

Scientific Data Mining for Spatio-Temporal Hydroacoustic Data Sets

Bart Buelens

**Submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy**

**University of Tasmania
November 2008**

Morris
Thesis
BUELENS
PhD
2008

A 7002 2073337B

857651

Declaration

This thesis contains no material which has been accepted for the award of any other higher degree or graduate diploma in any tertiary institution. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference has been made in the text of the thesis, nor does this thesis contain any material that infringes copyright.

A handwritten signature in black ink, appearing to read 'Bart Buelens', with a long horizontal stroke extending to the right.

Bart Buelens

Statement of Authority of Access

This thesis is not to be made available for loan or copying for two years following the date this statement was signed. Following that time the thesis may be made available for loan and limited copying in accordance with the Copyright Act 1968.

A handwritten signature in black ink, appearing to read 'B. Buelens', with a long horizontal stroke extending to the right.

Bart Buelens

Date 16 November 2008

Statement of co-authorship

The publications of the work undertaken as part of this thesis are the following:

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2003). "Midwater acoustic modeling for multibeam sonar simulation," 146th ASA Meeting (Austin, Texas), The Journal of the Acoustical Society of America 114, p. 2308.

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2004). "A framework for scientific data mining in hydroacoustic data sets," 2nd International Conference on Artificial Intelligence in Science and Technology (AISAT) (Hobart, Tasmania, Australia), pp. 104-108.

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2005). "Model inversion for midwater multibeam backscatter data analysis," IEEE Oceans '05 Europe (Brest, France), pp. 431-435.

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2005). "A scientific data mining approach to midwater multibeam echosounding for fisheries applications," 1st International Conference on Underwater Acoustic Measurements: Technologies & Results (UAM) (Heraklion, Crete, Greece).

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2006). "Computational challenges in processing and analysis of full-watercolumn multibeam sonar data," 8th European Conference on Underwater Acoustics, edited by S. M. Jesus, and O. C. Rodríguez (Carvoeiro, Portugal), pp. 799-804.


Buelens, B., Pauly, T., Williams, R., and Sale, A. (in press). "Kernel methods for detection and classification of fish schools in single beam and multibeam acoustic data," in ICES Journal of Marine Science, Special Issue on the Ecosystem Approach with Fisheries Acoustics and Complementary Technologies.

The proportion of the work undertaken for each of the manuscripts is as follows:


- Mr. Bart Buelens (70%) is the primary author. He conducted the research and prepared the material for publication.
- Dr. Ray Williams (10%) and Prof. Arthur Sale (10%) of the School of Computing, University of Tasmania, provided general guidance as supervisors and contributed to the general ideas and textual presentation of the material.

- Dr. Tim Pauly (10%) of Myriax Pty Ltd contributed to aspects relating to underwater acoustics and fisheries research and offered comments on manuscript drafts.

We the undersigned agree with the above stated proportion of work undertaken for each of the above published or submitted manuscripts contributing to this thesis.

Signed: 

 Dr. Ray Williams
 Supervisor
 School of Computing and
 Information Systems
 University of Tasmania



 Dr. Julian Dermoudy
 Head of School
 School of Computing and
 Information Systems
 University of Tasmania

Date: 2/12/2008

1/12/2008

Abstract

Managing natural marine resources for sustainable exploitation of the oceans and the flora and fauna they contain is a challenging task. Decisions by policy makers are based on advice from the scientific community. Through surveying and monitoring programs, scientists study the marine environment to gain insight into its structure and function. Employing acoustic techniques, sonar systems are often the best tools available to effectively observe aquatic environments. Important applications include fisheries and seafloor mapping. Fish stock assessments are typically conducted using single beam echosounders, while bathymetric surveys are conducted with multibeam sonar.

Multibeam sonar instruments that are capable of collecting samples for the complete water column are an emerging technology. Since they collect acoustic data over much greater sampling volumes than single beam instruments, significant improvements in fisheries studies are expected. The combined collection of seafloor and water-column data will lead to survey cost savings and to an integrated, ecosystem-based approach to monitoring the ocean environment. While standard data analysis procedures are established for single beam fisheries and standard multibeam bathymetric applications, this is not the case for full water-column multibeam sonar data.

