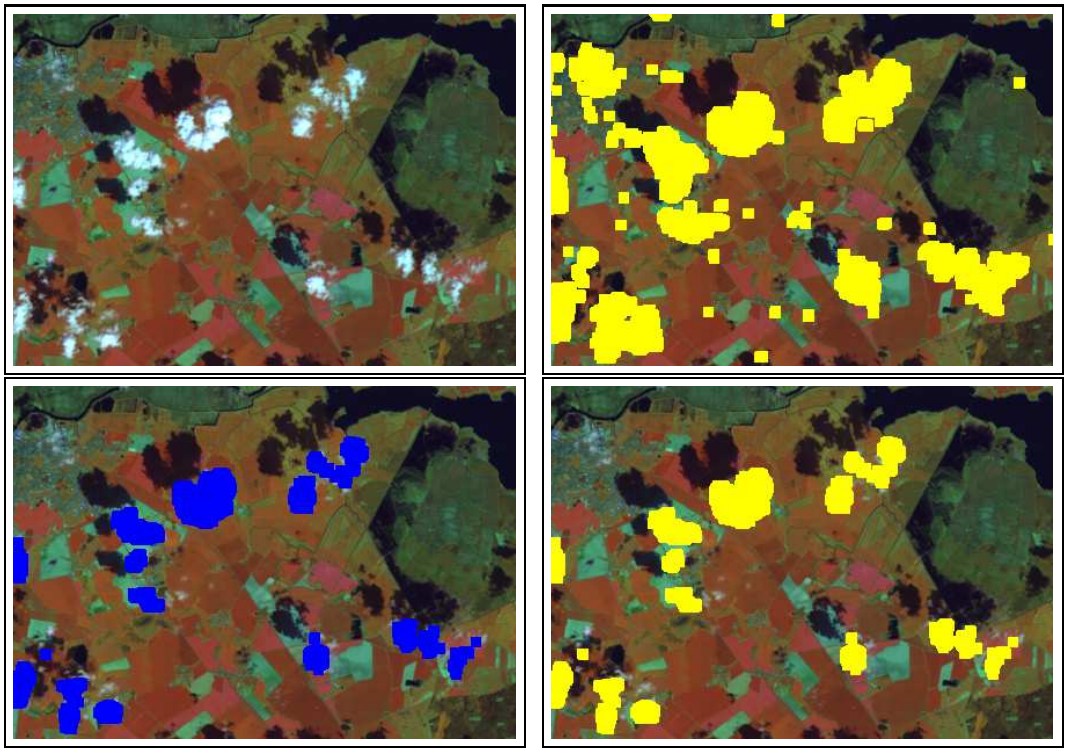


Figure 3.13: Example for the cloud detection. Top left: original cloud-covered clip of a TM frame (196/022;2001) from Eastern Germany near the Baltic Sea. Top right: mask of all pixels brighter than the thresholds for channel 3 and 4 of TM. Bottom left: cold mask derived from the TIR band of TM by masking out all pixels colder than the temperature threshold. Bottom right: final cloud mask created from Bright and Cold mask. Note that neighbouring pixels in a distance of 150 m (5 pixel) are also declared as cloud-covered, bright or cold pixels. Original data: LANDSAT ETM+ ©ESA, 2001. Distributed by Eurimage.

(red) and 4 (NIR) are selected to identify bright pixels; each pixel that exceeds a certain radiance in both of these channels is added to the bright mask. The upper right clip in figure 3.13 shows the result for this bright mask. This clip, also contains other bright objects such as industrial areas with concrete. Other surface types that exceed the threshold are sand at beaches or sand pits. Thus another physical property of clouds is necessary for an unambiguous identification.

### Cold Mask

Clouds are usually higher (about 2000 m for fair weather conditions) than the earth surface. In fair weather conditions the temperature for the atmosphere decreases by about 0.7 K every 100 m. Therefore the top of these clouds is about 14 K colder than the surface temperature. The temperature of the surface and at the cloud top can be measured by using the TIR channel of the TM sensor (see Section 1.4.1, p. 6). The right clip in Figure 3.12 shows the thermal channel of a TM image. Colder regions appear darker than warmer ones. Thus, clouds appear black in this clip and the TIR channel of TM can be used to construct the cold mask. An example for the cold mask is displayed in the lower right clip in figure 3.13.

As stated in Section 1.4.1 (p. 6) the emissivity in the TIR is usually about 0.98 to 1.0 for the surfaces in the study area. An exception are water surfaces which can have an emissivity of 0.95 and below. This can result in an underestimation of the temperature and thus to a misclassification of water pixels. Therefore, the cold mask is also not sufficient to identify clouds on its own and only a combination of cold and bright masks allows to identify cloud-covered regions unambiguously.

### Bright and Cold Thresholds

The thresholds for these masks have to be determined since they depend on the season and the geographical region. Stowe et al. (1991); Di Vittorio (2002); Saunders and Kriebel (1988) used a statistical approach to determine these thresholds automatically. Since the images used in this study are obtained in fair weather conditions in the time span from spring to early summer and from the same geographical region, it is not necessary to adapt the thresholding values for different satellite images. Therefore these values are determined by visual inspection of 8 partly cloud-covered satellite images.

Every pixel with a radiance in the red channel of more than $50.73\,\mathrm{W/(m^2 sr \mu m)}$ and a radiance in the NIR channel of more than $129.30\,\mathrm{W/(m^2 sr \mu m)}$ is added to the bright mask. The cold mask consists of pixels with colder temperature than 283.85 K. The temperature is determined with the TIR channel and equation (2.3) in Section 2.2.1 (p. 31).

In order to mask out the thinner borders of the cloud and to take the lower spatial resolution of the thermal infrared channel (see Table 2.4 (p. 30) into account, pixels that are next to cloud-covered pixels are also added to the cloud

mask. This is necessary since the high reflectivity of the clouds can overshine neighbouring pixels (see Section 1.4.2, p. 11). Thus, all pixels that are in the range of 150 m (5 pixels for TM) from a cloud-covered pixel are also declared cloud-covered.

#### Cloud Mask

The final cloud mask is the combination of both masks. Each pixel that is present in both, the cold and the bright masks is added to the cloud mask. The lower right clip in Figure 3.13 shows an example of the cloud detection with this method. It illustrates that all larger clouds in this image are identified by the thresholding method. There are some smaller clouds and some cloud borders that could not be detected. This is caused by the lower resolution of the TIR band of TM, that averages the temperature over four regular TM pixels, and by clouds that can be categorized as haze. These clouds are detected with the haze detection scheme described later.

### 3.2.3 Cloud Shadow Detection

The determination of the cloud shadows is not possible with a simple thresholding method. Nonetheless, thresholds are necessary to identify regions that are probably cloud shadows. Similar to the cloud mask, these regions are also delineated with two attributes. These are used to construct two further masks, the dark and the water mask. Similar to the method used by Simpson and Stitt (1998) for AVHRR images these two masks can be used to calculate the projection from sunlight of clouds on the surface, i.e. cloud shadows.

#### Dark Mask

Cloud shadows are dark, which is also a result of the strong scattering of light by clouds. The direct sunlight is scattered by the cloud and does not reach the surface. The light reflected from shaded regions originates from indirect sunlight scattered in the atmosphere. Therefore, all pixels having low brightness are added to the dark mask (see below). An Example for the dark mask is given in the upper left clip of figure 3.14. This dark mask still includes a number of regions that are not shaded but are dark themselves. These regions are dark water surfaces as seen in the upper right of this clip or dark woods, mainly coniferous woodland, as visible in the lower left of this clip.

#### Water Mask

Since dark waters cover large areas of the satellite images, these surfaces have to be identified. The water surfaces can easily be detected by the normalised difference vegetation index (NDVI) (Asrar, 1989, p. 120). Actually, the purpose of the NDVI is to quantify the photosynthesis with satellite data. But since water has the lowest NDVI of all surface types in a satellite image, it

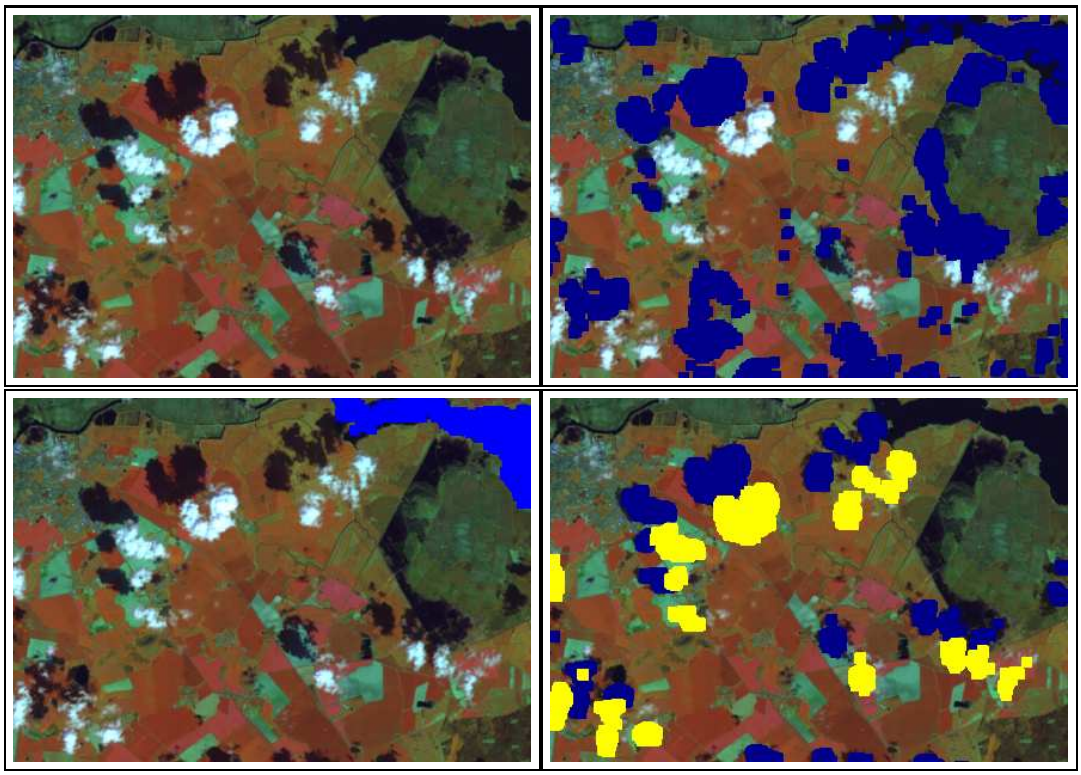


Figure 3.14: Determination of the cloud shadow mask. Top left: original image as displayed in Figure 3.13. Top right: dark mask (blue) including all pixels darker than the threshold in channel 3. Bottom left: the water mask; all pixels with an NDVI below the threshold. Bottom right: Final cloud and cloud shadow mask; created from the determined azimuth angle and the calculated shifting vector $d$ between cloud and initial cloud shadow mask. Original data: LANDSAT TM ©ESA, 2001. Distributed by Eurimage.

is perfectly suitable to identify water surfaces. The NDVI is the normalized ratio of red and NIR radiance. Practically, the NDVI is sensor dependent and the NDVI for TM can be calculated with the radiances $I_{\mathrm{ch3}}$ and $I_{\mathrm{ch4}}$ of the channels 3 and 4:

$$\mathrm{NDVI}_{\mathrm{TM}} = \frac{I_{\mathrm{ch4}} - I_{\mathrm{ch3}}}{I_{\mathrm{ch4}} + I_{\mathrm{ch3}}} \tag{3.29}$$

In oder to produce a water mask, all pixels with an NDVI below a threshhold are added to a water mask. The lower left clip in Figure 3.14 shows such a water mask.

### Dark and Water Thresholds

As for the bright and cold masks, the dark and water thresholds are also determined by visual inspection of the 8 images used for the determination of the cloud threshhold. Each pixel that has a NIR radiance below $23.70\,\mathrm{W}/(\mathrm{m}^2\mathrm{sr}\mu\mathrm{m})$ is assigned to the dark mask. Pixels with a $\mathrm{NDVI}_{\mathrm{TM}}$ smaller than -0.4 are assigned to the water mask.

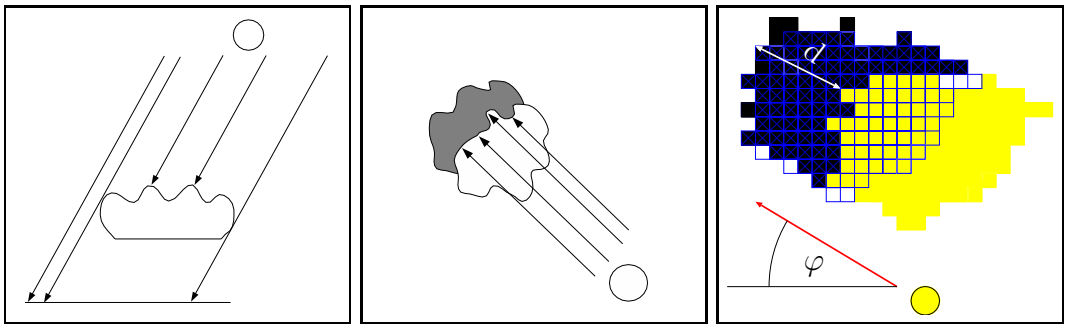



Figure 3.15: Sketch of the illumination conditions for clouds and the underlying surface. The position of the cloud shadow depends on the incident angle of the sunlight. Left: the distance of the cloud shadow from the cloud position depends on the sun declination and the height of the cloud top. Middle: the direction of the cloud shadow depends on the sun azimuth angle. The shape of the shadow is depending on the horizontal shape of the cloud and to a lesser extent on the vertical structure. Right: Sketch of the determination of the cloud shadow distance $d$. The cloud mask (yellow) is shifted pixelwise in the direction of the sun azimuth angle $\varphi$ on the dark mask (black) until the number of coincident pixels (marked by the blue crosses) reaches a maximum.

### Shadow Mask

The dark and water mask are used to determine an initial shadow mask by removing the water pixels from the dark mask. This initial shadow mask still includes other dark surface types, mostly coniferous forests. These remaining surface types can not easily be distinguished from cloud shadows by their spectral properties.

To separate these surface types, the cloud mask can be used, since every cloud shadow depends on a cloud and the incoming sun light. The two left sketches in Figure 3.15 illustrate the interrelation of sun, cloud and cloud shadow. The cloud shadows are a projection of the horizontal cloud shapes (see Figure 3.15) in the direction of the sun azimuth $\varphi$. Likewise, it can be seen, that the distance $d$ between cloud and cloud shadow depends on the sun elevation $\Theta$ and the cloud top altitude $h_{\mathrm{CT}}$. For a plane earth surface this altitude can be calculated by:

$$h_{\mathrm{CT}} = d \tan \Theta \tag{3.30}$$

The position of the clouds are known from the cloud mask discussed above. The sun azimuth and declination are available from in the header information of the satellite data. Merely the information on cloud top height has to be determined. This information can be obtained by comparing the cloud mask and the shadow mask. Since the shapes of the cloud shadows are a projection of the horizontal cloud shapes (Simpson et al., 2000). For nadir looking sensors that have a small FOV, the shadow mask is primarily a cloud mask shifted in the direction of the sun azimuth angle. Wen et al. (2001) used this to generate a cloud shadow mask for a single TM image by determining $d$ through visual

inspection.

In this study, an automated solution is required; The distance $d$ can be determined by shifting the cloud mask over the dark mask in the direction of the sun azimuth and counting the number of coincident pixels between these masks. The right sketch in Figure 3.15 shows the principle of this method. The cloud mask (yellow pixels) is shifted pixelwise in the direction $\varphi$. For each shift, the number of coincidences of the shifted cloud mask (the blue grating) and the shadow mask is counted. Coincidences are marked by blue crosses within the pixels. The shift that yields most coincidences between shifted cloud mask and shadow mask is the approximate distance between cloud and cloud shadow since the cloud height is not equal for all clouds.

This method results in a cloud top height that is an average for all clouds. This assumption is only applicable for a mostly constant cloud top height for all clouds. More distinct values can be obtained by applying this method to parts of the image or to single clouds but these improvements require to solve a number of problems, e.g., the differences of cloud and cloud shadow shape resulting from the variation of cloud top hight (see Figure 3.15) and the overlapping of cloud shadows and other dark surfaces. Therefore and since the shifting of the complete image gives satisfying results, this was not performed.

The final cloud shadow mask is determined by shifting the cloud mask $d$ pixels in the direction $\varphi$. The lower right clip in Figure 3.14 shows the result of the cloud and cloud shadow detection. It can be seen that the shift $d$ estimated with this method is correct.

### Conclusion

The above described methods allow to identify cloud-covered and shaded regions for fair weather conditions. Nonetheless, there are still errors in the detection as can be seen in Figure 3.14. These errors are:

1. Smaller clouds are not detected. For instance, the smaller clouds in the upper left part of the clips in figure 3.14 are not identified.

2. The borders of the clouds are not always identified correctly. Even with the 5 pixel border added to each clouded pixel, there are still some remaining cloud pixels (see upper right cloud in figure 3.14).

3. The shadow borders are not identified correctly since the cloud shadows are usually larger than the clouds. The reason for this are the thinner borders of the clouds that have a stronger influence on the shadow than on the clouds[6].

[6]The radiation is scattered by aerosol particels in all directions and this radiation is missing at the surface; Only a smaller part of the scattered radiation is reflected to the satellite. Thus thin clouds have a stronger influence on the effect from cloud shadows than on the effect from clouds.

The reasons for these errors are the lower resolution of the thermal infrared channel and the influence of thin clouds on the cloud shadows. Thus the detection of thin clouds will provide a method to correct these errors and will be described below.

### 3.2.4 Haze Detection

Physically, the difference between clouds and haze results mainly from the amount of scattering. Haze is transparent and the underlying surface is still visible. Therefore, this surface can be classified if the effect of haze on the radiation reflected from the surface can be compensated, i.e., the haze is thin enough. Thus the pixels in a satellite image can be divided into three classes:

- Clear: Pixels that can be classified without any adaptation of the classification algorithm since they are only slightly affected by aerosol scattering.
- Hazy: Pixels that can be classified with an adaption of the classification algorithm to the scattering.
- Cloud-covered: Pixels that have to be excluded from a classification since the effect of aerosol scattering is too strong for a reasonable classification.

These three types of pixels require different treatment in order to be classified. Therefore, the haze algorithm has to accomplish the following tasks:

- hazy regions have to be identified in order to decide if a correction is necessary or possible.
- The influence of the haze has to be quantified in order to adapt the classification algorithm to different extents of haze influence.

Since haze is present in most multispectral satellite images, there have been a number of approaches to identify and correct the influence of haze in these images. The most common method is the dark object subtraction (DOS) method (Schowengerdt, 1997; Liang et al., 2001; Holben et al., 1992). The DOS requires dark surfaces, i.e., dark woods or deep waters. Since these types of surfaces are limited in a satellite image, the DOS method can only be applied to large areas with homogeneously distributed aerosols. This is usually not the case as visible in figure 3.11. Most other haze correction algorithms, e.g., the pseudo-invariant features (PIF) method (Du et al., 2002), have a similar drawback, since they are based on special surfaces which showinvariant spectral properties and are also limited in an image. Therefore these methods are usually only suitable for the radiometric correction for the effects of constantly distributed aerosols in satellite images (Richter, 1996). In this study, the radiometric correction is not of importance (see Section 3.2, p. 60). What is required is the detection and correction of relative differences within one

image resulting from heterogeneously distributed haze. Thus, a pixelwise haze detection algorithm is required.

Recently, Zhang et al. (2002b) developed a new haze detection algorithm, the HOT method which allows this type of quantification. The method has already been applied successfully in various contexts (Zhang et al., 2002a; Guindon and Zhang, 2002; Cihlar et al., 2003).

### The Basic Concept of the HOT Method

Zhang et al. (2002b) used a radiative transfer model (Moderate Resolution Transmittance (MODTRAN)) to simulate the measurement of the TM channels 1 and 3 for various surface types. The surface reflectances were obtained from independent measurements by an airborne spectrometer. These surface types are representative for landscapes in Canada and also typical for most areas of northern Germany. Figure 3.16 shows the result from these calculations; the results of the different surface types are displayed as capital letters. It can be seen that the various surface types are located on one line. Zhang et al. (2002b) named this line the clear line (CL).

The effects of haze have been modeled by including clouds of different optical depth $\tau$ into the model calculations. These results are also displayed in Figure 3.16 indicated by numbers adjacent to the letters. These numbers form branches spreading from the CL. It can be seen that the perpendicular displacement is constant for a certain optical thickness. According to Zhang et al. (2002b) the influence of haze can be quantified by this displacement that can be calculated by the following formula:

$$\mathrm{HOT} = q_{\mathrm{ch1}} \sin\Theta - q_{\mathrm{ch3}} \cos\Theta. \tag{3.31}$$

Here, $\Theta$ is the angle between the clear line and the abscissa, and $q_{\mathrm{ch1}}$ and $q_{\mathrm{ch2}}$ the original DN values for the TM channels 1 and 3, respectively (see Section 2.2.1, p. 31).

Unfortunately, the above equation is only valid if the CL is a line through the origin, which is not the case as obvious from figure 3.16. Thus equation (3.31) is not applicable in this context and the interception with the ordinate, $q_{\mathrm{ch3,0}}$, has to be taken into account. Therefore a corrected HOT value, $\mathrm{HOT}_c$, can be defined by transforming equation (3.31) to:

$$\mathrm{HOT}_c = q_{\mathrm{ch1}} \sin\Theta - (q_{\mathrm{ch3}} - q_{\mathrm{ch3,0}}) \cos\Theta \tag{3.32}$$

This modification can simply be interpreted as a shift of the CL to the origin of the diagram in figure 3.16. Since the different sensors have different calibration coefficients, the use of the DN values is not meaningful. Comparable results for the different sensors can be obtained by using the radiances instead of the DN of channel 1 and 3, thus the HOT value $\mathrm{HOT}_{cc}$ used in this study has the following form:

$$HOT_{cc} = I_{\mathrm{ch1}} \sin\Theta - (I_{\mathrm{ch3}} - I_{\mathrm{ch3,0}}) \cos\Theta. \tag{3.33}$$

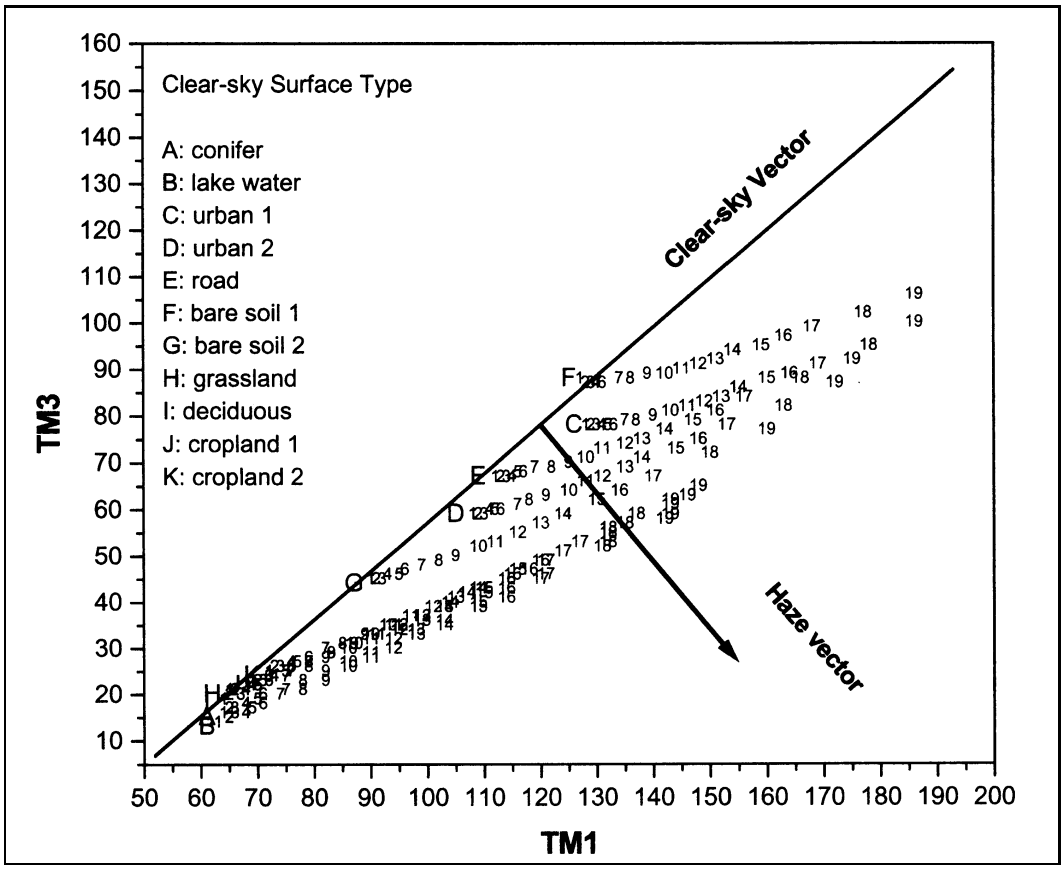



Figure 3.16: Schematic diagram of the simulation results performed by Zhang et al. (2002b) for various surface types and haze conditions. The axes show the simulated DN values $q_{\mathrm{ch1}}$ and $q_{\mathrm{ch3}}$ for the TM Channels 1 and 3 (see Section 2.2.1, p. 31). Under clear conditions, the various surface types (indicated by the capital letters) are located on the Clear Line named Clear-sky Vector in this diagram. The optical depth of the clouds has been varied between 0 and 6.7 in 18 equal steps (indicated by the numbers adjacent to the capital letters) at a wavelength of 0.55 $\mu$m (adapted from Zhang et al., 2002b).

### Application to Satellite Data

The theoretically determined relation between the CL and the radiance of the various surface types has to be verified with real satellite data, mainly since the model results are only based on a limited number of surface types. In addition, this is necessary since there are possibly other effects of the atmospheric constituents on the radiation such as the modelled aerosol scattering. Zhang et al. (2002b) performed this for various TM images in Canada. They selected haze-free regions in these images and calculated the correlation $\gamma$ between both channels. This correlation was at least 0.87 but generally above 0.9. This confirms that the radiances for channels 1 and 3 are linearly dependent for the majority of surface types and are thus located on the CL.

Another result from these calculations was that $\Theta$ is not constant for all images but varies from image to image. Therefore, this parameter has to be

adapted for every image. This done by selecting haze and cloud-free regions in the images and using these cloud-free pixels to obtain the CL and $\Theta$ with a linear regression.

As stated above, the $\mathrm{HOT}_{cc}$ value also depends on $I_{\mathrm{ch3,0}}$. This was not taken into account by Zhang et al. (2002b), yet results from calculations in this study (see below) show that this is mandatory.

Since the surface types in northern Germany might differ from the ones in Canada, the HOT method has to be validated for this study. Figure 3.17 shows three scatterplots for the radiance of channels 1 and 3 of all pixels from a mostly haze and cloud-free TM image. The upper diagram in Figure 3.17 shows a scatterplot for all pixels of the image 196/023; May 11, 2001. The majority of pixels show the expected linear dependency for the radiances. Nonetheless, there is a branch in this diagram that is not in agreement with the CL. Therefore, at least one surface type is not located on the CL. This surface type can be identified as flowering canola by anticipating the results of the canola classification in Section 4 (p. 87) which is also confirmed by the data shown in Figure 3.19. In this figure pixels of canola fields have negative HOT values where regions with negative HOT values can be identified as flowering canola fields. Further proof is given in the two lower diagrams of figure 3.17, in which the canola pixels are separated from the pixels with other surface types with the aid of the algorithm. The left diagram shows the canola pixels and the right one the non-canola pixels. From these diagrams it is clear that the branch in the upper diagram results from flowering canola.

The influence of flowering canola has to be taken into account for the selection of the training data set. Pixels of flowering canola have to be excluded from the determination of the CL, since the canola pixels influence the linear regression. This is be done by applying the canola classification on the training regions for the HOT method which is possible since the classification of canola is reliable in clear sky regions.

### Selection of Training Data Sets

Zhang et al. (2002b) simply used haze and cloud free regions in the image that were selected by visual inspection. Figure 3.18 shows two examples of rectangles that are used to calculate the parameters for the CL. Similarly, rectangular cloud-free regions are selected for all images available in this study.

Since flowering canola does not satisfy the CL condition, pixels from this surface type have to be excluded from the training data set. This is accomplished by the classification algorithm described in Section 4 (p. 87), which can be applied to clear sky regions without problems.

The parameters $\Theta$, $I_{\mathrm{ch3,0}}$ and the correlation coefficient $\gamma$ are calculated for all images with a linear regression between the radiance of channel 1 and 3 for all non-canola pixels in the cloud-free rectangles. Table 3.4 displays the range for the determined parameters. The correlation is always higher than 0.9, which is higher than the correlations calculated by Zhang et al. (2002b).

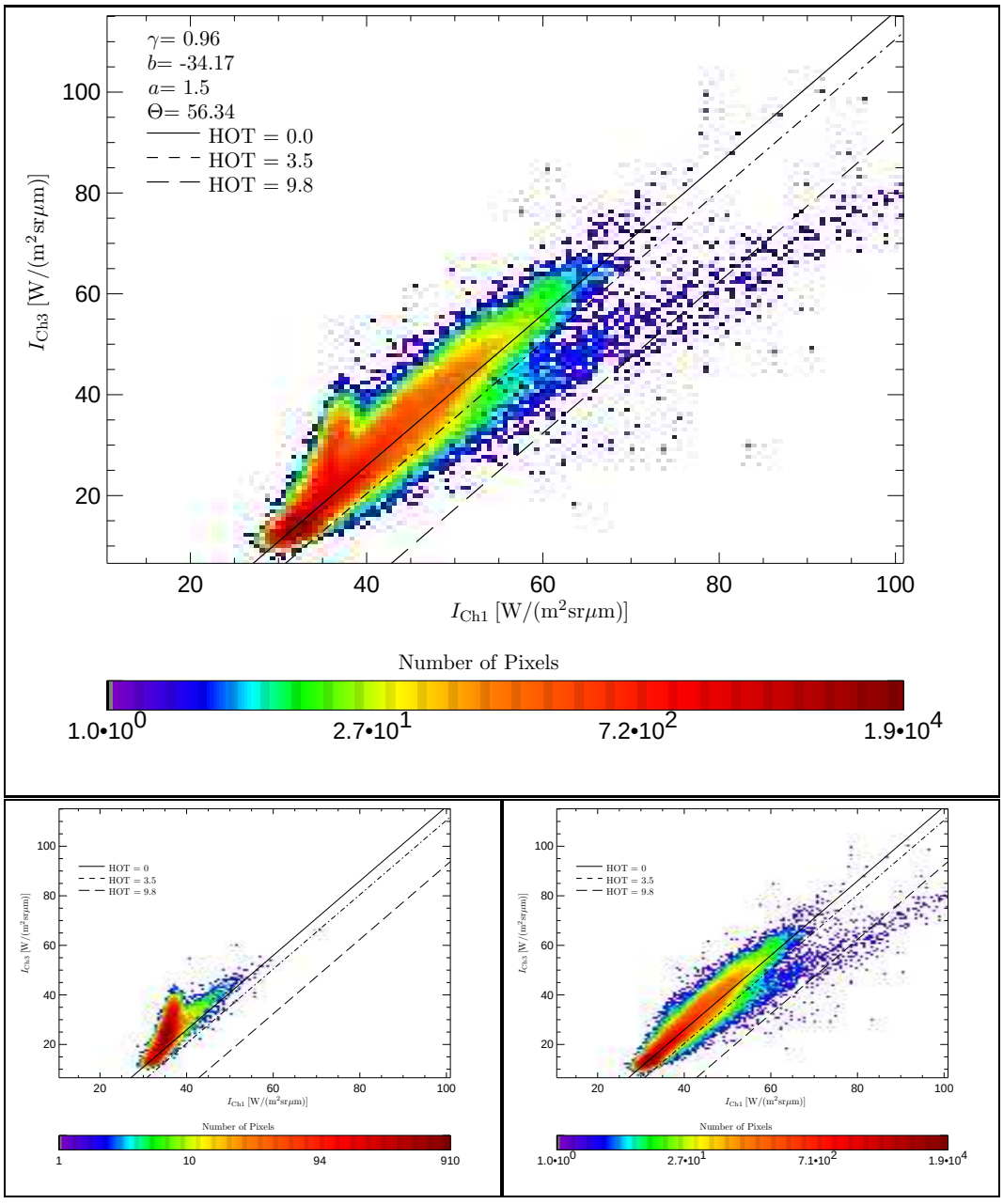



Figure 3.17: Scatterplots for the mostly cloud free image 196/023; May 11, 2001. Note that the number of pixels is displayed on a logarithmic scale. Top: Scatterplot of the radiances $I_{\mathrm{ch1}}$ and $I_{\mathrm{ch3}}$ for all pixels; observe the small branch pointing upward of the scatter cloud. Bottom left: Scatterplot for the radiances of all pixels that have been identified as canola (see Section 4, p. 87). Bottom Right: Scatterplot for the radiances of all remaining pixels. The lines display the pixels of equal HOT values; HOT= 0.0 for the CL, HOT=3.5 for the haze-covered HOT threshold and HOT=9.8 for the cloud-covered threshold. Note that the CL (HOT=0) displayed in all plots is calculated excluding the canola pixels.

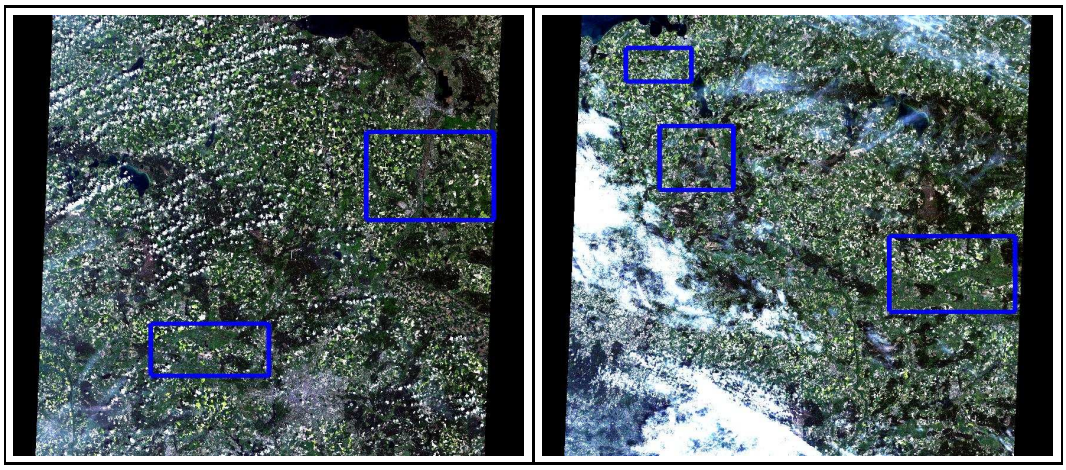


Figure 3.18: Example for the training data set selection for the HOT method. The blue rectangles represent the areas chosen to determine the CL. Left: image 193/023; May 13, 2001. Right: image 194/023; May 12, 1999. Original data: LANDSAT TM ©ESA, 1999 and LANDSAT ETM+ ©ESA, 2001. Distributed by Eurimage.

Important for the comparability of the haze detection in different images is the error of the HOT value caused by neglecting $I_{\mathrm{ch3,0}}$. If $I_{\mathrm{ch3,0}}$ is constant for all images, the results of the original method by Zhang et al. (2002b) can be compared since the error for the HOT value is constant for all images. As obvious from Table 3.4, the value of $I_{\mathrm{ch3,0}}$ shows a large range, which can result in an error of about 25 and as the threshold chosen for the identification of haze and cloud-cover is merely 3.2 and 9.8 (see below), this parameter has to be taken into account.

Table 3.4: The minimum and maximum value for angle $\Theta$ and interception, $I_{\mathrm{ch3,0}}$, of the CL with the ordinate and for the HOT method. Also displayed are the values for the correlation coefficient $\gamma$. Note the large range for $I_{\mathrm{ch3,0}}$.

| | Min | Max |
|---|---|---|
| $\Theta$ | 54.2° | 60.3° |
| $I_{\mathrm{ch3,0}}$ | −55.1 | −29.3 |
| $\gamma$ | 0.91 | 0.98 |

The remaining parameter, the angle $\Theta$ shows a variation that is comparable to the results from Zhang et al. (2002b). According to them this variation is the result of scattering from aerosols, which are distributed homogeneously in the image. Also likely is the influence of the gaseous absorption of water vapour and ozone. In particular ozone may have an effect on the angle $\Theta$ since it mainly absorbs radiation measured by Channel 3 of TM but this requires further investigation.

### Result of the HOT Method

With these adjustments of the training data set and equation (3.33), the HOT value can be calculated for the available data. Figure 3.19 shows an example for the application of the HOT method. Displayed are images of the HOT value (top), the true colour composite (bottom left) and the false colour composite (bottom right) of a hazy region for the clip near Braunschweig.

The HOT value in the upper image allows to identify cloud and hazy regions and to quantify the modification of the signal by aerosol scattering. For instance, the cloud-covered region on the left of the clip can be identified by HOT values above 9.8. The thinner fringe regions of this cloud and the thinner clouds at the bottom of the clip shows, that the HOT values can be identified if they are in the range of 3.5 to 9.8. Note that the HOT method is capable of detecting haze-cover that is barely visible in the true colour composite, e.g., the vertical thin cloud at the bottom of the clip, whose extension to the centre of the clip is only visible in the HOT value image. Also remarkable is the vertical contrail on the left of the clip, which can also be identified with this method. A similar result can be seen in the diagrams displayed in Figure 3.20. The upper diagram shows the HOT values in a scatter plot of the radiances $I_{\mathrm{ch1}}$ and $I_{\mathrm{ch3}}$ of the same region. The effect of haze and cloud-cover is clearly visible in this diagram by the increasing spread of the scatter cloud with increasing brightness of the radiances for both channels. This shape of the scatter cloud is expected for cloud-covered regions from the model calculations by Zhang et al. (2002b) (see Figure 3.16) and is confirmed by a comparison with the lower diagram in that figure that shows a scatter plot from a cloud-free region which demonstrates the linear dependency that is predicted for clear sky regions.

An additional example is displayed in Figure 3.21 that shows the HOT value and the true colour representation for a region near the Baltic Sea. In this clip, haze and cloud-cover show the same range of HOT values as in Figure 3.19. For instance, the small thin cloud marked with the yellow circle in the right image can be identified as haze by its HOT values. This shows that the haze is reliably estimated by the HOT method and this result can be confirmed for all images used in this study.

