



TAMPERE UNIVERSITY OF TECHNOLOGY

TEO KANNIAINEN

FEATURE EXTRACTION AND CLASSIFICATION OF THE  
FOREWINGS OF THREE MOTH SPECIES BASED ON DIGITAL  
IMAGES

Licentiate thesis

Examiners: professor Jouko Halttunen  
and PhD Irene Vänninen  
Examiners and topic approved in the  
Council meeting of Faculty of  
Automation, Mechanical and Materials  
Engineering on 5 October 2011

# Tiivistelmä

TAMPEREEN TEKNILLINEN YLIOPISTO

**KANNIAINEN, TEO:** Feature extraction and classification of the forewings of three moth species based on digital images

Lisensiaatintutkimus, 59 sivua, 5 liitesivua

Joulukuu 2011

Automaatio-, kone- ja materiaalitekniikan tiedekunta

Systeemitekniikan laitos

Tarkastajat: Professori Jouko Halttunen ja FT Irene Vänninen

Avainsanat: Lepidoptera, *Cydia*, omenakääriäinen, etusiipi, piirreirrotus, digitaalikuva, askeltava regressioanalyysi, lineaarinen regressioanalyysi, hierarkinen kokoava ryvästys

Tutkimuksessa haluttiin selvittää, voidaanko omenakääriäinen erottaa muista lähilajeista etusiivestä otetun digitaalikuvan avulla. Lisäksi haluttiin selvittää ne etusiiven alueet, joista lajitunnistus kannattaisi tehdä.

Tutkimuksessa käytettiin digitaalisia RGB-kuvia kolmen *Cydia*-lajin lajitunnistukseen. Tutkimukseen valittiin kohdelajiksi omenan tuholainen, *Cydia pomonella*, sitä ulkonäöltään läheisesti muistuttava *Cydia splendana* sekä näistä kahdesta ulkonäöltään selvästi erottuva *Cydia strobilella*. Etusiivistä valittiin alueet, joissa tekstipohjaisen tiedon perusteella sijaitsivat tyypilliset lajituntomerkit.

Tutkimukseen otetuista 12 etusiivestä määritettiin 6 aluetta, joista kaikista määritettiin 168 piirrettä. Piirteisiin kuului muun muassa paikallisia pikselikohtaisia intensiteettejä sekä suodatettuja mittaustuloksia.

Piirteiden määrän vähentämiseksi käytettiin askeltavan regressioanalyysin algoritmeja. Valittujen piirteiden perusteella muodostettiin lineaarisia malleja, jotka testattiin lineaarisella regressioanalyysillä ja hierarkisella kokoavalla ryvästyksellä.

Tutkimuksen perusteella *Cydia pomonella* –laji pystytään erottamaan kahdesta muusta *Cydia*-suvun lajista etusiivistä otettujen digitaalikuvien perusteella. Kaikkien kolmen *Cydia* –suvun lajin lajitunnistus oli luotettava, kun kuvien 6 tutkittua aluetta otettiin mukaan analyysiin. *Cydia pomonellan* etusiivet pystyttiin erottamaan kahdesta muusta *Cydia* –suvun lajin etusiivistä myös vain kolmen alueen perusteella. Lajitunnistus kannattaa tehdä siiven keskiosan juovikkaan alueen sekä siiven päädyssä olevan pronssinvärisen ovaalin alueen perusteella, mutta luultavasti ei siiven keskiosan tummanruskean viirun perusteella.

Erottelevimmat piirteet saatiin 21 x 21 ja 9 x 9 –kokoisilla suotimilla suodatetuista alueista, jotka selittivät paremmin lajien välistä eroa kuin pikselikohtaiset intensiteetit.

## Abstract

TAMPERE UNIVERSITY OF TECHNOLOGY

**KANNIAINEN, TEO:** Feature extraction and classification of the forewings of three moth species based on digital images

Licentiate Thesis, 59 pages, 5 Appendix pages

Month and year of thesis completion: December 2011

Faculty of Automation, Mechanical and Materials Engineering

Department of Automation Science and Engineering

Examiners: Professor Jouko Halttunen and PhD Irene Vänninen

Keywords: Lepidoptera, codling moth, *Cydia*, forewing, feature extraction, digital image, stepwise regression, linear regression, hierarchical agglomerative clustering

The main objective of this research was to find out the possibility to use digital images of forewings in the species identification of codling moth. The suitability of different areas of forewing as identification marks was also determined.

Digital RGB images were used to determine the features of forewings of three different *Cydia* species (Lepidoptera, Tortricidea). The chosen species were *Cydia pomonella*, *Cydia splendana* and *Cydia strobilella*. Text-based descriptions of the visual appearances of the moth species were used in feature selection. Image processing methods were applied on 6 different areas of 12 different forewings. 168 local features were calculated for each area. Features included direct pixel-wise intensity values and spatially filtered values.

Stepwise regression was performed in order to reduce the number of features in linear models. The models were tested with linear regression analysis and hierarchical agglomerative clustering.

Based on this research, *Cydia pomonella* can be identified from the two other *Cydia* species by forewing images. The identification was more reliable when the features of all 6 target areas were included compared to the case that the features of only 3 target areas were included. However, the forewings of *Cydia pomonella* were separated correctly from the forewings of two other *Cydia* species with 3 visible areas.

Identification of sitting *Cydia pomonella* can be based on the measured or calculated features in the white-brown veined area in the middle of the forewing and in the bronze coloured oval in the sub marginal area but possibly not in the dark brown stripe in the inner margin of the forewing. To have distinctive features in regression models, it is recommended to use 21 x 21 -sized or 9 x 9 -sized filtered values rather than direct pixel-wise measurements.

## **Preface**

The research presented in this thesis has been carried out at the Department of Automation Science and Engineering at Tampere University of Technology (TUT), Finland, during the year 2011. It is a pleasure to thank the people who have made this thesis possible.

First, I would like to express my gratitude to my supervisor, Professor Jouko Halttunen for his guidance and support. I am grateful for the possibility to study at the department and to get involved with this interesting field of research. I am especially grateful to my instructor, laboratory engineer Heimo Ihalainen, for his valuable contribution. I greatly appreciate the technical expertise and help from him. He had always time slots for me when needed. I thank researchers Mr Kalle Marjanen and Ms Marja Mettänen for their technical help.

I am thankful for another of the inspectors of this thesis, PhD Irene Vänninen, for her valuable contribution.

I would like to thank collection manager, MSc Jaakko Kullberg for his help to confirm the identification of the moths and possibility to take excellent pictures of the moths.

The grammar of this thesis was evaluated and corrected by Mr Niall O'Donoghue. I am grateful of that.

Finally, I would like to thank my wife Liisa and family for support and understanding. You enabled me to find time to write this thesis.

## Contents

Tiivistelmä.....	i
Abstract.....	ii
Preface.....	iii
Contents .....	iv
List of acronyms .....	vi
1 Introduction.....	1
1.1 The possibilities of digital image-based identification of harmful moths .....	1
1.2 General description of codling moth ( <i>Cydia pomonella</i> (L.)) and its influence on horticultural production in Finland .....	2
1.3 Monitoring and control of codling moth in horticultural production .....	3
1.3.1 Use of semiochemicals as attractants.....	3
1.3.2 Use of adhesive traps in monitoring .....	4
1.3.3 Control of codling moths population; mating disruption, sterile insect technique, use of insecticides, biological and microbial control.....	4
1.4 Morphology and appearance of forewings of <i>Cydia</i> -genus .....	5
1.4.1 Structure and texture .....	5
1.4.2 Geometric positioning and form of forewings.....	6
1.4.3 Ageing of wings .....	7
1.4.4 Visual differences of forewings .....	7
1.5 Use of 2-D digital colour images in the description of different object properties.....	8
1.5.1 Physical and mathematical background in image acquisition .....	8
1.5.2 Colour in digital images .....	9
1.5.3 Lighting of an object.....	10
1.5.4 Colour temperature and white balance .....	10
1.5.5 Optics in digital macrophotography .....	11
1.6 Methods used in feature extraction and classification of insects.....	12
2 Research problem and objectives .....	14
3 Material .....	15
4 Methods .....	17
4.1 Description of image acquisition tools and settings .....	17

4.2	Data preparation for analysis.....	18
4.3	Selection of key feature areas.....	20
4.4	Methods used in feature extraction .....	23
4.4.1	Measured features .....	23
4.4.2	Data sets.....	26
4.4.3	Normalization of data.....	26
4.4.4	Reduction the amount of features.....	26
4.4.5	Classification of data.....	28
5	Results .....	30
5.1	Data set 1 .....	30
5.2	Data set 2 .....	33
5.3	Data set 3 .....	36
5.4	Data set 4 .....	41
5.5	Data set 5 .....	44
5.6	Data set 6 .....	47
6	Discussion .....	52
6.1	Properties of the Camera and optics.....	52
6.2	Effect of the lighting on the reflected image.....	52
6.3	The amount and quality of forewings .....	52
6.4	Orientation of forewings .....	53
6.5	Normalization .....	53
6.6	Feature selection.....	53
6.7	Data set 1 .....	54
6.8	Data set 2 .....	55
6.9	Data set 3 .....	55
6.10	Data set 4 .....	55
6.11	Data set 5 .....	56
6.12	Data set 6 .....	56
7	Conclusions.....	57
	References.....	58
	Appendix.....	i

## List of acronyms

RMSE	Root mean square error
<i>k</i> -means	Clustering method in which each observation belongs to the cluster with the nearest mean
STD or std	Standard deviation
SLR	Single-lens reflex
DoF	Depth of field
CoC	Circle of confusion
LED or led	Light emitting diode
USM	Ultrasonic motor
CMYK	Cyan, magenta, yellow, black (colour space defined by the CMYK Colour Model)
HSI	Hue, Saturation, Intensity (colour space defined by the HSI Colour Model)
RGB	Red, Green, Blue (colour space defined by the RGB Colour Model)
CIE L*a*b	Colour space specified by the International Commission on Illumination

## **1 Introduction**

### **1.1 The possibilities of digital image-based identification of harmful moths**

Digital images consist of definite measurements of reflected light intensities of objects and digital image processing methods are based on these measurements. Recently, the size and pixel densities of the semiconductive cells of cameras have increased. The processing time of digital data is remarkably faster and more reliable than it was 10 years ago. Small-sized camera modules are available at relatively cheap cost. Digital analyzing methods have developed for helping humans to recognize different things in the surrounding environment. Important applications in digital image processing have developed for the identification of humans, bacteria and other living organisms. In industrial engineering, several optical, computer-based applications have developed.

The recognizing of pests in horticultural production is often based on human visual recognition and identification. Pests are attracted to a certain trap and then identified and counted. The methods could be laborious and there is usually some uncertainty due to the human factor. Sometimes dangerous pests are recognized too late resulting in decreased yield. Modern high pixel-density cells with a high speed processor can help to recognize and even identify pests just on time, thereby reducing the amount of applied pesticides and timing plant protection actions right. There have been reports on automated identification of different insects, especially butterflies. However, for commercially important pests, automated digital image-based identification methods and systems have not been reportedly developed in Finland.

Several species of moths cause injury in open field horticultural production. One of the most monitored common pests in apple orchards is the codling moth. There are several text-based descriptions of the visual appearance of the codling moth as a basis for species identification. There have not been reports on attempts to describe the visible features of codling moth based on digital images.

Monitoring traps are not absolutely specific for catching a certain species. It is possible and in practice probable that other species will be caught too. There exist a couple of visually quite similar species which could be misidentified as codling moth and there are also visually totally different species caught by the same traps. It is therefore important to define the codling moths reliable

identification characteristics in order to automate the recognizing of it from other species.

## **1.2 General description of codling moth (*Cydia pomonella* (L.)) and its influence on horticultural production in Finland**

Codling moth is one of the most monitored and destructive insects in apple orchards because its larva causes serious injuries to apple fruits. Sometimes it is also observed to cause injury to other host plants, such as pear and plum. The most common scientific name of the codling moth is *Cydia pomonella* described by Linnaeus in 1758 but also *Laspeyresia pomonella* is used, probably because the genus *Cydia* was described later by Hübner (1825). The taxonomy of the moths is somewhat inconsistent because of the use of several taxonomical systems. The most common classification is that the *Cydia pomonella* belongs to the microlepidoptera's large family of Tortricidae (tortrix moths), which is composed of circa 9800 species worldwide, around 400 of them being found in Finland (Kullberg et al 2010, Baixeras et al 2010). The family of Tortricidae is divided into three subfamilies: Tortricinae, Olethreutinae and Chlidanotinae. Subfamily Olethreutinae consists of 6 tribes, which are common in the Northern hemisphere. One of those tribes is Grapholitini consisting of around 600 species, with circa 10 % of those being found in Finland. Tribe Grapholitini consist of 6 genera, of which the genus *Cydia* includes 28 species (found in Finland). Dedicated taxonomists identify the species accurately by microscope by looking at the shape and morphology of the genitals of the species. If, however, the moth individual is in good condition, it may be possible to identify the species within the genus *Cydia* visually based on their forewing features. Both male and female moths seem to have visually similar habitus, differing from each other on the basis of the microscopic structure of their antennal lobes and genitals. Studies on grape wine moths (*Lobesia botrana*) have shown that usually both sexes have similar responses to a host plant's odour, but different responses to sex pheromones (Masante-Roca et al 2002).

The distribution of a *Cydia pomonella* is quite large covering generally the Northern hemisphere. The life cycle of *Cydia pomonella* is quite well known. In Finland, it is usually overwinters as a fully developed larva and in early spring it develops into the pupa stage. Hatching time of the pupae depends on the heat summation of the growing period and can be predicted by calculating the heat summation. In Southern Finland, hatching of the pupas usually takes place in the end of May or in June, just after the flowering period of apples. During some warm growing periods, a second generation is also partially developed. The flying-time of an adult is in the evening twilight when the temperature is over 15 degrees Celsius. The adults mate and lay the eggs on leaves, young fruits or

small branches. The eggs hatch to larvae after 10 to 12 days after the peak of the flying time. Depending on the chemical plant protection agent, the plant protection is performed just after the flowering time or during the hatching of the eggs.

During the first few days, the larva walks around the fruit peel and tries to find a point where the peel is a little bit damaged or thinner than the surrounding areas. Very often the larva chooses the stalk or stalk-eye. When a thinner or damaged point is found, the larva penetrates the fruit and digs into the core of the fruit. The larva eats the fruit and seeds until it is fully developed. After the eating-period it leaves the fruit, crawls down into the top soil under leaf litter or the bark of the tree-trunk and weaves a cocoon as an outer cover to protect itself. The damaged fruit drops earlier than the undamaged fruit. Because of the injury, it is also unmarketable (figure 1).



Figure 1. Images of apple fruit affected by larva of *Cydia pomonella*.

### 1.3 Monitoring and control of codling moth in horticultural production

#### 1.3.1 Use of semiochemicals as attractants

Semiochemicals are organic compounds that can be associated with the communication of certain species or between species. Semiochemicals divide into several groups: pheromones, allomones, kairomones, attractants and repellents. Codling moth is monitored mostly by sex pheromones although kairomones are also found to be good attractants (Light et al 2001, Light and Knight 2005) Kairomones are for instance host-plant volatiles and attract both

males and females by odour. Sex pheromones are quite short-chain organic compounds, which carry a message that attracts the opposite sex to mate. The sex pheromones are released by adult female codling moths. Sex pheromone compositions of different moths are quite well known. The molecular weights of sex pheromones of the family of Tortricidae are approximately 200-300 g mol<sup>-1</sup>. The sex pheromone of codling moth consists of (E,E)-8,-10-Dodecadienal, (E,E)-8,-10-Dodecadien-1-ol, (Z,E)-8,-10-Dodecadien-1-ol and (E)-8-Dodecen-1-ol. All of these compounds also attract other species (Al-Sayed 2011). Those species are usually taxonomically nearby the target species but sometimes the attracted species can be quite different and taxonomically distant from the target species (Peltotalo and Tuovinen 1986). Some of those attracted species are not harmful to horticultural production but at least one is known to be harmful, *Pammene rhediella* (fruitlet mining tortrix), which is yet much smaller than codling moth and also coloured differently.

### **1.3.2 Use of adhesive traps in monitoring**

The monitoring of codling moth is usually based on evaluation of the number of flying adults. Adults are attracted to blue glue traps by volatile sex pheromones. The traps are placed in an apple orchard just before the flight of young moths. Young hatched male moths, attracted by sex pheromones, find their way to the trap. The adhesive glue sheet is placed in the trap. When the moth flies into the trap it will become stuck to the adhesive sheet. The amount of trapped moths will be used for evaluating the need for plant protection actions. Trapped insect recognition, identification and counting are made manually and are based on human vision.

The position of trapped insects in adhesive sheets varies a lot, because the insect can be caught on the sheet with any part of its body. Relatively big insects also can move on the sheet, releasing scales or legs before they die. The sheets also trap occasional flying insects, which are probably not attracted by the pheromones but are only seeking a suitable place to sit, relax or overnight.