In this thesis, a data mining approach for handling such data is proposed. The developed method consists of a preprocessing algorithm based on an inversion technique, followed by a pattern analysis algorithm using kernel clustering methods. The preprocessing algorithm applies a deconvolution as a model inversion method to reduce the data set in size and to convert the acoustic measurements into a generic vector representation. Each vector has a spatial and a temporal component as well as a number of additional features typically relating to the acoustic backscatter energy. These spatio-temporal vectors are then subjected to pattern analysis algorithms. Two clustering algorithms are selected: a density based spatial clustering algorithm, and a clustering algorithm based on kernel methods. A new method is developed to allow the kernel clustering algorithm to make use of the spatial and non-spatial components of the data in a combined fashion. This results in a powerful, flexible and versatile clustering procedure. The outcome is a segmentation of the data into coherent structures, for example fish schools and the seabed. Classification is achieved through the differentiation between data clusters indicative of different fish species or seabed habitats. The effectiveness of the data mining methods is demonstrated in a number of case studies.

It is hoped that the developed approach will facilitate routine use of water-column multibeam sonar data for fisheries applications in particular, and for ecosystem studies and marine resource management in general.

Acknowledgements

I am very grateful to my supervisors Ray Williams and Arthur Sale at the University, and Tim Pauly at Myriax Pty Ltd, for the support they have given me during the six years of candidature. They have been very responsive to questions, provided me with useful advice and feedback, and have shown a great interest in my project. I appreciate the effort they have put in maintaining their supervisory roles also while I was overseas during the last two years of candidature.

I wish to thank Myriax Pty Ltd (formerly known as SonarData Pty Ltd) for providing the funding for my PhD project. Such an investment carries a certain level of risk and I am grateful that they have put their trust in me.

The feedback I received from Matt Wilson, David Millington and Toby Jarvis who have proof read my thesis has been very valuable, for which I am thankful.

The following people have provided me with sonar data sets that I could use in the context of my PhD research: John Anderson, Northwest Atlantic Fisheries Centre, Department of Fisheries and Oceans, St John's, NF, Canada; Ken Foote and Dezhuang Chu from Woods Hole Oceanographic Institution, Woods Hole, MA, USA; John Horne, School of Fisheries, University of Washington, Seattle, WA, USA; Toby Jarvis, Australian Antarctic Division, Kingston, Tasmania, Australia; Chris Malzone, Reson Inc., Goleta, CA, USA; Tom Weber, University of New Hampshire, Center for Coastal and Ocean Mapping, Durham, NH, USA. I appreciate their efforts in making the data files available. Where such data files are used in examples or case studies in this thesis, the source of the data is acknowledged in the text.

Finally I thank my wife Lieve for giving me the time and space to pursue this PhD, and generally for putting up with me in particular during the more stressful stages of candidature. Siska-Lut and Korneel, my two children, were born after I commenced this project, so they haven't known their dad not doing a PhD, but they will hopefully notice a difference after I have submitted this thesis.

Bart Buelens

CONTENTS

1	INTRODUCTION	1
1.1	MOTIVATION	1
1.2	CONTEXT	2
1.3	PROBLEM DESCRIPTION	5
1.4	RESEARCH OBJECTIVES	5
1.5	THESIS SYNOPSIS	6
2	BACKGROUND.....	7
2.1	UNDERWATER ACOUSTICS.....	7
2.1.1	<i>Underwater acoustic measurements</i>	<i>7</i>
2.1.2	<i>Sonar instruments.....</i>	<i>12</i>
2.1.3	<i>Multibeam sonar for water-column measurements.....</i>	<i>16</i>
2.2	DATA MINING AND PATTERN ANALYSIS	18
2.2.1	<i>The data mining process.....</i>	<i>18</i>
2.2.2	<i>Spatio-temporal hydroacoustic data</i>	<i>20</i>
3	DATA PREPROCESSING.....	23
3.1	OBJECTIVES.....	23
3.2	ACOUSTIC MODELING	24
3.2.1	<i>Concept.....</i>	<i>24</i>
3.2.2	<i>Model input.....</i>	<i>25</i>
3.2.3	<i>Acoustic ray tracing</i>	<i>26</i>
3.2.4	<i>Modeling multibeam sonar.....</i>	<i>27</i>
3.2.5	<i>Model output.....</i>	<i>29</i>
3.2.6	<i>Model validation.....</i>	<i>29</i>
3.3	MODEL INVERSION	34
3.3.1	<i>Concept.....</i>	<i>34</i>
3.3.2	<i>Model approximation</i>	<i>34</i>
3.3.3	<i>Deconvolution for real-world data.....</i>	<i>37</i>
3.3.4	<i>Deconvolved multibeam sonar data</i>	<i>38</i>
3.4	SCATTER NODES	43
3.4.1	<i>Definition.....</i>	<i>43</i>
3.4.2	<i>Feature extraction</i>	<i>44</i>
3.4.3	<i>Bathymetric soundings as scatter nodes</i>	<i>47</i>
3.4.4	<i>Scatter nodes from single beam sonar data</i>	<i>48</i>
3.5	OUTCOMES	49
3.5.1	<i>Data compactness.....</i>	<i>50</i>
3.5.2	<i>Usability</i>	<i>51</i>