Besides the recognition of clouds, the effect of the different surface types on the HOT value must by small compared to the effect of aerosols. As visible in the two lower clips in Figure 3.19, the displayed region consists of typical land surface found in northern Germany like, e.g., various types of field crops, woods, streets and urban regions. In this image, it can be seen that most underlying surfaces have a HOT value of about 0. Exceptions are:

- Bare soil with a HOT value below 0 which appears turquoise in the false colour image on the bottom right.
- Settlements or industrial areas with a high HOT value, visible at the top right of the clip as grey or white regions in the true colour image.
- Broad roads like the highway that leads from the left to the right side

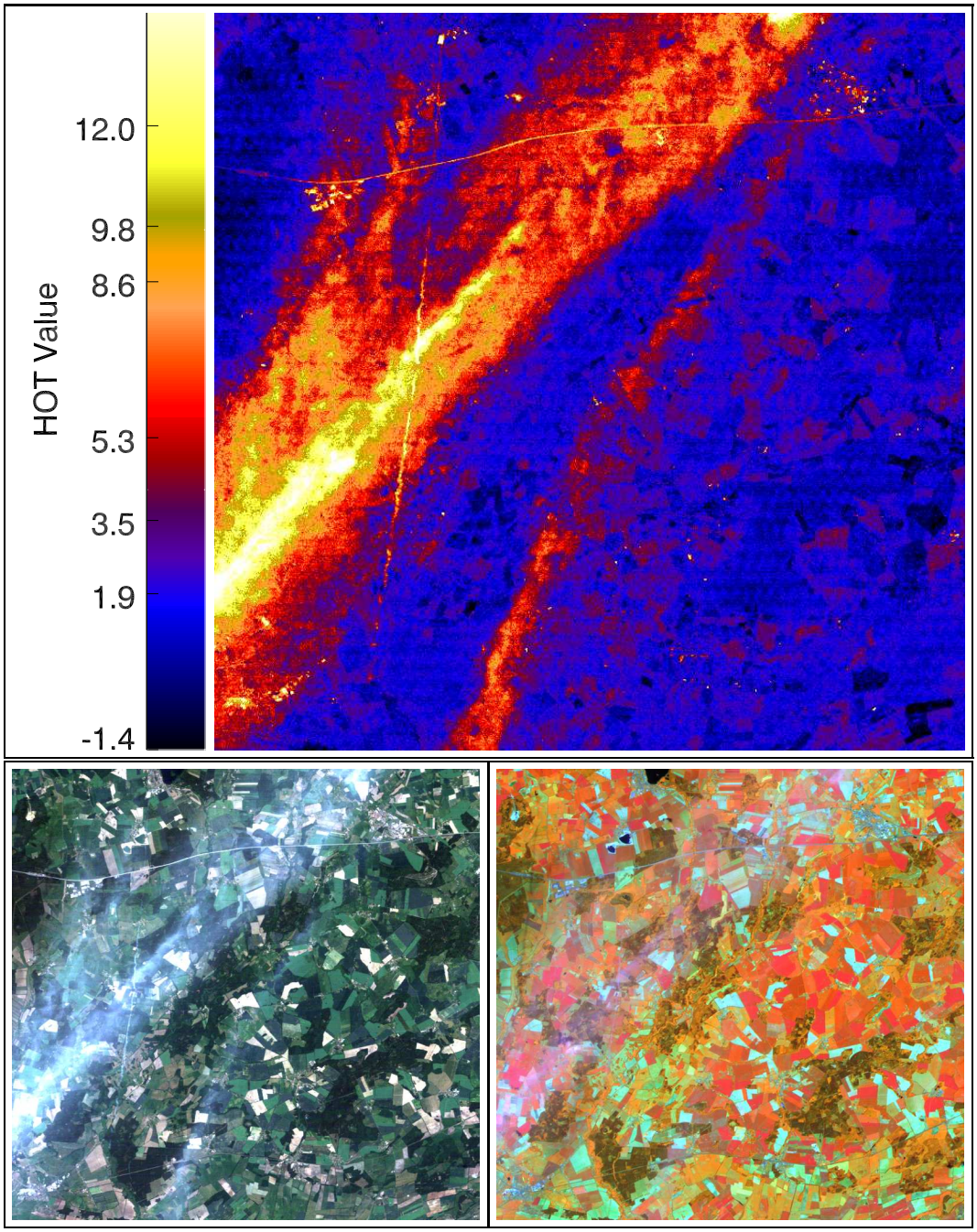
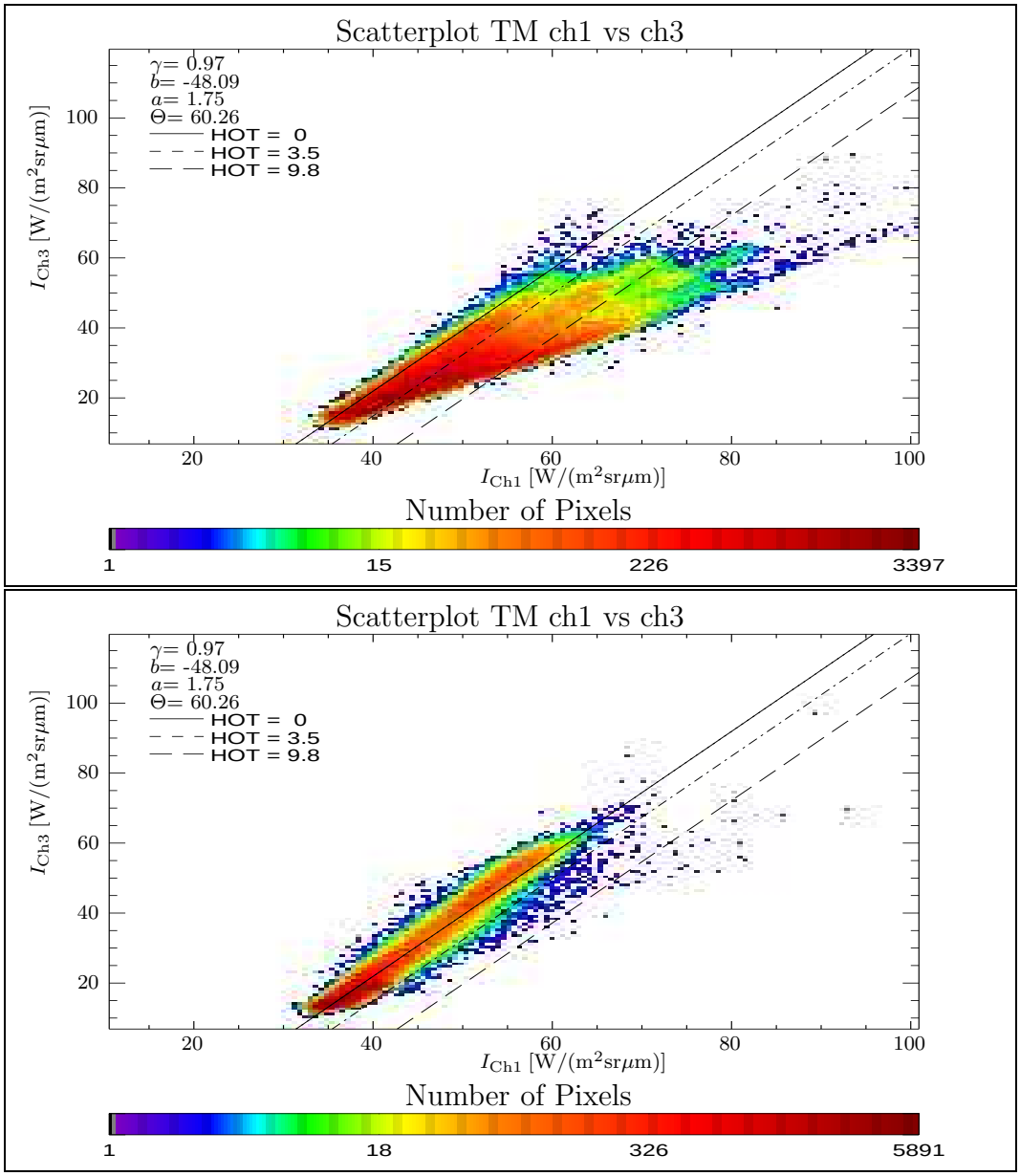



Figure 3.19: Top: Result of the haze detection, i.e., the HOT value calculated with equation (3.33). Bottom: The original data in true colour (left) and false colour (right) representation with channels 4, 5 and 3 (right). Displayed is a clip near Braunschweig from the TM image 193/023; June 5, 1998. The upper image demonstrates that the cloud and hazy regions are clearly separable from each other and from the clear regions by the HOT value. The colour bar for the HOT value is chosen such that haze (3.5<HOT<9.8) will appear in orange and clouds (HOT>9.8) in yellow. Moreover it shows that the majority of surfaces have a HOT value of about 0 and thus do not influence the detection of cloud and haze. Original data: LANDSAT TM ©ESA, 1998. Distributed by Eurimage.



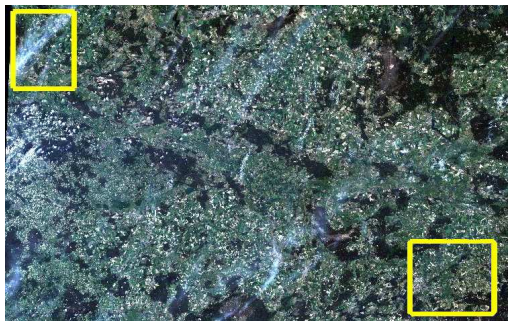



Figure 3.20: Top: HOT scatterplots of a hazy and cloud-covered region (upper left rectangle in the small image) which shows data from Figure 3.19. Bottom: HOT scatterplot of a clear region (lower right rectangle in the small image). The influence of haze and clouds is clearly visible in the upper diagram. The lower diagram shows the expected linear dependency (CL) of the radiances. The lines for the HOT threshold in both plots demonstrate the selection of clear, hazy and cloud-covered pixels. Original data: LANDSAT TM ©ESA, 1998. Distributed by Eurimage.

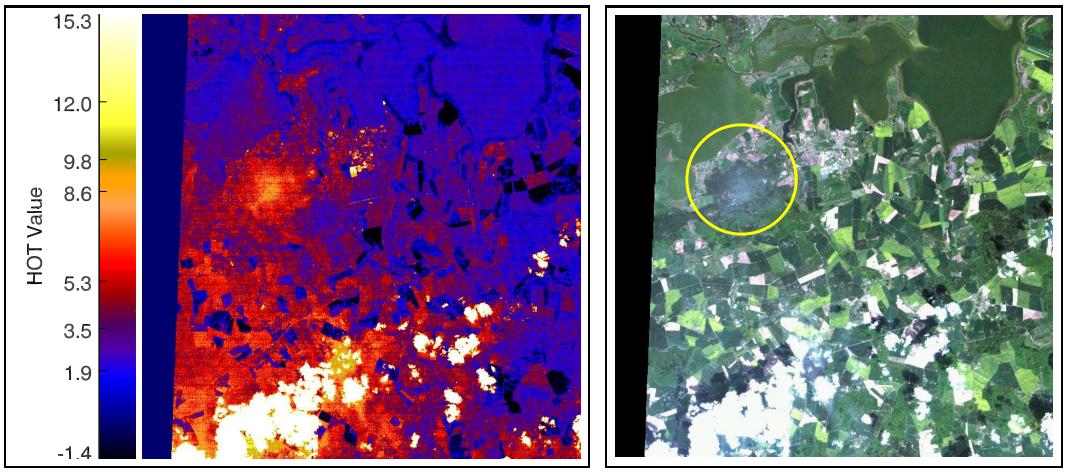



Figure 3.21: Result of the haze detection for image 193/022; May 13, 2001 for an image clip west of Stralsund at the Baltic Sea. Left: The HOT value; Right: a true colour composite of the same clip. Note that the small hazed regionmarked by the yellow circle in the true colour image can be identified clearly as haze. Also of interest are the areas with negative HOT value, which can be identified as flowering canola fields by the yellowish colour in the true colour image. Original data: LANDSAT TM ©ESA, 2001. Distributed by Eurimage.

of the clip in Figure 3.19 and have the same colour as settlements in the true colour image.

- Sea surfaces with a high content of sediments (not displayed in Figure 3.19).

As discussed above, flowering canola is also an exceptional surface type with a negative HOT value. This effect is not visible in Figure 3.19 since the TM image in Figure 3.19 was acquired in June two weeks after the flowering period of canola in that year. This effect is demonstrated by the TM image displayed in Figure 3.21 which was acquired during the maximum of the flowering period in 2001. This is also visible by the yellowish fields in the true colour composite. For this image, the HOT value is clearly negative for the canola fields visible in this clip. Note that the strength of the flowering is not constant but changing from field to field. Moreover, the lower HOT value of fields in hazed regions indicate that the effect of canola flowering and aerosols on the HOT value cancel each other.

These results demonstrate that the HOT method allows the quantification of haze and thin clouds in TM satellite images. Thus it can be used to adjust the classification algorithm in hazy regions and to improve the cloud and cloud-shadow detection. Nonetheless, these two tasks require further adjustments of the haze algorithm for automatic processing.

### Application to the Detection of Canola

As demonstrated above, the quantification of haze is possible for the majority of surfaces in the images. Unfortunately, just the surface type that is the objective of this project, flowering canola, is an exception. The yellow canola blossoms have the opposite effect as haze and the HOT value decreases with an increasing number of blossoms. This results in a negative HOT value for clear-sky regions and a reduced HOT value for hazy regions. Since the density of flowering is usually not known, it is not possible to retrieve correct HOT values for this surface type directly. Thus the haze correction for the classification algorithm requires to take surrounding pixels into account that satisfy the CL requirement. Practically, this is only possible with a recursive solution applying the classification algorithm. The description of this method requires knowledge on the classification algorithm and thus will be described later in Section 4 (p. 87). Another effect of haze might be the misinterpretation of non-canola surfaces. Since it is possible to detect haze over the majority of surfaces, this can be prevented. The exceptions are bare soil, settlements, roads and some sea surfaces. These surfaces can be treated in the same manner as flowering canola by using the HOT values of adjacent pixel to interpolate the HOT value for these surfaces. Practically, this is not necessary since the spectral properties of these surface types are not altered by haze in the manner to be misinterpreted as canola (see Section 4, p. 87). Note that this might be necessary for other types of agricultural plants.

### Improvement for the Cloud Detection

Besides the quantification of canola, the haze detection can also improve the cloud detection algorithm described in Section 3.2.2. Clouds can also be identified by their high HOT value; visual inspection of several images showed, that a HOT value of 9.8 is a good threshold for cloud cover. This new cloud-cover mask give better results for small clouds and for the fringes of clouds. Nonetheless, this new cloud-cover mask can not detect thick clouds since the radiance reflected by these clouds exceeds the intensity range for channel 1 of the TM sensor. This saturation results in an HOT value below the cloud threshhold.

The best result for the cloud detection is achieved by combining both masks to a new mask (see Section 3.2.1, p. 62), which allows a better identification of cloud edges and smaller clouds but also includes thick clouds. Figure 3.22 shows a comparison of the old and new masks for the cloud detection. Five different clouded regions have been marked by yellow circles in the upper image of that figure; all clouds have been identified correctly as visible in the lower image. These results show the high quality of the cloud detection.

Nonetheless, the lower image in Figure 3.22 also shows that some areas are wrongly declared as clouds. For instance, the region left of the clouded region marked as cloud covered is actually a village. Thus, there are surface types left with HOT values similar to those of clouds. These surface types

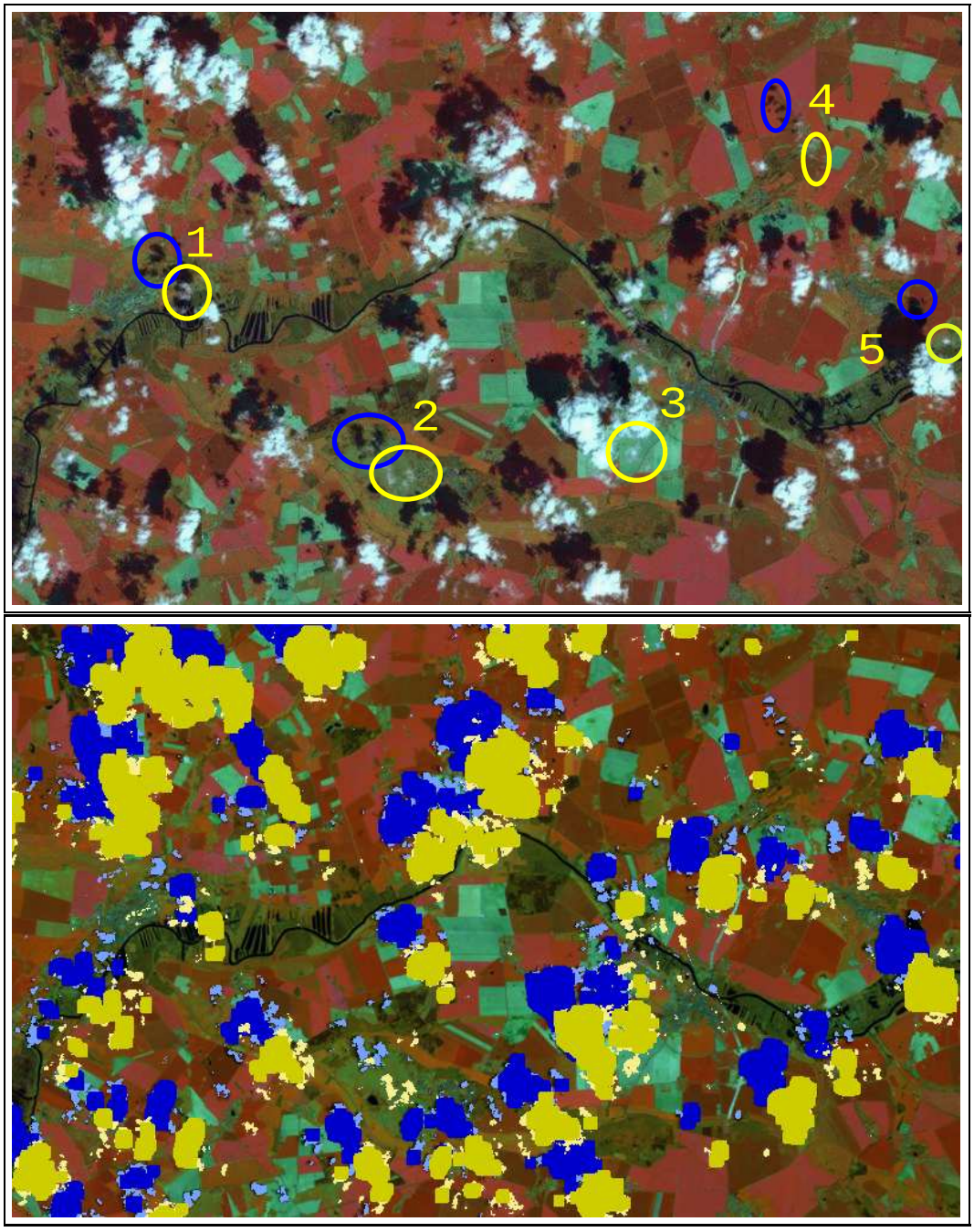



Figure 3.22: Final result of the cloud and cloud shadow detection. Displayed is a clip of the image 196/022; May 2, 2001, south of Stralsund. Top: Original image in false colour representation of the channels 3,4 and 5. The circles indicate clouds (yellow) or cloud-shadows (blue) which could not be detected by the threshold cloud detection. Bottom: Comparison of the cloud and cloud-shadow detection with the threshold (dark yellow/blue) and HOT method (bright yellow/dark blue). It can be seen that all borders of clouds and smaller clouds are much more accurately identified by the HOT method. Original data: LANDSAT ETM+ ©ESA, 2001. Distributed by Eurimage.

are mainly roads, settlements and beaches and remain mostly unchanged for a longer period of time, usually many years. This can be exploited to distinguish these surfaces from clouds by comparing the results of the HOT method from a cloud free image with the one of the clouded image. The HOT values for a cloud-free image of the same frame are calculated and every pixel that has a HOT value above 9 and its neighbouring pixels are removed from the cloud mask of the original image. Using the image-to-image registration described in Section 3.1.4 (p. 55), the corresponding pixels in the clouded image can be accurately identified and masked out.

Another surface type that is misinterpreted as cloud is the sea surface with high sediment concentration. This surface type is not constant in time but fortunately the sea surface itself is constant for a longer period and can be masked out easily.

Such an improved cloud mask is also used to generate a more accurate cloud-shadow mask. As shown in Figure 3.22 all four marked cloud shadows that have not been detected by the threshold algorithm can be identified with the new method. Also obvious from this figure is the necessity to remove surface types with high HOT value from the cloud-cover mask since misclassified clouds are naturally also misclassified as cloud-shadows.

### 3.2.5 Radiance Correction with the HOT Value

The influence of haze on the radiance can be corrected to some extent. The largest influence of aerosol scattering on the radiation received by the sensor results from the sun irradiance scattered by these aerosols to the sensor. This increases the measured radiation independently of the radiation reflected by the surface and thus is a constant radiation offset for equal haze conditions (Chavez, 1996; Moran et al., 1992).

The DOS method (Schowengerdt, 1997) uses dark objects in the image to estimate the amount of this radiation by observing the light reflected by dark surfaces. Under the assumption that those dark objects reflect no or only a small amount of radiation (i.e., have a reflectance close to zero), the radiation originating from the dark objects must be the radiation scattered in the atmosphere, i.e., mostly by aerosols in haze regions.

Under this assumption, the radiance at the sensor for objects with higher reflectance is then corrected by subtracting the radiance measured over the dark object. Originally, the dark objects are selected manually but an automated selection is possible by selecting dark object with the aid of a radiance histograms selecting those 1 % of all pixels that are darkest (Schowengerdt, 1997).

The drawback of the DOS method is the requirement for either a homogeneous haze cover or a large number of dark objects, which are seldom present or available, respectively. The improvement possible by using the HOT method is the correction of haze independently of nearby dark objects in the image. Similar to Zhang et al. (2002b), this is achieved by calculating histograms

Table 3.5: Mean $\bar{q}_i$ and dark $q_{i,1\%}$ sensor measurement of the TM channels for different HOT-Values.

| Channel $i$ | Clear sky | | 3 to 5 | | 5 to 7 | | 7 to 9 | | 9 to 11 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $q_{i,1\%}$ [DN] | $\bar{q}_i$ [DN] | $q_{i,1\%}$ [DN] | $\bar{q}_i$ [DN] | $q_{i,1\%}$ [DN] | $\bar{q}_i$ [DN] | $q_{i,1\%}$ [DN] | $\bar{q}_i$ [DN] | $q_{i,1\%}$ [DN] | $\bar{q}_i$ [DN] |
| 1 | 80 | 94 | 92 | 114 | 101 | 129 | 111 | 146 | 118 | 161 |
| 2 | 31 | 42 | 35 | 53 | 39 | 62 | 43 | 73 | 46 | 83 |
| 3 | 27 | 43 | 31 | 56 | 36 | 68 | 41 | 82 | 44 | 94 |
| 4 | 33 | 88 | 41 | 99 | 46 | 107 | 50 | 114 | 54 | 120 |
| 5 | 26 | 69 | 31 | 79 | 36 | 88 | 37 | 98 | 38 | 108 |
| 7 | 10 | 30 | 12 | 36 | 14 | 41 | 17 | 48 | 16 | 54 |

in dependence of the pixel's HOT value, i.e., the radiance of pixels within a specific HOT value range are used to produce a histogram. As an example, Figure 3.23 shows the radiance[7] distribution for each channel for different HOT-value ranges. With an increasing HOT-value the distributions are shifted to higher radiances and as expected, the shift decreases with increasing wavelength (see Section 1.4.2, p. 11).

Table 3.5 lists the mean values for the different HOT value ranges and confirms the observed shift of the histograms by the increasing mean values for the distributions. Additionally listed are the mean radiances $I_{i,1\%}$ for the 1 % darkest pixels , which likewise show a shift to higher radiances.

The $I_{i,1\%}$ are displayed in Figure 3.24 and it can be seen that the HOT value and this radiance are linearly related. This is approximated by the linear regression of $I_{i,1\%}(\mathrm{HOT})$ and allows to define a HOT-dependent dark radiation $I_\mathrm{d}(\mathrm{HOT})$ with:

$$I_\mathrm{d}(\mathrm{HOT}) = a \cdot \mathrm{HOT} + b \tag{3.34}$$

The calculated regression lines are also displayed in Figure 3.24. With this linear relationship, each pixel can be corrected according to the HOT value determined by Equation 3.33. The $I_d(\mathrm{HOT})$ is calculated for each image to allow a correction of the influence of haze, unless the haze cover is below 10 % for the complete image. Unfortunately, the HOT values of canola are influenced by the flowering of canola, which has to be corrected before applying this method successfully. This will be discussed later in Section 4.2.3 (p. 113).

[7]Note that the values displayed in the Table 3.5 and in the Figures 3.23 and 3.24 are the uncalibrated sensor measurements $q_i$. The reason for this is the easier presentation of all channels in single diagrams. The actually applied algorithm uses the calibrated radiances $I_i$.

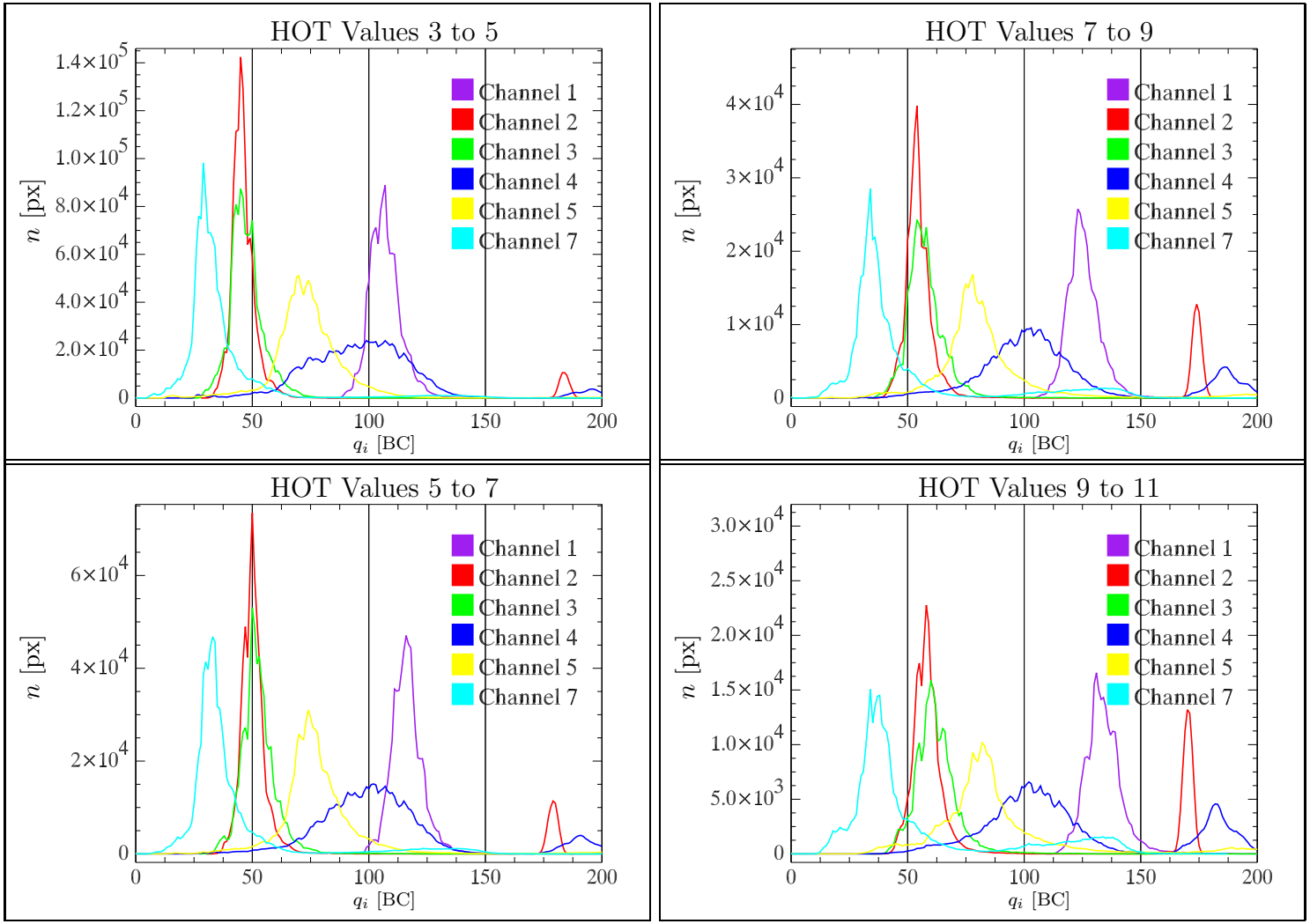



Figure 3.23: Histogram of TM data for different HOT values. Note that all channels are shifted to the right with increasing HOT value. This effect is stronger for the shorter wavelengths. The histograms have been exemplarily calculated from the TM image 193/024; May 02, 2000.

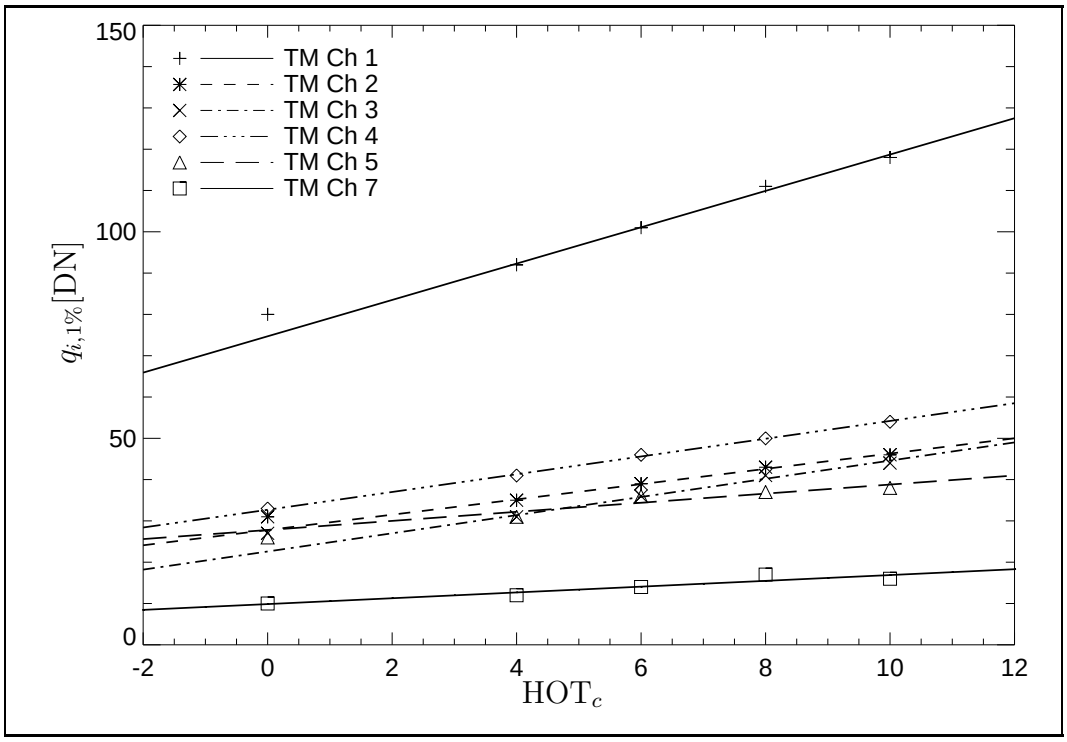



Figure 3.24: Uncalibrated sensor measurement $q_{i,1\%}$ of the 1% darkest pixels in dependence of the HOT-Value the TM Channels 1,2,3,4,5 and 7. Displayed are the regression lines for the different channels. Note that the $q_{i,1\%}(\mathrm{HOT} = 0)$ has been ignored for the calculation of the regression lines.

### 3.2.6 Conclusion

The method described above allows a reliable identification of clouds and cloud shadows which is essential in order to identify the regions in the image which allow a classification and those which do not. This information is mandatory for the generation of cultivation statistics. The method can be applied autonomously. The only exception is the selection of the cloud-free training areas for the HOT method which has to be done by a human operator.

Moreover, the haze quantification with the HOT method allows to extend the classification to regions with inhomogeneous haze cover by correcting the influence of haze with a method based on a histogram and the HOT values. Thus, the satellite data can be used more effectively.

# Chapter 4

# Classification

Classification is the assignment of a label to regions in a satellite image consisting of similar surface types. The classification method can either be based on the reflectances of a single pixel or those of a segment. A segment is a collection of neighbouring pixels. The assignment is usually based on the comparison of the pixel's reflectances $\sigma_i$ for each channel $i = 1, 2, 3, 4, 5, 7$ with a known distribution of reflectances of several surface type classes[1]. In this study, we will investigate the reflectances of canola fields and that of other surface types classes present in a satellite image in order to decide whether and how to distinguish the different surface type classes.

An obvious example of the different reflectances from surface type classes is the yellowish colour of flowering canola. In principle, this could be used to identify canola fields with a simple threshhold algorithm. Unfortunately, canola does not show a constant flowering over the complete image and even within fields there can be variations not to mentrion satellite images from acquired earlier or later than the flowering of canola.

Therefore, a more sophisticated method is necessry which also takes possible vaiations in canola fields into account (Colwell, 1983, Chapter 18). Moreover, it is important to get a comparison to the variations of other surface type classes.

Agricultural plant are of special interest in this context and their reflectances will be investigated with the Quillow mapping data set (see Section 2.2.2, p. 34). This data set provides information on the crops planted in an area of 20,000 ha in the federal state of Brandenburg, in particular the field edges and the crop type grown. This information will allow to determine the radiances reflected by fields of the most common agricultural plants in Northern Germany for three TM images from the years 1999, 2000 and 2001, taken at different growth stages of canola. Besides a first investigation of the separability of agricultural plants, this discussion will show that not all TM channels are necessary to separate canola fields from fields of other agricultural plants in a TM or LISS/3 satellite image.

[1] In satellite remote sensing, one generally speaks of surface types classes. These describe a collection of similar surface types or surface type mixtures

The quality of crop type identification depends on the classification algorithm used. The commonly used MLC is applied to the data. The results are evaluated with the Quillow mapping data set and provide a first quantitative estimation of the separability of classes (Richards, 1986, Chapter 10) and the expected classification accuracy.

Although the MLC is common used in surface type classification, another classifier, the MDC, is more suitable in this study since it requires only training data for one class, i.e., the canola class, whereas the MLC requires training data for all classes. But since the MDC is not as accurate as the MLC, the results have to be improved after a first classification by combining neighbouring classified pixels to segments and using the statistics of the segment's radiances to perform region growing and to decide if the segment really is canola. The segmentation also allows to vectorise the data, in order to reduce the amount of data and to obtain additional information on the canola fields, e.g., the orientation with respect to the main wind direction.

As discussed before the classifier has to be adapted in haze covered regions of the images. This is achieved by a histogram shift based on the HOT method (see Section 3.2.4, p. 71). The variation of reflectivity during the canola flowering is compensated by an adaption of the classification algorithm. This is necessary to achieve a homogeneous quality of accuracy over the complete satellite image.

Finally, the classification has to be applied to all available images which requires an automated selection of training data from neighbouring fields and the selection of a priori training data sets for years without ground truth data are available. The results of the complete classification are then compiled to global data which are used for further studies on the acreage of canola. A selection of these will be presented in Chapter 5.

## 4.1 Spectral Properties of Surfaces in Northern Germany

The reflectance represented in a single pixel is usually a mixture of reflectances of various materials. Thus, a surface type class in a satellite image does not represent one specific material but rather a typical mixture of materials. Such surface type classes generally depend on the sensor's resolution: Red roofs can be a surface type in an aerial photograph but not in lower resolution satellite data. An adequate label for a corresponding TM data surface type class might be settlement or urban.

In general, the reflectances of surface type classes are determined from the image. Moreover, the reflectances are seldom constant in time, predominantly because plant cover is a part of most of these surface types classes. The reflectivity of plant cover depends on the growth and health state of the plants. Nonetheless, the reflectances can be assumed to be constant for one TM image, since the growth stages of plants can be assumed to be comparable within one

image. In order to distinguish different surface type classes by their spectral reflectances, the reflectance of the various surface type classes in Northern Germany have to be discussed.

### 4.1.1 Surface Types in Northern Germany

Potentially, numerous surface types classes are distinguishable with TM or LISS/III and it is difficult to identify them all. Nonetheless, most surface types have reflectance properties quite distinct from canola since they contain less vegetation, such as urban areas, or are much darker, such as woods and waters. Moreover, some surface type classes can also be distinguished by their shape (roads) and extent (villages, fields and woods). Taking this into account, the most likely surfaces types to be mistaken for canola are fields of other agricultural crops which will therefore be discussed here in detail.

### 4.1.2 Agricultural Plant Covers

The reflectance of plant cover is not constant but rather depends on growth stage, fertilisation and past weather conditions but these parameters are seldom available. If these parameters were present, resulting reflectance can be modelled (Asrar, 1989; Sarkar et al., 2002; Rudorff and Batista, 1990; Toll et al., 1997; Dawson et al., 1998). This is not the case in this study.

Another method to determine the reflectances of canola is the measurement with a ground based radiometer (Martonchik, 1994; Richardson et al., 1992; Gates et al., 1965). These measurements have to be made contemporaneous with the satellite data because of the dependence on the growth stage. Additionally, both methods require a correction for the atmospheric influence in order to allow a comparison with the satellite data (Colwell, 1983, Chapter 18). This requires additional information on the atmospheric conditions, which is difficult to obtain.

A much more efficient method is to collect the reflectivity for the different surface type classes from the satellite image itself. And, as discussed in Section 3.2 (p. 60), the radiance at the sensor can be used equivalently to the reflectances for comparisons within the same image.

Therefore, the radiances or reflectances of crops that are likely to be mistaken for canola have to be investigated and it is necessary to identify pixels representative for these crops. The radiances of the pixel of agricultural crops will be used to investigate their distribution in channel space (Schowengerdt, 1997, Chapter 9) and to acquire a training data set for the evaluation of the classification algorithms used in this study.

#### Selection of Training Data

The Quillow mapping data set (see Section 2.2.2, p. 34) allows to identify fields of agricultural plants for the years 1999, 2000 and 2001. Although it


Figure 4.1: Selection of representative pixels for the different agricultural plants by identifying homogeneous regions within the different fields mapped in the Quillow mapping data set; displayed is a clip of the image 193/023; May 15, 2001. Left: Fields emphasised by the coloured edges that where taken from the Quillow Mapping data set; the colours correspond to the legend in the right clip. Note that the fields are frequently containing more than one surface type. Right: Manuallay selected training pixels for each agricultural crop displayed as filled polygons, the corresponding crop is visible in the legend on the right of this clip. Inhomogeneous regions and obviously falsely mapped fields have been omitted in this selection. Note that the training data set has been generated for the complete Quillow data set and not only for this clip. Original data: LANDSAT ETM+ ©ESA, 2001. Distributed by Eurimage.

covers about 20,000 ha which is only about 0.2 % of the complete investigation area, it allows to investigate the reflectance properties of the most common agricultural plants in Northern Germany.

As discussed in Section 3.1.4 (p. 57) the georectification has been improved to one pixel deviation. Nonetheless, the mapped field boundaries available cannot be used directly to select representative pixels by simply using all pixels within them. The two reasons for that can be seen in the left clip of Figure 4.1. This clip shows a false colour TM image overlaid with the edges from the Quillow mapping data set. The different colours of the edges indicate the crop grown on the field enclosed by a field border.

First, this clip shows that there are variations within the fields. These result predominantly from small forest patches or from variations in the field partitioning that may have changed since the ground truth field edges have been taken. This is obvious from the similarity of the colour of these variations with image areas of forest (lower left corner) and other crops (lower right corner) in the left clip of Figure 4.1.

Differences in seeding density, water supply or fertilisation might also ex-

plain such variations, but they usually have a smaller influence on the radiance than the mapping inaccuracies mentioned above and would not appear in such an obviously different colour.

Secondly, the labelling of the fields is incorrect for some fields. These errors can be identified by comparing the colours of fields for different crops to those of other crops. These errors mainly results from the accuracy of the mapping data and will be discussed in detail in Appendix B since they have an influence on the accuracy assessment for the classification. Obviously, these fields must not be used for a training data set since they represent different crops.