### **1.3.3 Control of codling moths population; mating disruption, sterile insect technique, use of insecticides, biological and microbial control**

There are several common methods of protecting the yield against injury caused by codling moths. In Finland, the most common way is to monitor the amount of trapped moth adults, and if the threshold value is exceeded the sprayed chemical insecticide is applied. Another possibility is to try to disrupt the mating by spraying the sex pheromone or kairomone over a larger area in

the apple orchard. The idea of the disruption is that the males and females don't find each other in order to mate. As a method, mating disruption has been tested also with other species, for instance with the summer fruit tortrix moth (*Adoxophyes orana*) (Milli and de Kramer 1991).

The sterile insect technique involves the colonization and mass rearing of the moths, which are exposed to gamma radiation and thereby become sterilized. Furthermore, fertilization will be attempted by sterilized sperm and hence no descendants will develop. The method has been successfully tested by FAO and IAEA and, for instance, in South Africa (Addison 2005).

There has been increased interest in protecting apples from codling moth by biological and microbial methods. One protection method is to spread codling moths' natural parasites in the orchard. For instance there have been successful trials with *Trichogramma platneri* (minute parasitic wasp), which parasites the eggs of codling moth. Also a couple of other species from the genus *Braconidae*, *Ichneumonidae*, *Pteromalidae*, and *Tachinidae* have been reported to be parasites of codling moth. However, maintaining the high density parasitic species population has been reported to be difficult because the natural size of the population of codling moth is usually low, and even couple of moths can cause large injury to the yield. Adult moths can also migrate long distances but the plant protection actions can be applied only locally.

Several bacteria, viruses, fungi and nematodes have reportedly been used against codling moth. A granulosis virus has been reported to be an efficient controller of the codling moths' population (Ballard 2000). It has been identified in codling moths' larvae, then derived and developed into a commercial pathogen product. The advantage of this product is that it seems not to interfere with the natural control of other pests. Trials with a fungi *Beauveria bassiana* and a nematode *Neoplectana carpocapsae* have also been noticed to be promising for plant protection against codling moth.

## **1.4 Morphology and appearance of forewings of *Cydia*-genus**

### **1.4.1 Structure and texture**

The wing of a codling moth consists of two chitinous layers accompanied with tubular veins. The layers with veins constitute a net to which the colourful scales are attached. The scales consist of chitin and are outgrowths of the body wall. Some of the scales (usually in the margin areas) are developed into hairs. The function of the hairs is probably to soften and minimize the noise emitted during the flight. The back side of the wing usually has a less coloured pattern.

### 1.4.2 Geometric positioning and form of forewings

During flight, the wings are flexible. If looked from the side view, the wings move up and down drawing a figure-of-eight diagram. The vortex of a forewing is always the leading part of the flight drawing that loop and hind wings follow the same track as forewings. The wings are seldom oriented totally flat as they are in the case of museum collections. During flight they are mostly in concave- or convex-shaped form.

When a codling moth is in sitting position, it sets the overlapped forewings lengthwise over the hind wings. The backside view looks like semi-circle (figure 2). The sitting position of moths differs between species. Within families, the sitting position of the wings is usually uniform but between the subfamilies they differ. For instance, *Argyresthia conjugella's* (apple fruit moth) forewings do not much overlap and the apices of the wings are vertically against each other.



Figure 2. Image of a shape of the sitting codling moth.

### 1.4.3 Ageing of wings

The flying period of an adult codling moth is from two to three weeks. During that period, the wings erode. The colours fade the margins fray and break up, and some of the scales fall away. The moths cannot regenerate the damaged parts of the wings. This sets a challenge to image-based measurements.

### 1.4.4 Visual differences of forewings

The wing span of an adult codling moth is from 14 mm to 22 mm. The distinctive marks of the forewing are an oval-shaped large bronze colour spot, a dark brownish stripe in front of the spot, and a veined white-brownish area in front of the strip. However, the oval bronze spot is glossy when viewed in from a certain direction, whereas viewed from other directions it seems to reflect as a matt and yellow-brownish colour. Under indirect lighting, the deep bronze colour is not noticeable at all. The stripes or bands of white-brown area are to some extent spatially and randomly distributed, are often non-continuous, and the intensity of the background of the veined area varies. Hind wings are uniformly light-brownish in colour without any other special distinctive marks. The backside colour patterns of forewings and hind wings are uniformly pale brown or pale beige. There is no evidence, in genus *Cydia*, that there exists a visual difference between the forewings of the male and female.

Visually, the nearest species of the codling moth seems to be *Cydia splendana* (chestnut tortrix), which is about of the same size, and also carries the oval spot in its forewing. Furthermore, the general colouring seems to be close to the codling moths' colouring. Its host plant is oak (*Quercus sp.*) and its distribution is connected to the distribution of oaks, which overlaps the distribution of codling moths. The noticeable visual difference between codling moth and chestnut tortrix is in the costa of forewing. The chestnut tortrix's costa is more strongly white-brown striped containing four dark dashes between the bright white stripes. A similar strong distinguishable striping of the costa of the forewing is also a characteristic distinctive mark of many other common *Cydia* species, for instance *Cydia strobilella*, *Cydia compositella*, *Cydia cosmophorana* and *Cydia nigricana*.

## **1.5 Use of 2-D digital colour images in the description of different object properties**

### **1.5.1 Physical and mathematical background in image acquisition**

Digital images have been successfully used to describe different objects. In horticultural production, digital images are commercially used for the optical sorting of greenhouse-grown tomatoes and roses. In open field conditions, there have been attempts to recognize weeds in seed lines in the same way.

A digital 2-D image consists of a certain finite amount of usually small square elements called pixels. A pixel can be determined as the point in a matrix, where a row and column intersect. A pixel contains information on the place and the intensity, and it has a discrete mathematical behaviour. In practice, this means that every pixel contains a finite value. Usually the target object itself has a continuous behaviour.

Acquisition of the RGB image is usually done by light-sensitive semiconductor cells via optics. The most common cell type is the so-called Bayer matrix where different pixels sense different colour intensities. The pixels together form a regularly organized mosaic layer, where one fourth of the pixels sense a colour red, one fourth blue, and usually the two fourths green colour intensities of an image. On the surface of the cell, this four-pixel structure is repeated horizontally and vertically to the size of the cell. Based on the intensity values of these four pixels and their neighbouring pixels, the camera's processor interpolates the intensity values for all pixels and all colour layers so that every output pixel contains full colour and intensity information. Such images are called, for instance, TIFF images. Some camera models can also produce un-interpolated images, where all those four colours exist separately. Depending on the camera producer, the extension of the image file is, for instance, .crw and .cr2 for Canon, .x3f for Sigma, .raw and .rw2 for Panasonic, and .nef and .nrw for Nikon. RGB colour images can include more information than gray-scale images, because the object is expressed with three different intensity layers instead of one. It is possible to extract features based directly on one of these layers intensities or by transforming the RGB image to other colour spaces for instance the HSI or CMYK colour space.

Digitizing the spatial coordinate values is called sampling and digitizing the intensity values is called quantization. The size and the pixel density of the Bayer matrix affect the spatial resolution of the output image and quantization affects the amount of gray-levels. Because of data processing and the physical behaviour of semiconductors, the number of gray levels is usually set to an integer power of 2. The image of an object is acquired, sampled and quantized

resulting in the digitized image. Mathematically, the digitized image output consists of a matrix or matrixes with columns and rows. Each pixel in the matrix represents a local intensity value. For a gray level image, the matrix is two-dimensional and for RGB image, the matrix is three-dimensional.

Spatial filtering is a group of methods useful for performing mathematical operations directly on the pixels. Those methods include, for instance, smoothing, sharpening, edge detection and convolution.

In a frequency domain, periodic signals can be expressed as the sum of sines and cosines of different frequencies. This sum (Fourier series) can always be formed when the signal is periodic, but even many non-periodic functions can be expressed using integrals of sines and cosines and then multiplying by a weighting function if the signal is finite (Fourier transform). One of the most powerful properties of Fourier series and Fourier transform is that they can be reconstructed between domain changes without information loss. Fourier transform enables, for instance, efficient use of different low-pass, band-pass and high-pass filters as well as high-level image processing methods.

The most common modern digital cameras produce RGB colour images by the so-called Bayer matrix cell, where each pixel is sensitive for only one of three colours. Four pixels are divided to one red-sensitive pixel, two green-sensitive pixels and one blue-sensitive pixel.

### **1.5.2 Colour in digital images**

A typical RGB colour image consists of the mixture of the primary colour intensities of red, green and blue. Any of the three intensity layers can be used separately, because each of these colour intensity layers includes a two-dimensional chart (matrix) of pixels.

A colour space is a specified standard way to describe colours. In theory, many of the different colour spaces do not describe all possible colours, but a certain proportion of colours (gamut) which is sufficient for certain purposes of use. SLR cameras process images usually by an RGB colour space or a slightly modified RGB colour space. For printed media, the use of a CMYK colour space is dominant. The HSI colour space is related to the way a human describes the colours.

An RGB colour space is based on the Cartesian coordinate system which can be illustrated by an RGB colour cube. The corners of the cube consist of the primary colours red, green and blue and complement colours cyan, magenta

and yellow. The remaining two corners are black and white. The grey scale axis is set diagonally through the cube between the black and white corners.

Humans view and describe a colour object by hue, saturation and intensity (brightness). The HSI colour space is often used to extract hue or saturation from colour information. It is useful, for instance, in colour segmentation and target object extraction from an rest image. The RGB colour space conversion to the HSI colour space is not mathematically linear but computationally practicable. However, when intensities are low, the use of hue as a strong feature is questionable, because it is very sensitive to intensities near the black point (Gonzales and Woods 2002, p. 298).

Because reflected colour depends on many factors, such as the illumination source and angle, the exact colour measurement of an object is difficult. Reflected colour could be measured accurately with a spectroradiometer, which is quite an expensive device.

Because one colour could be reconstructed from more than only one combination of different colours, the reflected colour from a target object is not always unambiguous.

### **1.5.3 Lighting of an object**

Light source has a key role in photography. Reflected light is composed via the optics into the semiconductor cell of camera. Part of the light stream is absorbed into the object, part is reflected into the cell and part is reflected elsewhere. If the object is glossy, the reflection intensity is high and only a small amount of light is absorbed. In general, glossy objects are a challenge for photography. The glossy areas could be reflective only to a narrow spatial direction and are not necessarily noticed by the sensing cell. In laboratory circumstances, this problem can be minimized by an indirect light source. Reflection of an indirect light source can reduce the possible indication of glossiness or soften it remarkably.

### **1.5.4 Colour temperature and white balance**

Reflected light is sensed by a semiconductive cell. The lighting environment of the object affects the sensed colours. If the information on the lighting conditions is not known by the camera, the produced image can have unrealistic colours. The type of lighting source is adjusted in a camera automatically or manually by setting the colour temperature. Colour temperature describes the colour spectrum radiated from "Planck's black body" in certain

surface temperatures. When a black body's surface temperature rises, the colour temperature also rises. Different light sources have different colour temperatures, which have to be known in order to have realistic colours to image. Colour temperature is set in a camera by adjusting the white balance to the correct value. Typically a digital camera's colour temperature adjustment ranges from candle light 1000 K ("warm"), via typical day light 5200 K ("neutral"), to the shade light 7000 K ("cold"). Another possibility to adjust colour temperature is the use of colour correction cards.

### **1.5.5 Optics in digital macrophotography**

Macrophotography is a type of photography, which is concentrating on close-up objects. It is widely used in insect photography because of its ability to have well qualified photos with high magnification. Macrophotography refers to a finished photograph of a subject at greater than life size. The ratio of the subject size on the image sensor plane to the actual subject size is known as the reproduction ratio. In macrophotography, reproduction ratios are greater than 1:1, although it rarely exceeds 1:1. Another property of macrophotography is focusing distance, which could be from several millimetres to infinite.

In macrophotography, the depth of field (DoF) is extremely low because of the limitations of the optics. This problem could be solved by focus stacking programs, but for measurement purposes the method is doubtful because of the changing focus distances. Depth of field is normally increased by stopping down aperture, but beyond a certain point, stopping down causes blurring due to diffraction, which counteracts the benefit of being in focus. In practice, 8 or 11 are often shown to be smallest acceptable apertures. The depth of field could be measured when focal length, aperture and circle of confusion (CoC) are determined. The circle of confusion is the part of an image that is acceptably sharp. For example, for a Canon 550D with a 60 mm focal length, the CoC is 0,019 mm. There are software based, online depth of field calculators available in the internet, for instance at [www.dofmaster.com](http://www.dofmaster.com).

Pixel density and cell size determine the output spatial resolution of the image. A large cell with a small aperture causes diffraction. Therefore aperture has to be bigger for large cells to avoid extra noise caused by diffraction. The advantage of big aperture is that it also enables a faster shutter speed and a less noisy image. On the other hand, long focal length and big aperture decreases the depth of field. The sensitivity dynamics of the cell affect the amount of noise induced to an image. A high S/N –ratio enables a faster shutter speed. However, by increasing the cell sensitivity (ISO value), usually the amount of noise also increases.

High quality, low-noise and high depth of field digital macro images are achieved by low ISO value, small focal length, small cell size with high pixel density and using medium size aperture. These properties exist in camera modules used, for instance, in mobile phones.

## **1.6 Methods used in feature extraction and classification of insects**

Image analysis is widely used in human recognition and identification applications. Some researchers have also concentrated on other insect and butterfly identification, however a rather small amount of publications exist of image analysis of moths.

The basic methods for image processing are well described in textbooks of Gonzales and Woods (2002), Nixon and Aguado (2006), Costa and Cesar (2001) and Dryden and Mardia (1998). In these books, the theoretical base of spatial and frequency domain filtering, shape detection, image segmentation, and object recognition, are widely explained with informative examples.

Many of the methods seem to lay on the basis of low-level local-feature measurements. Features usually consist of colour, histogram, form, place or other simply computed local descriptors. In recent years, several high-level classification methods have been developed. They usually rely on neural networks in classification. In these models, features include also texture descriptors or transformed information, such as wavelets or Fourier descriptors.

Delgado (2010) developed methods for classification of four stonefly species in water streams by local feature extraction. Local features were entered as descriptors into histograms and classified by a logistic model tree. Arbuckle et al (2001) used local feature extraction methods to identify bees. Local features were extracted from forewings and venation was identified. The geometric features of local areas were also calculated. The classification of features was based on Support Vector Machines and Kernel Discriminant Analysis. Pantofaru et al (2006) used region-based context features in the classification of a few butterflies. They also used texture parameters as features. Schmid et al (2004) used local invariant features in pattern recognition. They explained how to extract scale and affine-invariant regions and how to obtain discriminate descriptors for these areas. They used images of butterflies as examples.

Digital Automated Identification System (DAISY) is a system that is based on detection via eigen-images. The classifier is based on two methods: the random n-tuple classifier and plastic self organizing maps. DAISY is reported to be capable of handling hundreds of taxa with high efficiency. However, the method is interactive and needs user's guidance in order to work well.

Neural networks have been tested for the classification of spiders. The classification is based on the wavelet encoded images, and it uses images of spiders' external genitalia for classification. This method also needs user activity and is therefore not automated for practical purposes. Zhu and Zhang (2010) recognized some butterflies by integrated region matching and dual tree complex wavelet transform. They also used k-means algorithms for the classification of regions of interest.

Pun and Cong (2010) used contour simplification and tangent function in shape classification. They had examples of butterfly image shapes which were scaled, rotated and then described by tangent function.

There has been some success in developing place, shift, direction and rotation –invariant transforms. Generalized Hough Transform has been in some cases effective. Unfortunately, when this transform is performed, it usually increases noise to the model and the results become difficult to interpret (Nixon and Aguado 2006, p. 213).

## 2 Research problem and objectives

There are several text-based descriptions of the visual appearance of moths as a basis for species identification. There have not been reports of attempts to describe the visible features of *Cydia pomonella* based on digital images. Knowledge of the measured image-based properties of *Cydia pomonella* is necessary for computer-based recognizing of *Cydia pomonella* in a trap. It is not known if the text-based descriptions are suitable for feature extraction based on digital images. Therefore a description of *Cydia pomonella* based on digital images is needed.

The first objective of the research is to test whether the differences between forewing features of *Cydia pomonella*, *Cydia splendana* and *Cydia strobilella* could be recognized by digital image processing methods or not. The second objective is to identify the areas in the forewings that could be the basis for the recognition of the species. The third objective is to test a classification method with the forewing data of the three species.

### 3 Material

The basic data consists of photographs of genus *Cydia* moths found in Finland. 28 *Cydia* species were photographed in laboratory circumstances. Three of them were chosen for this study. Because the objective was to determine the features of the codling moth, it was the target species of the research. Another objective was to test if the distinguishable features seen in digital images of the codling moth differ from the species visually close to the codling moth. Therefore the chestnut tortrix was chosen for comparison. The third species *Cydia strobilella* was chosen because it is visually distinct from the codling moth and therefore provides data that differs visually distinctly from the codling moth.