- 4 PATTERN ANALYSIS 53
 - 4.1 OBJECTIVES 53
 - 4.2 EXPLORATORY DATA ANALYSIS 53
 - 4.2.1 *Concept* 53
 - 4.2.2 *Visualizing scatter nodes*..... 54
 - 4.2.3 *Echoview*..... 56
 - 4.2.4 *Eonfusion*..... 59
 - 4.2.5 *Other packages*..... 59
 - 4.3 SPATIAL CLUSTERING..... 60
 - 4.3.1 *Concept* 60
 - 4.3.2 *Overview of clustering methods*..... 61
 - 4.3.3 *Spatial clustering with DBSCAN* 69
 - 4.4 KERNEL METHODS FOR CLUSTERING 80
 - 4.4.1 *Concept* 80
 - 4.4.2 *Statistical learning theory*..... 83
 - 4.4.3 *Kernel methods*..... 85
 - 4.4.4 *The Hahn-Banach theorem* 89
 - 4.4.5 *Kernels for spatio-temporal feature vectors* 90
 - 4.4.6 *Clustering with kernels* 92
 - 4.4.7 *Clustering scatter nodes using kernel methods*..... 98
 - 4.5 UNDERSTANDING SCATTER NODE PATTERNS..... 100
 - 4.5.1 *Segmentation*..... 100
 - 4.5.2 *Classification*..... 100
 - 4.5.3 *Visualizing patterns*..... 102
 - 4.5.4 *Measuring patterns* 103
 - 4.5.5 *Assessing pattern quality*..... 104
 - 4.6 OUTCOMES..... 106

- 5 CASE STUDIES 109
 - 5.1 MODELED DATA 109
 - 5.1.1 *Description of the data set* 109
 - 5.1.2 *Analysis* 110
 - 5.1.3 *Results* 112
 - 5.2 SALMON BANKS..... 112
 - 5.2.1 *Description of the data set* 112
 - 5.2.2 *Analysis* 113
 - 5.2.3 *Results* 118
 - 5.3 LAKE OPEONGO 120
 - 5.3.1 *Description of the data set* 120
 - 5.3.2 *Analysis* 120
 - 5.3.3 *Results* 123
 - 5.4 SOUTHERN OCEAN 123
 - 5.4.1 *Description of the data set* 123
 - 5.4.2 *Analysis* 124
 - 5.4.3 *Results* 126

6 CONCLUSIONS..... 131

6.1 SPATIO-TEMPORAL HYDROACOUSTIC DATA MINING 131

6.2 AN EXTENSIBLE FRAMEWORK 134

6.3 SUMMARY 136

APPENDIX: ABSTRACTS OF PUBLICATIONS..... 137

REFERENCES 141

1 INTRODUCTION

1.1 MOTIVATION

In the 1990s, the first results of conducting fisheries research studies using multibeam sonar were published (Misund and Aglen, 1992; Soria *et al.*, 1996; Gerlotto *et al.*, 1999; Noettestad and Axelsen, 1999). By the turn of the century, it was clear that this new approach offered new possibilities and would lead to significant advances in fisheries research. While standard data processing and analysis methods were established for data collected using single beam sonar, no such methods were available for multibeam sonar data. In 2002, the research project that has led to this thesis was started, with the aim of developing a data processing and analysis methodology for multibeam sonar data. Such methods must be capable of handling the large data volumes that multibeam sonars collect, and be applicable to data from a wide range of instruments. The methods should derive useful information from the data, in a fashion that facilitates the combination of multibeam data with other data sets, for an integrated, ecosystem-based approach to the study of aquatic environments.