A smaller error is caused by the inaccuracy of the border locations, either of the mapping or of the georectification of the satellite data (see Section 3.1.4, p. 57). Additionally, pixels at the edge of the fields might also include a fraction of other plants (mixed pixels) and are not usable as training data set. For these reasons, representative pixels have to be identified manually.

The right clip in Figure 4.1 shows an example of these selections. These pixels have been obtained by selecting homogeneous areas in fields of the different crops common in Northern Germany. Additionally, fields that obviously have a wrong label are also omitted from the selection of such representative pixels.

The training data include pixels for ten common crops. These representative pixels have been selected from the majority of fields available from the Quillow data set for the years in which the mapping data is available. The number of pixels depends on the acreages of the crop in question and ranges from more than 8000 for wheat to 700 for peas for each year, which is sufficient for a representation of the reflectance properties of these crops.

### Distribution of Radiances in Channel Space

The resulting distribution of the radiances of these pixels are scatter clouds in a four (for LISS/III data) or six dimensional (for TM data) channel space. An identification of the different plants with the satellite data is only possible if the scatter clouds for the different plants do not or only slightly overlap. The quality of the separability depends on the size of this overlap. Therefore, Figure 4.2 shows the distribution of all plants for the TM channel combination 3 (red) and 4 (NIR). Two further diagrams, one for the optical TM channels 1 (blue) and 2 (green) and one for the MIR channels 5 and 7 are displayed in Figure 4.3. This representation is equivalent to the projection of the remaining channels onto the plane of the displayed channels. Note that in spite of scatter clouds overlapping in two channels, they might by separable with the aid of a third or fourth channel.

There are two principal clouds in all three diagrams. The rightmost cloud in Figure 4.2 and the upper one in the diagrams in Figure 4.3, consist of radiances for peas, sugar beet and maize pixels. The lower cloud is formed by the cereals[2], clover and canola. These two principal clouds can be explained

[2]Although maize is also a cereal, the notion of cereals in this context is used for barley,

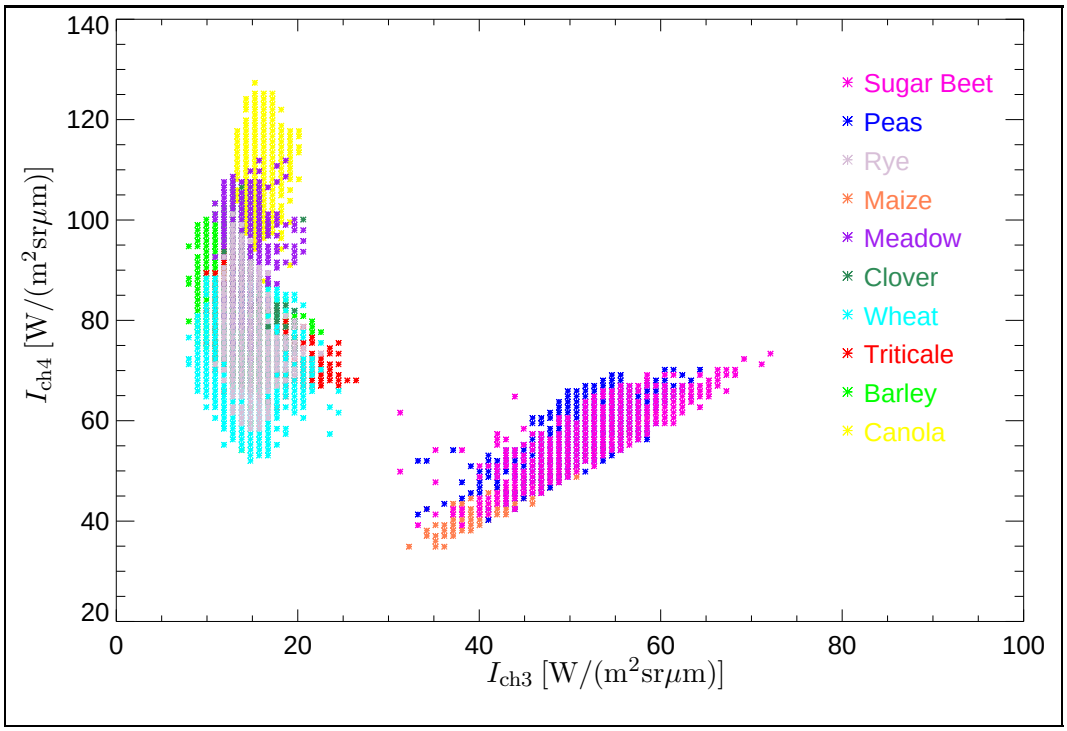



Figure 4.2: Scatter diagram of the radiances for the selected representative TM pixels (see Figure 4.1) plotted for the channels 3 and 4 for 10 different agricultural plants. Note that the radiance for canola has been shifted by the size of one size to the upper right in order two allow to display all canola pixels. Pixels for all other plants are simply overplotted in order of their appearance in the legend on the right. This plot shows that canola can be distinguished from cereals (barley, triticale, rye and wheat) by using channel 4. Both channels allow the distinction of canola from maize, sugar beet and peas.

by the lack of a dense vegetation cover for sugar beet, peas and maize and thus are or appear as bare soil in the satellite image that have been acquired in spring and early summer. Since they are harvested later than the other crops these plants will be called late crops hereafter.

Fields of these plants are easily distinguishable from canola and other early growing plants that already show a dense vegetation cover in spring and early summer. All early growing crops are located in the same principal cloud and are more difficult to distinguish. Figure 4.2 demonstrates that channel 4 allows to separate canola from the majority of these crops since canola is brighter than the other crops in this channel. The crops located closest to canola in this figure are barley, meadow and rye. Obviously, these plants are most likely mistaken for canola.

Two further diagrams of different channel combinations are shown in Figure 4.3. Both diagrams show scatter clouds similar to the one in Figure 4.2.

rye, wheat and triticale since they are diffcult to distinguish.

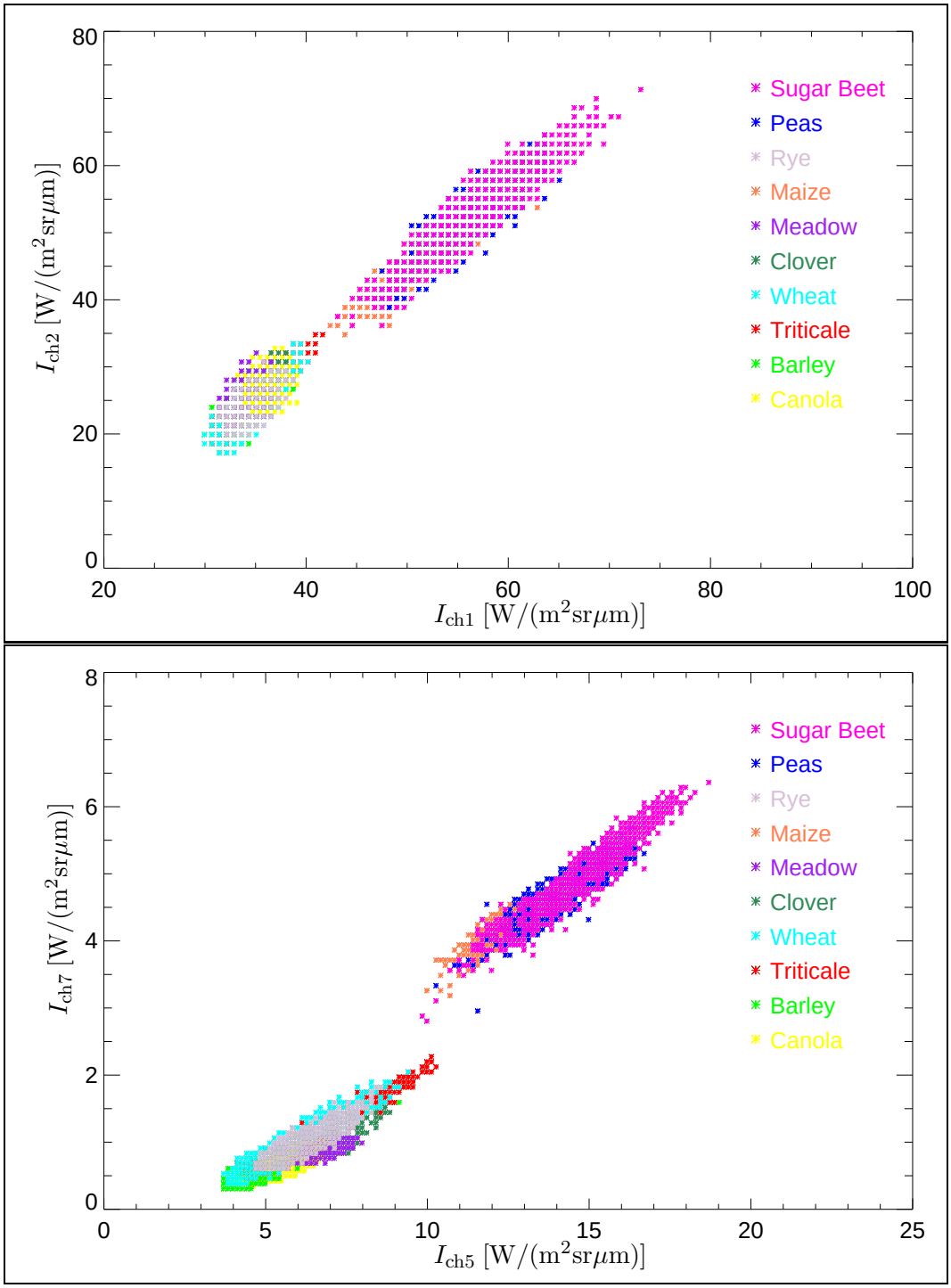



Figure 4.3: Scatter diagram for the radiances in two VIS channels (upper diagram) and two MIR channels (lower diagram) for 10 different agricultural plants. The diagrams are displayed as in Figure 4.2. Both diagrams show two scatter clouds that consist either of the radiances for early growing crops including canola (lower cloud in both diagrams) or late growing crops (upper cloud). Canola, being within the lower cloud, is difficult to distinguish from cereals and meadows using these channels.

The upper diagram shows that the early growing plants in the lower cloud are difficult to distinguish from each other with the TM channels 1 and 2, although both channels still allow a discrimination of early and later crops. This is also true for the MIR channels 5 and 7 displayed in the lower diagram of Figure 4.3.

Nonetheless, these two-dimensional plots do not represent the multidimensional shape of the clouds and the use of additional channels will improve the separability. A measure for the separability with the aid of the MLC will be presented in Section 4.2 (p. 95).

### 4.1.3 Channel Selection

The volume of data for multispectral satellite images from sensors like TM and LISS/3 is very large: A complete TM image, e.g., contains about 400 MB of data. Since these data have to be kept in the working memory of the computer for each processing step, it is of advantage to reduce the number of channels used for a classification. This will also result in a faster classification since fewer parameters have to be processed. Additionally, channels that are more sensitive to atmospheric variations might be ignored.

A reduction of channels is possible if the satellite data include redundant information in the available channels. For instance, the distribution of pixels in the upper diagram in Figure 4.3 indicates that the radiances of the TM channels 1 and 2 are linearly dependent, i.e., the radiance for TM channel 1 can be derived directly from the radiance of channel 2. This is also the case for the distribution of radiances for the MIR channels 5 and 7 in the lower scatter diagram of Figure 4.3, especially when focusing on the scatter cloud for the early plants.

This linear dependence is confirmed by the correlation coefficients that have been calculated for each channel combination and are listed in Table 4.1. The three VIS channels have a very high correlation of 0.99. This high correlation was already observed for channels 1 and 3 in the discussion of the HOT method (see Section 3.2.4, p. 71). Moreover, the distribution of the canola radiances for the channels 1 and 2 in Figure 4.3 is located within the scatter clouds of the cereals (rye, triticale, wheat), meadows and clover.

Therefore, the use of more than one VIS channel will not improve the discrimination of canola from the other crops. The channel that was selected is the TM channel 3, the one least influenced by aerosol scattering and thus the most suitable VIS channel for a surface type classification.

Moreover, the TM channel 3 has a similar counterpart on the LISS/3 sensor, and a comparable combination of channels is of advantage to implement the classification to LISS/3 and evaluating the classification results based on data from different sensors.

Similar to the VIS channels, the two MIR Channels are also highly correlated with a correlation coefficient of 0.98 and thus likewise only one MIR channel is necessary. The channel that was chosen is the channel 5 since it is less affected by water vapour absorption (Asrar, 1989, p. 340) and a similar

Table 4.1: Correlation coefficients for all TM channel combinations. The correlation coefficients (see Section 3.1.4, p. 57) have been calculated from the quillow training data set of 1999 for all ten selected agricultural plants. Note the very high correlation of 0.99 between the three VIS channels which indicate that only one of the VIS channels is necessary for a classification. The same can be observed for the two MIR channels which can also be represented by only one of both.

| Spectrum | VIS | | | NIR | MIR | |
|---|---|---|---|---|---|---|
| Channels | 1 | 2 | 3 | 4 | 5 | 7 |
| 1 | 1.00 | 0.99 | 0.99 | -0.40 | 0.96 | 0.96 |
| 2 | 0.99 | 1.00 | 0.99 | -0.37 | 0.96 | 0.96 |
| 3 | 0.99 | 0.99 | 1.00 | -0.48 | 0.97 | 0.99 |
| 4 | -0.40 | -0.37 | -0.48 | 1.00 | -0.50 | -0.58 |
| 5 | 0.96 | 0.96 | 0.97 | -0.50 | 1.00 | 0.98 |
| 7 | 0.96 | 0.96 | 0.99 | -0.58 | 0.98 | 1.00 |

channel is present on the LISS/3 sensor.

The remaining NIR channel shows only small correlation with the other channels and thus contains information not present in the other channels. An example is the high radiance of canola in Figure 4.2 for this channel.

The correlation between the VIS and MIR channels is quite high with a value of 0.96. Nonetheless, both channels will be used to identify the different crops.

Therefore, the combination of the TM channel 3, 4 and 5 is the one used to distinguish different types of plant covers in this study. The corresponding channels for LISS/II are also 3, 4 and 5. These channel combination is also used to display the satellite images as false colour image. Channel 4 is assigned to the red component of the image, channel 5 to green and channel 3 to blue .

Since the above consideration was only a qualitative one, this will be tested and discussed by using this selection on the one hand and on the other hand the complete set of TM channels for a classification with the MLC and then comparing the results of the classification with the Quillow mapping data set in the next section.

## 4.2 Pixel-Based Classification

The above discussion shows that the radiances of canola pixels are quite distinct from those of other agricultural crops which suggests a pixel-based classification. This type of classification is commonly used in the identification of agricultural crops (Richards, 1986; Schowengerdt, 1997). Nonetheless, the

accuracy of such a classification is estimated best by actually applying it.

The MLC described in (Richards, 1986, Chapter 8) is a very common classifier and frequently used in the identification of different crops with multispectral satellite images (Richards, 1986; Lillesand, 2000). The MLC is based on the comparison of probability density functions approximated with sets of training data for each surface type class. A reliable result from this method usually requires training data sets for all classes in the image. A complete set of surface classes is difficult to obtain since a TM or LISS/3 image contains numerous different surface types. These are difficult to determine and thus, the MLC is not applicable for the complete classification of the available satellite images.

If the classification is limited to the crops known from the Quillow mapping data set and the results of the classification is only compared to the agricultural fields mapped, the MLC allows to estimate the separability of the different crops by a pixel-based classification. Furthermore, the limitation of a pixel-based classification and the sensor properties can be discussed without the selection of additional non-agricultural surface classes or a probability threshhold.

In Section 4.2.2, p. 102 we will compare the results of the MLC classification with the Mahalanobis distance classifier (MDC) since this classifier has the advantage of only depending on one single training data set (Richards, 1986, Chapter 8). The results of the MDC with this classifier will also be tested with the Quillow mapping data set.

### 4.2.1 Maximum Likelyhood Classification

The MLC is trained with the data set for the ten major crops in the Quillow data set by calculating the probability density functions for the MLC for the years 1999, 2000 and 2001.

#### Reduced and Complete Channel Set Classification

After the classification, the original fields from the Quillow mapping data set are compared to the result from the MLC in order to estimate the accuracy of the classification. This classification has been performed with the reduced set of channels proposed above and with the complete set of available channels in order to demonstrate the capability of the reduced set of channels. The results of these classifications are used to calculate confusion matrices. The confusion matrices of the year 2001 are shown in Table 4.2. They list all classified crops.

Most obvious is the high fraction of canola detected by the MLC (86 %)[3]. Using the complete set channels for the classification achieves only 0.32 %

[3]Note that this value is obtained from the original Quillow data set. The acreage identified in the corrected Quillow mapping data set (see Appendix B) is 95 %.

Table 4.2: Confusion matrix of the MLC classification result for the Quillow data set. The satellite image used was the TM image 193/023; May 13, 2001. Listed are the fractions of pixels for the classification that are in accordance with the crop on the mapped fields, i.e, the diagonal elements shows the correct classifications. Shown are the results for the classification with the three selected (upper part of the table) and all TM channels (lower part of the table) available. The entry representing **canola** has been emphasised by bold typesetting. The remaining field crops have been sorted into two principal classes: `cereals` and *late crops* which are also emphasised by different fonts. The last row lists the fraction for each crop. The size of one pixel is 900 $m^2$.

| | | | Ground Truth [%] | | | | | | | | | | crop fraction |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | class [%] | canola | cereals | | | | late crops | | | clover | meadow | of ground truth area [%] |
| | | | | barley | triticale | wheat | rye | maize | peas | sugar beet | | | |
| Maximum Likelyhood Classifier on TM data from 2001 | TM Channels 3, 4 and 5 | canola | **86.28** | 2.58 | 1.25 | 2.13 | 3.54 | 0.67 | 0.34 | 4.24 | 3.04 | 1.30 | 19.78 |
| | | barley | 1.19 | `54.03` | `21.51` | `6.01` | `4.33` | 0.03 | 0.68 | 0.35 | 0.00 | 0.20 | 9.59 |
| | | triticale | 0.70 | `6.82` | `29.82` | `7.28` | `11.74` | 0.14 | 0.98 | 1.26 | 5.37 | 4.36 | 6.82 |
| | | wheat | 3.08 | `8.09` | `13.47` | `52.95` | `12.03` | 2.04 | 12.56 | 3.40 | 3.25 | 6.76 | 21.19 |
| | | rye | 4.31 | `17.84` | `26.22` | `24.82` | `55.15` | 8.15 | 11.66 | 12.38 | 1.75 | 7.24 | 17.75 |
| | | maize | 0.57 | 1.97 | 0.31 | 0.50 | 2.26 | *50.56* | *13.92* | *19.92* | 0.83 | 0.51 | 6.26 |
| | | peas | 1.12 | 3.85 | 1.04 | 1.13 | 1.52 | *16.64* | *56.15* | *9.82* | 1.55 | 1.02 | 4.01 |
| | | sugar beet | 0.44 | 1.38 | 0.09 | 0.25 | 3.63 | *20.58* | *1.92* | *47.46* | 0.21 | 0.67 | 4.59 |
| | | clover | 0.10 | 0.81 | 0.31 | 0.82 | 0.13 | 0.06 | 0.11 | 0.15 | 46.18 | 10.35 | 1.67 |
| | | meadow | 2.20 | 2.64 | 5.98 | 4.11 | 5.67 | 1.15 | 1.69 | 1.03 | 37.82 | 67.59 | 8.33 |
| | | Total [%] | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | TM channels 1, 2, 3, 4, 5 and 7 | canola | **86.62** | 2.22 | 1.20 | 2.25 | 3.97 | 0.67 | 0.58 | 4.23 | 3.10 | 1.35 | 20.03 |
| | | barley | 1.55 | `57.39` | `21.35` | `3.77` | `3.97` | 0.07 | 0.58 | 0.36 | 0.00 | 0.43 | 9.11 |
| | | triticale | 1.30 | `10.98` | `36.36` | `12.45` | `12.75` | 1.95 | 2.14 | 4.76 | 4.85 | 4.56 | 9.89 |
| | | wheat | 2.80 | `6.42` | `15.76` | `63.94` | `7.74` | 2.69 | 11.63 | 4.61 | 3.64 | 8.07 | 24.49 |
| | | rye | 3.25 | `12.92` | `20.20` | `12.31` | `61.51` | 6.11 | 10.87 | 7.69 | 0.63 | 2.37 | 12.01 |
| | | maize | 0.64 | 2.24 | 0.40 | 0.46 | 2.29 | *51.73* | *14.20* | *14.30* | 1.26 | 2.10 | 6.27 |
| | | peas | 1.04 | 3.42 | 0.91 | 1.01 | 1.32 | *12.31* | *56.59* | *7.92* | 1.02 | 1.57 | 3.48 |
| | | sugarbeet | 0.42 | 1.65 | 0.09 | 0.29 | 3.69 | *23.44* | *2.64* | *54.93* | 0.24 | 1.03 | 5.27 |
| | | clover | 0.26 | 1.17 | 1.63 | 1.66 | 1.00 | 0.70 | 0.29 | 0.97 | 48.81 | 11.89 | 2.47 |
| | | meadow | 2.13 | 1.58 | 2.09 | 1.86 | 1.77 | 0.32 | 0.47 | 0.25 | 36.45 | 66.65 | 6.98 |
| | | Total [%] | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | | Total [px] | 36787 | 17004 | 15360 | 55752 | 9086 | 15446 | 2659 | 8068 | 1938 | 12622 | 174722 |

better classification accuracy than using the reduced number of channels[4].

This shows that the reduced channel set is sufficient for the classification of canola. In contrast, the classification with all channels provides better results for cereals[5] and sugar beet since the accuracy achieved with the full set of channels is about 5 % higher than with the reduced set of channels. For the remaining crops, the classification with the complete set of channels yields slightly better results of 1 %.

Consequently, the use of the complete data set hardly improves the classification of canola acreage in a TM image. Considering the increased amount of computational time and computer memory and the higher sensitivity to the atmosphere of the omitted channels, the use of all channels poses more disadvantages than advantages. Therefore, the selected channels 3, 4 and 5 is the most suitable combination for the detection of canola. For cereals or sugar beet, however, the use of the complete channel set allowed to classify 7 % more of the acreage mapped in the ground truth data.

### Separability of the Different Crops

Besides the justification to reduce the number of channels, this result allows to discuss the separability of different crops with multispectral satellite data using the example of the TM sensor. For instance, the agreement of the satellite classification with the ground truth data is quite good for canola, which was expected from the qualitative discussion on the channel space distribution of the different crop classes. The crops mainly misinterpreted as canola (see the first row in table 4.2) are sugar beet (4.24 %), rye (3.54 %), clover (3.04 %) and barley (2.58 %). The amount of sugar beet identified as canola is surprisingly high since the scatter diagrams show a large distance in channel space between these crops; it probably results from errors in the Quillow mapping data set described in Appendix B. The confusion of the cereals and also clover is in agreement with their adjacency in the channel space (see Figures 4.2 and 4.3).

The fraction of 86 % canola acreage identified with the MLC is satisfying. Moreover, the selected TM channels have a direct counterpart in the LISS/3 sensor which should thus also be suitable for the classification of canola.

A similarly high agreement is not found for any of the other crops. This results mainly from the similarity of the plants or their growth state, especially for cereals, which are difficult to distinguish at this development stage even for an observer on the ground.

Another example are maize, sugar beet and peas, which have not reached a growth stage with a sufficient plant cover to be detectable by the satellite sensor and simply appear as bare soil in the satellite image (see also Figure 4.1). These similarities allow to merge these classes into two new principal classes: cereals

[4] Accuracy in this context is the percenatge of mapped acreages identified with the satellite image classification.

[5] Although maize is also a cereal, in this context the name cereal is used for barley, rye, triticale and wheat since these plants are planted much earlier than maize.

and late crops. These principal classes are already emphasised in Table 4.2 by the title of the columns and the different fonts used for the numbers in the confusion matrix.

### Different Satellite Acquisition Dates

The results for these merged classes are displayed in Table 4.3 for the three years investigated. The merging of the crop classes to two principal classes results in classification accuracies comparable to the one of canola. Thus a classification of cereals and late crops with a pixel-based classification would achieve results comparable to that of canola, assuming that the principal classes do not have to be broken down into the actual crop classes.

Similar to the discussion on the classification result of all ten classes for 2001, the most common plants to be mistaken for canola in 1999 and 2000 are also cereals. Generally, the results for 1999 and 2000 show a less accurate result for canola. The fraction of identified canola acreage is only 65 % for 1999 but still 80 % for 2000. Again, the high amount of late crops misinterpreted as canola suggests that at least 5 % of this misclassification results from wrong mapping. This assumption is confirmed by the discussion in Appendix B.

Besides the errors of the mapping data set, another important source of class error is the confusion of cereals and canola, which cannot be explained by the wrong mapping. Therefore, a clip of the classification result of the MLC is shown in the left clip of Figure 4.4: a lot of canola pixels have not been identified within the mapped field. Some pixels are identified as meadow but all remaining pixels in these fields are identified as cereals (not shown in Figure 4.4). The most likely explanation for this misclassification is obvious from the right clip of Figure 4.4, which shows that the canola fields do not have the yellowish colour that can be expected for flowering canola (see Figure 3.21, p. 80). The missing of this unique attribute of canola reduces the differences between the spectral signature of the different classes, especially between cereals and canola. The reason for this missing attribute is the acquisition date of the image: it was acquired on April 30, 1999 well before the main flowering period.

Nonetheless, the left clip in Figure 4.4 also shows that in all canola fields at least half of the canola pixels could be identified. This is also true for all other fields of the mapped area. Therefore, the pixels found can be used to make use of the neighbourhood to identify the remaining pixels of these fields, which will be described later in the discussion on segmentation and region growing.

### Drawbacks of the Maximum Likelyhood Classification

The above discussion on the MLC showed that a pixel-based classification gives acceptable results for the identification of canola, especially when compared to the corrected ground truth data (see Appendix B). Unfortunately, the MLC is not suitable for the classification of the complete data set as it can only be

Table 4.3: Confusion matrix for the comparison of the MLC for the years 1999, 2000 and 2001 with the ground-truthing provided by the Quillow data set. The channels 3,4 and 5 with training data set for the same classes as listed in table 4.2. Listed are only the merged new classes cereals and late crops. These classes are emphasised by the fonts used: cereals in typewriter (e.g., `99.99`), latecrops in italics (e.g., *99.99)* and canola in bold face (e.g., **99.99**). This table shows that the agreement for canola in years 2000 and 2001 is more than 80 %. The majority of the remaining mapped canola is misinterpreted as cereals. In 1999, the identified canola reaches only 65 % and 18 % are interpreted as cereals. Note the high value of 5.4 % canola pixels identified as late crops.

| | | | Ground Truth [%] | | | | | crop fraction |
|---|---|---|---|---|---|---|---|---|
| Maximum Likelyhood Classifier applied to TM channel 4, 5 and 5 | Date | Class [%] | canola | cereals | late crops | clover | meadow | of ground truth area [%] |
| | May 15, 2001 | canola | **86.28** | 2.20 | 1.73 | 3.04 | 1.30 | 19.78 |
| | | cereals | 9.29 | `89.58` | 14.10 | 10.37 | 18.55 | 55.36 |
| | | late crops | 2.13 | 3.26 | *82.91* | 2.58 | 2.20 | 14.87 |
| | | clover | 0.10 | 0.67 | 0.09 | 46.18 | 10.35 | 1.67 |
| | | meadow | 2.20 | 4.30 | 1.17 | 37.82 | 67.59 | 8.33 |
| | | Total [%] | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | | Total [px] | 36787 | 97202 | 26173 | 1938 | 12622 | 174722 |
| | May 2, 2000 | canola | **80.23** | 2.98 | 1.80 | 1.68 | 2.09 | 14.15 |
| | | cereals | 9.51 | `70.32` | 5.56 | 16.41 | 21.84 | 46.11 |
| | | late crops | 2.26 | 3.70 | *84.35* | 3.49 | 6.52 | 15.50 |
| | | clover | 6.56 | 17.68 | 7.56 | 38.37 | 28.89 | 15.99 |
| | | meadow | 1.44 | 5.31 | 0.73 | 40.05 | 40.66 | 8.25 |
| | | Total [%] | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | | Total [px] | 33382 | 132511 | 32719 | 2377 | 24262 | 225251 |
| | April 30, 1999 | canola | **65.71** | 1.18 | 1.53 | 0.79 | 0.44 | 12.42 |
| | | cereals | 17.90 | `91.14` | 10.73 | 26.52 | 54.67 | 60.81 |
| | | late crops | 5.40 | 4.71 | *86.59* | 3.16 | 4.76 | 19.10 |
| | | clover | 0.97 | 0.89 | 0.37 | 47.84 | 18.15 | 2.66 |
| | | meadow | 10.02 | 2.08 | 0.78 | 21.69 | 21.98 | 5.02 |
| | | Total [%] | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | | Total [px] | 38431 | 123906 | 38484 | 2153 | 17699 | 220673 |

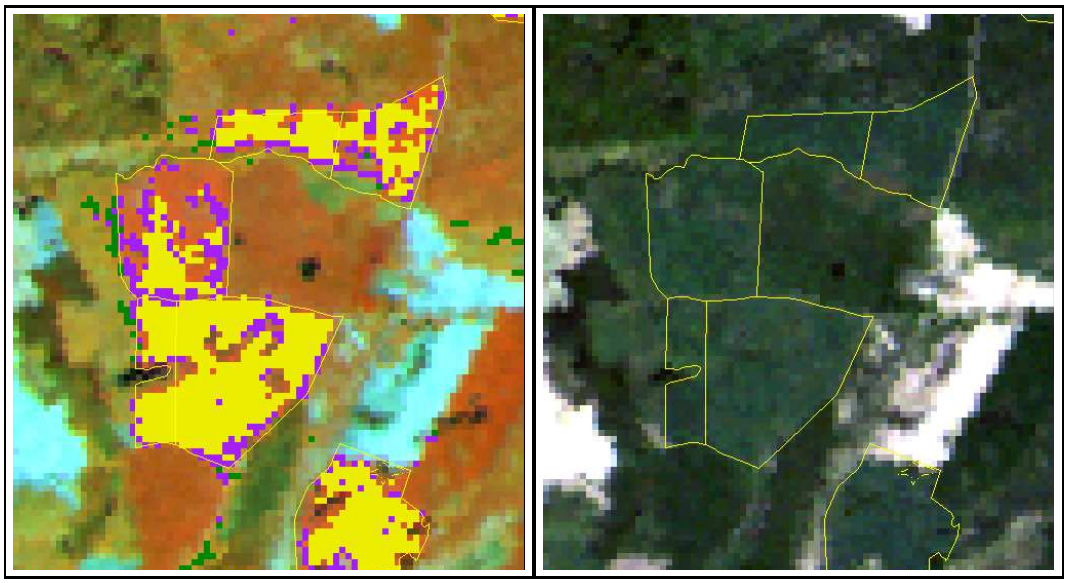


Figure 4.4: Clip of the Result from the MLC for 1999. Left: False colour representation overlaid with the classification result. Yellow pixels represent identified canola pixels, green pixels represent clover and purple pixels represent meadow. Note that pixels that are identified as cereals are not displayed in this figure since most of the other pixels in the clip are classified as cereals. The canola fields in the Quillow mapping data set are represented by the yellow lines. Right: True colour representation of the same clip, also overlaid with the mapped fields. The missing yellowish colour of the canola fields (see Figure 3.21) indicate that this image was acquired before flowering. Original data: LANDSAT TM ©ESA, 1999. Distributed by Eurimage.

applied if the majority of surface type classes in the image are known and an appropriate set of training data is available for each of them.

- The effort for a limited number of images, like the three TM images used for the investigation on the separability is manageable since it requires only to manually identify sufficiently large training data sets for ten different agricultural crops. This manual selection is no longer appropriate if a much larger volume of data has to be processed. The 50 images available in this study would require to manually select 500 sets of training data only for the agricultural crops. Moreover, the comparison with the validation data concentrated on the agricultural crops and a real classification would require to add additional classes like urban regions, woods or water.

- The automated selection of training data sets for canola that is used for overlapping frames (see Section 4.4.1, p. 126) might solve this problem, since it can also be applied for other surface type classes. However, even for an automated training data selection, the restriction to only one surface type class allows to limit the effort and also to reduce the error.

- The flowering of canola also reduces the applicability of the MLC since it can change the spectral signature of canola within a very short period of time (i.e., days) or within shorte distances (a few kilometres), which has been discussed above. The MLC can take this into account by selecting additional training data, which includes pixels that show flowering canola with a similar intensity. This type of training data might not be available if the overlap of the images does not contain fields with intensely flowering canola. Moreover, the pixels of these fields might be too rare to be of significance in the statistics and thus be ignored in the classification.

These shortcomings of the MLC can be overcome when using classifiers that can be applied if only one surface class is known, like the parallelepiped classifier (PPC) and the MDC for the classification.

### 4.2.2 Single-Class Classification

The distribution of the pixel radiances in the TM channel 3, 4 and 5 for canola is shown in Figure 4.5. A classification algorithm based on the information of the radiance distribution for a single class has to represent the shape of this distribution accurately. The geometrical shape describing the boundaries between the radiances of pixels assigned to this single class and pixels not assigned to it is called a decision surface.

#### The Parallelepiped Classifier

The most simple solution in the case of a classification with three channel is a three dimensional box. The classification with this decision rule is called PPC and is represented by a box centred at the mean radiances $\bar{I}_i$. The length of the edge in the direction of the radiances for channel $i$ of the box can be estimated by the variance $\sigma_{ii}$ of that channel. Usually, a multiple $k$ of the standard deviation $\sqrt{\sigma_{ii}}$ is used and the pixel is classified as canola if the following equation is true for all selected channels $i$:

$$\bar{I}_i - k\sqrt{\sigma_{ii}} < I_i < \bar{I}_i + k\sqrt{\sigma_{ii}} \quad \text{for} \quad i = 3, 4, 5 \tag{4.1}$$

An example of the PPC is marked by the dashed rectangles in the diagrams of Figure 4.5, for which $k = 3$ has been selected. The distribution of canola is not represented well by the rectangles and consequently not by the box. This is especially obvious from the left diagram which shows a slanted distribution of radiances[6] in channel 5. Therefore, the PPC is not suitable to represent the radiance distribution for canola and a better approximation of the shape of the distribution has to be found.

[6] Actually Figure 4.5 shows the original data $q_{\text{cal}}$ in DN which are linearly related to the radiances.

### The Mahalanobis Distance Space

The MLC is based on the assumption of a known type of distribution for the data (Richards, 1986, Chapter 8). Generally a normal distribution is assumed. Furthermore, it was shown that the surface of equal probability of an $n$-dimensional normal distribution is an $n$-dimensional ellipsoid. This ellipsoid can be obtained directly from the covariance matrix $\Sigma$ and the mean of the radiances $\bar{\mathbf{I}}$. In the case of the reduced channel set used for the classification of canola, covariance matrix and the classes mean radiances are:

$$\Sigma = \begin{pmatrix} \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{53} & \sigma_{54} & \sigma_{55} \end{pmatrix} \quad \bar{\mathbf{I}} = \begin{pmatrix} \bar{I}_3 \\ \bar{I}_4 \\ \bar{I}_5 \end{pmatrix} \tag{4.2}$$

where $\sigma_{ij}$ denotes the variances ($i$=$j$) and covariances ($i \neq j$) calculated for the distribution of the radiances $I_i$ and $I_j$ for channel $i$ and $j$.

The direction and length of the principal axes of this ellipsoid can be calculated from the covariance matrix with an eigenvalue analyses (Richards, 1986). The eigenvectors $\mathbf{e}_i$ determine the direction of the principal axes and the eigenvalues $\lambda$ indicate their length. This eigenvalue analysis thus allows to obtain information on the shape and attitude of the ellipsoid describing the surface of equal probability.

Slices of this ellipsoid are shown in the three scatter diagrams in Figure 4.5 by the ellipses enclosing the radiance distribution. Additionally, the projections of the eigenvectors multiplied with the eigenvalues determine the size of the ellipsoid with the confidence factor $k = 4$. The figure shows that the ellipses represent the distribution of the radiances much better than the rectangles used by the PPC. Note that the size of the ellipse in the direction of channel 5 appears to be too small. This results from the inclination of the scatter cloud in channel 4 and only represents a slanted slice of the ellipsoid centre.

Consequently, this ellipsoid is a good decision surface for the identification of canola pixels. Unfortunately, the decision surface is a slanted ellipsoid with principal axes of three different lengths. A much simpler representation can be found by transforming the radiances to a new coordinate system with the origin at the mean radiances $\bar{\mathbf{I}}$ for the canola class, the axes transformed to the coordinate system described by the eigenvalues of the eigenvectors. These coordinates are scaled by the eigenvalues so that the class ellipsoid is transformed to a sphere.

This can be achieved by using the eigenvectors to transform the radiance vector $\mathbf{I}$ to a new vector of transformed radiances $\mathbf{I}'$ scaled with the square roots of the eigenvalues. The components of this vector are:

$$I'_i = \frac{\mathbf{e}_i(\mathbf{I} - \bar{\mathbf{I}})}{\sqrt{\lambda_1}} \quad \text{for} \quad i = 1, 2, 3 \tag{4.3}$$

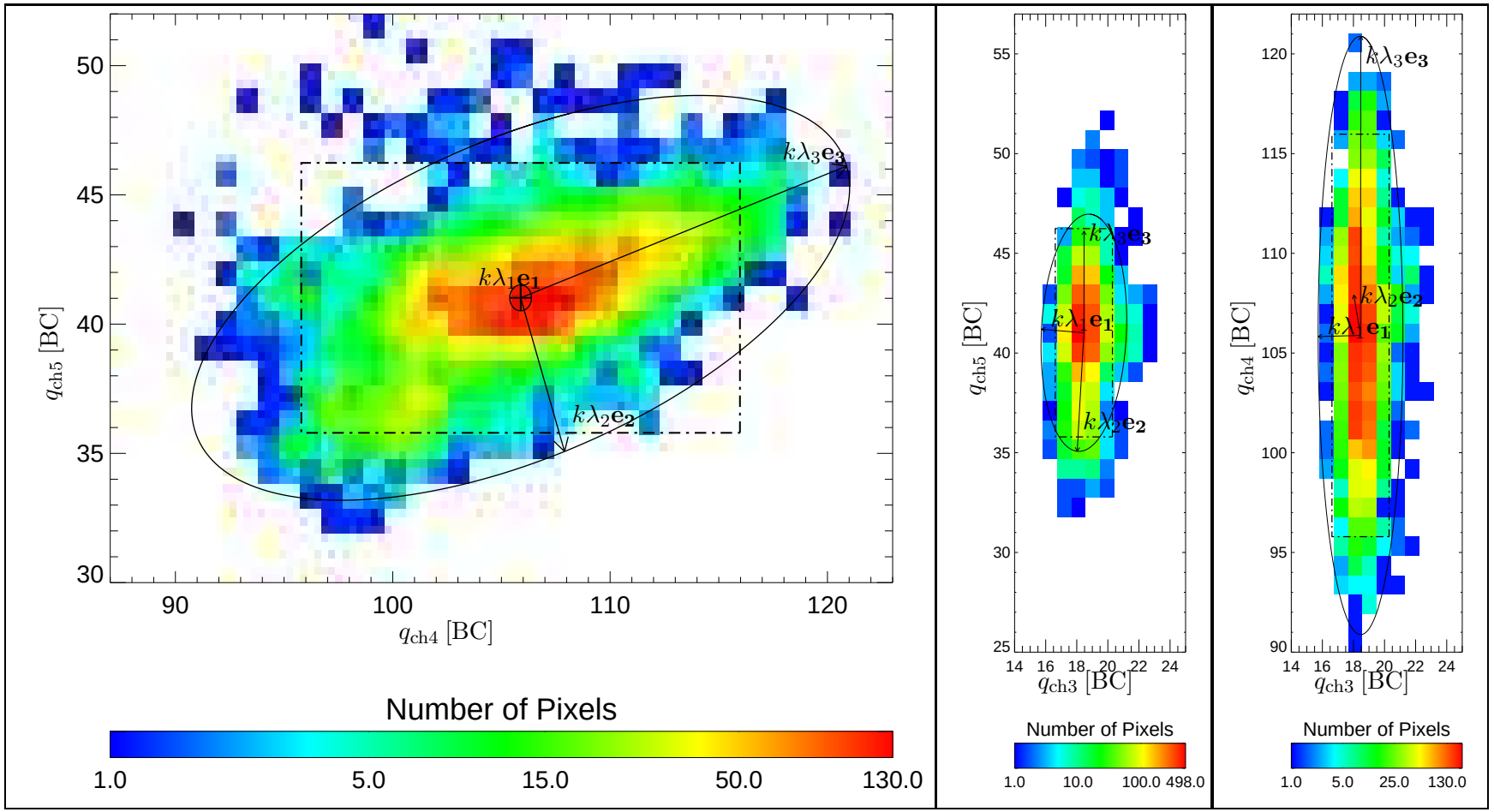



Figure 4.5: Logarithmic density scatter diagram of the DN values $q$ for the TM channels for the canola training data set (see Figure 4.1) from image 193/023; April 30, 1999. Left: Channels 4 and 5; Middle: Channels 3 and 5; Right: Channels 3 and 4. The dashed rectangle represents the tripled standard deviation for the corresponding channels and a possible decision surface for the PPC. The ellipse indicates the ellipsoid formed by the surface of equal probability with $k = 4$ projected onto the plane of the displayed axes, i.e., a possible Mahalanobis distance decision rule. It can be seen that the distribution is rotated with respect to the band axis and thus the transformation with the MDC gives better results than a parallelepiped classifier.