The species were photographed in the Finnish Museum of Natural History. Each individual moth was selected from the National collection of Lepidoptera. Because there were a couple of individuals from each moth species, the photographed individuals were chosen as visually representative samples. Four individuals of *Cydia pomonella*, two of *Cydia splendana* and one of *Cydia strobilella* were photographed. The identification of the photographed individuals was confirmed by the collection manager, M.Sc. Jaakko Kullberg.

Because the target specie of this research was *Cydia pomonella*, fewer individuals were photographed for the other two species involved in this research. For statistically more reliable models, more photographs should be taken. The models presented in this research are based on four *Cydia pomonella*, two *Cydia splendana* and one *Cydia strobilella* individuals.

From these 7 photographs, 6 *Cydia pomonella* wings, 4 *Cydia splendana* wings and 2 *Cydia strobilella* wings were extracted for further data processing. Two of the 8 *Cydia pomonella* wings were excluded because these wings were either too outworn or there was a needle overlapped on the forewing image (figure 3).



Figure 3. Images of four *Cydia pomonella*, two *Cydia splendana* and one *Cydia strobilella* included in the study. Two excluded forewings are marked with a criss-cross.

## **4 Methods**

### **4.1 Description of image acquisition tools and settings**

The moths were photographed using a Canon 550D SLR camera equipped with an EF-S 60 mm F 2.8 Macro USM lens. The focal length 60 mm corresponds to 96 mm in full-frame 35 mm size. The shooting distance was 19 cm, which produced 1:1 images to the cell. The cell of the camera was a mid-size (22,3 mm × 14,9 mm, 5184 × 3456 pixels) CMOS Bayer matrix cell, one pixel corresponding  $4,3 \times 10^{-3} \text{ mm}^2$  and one mm corresponding to 232 pixels. In this research, all measurements were performed pixel-wise. Aperture was chosen as f/8, ISO speed 100, and exposure time 0.5 seconds. A tripod was used.

A special device for lighting was prepared (figure 4). It consists of a white half dome plastic diffuser, a circle led strip light source consisting of 45 single leds and having total power of 3 watts. The diffuser was lit by white leds. The colour temperature of the leds was 5500 K corresponding to nearly direct sunshine. A black paper cylinder was installed to isolate the target from light coming from the outside surroundings of the device.

The grey background of the moths was used as a white balance reference. The grey colour background was a pigment sheet, which was cut from QP Card 101 and CIE L\*a\*b values were 48\*0\*0. The white balance was corrected afterwards in the images.

The images were saved as .cr2 format and for data analysis were transformed using the Canon Digital Photo Professional program (version 3.8.0.0) into .tif format.



Figure 4. Images of the lighting device and its installation to the camera.

#### 4.2 Data preparation for analysis

Forewings were extracted visually by hand from the rest of the body of the imaged moths. The left forewings were transformed to “right forewings” by mirror image of the moth (figures 5, 6, 7 and 8). After that the forewings of each moth species were transformed to similar spatial place by spatial transformation. The spatial transform action was controlled by control point pairs from each wing (figure 9). From four to six similar control points were chosen from each of the moths. The spatial transform was based on nonreflective similarity which is used in cases when shapes in the input image are unchanged, but the image is possibly distorted by some combination of translation, rotation, and scaling. The transformation was applied separately to each species. Examples of prepared forewings are presented in figures 10 and 11.

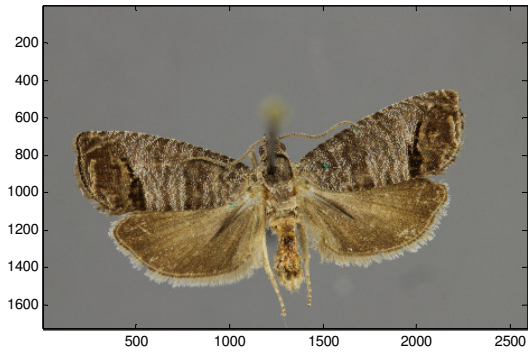


Figure 5. Original image of *Cydia pomonella*.

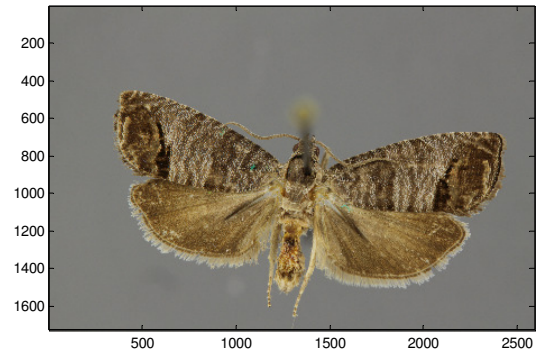


Figure 6. Mirror image of *Cydia pomonella*.

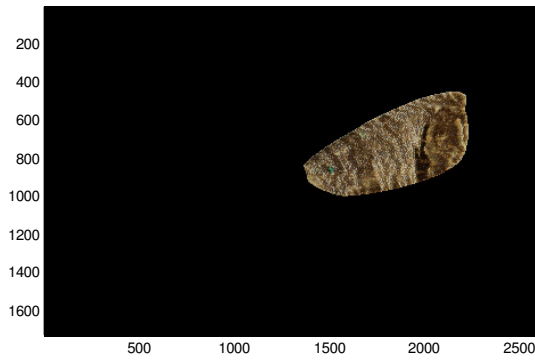


Figure 7. Wing 1 extracted from the original image.

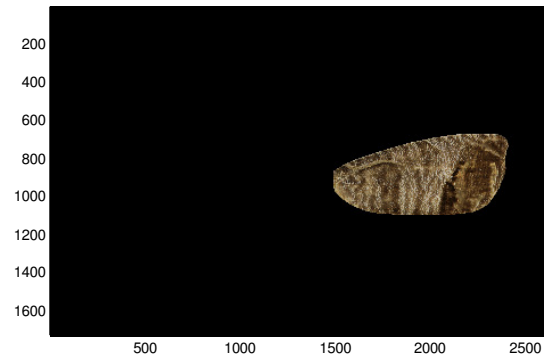


Figure 8. Wing 2 extracted from the mirror image.

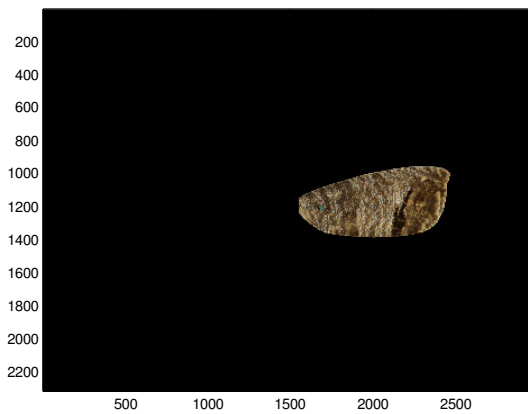


Figure 9. Wing 1 transformed to the same alignment as wing 2.

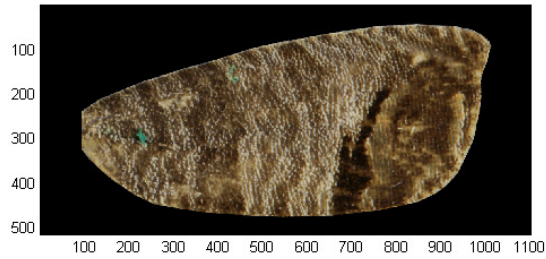


Figure 10. Wing 1; result of diminishing the black margins and transformed to defined size and place. This wing is ready for analysis.



Figure 11. Wing 2; result of diminishing the black margins and transformed to defined size and place. This wing is ready for analysis.

The original images consisted of intensities of red, green and blue channels. In order to have more colour information, HSI transform action was performed on the images resulting in hue, saturation and intensity channels.

### 4.3 Selection of key feature areas

Six feature areas were selected to the data analysis. The areas were selected visually by text-based distinctive identification marks of the moth species (figures 12, 13 and 14).

Area 1: Striped Costa (*Cydia splendana* and *Cydia strobilella*)

Area 2: Veined brown-white area in the middle of forewing (*Cydia pomonella*)

Area 3: Large dark brown spot (or stripe) in inner margin (*Cydia pomonella* and *Cydia splendana*)

Area 4: White spot in upper outer margin (*Cydia Strobilella*)

Area 5: White spot in lower outer margin (*Cydia Strobilella*)

Area 6: Oval bronze-coloured area in sub marginal area (*Cydia pomonella* and *Cydia splendana*)

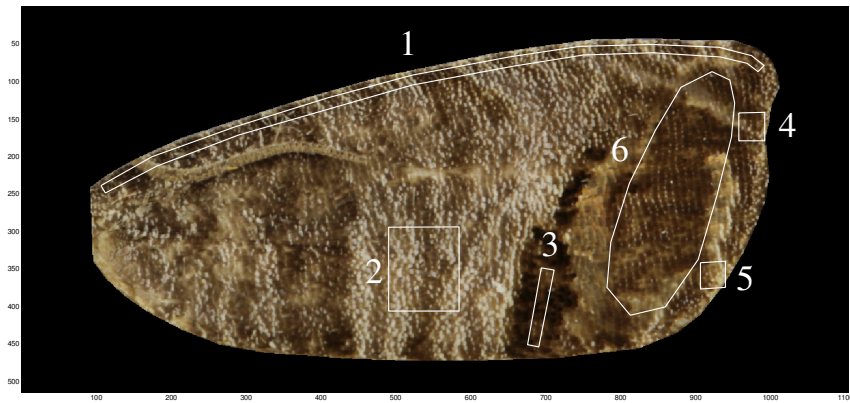


Figure 12. Image of *Cydia pomonella*'s wing and key feature areas.



Figure 13. Image of *Cydia splendana*'s wing and key feature areas.

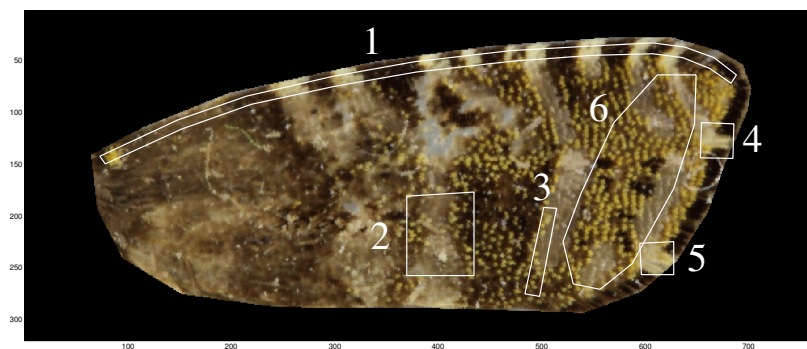


Figure 14. Image of *Cydia strobilella*'s wing and key feature areas.

Special interest was focused on the areas 2, 3 and 6. These areas can be fully viewed in photographs that are taken from above sitting moths (figure 15). Areas 1, 4 and 5 are not at all or are only partially visible from this view.



Figure 15. Images of four codling moths in the sitting position. The magnification of the images is different and they are therefore scale wise not comparable with each other.

#### 4.4 Methods used in feature extraction

The selected areas were measured for 42 local features using four channels resulting in 168 features per area per wing. Convolution and filtering with a standard deviation filter were measured for 4 different spatially oriented filters ( $0$ ,  $\pi/2$ ,  $\pi/4$  and  $-\pi/4$ ). The length of the convolution filter was set to 9 and 21 pixels. The length of the convolution filters was chosen visually to correspond roughly to the distances of the horizontal intensity maximums of the veined area of *Cydia pomonella*.

Local entropy as an example of texture feature was determined but there were no differences between wings or species. Therefore results of local entropy are not included in this research.

Saturation and hue channels were measured but the results are not published in this research. There were differences in hues between wings and species. However, the biggest differences in hues were found in the dark brown stripe area in which the intensities were low. Because hue is very sensitive to changes in low intensities, it was questionable to use it as a strong feature in this research.

##### 4.4.1 Measured features

The measured features were as follows:

Mean of

1. Pixel intensities
2. Local mean of 9 x 9 -sized filter
3. Local mean of 21 x 21 -sized filter
4. Local standard deviation of 9 x 9 -sized filter
5. Local standard deviation of 21 x 21 -sized filter
6. Values of convolution with  $0$ ,  $\pi/2$ ,  $\pi/4$  and  $-\pi/4$  radian direction of 9 pixel-sized line filter.
7. Values of convolution with  $0$ ,  $\pi/2$ ,  $\pi/4$  and  $-\pi/4$  radian direction of 21 pixel-sized line filter.
8. Values of filtering with 9 pixel-sized standard deviation line filter of  $0$ ,  $\pi/2$ ,  $\pi/4$  and  $-\pi/4$  radian direction
9. Values of filtering with 21 pixel-sized standard deviation line filter of  $0$ ,  $\pi/2$ ,  $\pi/4$  and  $-\pi/4$  radian direction

### Standard deviation of

1. Pixel intensities
2. Local mean of 9 x 9 -sized filter
3. Local mean of 21 x 21 -sized filter
4. Local standard deviation of 9 x 9 -sized filter
5. Local standard deviation of 21 x 21 -sized filter
6. Values of convolution with 0,  $\pi/2$ ,  $\pi/4$  and  $-\pi/4$  radian direction of 9 pixel-sized line filter.
7. Values of convolution with 0,  $\pi/2$ ,  $\pi/4$  and  $-\pi/4$  radian direction of 21 pixel-sized line filter.
8. Values of filtering with 9 pixel-sized standard deviation line filter of 0,  $\pi/2$ ,  $\pi/4$  and  $-\pi/4$  radian direction
9. Values of filtering with 21 pixel-sized standard deviation line filter of 0,  $\pi/2$ ,  $\pi/4$  and  $-\pi/4$  radian direction

All six areas were measured for red, green, blue and intensity channels separately. Totally, 1008 feature vectors were established for each forewing: 504 of the features were means and 504 of the features were standard deviations. Each defined area of the forewings was calculated for 168 features: 84 of the features were means and 84 of the features were standard deviations. A description of the filters is attached in appendix 1. Visual examples of measurements are presented in figures 16 and 17. Values in the tables are presented to an accuracy of four significant decimals.

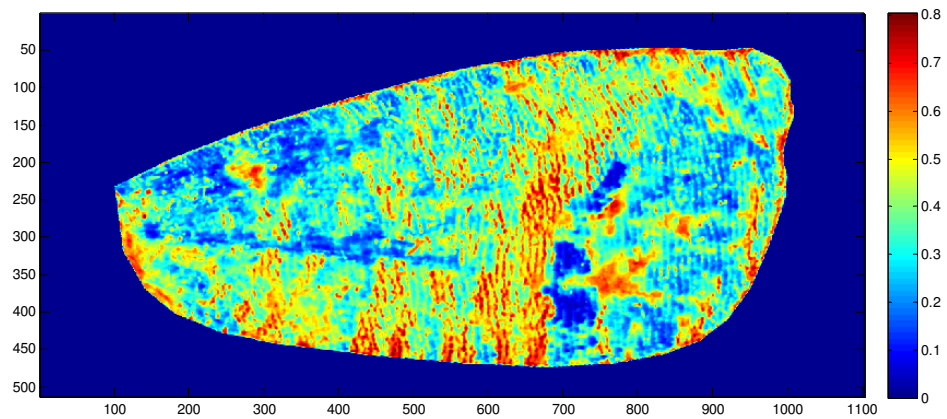


Figure 16. Example of red channel intensity of one forewing of *Cydia pomonella*.

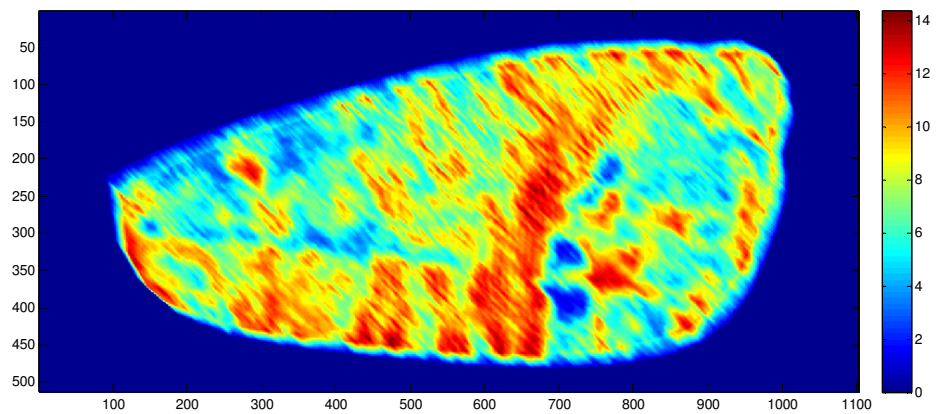


Figure 17. Example of forewing of *Cydia pomonella* convolved with  $-\pi/4$  radian-oriented 21 pixel-sized line filter. In this figure, convolution is performed to red channel intensities of the original image. Note the higher responses compared to values in figure 16.

#### 4.4.2 Data sets

The features were divided in 6 different data sets for analysis. The data sets and performed operations are presented in table 1.

Table 1. Different data sets and performed operations.