With this transformation, the covariance matrix for those new coordinates becomes the identity matrix:

$$\Sigma' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.4}$$

This new coordinate system, called Mahalanobis distance space (MDS), allows a much easier evaluation of the separability of classes than the original channel space. One reason is that in this space the distance of pixels is measured by the normalised standard deviation of the canola class. Moreover, the distribution of the other agricultural classes are also normally distributed and thus can be described as ellipsoids in channel space similar to the one of canola. These ellipsoids are also transformed to the MDS and remain ellipsoids in the new space. In the MDS these ellipsoids only have to be compared to a sphere to obtain information on the relations of the canola class and other classes for other crops.

### The Mahalanobis Distance Classifier

The relations from the previous section are used to adapt the Mahalanobis distance classifier to the distribution of crop classes. The decision rule for the classification is based on the canola sphere in the MDS and the confidence factor $k$:

$$I'^2_1 + I'^2_2 + I'^2_3 \leq k^2, \tag{4.5}$$

The variation of the confidence factor allows to adapt this rule to the distribution of other classes and since the equation describes the distance of a pixel in the MDS, the classification is named Mahalanobis distance classifier. The selection of $k$ is based on the following aspects:

- The fraction of canola pixels within the sphere in the MDS, i.e., the class statsitics.
- The distance in the Mahalanobis space to the other surface types distributions, i.e., the position and the extent of the ellipsoids for the pixels of other agricultural crops.

The first issue can be answered according to Press et al. (Chapter 15, p. 697, 1992) who state a factor of $k^2 = \Delta\chi^2 = 8.02$ or $k = 2.83$ to include 95.4 % of the canola pixels in this 3-dimensional ellipsoid.

This estimation is only based on the statistics of the radiances for the canola class. The other crop classes have not been taken into account for the selection of this confidence ellipsoid. Now, the ellipsoids of the other classes are transformed to MDS, are used to investigate their position, extent and attitude relative to the canola class.

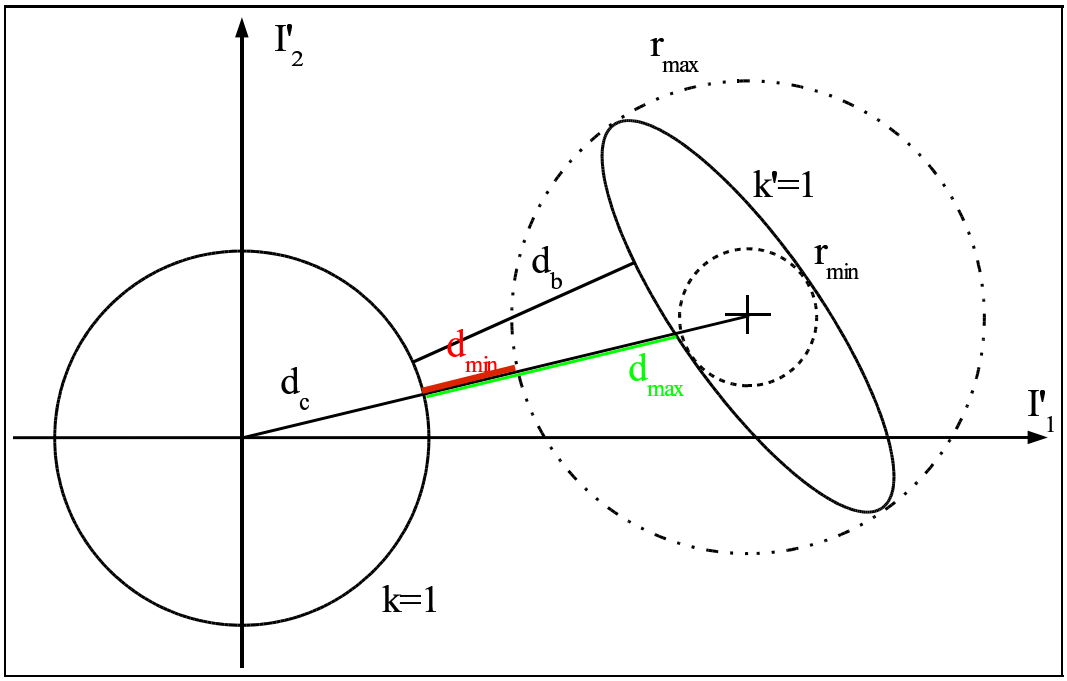



Figure 4.6: Sketch of the relations of canola and other agricultural crops class ellipsoids in the canola Mahalanobis space. The sphere on the left represents the ellipsoid of equal probability for canola with a radius of $k = 1$. The ellipsoid on the right illustrates the shape and position of the ellipsoid for another crop, e.g., rye or triticale. The ellipsoid on the right is described by to main axis that are illustrated by the dashed circles with the smallest $r_{\min}$ and largest radius $r_{\max}$. There are four different distances shown in this figure: $d_c$ is the distance between the origin and the class mean of the non-canola class; $d_b$ is the shortest distance from the sphere to the ellipsoid; $d_{\max}$ the largest distance from sphere to ellipsoid and $d_{\min}$ the shortest possible distance. The last two distances are approximations since the $d_b$ depends on the attitude of the ellipsoid. These parameters have been calculated for all classes and are shown in Table 4.4

.

A sketch of the situation in the MDS is shown in Figure 4.6 to illustrate the relation of the class ellipsoid and the canola sphere. A good estimation for the optimal size of the canola sphere is to evenly enlarged both, the sphere and the ellipsoid, by a factor $k$ until both merely touch in one point. Then $k$ would describe the maximum extent of the sphere and ellipsoid with no overlap.

Since it is complicated to calculate this optimal enlargement analytically, it is estimated using the parameters described in Figure 4.6. They have been calculated for the ground truth data from the Quillow data set (see Table 4.4). The first parameter important for the separability is the distances $d_c$ of the ellipsoids centre, or from the non-canola ellipsoid centre. As expected from the previous discussion, the closest ellipsoids are the ones of the cereals and meadow with a distance of 4 to 6.

Table 4.4: Distance and extent of the different probability ellipsoids for $k = 1$ for the training data sets of the year 1999 transformed to the canola mahalanobis space. $\phi$ is the angle between the longest axis of the ellipsoid and the conection line of origin and class centre.All other sympbols are explained in Figure 4.6

| | class | $d_c$ $[\sigma_c]$ | $r_{\max}$ $[\sigma_c]$ | $r_{\min}$ $[\sigma_c]$ | $d_{\min}$ $[\sigma_c]$ | $d_{\max}$ $[\sigma_c]$ | $\phi$ [°] |
|---|---|---|---|---|---|---|---|
| May 15, 2001 | canola | 0.00 | 1.00 | 1.00 | - | - | - |
| | barely | 4.18 | 1.79 | 0.41 | 2.39 | 3.77 | 83.39 |
| | triticale | 4.86 | 2.91 | 0.41 | 1.95 | 4.45 | 92.17 |
| | rye | 6.13 | 2.27 | 0.40 | 3.87 | 5.73 | 92.92 |
| | wheat | 6.25 | 2.86 | 0.45 | 3.39 | 5.80 | 94.82 |
| | clover | 10.04 | 5.51 | 0.60 | 4.53 | 9.44 | 42.21 |
| | maedow | 12.00 | 3.43 | 0.81 | 8.57 | 11.19 | 26.11 |
| | peas | 41.57 | 5.54 | 1.29 | 36.02 | 40.28 | 21.14 |
| | maize | 41.83 | 4.06 | 0.68 | 37.77 | 41.15 | 91.54 |
| | sugar beet | 47.63 | 3.65 | 0.35 | 43.98 | 47.28 | 32.34 |
| May 2, 2000 | canola | 0.00 | 1.00 | 1.00 | - | - | - |
| | triticale | 4.37 | 1.62 | 0.32 | 2.75 | 4.05 | 71.65 |
| | rye | 4.54 | 1.59 | 0.33 | 2.95 | 4.21 | 81.47 |
| | wheat | 4.67 | 2.06 | 0.36 | 2.60 | 4.31 | 86.54 |
| | barley | 5.47 | 2.03 | 0.31 | 3.44 | 5.16 | 58.04 |
| | clover | 5.68 | 2.76 | 0.53 | 2.91 | 5.15 | 68.49 |
| | maedwo | 6.46 | 1.71 | 0.55 | 4.75 | 5.91 | 121.13 |
| | peas | 16.36 | 4.20 | 0.72 | 12.16 | 15.64 | 52.03 |
| | maize | 18.39 | 3.91 | 0.85 | 14.49 | 17.54 | 49.50 |
| | sugar beet | 20.97 | 4.46 | 0.51 | 16.51 | 20.46 | 40.27 |
| April 30, 1999 | canola | 0.00 | 1.00 | 1.00 | - | - | - |
| | barley | 6.28 | 2.32 | 0.95 | 3.95 | 5.32 | 80.49 |
| | maedow | 7.45 | 1.55 | 1.48 | 5.90 | 5.97 | 117.20 |
| | clover | 8.19 | 1.40 | 1.09 | 6.80 | 7.10 | 48.35 |
| | rye | 8.74 | 2.00 | 0.94 | 6.74 | 7.80 | 140.96 |
| | triticale | 8.77 | 5.17 | 0.76 | 3.60 | 8.02 | 57.07 |
| | wheat | 9.39 | 3.15 | 1.30 | 6.24 | 8.09 | 69.13 |
| | maize | 41.64 | 3.93 | 0.63 | 37.71 | 41.01 | 32.99 |
| | peas | 51.41 | 4.80 | 0.98 | 46.61 | 50.42 | 21.78 |
| | sugar beet | 54.33 | 7.56 | 1.75 | 46.77 | 52.58 | 21.78 |

The extent of the distribution is best described by the radii of the enclosing and enclosed spheres and the distance to the centre. The necessary enlargement $k$ for the two spheres to touch each other can be calculated from the relations:

$$k_{\max} + k_{\max} r_{\max} = d_c \quad \text{and} \quad k_{\min} + k_{\min} r_{\min} = d_c,$$

where $k_{\max}$ represents the confidence factor for the enclosing sphere and $k_{\min}$ the confidence factor for the enclosed sphere.

Assuming a minimum distance of 4.18, which is the smallest distance found in Table 4.4 and the largest sphere, $k$ has a value of 1.23, which is very small and would, according to Press et al. (Chapter 15 1992), a $k_{\min}$ of 1.49 and include only 68 % of the canola pixels. This is only the case, if the ellipsoid is directly pointing in the direction of the origin. But since the angles also listed in Table 4.4 found for the cereals are all nearly perpendicular to the connecting lines of the class centres the smaller, enclosed sphere is the one more likely to provide the relation between the class distributions. By taking the radius of 0.41, the confidence factor becomes $k_{\max} = 2.96$ which will include more than 99.73 % of the canola pixels.

This value is still an estimation and therefore, three different values for $k$ will be tested. Figure 4.7 shows the result of the classification for $k$=4 and $k$=5 compared to the result of the MLC classification result for canola. From those clips, it can be seen that the result for the MDC gives nearly as good results as the MLC. We also recognize that a $k$ of 4 misses some pixels of canola, mostly inside the fields. Some of these can be identified by incerasing $k$ to 5. However, this as expected increases the number of misclassifications outside these fields. This figure shows that the canola pixels not identified by the MDC are within the fields and are thus detectable by a region growing (see Section 4.3.2, p. 119). Similarly, the positive misclassifications are usually single pixels or small segments. From this apriori knowledge, these errors can be corrected in a postprocessing step.

A quantitative comparison of the classification is shown in Table 4.5. The MLC delivers better classification results than the new MDC algorithm since it misses more than 15 % of the mapped canola pixels found by the MLC. The larger confidence factor allows a better identification of canola but also increases the misinterpretation of other plants as canola.

This result can be expected from the training of the neighbouring surface type classes, that has been ommitted in the MDC algorithm. Nonetheless, most of these inaccuracies can be compensated with some further additional processing that will be presented hereafter.

### Impact of Canola flowering

One reason for the misclassification of canola results from canola flowering. Although this factor facilitates the identification of canola because of its distinct reflectance, it impedes the selection of training data sets because of the variation of reflected radiances caused by the variation of flowering. It is mainly

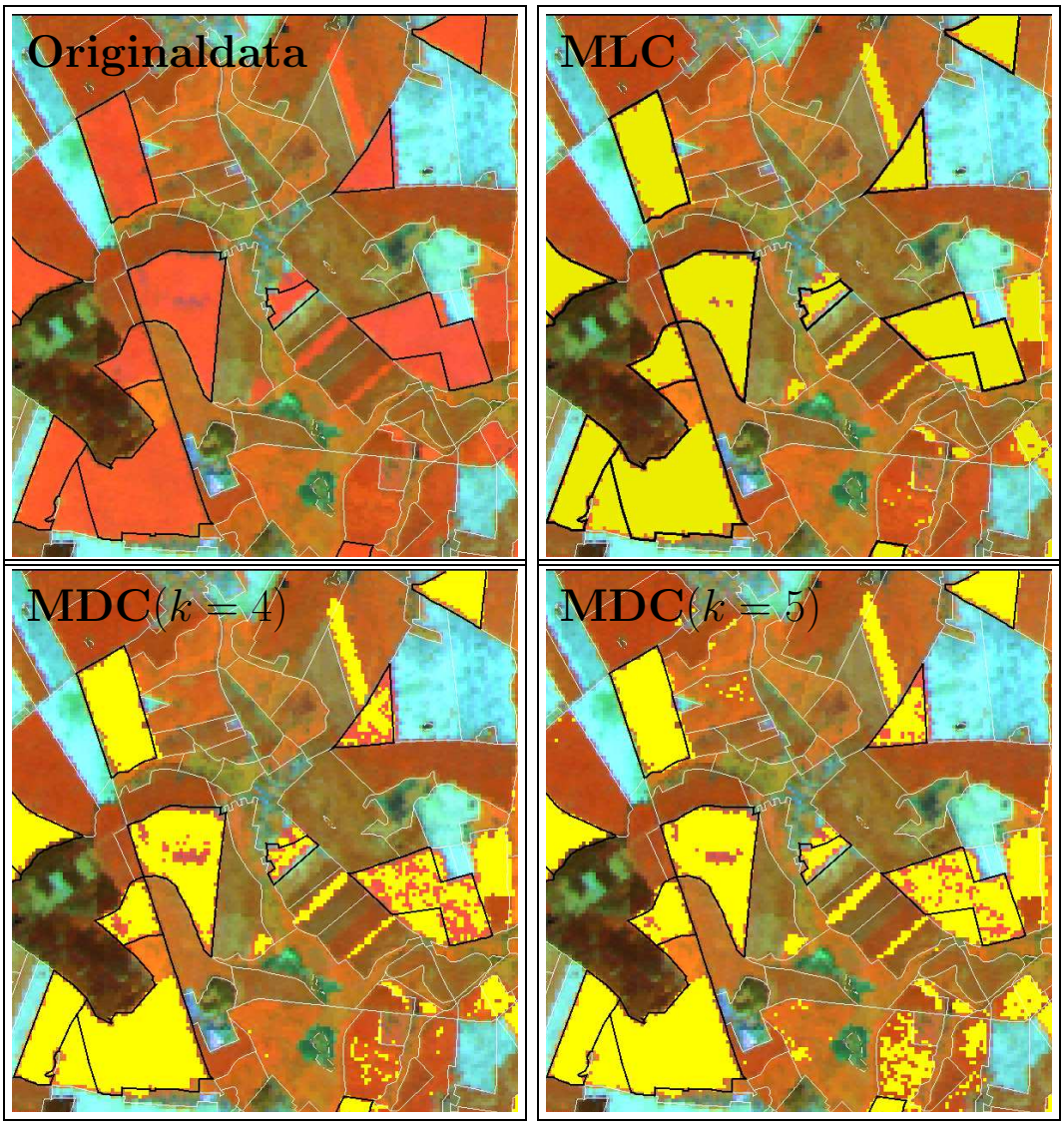



Figure 4.7: Comparison of the classification result for the MLC and the MDC with two different threshold parameters for the TM image 193/023; May 13, 2001. Top left: original data in false colour representation, canola fields are marked by the thick black lines, other fields by white lines. Top Right: Result for the identification of canola with the MLC. Bottom Row: MDC with $k = 4$ (left) and $k = 5$ (right). The best result is achieved with the MLC classification which is obvious from the pixel in a canola field, that have been identified almost completely. Moreover, the number of misclassifications is quite small (see lower right field in the upper right clip). Also visible are several errors in the Quillow mapping data set, e.g., the two narrow field in the center are not being mapped and the field in the right middle has only been partly mapped. Original data: LANDSAT TM ©ESA, 2001. Distributed by Eurimage.

Table 4.5: Comparison of the MLC classification result with the result of the MDC for three different $k = 3.5$, 4 and 5 validated with the quilow mapping data set. It can be seen, that the fraction of identified pixels achieved with the MLC cannot be achieved with the MDC unless the confidence factor is enlarged to more than 5.

| | | Satellite Based Classification | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | classifier | MLC | | MDC $k = 3.5$ | | MDC $k = 4$ | | MDC $k = 5$ | |
| | | class | canola | non-canola | canola | non-canola | canola | non-canola | canola | non-canola |
| Ground Truth | 2001 | canola | 86.28 | 2.57 | 68.15 | 4.48 | 71.07 | 7.66 | 75.17 | 16.45 |
| | | non canola | 13.72 | 97.43 | 31.85 | 95.52 | 28.93 | 92.34 | 24.83 | 83.56 |
| | 2000 | canola | 80.23 | 2.65 | 71.63 | 1.74 | 75.37 | 2.31 | 82.46 | 21.92 |
| | | non canola | 19.77 | 97.35 | 28.37 | 98.26 | 24.63 | 97.69 | 17.54 | 77.08 |
| | 1999 | canola | 65.71 | 1.18 | 60.45 | 28.93 | 63.72 | 1.19 | 68.53 | 1.97 |
| | | non canola | 34.71 | 98.82 | 4.64 | 95.36 | 34.28 | 98.81 | 31.47 | 98.03 |

caused by an increase of the number of blossoms and does not conform to the normal distribution determined for the MDC from region with less strongly flowering fields. Thus selecting training data in regions with generally weaker flowering canola can lead to misclassification resulting from variations of the flowering in other parts of the satellite image.

Note that differences in strength of flowering also occures inside fields, as visible in Figure 4.8. This a possible explanation for the gaps in some field for the classification result in Figure 4.7, which are likely the result of such variations.

This effect can already be observed in the relatively small area in which the MLC is tested, but it becomes more important if the classified area and the training data set are separated by a longer distance, i.e., 250 km at maximum, if the training data set is located diagonally in the other corner of the satellite image.

Figure 4.8 shows an example for such a situtation. The displayed area is located in the same frame as the Quillow ground truth data which is used as training data set, but located southeastwards at about 150 km distance. The upper left clip shows the MDC classification result. The pink pixels are canola which is easily verified in the upper right clip with the true colour representation. The pink pixels of the false colour clip appear in the typical yellowish green colour of flowering canola. Therefore, the pink pixels represent canola. The reason for their misclassification is obvious from the right clip: the pixels that are not identified as canola show a brighter yellow colour than the corretly identified pixels and indicate a stronger flowering in these fields. This stronger flowering is probably due to the climate in this region.

In the lower left clip, the HOT value for this image is displayed, in order to exclude aerosol scattering as a possible cause for the misclassification. The low HOT values of less than 3.5 indicate that the misclassifications are not resulting from aerosol scattering.

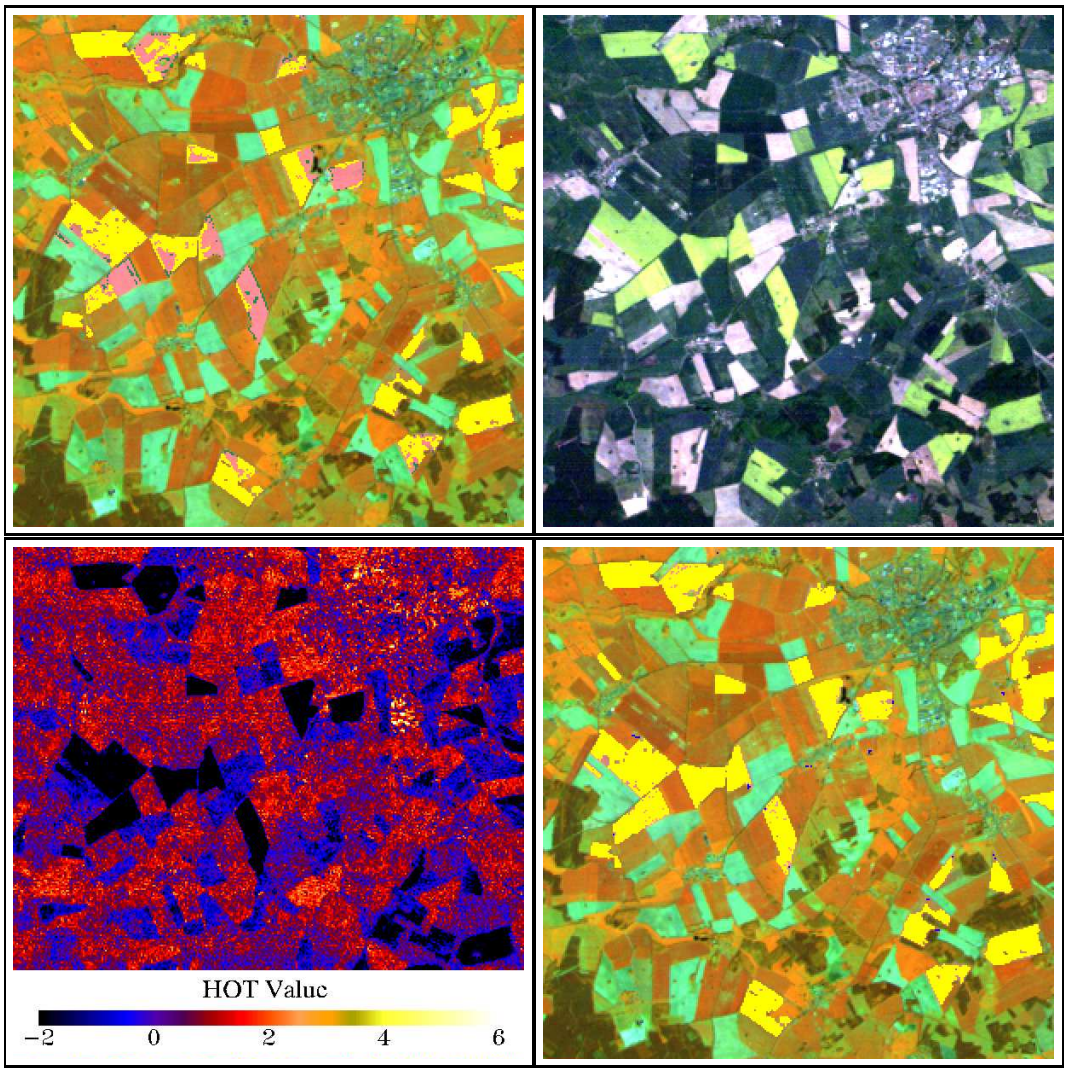



Figure 4.8: Example of missed classification resulting from intensely flowering canola. Top left: False colour image overlaid with the MDC result: identified canola pixel are masked with yellow and dark green. Top right : True colour image of the same clip. Note the pink area in the left upper clip which have not been classified corectily although false and true colour image both indicate, that these are canola fields. Bottom left: HOT value of this clip to demonstrate that the missclassification is not resulting from aerosol scattering. Bottom right: Classification result of the MDC adapted to the flowering of canola. Original data: LANDSAT TM ©ESA, 2000. Distributed by Eurimage.

Therefore, it is necessary to account for this variations of the flowering intensity of canola in the complete satellite image. A simple approach to overcome this problem is to enlarge the classification sphere discussed in Section 4.2.2, p. 105. However, this would increase the misinterpretation of other crop pixels as canola which can be observed in Figure 4.7. Therefor, it is necessary to investigate the influence of the canola flowering on the reflected radiances. This is done by selecting an additional set of training data (8000 pixels for the

image 193/023; May 2, 2000) from pixels that are missed by the classification. This new surface type class of "flowering canola" can be used to investigate its relations to the original canola class and the classes of other crops. The position of the class centre of the new class in the channel space is located

$$\Delta\bar{\mathbf{I}} = \begin{pmatrix} \Delta\bar{I}_3 \\ \Delta\bar{I}_4 \\ \Delta\bar{I}_5 \end{pmatrix} = \begin{pmatrix} 0.82 & \cdot 13.03 \\ 0.06 & \cdot 5.01 \\ 0.11 & \cdot 6.47 \end{pmatrix} \frac{\mathrm{W}}{(\mathrm{m}^2\mathrm{sr}\mu\mathrm{m})}$$

from the centre of the original canola class[7]. Thus, the flowering of canola results mainly in an increase of the reflected red light. This is to be expected from the yellow colour of the blossoms. An smaller increase of mean radiances can also be observed for the two infrared channels.

Additionally to the mean radiance of the flowering canola distribution, the shape of this distribution is also of interest. This is best described by the ellipsoid of equal probability. Therefore, the ellipsoids used to characterize the various crops classes in comparison to the canola class is also used to describe the relation between the classes of canola, flowering canola and the other crops.

The evaluation of the floweing canola ellipsoid shows that its distance from the origin $d_c$ is 3.73 $\sigma_c$. This clearly explains the missclassifications because this value is near the boundary of the classification sphere defined by $k = 4$. The extent of the flowering canola ellipsoid is smaller than the original canola sphere, which can be seen from radii of the enclosing and enclosed spheres $r_{\max} = 1.1$ and $r_{\min} = 0.68$ and thus a large part of the flowering canola pixels are not detected by the MDC.

Consequently, the MDC has to be adapted for the flowering of canola by enlarging the decision sphere so that it also includes flowering canola pixels and excludes non-canola pixel. As discussed above, a symmetric enlargement would also increase the misclassifications (see Figure 4.7). Therefore, the enlargement of the sphere has to be asymmetric and accommodated especially to the flowering canola.

It is found, that the class closest to flowering canola is rye, whose class centre is located at a distance of 7.04. This large distance to non canola classes in the MDS are common for the three images investigated. Unfortunately, it is has not been possible to find a direct relationship between the flowering and the original canola class that could be exploited to identify flowering canola with an acceptable accuracy by adapting the original class. The reason is probably the different flowering states of the original canola class, which is also changing with time and therefore has no constant relationsship to the flowering canola class.

Therefore, it is not possible to automatically identify the flowering canola fields without the identification of additional training data sets. Rather it is

[7]The first numbers for the components are the gain of the calibration and the second number the mean DN calculated.

necessary to select such data sets for every from the image by visual inspection. This additional training data can be identified easily by selecting the pink coloured pixels in the false colour image that have been missed by the classification.

This new learning data is used to train the MLC for flowering canola and is applied in conjunction with the original canola MLC. The lower right clip of Figure 4.8 shows the classification result for the image clip with this additional flowering canola class. All fields now have been identified properly.

The manual selection of additional class of flowering canola is a drawback for automatic classification aimed for in this study. and it would be desirable to obtain this training data automatically.

### 4.2.3 Haze Correction

The above described classification assumes that the pixel are not covered or shaded by clouds. Cloud-covered and shaded pixels are simply omitted from the classification with the cloud and cloud shadow masks presented in Section 3.2.4 (p. 81) and pixels adjacent to clouds or cloud-shadows are marked as cloud neighbours (see Section 4.3.3).

Pixels covered only by thin clouds still can be classified if the influence of the aerosol scattering can be compensated for the HOT-Value (see Section 3.2.4, p. 71) is used to quantify the haziness of a pixel.

In principle the HOT value is used to correct the radiances with a term derived from the comparison of histograms (see Section 3.2.5, p. 83), but as known from that section, the HOT value is not calculated correctly for flowering canola fields. Therefore, it is necessary to distinguish two different effects of haze on the classification:

**False alarm:** A pixel is identified as canola although it contains a different surface type.

**False all-clear:** A pixel is not identified as canola although it contains enough canola to be classified under clear sky conditions.

Examples for both cases can be seen in Figure 4.9, which shows a comparison of TM images showing the same area in Mecklenburg-Vorpommern from the same year but acquired under cloud free conditions (upper right clip) and under hazy conditions (lower left clip). Both images are overlaid with the classification result of the MDC without atmospheric correction. The classification result is already segmented (see below) and small segments (green) have been marked additionally to the identified canola pixels (yellow). Since both clips are acquired in the period of about 30 days, the canola fields identified in the images should be identical, which is obviously not the case. Since the left clip is acquired under cloud free conditions, this classification is used as reference to evaluate the influence of the haze. The two types of misclassification and their correction will be discussed separately hereafter.

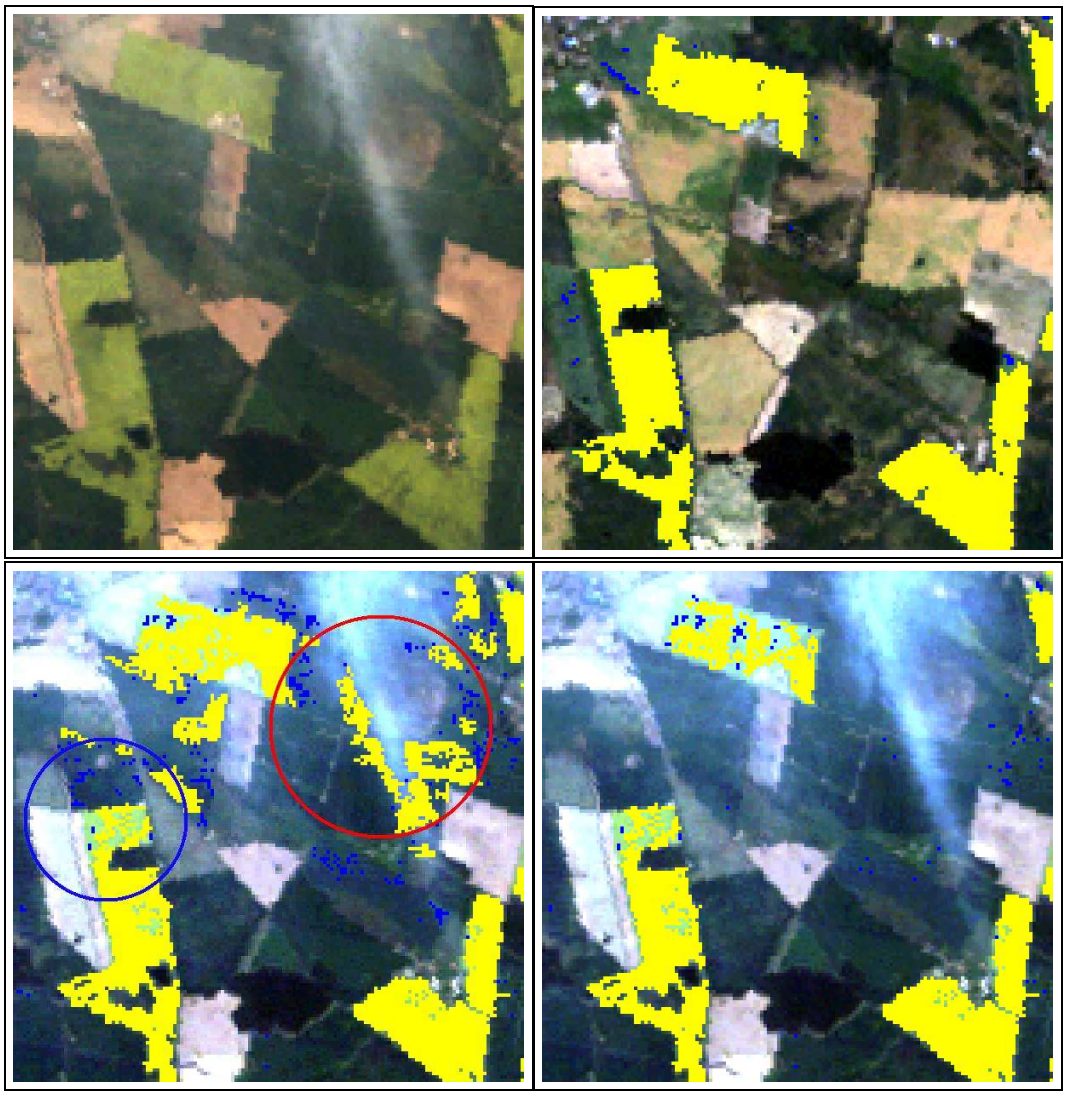


Figure 4.9: Example of the influence of haze on the classification of canola. All clips show the same region in the same year but are taken from two different images, one from the cloud free image 194/023; June 10, 2000 and the other from the partly clouded image from 193/023; May 8, 2000. Upper Left: True colour representation of the hazed image. Top right: Classification result of canola with the MDC under clear sky conditions. Bottom left: The same clip classified by the MDC in hazy and cloudy conditions. These images show the two different errors induced by haze cover. The blue circle shows a false-all clear, i.e., a field that has been partly missed by the MDC. The red circle shows an example of a false alarm, i.e., pixels that are incorrectly identified as canola by the MDC. The blue pixels indicate undersized segments. Bottom right: classification result of the MDC with additional HOT based haze correction. Note that all false alarm pixels are identified but the false-all clear pixels are corrected. Original data: LANDSAT TM ©ESA, 2000. Distributed by Eurimage.

### False Alarm

Some examples of this misclassifications are marked by the red circle in the lower left clip in Figure 4.9. This error occurs if the radiances reflected from the field are altered by aerosols resulting in signal resembling the one of canola, i.e., the signature is shifted into the decision ellipsoid presented in Section 4.2.2. From Figure 4.9, it can be seen that this error occurs quite frequently under hazy conditions. Most of the misclassifications are already masked out since they do not meet the minimum size criterion for canola fields (see below). These pixels are marked in blue.

The surface below that cloud is thus not a canola field and therefore, the HOT values for these pixels are correct. A correction for this misclassification is thus achieved by applying the shift suggested in Section 3.2.5, p. 83. Practically, this is done by applying this correction every time the HOT value exceeds 3.5. This threshold value is determined from observations of the missclassification of the clear sky classification algorithm. The results of the correction are shown in the lower right clip of Figure 4.9. It can be seen that all pixels incorretly identified as canola have been removed from the classification.

### False All-Clear

This type of errors occurs if the cloud alters the signal from a canola field in the manner, that the radiances is no longer identified as canola by the classification algorithm. An example is also shown the lower left clip in Figure 4.9 and marked by the blue circle.

This type of error is more difficult to correct than the false alarm, mainly because flowering canola impacts the HOT value and the haziness is underestimated over such fields.

There are three possible solutions for this problem. The first and optimal solution would be to correct the influence of the flowering canola in an iterative procedure. First experiments with a recursive method showed promising results, but needs to be improved to allow an automated application which is essential for the processing of the data in this project.

Another possibility is to smooth and filter the image formed by HOT values (Castleman, 1996, Chapter 11). This is not applicable since the haze in the image very inhomogeneous as seen in Figure 4.9 and a filtering or smoothing that interpolates the HOT value in order to remove the influences of the canola fields would also eliminate the variations in the cloud cover. Moreover, resulting from the size of the TM images, these algorithms are time consuming.

Another solution is to assume the quantity of flowering to be more or less constant over the complete image. In this case, the change to the HOT values originating from the canola flowering can be estimated from clear regions and added as canola offset to the HOT value over canola fields. This method would allow a fast classification and leave the small scale structure of the thin clouds untouched. But, since the assumption of constant flowering of canola over the

complete satellite image is not true in all casses, the results of this flowering adaption are not very reliable.

Nonetheless and in lack of an alternative, this offset to the HOT value over canola fields and the corrected radiances are used for the classification. An exampel of this correction is shown in the Figure 4.9 and compared to the original result of the classification. This comparison shows no improvement to the original classification for the majority of fields. Thus, the simple adjustment suggested above is not sufficient for a compensation of the flowering effects of canola and it is necessary to adapt this algorithm according to the method proposed above.

However, the haze correction for images with less strongly flowering canola works much better. Actually, the image used to demonstrate the haze correction is the same as the one used to investigate the impact of flowering on canola.

**Conclusion**

The above discussion has shown, that the HOT method gives a very good possibility to exclude the false alarm misclassification. Since these are frequent in hazy regions, this is an important progress. On the other hand, the HOT method cannot be applied easily over flowering canola fields. In order to allow an accurate classification as under clear sky coniditions, the effects of flowering on the HOT value has to be compensated for which might be possible by recursively correcting the influence of haze and canola flowering. Further development of method to correct the haze identification over canola fields would have exceeded the frame of this study.

## 4.3 Segmentation and Post Classification

The results of the pixel-based classification does not yet contain information on the belonging of pixels to individual fields. This information can be obtained by a segmentation of the results of the pixel based classification.

A segmentation merges neighbouring pixels belonging to the same class into one segment. In remote sensing of agricultural fields, these segments are mostly equivalent to individual fields. The segmentation will needed for the following purposes:

1. Remove Pixels misinterpreted as canola because of the spectral mixture of other surface types by simply removing undersized segments.

2. The application of a region growing procedure to compensate for smaller variation of the spectral signature of canola.

3. Identifying pixels adjacent to the segment and which might partly contain canola. Adjacent pixels of the segment can be identified with the segments border pixels since these might also contain canola.

4. Derive shape, size and location parameters the identified fields, e.g., the field centre, size, orientation with respect to North or border length.