Data set	Areas included	Grouping	Features /Group	Stepwise regression performed	Features in model after stepwise regression	Hierarchical clustering performed
Data set 1	1 2 3 4 5 6	12 wings	1008	No	1008	yes
Data set 2	1 2 3 4 5 6	12 wings	1008	yes	10	no
Data set 3	1 2 3 4 5 6	3 species	1008	yes	10	yes
Data set 4	2 3 6	12 wings	504	yes	3	no
Data set 5	2 3 6	3 species	504	yes	1	yes
Data set 6	2 3 6	2 groups of species	504	yes	5	yes

#### 4.4.3 Normalization of data

The data was normalized within each feature vector. The mean of the values was removed and the resulting values divided by the standard deviation of the feature values. For data set 1, both normalized and non-normalized data was used.

#### 4.4.4 Reduction the amount of features

The data consisted of a large amount of features and there was redundant information between the features; some of the features correlated with each other strongly. Therefore a selection of the features according to visualization and classification was needed.

Multiple linear regression models are built from a potentially large number of predictive terms. The number of interaction terms, for example, increases exponentially with the number of predictor variables. For multiple-term models, it is possible to include redundant terms in a model that confuse the identification of significant effects. To solve these problems, a stepwise regression was performed on the data.

Stepwise regression is a systematic method for adding and removing terms from a multilinear model based on their statistical significance in a regression. The method begins with an initial model and then compares the explanatory power of incrementally larger and smaller models. In each step, the p-value of an F-statistic and the RMSE value are computed to test models with and without a potential term. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the term is added to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the term is removed from the model. The used stepwise regression algorithm proceeds as follows:

Initial terms will be fitted to the model. In this case, no terms (features) were included in advance.

Calculation of the F- and p-values of the data. If any terms (features) not in the model had p-values less than the entrance tolerance, the term with the smallest p-value was added to the model. This step was repeated as long as there existed terms which had p-values smaller than the entrance tolerance. The entrance value for p was set to 0.05. The first feature to be included into the model was the one having the smallest p-value and which, at the same time, decreased the RMSE value (Root Mean Square Error) most of all.

If any terms in the model had p-values greater than an exit tolerance, the term with the largest p-value was removed. The exit tolerance value for p was set to 0.10.

The algorithm stopped when there was no improvement in the model and the variation in the data was completely explained by linear combination of the features and the rest as RMSE value.

Depending on the terms included in the initial model and the order in which terms are moved in and out, the method could build different models from the same set of potential terms. There was no evidence, however, that a different initial model would have led to a better fit. In this sense, stepwise models were locally optimal.

The resulting reduced set of features was chosen by the stepwise regression algorithms in order to have a sufficient fit of the linear combination model. If the

grouping was made by species or species groups, resulting linear models were tested and entered to hierarchical agglomerative clustering analysis.

The chosen features were entered to linear models and the models were tested on linear regression analysis. In linear regression, the model specification was that the dependent variable was a linear combination of the feature values. Linear regression models (equations) were fitted using the least squares approach. The R squared –value indicated the explanatory power of the model to the predicted dependent variable. The root mean square error (RMSE) indicated the unexplained variation of data in the model.

The validation of the different linear regression models was evaluated only with the original data. Therefore the explanatory power of the models is concerning only this original data. For more reliable and general prediction, the models should be tested with other independent data sets. However, this was not done because of the lack of material.

#### **4.4.5 Classification of data**

The data sets 1, 3, 5 and 6 were tested with two clustering methods: *k*-means clustering and hierarchical agglomerative clustering. Because the *k*-means clustering method resulted in similar scores to the hierarchical agglomerative clustering, only the results of the latter method were published.

In the hierarchical agglomerative clustering method, several algorithms are used. In a hierarchical cluster tree, any two objects in the original data set were linked together at some level. The method with this feature vector data was as follows:

All feature vectors were determined as “initial clusters”.

The “initial clusters” were combined into pairs. The distance of different objects was calculated according to the Euclidean distance between all possible pairs. The result of this calculation was a distance matrix. The most proximity objects were combined as pairs based on their similarity.

The objects were paired into binary clusters.

The newly formed clusters were grouped into larger clusters by distance matrix until a hierarchical tree was formed.

The branches (smaller clusters) were connected together with links. The height of the link (linkage distance) represented the distance between the two clusters describing the consistency.

## 5 Results

### 5.1 Data set 1

Box plot figures were used to generally describe the levels and variety of the non-normalized feature values of the forewings (figures 18 and 19). On each box, the red band inside the box is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually with red cross. If the data is normally distributed, the single value is determined as an outlier if it is not included into 99,3 % coverage of data points (approximately  $\pm 2,7\sigma$ ). In this data set, an outlier is an outlier only by descriptive means; no data points were excluded because of the outlier status.

Most of the features had relatively small means and standard deviations. However, some of the features had rather high means and standard deviations, which resulted in remarkably skewed distributions of the data and made the interpretation difficult. The low values were the results from features, which were associated with the direct channel values of intensities between 0 and 1. The higher values were mainly associated with the features that were filtered outputs of different channels.

One feature consists of measurement or calculation of pixel intensities in defined area and channel. For example, the feature number 1 is a mean of intensity values of pixels in red channel in area 1 and the feature number 367 is a standard deviation of convolution response values in blue channel in area 2 (convolved with 9 pixel –sized line filter in  $+\pi/4$  radian direction).

The indexes in tables refer to the following data:

1. *Cydia pomonella*, wing 1
2. *Cydia pomonella*, wing 2
3. *Cydia pomonella*, wing 3
4. *Cydia pomonella*, wing 4
5. *Cydia pomonella*, wing 5
6. *Cydia pomonella*, wing 6
7. *Cydia splendana*, wing 1
8. *Cydia splendana*, wing 2
9. *Cydia splendana*, wing 3
10. *Cydia splendana*, wing 4
11. *Cydia strobilella*, wing 1
12. *Cydia strobilella*, wing 2

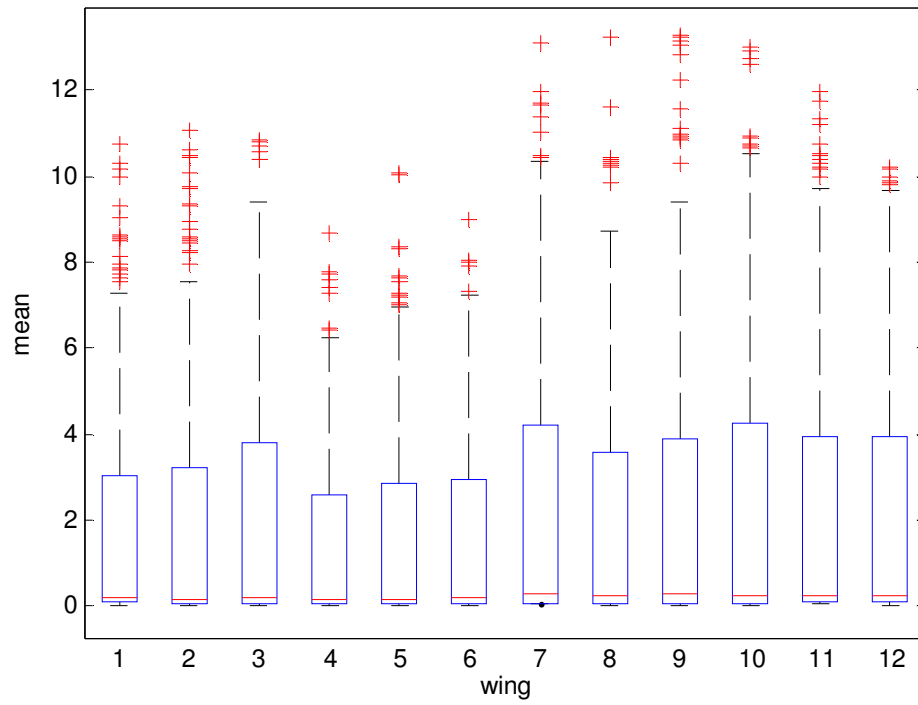


Figure 18. Means of non-normalized features. Each box contains 504 features including measured values and calculated responses.

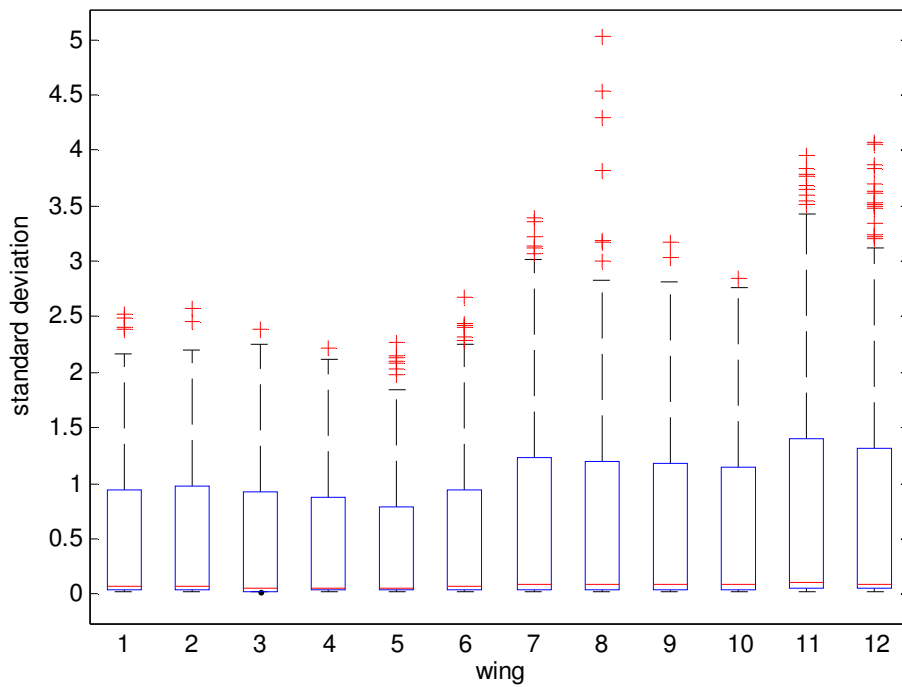


Figure 19. Standard deviations of non-normalized features. Each box contains 504 features including measured values and calculated responses.

All 1008 features were classified and clustered into 12 groups by hierarchical agglomerative clustering. The clustering tree for non-normalized data is presented in figure 20 and for normalized data in figure 21. The height of the link (linkage distance) represents the distance between the two clusters describing the consistency. The high linkage distance between links (clusters) indicates that they are quite far from each other (dissimilar).

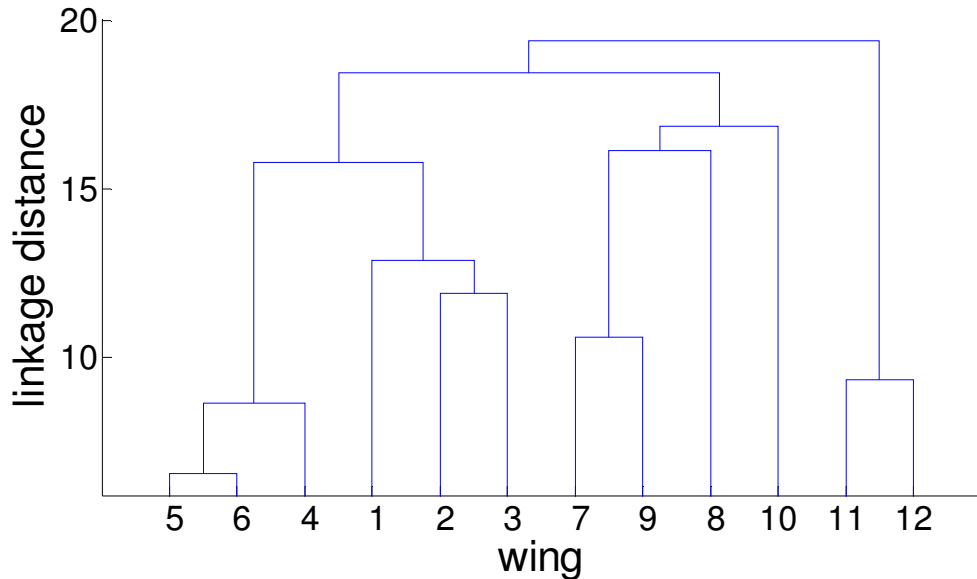


Figure 20. Hierarchical agglomerative clustering tree formed by all 1008 non-normalized features.

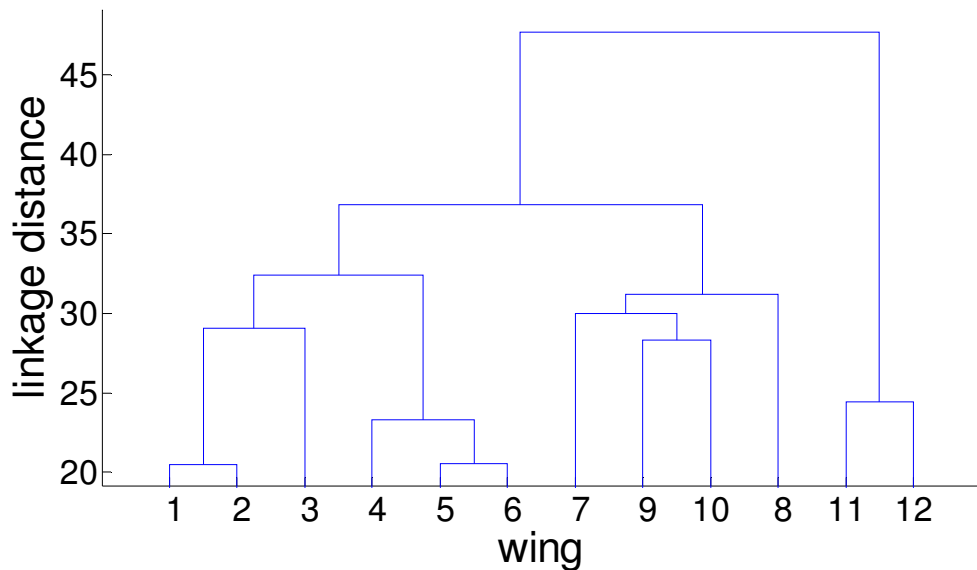


Figure 21. Hierarchical agglomerative clustering tree formed by all 1008 normalized features. Note the difference in linkage distances compared to non-normalized data in figure 20.

## 5.2 Data set 2

The data was divided into 12 groups; each group represented one wing. The data of all 6 areas were included. The features of models were chosen by stepwise regression algorithms. The first chosen feature decreased the explained variation in the model for the most part. The chosen features and feature values are presented in table 2 and in table 3. The RMSE and R squared values are presented as a function of the sum of included features in model (figure 21).

Table 2. Data set 2.

Feature	Area	Channel	Value	of Value	Filter	Orientation of filter
Feature 1	2	Intensity	Std	local std	21 x 21	no
Feature 2	1	Red	Std	local mean	21 x 21	no
Feature 3	4	Green	mean	local std	9 x 9	no
Feature 4	3	Blue	mean	std	9 x 9	+ $\pi/4$
Feature 5	2	Red	Std	local std	21 x 21	no
Feature 6	4	Blue	Std	std	9 x 9	- $\pi/4$
Feature 7	6	Intensity	mean	std	21 x 21	- $\pi/4$
Feature 8	4	Blue	Std	local std	9 x 9	no
Feature 9	4	Green	Std	std	21 x 21	vertical
Feature 10	5	Red	mean	convolution	21 x 21	- $\pi/4$

Table 3. Values of the features in Data set 2.

Group (wing)	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
1	0,0151	0,0751	0,0881	0,0222	0,0151	0,0451	0,0642	0,0383	0,0422	9,3189
2	0,0152	0,0856	0,0838	0,0201	0,0151	0,0366	0,0644	0,0313	0,0279	9,2895
3	0,0165	0,0786	0,0667	0,0326	0,0165	0,0411	0,0669	0,0321	0,0304	9,3375
4	0,0183	0,0822	0,0757	0,0337	0,0183	0,0344	0,0700	0,0327	0,0415	6,4222
5	0,0190	0,0850	0,0646	0,0322	0,0190	0,0294	0,0632	0,0260	0,0348	6,8720
6	0,0195	0,0978	0,0683	0,0338	0,0194	0,0284	0,0679	0,0382	0,0467	7,1259
7	0,0212	0,0988	0,0898	0,0406	0,0212	0,0285	0,0763	0,0262	0,0456	11,0150
8	0,0239	0,1016	0,1027	0,0368	0,0240	0,0527	0,0768	0,0454	0,0657	9,8443
9	0,0266	0,0929	0,0788	0,0406	0,0266	0,0324	0,0777	0,0401	0,0609	12,2170
10	0,0226	0,1050	0,0437	0,0510	0,0226	0,0143	0,0704	0,0234	0,0440	12,5778
11	0,0246	0,1350	0,1398	0,0553	0,0246	0,0441	0,1034	0,0404	0,0619	11,3132
12	0,0242	0,1506	0,1283	0,0337	0,0242	0,0415	0,0853	0,0398	0,0644	9,8551

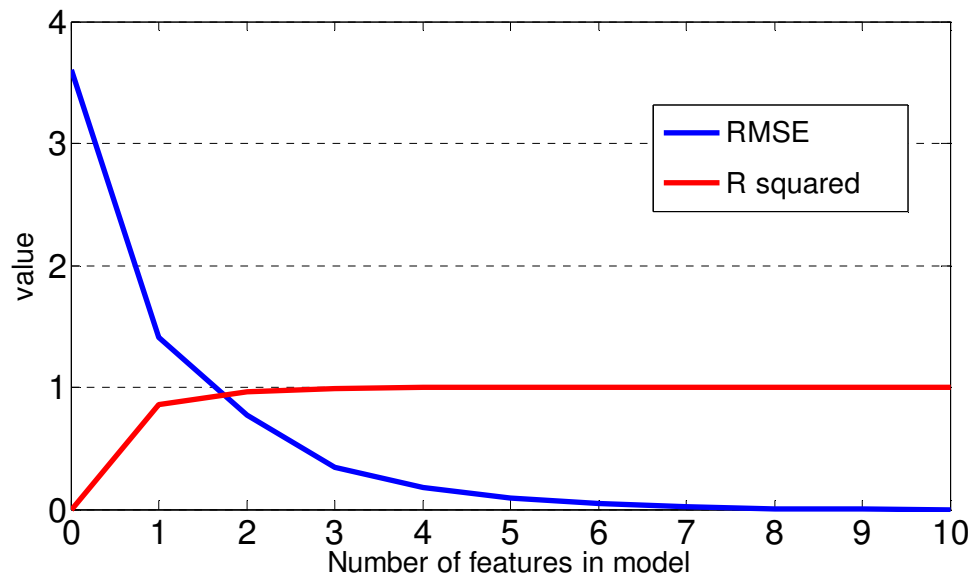


Figure 22. RMSE and R squared values and number of features in the model, Data set 2.

Figure 22 showed that a linear combination with only few features explained most of the variation between different wings. 10 features in the model explained completely the variation between different wings.

Regression analysis was performed on the following models:

$$Y_{fit1} = \text{constant1} + \text{constant2} * \text{feature1}$$

$$Y_{fit2} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2}$$

$$Y_{fit5} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2} + \text{constant4} * \text{feature3} + \text{constant6} * \text{feature4} + \text{constant6} * \text{feature5}$$

$$Y_{fit10} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2} + \text{constant4} * \text{feature3} + \text{constant5} * \text{feature4} + \text{constant6} * \text{feature5} + \text{constant7} * \text{feature6} + \text{constant8} * \text{feature7} + \text{constant9} * \text{feature8} + \text{constant10} * \text{feature9} + \text{constant11} * \text{feature10}$$

The resulting models tested with data are presented in figures 23, 24, 25 and 26. The R squared value and estimate of error variance of each model are presented in table 4. The blue marks refer to responses of the different wings in the model and the red line describes the regression equation.

Table 4. R squared values and error variance estimates of models.

	Yfit1	Yfit2	Yfit5	Yfit10
R squared	0,8611	0,9622	0,9996	1,0000
Estimate of error variance in model	1,8055	0,4918	0,0046	0,0000

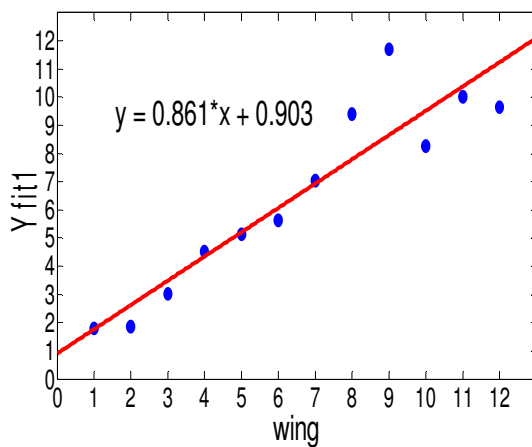


Figure 23. Data set 2 tested with model Yfit1.

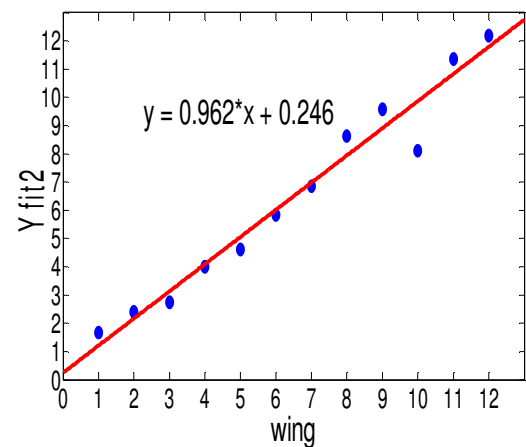


Figure 24. Data set 2 tested with model Yfit2.

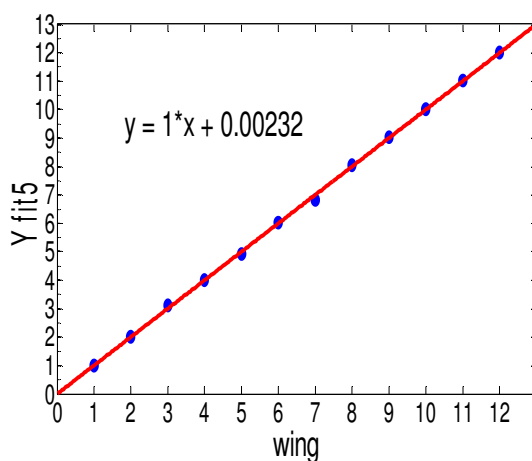


Figure 25. Data set 2 tested with model Yfit5.

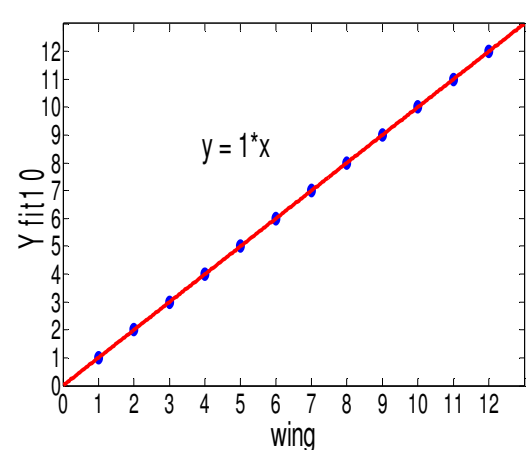


Figure 26. Data set 2 tested with model Yfit10.

Figures 23, 24, 25, 26 and table 4, show that 5 features in the linear model sufficiently explain the variation between different wings.

### 5.3 Data set 3

The data was divided into three groups; each group represented one of the species. The data of all 6 areas were included. The features of models were chosen by stepwise regression algorithms. The chosen features and feature values are presented in tables 5 and 6. The RMSE and R squared values are presented as a function of the sum of included features in the model (figure 27).

Table 5. Data set 3

Feature	Area	Channel	Value	of Value	Filter	Orientation of filter
Feature 1	1	Blue	std	local mean	21 x 21	no
Feature 2	6	Red	std	std	9 x 9	$-\pi/4$
Feature 3	2	Green	mean	std	9 x 9	vertical
Feature 4	6	Green	std	std	9 x 9	$-\pi/4$
Feature 5	5	Blue	mean	std	9 x 9	vertical
Feature 6	2	Blue	std	std	9 x 9	$-\pi/4$
Feature 7	4	Intensity	std	std	21 x 21	vertical
Feature 8	4	Red	std	convolution	9 x 9	$-\pi/4$
Feature 9	6	Green	std	local mean	9 x 9	no
Feature 10	6	Intensity	std	convolution	9 x 9	horizontal

Table 6. Values of the features in Data set 3.

Group	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
1	0,0602	0,0294	0,0645	0,0306	0,0708	0,0414	0,0365	0,8537	0,0991	0,9849
1	0,0704	0,0284	0,0651	0,0299	0,0468	0,0421	0,0583	1,0059	0,1006	0,9990
1	0,0670	0,0257	0,0449	0,0294	0,0398	0,0327	0,0200	0,8076	0,1155	1,0325
1	0,0562	0,0342	0,0742	0,0353	0,0406	0,0341	0,0313	0,9713	0,0885	1,0039
1	0,0561	0,0280	0,0467	0,0291	0,0272	0,0365	0,0324	0,7553	0,0684	0,7324
1	0,0618	0,0279	0,0508	0,0299	0,0343	0,0369	0,0431	0,9360	0,0844	0,8874
2	0,0833	0,0344	0,0503	0,0369	0,0653	0,0337	0,0390	1,1803	0,1289	1,2455
2	0,0829	0,0339	0,0549	0,0346	0,0688	0,0343	0,0797	2,0636	0,0983	1,0119
2	0,0862	0,0324	0,0530	0,0327	0,0602	0,0345	0,0515	1,3843	0,1108	1,1020
2	0,0892	0,0300	0,0406	0,0319	0,0529	0,0313	0,0522	1,1399	0,1195	1,1951
3	0,1033	0,0443	0,0688	0,0445	0,0815	0,0258	0,0684	1,6732	0,0914	0,9589
3	0,1110	0,0380	0,0582	0,0365	0,0751	0,0292	0,0617	1,7093	0,0835	0,8755

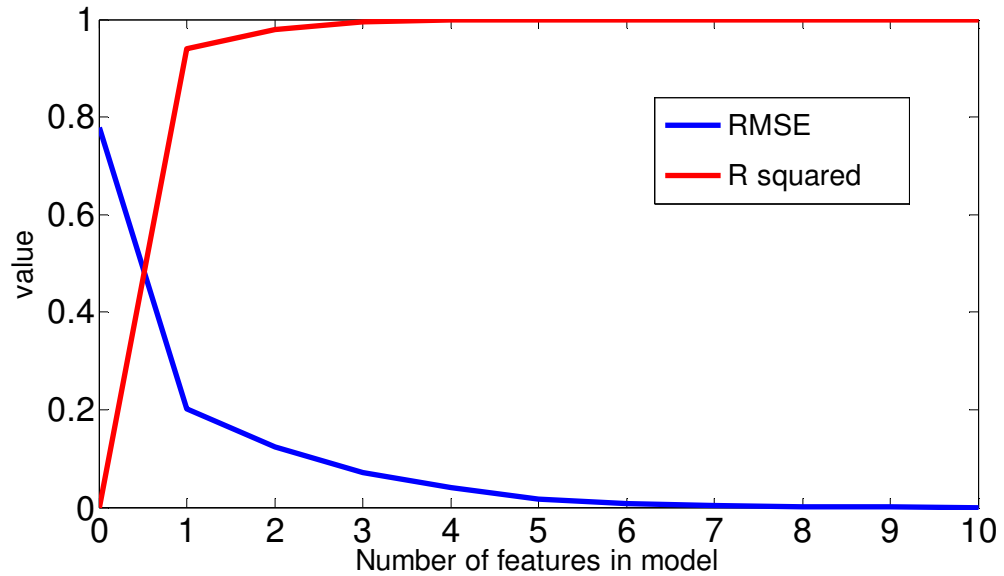


Figure 27. RMSE and R squared values and number of features in the model, Data set 3.

Figure 27 showed that the linear combination with only few features explains most of the variation between the moth species. 10 features in the model completely explained the variation between the moth species.

Regression analysis was performed on the following models:

$$Y_{fit1} = \text{constant1} + \text{constant2} * \text{feature1}$$

$$Y_{fit2} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2}$$

$$Y_{fit5} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2} + \text{constant4} * \text{feature3} + \text{constant6} * \text{feature4} + \text{constant6} * \text{feature5}$$

$$Y_{fit10} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2} + \text{constant4} * \text{feature3} + \text{constant5} * \text{feature4} + \text{constant6} * \text{feature5} + \text{constant7} * \text{feature6} + \text{constant8} * \text{feature7} + \text{constant9} * \text{feature8} + \text{constant10} * \text{feature9} + \text{constant11} * \text{feature10}$$

Resulting models tested with data are presented in figures 28, 29, 30 and 31. The R squared value and estimate of error variance of each model are presented in table 7. The blue marks refer to responses of the different wing groups in the model and the red line describes the regression equation.

Table 7. R squared values and error variance estimates of models.

	Yfit1	Yfit2	Yfit5	Yfit10
R squared	0,9390	0,9792	0,9997	1,0000
Estimate of error variance in model	0,0370	0,0126	0,0002	0,0000

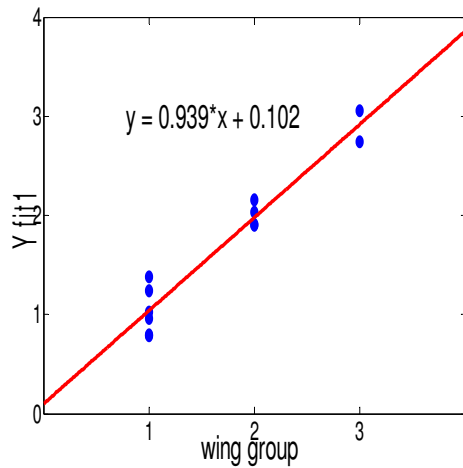


Figure 28. Data set 3 tested with model Yfit1.

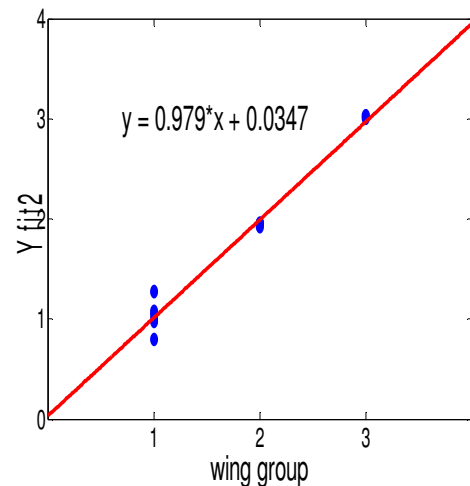


Figure 29. Data set 3 tested with model Yfit2.

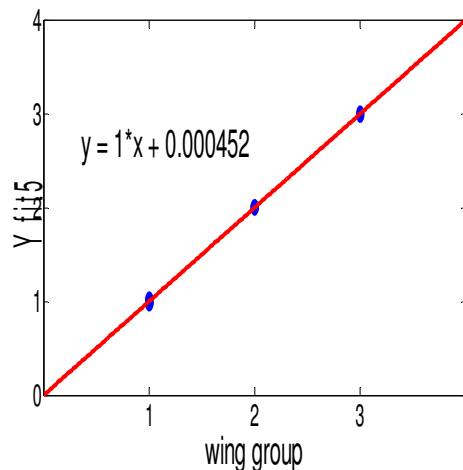


Figure 30. Data set 3 tested with model Yfit5.

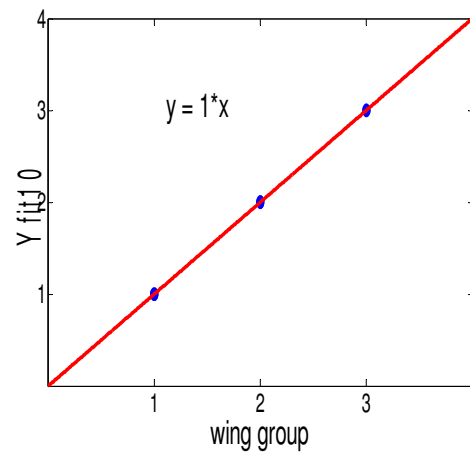


Figure 31. Data set 3 tested with model Yfit10.

Figures 28, 29, 30, 31, and table 7, show that only one or two features in the linear model explain the variation between species sufficiently. 10 features completely explained the variation between the species.

The models were tested for hierarchical agglomerative clustering. The results are presented in figures 32, 33, 34 and 35.

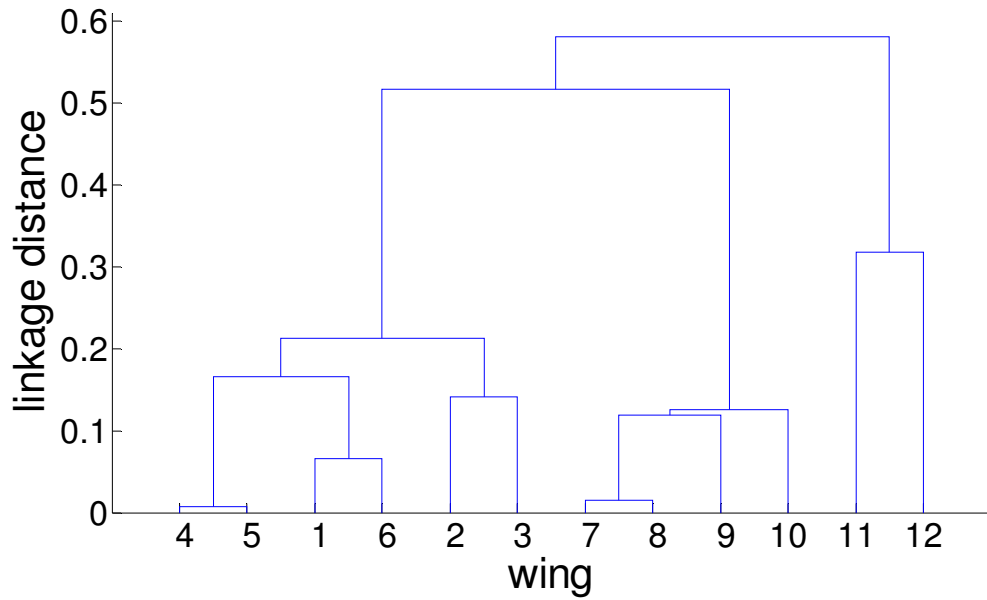


Figure 32. Hierarchical cluster tree formed by Yfit1.