5. Reduce the amount of data necessary to represent the result by a vectorisation of the segments. This is also important for a further processing in a GIS and the compilation of the acreage statistics for the complete investigated area.

### 4.3.1 Segmentation of the Pixel Based Result

#### Neighbourhood Definition

A segmentation is based on the neighbourhood of pixels. There are two main definitions of neighbouring pixels in digital images:

- The **4-neighbourhood** is defined by the direct neighbours of the pixels, i.e., the upper, left, lower and right pixel.
- The **8-neighbourhood** additionally includes the diagonal neighbours.

The 4-neighbourhood has the advantage that it is less likely to merge neighbouring fields together and thus gives a more accurate result for the field size distribution. The 8-neighbourhood merges neighbouring fields more frequently to one single segment than the 4-neighbourhood. This leads to an overestimation of larger fields. The advantage of the 8-neighbourhood is that it also allows to identify narrow fields that are oriented diagonally to the sensor's scan direction. These fields would be overlooked by using a 4-neighbourhood rule since such pixels are frequently not joined to a segment although the field is larger than the minimum field size (see below).

In this study, it is more important to identify the correct acreages and field number of canola than to identify the correct field size. Therefore, the 8-neighbourhood is used to merge the pixels to segments.

#### Border Pixels and Undersized Segments

The segments identified are used to correct some errors of the classification result. One source of error for the pixel-based classification are mixtures of radiances from different surface types within one pixel that result in a radiance similar to those of canola. These mixtures do not occur very frequently and are usually limited to single pixels. Therefore they can be removed by defining a minimum field size. Another reason for a misclassification are surfaces that have not been taken into account by the discussion on the surface type separability. This is especially important for non-agricultural surfaces since these have not been investigated here in detail because the number of the possible surfaces is to large. Anyhow, these surfaces are most likely much smaller than agricultural fields and thus can also be filtered out by a minimum field size.

As discussed in Section 2.1.1, p. 17, a typical field in Northern Germany is usually larger than 1 ha and segments smaller than 1 ha are removed from the

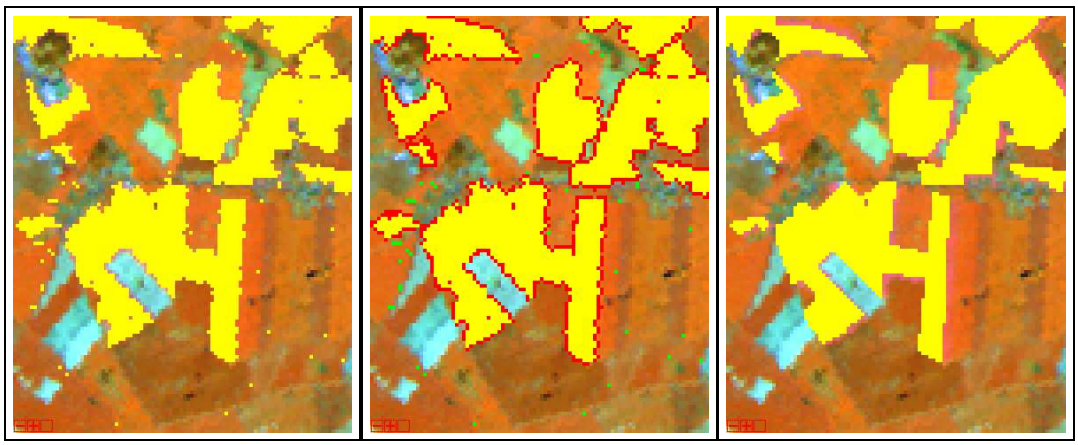


Figure 4.10: Example of the corrections applied to the classification with the aid of the segmentation. Displayed is a clip of the image 193/023; May 2, 2000. Left: Original data overlaid with the MDC result in yellow. Centre: Original data overlaid with the segmentation result (Yellow: canola pixel, red: segment border pixel, green: undersized segments ($< 1$ ha)). Right: Original data overlaid with a pixel map generated from the Quillow mapping data set. Original data: LANDSAT TM ©ESA, 2000. Distributed by Eurimage.

classification. This amounts to the number of 11 pixels for TM and 16 pixels for LISS/3.

Generally, pixels at the field border are mixed with other surface types. In order to obtain an estimation of these acreages missed by the classification, every neighbouring pixel is identified and assigned to the segment as border pixel. Since the ratio of border and field pixels decreases with growing field size, this is parameter is especially important for small fields.

The results of the segmentation and a comparison with the ground truth data is shown in Figure 4.10. Pixels masked out because of their small size (undersized segments) are marked in green and the border pixels are marked in red. The classified canola pixels, i.e., pixels of segments with sufficient size, are marked in yellow. It can be seen that the removal of the undersized segments can be justified by the field structure and the comparison with the mapped fields. The comparison with the field structure visible in the clip with the original data shows the improvement of the classification with removal of undersized fields. The inclusion of the segment border pixels also shows an improvement the matching with field edges still visible left clip.

The Table 4.6 show a comparison for the complete Quillow mapping data of the results for the MLC, the MDC and the result of the MDC, also taking the adjustments of the segmentation into account. Note that the classification results compared in this table are already processed with the region growing presented below since these results are the ones actually used for the final result of the classification. The rightmost column in Table 4.6 shows the changes of the classification result based on the removal of undersized segments. The

influence of this removal on the total acreage identified is very small with only 0.1 % improvement for non-canola surfaces in 2000 and even only 0.1 % in the other years. In contrast, the accuracy of the canola classification is reduced by up to 0.3 % in 2000. It is important, that the acreage of the non-canola fields is larger than the canola acreage, since a small relative errors for non-canola acreage results in a larger absplute error than a comparable error for the canola acreages. According to Table 4.3 it ranges from 12 to 20 % of the total acreage.

The other comparison shown in Table 4.6 is that between the MDC with and without including the adjacent pixels. As discussed above, the border pixels give an estimation of the influence of mixed pixels at the border of the fields. Here, the results for the years 1999 and 2000 show an improvement of about 3 % by simply adding the adjacent pixels to the segment. Nonetheless, for the year 2000 this results in a larger error for the non-canola surfaces of about 3 % of a much larger area. These errors result from overshown pixels (see Section 5.3.1, p. 153) which are also visible by comparing the mapped and classified data in Figure 4.10 which shows smaller mapped fields than classified fields. This errors will be discussed in detail in Section 5.3, p. 152.

### 4.3.2 Segment Based Region Growing

As visible in Figure 4.7 and Table 4.5, both pixel-based classification algorithm fail to classify some of the pixels obviously belonging to the canola fields. Some of these pixels can be identified with the adaption of the classifier to the flowering of canola, but others are still missed by the classification.

A possible improvement of the classification is suggested by the upper left clip of Figure 4.7 by the canola fields that appear as generally homogeneous region. Thus, a region growing procedure, applied to the segmentation results can yield the classification of the complete field. A region growing algorithm is an algorithm that uses information on the properties of adjacent pixel to identify homogeneous regions in an image. Usually, this is done recursively until no more pixel belonging to the segment are identified (Castleman, 1996, Chapter 18). Typically, region growing is used on the complete image. Region growing using the identified segments as starting point for the growing, allow to reduce the computational burden. This is only possible, if at least some of the pixels in the fields have been identified corretly by the pixel-based classification. The evaluation of the pixel-based classification showed, that all fields have been identified correcly in the Quillow mapping data set.

Therefore, only missed pixels neigbouring already identified segments have to be identified. This is accomplished by applying a region growing algorithm on the pixels adjacent to the segments. The decision of the belonging of a neigbouring pixel to a segment is decided by the modified MDC by simpy shifting the origin of the MDS to the mean of the segments radiances and adding pixels within this new MDC sphere[8].

[8]Actually, this is equivalent to the defition of a pixel based gradient, that includes pixels

This automatic selection of training data requires some precautions since the newly determined classifier might as well be based on misclassified pixels and cause classification errors, e.g., by the classification of a completely different surface type or by an uncontroled sprawl of the region growing. Most of these corrupted training data can be excluded by the above rule for undersized segments which are ignored for the classification. To make sure that the region growing process only starts from correctly classified canola pixels, only segments with a minimum size of 2 ha are used. Another test is applied on complete resulting segment by comparing its mean radiances to the MDC with a tighter confidence factor of $k = 3$. With these two tests the uncontroled sprawl of the region growing can be prevented. This was tested by the visual inspection of the classification result.

The MDC rule can be applied to all pixels adjacent to the segment by adding every pixels that confirms it. The following procedure is performed for each segment until the segment is not changed further by the region growing procedure.

1. Shift the MDC to the mean radiance of the segment.
2. Identify all pixels adjacent to the segment.
3. Clarify all adjacent pixels with the modified MDC and include those to the segment.
4. Identify and clarify pixels that are adjacent to the newly identified pixels.
5. Repeat with the last step until no more adjacent pixels are identified as "canola".

In order to assure that the newly selected pixels are truly canola, the mean of all added pixels is calculated and also evaluated with the MDC reduced to $k = 3$. Note that the pixels added are not used for a new training data set since this also increases the risk to sprawl of the region growing and therefore severe errors in the classification. Segments at the border of hazed regions have to be excluded since there it is not possible to prevent the region growing from sprawling.

An example for the improvement achieved by the region growing is shown in Figure 4.11, especially in the segments near the lower centre of this clip. The quantitative comparison of the region growing results with the MLC in Table 4.6 shows, that the it allows for an accuracy comparable or better than the classification with the MLC.

### 4.3.3 Vectorisation

The results of a pixel-based classification is generally a mask indicating the class membership of the pixel in question. This type of representation has

with a gradient smaller than the distance to the next agricultural class.

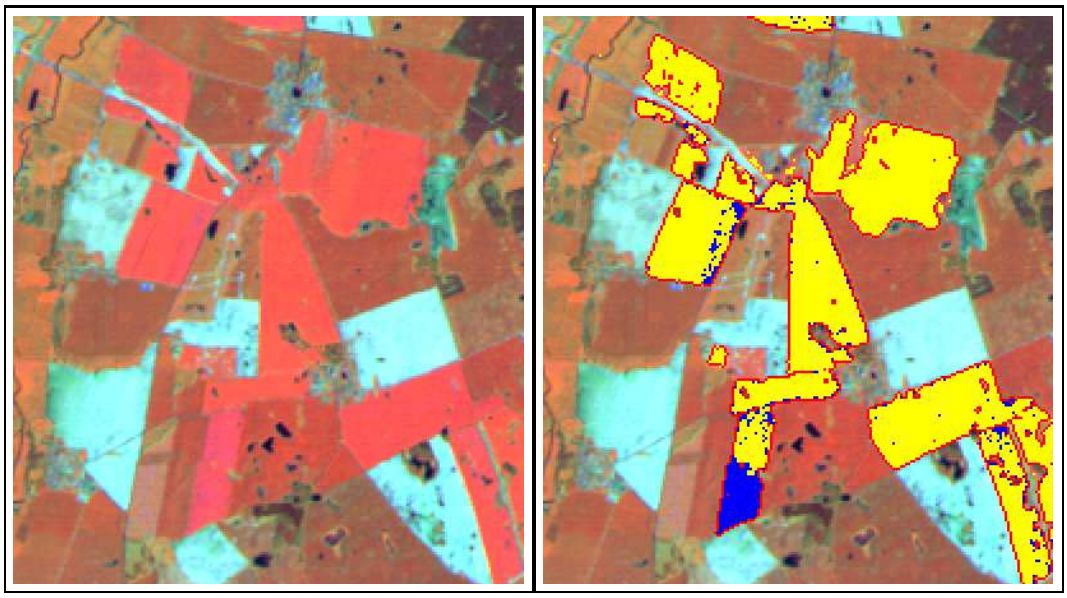


Figure 4.11: Example for the result of the region growing applied to the segments. Left: Original false colour image acquired April 30, 1999. Right: Result of the MDC (yellow pixels) and the segment-based region growing (blue pixels). The red pixels indicate the pixels adjacent to the segment. This image shows that the classification can be improved by the application of the segment based region growing. Original data: LANDSAT TM ©ESA, 1999. Distributed by Eurimage.

disadvantages for further processing of the data. A commonly used simplified representation is a vectorisation of the data, i.e., the translation of the segments to a geometric form that represents the segment. This generally allows a faster and easier processing than a pixel based raster results.

### Approximating Rectangles

Usually, fields and segments are represented as polygons formed by the pixels. In this study, the canola fields are approximated by rectangles that have the same area, orientation and long to short edge ratio. An example of this approximation is shown in Figure 4.12. A detailed description on the method is presented in Appendix A.

The representation as rectangles for the vectorisation has the following advantages:

- An rectangle can be described by only four lines which reduces the amount of memory compared to the representation as a polygon. This is especially important since the number of fields for a complete data set is quite large, e.g, the data set for 2001 contains 300,000 canola fields and a further processing would be difficult with the computers available in this study.

Table 4.6: Comparison of the MLC classification accuracy with the accuracy obtained by the MDC with the additional usage of the region growing algorithm. The column headlines indicate: MLC accuracy of the MLC classification. MDC: MDC accuracy with additional region growing. MDC case 1: all pixels identified as canola plus the segment border pixels including region growing; MDC case 2: same as case 1, but the undersized segments removed (see also Figure 4.10).

| | | Classification | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | classifier | MLC | | MDC | | MDC case 1 | | MDC case 2 | |
| | | class | canola | non-canola | canola | non-canola | canola | non-canola | canola | non-canola |
| Ground Truth | 2001 | canola | 86.28 | 13.72 | 86.45 | 13.55 | 86.25 | 13.75 | 86.04 | 13.94 |
| | | non canola | 2.57 | 97.43 | 2.43 | 97.57 | 3.66 | 96.34 | 3.65 | 96.35 |
| | 2000 | canola | 80.23 | 19.77 | 81.18 | 18.18 | 84.22 | 15.75 | 83.88 | 16.12 |
| | | non canola | 2.65 | 97.35 | 7.08 | 92.92 | 10.75 | 89.25 | 10.65 | 89.35 |
| | 1999 | canola | 65.71 | 34.71 | 75.07 | 24.93 | 78.26 | 21.24 | 78.38 | 21.62 |
| | | non canola | 1.18 | 98.82 | 1.05 | 98.95 | 2.07 | 97.93 | 2.06 | 97.94 |

- It is easier to identify corresponding fields in different images of the same region with the vector information, i.e., the centre of the rectangles and the extent.
- The simple shape of the fields allow to extract and compare further parameter like, e.g., the orientation to North or the ratio of long to short edge.
- The vectorised rectangles can easily be transfered to a GIS which is necessary for the further use of the data within the project GenEERA.

### Additional Field Information

The lower clips in Figure 4.12 shows the approximated rectangles overlaid over a false colour image from the region at the River Elbe north of the city of Hamburg. The following parameters have been extracted from the satellite data for each rectangle and can be used for further statistics and investigations:

- Field size in ha.
- The aera covered by the neighbouring pixels, i.e., the number of border pixels multiplied with the sensor resolution.
- The Position of the field in the satellite image coordinates.
- The Position, i.e., the centre coordinate of the field in UTM Zone 32 coordinates.
- The length of the two main axis of the rectangle.

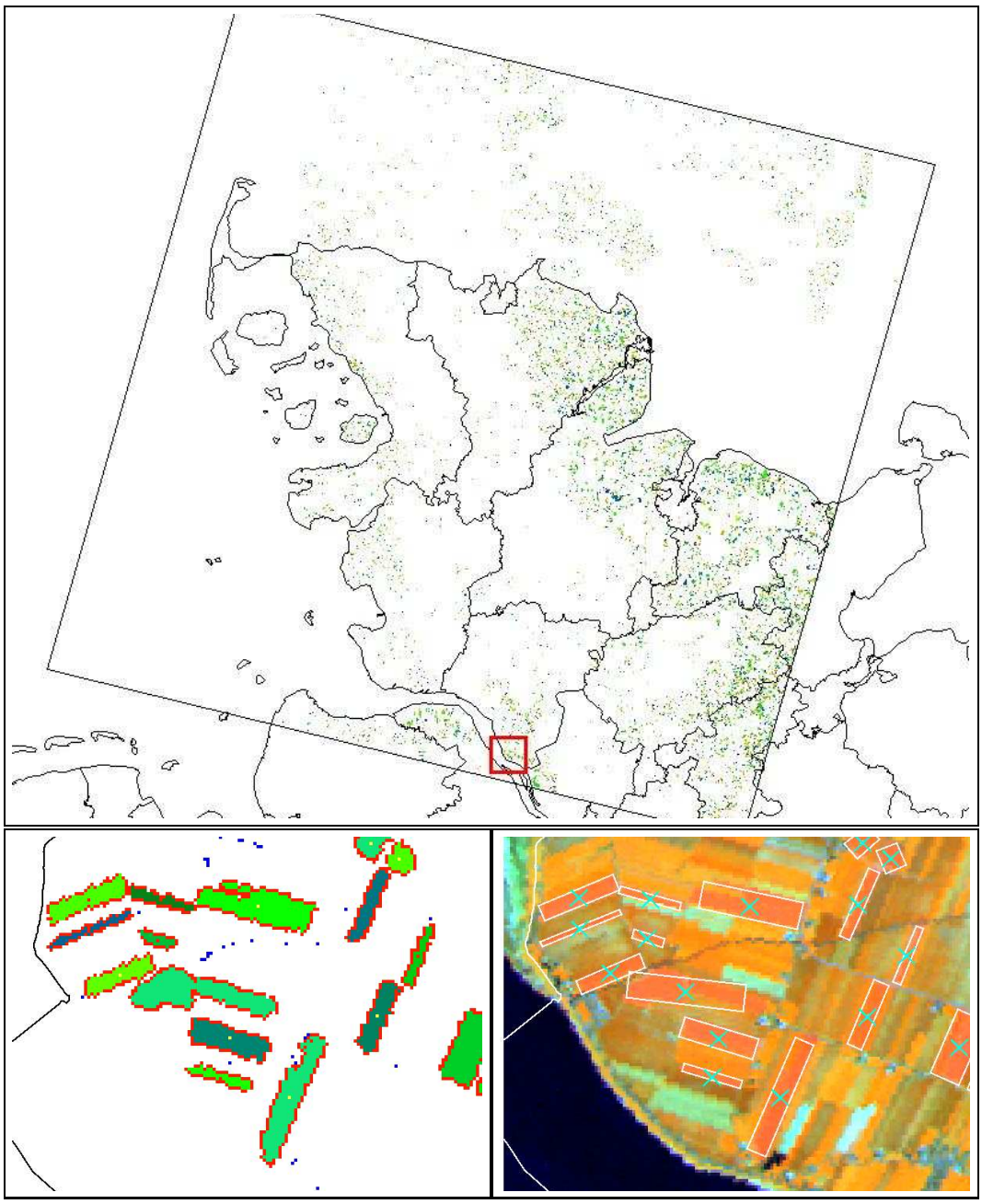


Figure 4.12: Image-based raster classification derived from the TM image 196/022; April 24, 1995. Top: Classification result for a complete TM image. The slanted rectangle marks the edges of the satellite image and the polygons represent the county borders. The red rectangle shows the position in the satellite image and on the map. Different shadings of green indicate different segments. Bottom left: Result of the raster classification. The different shades of green pixels indicate different segments, red pixels indicate the neighbouring pixels and yellow pixel the segments centre pixel. Note that the centre pixels are sometimes skipped by the resampling. Bottom right: False colour image overlaid with the result of the vectorization show by the white rectangles and the green crosses. Original data: LANDSAT TM ©ESA, 1995. Distributed by Eurimage.

- The mean and standard deviation for the radiances of the segment.
- The orientation of the long edge of the field with respect to North.
- The ratio of the field's longer to shorter edge.
- The information if a cloud or a cloud shadow is adjacent to that field, marking that the field size is probable underestimated.
- The information if the region of the field is hazed, indicating, if the information of this field is less reliable than in clear sky regions.
- The length of the borders of the field. This information is important since most transfer of genes can happen at the border of the fields.

### 4.3.4 Image Classificaton Result

Figure 4.12 shows an example of the results for the classification of one single satellite image. The top image shows the pixel based result for the complete image 196/022; April 24. The lower left clip shows an enlarged clip from this image. The various segments can be identified by their different shades of green. The centre pixels of the fields are marked by yellow pixels[9] and the neigbouring pxiels are marked in red. The blue pixels are rejected by the undersized segments rule.

The lower right clip in Figure 4.12 shows the same enlarged clip in the vectorized representation, shown are the representative rectangels and the centre of mass for each field in the clip. The frame-based vectorized data are also stored in one file per frame and look very similar to the upper image in this figure, if displayed completely.

## 4.4 Complete Dataset Classification

The presented algorithms for the classification of canola and clouds require individual training data sets for each satellite image and since single images of TM or LISS/3 which only cover a small part of the investigation area, the classification of the complete investigation area requires mulitple training data sets. These training data sets have to be extracted from identical fields in neighbouring satellite images. This is especially important since the images were acquired under different atmospheric conditions and at different growth stages.

Moreover, the results of the classification and the cloud detection for the single images have to be compiled into results for the complete year. Two different results are desired for the identification of the canola acreage: A vector data set containing all identified segments with the additional information

[9]Note that some of these centre pixels have been left missed by the nearest neigbour resampling, which is also the case for some neighbouring pixels.

available for these fields (last section) in a GIS compatible data format. This compilation of the vector data is necessary to allow further processing by the partners in the project GenEERA. Since the overlap of the images allows to select the image, where the classification result provides the most reliable classification results, e.g., least cloud or haze cover.

This data set produced from the satellite image classification is quite large, e.g., for 2001 it contains about 300,000 segments. Such a large data set is not very practical if statistics for the complete area are needed, mainly because of the large amount of computer memory necessary to store the information on all fields. Therefore, as a second procedure, the obtained field information is used to construct statistical results with a much lower spatial resolution.

### 4.4.1 Collection of Training Data Sets

Training data have so far only been identified for the three images covering the Quillow mapping data set. The obtained training data sets cannot be used to classify other TM images or LISS/3 image since the appearance of canola in the satellite image depends on the growth stage and health state of the plants. Therefore, it is necessary to determine training data for each of the 50 images available. However, the amount of ground truth data that allows to identify the training data directly is limited.

The best training data is available for the frame 193/023 in the east of Germany for the images of the years 1999 to 2001, since it allows to identify canola fields from ground truth information. A similar ground truth data set is only available in the region of Bremen covered by the frames 195/023 and 196/023. Unfortunately, this data set is only available for 2001 and 2002 and does not cover such a large area.

Another ground truth data set is available for the experimental farm in Sickte near Braunschweig covered by the frame 195/024. This data set covers a larger time period (1996 to 2001) but only consists of about three to five relatively small canola fields present per year. This is too small to obtain a sufficient number of training pixels.

The conclusion of this inventory is that only the minority of images allow to select training data sets with ground truth information. Consequently, another way has to be found to identify canola fields to train the classification algorithm.

#### Neighbouring Satellite Images

Figure 2.6 (p. 28) shows the coverage of the available TM images. They generally overlap, especially in the years 1998, 1999, 2000 and 2001. The overlap in combination with the ground truth data available for the images of the years 2001, 2000 and 1999 for frame 193/023 will be used to extend the training data for the images to images where no ground truth data is available.

This is accomplished by previously classifying and segmenting the images

from frame 193/023 as discussed above. The pixel-accurate image-to-image registration allows to identify training data in the neighbouring image.

### Automated Selection from Overlapping Images

The selection of training data cannot be performed by simply selecting corresponding pixels for two reasons. First, the image-to-image registration still has an error of about one to two pixels (see Section 3.1.4, p. 55). Therefore, the edge pixels of the fields are removed in order to identify only canola pixels. This is also important since canola fields have a tendency to appear larger than they are in nature resulting from the bright colour that is overshining[10] neighbouring pixel of darker surfaces types (see Section 5.3.1, p. 153). Therefore three layers of edge pixels are removed from the segment by applying an erosion algorithm (Castleman, 1996). If more than 10 adjacent pixels ramain, they are used to identify training data in the overlapped image.

The other reason is that the crop on the observed field may have changed in the time between the acquisition of the satellite images. This can happen if the canola plants on that field are damaged, e.g., by drought or storms, and the farmer plough these fields. Although this does not happen frequently it has to be taken into account. As addtitional difficultiy the segments do not necessarily correspond to fields and only part of the segment may have changed. Therefore, it is necessary to find a criterion to identify fields where the crop has completely or partly changed.

The criterion is based on simple statistics by calculating the mean radiances and standard deviations for each segment. Radiance and standard deviation are avreaged and segments with a mean radiance that differs more than three times the averaged standard deviation are rejected.

This method allows to identify training data indirectly for images that overlap each other and could be applied for the images of the years 1999, 2000 and 2001. However, in the years 1995 to 1998 there was no ground truth based traing data available and the training data has to be colected in a different way.

### Manual Selection of Training Data

In the years 1995 to 1998 and 2002, the situation is less promising. The mapping information available from the experimental farm in Sickte (see Section 2.2.2, p. 34), with three to five fields is too small to select a representative training data set.

In order to allow a classification of the images for the years 1995 to 1998, it is necessary to identify canola fields by visual inspection. This is possible

[10]The sensor's point spread function is influenced by light reflected from areas in the neighbourhood of the pixel. Bright areas, like flowering canola fields, thus appear larger in regions of surrounding darker pixels.

because of the distinct appearance of canola fields in satellite images during flowering and the experience gained in the years with mapping data available.

The most unique appearance of canola is found during flowering: a pink to purple colour in the false colour representation or a greenish yellow in the true colour representation. Thus one image acquired during the flowering period is selected for each year. Note that two images were necessary for 1995 and 1996 since there is a missing overlap between groups of images (see Figure 2.6, p. 28).

In these images 10 to 30 fields (the number is larger in the west of the investigation area since the fields are smaller) are selected and pixels from homogeneous regions within these fields (see Section 4.1.2, p. 89) are used as training data for the MDC.

The classification result is used in the same manner as the one obtained with the ground truth data to identify more training data in adjacent images.

### Flowering Canola

As discussed in Section 4.2.2 (p. 108), the flowering of canola needs also to be obtained by visual inspection. Although this has to be done manually, it is simpler than the identification of the non-flowering canola, since flowering canola can easily be identified by its colour in the true colour image.

### The TM Image Acquired in 2002

These data present two specific difficulties. First, this frame was acquired very early on the 3rd April, 2002 and there is no image available for comparison with this early growth stage. Second, the grond truth data used for the selection of training data in this image does not contain information on the field edges as the Quillow data set. Mapped are information on habitat type at a single postion. Since the habitat types "canola field" and "adjacent to canola field" have been mapped, one of the fields adjacent or surrounding the postion of these habitat types has to be a canola field. An example of this selection is shown in Figure 4.13.

Therefore, the selection of training data for this image have to be a combination of the selection strategies for the ground truth and the visual inspection methods. The canola fields are identified by choosing the one field near the mapping position, indicated by the yellow flags in Figure 4.13), that seems most likely to be canola by comparing its colour to other fields adjacent to canola habitats.

The fields identified with this method were then used as training data for the classification algorithm.

### Conclusion

Using image overlap and visual inspection allow to identify training data sets for each image although only limited ground truth data was available. More-

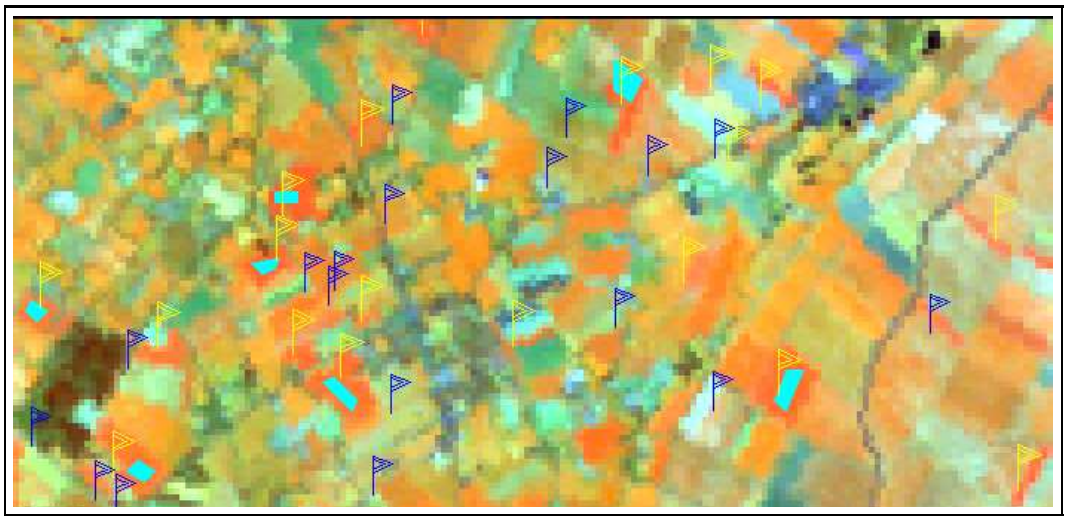


Figure 4.13: Example for the training data set selection based on the single point based habitat mapping data set from Bremen in 2002 (see Section 2.2.2, p. 34). Displayed is a clip from image 196/023; April 3, 2002 from the south of Bremen. Yellow flags: canola field habitat type. Blue flags: Other habitat type, e.g. street border or field of other crops. The turquoise polygons indicate pixels acquired as training data set for the classification algorithm with the aid of the habitat mapping. Original data: LANDSAT ETM+ ©ESA, 2002. Distributed by Eurimage.

over, the manual selection of training data sets could be limited to one image per year since the selection of training data with the overlapping image method works autonomously after a starting image is classified. An exception are the years 1996 and 1997, in which satellite data with no overlap to the other images was present.

The selection of training data for the year 2002 is not very reliable since it had to be guessed which of the fields are actually canola since the field edges were not mapped. Moreover, it is not possibility to compare the appearance of canola in that early growth stage to the data from other images since no image from a similarly early date available.

### 4.4.2 Compilation of the Classification Results

The results of the classification of canola are the basis for a further investigation concerning the biology of pollen transfer. Therefore, the results of the georectification, the cloud-mask and the classification have to be combined.

There are two different types of data sets available, the result for the individual images (image-based results) and the results for the entire investigation area for a specific year (year-based results) .

### Image-Based Results

The image-based results are of interest for institutes that perform investigations on the regional agricultural and botanical situation and allow to compare the classification result with maps like, e.g., habitats for breeding partners. Therefore, the location of each canola pixel is of interest and the raster data have to be converted to the vector information described above. The raster data are converted to georectified grey-level raster graphics (GeoTiff). In these raster graphics the classification including border pixels and undersized fields are stored. The border pixels and the undersized fields can be identified by their pixel value and equal values for pixels indicate their belonging to the same segment.

In a second file, a geocoded grey-level raster graphics is provided for the cloud cover mask, indicating clouded, shaded and hazed pixel by different grey-levels.

Moreover, the vectorised data, i.e., the representative rectangles are also provided in a standard geographical Vector format (ESRI$^{TM}$ArcView shapefile). This shapefile contains the vector information listed in Section 4.3.3 for each segment identified for the image.

### Year-based Results

The evaluation of the classification results for one whole year is difficult when using the image based results. The main reason is the thata fields in overlapping regions appears twice. The representation of the year-based results needs to fulfil the following requirements:

1. All identified segments of one year are present as representative rectangles.

2. The area evaluated with the satellite data is noted.

3. Individual fields only appear once in the data set.

Most of these three requirements are fulfilled by extracting the rectangles from the image-based results and writing them to the new data set for the entire investigation area. The exceptions are the regions where to images overlap. In these regions, the cloud/cloud-shadow identification is taken into account to select the more reliable classification result. The cloud cover mask has to be evaluated in order to select the most appropriate segment. This is obtained by dividing the overlapping area into rectangles of $5 \times 5\text{km}^2$ based on the geographic coordinates and determining the area covered by cloud, cloud-shadows or haze. Note that fields from haze covered regions are only used if the area in the other image is cloudy.

The comparison of the cloud cover for each rectangle is used to select the segment to be added to the year based data set. Also taken into account is the coverage of the satellite data since fields touching the border in one image

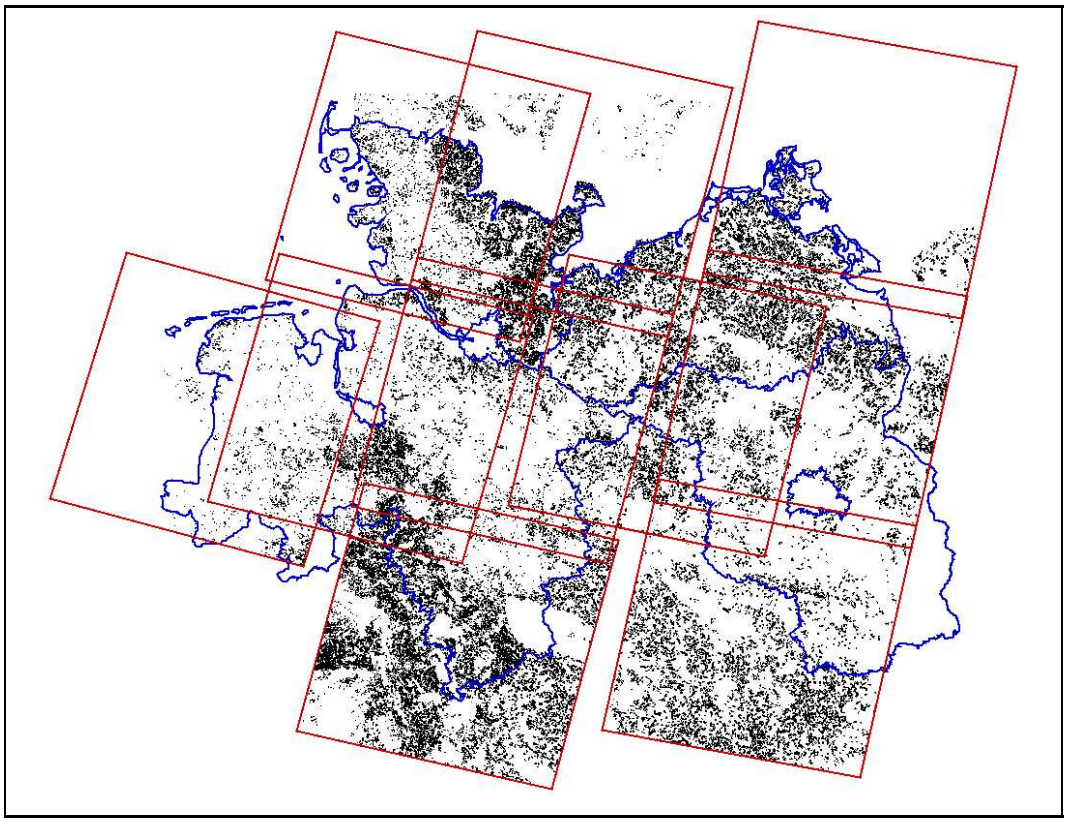


Figure 4.14: Example for the vectorized classification result for the year 2001. Displayed are the frames used for the classification (red lines), the boundaries of the federal states (blue lines) and every field found (black dots). A zoom would reveal the same representation as for the image-based vector results shown in Figure 4.12.

might be better represented in another one. Another error, although it seldom happens, for this method is that a field might be added twice if the field centres are located on the edge of the rectangles. Therefore, each rectangle is tested for its centre being located within another rectangle.

Besides the information on the identified canola fields it is necessary to provide the coverage of the satellite data resulting from frame coverage and cloud-cover. The representation of the frame coverage can simply be obtained by identifying the frame borders and writing them to another shapefile (see also Figure 2.6, p. 28).

The cloud coverage is more difficult to describe since it is mainly available as a pixel-based result. Therefore, the above described $5 \times 5\,\mathrm{km}^2$ rectangles are used to represent the cloud coverage by building polygons enclosing the rectangles with a cloud cover of more then 20 %.

With this data set the results can be processed with a usual GIS software providing all required information in a data set of reduced size and ambigiouties clarified. Figure 4.14 gives an impression on this data set for 2001. The size of this figure is not adequate to identify details, but this file contains exactly

the same information for each field as the frame-based vector files shown in Figure 4.12. An example for the coverage information is shown in Figure 5.2 (p. 136).

### Compiling Acreage Statistics for the Investigation Area

The above described data needs much storage space, e.g, 1 GB for the year 2001. However, a simple statistic of the field size distribution or total acreage is sufficient. Therefore, statistics for a $5 \times 5\,\mathrm{km}^2$ grid are extracted from the classification result. The following results are generated:

- Total acreage of canola, for the vectorised result and for the raster result (including the undersized fields).
- Fraction of canola fields compared to the complete evaluated area.
- Evaluable area, i.e., fraction of the rectangle that is covered by the satellite data and not covered by clouds.
- Cloud-/cloud-shadow-cover fraction.
- Mean field size.
- Mean orientation with respect to the North.
- Mean ratio of long and small rectangle sides.

Some of these statistics will be presented in the following chapter and all results are available on DVD (see Appendix C).

## 4.5 Conclusion

A classification algorithm has been presented that can identify the fields of canola larger than 1 ha with an accuracy of about 80 %. The accuracy is achieved by applying an additional region growing. Moreover, the reflectance properties of flowering canola had to be acounted for by an additional classification. The adaption of the flowering canola could not be automised and still requires the manual selection of training data sets. Nonetheless, the classification with the automatically selected training data sets simplyfies the selection of these flowering canola training data sets.

A correction for thin cloud cover is presented which allows to classify canola pixels below it to some extent and prevents the confusion of other surface types with canola .

From the results of the classification segments of neighbouring canola pixels are constructed. They are used to calculate representative rectangles to produce a vector based data set.

Since the objective of this study is the evaluation of the canola cultivation in entire Northern Germany and the training data sets are only available for a

few images, a method has been developed, that allows to extract training data sets from overlapping images with the help of the classification result for one of the images.

The results obtained for the different satellite images are compiled to data sets for the complete investigation area. One data set contains information on various parameters, e.g., position and size for all individual identified fields.

All methods described in this chapter are developed under the requirement of a mostly automated data processing to reduce the effort necessary by a human operator. This has been achieved for most of the processing steps.

# Chapter 5

# Results and Validation

The methods described in the previous two chapters have been applied to all 48 satellite images. In the first section of this chapter the results of the classification are presented. The focus will be the statistical description of the classification results. In Section 5.2 the validation of the classification results is carried out with positions of known canola fields and a comparison with cultivation statistics from 1995 and 1999. Section 5.3 discusses the error sources for the identified field sizes and the likelyhood of overlooking small fields. The chapter is closed with a final appraisal of the results from this study.

## 5.1 Classification Results

As presented in the last chapter, the results are either based on a single image or are compiled for the complete investigation area. The results are provided in two different forms. A detailed form, including all fields with additional information for each of them, and a compiled form that totals or averages the parameters for all fields within $5 \times 5\,\mathrm{km}^2$ squares of a grid overlaid over the investigation area.

### 5.1.1 Detailed Field Information

The per-field information and its structure has already been presented in Section 4.4.2, p. 128. These data are useful for the investigation of the local cultivation characteristics of canola. Moreover, they contain detailed information on the cloud situation in which the classification has been performed, which allows the estimation of the local reliability of the classification result.

The image-based results consist of the individual classification results for each satellite image. There are 48 such data sets available, one for each satellite image (see Figure 2.6, p. 28). Additionally, a set of all fields identified is compiled into one vector data set for each year. This substantial amount of data is too large to be displayed in detail in this thesis, e.g., for the year 2001 more than 300,000 canola fields have been identified.