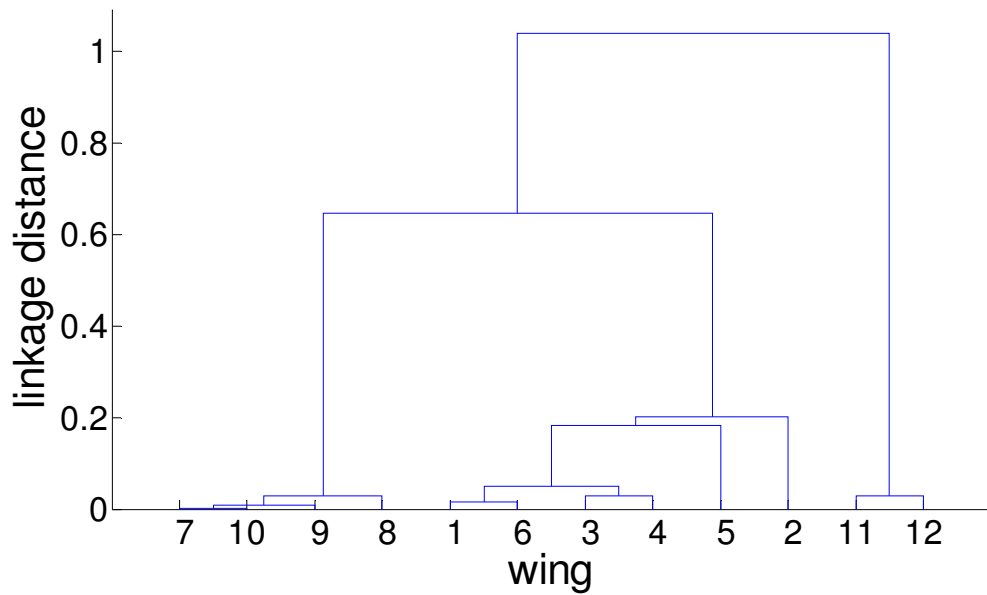


Figure 33. Hierarchical cluster tree formed by Yfit2.

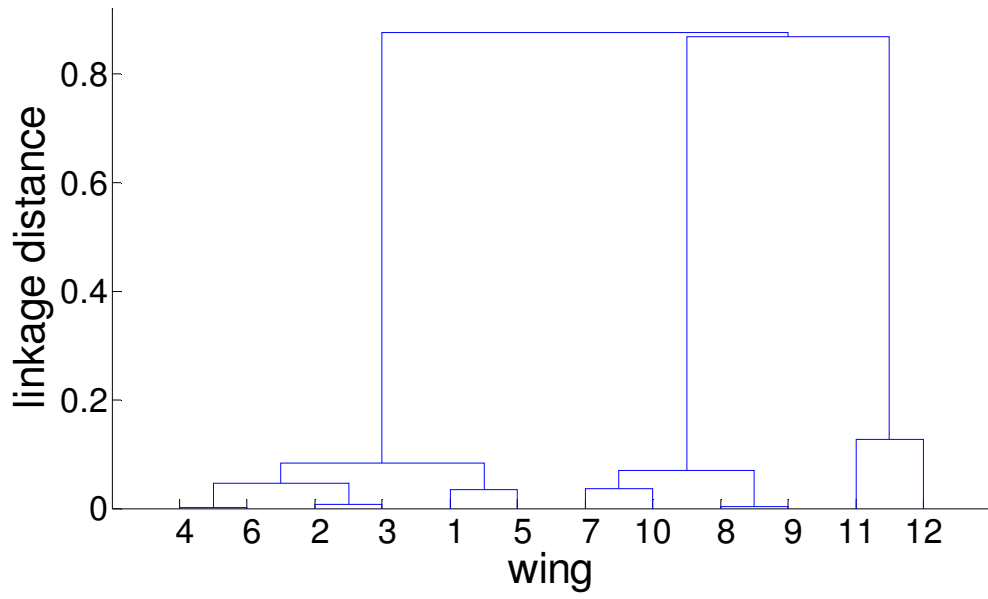


Figure 34. Hierarchical cluster tree formed by Yfit5.

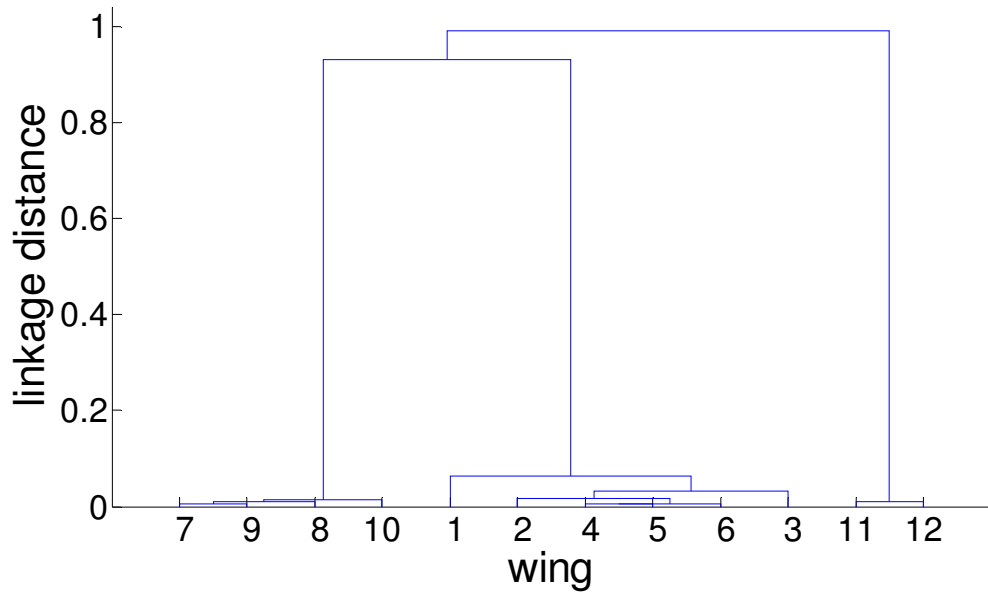


Figure 35. Hierarchical cluster tree formed by Yfit10.

## 5.4 Data set 4

The data was divided into 12 groups; each group represented one wing. The data of areas 2, 3, and 6 were included. The features of models were chosen by stepwise regression algorithms. The chosen features and feature values are presented in tables 8 and 9. The RMSE and R squared values are presented as a function of added number of the sum of included features in model (figure 36).

Table 8. Data set 4.

Feature	Area	Channel	Value	of Value	Filter	Orientation of filter
Feature 1	2	Intensity	std	local std	21 x 21	No
Feature 2	2	Blue	mean	std	21 x 21	$-\pi/4$
Feature 3	6	Intensity	std	std	21 x 21	$-\pi/4$

Table 9. Values of the features in Data set 4.

Group	Feature 1	Feature 2	Feature 3
1	-1,4034	1,6054	-0,3593
2	-1,3900	1,7387	-0,1965
3	-1,0388	-0,2688	-1,6056
4	-0,5875	0,4483	0,3317
5	-0,4086	0,0592	-1,6160
6	-0,2604	-0,2381	-1,0596
7	0,1593	0,1878	0,7200
8	0,8631	0,3617	0,4619
9	1,5519	-0,2948	0,4823
10	0,5248	-1,0559	0,6488
11	1,0503	-0,7905	1,6536
12	0,9392	-1,7531	0,5387

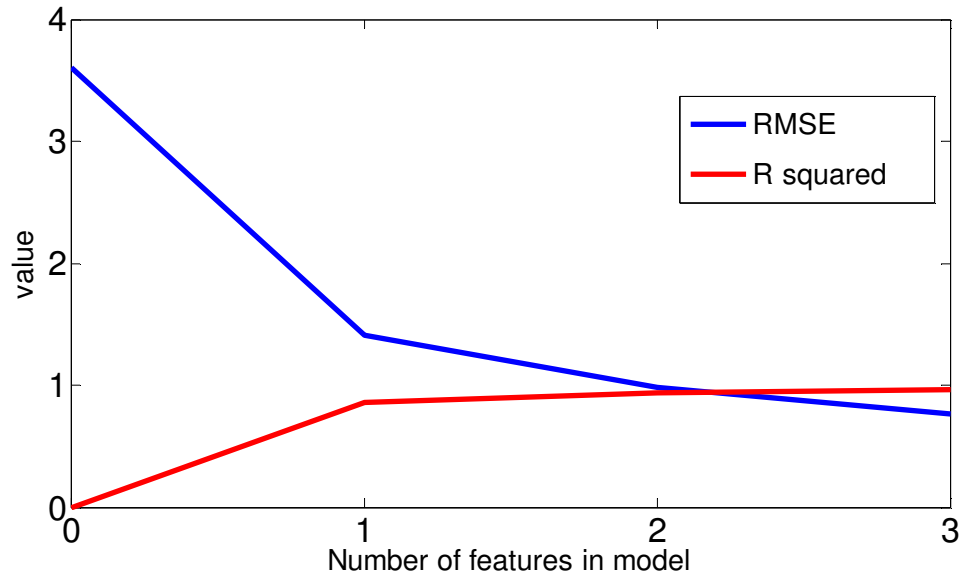


Figure 36. RMSE and R squared values and number of features in the model, Data set 4.

Figure 36 shows that a linear combination with three features explains most of the variation between the wings.

Regression analysis was performed on the following models:

$$Y_{fit1} = \text{constant1} + \text{constant2} * \text{feature1}$$

$$Y_{fit2} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2}$$

$$Y_{fit3} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2} + \text{constant4} * \text{feature3}$$

The resulted models tested with the data are presented in figures 37, 38 and 39. The R squared value and estimate of error variance of each model are presented in table 10. The blue marks refer to responses of the different wings in the model and the red line describes the regression equation.

Table 10. R squared values and error variance estimates of models.

	Yfit1	Yfit2	Yfit3
R squared	0,8611	0,9387	0,9672
Estimate of error variance in model	1,8055	0,7969	0,4265

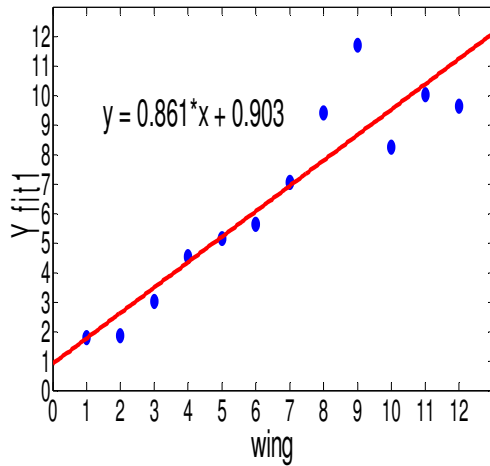


Figure 37. Data set 4 tested with model Yfit1.

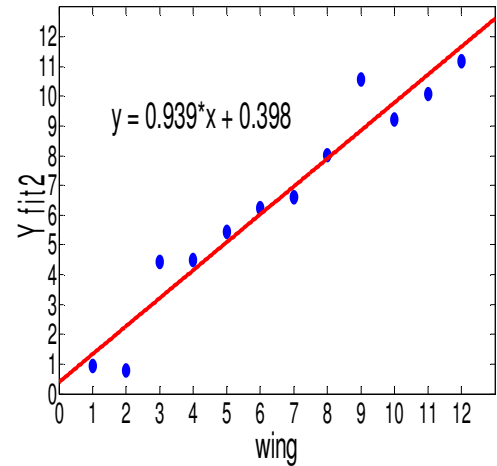


Figure 38. Data set 4 tested with model Yfit2.

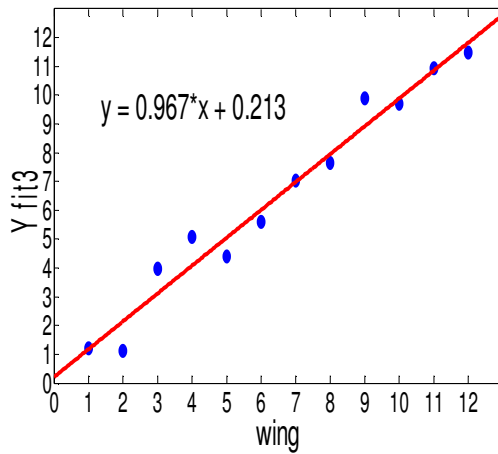


Figure 39. Data set 4 tested with model Yfit3.

The figures show that 3 features in the linear model explain the variation between wings quite well. However, the error variance estimate for Yfit3 is rather high, indicating an unexplained proportion of variance is high.

## 5.5 Data set 5

The data was divided into 3 groups; each group represented one of the species. The data of areas 2, 3, and 6 were included. The features of models were chosen by stepwise regression algorithms. The chosen feature and feature value are presented in tables 11 and 12. The RMSE value is presented as a function of the sum of included features in model (figure 40).

Table 11. Data set 5

Feature	Area	Channel	Value	of Value	Filter	Orientation of filter
Feature 1	6	Red	mean	local std	21 x 21	no

Table 12. Values of the feature in Data set 5.

Group	Feature 1
1	-0,9284
1	-0,8756
1	-0,4476
1	-0,5065
1	-1,1613
1	-0,5814
2	0,5307
2	0,0426
2	0,4945
2	0,0985
3	2,4193
3	0,9152

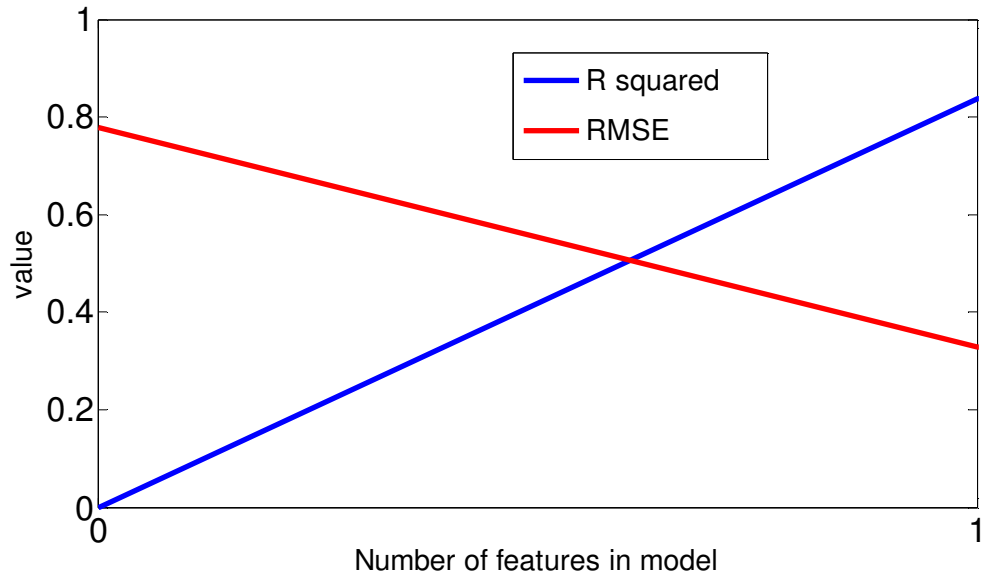


Figure 40. RMSE and R squared values and number of features in the model, data set 5.

The figure 40 shows that a linear combination with one feature explains most of the variation between the species.

Regression analysis was performed on the following model:

$$Y_{fit1} = \text{constant1} + \text{constant2} * \text{feature1}$$

The resulting model tested with data is presented in figure 41. The R squared value and estimate of error variance of model are presented in table 13. The blue marks refer to responses of the different wing group in the model and the red line describes the regression equation.

Table 13. R squared value and error variance estimate of model.

	Yfit1
R squared	0,8371
Estimate of error variance in model	0,0987

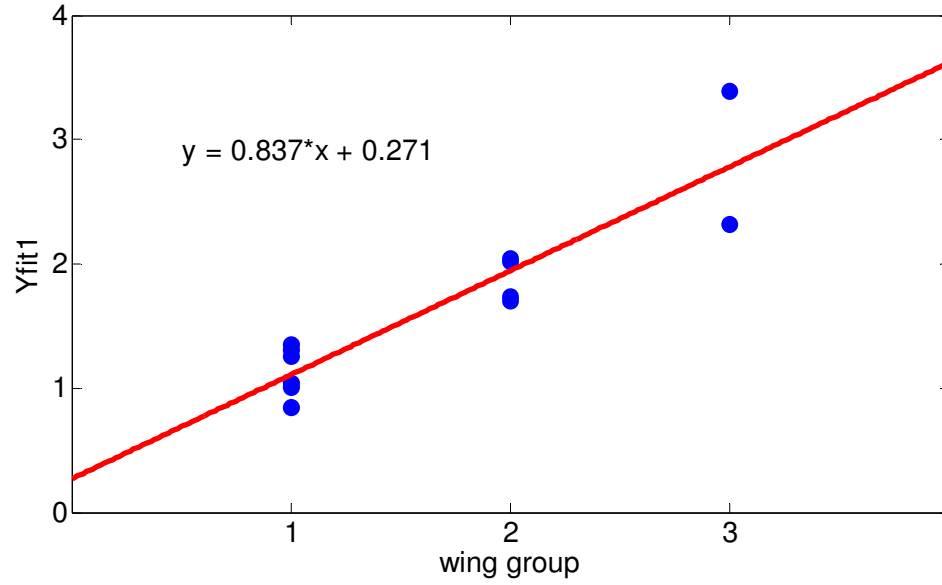


Figure 41. Data set 5 tested with model Yfit1.

The model Yfit1 was tested for hierarchical agglomerative clustering. The results are presented in figure 42.

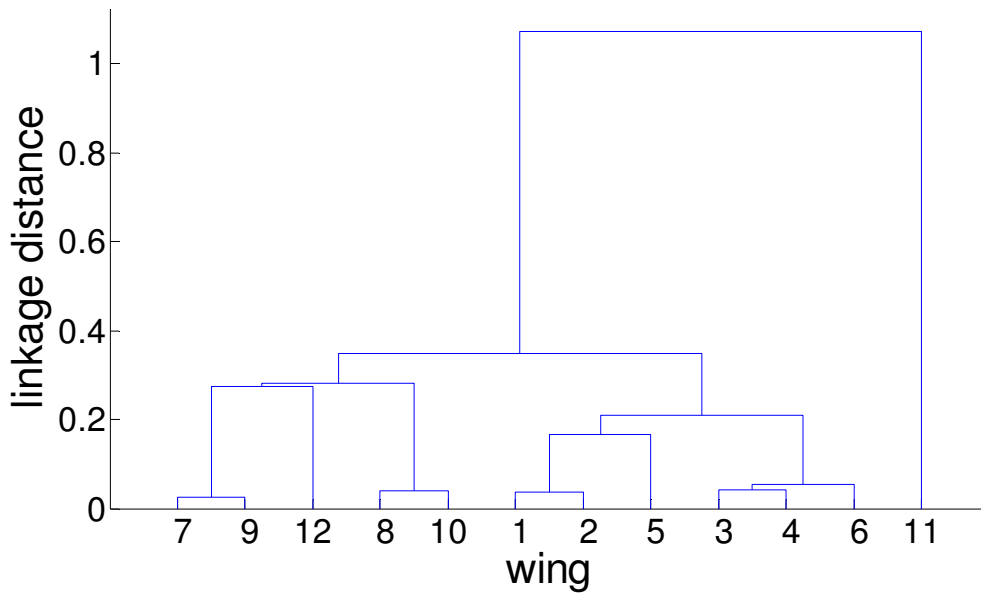


Figure 42. Hierarchical cluster tree formed by model Yfit1.

## 5.6 Data set 6

The data was divided into two groups. The first group consisted of the features of *Cydia pomonella* and the second group consisted of the features of *Cydia splendana* and *Cydia strobilella*. The data of areas 2, 3, and 6 were included. The features of models were chosen by stepwise regression algorithms. The chosen features and feature values are presented in table 14 and 15. The RMSE and R square values are presented as a function of the sum of included features in model (figure 43).

Table 14. Data set 6

Feature	Area	Channel	Value	of Value	Filter	Orientation of filter
Feature 1	2	Red	std	local std	21 x 21	no
Feature 2	2	Blue	std	std	21 x 21	horizontal
Feature 3	6	Green	std	local std	21 x 21	no
Feature 4	2	Green	std	std	21 x 21	$-\pi/4$
Feature 5	6	Red	mean	convolution	21 x 21	horizontal

Table 15. Values of the features in Data set 6.

Group	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
1	-1,3961	0,3813	0,4649	-0,5720	-0,5819
1	-1,3948	0,6213	0,4638	0,2438	-0,2819
1	-1,0329	0,0197	-0,9494	-0,7443	0,6552
1	-0,5837	0,3930	0,0485	1,4080	-1,2350
1	-0,4058	1,2147	-1,8885	0,1487	-0,9439
1	-0,2870	1,7720	-1,4464	0,2311	-0,7550
2	0,1602	-1,5217	1,5503	-1,4838	-0,2731
2	0,8815	-0,1115	0,5960	-0,4943	-0,8991
2	1,5464	-0,2849	0,2601	1,5194	0,3556
2	0,5237	-1,1943	0,8157	-0,8325	0,5432
2	1,0470	0,0732	0,5480	1,3806	1,8258
2	0,9414	-1,3628	-0,4632	-0,8047	1,5900

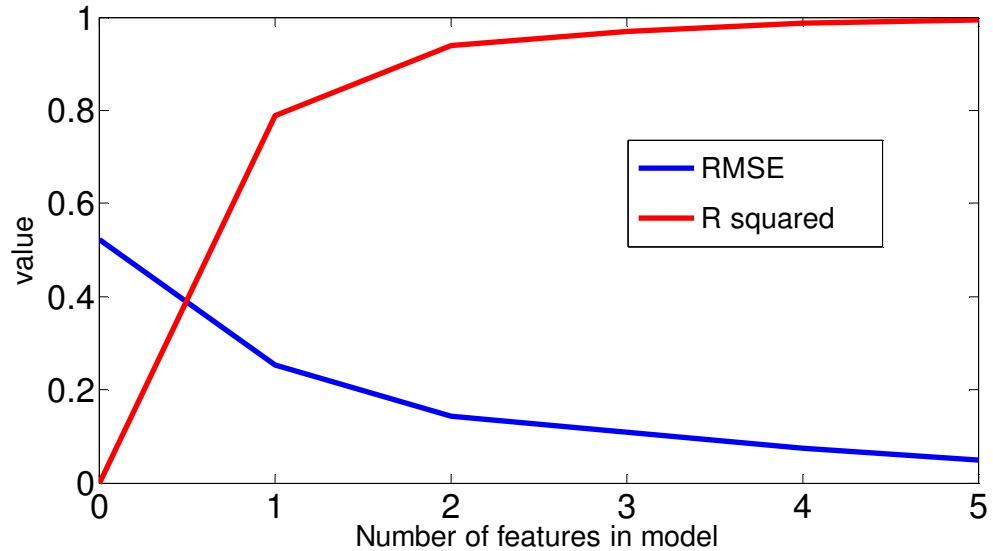


Figure 43. RMSE and R squared values and number of features in the model, Data set 6.

The figure 43 shows that a linear combination with only few features explains most of the variation between the moth species.

Regression analysis was performed on the following models:

$$Y_{fit1} = \text{constant1} + \text{constant2} * \text{feature1}$$

$$Y_{fit2} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2}$$

$$Y_{fit3} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2} + \text{constant4} * \text{feature3}$$

$$Y_{fit5} = \text{constant1} + \text{constant2} * \text{feature1} + \text{constant3} * \text{feature2} + \text{constant4} * \text{feature3} + \text{constant5} * \text{feature4} + \text{constant6} * \text{feature5}$$

The resulted models tested with data are presented in figures 44, 45, 46 and 47. The R squared value and the estimate of error variance of each model are presented in table 16. The blue marks refer to responses of the different wing groups in the model and the red line describes the regression equation.

Table 16. R squared values and error variance estimates of models.

	Yfit1	Yfit2	Yfit3	Yfit5
R squared	0,7883	0,9394	0,9683	0,9950
Estimate of error variance in model	0,0577	0,0165	0,0086	0,0014

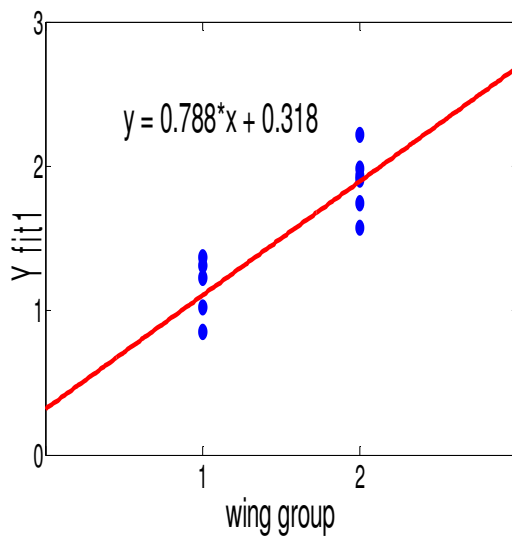


Figure 44. Data set 6 tested with model Yfit1.

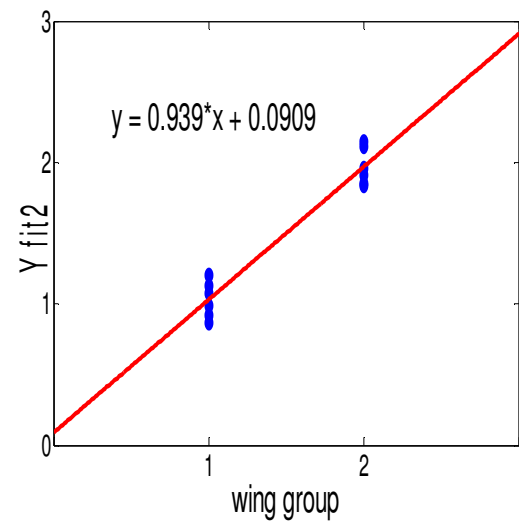


Figure 45. Data set 6 tested with model Yfit2.

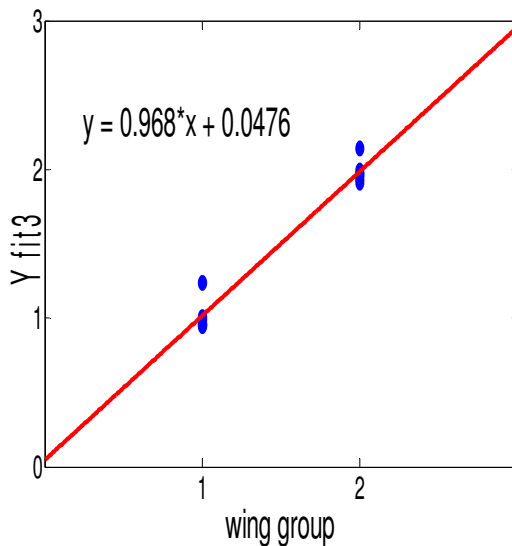


Figure 46. Data set 6 tested with model Yfit3.

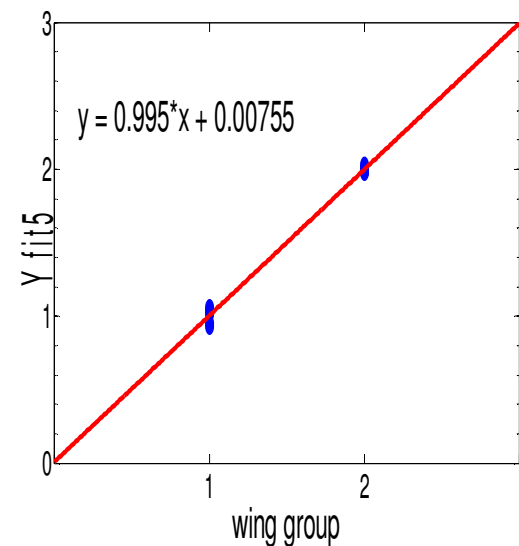


Figure 47. Data set 6 tested with model Yfit5.

Figures 44, 45, 46, 47, and table 16, show that two features in a linear model sufficiently explain the variation between species groups.

The models were tested for hierarchical agglomerative clustering. The results are presented in figures 48, 49, 50 and 51.

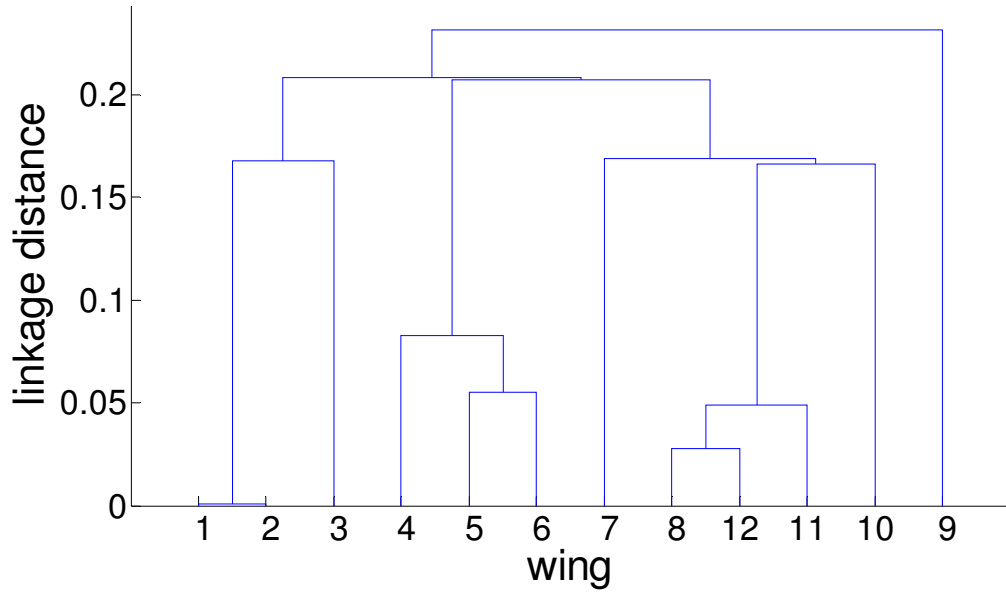


Figure 48. Hierarchical cluster tree formed by Yfit1.

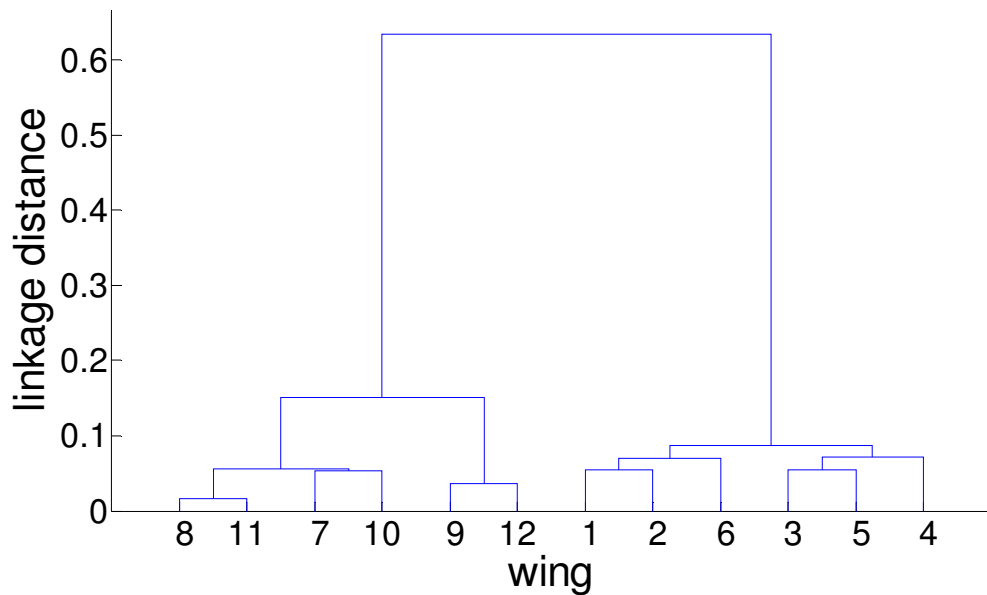


Figure 49. Hierarchical cluster tree formed by Yfit2.

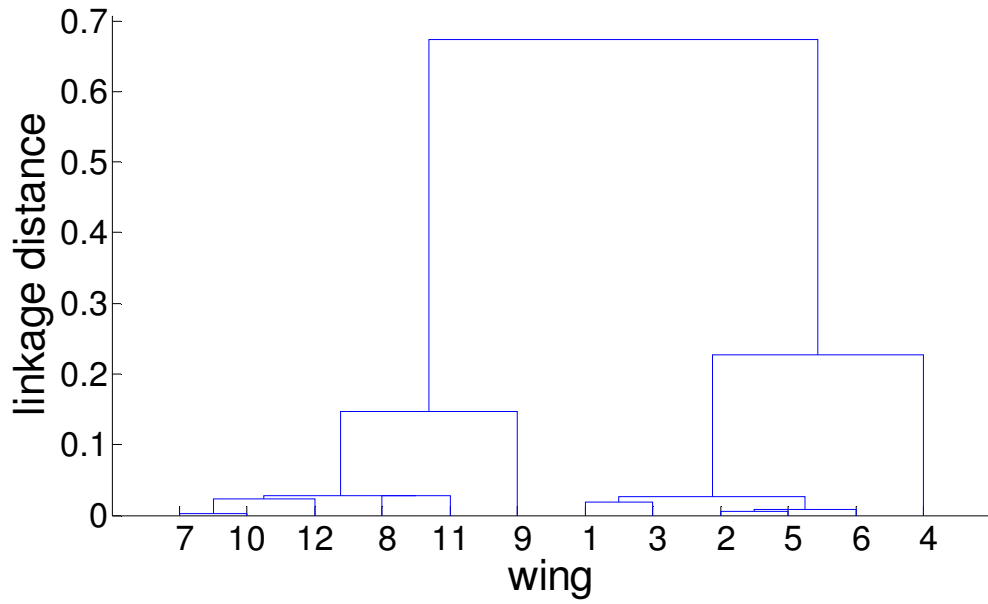


Figure 50. Hierarchical cluster tree formed by Yfit3.