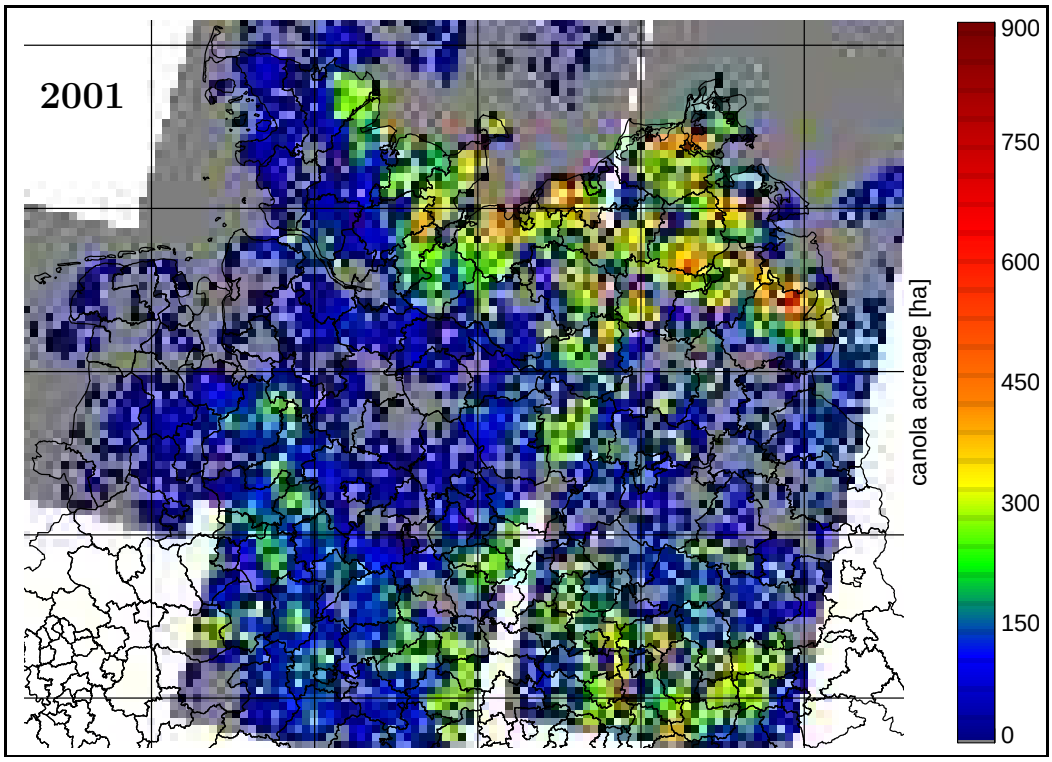



Figure 5.1: Map of the total canola acreage for 2001 represented as a $5 \times 5\,\mathrm{km}^2$ raster. The black grid lines represent the 105 km lines of the UTM Zone 32 grid and the irregular lines the county borders. Regions with no satellite data available are displayed in white. The main cultivation areas of canola in the northeast of Germany (upper right of the map), in the east of Schleswig-Holstein (upper centre) and south of Berlin (lower right) are clearly visible.

Although these detailed data sets are one of the main outcomes of this study, they are only intermediate results and have to be analysed further in order to estimate the gene transfer from genetically modified canola to non-modified canola or other breeding partners of canola on a small scale in various regions of the investigation area. This is not in the scope of this thesis and will not be discussed here. The cultivation characteristics for the complete investigation are discussed here instead.

### 5.1.2 Averaged Results

The averaged results contain the averaged or total of the parameters listed in Section 4.4.2 (p. 131). Some of the parameters will be presented here and their importance for the risk assessment of gene transfer will be discussed exemplarily.

### Canola Acreage in Northern Germany

The acreage of canola for the year 2001 is shown in Figure 5.1. Clearly visible are the main cultivation areas in northeastern Germany, eastern Schleswig-Holstein and south of Berlin. The main cultivation areas are already known from the agricultural statistics and show a good agreement. In these regions, the canola acreage is larger than 500 ha, i.e., 20 % of the raster square area of 2500 ha ($5 \times 5\,\mathrm{km}^2$) is covered with canola. The largest canola acreage was found in the northeast of Germany with an acreage of 1050 ha, which corresponds to 42 %.

The regions with little canola cultivation are mostly located in northwest Germany and west of Schleswig-Holstein in the coastal areas near the North Sea.

### Canola Cultivation Density

The total acreage is only meaningful in raster squares completely covered with cloud free satellite data. In raster squares at the border of a satellite image or in squares with cloud cover, it is likely that not all fields in the square are identified. Nonetheless, a relative coverage can be obtained by identifying the observable area, i.e., the area in without cloud cover and the satellite data available. The upper map in Figure 5.2 shows the observable area for the year 2000. The coverage of the investigation area was nearly complete, with the exception of significant cloud cover in the eastern regions. The cultivation density is the ratio of the identified canola acreage and the observable area and is shown for the year 2000 in the lower diagram of Figure 5.2.

The comparison of this diagram with the total cultivation area from 2001 in Figure 5.1 shows a good agreement for the main cultivation areas. Moreover, the colour scale of the total acreage and the cultivation density is equivalent for cloud free regions and a comparison with the amount of canola identified in 2000 and 2001 is very similar. This suggests that neither the main cultivation of canola changed within the years nor the classification in 2000 and 2001 changed noticeably.

The cultivation density for the years 1995 to 1999 is displayed in Figure 5.3 and shows good agreement with those 2000 in Figure 5.2 (bottom). The results for the investigation period showed that the main cultivation area also did not change noticeably within this longer period.

The cultivation density for 2002 is also shown in Figure 5.3 and allows to identify the cultivation area south of Bremen. Nonetheless, it misses the canola cultivation near the river Elbe. This is caused by strong cloud and haze cover in that part of the image. Moreover, the early acquisition date makes the identification less reliable (see also Section 5.2.1, p. 144).

Pollen and seed dispersal is more likely in regions with high canola cultivation density simply because of the larger number of canola plants, i.e., in the main cultivation areas mentioned above.

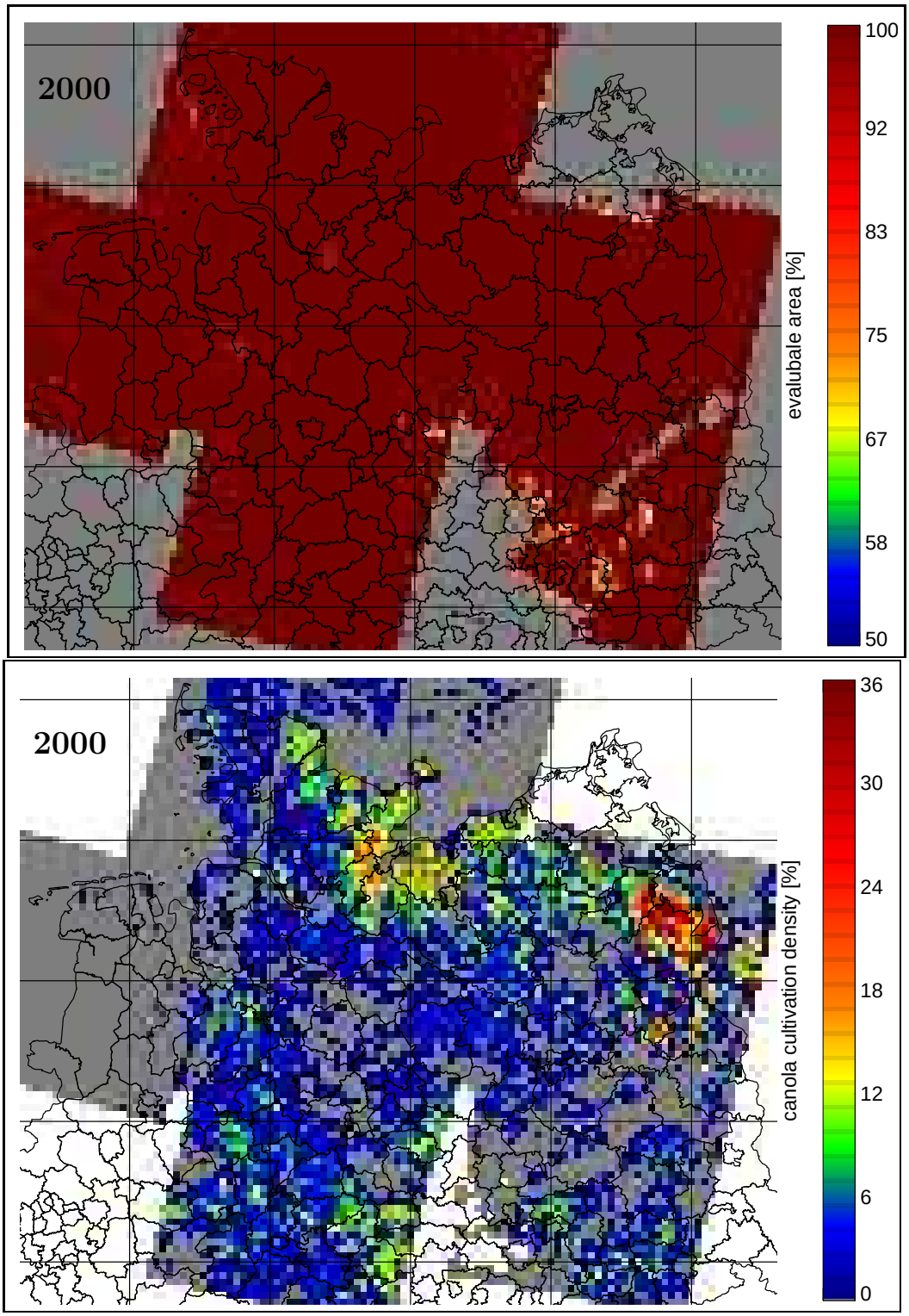



Figure 5.2: Top: Observable area of the investigation area for the year 2000. Grey pixels indicate a coverage below 50 %. Bottom: Cultivation density for the complete investigation area for the year 2000. White pixels are those with an observable area of zero. In cloud free areas, the largest displayed cultivation density of 36 % corresponds to 900 ha total acreage.

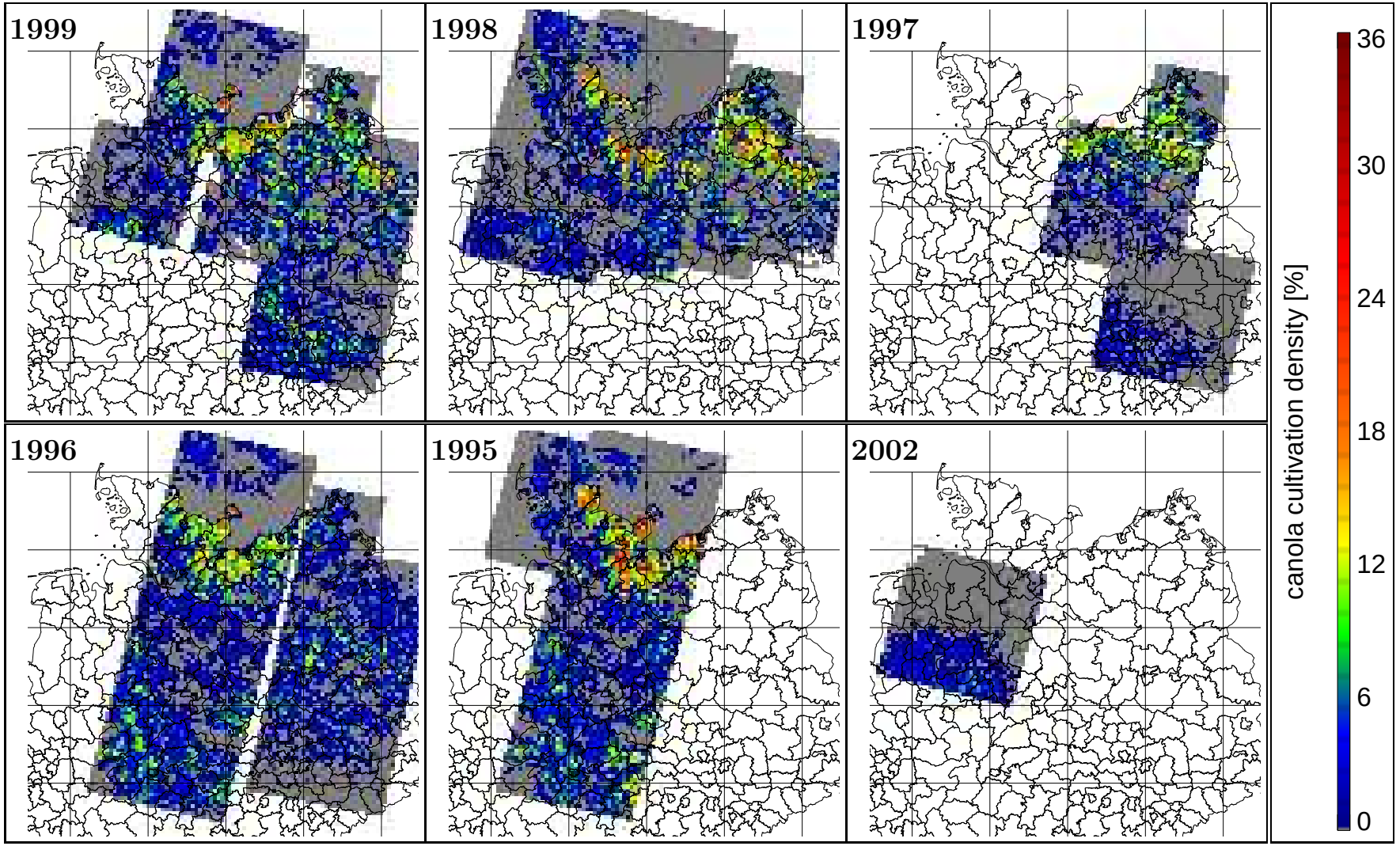



Figure 5.3: Canola cultivation density for the complete investigation area for the years 1995 to 1999 and 2002, same representation as in Figure 5.2 (bottom). The main cultivation areas can be identified in most years.

### Field Size Distribution

The size of canola fields is another important parameter, since smaller fields share a longer edge with the surrounding non-canola vegetation for the same total cultivation area than larger ones. Consequently, a larger number of canola plants is close to potential habitats of interbreeding partners.

Moreover, a smaller field size often indicates a larger number of farms and thus a higher probability of the cultivation of different breeds of canola, including GM canola in the future.

The mean field size is obtained by separately averanging the size of canola fields in each of the $5 \times 5\,\mathrm{km}^2$ raster squares. Figure 5.4 shows this parameter obtained from the classification result for the year 2001. It clearly shows the differences in field size distribution between the western and eastern federal states, which is the result of different historical developments of the two regions. The range of field sizes found is about 2 to 20 ha in the western regions. The large fields in the northeast of Germany have a size ranging from 30 to 50 ha. The largest field found in that area was 200 ha and is located in Brandenburg. Additionally, it can be seen that the mean field sizes in the main cultivation areas in west Germany are also larger than in areas with less canola cultivation.

Fields in eastern Germany are generally larger than fields in western Germany and fields outside the main cultivation areas are also smaller than those within. Consequently, the potential of gene transfer from field to habitats surrounding the field is more likely in the western regions. Additionally, the number of different breeds in these regions is potentially higher since there are more fields and therefore the probability of a GM canola field is higher.

### Ratio Border/Field Pixels

The length of the contact zone between the field and the surrounding vegetation depends not only on the size of the field but also on its shape, i.e., an irregularly shaped field generally has a longer contact length than a rectangular one[1]. Therefore, the lower diagram in Figure 5.4 shows the distribution of the number ratio of border and field pixels[2]. A comparison of this ratio and the field size distribution allows to identify regions with irregularly shaped fields, i.e., regions with a high border to field ratio and a large mean field size, which is very likely resulting from the fractal shape of the fields in such regions. Most obvious are regions in the northeast of the investigation area, northern Brandenburg and Mecklenburg-Vorpommern. The region in the southeast (Sachsen-Anhalt) shows much lower values for this ratio, which means that

[1] A irregularly circular field has a longer contact length than a narrow rectangular one of the same size. However, the approximation by rectangles conserves the ratio of long to short field extent.

[2] This parameter is only applicable to classification results from years consisting only of TM data, since LISS/3 has a different spatial resolution and the results are not directly comparable.

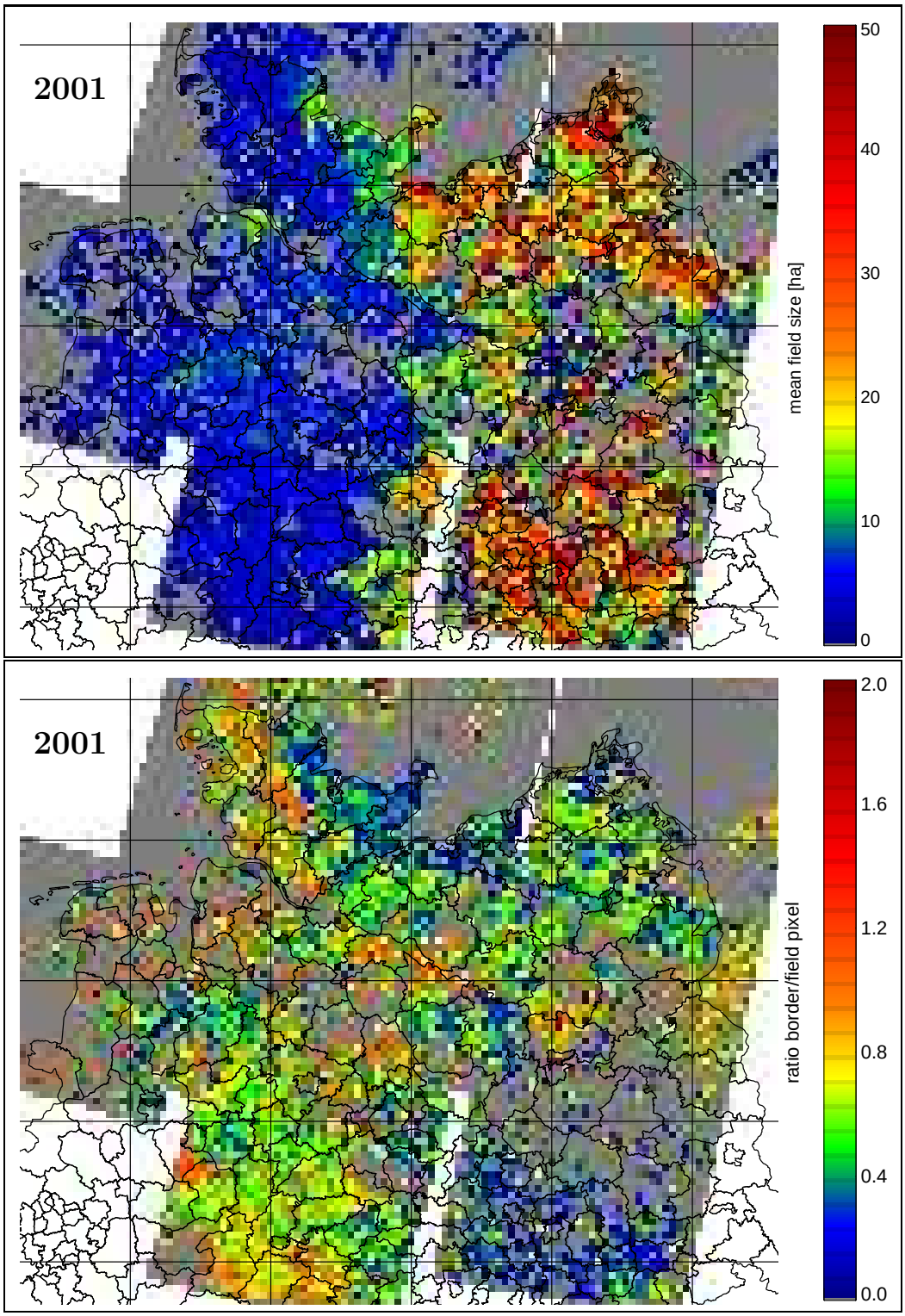



Figure 5.4: Top: Mean field size for 2001 averaged over $5 \times 5\,\mathrm{km}^2$ squares. Bottom: Number ratio of border and field pixels.

it mainly consists of fields with more regular shapes. The high values in the western parts result from the smaller field sizes.

Besides the importance for the contact length, this value also allows to give an estimation of the classification quality. Areas with high values have a large number of neighbouring pixels and therefore misclassifications are more likely (see Section 5.3, p. 152).

### Border Pixel Area

An absolute estimation at the contact length of the fields can be given by the area of covered by border pixels. The upper diagram in Figure 5.5 shows the total area of pixels neighbouring canola pixels.

As discussed in Section 4.3.1 (p. 117) neighbouring pixels are important to estimate the quality of the classification. Additionally, this parameter can be used to estimate the area of the pixels that contain canola as well as other vegetation. Therefore, the lower diagram in Figure 5.5 shows the total area of border pixels for each square. Although this parameter is simply based on the satellite's spatial resolution, it allows to give a quantitative estimation of the direct contact area. As expected, the largest values are obtained in the main cultivation areas of northeastern Germany. Remarkable is the large number of border pixels in the areas south of Bremen which is comparable to the number of border pixels in the main cultivation areas although the amount of grown canola is lower.

The potential gene transfer from canola fields to surrounding vegetation would be high in the main cultivation areas in northeast Germany, northeast Schleswig-Holstein but also in the region south of Bremen.

Note that for an exact assessment of pollen transfer to the surrounding vegetation, detailed assumptions on the pollen transport should be included. Generally, it can be stated that the importance of the field shape is decreasing with increasing pollen transport range.

### Length of Contact Line

Under the assumption of regular fields, a more meaningful parameter can be determined from the approximated rectangle: The "length of the contact line" is the total length of all field borders. It gives the length of the line where canola is in contact with other plants. An approximation can be obtained from the representative rectangles by adding the length of their borders. The total of these borders for all fields is shown in the lower map in Figure 5.5. This shows quite similar results to the border pixel area. High values of up to 5 km are found in the main cultivation areas. Unlike the border pixel area, the length of cantact line in the southeast is as high as in the northeast of the investigation area. This results from the approximation as rectangles which ignores the longer border length of irregularly shaped fields (common in the northeast). The contact line is much shorter in the west of Germany than in the East.

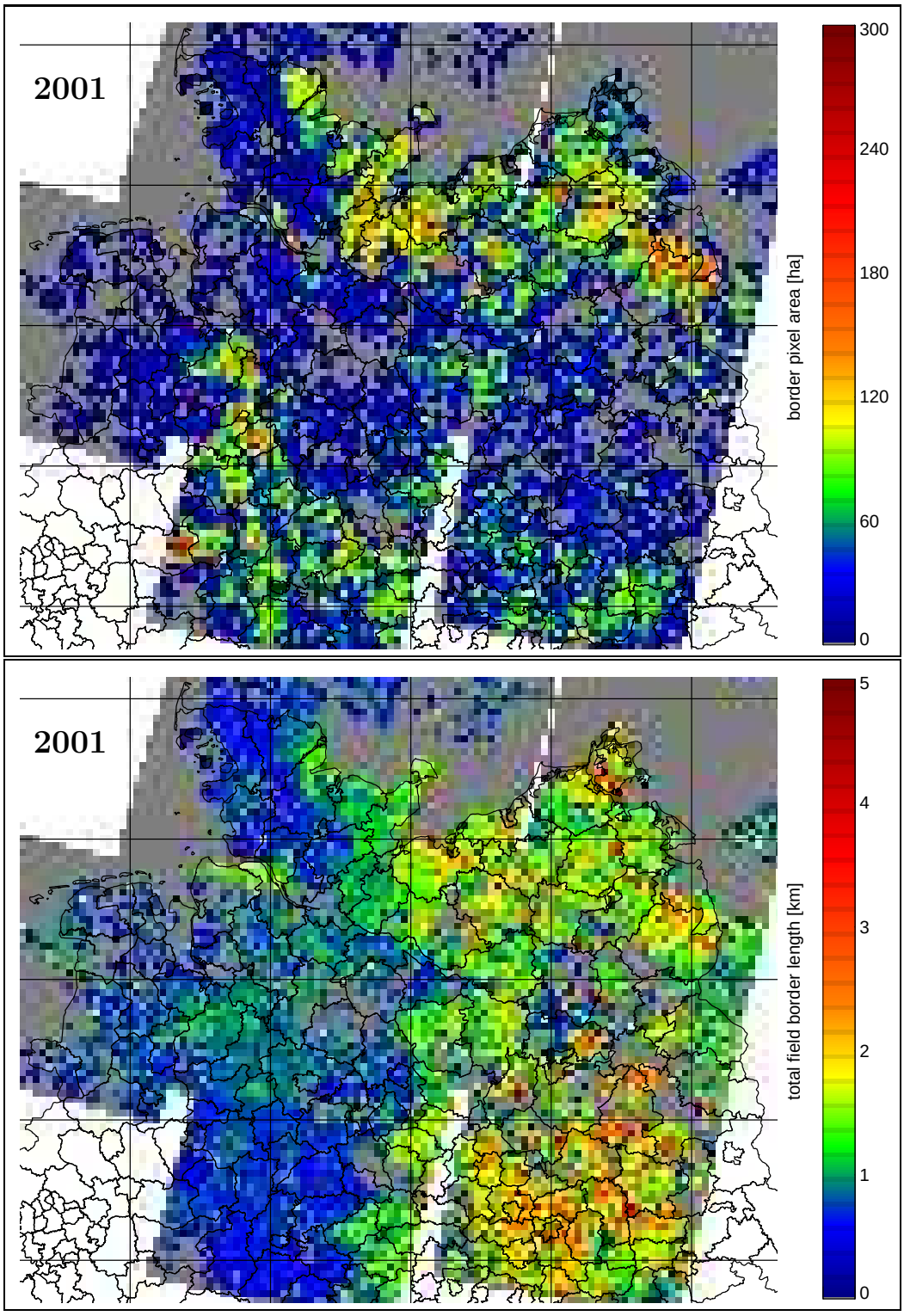



Figure 5.5: Top: The total area of the border pixels, i.e., pixels that probably contain canola as well as other vegetation. Bottom: Approximated length of the contact lines for all canola fields, estimated from the edges of the approximated rectangles.

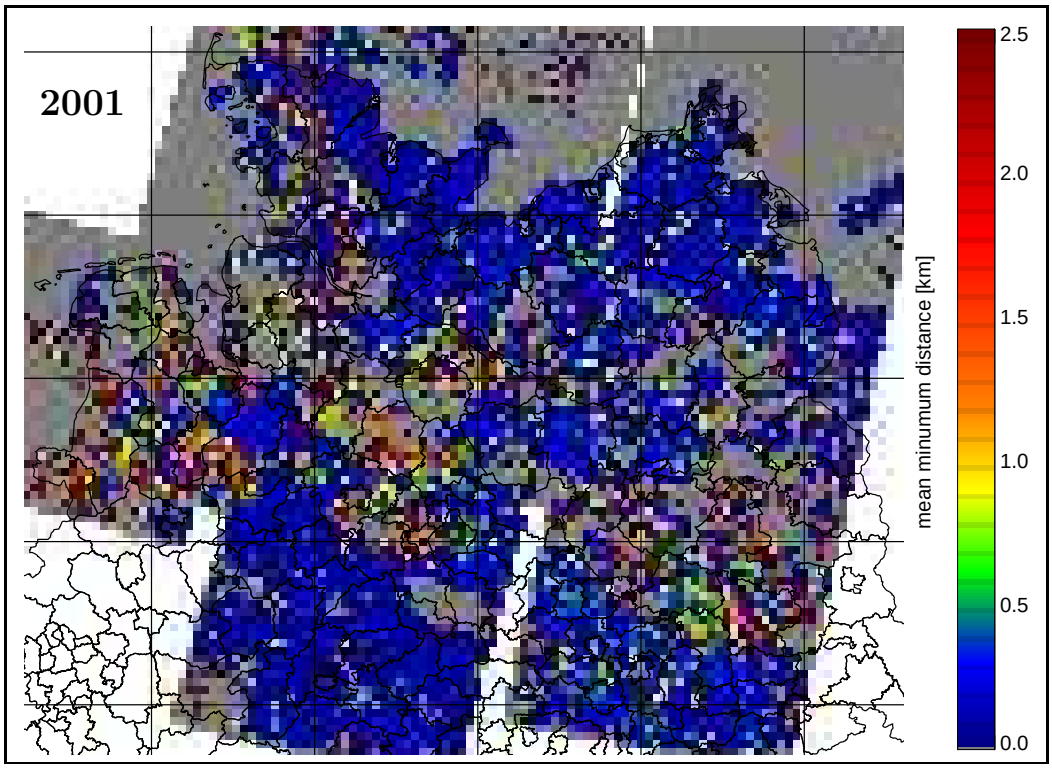



Figure 5.6: Mean minimum distance between canola fields on the example of the result for 2001. Grey pixels included less than two canola fields and therefore did not allow to calculate a mean distance.

According to this parameter, hybridisation would be much more likely in the east of the investigation area.

### Distances Between Field Borders

The most important parameter to estimate the potential breeding of non-modified and GM canola in connection with the pollen transport range is the distance between fields. Obviously, the nearest fields are of interest and therefore, the shortest distance to the next canola field has been calculated and averaged for all fields. The result is shown in Figure 5.6 and indicates that the mean distance is usually below 500 m in the main cultivation areas. Because of the limitation to the raster grid of $5 \times 5\,\mathrm{km}^2$ the maximum distance displayed has been limited to 2.5 km. Nonetheless, the grey regions can be taken as an indicator for regions of distances above 2.5 km. Considering the transport range of up to 4.5 km, the transport of pollen from field to field is very likely in the main cultivation area of northern Germany.

**Conclusion**

The two potential risks linked with GM canola cultivation are hybridisation and breeding with non-modified canola, also called contamination. Considering the above parameters presented, the following qualitative assessment of these risks can be made for northern Germany:

**Contamination:** The contamination can be expected to be very likely in the main cultivation areas in Schleswig-Holstein, Mecklenburg-Vorpommern and Brandenburg because of the low field distance of only 0.5 km which is below the expected pollen transport range (Breckling et al., 2003; Rieger et al., 2002). The short distance between fields can also be found in the areas with smaller canola acreage like the region south of Bremen or west of Hanover. In this context, the small field sizes in regions with high canola acreage can also indicate an increased risk because of the possibly increased number of different canola breeds.

**Hybridisation:** The hybridisation mainly depends on the neighbourhood of canola with other types of vegetation. This can be best estimated with the total field border length. This parameter shows values of up to 5 km for eastern Germany, but also values of up to 2 km in the cultivation areas of western Germany. Because of the approximation of the field border length which ignores the regularity of the field shapes, the number of border pixels can also be used to estimate the contact length to the surrounding vegetation. Surprisingly, this parameter showed comparably high values for northeastern Germany but also eastern Schleswig-Holstein and the region south of Bremen. The importance of these two parameters depends on the assumed pollen transport range[3] and the distance to the habitats of interbreeding partners. The pixel border area represents the short distanc transport more successfully since it accounts for variations in the shape of the field, which is less important for higher distances.

## 5.2 Validation

The estimation of the quality of the information gained on the cultivation of canola requires a comparison with ground truth data. As described in Section 2.2.2 (p. 33), some sets of validation data are available for this study. The Quillow mapping data set has already been evaluated in the description of the MDC. This comparison showed an accuracy of 78 to 86 % identified canola acreage and an identification of 100 % of the present canola fields by means of identifying the presence of a field. Nonetheless, this validation is not representative for the complete classification, since:

- The training data set has been derived from the region classified and can not describe the variations potentially present for larger distances from

[3]The correct pollen transport range is still been discussed (Rieger et al., 2002).

the training data. The reflectance variations of canola flowering is one example of such effects.

- The fields in eastern Germany are much larger than in the West. Therefore, the number of fields in the west completely missed by the classification is potentially higher than in the region of the Quillow data set.

- The training data for the fields located in images without ground truth data available might be corrupted by the automatic training data selection (see Section 4.4.1, p. 126).

Two different types of validation are applied. The first one is based on the position and, if available, the edges of known canola fields in other regions than the Quillow region. The second is the comparison of the identified total acreage with the agricultural statistics available for Schleswig-Holstein and entire Germany (see Section 2.2.3, p. 36).

### 5.2.1 Known Field Positions

The comparison with known field positions allows to validate the local results of the classification. This has the advantage that the reasons for misclassification, e.g., small field size or strong flowering, can be directly investigated. Moreover, a classification error can not compensated and thus be masked by other types of misclassifications, which is possible if only the total of canola acreage is compared as in the comparison with the agricultural statistics.

For a small number of fields, information on all corner coordinates, i.e., the field sizes, was available and for a larger number only one coordinate and the information if a canola field is present or not. While the first type of validation data only contains very few fields, the second validation type contains a much larger number of samples and is therefore presented first.

#### Mapping of Interbreeding Partners

The mapping of interbreeding partners described in Section 2.2.2 (p. 34) gives ground truth information on the position of canola fields for the years 2001 and 2002. The position of canola fields and non-canola fields could be compared to the classification result for the TM images 196/023; May 10, 2001 and 195/023; May 11, 2001. The frame 196/023 covers the complete area mapped and can be compared to the complete mapping data set in 2001 and 2002. The image 195/023 only covers a smaller part in the eastern region and the number of comparable samples is reduced.

The mapping of interbreeding partners contains information on the habitat type in which the interbreeding partner is found. The habitats usable for the comparison with the results those marked as “is canola field” or “neighbours canola field”. Additionally, the fields of other crops are also listed as habitat type, e.g., marked as “is cereal field” or “is corn field”.

Table 5.1: Validation of the classification with the mapping of interbreeding partners. Listed is the number of canola field habitats within 120 m of pixels classified as canola and the habitats of fields of other crops that are within 60 m of pixels classified as canola. The number in brackets show the result for the visually tested habitats.

| frame | date | canola field | | | non-canola fields | |
|---|---|---|---|---|---|---|
| | | total | found | missed | total | misclassified |
| 196/023 | May 10, 2001 | 132 (102) | 87 | 45 (15) | 60 | 2 (1) |
| 195/023 | May 11, 2001 | 97 (65) | 53 | 44 (12) | 41 | 2 (0) |
| 196/023 | April 3, 2002 | 297 (243) | 140 | 157 (103) | 141 | 5 |

These habitat types are compared to the pixel-based classification. The limited mapping accuracy – the database has been collected without the aid of a GPS receiver – requires to define a neighbourhood in which the habitat types are judged as found if a classified canola pixel is present. A distance of 120 m has been selected as direct neighbourhood of the habitat of types "is canola field" or "neighbours canola".

Since the habitat types "is field of other crops" are available, a negative test is also applied by comparing non-canola field habitats with the classification result. The mapped position is marked as "misclassified".

The result of this validation is shown in Table 5.1 for the two different images available in 2001 and for the one of 2002.

The validation of the image 196/023; 2001 show that the 87 of 132 fields could be identified. Nonetheless, the number of 45 fields missed is larger than expected. Therefore, the missed fields have been visually inspected in the satellite image, which showed that 30 of those fields could not be identified visually in the satellite data.

The reason for this is probably errors in the mapping or fields that have been ploughed in in the meantime. Besides errors in the mapping of the habitats, this results from spring sown canola which is not asssigned a special habitat type and was not considered in this study. Of the remaining 15 fields, 10 fields were to small to be identified and 5 fields showed strong flowering. Note that the latter fields are all partly identified but not at the location of the mapped habitat.

The situation for image 195/023; 2001 is similar and the majority of habitats could be identified: 53 of 97 fields in the classification result. Nonetheless, the classification misses 12 fields in this region, of which 8 were too small and 4 were only partly classified.

As described in Section 4.4.1 (p. 127), the selection of training data was not very reliable for the image acquired in 2002. This can be confirmed by the identification of habitats since only half of them could be identified.

### GPS Mapped Fields

In addition to the habitat mapping, 15 canola fields were visited on May 15, 2001 in the area of Bremen. For each field, the position of at least two corner points and the field orientation with respect to North have been measured with the aid of a GPS receiver. For three of these fields, all corner coordinates have been determined.

All of the present 15 fields were identified by the classification. The field shapes of the three fields of which all corner coordinates were known have been used to compare the field size from the classification result with the field size derived from the corner coordinates and resulted in 92 %, 87 % and 69 % of these fields.

Additionally, 12 fields identified in the satellite image 196/023; May 11, 2001 were visited afterwards at the end of June and all fields could be confirmed as canola fields.

### Seed Growing Fields

Further positions of canola fields could be obtained for fields used for canola seed production. The fields were visited with an associate of a seed producer *Deutsche Saatveredelung* (German Seed Refinement). The positions and arrangements of fields and surrounding surface objects like woods, farm houses or roads have been noted, which allowed to identify the correct field at the position marked with a GPS receiver.

These included 9 fields for the year 2002, 5 fields for 2001, one for 2000 and one for 1999. The comparison with the classification showed, that 6 of the 9 fields in 2002, 4 of those in 2001 and both fields from 1999 and 2000 could be identified by the satellite data.

### Experimental Fields

Another validation data set is available from the experimental farm in Sickte near Braunschweig. It also contains information on the field size which allows to compare the area of the segments with the field sizes. The field size of the classification was corrected for mixed pixels by adding the halved area of the border pixels. Figure 5.7 shows the positions and shapes of those experimental fields on which canola was grown sometime in the period of 1995 to 2001. The fields with the white border were canola fields in 1995. Generally two to three of these fields were sown with canola each year.

Table 5.2 shows a comparison of these fields with the classification results for the images available for this area. It can be seen that except for two fields in 2000 and one in 1996 (cf. the zeros in the right part of Table 5.2), all fields have been identified. The area identified generally ranges from 70 to 130 %. In this context it is important to note that these fields are quite small and an misclassification of a single pixel has large influence on the accuracy. One field was overestimated by 90 % in 2001, because there was a canola field

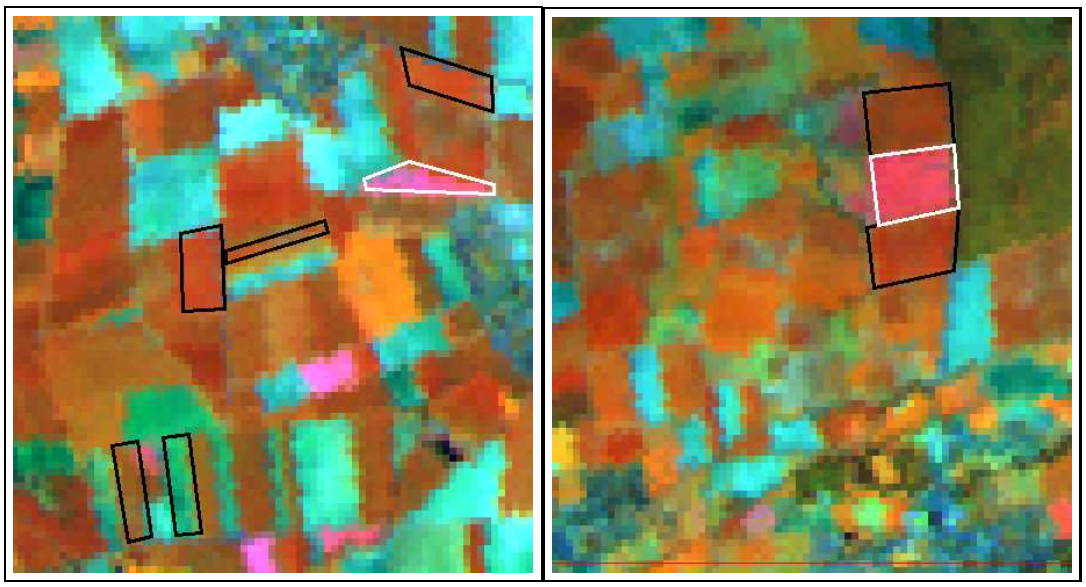


Figure 5.7: Field used for canola cultivation at one time during the period of 1995 to 2001 overlaid over the images from 195/023 and 195/024 from May 3, 1995. Left: Fields near the village Sickte. Right: Fields near the village Wendhausen. Both clips are located southeast of the city of Braunschweig. Note, that the fields marked by the white borders were canola fields in 1995, which can also be seen from the pink colour of the false colour representation. Original data: LANDSAT TM ©ESA, 1995. Distributed by Eurimage.

directly adjacent to the experimental field. Both fields have been joined by the segmentation algorithm and therefore, the area of the experimental field is overestimated.

Consequently, these results generally confirm the accuracy obtained from the Quillow mapping data set.

#### Conclusion

The above evaluation data sets showed, that the majority of fields could be identified by the classification algorithm. The addition of all available field positions under the exclusion of the comparisons for 2002, yield that 189 of 224 fields could be identified which is 84,3 %. Only one field was misclassified as canola. The available field sizes showed a similar result with agreements of 70 to 130%.

### 5.2.2 Agricultural Statistics

In contrast to the spot tests with the positions of known fields, the agricultural statistic allows to compare the global results of the classification. Two types of statistics are available in this study, the county statistics collected in 1995

Table 5.2: Ratio of the area of mapped experimental canola fields $a_m$ and the area identified by the classification $a_c$. Listed are the percentages for fields that were seeded with canola in the particular year (see Figure 5.7). The order of entries in the last coloumn is depending on the size of the mapped field. Note that there were nine fields available in total, however only fields that were seeded with canola in that year and appear in the satellite image, are listed.

| Image | Fields | $a_c/a_m$ [%] | | | |
|---|---|---|---|---|---|
| 195/023; May 11, 2001 | 1 | 72 | – | – | – |
| 195/024; May 3, 2001 | 3 | 105 | 90 | 190 | – |
| 195/023; May 16, 2000 | 1 | 69 | – | – | – |
| 195/024; June 9, 2000 | 2 | 0 | 0 | – | – |
| 195/023; May 11, 1998 | 2 | 99 | 121 | – | – |
| 195/023; June 6, 1996 | 3 | 121 | 99 | 100 | – |
| 195/024; June 6, 1996 | 4 | 126 | 112 | 121 | 0 |
| 195/023; May 3, 1995 | 1 | 69 | – | – | – |
| 195/024; May 3, 1995 | 2 | 101 | 72 | – | – |

for all counties in Germany and the township statistics from 1995 and 1999 for the federal state of Schleswig-Holstein (see Section 2.2.3, p. 36).

In order to compare statistics and classification, the acreage of identified fields within a county or township, respectively, has been totalled and was then compared to the statistical estimation. The ratio of satellite-derived acreage and acreage from statistics is expressed in percent. The error of the field size resulting from mixed pixels at the borders of the fields is taken into account by adding the halved area of the neighbouring pixels to the field sizes.

Additionally, the area covered by the satellite data has been determined and only counties or townships with a satellite coverage of 90 % were evaluated.

### County Statistic

Figure 5.8 shows the comparison of the statistics and classification. The differences are indicated by the fill colour of the different counties. Additionally shown are the edges of the satellite frames available for 1995, to indicate the area covered by satellite data.

Figure 5.8 shows good agreement of 70 to 110 % identified acreage for most counties. Moreover, the agreement in the main cultivation area in the east of Schleswig-Holstein is 90 to 110 % for 4 counties and 70 to 90 % for 7 further counties. Results of 90 to 110 % agreement are also achieved in the counties Verden and Nienburg located south of Bremen and also show a dense canola cultivation.

This was not valid for the county of Harburg, located directly south of the

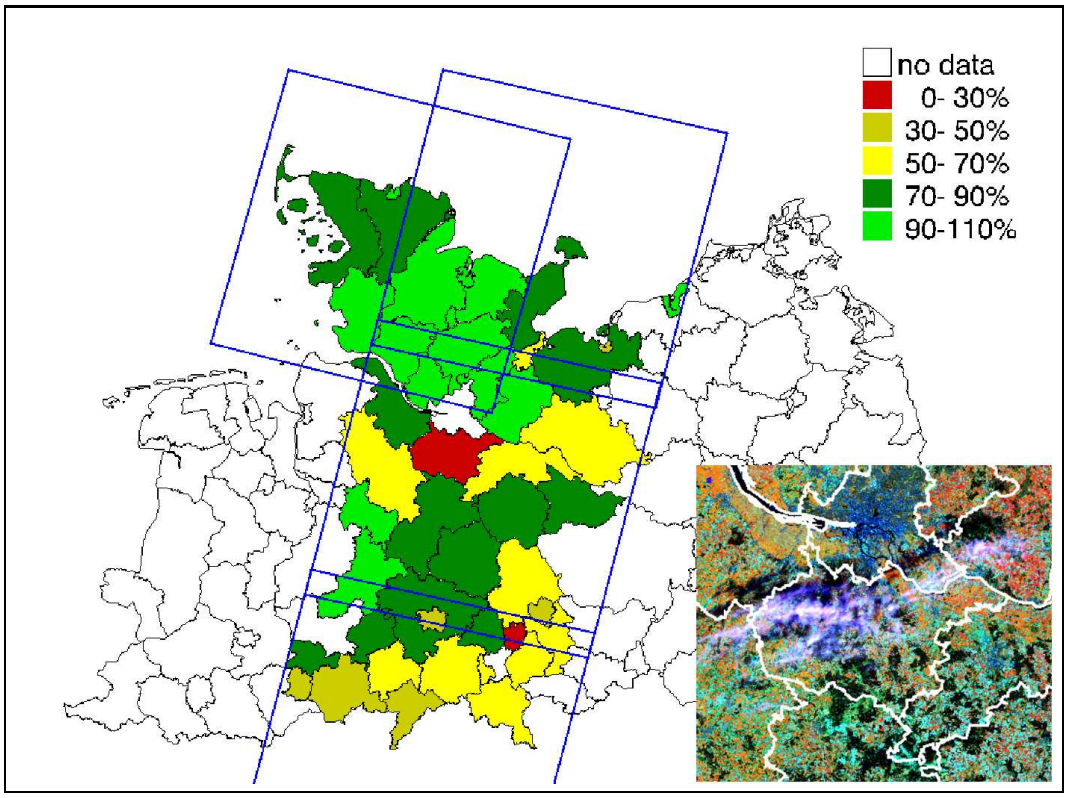



Figure 5.8: Comparison of the classification results and the cultivation statistics collected in 1995 for the counties. The blue rectangles show the coverage of the four available satellite images. The counties are marked by the thin black lines. The fill colours of counties indicate the comparison result. No fill colour is used for counties where the satellite coverage was below 90 % or where no statistical data were available. The image in the lower right shows the cloud cover which prevented a classification in the county of Harburg, south of Hamburg. Original data: LANDSAT TM ©ESA, 1995. Distributed by Eurimage.

city of Hamburg, where less than 30 % of the canola could be identified by the satellite data. This error results from cloud and haze cover in this county which is shown in the lower right clip in Figure 5.8.

According to the applied cloud algorithms 12 % of the area were covered by thick clouds and 37 % by thin clouds. Additionally, Figure 5.2 shows that in this area canola is not cultivated widely. Moreover, most of the canola fields are located in the northern part of this county which is cloud-covered (see lower right image in Figure 5.8.

The southernmost counties also show a less successful identification of canola acreage with mostly 50 to 70 % but also only 30 to 50 % for five of the townships. A possible explanation is the low amount of canola acreage in these counties. Note that the two smaller counties located northwards are the cities of Hannover and Braunschweig where the canola acreage naturally was also very low.

### Township Statistics

A more detailed comparison is possible with the township statistics from Schleswig-Holstein. Figure 5.9 shows the results of this comparison for the statistics of 1995 and 1999. In 1995, the classification of the complete area of Schleswig-Holstein could be compared to the statistics because of the complete coverage with satellite data. In 1999 only about half of the area was covered and could be evaluated. The frame coverage for that year can be seen in Figure 5.8. Note that, for the statistics, all fields are assigned to the township of residence of the owner or tenant, even if fields are in another township. This is especially important for the townships neighbouring Mecklenburg-Vorpommern since farmers from these townships have frequently acquired acreage in the neighbouring federal state. The results of this comparison are discussed separately:

**1995:** The left map in Figure 5.9 shows the comparison for this year. The comparison shows good agreement for the main cultivation areas in the east of Schleswig-Holstein. Generally the comparison shows 70 to 130,% agreement for this area. Very high or very low percentages of above 130 % or below 70 % in some townships can be explained by the farm-based assignment described above. The comparison of the western regions of Schleswig-Holstein show a higher differences in the comparison, which result from the much lower cultivation density in these regions, which lead to higher differences in the relative comparison. Nonetheless, classification and statistics show good agreement for the townships without canola acreage marked in blue. Here, also the farm based assignment of acreage can be the reason for some of the differences.

**1999:** The right map in Figure 5.9 shows the result of the comparison for 1999. Apparent is the overestimated amount of canola acreage on the island of Fehmarn and on the adjacent peninsula. This can be explained by strongly flowering canola observed in the satellite data. Directly below this region the comparison shows very low amounts of identified canola. This is the consequence of a thin cloud cover in this region which could not be compensated by the haze correction because of the strong flowering. The same effect is the reason for the underestimation in the northernmost part of this map. In the western part of Schleswig-Holstein, there are variations which can again be explained from the very low cultivation density in this region.

### Area Comparison

In addition to the discussion on the acreage for single townships or counties, respectively, statistic and classification are also compared for the complete overlap of statistics and classification.

The results of this comparison for area of completely covered townships or counties are shown in Table 5.3. The list shows the ratio of the canola acreage

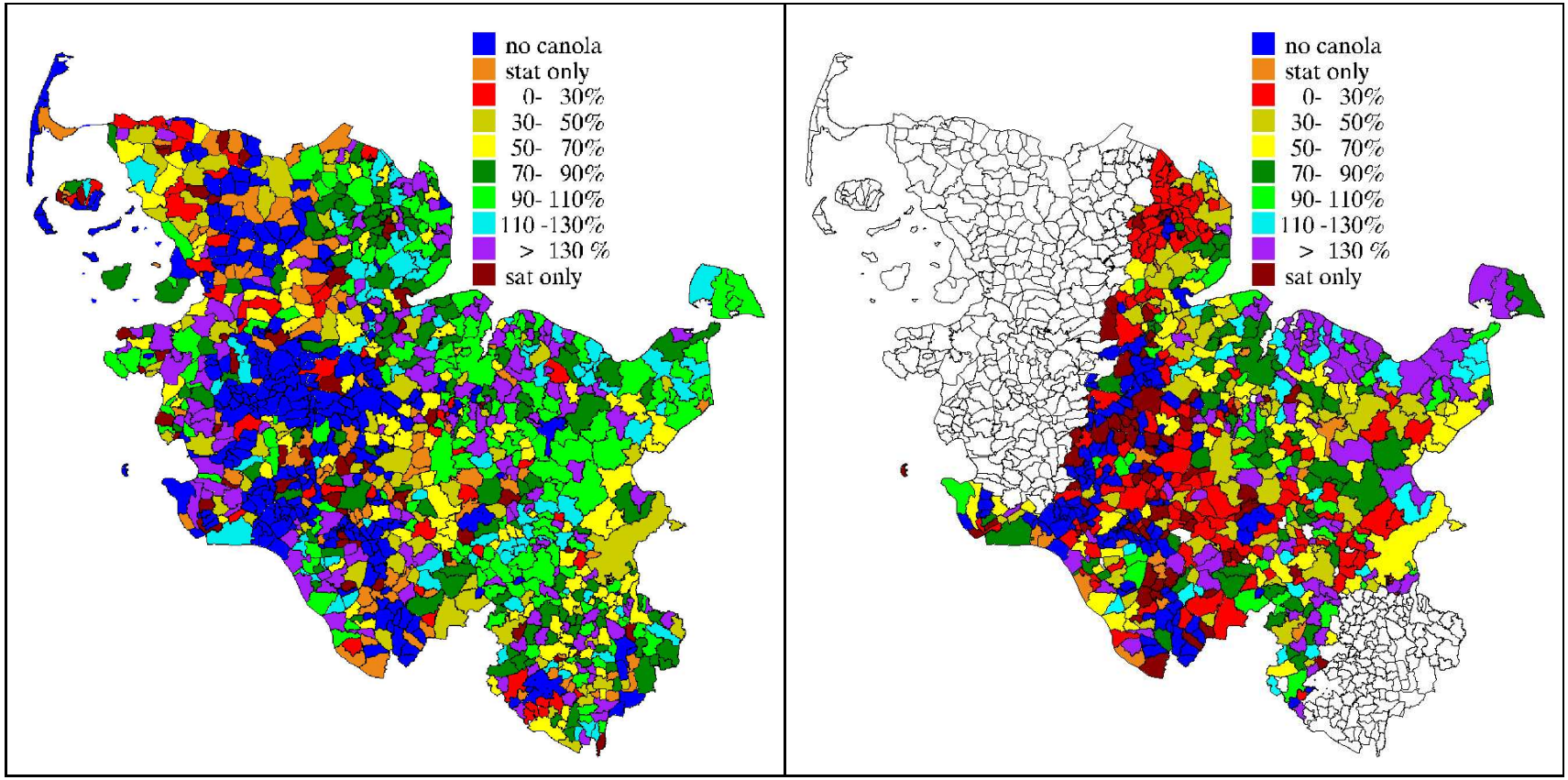



Figure 5.9: Comparison of the classification result for the township statistics of Schleswig-Holstein for 1995 (left) and 1999 (right). The colour scale is similar to the one used in Figure 5.8, with the exception of three additional colours: "no canola" indicates townships with no canola acreage, neither in the satellite classification nor in the statistics, "stat only" indicates townships with only canola acreage noted in the statistics and "sat only" indicates townships where canola acreage has only been found in the satellite data.

Table 5.3: Comparison of the identified canola acreage $A_C$ and the acreage estimated by the different statistics $A_S$. Additionally included is the area of the border pixels $A_p$. The results are shown for the township statistics in Schleswig-Holstein (SH) and for the county statistics in northern Germany (NG).

| Statistic | Year | eval. area [ha] | ratio [%] | | |
|---|---|---|---|---|---|
| | | | $A_C/A_S$ | $(A_C + 0.5A_B)/A_S$ | $(A_C + A_B)/A_S$ |
| Township SH | 1995 | 1,009,204 | 81.16 | 102.49 | 123.82 |
| Township SH | 1999 | 491,501 | 68.04 | 88.04 | 108.05 |
| County NG | 1999 | 4,497,541 | 62.45 | 82.88 | 103.328 |

found by the satellite and the acreage statistically estimated. Additionally, the influence of the border pixels is included into this comparison.

From this table, the acreage of canola is underestimated by 20 to 40 % if the border pixels are not taken into account. Better results can be obtained by adding the halved border pixel area to the field sizes. In this case, the classification from 1995 showed only 2 % deviation from the statistic. The comparison for the year 1999 still showed a difference of 20 to 10 %. This error can be compensated by adding the full size of the border pixels, which leads to an overestimation of 23 % for the statistic from 1995.

These differences in classification accuracy can be explained by a thin cloud cover that was present in 1999 and not in 1995. Nonetheless, cloud covered areas were small compared to the complete evaluated area. Another possible explanation is the strong flowering that could be observed in 1995 and allowed an easier separation of canola than those of 1999. On the other hand, this effect can also be responsible for an overestimation of canola acreage because of the misinterpretation of mixed pixels (see below).

**Conclusion**

The comparison with the statistics confirms the amount of 80 % to 90 % of identified canola obtained from the classificaton in the Quillow mapping data set. Similar the comparison to other known fields showed similar fraction of field sizes identified by the classification.

## 5.3 Error Sources for the Classification

This section will discuss the different sources for errors in the final result of the identified canola fields. The identified acreage depends on the number of correctly identified canola pixels and on border pixels, i.e., pixels which additionally include other surface types than canola (mixed pixels).

Besides this obvious error of the total identified acreage of canola, classification errors can have a stronger influence on the other parameters derived

in this study. Therefore, this section will give a short overview over the error sources for the derived parameters.

### 5.3.1 Classification Errors

Most obvious are the pixels that are wrongly interpreted because of errors in the classification algorithm and the chosen training data. According to Section 4.3.2 (p. 119), this canola acrage that can be identified with remote sensing is about 80 % to 90 % for the classification method used in this study. Remarkable is the high amount of canola found in the images which showed strong canola flowering. One effect responsible for this is the increased separability between canola and other vegetation, which allows improves classification accuracy.

Another effect of the strong flowering on the classification result is the misclassification of mixed pixels. Mixed pixels are pixels that cover more than one surface type. In this study, the amount of canola acreage in these mixed pixels at the canola field borders is estimated by identifying border pixels and adding half of their area to the field size. This estimation might be influenced by the strength of canola blooming because of the bright yellow flowers. The effect can be illustrated as follows: the reflected radiation of a canola field is a combination of the radiance reflected by flowers and green leaves. In bordering pixels more green from non-canola plant cover is mixed into it. So, a border pixel of a field with strongly flowering canola most likely resembles a fully canola covered pixel with less intense flowering. Thus an only partly covered pixel is interpreted as fully covered by canola and thus overestimate the amount of identified canola is overestimated. This simplified explanation can also be applied to the three channels used for the classification since the signature of flowering canola is also distinct in the false colour representation and will most likely also lead to an overestimation of canola. Note that the optics of the sensor also have an influence on the classification of pixels neighbouring canola fields since the radiance measured for one pixel originates also from areas directly adjacent to the pixel, which is described by the point spread function (PSF). This effect is called adjacency or overshine effect. An accurate estimation of this effect requires the knowledge on the spectral properties of the surface types within the pixel. This effect is difficult to estimate without additional digital maps or higher resolved satellite or aircraft data. Further information on mixed pixels and adjacency effects can be found in Cracknell (1998); Huang et al. (2002); Townshend et al. (2000); Ren and Chang (2002); Kerkes and Baum (2002); Garcia-Haro et al. (1999).

### 5.3.2 Splitting and Merging of Fields

The field size is of special interest in this study. Therefore, it is necessary to discuss the comparability of segments derived from satellite data with field shapes on the ground. In this context, there are two possible sources of errors:

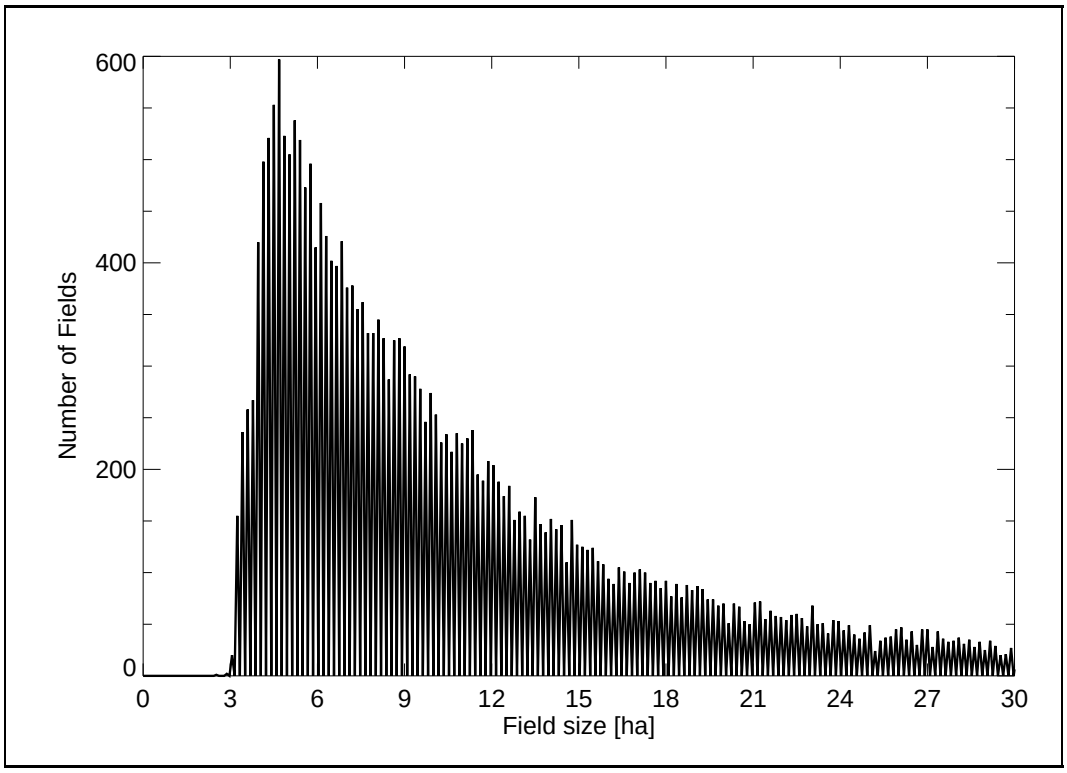



Figure 5.10: Field size distribution for northwest Germany (West of 10°E) displayed as a histogram for the field sizes for the year 2000. The majority of fields have sizes of about 4 ha. Note that the field size is not truncated at the minimum field size of 2 ha since the halved border pixel area is added to the field size.

**Merging of Fields:** Adjacent fields are joined two one segment. Depending on the neighbourhood definition, this is already the case if two pixels are diagonally adjacent.

**Splitting of Fields:** Narrow fields or fields that show high variation in their reflectance can be misinterpreted as two individual fields if pixels of the fields are missed by the classification algorithm.

These two error sources have an effect on the field size estimation. On the one hand, if fields are joined, they are interpreted to be larger than they truly are; on the other hand, if a smaller and narrower field is interpreted as two individual fields, the number of smaller fields is wrongly increased.

### 5.3.3 Undersized Fields

Another source of errors is the definition of a minimum field size (undersized segments) since fields smaller than 2 ha are removed. Obviously, the field size is limited because of economical considerations, i.e., larger fields have a better cost-value-ratio than smaller ones. Nonetheless, the minimum field size has

to be evaluated. A first exemplary evaluation has been undertaken with the comparison of the spatial resolution of an aerial photograph and of the TM sensor in Section 2.1.1, p. 17.

A further evaluation is possible with the classification results yielded from this study by evaluating the distribution of field sizes. Figure 5.10 shows the distribution of canola fields in northwest Germany in the year 2000. The northwest has been chosen since the fields in western Germany are generally smaller than in the east. It can be seen from this figure that the majority of fields have a size of about 4 ha. The number of fields smaller than this size decreases rapidely.

## 5.4 Conclusion

In this chapter, a selection of results from this study has been shown. The main cultivation areas have been identified successfully and also the field size distribution yielded reasonable results for the years 1995 to 2001. The results from 2002 showed a large amount of misclassifications due to cloud cover but also due to the early acquisition date of the satellite data in the beginning of April. The validation confirmed the accuracy of the identification of about 70 to 90 %, which is nearly as good as in the Quilllow mapping area where the accuracy is 78 to 86 %. The discussion on the error sources showed, that further investigations on the effect of flowering on the classification of neighbouring pixels is necessary, especially since the flowering is responsible for a difference of up to 20 % of the classification accuracy. Additionally, the determination of field sizes can be improved with a more sophisticated vectorisation of the classification result and the integration of an electronic map (Evans et al., 2002; Le Moigne and Tilton, 1995; Huang et al., 2002).

# Chapter 6

# Summary and Outlook

The first part of this chapter recapitulates the achievements of the methods in preparing and classifying the satellite data using the methods presented in this thesis. Moreover, it gives a short overview of the information gained on the canola cultivation in northern Germany regarding potential gene transfer to non-modified canola or wild relatives of canola.

The second part gives an outlook of the possible improvements of some of the applied methods presented in this thesis. Moreover, it describes the possible usage of these data for further studies on the risk assessment of GM plants.

## 6.1 Summary

The aim of this work was the identification of canola fields in northern Germany for the period from 1995 to 2002, analysing a total of 47 LANDSAT TM and 1 IRS LISS/3 satellite images. This large amount of data required the development of mostly automated methods for the georectification, the atmospheric correction and the classification. A detailed assessment of these methods has already been given at the end of the according chapters. Thus, here, only a brief overview will be given:

**Data selection:** The most appropriate sensors currently available for the identification of canola fields in the complete area of northern Germany are LANDSAT TM/ETM+ and IRS LISS/3.

**Georectification:** The investigation of the mapping accuracy showed that a georectification with a polynomial approximation of first order is sufficient for a mapping accuracy of about 120 m. This result is probably not valid for regions with higher elevations than the flat regions of northern Germany. An application of the georectification in other regions might therefore require the use of a DEM. An additional correction based on image correlation achieved an accuracy of 30 m for the identification of corresponding pixels in overlapping satellite images.

**Atmospheric Correction:** A number of atmospheric influences on the classification can be compensated by selecting training data from the satellite images themself. This is however not possible for effects of clouds. Opaque clouds and their shadows could be identified using a threshhold for brightness and cloud top temperature. The corresponding cloud shadows wre identified by estimating the distance and direction of clouds and cloud shadows, which could be determined by a comparison of cloud cover and dark surfaces. Thin clouds or haze are detected by the HOT method proposed by Zhang et al. (2002a) which has been adapted to regions with flowering canola fields. In combination with a histogram-based correction, the scattering effects of thin cloud-cover on the radiance received by the satellite sensor were corrected. Note that, this correction was not possible over flowering canola fields since the flowering prevented a correct estimation of the haze amount by the HOT method.

**Classification:** The Mahalanobis distance classifier (MDC) has been compared to the maximum likelyhood classifier for various agricultural plants in order to adjust the classification range of the MDC. The classification accuracy of the MDC ranged from 63 to 71 %. The adjusted MDC was applied to all images, since it only depends on the training data set for one single surface type here, canola. This allowed a partly automatic selection of training data from the overlap of satellite images. The MDC had to be manually adapted to variations of the flowering strength. Additionally, the haze correction has been included in the algorithm. Note that, this haze correction failed over flowering canola fields because the HOT method does not work over such fields (see above).

**Postclassification:** The classification result was used to construct segments to identify individual fields and also to apply a correction of the classification by including adjacent pixels with similar reflectance properties. This improved the classification accuracy to 78 to 86 %. Moreover, the segments were approximated as rectangles for a vectorised representation of the classification result.

The results of the classification were compiled in various data sets and were distributed to the participants of the GenEERA project. Additionally, averaged and totalled parameters have been presented which allow a qualitative discussion on the hybridisation and contamination probabilities for GM canola.

It could be seen that an increased probability of contamination is to be expected in the main cultivation areas in eastern Schleswig-Holstein, Mecklenburg-Vorpommern and northern Brandenburg, mostly because of the large cultivation density of up to 42 % and the short minimum distances between canola fields with a mean of about 0.5 km for most regions.

The hybridisation probability is increased because of the high cultivation density, but also because of the frequently irregularly-shaped fields in Mecklenburg-Vorpommern which results in an increased contact line length

with other vegetation. Another region with an increased contact length is located south of Bremen because of numerous small fields.

## 6.2 Outlook

This work was part of the joint research project GenEERA. The data presented here were used by the project partners for further comparison with ecological and agricultural parameters. These results will be presented in the final report of the GenEERA project and are already partly published in Breckling et al. (2003).

The identification of surface types is a major task of satellite remote sensing. Especially the identification of crops is important for agricultural aid control and yield predictions. Therefore, an important extension of the methods described in this thesis is its application to other agricultural crops. This is also important for the risk assessment of other GM plants.

Additionally, there are possible improvements for the different methods used in this study:

**Georectification:** The georectification will require an additional DEM when applied to other than the flat regions of northern Germany. Moreover, the correlation method can be used to achieve a pixel accurate georectification by georectifying a reference image for each frame to a map with this accuracy.

**Cloud Identification:** The HOT method may be adaptable to the flowering of canola by identifying a direct relationship of the reflectances for the TM Channels 1 and 3 and allow to develop a correction depending on the strength of flowering.

**Classification:** A major improvement for the classification would be a better estimation of the border pixels of the fields. Especially for northwestern Germany with its mostly small fields the accuracy of the classification could be improved. A possible solution is the use of higher resolved satellite data. Unfortunately, sensors with a higher spatial resolution and a coverage comparable to that of LANDSAT TM are neither available nor planned. A possible improvement of the accuracy might be gained by identifying the fraction of different surface types in the border pixel. Such methods have been proposed by Huang et al. (2002); Townshend et al. (2000); Ren and Chang (2002). Nonetheless, these methods will have to compensate the influence of brightly flowering canola.

**Postprocessing:** The postprocessing can be improved by a more sophisticated approximation of the field shapes. This can be achieved by identifying all agricultural crops on the field, and if possible the fraction at the border pixels. A possible solution was presented by Janssen and Molenaar (1995); Smith and Fuller (2001). In this context, also an additional

vector data set, e.g. polygons instead of rectangles, can give further information on the exact field shapes. Nonetheless, it has to be stated that these improvements will require either a much longer processing time or expensive mapping data.

This work showed that a mostly autonomous crop type identification over large areas with high resolution sensors is possible, which can improve the various applications of remote sensing in agriculture and ecology. Moreover, the usage of auxiliary data could be restricted to a minimum, which is essential for the evaluation of data form larger areas.

# Appendix A

# Approximated Rectangles

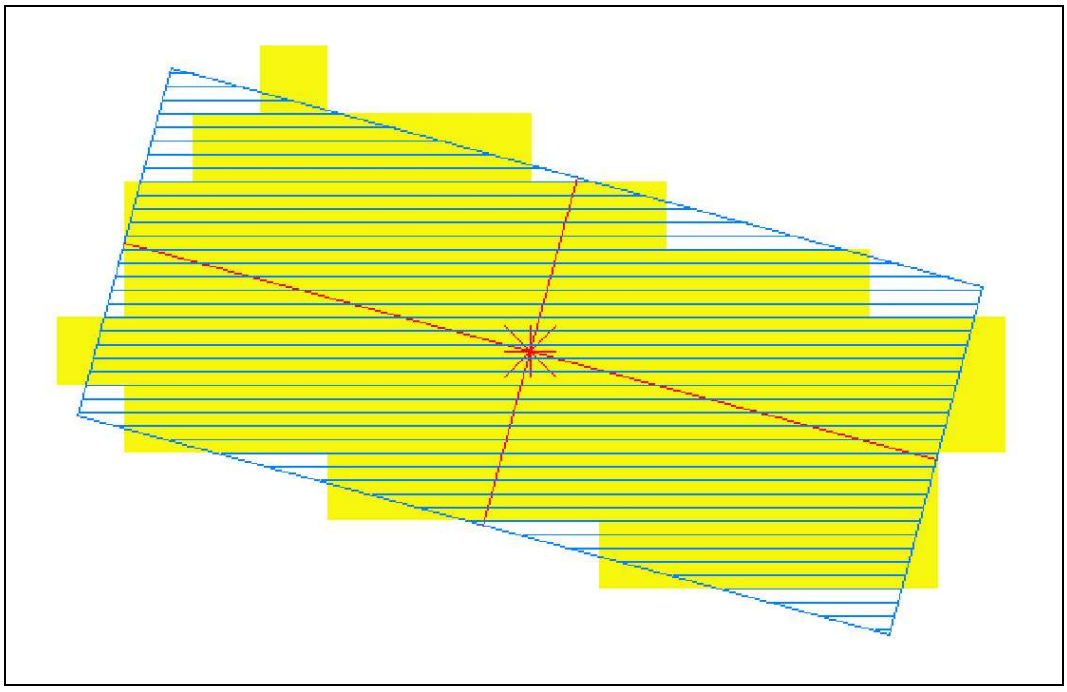


Figure A.1: Example for the approximation of a segment of pixels by a representative rectangle (blue hatch). The star indicates the centre of the segment calculated by the centre of mass relation. The red lines indicate the directions of the eigenvectors.

Classification results of satellite or computer images are usually pixel based raster data. Vectorised data have the advantage of being easier to process, e.g., to calculate size, determine the intersection with other objects or to derive distances between objects.

There are different possibilities to generate a vectorisation from raster data. Generally, a segmentation is performed first to identify pixels belonging together. One possibility to derive vector information from the pixels of a segment is to use the outlines of the pixels at the segment border. This increases

the necessary memory and also does not allow an easier processing. A better possibility is to approximate the segments with a simple geometric form. Since fields frequently have a rectangular shape, rectangles are chosen as representation for the fields. This appendix briefly describes the mathematical background of this approximation.

## A.1 Centre of Mass

Important is the position $\bar{\mathbf{r}}$ of the rectangle. The centre position, named here centre of mass (CM), can be calculated with the centre of mass equation:

$$\bar{\mathbf{r}} = \frac{\sum_i \mathbf{r}_i A_p}{\sum_i A_i} \tag{A.1}$$

Here, $A_p$ is size of a pixel and $\mathbf{r}_i$ the centre position of the of the pixel $i$.

## A.2 Principal Axes

Starting from the centre of mass, the principal axes can be calculated. The inertia tensor $\mathbf{M}$ can be derived from the pixel position by:

$$\begin{aligned} \mathbf{M} &= \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \\ &= A_p \sum_i \begin{pmatrix} (y_i - \bar{y})^2 & (x_i - \bar{x})(y_i - \bar{y}) \\ (x_i - \bar{x})(y_i - \bar{y}) & (x_i - \bar{x})^2 \end{pmatrix} \end{aligned} \tag{A.2}$$

$$\tag{A.3}$$

The principal axis can be identified with an eigenvalue analysis. The eigenvalues $\lambda$ can be determined by solving the following equation:

$$\det(\mathbf{M} - \lambda \bar{I}) = \begin{vmatrix} m_{11} - \lambda & m_{12} \\ m_{21} & m_{22} - \lambda \end{vmatrix} = 0 \tag{A.4}$$

$$\lambda^2 - (m_{11} + m_{22})\lambda + m_{11}m_{22} - m_{12}m_{21} = 0 \tag{A.5}$$

Solving this equation leads to the two eigenvalues $\lambda_+$ and $\lambda_-$, which are squares of the length of the principal axes.

$$\lambda_+ = \frac{m_{11} + m_{22}}{2} + \sqrt{m_{12}m_{21} + (m_{11} - m_{22})^2/4)} \tag{A.6}$$

$$\lambda_- = \frac{m_{11} + m_{22}}{2} - \sqrt{m_{12}m_{21} + (m_{11} - m_{22})^2/4)} \tag{A.7}$$

The eigenvalues are used to calculate the corresponding eigenvectors $\mathbf{e}_+$ and $\mathbf{e}_-$ by solving:

$$(\mathbf{M} - \lambda_\pm \mathbf{I})\,\mathbf{e}_\pm = \begin{pmatrix} m_{11} - \lambda_\pm & m_{12} \\ m_{21} & m_{22} - \lambda_\pm \end{pmatrix} \mathbf{e}_\pm = \mathbf{0} \tag{A.8}$$

These eigenvectors indicate the direction of the main axis and one belonging to the higher eigenvalue can be used to determine the orientation with respect to the North.

## A.3 Area-Conserving Rectangles

Agricultural fields are frequently rectangular. Thus, it is the shape most appropriate for an approximation. The orientation of the rectangle is represented by the eigenvectors. Now, the length of the sides have to be determined from the eigenvalues. To conserve the ratio of the two main axes, the following equations can be written for a rectangle with the area $F$ and the sidelengths $a$ and $b$:

$$F = ab \quad \text{with} \quad \frac{a}{b} = \sqrt{\frac{\lambda_+}{\lambda_-}} \tag{A.9}$$

$$\Rightarrow a = \frac{F}{b} \tag{A.10}$$

$$b = \sqrt{\frac{\lambda_-}{\lambda_+}}a \tag{A.11}$$

$$a = \frac{F}{\sqrt{\frac{\lambda_-}{\lambda_+}}a} \tag{A.12}$$

$$b^2 = \frac{\lambda_-}{\lambda_+}a^2 \tag{A.13}$$

$$a^2 = \sqrt{\frac{\lambda_+}{\lambda_-}}F \tag{A.14}$$

$$b^2 = \frac{\lambda_-}{\lambda_+}\sqrt{\frac{\lambda_+}{\lambda_-}}F \tag{A.15}$$

Therefore, the length of the rectangle can be determined by the following equations:

$$a = \sqrt[4]{\frac{\lambda_-}{\lambda_+}}\sqrt{F} \tag{A.16}$$

$$b = \sqrt[4]{\frac{\lambda_+}{\lambda_-}}\sqrt{F} \tag{A.17}$$

The size of the rectangle can simply be determined from the sensor's resolution and the number of pixels in the segment.

# Appendix B

# Mapping Errors in the Ground Truth Data Set

In order to identify representative areas in the satellite images it is necessary to inspect the quality of the mapped data for true colour and false colour (channels 4, 5 and 3) representation. The image clips in figure B.1 show a clip of a TM image overlaid with the quillow mapping data set. Examplarily emphasized are the edges of canola (yellow) and wheat (turquoise) fields.

This can be seen from the upper left canola field in the clips in figure B.1, that both indicate that there are two different agricultural plants cultivated on this fields. Moreover, a comparison with other canola fields suggests that neither of these plant covers are actually canola. This inaccuracy of the mapping results from the different partitioning of fields by the farmers which are not included in the quillow data set, since the field borders are mapped only once and the plant is later determined by interogating the farmers. Another possible reason for this faulty labeling might be explained with damages caused by freezing or drought. In these cases, farmers at times plough the fields in order to plant a new plant, e.g. spring sawn canola or maize.

Figure B.1: Example for the inaccuracies of the agricultural plant mapping at the river Quillow. Left: true colour representation of a clip of the TM image 193/023; 2001. Right: false colour representation of the same clip. Both clips are overlaid with the mapped field edges from the quillow data set. Shown are canola (yellow) and wheat (turqoise) field egdes that has been marked as canola. Most obvious is the bare soil on the left side of the upper right field. It apears brown in the true color and turqoise in the false colour representation. A comparison to other fields mapped as canola reveals that the right side of this fields is also not canola. Also visible are smaller areas with different colour within the other fields. These are most propably small woods or areas that are also used to plant different crops.

# Appendix C

# Available Data Set DVDs

This appendix will give an overview of the structure of folders and files on the digital versatile disks (DVDs) that have been created with the classification results from this project. It mainly addresses the project partners of GenEERA but also other parties interested in canola cultivation in northern Germany might contact the author of this thesis at this emailaddress: *hlaue@iup.physik.uni-bremen.de*.

In this thesis all results are projected onto the UTM grid for zone 32 North with the WGS84 ellipsoid. Nonetheless, the data can also be obtained in the following projections:

- *Gauß-Krüger* (german grid) for the 3rd, 4th and 5th stripe.
- UTM for zone 33 for zone 33 North.
- European Terrestrial Reference System 1989 (ETRS89) for zone 33 North.

## C.1 Structure of Folders

A sketch of the folder structure is shown in Figure C.1. The uppermost folder on the DVD indicates the type of projection used, e.g., "utm32" stands for UTM zone 32 and "gk3" for Gauß-Krüger third stripe. Below this folder, the left branch in Figure C.1 containes the image based results and the right folder the years based results.

### C.1.1 Image-Based Results

In the branch of the image-based results, the name of the folders below "frame" indicates the type of data. The left branch "raster" contains the pixel-based results and "vektor" the vectorised data. Each of this two folders contain the same structure indicating the image stored in them, i.e., the year of acquisition, which itself contains folders for each row/path combination available in this year, e.g., folder "196023" contains the results for path 196 and row 023.