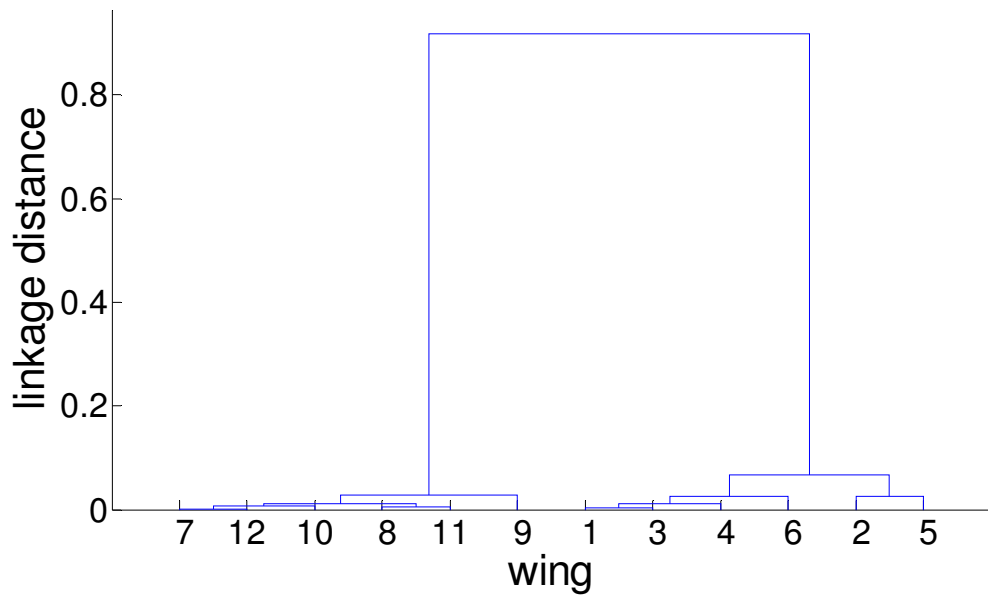


Figure 51. Hierarchical cluster tree formed by Yfit5.

## **6 Discussion**

### **6.1 Properties of the Camera and optics**

High quality, low-noise and high depth of field digital macro images are achieved by low ISO value, small focal length, small cell size with high pixel density and using medium size aperture. In this research, the photographs were taken with an SLR camera which had a medium-size cell, a medium focal length and a low ISO value. However, the depth of field was long enough to produce sharp photographs because the wings of the moths were in a flat position. In laboratory circumstances it was possible to adjust the camera settings to produce reasonably qualified photographs. In practice, the depth of field was from two to three millimetres. This is not sufficient for practical open field purposes. For open field image acquisition, a camera module with smaller cell and larger depth of field properties is recommended.

### **6.2 Effect of the lighting on the reflected image**

Light emitting diodes were used for object lighting. It is possible that some of the properties may not have been recorded well because the led's emitted light spectrum is limited. However, it was possible to measure features that clearly differed to each other. Therefore white leds were reasonable for lighting.

### **6.3 The amount and quality of forewings**

For this research, only a few moths were photographed. This was a limit for generalized interpretation of results, especially the regression models. However, the species were inspected and confirmed by M.Sc Jaakko Kullberg, the collection manager of the Finnish Museum of Natural History. Special attention was also focused on selecting process of the moths. The individuals were selected based on their visual representativeness of the species. More individuals are needed for a generalization of the results.

#### **6.4 Orientation of forewings**

Usually moth wings are in moth collections in a flattened position. It enables the determination of several features. In open field conditions, especially in pheromone traps, they usually are in a sitting position and some of the features of forewings are not visible. Therefore features for image-based purposes should be chosen by visible features. In this research, special interest was focused on areas 2, 3 and 6, which are visible in the sitting position of moths.

Moths adhere to glue traps in different positions. For image-based measurements, a better type of trap should be developed; the visible position of moths has a key role.

Another possibility could be to take many pictures of the moving moth, one after the one, with high frequency, and analyze the features from the series of the pictures. If the target specie exists in the pictures, it is more or less probable, that the key features of the forewings exist in some of the pictures.

#### **6.5 Normalization**

Normalization of data affected clustering. Linkage distances were longer with normalized data. Normalization scaled feature values into the same range. Without normalization, higher values will have more weight in clustering than lower values. Normalization improved the clustering method to classify wings more accurately. For regression analysis, normalization had no effect. This was also tested but results are not published in this research.

#### **6.6 Feature selection**

Stepwise regression algorithms were used to reduce the amount of features. Because of the behaviour of the method, several combinations of features could be assessed for almost similar results. However, the features selected by stepwise regression will lead to the local optimum of a model. The method was reliable for reducing features. There was a lot of correlation between the features and stepwise regression algorithms were able to decrease

redundant information. In some cases, redundant information was so dominant that only one of the features fulfilled the criteria of stepwise regression and hence was associated as a feature in the explanatory model.

In this data, stepwise regression algorithms found local optimums for describing the differences between wings and wing groups. The dominating features in the models were filtered responses of intensities. In general, differences were mainly explained by combinations of 21 x 21 -size filter responses. In models, there were no features including direct pixel-wise intensities although they existed in the data. Based on this data, it is suggested to measure different filter responses instead of direct pixel-wise measurements.

Different channels were quite equally representative in models. Different channels were also highly correlated with each other and resulting models were strongly affected by the determination of initial terms. Because of the high correlation rate between channels, it could be possible to achieve almost as good a fit in models by calculating together different channel responses as those of separate channels. It could improve the reliability of models averaging the responses of different channels, and thus reduce the effect of outliers and noise on model features. However, this was not tested in this research.

The validation of the different linear regression models was evaluated only with the original data. Therefore the explanatory power of the models is concerning only the data existing in this research. For more general prediction, the models should be tested with other independent data sets or with splitting the data into two groups. Another group should be used to produce the models and another group to test the models. This kind of data arrangement should be statistically more reliable. However, this was not done because of the lack of the material.

## **6.7 Data set 1**

Hierarchical agglomerative clustering algorithms were able to classify the wings into right clusters without performing stepwise regression. However, the linkage distance was relatively small

between *Cydia pomonella* wings and *Cydia splendana* wings. On the other hand, *Cydia strobilella* wings were quite well distinguishable from the other species. If a target was to differentiate *Cydia pomonella* from the other species, it would not be reasonable to use all 1008 features in the classification.

## **6.8 Data set 2**

Stepwise regression algorithms found 10 features that could explain variation between wings completely. However, fewer features were needed for reliable classification. RMSE values in the models decreased clearly by adding features into the models. Five features were a sufficient number to explain the variation between the wings.

## **6.9 Data set 3**

Data set 3 was divided by species. Only one feature of 1008 was needed to separate the species from each other in hierarchical agglomerative clustering. The linkage distance increased by adding more features to model. The most effective features in the models were found in areas 1, 6 and 2. This one feature in the model resulted in an RMSE value of 0,0370 and two features in the model 0,0126, which are rather low values. The filtered responses were dominant in all models and variation between species was explained completely by features which consisted of either 9 x 9 – size or 21 x 21 -size filtered responses. If measurements consisted of all 6 areas, there was no doubt about using only a few feature parameters in the classification of these three species.

## **6.10 Data set 4**

Data set 4, 5 and 6 were assessed to reflect the visible areas of a moth in the sitting position. Therefore only half of the key feature areas (2, 3 and 6) were included. Stepwise regression algorithms reduced the amount of features in the models. For data set 4, the maximum of three parameters were included in the model. Regression equations show quite low R squared and quite high

RMSE values indicating that the remaining residual variance in models was high. There was not enough information in the data set's features to achieve a better fit for the models.

### 6.11 Data set 5

Stepwise regression algorithms were able to choose only one feature from area 6 of the model. The R squared value was low and the error variance estimate was relatively high. In this case, it was not possible to increase the model and clustering accuracy, because the data did not contain features which would decrease the RMSE of the model. A hierarchical agglomerative clustering method misclassified one of the *Cydia strobilella* wings to the wrong cluster. However, *Cydia pomonella* wings were classified correctly. Based on this research, it is not possible to classify reliably between these three species by the measured features of visible areas 2, 3 and 6 but some certainty could be noticed for separating *Cydia pomonella* from these two species.

### 6.12 Data set 6

In data set 6, the target species *Cydia pomonella* wings were grouped into a single group and the others formed another group. Two features in the model were sufficient for correctly classifying the wing groups. The RMSE values were low and the R squared values were high, with the exception of only one feature included in the model. It was possible to classify *Cydia pomonella* wings into their own group with a rather high certainty by including all found five features into the model. The best model included features from areas 2 and 6 but not from area 3. The features consisted of 21 x 21 –size filtered responses of red, blue and green channels. The reason for excluding the brown stripe from effective features was the similarity possibility in this area between *Cydia pomonella* and *Cydia splendana*.

## 7 Conclusions

Based on this research, the *Cydia pomonella* species can be identified by its forewing properties. If there were measurements available from all forewing areas, identification of the three *Cydia* species from each other was reliable. If identification was based on a sitting moth's visible areas, it was not as reliable as when it was based on all areas. However, the identification of *Cydia pomonella* from the other two *Cydia* species was still possible. Stepwise regression algorithms were a reliable method to reduce the number of features in feature sets that contained redundant information. The discriminating features were found to be quite evenly distributed between all examined areas, with the exception of the area 3. The identification of a sitting *Cydia pomonella* can be based on the measured and calculated features in the white-brown veined area (area 2) and the bronze coloured oval (area 6), but possibly not the brown stripe (area 3). To have discriminate features in regression models, it is recommended to use 21 x 21 -sized or 9 x 9 -sized filtered values rather than direct pixel-wise measurements.

To have more reliable and general models, the models developed in this research should be tested with other independent data sets. The current trap models should be improved to be more suitable for image analyzing purposes. Another possibility is to develop a new trap model, in which the moths could be photographed in such position, which enables the efficient use of image analyzing methods. With modern high-speed camera modules it could be possible to take pictures with high frequency thus reducing the requirements of moth positioning and enabling to take pictures of moving moth in different positions.

## References

- Addison, M. F. 2005. Suppression of codling moth *Cydia pomonella* L. (Lepidoptera: Tortricidae) populations in South African apple and pear orchards using sterile insect release. Proceedings of the IXth International Pear Symposium, Stellenbosch, South Africa, 1-6 February, 2004. *Acta Horticulturae* 671, p. 555-557.
- Arbuckle, T., Schroder, S., Steinhage, V. and Wittmann, D. 2001. Biodiversity informatics in action: identification and monitoring of bee species using ABIS. In Proc. 15<sup>th</sup> Int. Symp. Informatics for Environmental Protection, vol 1, pages 425-430. Zurich 2001.
- Baixeras, J., J. W. Brown & T. M. Gilligan. 2010. T@RTS: Online World Catalogue of the Tortricidae (Version 1.4.0). <http://www.tortricidae.com/catalogue.asp>, referred 31.8.2011
- Ballard, J., Ellis, D. J., Payne, C. C. 2000. Uptake of granulovirus from the surface of apples and leaves by first instar larvae of the codling moth *Cydia pomonella* L. (Lepidoptera: Olethreutidae). *Biocontrol Science and Technology* 10, p. 617-625.
- Costa, L., F. and Cesar Jr, R., M. 2001. *Shape Analysis and Classification*. 659 p. CRC press, Florida.
- Delgado, N., L. 2010. Local-Feature generic Object Recognition with Application to Insect-Species Identification. Doctoral thesis, University of Washington, department of Electrical Engineering.
- Dryen, I, L. and Mardia, K.V. 1998. *Statistical Shape Analysis*. 347 p. John Wiley & Sons, New York.
- El-Sayed A., M. 2011. The Pherobase: Database of Insect Pheromones and Semiochemicals. <http://www.pherobase.com> , referred 31.8.2011
- Gonzales, R., C. and Woods, R, E. 2002. *Digital Image Processing*. 793 p. Prentice Hall, New Jersey.
- Kullberg, J., Ahlberg, A., Kaila, L. and Varis, V. (2011). Checklist of Finnish Lepidoptera – an updated version. Suomen perhosten luettelo – päivitetty versio. Referred 31.8.2011. Original publication: Kullberg, J., Ahlberg, A., Kaila, L. and Varis, V. 2002 Checklist of Finnish Lepidoptera - Suomen perhosten luettelo. *Sahlbergia* 6(2):45-190
- Light, D.M., Knight, A.L., Henrick, C.A., Rajapaska, D., Lingren, B., Dickens, J.C., Reynolds, K.M., Buttery, R.G., Merrill, G., Roitman, J. and Campbell, B.C.

(2001). A pear-derived kairomone with pheromonal potency that attracts male and female codling moth, *Cydia pomonella* (L.). *Naturwissenschaften* (2001) 88:333–338. Springer-Verlag.

Light, D.M., Knight, A. (2005). Specificity of Codling Moth (Lepidoptera: Tortricidae) for the Host Plant Kairomone, Ethyl (2*E*,4*Z*)-2,4-Decadienoate: Field Bioassays with Pome Fruit Volatiles, Analogue, and Isomeric Compounds. *J. Agric. Food Chem.*, 2005, 53 (10), pp. 4046–4053. American Chemical Society.

Masante-Roca, I., Gadenne, C. and Anton, S. (2002). Plant odour processing in the antennal lobe of male and female grapevine moths, *Lobesia botrana* (Lepidoptera: Tortricidae). *J. Insect Physiol.* 48, p.1111 -1121.

Milli, R., de Kramer, J. J. (1991). Analysis of pheromone distribution in apple orchards where *Cydia pomonella* and *Adoxophyes orana* were controlled by mating disruption. pp. 397-400. In: I. Hrdy (ed.) *Proceedings, Insect Chemical Ecology*, 12-18 August 1990, Tabor, Czechoslovakia. Academia Praha and SPB Academic Publishing.

Nixon, M and Aguado, A. 2006. *Feature Extraction & Image Processing*. 343 p. Elsevier Ltd, Oxford, UK.

O'Neill, M. A., Gauld, I. D., Gaston, K. J and Weeks, P. 2000. DAISY: an automated invertebrate identification system using holistic vision techniques. In *Bio-NET-Intl. Group for Computer-Aided Taxonomy (BIGCAT)*, pages 13-22, Egham, 2000.

Pantofaru, C., Dorko, G., Schmid, G. and Hebert, M. 2006. Combining Regions and Patches for Object Class Localization. *The Beyond Patches Workshop in conjunction with the IEEE conference on Computer Vision and Pattern Recognition*, June, 2006, pp. 23 - 30.

Peltotalo, P. & Tuovinen, T. (1986). Specificity of pheromone preparates for Lepidopterous pests. *Annales Agriculturae Fenniae*, vol. 25: 139-146 (1986).

Schmid, C., Dorko, G., Lazebnik, S., Mikolajczyk, K. and Ponce, J. 2004. Pattern recognition with local invariant features. *Handbook of Pattern Recognition and Computer Vision*, 3rd edition.

C.H. Chen and P.S.P Wang editors, *World Scientific Publishing Co.*, 2005, pp. 71-92.

Zhu, L., Zhang, Z. 2011. Insect recognition based on integrated region matching and dual-tree complex wavelet transform. *Journal of Zhejiang University – Science C (Computers and Electronics)* 2011, vol 12(1), pages 44-53.

## Appendix

### Filters in feature extraction

```

0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0

```

9 x 9 –size line filter, horizontal direction. Used in conv2 –and stdfilt -functions of Matlab.

```

0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0

```

9 x 9 –size line filter, vertical direction. Used in conv2 – and stdfilt -functions of Matlab.

```

1 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0
0 0 0 0 1 0 0 0 0
0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 0 0 1

```

9 x 9 –size line filter,  $-\pi/4$  direction. Used in conv2 –and stdfilt -functions of Matlab.

```

0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 1 0 0
0 0 0 0 0 1 0 0 0
0 0 0 0 1 0 0 0 0
0 0 0 1 0 0 0 0 0
0 0 1 0 0 0 0 0 0
0 1 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0

```

9 x 9 –size line filter,  $+\pi/4$  direction. Used in conv2 –and stdfilt -functions of Matlab.

```

1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1

```

9 x 9 –size filter. Used in stdfilt –function of Matlab for calculating local standard deviation.

```

1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81
1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81
1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81
1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81
1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81
1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81
1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81
1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81
1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81 1/81

```

9 x 9 –size filter. Used in conv2 –function of Matlab for calculating local mean.