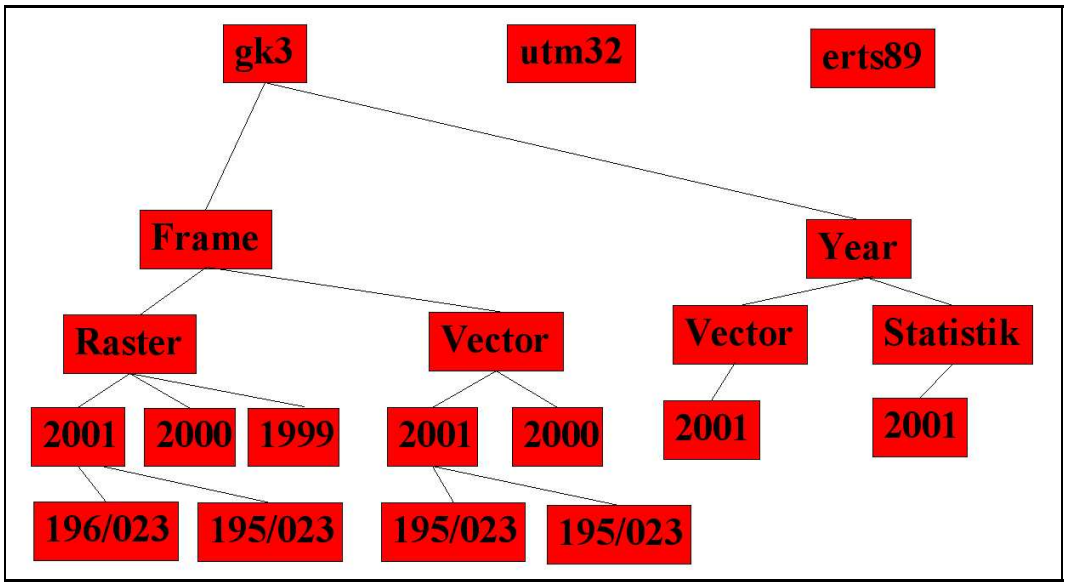



Figure C.1: Structure of folders on the data DVDs.

**Coverage and assignment of path and row numbers**

As introduced in Section 2.1.1 (p. 17), the frames used for satellite data are organised in path and rows. Therefore, the folders for the satellite data in the year folder are named according to the path and row of the satellite data. The positions of the path and rows is shown in Figure 2.2 (p. 19).

### C.1.2 Available Raster Files

For each image, the following files are present:

**cloud_SYSTEM_PATH_ROW_YEAR.tif** contains the geocoded cloud mask. Grey level values of 32 (dark grey) are hazed pixel, which can influence the classification (see Section 3.2.4, p. 71). Grey level values of 64 (a brighter grey) are cloud shadow pixel and grey values of 127 (light grey) are cloud covered pixel, q.v. Figure C.1.2.

**classification_SYSTEM_PATH_ROW_YEAR.tif** contains the canola classification. Grey values of 0 (black) are non-canola pixel. The different shades of grey allow to descriminate different segements. Addtitionally, a grey level value of 154 indicates the CM of the segment[1] and a grey level of 110 indicates border pixel. Additionally, pixel that have been added by the region-growing are are marked with a grey level value of 1900, cp. Figure C.1.2.

**sshift_SYSTEM_PATH_ROW_YEAR.tif** contains the shift in easting

[1]Note that due to the resampling some CM and border pixel are overlooked and therefor not present in the geocoded image

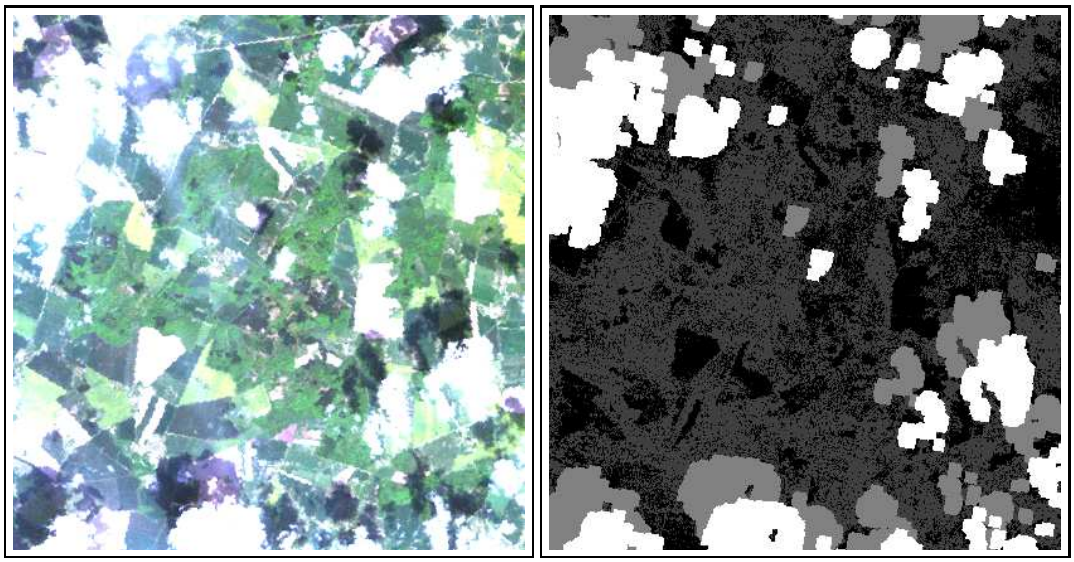


Figure C.2: Example for the cloud cover in the pixel based representation. Shown are clips from Mecklenburg-Vorpommern acquired in 2001. Original data: LANDSAT TM ©ESA, 2001. Distributed by Eurimage.

.

direction in pixel size obtained from the correlation method described in Section 3.1.4, p. 57.

**tshift_SYSTEM_PATH_ROW_YEAR.tif** contains the shift in northing direction in pixel size obtained from the correlation method described in Section 3.1.4, p. 57.

SYSTEM, PATH, ROW and YEAR are replaced by the information of geographic projection, e.g., "utm32", the path, the row and the acquisition year.

### C.1.3 Vectorized data

The vectorized data is available in two different types: a point based and a rectangle based shapefile. The point based shapefile is much smaller and can be processed easier and the rectangle based shapefile allows a better comparison to a map or to the raster classification result. Both data sets contain additional information on the segments stored into the tags of the shapefile. Table C.1 list the tag names and their meaning. Additional to the classification result, there are shapefiles containing the image border of the satellite image.

The following files are present in this folder:

**segmentshape_SYSTEM_PATH_ROW_YEAR_polygon.zip:** Zipped shapefile with the approximated rectangles for each segment.

Table C.1: Tag names and their description stored in the segment shapefiles.

| Tag | Description |
|---|---|
| ID | Sequential segment number |
| SATCMX | x-position of the CM in image coordinates |
| SATCMY | y-position of the CM in image coordinates |
| RECHTS | Easting of the CM [px] |
| HOCH | Northing of the CM [px] |
| SSIZE | Segment size [$m^2$] |
| SANGLE | Angle with respect to the North [°] |
| ACHSE1X | Easting component of the long principal axis [m] |
| ACHSE1Y | Northing component of the long principal axis [m] |
| ACHSE2X | Easting component of the short principal axis [m] |
| ACHSE2Y | Northing component of the short principal axis [m] |
| CH3_MEANRA | Mean radiance for the segment for channel 3 [$W/(m^2 sr \mu m)$] |
| CH3_SDVRA | Standard deviation of the radiance for channel 3 for the segment [$W/(m^2 sr \mu m)$] |
| CH3_MEANBY | Mean of the original satellite data for channel 3 [DN] |
| CH3_SDVBY | Standard deciation of the original satelliten data for channel 3 [DN] |
| CH4_MEANRA | like above for channel 4 |
| CH4_SDVRA | |
| CH4_MEANBY | |
| CH4_SDVBY | |
| CH5_MEANRA | like above for channel 5 |
| CH5_SDVRA | |
| CH5_MEANBY | |
| CH5_SDVBY | |
| CLOUDED | Segment neighbours cloud or cloud shadow |
| HAZY | Segment neighbours hazy region |
| BORDERSIZE | Area of neighbouring pixels [$m^2$] |

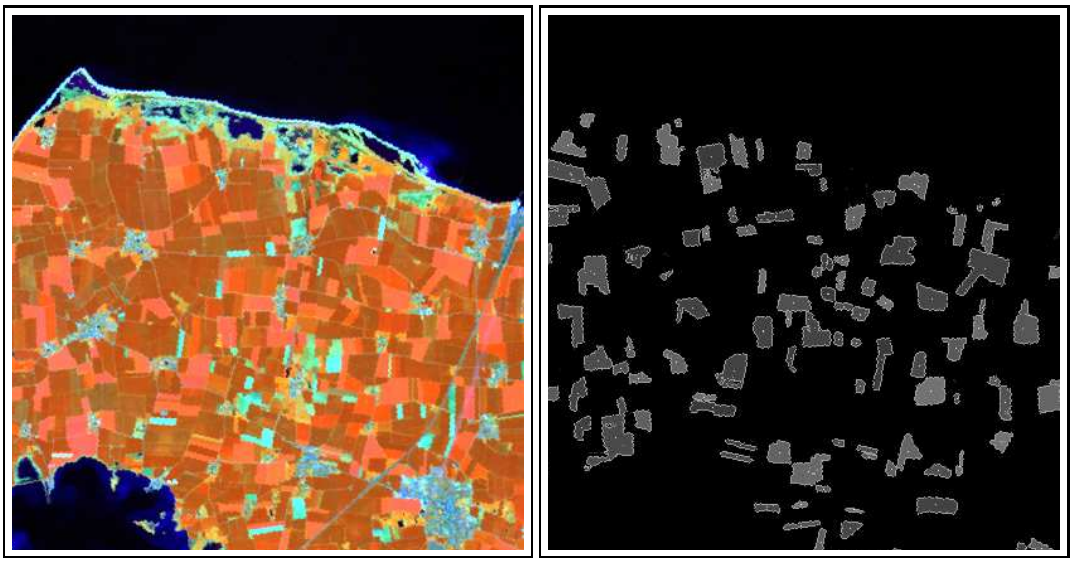


Figure C.3: Example for the identified canola segments in the pixel-based representation. Shown is the island of Fehmarn. Original data: LANDSAT TM ©ESA, 2001. Distributed by Eurimage.

**segmentshape_SYSTEM_PATH_ROW_YEAR_point.zip:** Zipped shapefile with points indicating the CM of the segments.

**bordershape_SYSTEM_PATH_ROW_YEAR_polygon.zip:** Zipped shapefile containing the image border.

**cloudshape_SYSTEM_PATH_ROW_YEAR_polygon.zip:** Zipped shapefile with information on the cloud coverage.

SYSTEM, PATH, ROW and YEAR are replaced by the information of geographic projection, e.g., "utm32", the path, the row and the acquisition year.

### C.1.4 Year-Based Results

The right branch in Figure C.1 contains the results compiled for each year. There are two types of data in this folder, indicated by the topmost folder, "vektor" and "statistik".

### C.1.5 Year-based Vectorized Data

"vektor" folder contains the compilation of the vectorised data for each year. Except for the size, they are equivalent to the image based vectorised data.

### C.1.6 Totaled and Averaged Information

The following data sets have been been calculated[2]:

**Anbauflaeche.tab:** The total acreage for the $5 \times 5\,\mathrm{km}^2$ squares obtained from the vectorized data.

**Anbauflaeche_pixel.tab:** The total acreage for the $5 \times 5\,\mathrm{km}^2$ squares obtained from the identified pixels.

**Beobachteteflaeche.tab:** The observable area, i.e., the area with satellite data available and not covered by clouds or cloud shadows.

**Beobachteteflaeche_clear.tab:** The observable area without also the haze-covered regions excluded.

**Anbaudichte_prozent.tab:** The cultivation density obtained from the vector-based results. The values have to be multiplied with 100 to obtain %.

**Anbaudichte_prozent_raster.tab:** Canola cultivation density obtained from the raster-based result. The values have to be multiplied with 100 to obtain %.

**Anbaudichte_prozent_raster_clear.tab:** Canola cultivation density obtained from the raster-based result with the exclusion of haze-covered regions. The values have to be multiplied with 100 to obtain %.

**Randflaeche.tab:** Area of neighbouring pixels in $\mathrm{m}^2$ obtained from the vectorized data.

**Randpixelflaeche.tab:** Area of neighbouring pixels in $\mathrm{m}^2$ obtained from the raster data.

**MittlerGroesse.tab:** Mean Field size in $\mathrm{m}^2$.

**MinimalerAbstand.tab:** Mean Minimum distance of the fields in m.

**Kantenlaengenverhaeltnis.tab:** Ratio of long and short principal axes: 1 indicates a quadratic field and smaller values a longish one.

**LaengederFeldgrenzen.tab:** Sum of all edges of the rectangles as an estimatinon of the field border length.

**MittlereAnbauFlaecheFehler.tab:** Fehler für die mittlere Anbaufläche (Anzahl der Randpixel pro Fläche).

**Mittlereorientierung.tab:** Mean orientation with repect to the North.

**MittlereMininmaleEntferung.tab:** Mean minimum distance, i.e., the averaged distance between each possible combination of fields in the square.

[2]Since the majority of users are german, the file have been named in german.

**rechtswerte.tab:** Easting of the upper left corner of the $5 \times 5\,\mathrm{km}^2$ square.

**Hochwerte.tab:** Northing of the upper left corner of the $5 \times 5\,\mathrm{km}^2$ square.

Values of -1 for relative values and of -9999.999 for absolute values indicate missing data.

There are two different representations for this data, indicated by the folder name. The file names are the same as listed above.

### Arrays

Array files are organised as matrix of values, with the rows indicating the values in the northing and the coloums in the easting direction. The position of each value can be obtained from the files "rechtswerte.tab" (easting) and "Hochwerte,tab" (northing).

### Lists

Lists are organies as lists with each row containing the postion and the value (see Table C.2).

Table C.2: Sturcuture of the list-files.

| Easting [m] (upper left corner) | Northing [m] (upper left corner) | size [$\mathrm{m}^2$] |
|---|---|---|
| 3.7456740000E06 | 5.7685760000E06 | 1.5390000000E05 |
| 3.7506740000E06 | 5.7685760000E06 | 2.1600000000E05 |
| 3.7556740000E06 | 5.7685760000E06 | 6.1290000000E05 |
| 3.7606740000E06 | 5.7685760000E06 | 6.9570000000E05 |
| 3.7656740000E06 | 5.7685760000E06 | 3.4650000000E05 |
| 3.7706740000E06 | 5.7685760000E06 | 5.9940000000E05 |
| 3.7756740000E06 | 5.7685760000E06 | 5.4000000000E04 |
| 3.7806740000E06 | 5.7685760000E06 | 0.0000000000E00 |
| 3.7856740000E06 | 5.7685760000E06 | 1.1700000000E04 |
| ⋮ | ⋮ | ⋮ |

# List of Acronyms

**ASTER** Advanced Spaceborne Thermal Emisssion and Radiation Radiometer

**ATKIS** *Amtliches Topographisch-Kartographisches Informationssystem* (Authoritative Topographic Cartographic Information System)

**BBA** *Biologische Bundesanstalt für Land- und Forstwirtschaft* (Federal Biological Research Centre for Agriculture and Forestry)

**BMBF** *Bundesministerium für Bildung und Forschung* (Federal Ministry of Education and Research)

**Canola** Canadian – Oil Low Acid

**CL** clear line

**CM** centre of mass

**DEM** digital elevation model

**DN** digital number

**DOS** dark object subtraction

**DSV** *Deutsche Saatveredelung* (German Seed Refinement)

**DVD** digital versatile disk

**ERS-SAR** European Remote Sensing Satellite-SAR

**ETM+** Enhanced Thematic Mapper+

**ETRS89** European Terrestrial Reference System 1989

**EU** European Union

**FOV** field of view

**GCP** ground control point

**GDR** German Democratic Republic

**GenEERA** *Generische Erfassung und Extrapolation der Rapsausbreitung* (Generic analysis and extrapolation of oilseed rape dispersal)

**GM** genetically modified

**GIS** geographical information system

**GOME** Global Ozone Monitoring Experiment

**GPS** Global Positioning System

**MODTRAN** MODerate resolution TRANsmittance

**HOT** haze optimized transform

**HRV** *Haute Résolution Visible*

**HRVIR** High Resolution Visible - Infrared

**IFOV** instantaneous field of view

**IRS** Indian Remote Sensing Satellite

**JERS-SAR** Japanese Earth Resource Satellite-SAR

**LISS/3** Linear Imaging Self Scanner/3

**MARS** monitoring agriculture with remote sensing

**MDC** Mahalanobis distance classifier

**MDS** Mahalanobis distance space

**MIR** middle infrared

**MLC** maximum likelyhood classifier

**MODIS** Moderate Resolution Imaging Spectroradiometer

**MODTRAN** Moderate Resolution Transmittance

**NASA** National Aeronautics and Space Administration

**NDVI** normalised difference vegetation index

**NIR** near infrared

**PAR** photosynthetically active radiation

**PIF** pseudo-invariant features

**PP** percentage points

**PPC** parallelepiped classifier

**PSF** point spread function

**RMS** root mean square

**SAR** synthetic aperture radar

**SCIAMACHY** Scanning Imaging Absorption Scatterometer for Atmospheric Cartography

**SPOT** *Système Pour l'Observation de la Terre*

**SRT** scale-rotate-translate

**SVD** singular value decomposition

**TIR** thermal infrared

**TM** Thematic Mapper

**UFT** *Zentrum für Umweltforschung und Umwelttechnologie* (Centre for Environmental Research and Environmental Technology)

**UTM** universal transverse Mercator

**VIS** visible

**ZALF** *Leibniz-Zentrum für Agrarlandschafts- und Landnutzungsforschung* (Leibniz-Center for Agricultural Landscape and Land Use Research)

# Acknowledgement

This work was funded by the BMBF and the University of Bremen. First, I wish to express my gratitude to my first examiner Prof. Dr. Klaus Künzi for his suggestions and his patience and to the second examiner, Dr. Broder Breckling from the UFT, also for his suggestions and patience but especially for the opportunity to participate in the multidisciplinary GenEERA project. In this context I would also like to thank Dr. Johannes Ranke, also a member of the UFT, since he was responsible for the first contact and consequently the cooperation between our institutes.

Moreover, I would like thank my colleagues and friends Dr. Christian Melsheimer, Dr. Ralf Schmidt, Dr. Georg Heygster, Christian Heidkamp, Dr. Jens Dannenberg, Dr. Norbert Schlüter, Sharon Kerr, Regina Krammer, Britta Beckmann and my brother, Dr. Carsten Laue for their suggestions and their helpful proofreading of this thesis.

A great help were the colleagues from the ZALF in Müncheberg, Dr. Michael Glemnitz, Dr. Angelika Wurbs, Dr. Uwe Heinrich and Bettina Funke who provided excellent evaluation data (the Quillow mapping data set). Moreover, I would like to thank them and Dr. Martin Wegehenkel for give me the opportunity to process the satellite images from frame 193/023 at their institute since these were only available at the ZALF.

Moreover, I would like to thank the colleagues from the UFT: Gertrud Menzel for the interesting excursions into botany and agriculture in the neighbourhood of Bremen and especially for the insight in the secrets of rapeseed and mustard fields. Carsten Borowy for his "brave" selection of the great amount of GCPs and Andreas Born for the help with the GIS data formats and suggestions on my work. All three of them provided great ground truth data with the mapping of the interbreeding partners, which I am also grateful of.

Ground truth was also available from seed-producing fields which were kindly visited with me by Mr. Wisloh from the local branch of the *Deutsche Saatveredelung* in Thedinghausen whom I also like to thank very much. Gratitude also to Mr. Meseke and Mr. Baar from the experimental farm of the BBA in Sickte who looked up the position of the experimental fields on which canola was grown with me.

Additional thanks to Dr. Ulrike Middelhoff from the Ecology Centre at the University of Kiel and Dr. Gunther Schmidt from the University of Vechta for the flowering periods of canola and their patience with the classification

results.

Special gratitude also to the people always helpful with administration and computers problems: Birgit Teuchert, Sabine Packeiser, Heiko Schellhorn and Claudia Vormann.

Moreover, I would like to thank my current and former colleagues of the IUP: Dr. Lars Kaleschke, Hong Gang, Gunnar Spreen, Dr. Nathalie Selbach, Nizy Mathew and Peter Mills for being such nice colleagues.

Additionally, I would like to thank all my friends not allready mentioned above: Ronald Bormann, Ines Koenen, Claudia Wienberg, Connie Eisenach, and Alex Drögmöller.

At last and most of all, I would like to thank my parents, Henry und Marita Laue, for their support and encouragement.

# Bibliography


Alados, I., Foyo-Moreno, I. and Alados-Arboledas, L., 1996: Photosynthetically active radiation: measurements and modelling. *Agricultural and Forest Meteorology*, **78**, 121–131.

Allen, J. D., 1990: A look at the Remote Sensing Applications Program of the National Agricultural Statistics Service. *Journal of Official Statistics*, **6**, 4, 393–409.

Asrar, G. (Ed.), 1989: *Theory and Applications of optical Remote Sensing.* John Wiley & Sons.

Bellow, M. and Ozga, M., 1991: Evaluation of clustering techniques for crop area estimation using remotely sensed data. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Atlanta, GA-USA: American Statistical Association.

Blackburn, G., 1998: Quantifying Chlorophylls and Carotenoids at Leaf and Canopy Scales: An Evaluation of Some Hyperspectral Approaches. *Remote Sensing of Environment*, **66**, 273–285.

Blackburn, G., 1999: Relationship between Spectral Reflectance and Pigment Concentrations in Stacks of Deciduous Broadleaves. *Remote Sensing of Environment*, **70**, 224–237.

Blaes, X., Holeck, F. and Defourny, P., 2001: Potential contribution of ENVISAT ASAR Alternating Polarisation and Wide-Swath modes images for crop discrimination at the regional scale. In *Proceedings of the 3rd International Symposium on "Retrieval of Bio- and Geophysical Parameters from SAR Data for Land Applications"*, University of Sheffield, UK: SCEOS,.

Breckling, B., Middelhoff, U., Borgmann, P., Menzel, G., Brauner, R., Born, A., Laue, H., Schmidt, G., Schröder, W., Wurbs, A. and Glemnitz, M., 2003: Biologische Risikoforschung zu gentechnisch veränderten Pflanzen in der Landwirtschaft: Das Beispiel Raps in Norddeutschland. In *GfÖ Arbeitskreis Theorie in der Ökologie:Gene, Bits und Ökosysteme.*, P. Lang Verlag, Frankfurt/Main.

Castleman, K. R., 1996: *Digital Image Processing.* 1st edn., Prentice Hall International, Inc.

Chavez, P. S. J., 1996: Image-based atmospheric corrections – Revisited and improved. *Photogrammetik Engineering & Remote Sensing*, **62**, 9, 1025–1036.

Cihlar, J., Guindon, B., Beaubien, J., Latifovic, R., Peddle, D., Wulder, M., Fernandes, R. and Kerr, J., 2003: From need to product: A methodology for completing a land cover map of Canada with LANDSAT TM. *Canadian Journal of Remote Sensing*, **29**, 2, 171–186.

Cihlar, J., Latifovic, R., Chen, J., Beaubien, J., Li, Z. and Magnussen, S., 2000: Selecting Representative High Resolution Sample Images for Land Cover Studies. Part 2. *Remote Sensing of Environment*, **72**, 2, 127–138.

Colwell, R. N. (Ed.), 1983: *Manual of Remote Sensing*, vol. 1: Theory, Instruments and Techniques. Falls Church, VA: American Society of Photogrammetry.

Cracknell, A. P., 1998: Synergy in remote sensing - What's in a pixel? *International Journal on Remote Sensing*, **19**, 11, 2025–2047.

Cramer, N., 1990: *Raps: Züchtung - Anbau und Vermarktung von Körnerraps.* Verlag Euler Ulm.

Davenport, I., Wilkinson, M., Mason, D., Charters, Y., Jones, A., Allainguillaume, J., Butler, H. and Raybould, A., 2000: Quantifying gene movement from oilseed rape to its wild relatives using remote sensing. *International Journal on Remote Sensing*, **21**, 18, 3567–3573.

Dawson, T., Curran, P. and Plummer, S., 1998: LIBERTY - Modeling the effects of leaf biochemical concentration on reflectance spectra. *Remote Sensing of Environment*, **65**, 50–60.

Di Vittorio, A., 2002: An automated, dynamic threshold cloud-masking algorithm for daytime AVHRR images Over land. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 8, 1683–1694.

Du, Y., Teillet, P. M. and Ciihlar, J., 2002: Radiometric normalization of multitemporal high-resolution satellite images with quality control for land cover change detection. *Remote Sensing of Environment*, **82**, 123–134.

Elachi, C., 1987: *Introduction to the Physics and Techniques of Remote Sensing.* John Wiley & Sons, Inc.

Erbertseder, T., Tungalagsaikhan, P., Bittner, M., Meisner, R., Schroedter, M. and Dech, S., 1999: Towards an operational atmospheric correction for AVHRR land surface products. In *Proceedings of the IGARSS 99*, Hamburg, Germany.

Eurimage, 2001: Price List, European. `http://www.eurimage.com/products/docs/price_euro.pdf`.

Euromap, 2001: European Price List. `http://www.euromap.de/products/prod_001.html`.

Evans, C., Jones, R., Svalbe, I. and Berman, M., 2002: Segmentating Multispectral LANDSAT TM Images Into Field Units. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 5, 1054–1064.

Forster, B., 1984: Derivation of atmospheric correction procedures for LANDSAT MSS with particular reference to urban data. *International Journal on Remote Sensing*, **5**, 5, 799–817.

Ganapol, B., Johnson, L., Hammer, P., Hlavka, C. and Peterson, D., 1998: LEAFMOD: A new Within-Leaf Radiative Transfer Model. *Remote Sensing of Environment*, **63**, 182–193.

Garcia-Haro, F., Gilabert, M. and Melia, J., 1999: Extraction of Endmembers from Spectral Mixtures. *Remote Sensing of Environment*, **68**, 237–253.

Gates, D., Keegan, H., Schleter, J. and Weidner, V., 1965: Spectral properties of plants. *Applied Optics*, **4**, 1, 11–20.

Genovese, G., 2004: MARS-STAT. `http://agrifish.jrc.it/marsstat/default.htm`.

Gitelson, A. and Merzlyak, M., 1997: Remote estimation of chlorophyll content in higher plant leaves. *International Journal on Remote Sensing*, **18**, 12, 2691–2697.

Gonzalez, R. C. and Wintz, P., 1987: *Digital Image Processing*. 2nd edn., Addison-Wesley Publishing Company, Inc.

Gross, H. and Schott, J., 1998: Application of spectral mixture analysis and image fusion techniques for image sharpening. *Remote Sensing of Environment*, **63**, 85–94.

Guindon, B. and Zhang, Y., 2002: Robust haze reduction: An integral processing component in satellite-based land over mapping. In *Procedings of the Joint International Symposium on Geospatial Theory, Processing and Applications*, Ottawa.

Hall, D. O. and Rao, K. K., 1999: *Photosynthesis*. 6th edn., Cambridge University Press.

Hall, F., Strebel, D., Nickeson, J. and Goetz, S., 1991: Radiometric rectification: Toward a common radiometric response among multidate, multisensor images. *Remote Sensing of Environment*, **35**, 11–27.

Hansen, M., Franklin, S., Woudsma, C. and Peterson, M., 2001: Caribou habitat mapping and fragmentation analysis using LANDSAT MSS, TM, and GIS data in the North Columbia Mountains, British Columbia, Canada. *Remote Sensing of Environment*, **77**, 50–65.

Holben, B., Vermote, E., Kaufman, Y., Tanré, D. and Kalb, V., 1992: Aerosol retrieval over land from AVHRR data-application for atmospheric correction. *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 2, 212–222.

Huang, C., Townshend, J., Liang, S., Kalluri, S. and DeFries, R., 2002: Impact of sensor's point spread function on land cover characterization: Assessment and deconvolution. *Remote Sensing of Environment*, **80**, 203–212.

Irish, R. R., 2000: LANDSAT 7 science data user's handbook. `http://lptpwww.gsfc.nasa.gov/IAS/handbook/handbook_toc.html`.

Janssen, L. and Molenaar, M., 1995: Terrain objects, their dynamics and their monitoring by integration of GIS and remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 3, 749–758.

Kalyanaraman, S., Rajangam, R. and Rattan, R., 1995: Indian remote sensing spacecraft 1C/1D. *International Journal on Remote Sensing*, **16**, 5, 791–799.

Kerkes, J. and Baum, J., 2002: Spectral imaging system analytical model for subpixel object detection. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 5, 1088–1101.

Kramer, H. J., 1996: *Observation of the Earth and Its Environment.* Berlin: Springer-Verlag.

Le Moigne, J. and Tilton, J. C., 1995: Refining Image Segmentation by Integration of Edge and Region Data. *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 3, 605–614.

Leo, O., 2004: MARS-PAC. `http://agrifish.jrc.it/marspac/olivine/default.htm`.

Liang, S., Fang, H. and Chen, M., 2001: Amospheric correction of LANDSAT ETM+ land surface imagery I: Methods. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 11, 2490–3498.

Liang, S., Fang, H., Morisette, J. T., Chen, M., Shuey, C. J., Walthall, C. L. and Daughtry, C. S. T., 2002: Atmospheric correction of LANDSAT ETM+ land surface imagery: II. Validation and applications. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 12, 2736–2746.

Lillesand, T., 2000: *Remote Sensing and Image Interpretation.* 4th edn., John Wiley & Sons.

Lopès, R., Fjørtoft, R. and Ducroft, D., 1999: *Edge detection and segmentation of SAR images in homogeneous regions*, chap. SAR Image Processing and Segmentation. World Scientific Publishing Co., 139–166.

Martonchik, J., 1994: Retrieval of Surface Directional Reflectance Properties Using Ground Level Multiangel Measurements. *Remote Sensing of Environment*, **50**, 303–316.

Menzel, G., Breckling, B. and Filser, J., 2003: *Monitoring der Umweltwirkungen transgener Kulturpflanzen in Bremen und im Bremer Umland*. Final report, UFT University of Bremen, Leobener Str.; D-28334 Bremen.

Michelson, D., Liljeberg, B. and Pilesjoe, P., 2000: Comparison of algorithms for classifying Swedish landcover using LANDSAT TM and ERS-1 SAR data. *Remote Sensing of Environment*, **71**, 1–15.

Mölders, N., Laube, M. and Raschke, E., 1995: Evaluation of model generated cloud cover by means of satellite data. *Atmospheric Research*, **39**, 91–111.

Moran, M., Bryant, R., Thome, K., Ni, W., Nouvellon, Y., Gonzales-Dugo, M., Qi, J. and Clarke, T., 2001: A refined empirical line approach for reflectance factor retrieval from LANDSAT-5 TM and LANDSAT-7 ETM+. *Remote Sensing of Environment*, **78**, 71–82.

Moran, S., Jackson, R., Slater, P. and Teillet, P., 1992: Evaluation of simplified procedures for retrieval of land surface reflectance factors from satellite sensor output. *Remote Sensing of Environment*, **41**, 169–184.

Myneni, R., Nemani, R. and Running, S., 1997: Estimation of Global Leaf Area Index and absorbed PAR using radiative transfer models. *IEEE Transactions on Geoscience and Remote Sensing*, **35**, 6, 1380–1393.

Müschen, B., Flügel, W., Hochschild, V., Steinnocher, K. and Quiel, F., 2001: Spectral and spatial classification methods in ARSGISIP project. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, **26**, 7-8, 613–616.

Ouaidrari, H. and Vermote, E., 1999: Operational atmospheric correction of LANDSAT TM data. *Remote Sensing of Environment*, **70**, 1, 1–127.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P., 1992: *Numerical Recipes in C: The Art of Scientific Computing*. 2nd edn., Press Syndicate of the University of Cambridge.

Price, J., 1994: How Unique Are Spectral Signatures. *Remote Sensing of Environment*, **49**, 181–186.

Ren, H. and Chang, C.-I., 2002: A Generalized Orthogonal Subspace Projection Approach to Unsupervised Multispectal Image Classification. *ieeetrgrs*, **38**, 6.

Richards, J., 1986: *Remote Sensing Digital Image Analysis*. Springer Verlag Berlin Heidelberg.

Richardson, A., Wiegand, C., Wanjura, D., Dusek, D. and Steiner, J., 1992: Multisite analysis of spectral-biophysical data for sorghum. *Remote Sensing of Environment*, **41**, 71–82.

Richter, R., 1996: A spatially fast atmospheric correction algorithm. *International Journal on Remote Sensing*, **17**, 6, 1201–1214.

Richter, R., 1997: Correction of atmospheric and topographic effects for high spatial resolution satellite imagery. *International Journal on Remote Sensing*, **18**, 5, 1099–1111.

Richter, R. and Lüdeker, W., 1999: Atmospheric water vapour retrieval from MOS-B imagery. *International Journal on Remote Sensing*, **20**, 6, 1133–1140.

Rieger, M. A., Lamond, M., Preston, C., Powles, S. B. and Roush, R. T., 2002: Pollen-mediated movement of herbicide resistance between commercial canola fields. *Science*, **296**, 2386–2388.

Rudorff, B. and Batista, G., 1990: Spectral Response of Wheat and Its Relationship to Agronomic Variables in the Tropical Region. *Remote Sensing of Environment*, **31**, 53–63.

Sarkar, A., Manoj Kumar Biswas, B. Kartikeyaqn, Vikash Kumar, K. L. Majumder and D. K. Pal, 2002: A MRF model-based segmentation approach to classification for multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, 5, 1102–1113.

Saunders, R. W. and Kriebel, K. T., 1988: An improved method for detecting clear sky and cloudy radiances from AVHRR data. *International Journal on Remote Sensing*, **9**, 1, 123–150.

Schimmeck, T., 2002: Feldzug im Grünen. *Die Zeit*, **39**.

Schlink, S., 1994: *Ökologie der Keimung und Dormanz von Körnerraps (Brassica napus L.) und ihre Bedeutung für eine Überdauerung im Boden*. Ph.D. thesis, Universität Göttingen.

Schopfer, P. and Brennicke, A., 1999: *Pflanzenphysiologie*. 6th edn., Springer-Verlag Berlin Heidelberg New York.

Schowengerdt, R. A., 1997: *Remote Sensing: Models and Methods for Image Processing*. 2nd edn., Academic Press.

Short, N., 2003: Remote sensing tutorial. `http:\\rst.gsfc.nasa.gov`.

Simon, B., 2004: Monsanto wins patent case on plant genes. *New York Times*.

Simpson, J., Jin, Z. and Stitt, J., 2000: Cloud shadow detection under arbitrary viewing and illumination conditions. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 2, 972–976.

Simpson, J. and Stitt, J., 1998: A procedure for the detection and removal of cloud shadow from AVHRR data over land. *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 3, 880–897.

Sims, D. A. and Gamon, J. A., 2002: Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and development stages. *Remote Sensing of Environment*, **81**, 337–354.

Smith, G. M. and Fuller, R. M., 2001: An integrated approach to land cover classification: an example in the Island of Jersey. *International Journal on Remote Sensing*, **22**, 16, 31–23–3142.

Snyder, J. P., 1984: *Map Projections Used by the U.S. Geolocgical Survey.* Washington: United States Gouvernment Printing Office.

Song, C., Woodcock, C., Karen C. Seto, Lennney, M. and Macomber, S., 2001: Classification and change detection using LANDSAT TM data: when and how to correct atmospheric effects? *Remote Sensing of Environment*, **75**, 230–244.

Space Imaging Eurasia, 2003: Price List of IKONOS Products. `http://www.sieurasia.com/eng/pricelist.pdf`.

Statistisches Bundesamt Deutschland, 2002: Struktuerhebungen in land- und forstwirtschaftlichen Betrieben. `http://www.destatis.de/basis/d/forst/forsttxt.htm`.

Stowe, I. L., McClain, E. P., Carey, R., Pellegrino, P., Gutmann, G. G., Davis, P., Long, C. and Hart, S., 1991: Global Distribution of Cloud Cover Derived from NOAA/AVHRR Operational Satellite Data. *Advances in Space Research*, **20**, 3673–3684.

Tanré, D., Holben, B. and Kaufman, Y., 1992: Atmospheric correction against algorithm for NOAA-AVHRR products: theory and application. *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 2, 231–248.

Thome, K., 2001: Absolute radiometric calibration of LANDSAT 7 ETM+ using the reflectance-based model. *Remote Sensing of Environment*, **78**, 27–38.

Toll, D., Shirey, D. and Kimes, D., 1997: NOAA AVHRR land surface albedo algorithm development. *International Journal on Remote Sensing*, **18**, 18, 3761–3796. Surface albedo plant cover.

Townshend, J. R. G., Hunag, C., Kalluri, S. N. V., Defires, R. S. and Liang, S., 2000: Beware of per-pixel characterization of land cover. *International Journal on Remote Sensing*, **21**, 4, 839–843.

Tucker, C., 1985: African land-cover classification using satellite images. *Science*, **227**, 369–375.

U.S. Geological Survey, 2003a: LANDSAT 7 SLC Anomaly Investigation. `http://landsat7.usgs.gov/updates.php`.

U.S. Geological Survey, 2003b: USGS LANDSAT project website: levels of processing. `http://landsat7.usgs.gov/l7_processlevels.html`.

Vincent, R. K., 1997: *Fundamentals of Geological and Environmental Remote Sensing*. Prentice Hall, Inc.

Wen, G., Cahalan, R., Tsay, S. and Oreopoulos, L., 2001: Impact of cumulus cloud spacing on LANDSAT atmospheric correction and aerosol retrieval. *Journal of Geophysical Research*, **106 1**, D11, 2129–12 138.

Wilkinson, M. J., Davenport, I. J., Charters, Y. M., Jones, A. E., Allainguillaume, J., Butler, H. T., Mason, D. C. and Raybould, A. F., 2000: A direct regional scale estimate of transgene movement from genetically modified oilseed rape to its wild progenitors. *Molecular Ecology*, **9**, 7, 831–1011.

Wright, G. G., 1985: Distribution and area of winter oilseed rape within eastern Scotland: a survey based on LANDSAT data. *Research and Development in Agriculture*, **2**, 1, 41–45.

Wright, G. G., 1994: The Application of Satellite Remote Sensing and Spatial Proximity Analysis Techniques to Observations on the Grazing of Oilseed Rape by Roe Deer. *International Journal on Remote Sensing*, **15**, 10, 2087–2097.

Wrigley, R., Spanner, M., Slye, R., Pueschel, R. and Aggarwal, H., 1992: Atmospheric Correction of Remotely Sensed Image Data by a Simplified Model. *Journal of Geophysical Research*, **97**, D17, 18 797–18 814.

Zhang, Y., Guindon, B. and Cihlar, J., 2002a: Development of a Robust Haze Removal Algorithm: Assessment Using Temporally Invariant Targets. In *Proceeding of the IGARSS*.

Zhang, Y., Guindon, B. and Cihlar, J., 2002b: An image transform to characterize and compensate for spatial variations in thin cloud contamination of LANDSAT images. *Remote Sensing of Environment*, **82**, 173–187.

Zhao, W., Tamura, M. and Takashi, H., 2000: Atmospheric and spectral corrections for estimating surface albedo from satellite data using 6S code. *Remote Sensing of Environment*, **76**, 202–212.

Zwiggelaar, R., 1998: A review of spectral properties of plants and their potential use for crop/weed discrimination in row-crops. *Crop Protection*, **17**, 3, 189–296.