In this introductory chapter, the context of the project is presented, followed by a description of the problem addressed, and the research objectives. The final section of the chapter gives a synoptic overview of the remainder of the thesis.

1.2 CONTEXT

Exploration and exploitation of the oceans and the natural resources they contain has been important for a long time, and affects many aspects of our society. Studying the dynamics of the water masses and the life they harbour contributes to our understanding of the global ecosystem and related issues, including climate change. Many important industries are based on ocean exploitation. These include the oil and gas industries and commercial fisheries. Increased human activity is putting pressure on the ocean environment. Sustainability is, therefore, a key aspect of contemporary marine resource management.

The latest edition of the United Nations Environment Programme (UNEP) publication *Global Environment Outlook* (UNEP, 2007) articulates a number of important messages with respect to aquatic ecosystems, including the following.

- Continued overexploitation of fish stocks affects human well-being. Implementation of policy responses to this issue enhances human health, socio-economic growth and aquatic environmental sustainability.
- The world's oceans are the primary regulator of global climate, and an important sink for greenhouse gases.
- Aquatic ecosystems continue to be heavily degraded, putting many ecosystem services at risk, including the sustainability of food supplies and biodiversity.
- A continuing challenge for the management of water resources and aquatic ecosystems is to balance environmental and developmental needs.

The main reasons for the decline in fish stocks (Figure 1.1) are a combination of unsustainable fishing, habitat degradation and global climate change. Declining fish stocks do not only cause loss of biodiversity, but they have serious implications for human well being too, with fish providing more than 2.6 billion people with at least 20 per cent of their average per capita animal protein intake (UNEP, 2007).

According to the United Nations Food and Agriculture Organization (FAO), a global shortage of fish supply is expected; fish prices are forecast to increase (FAO, 2006). While pollution, shipping, military activities and climate change threaten marine biodiversity and ecosystems, fishing currently presents the greatest threat (Gjerde, 2006).

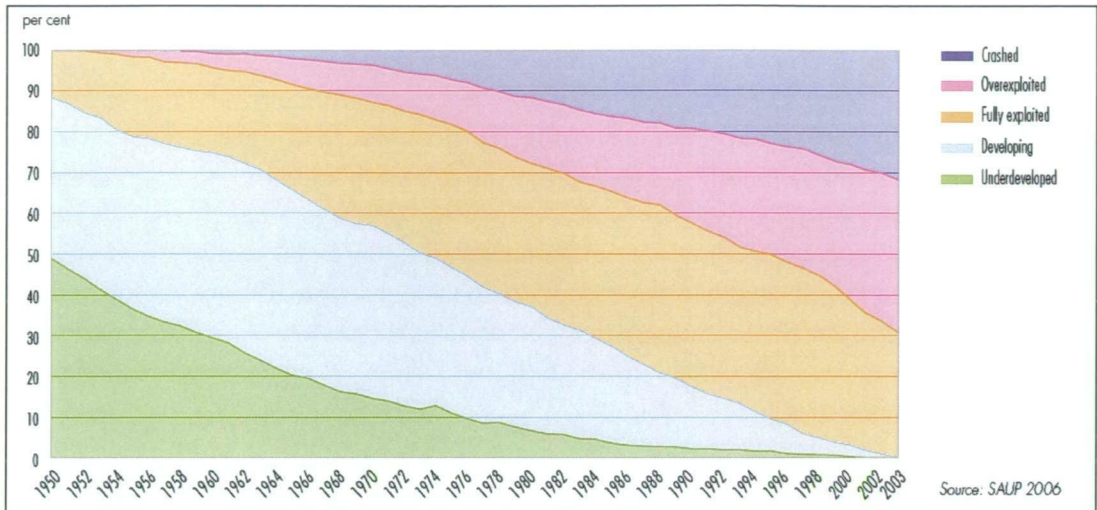


Figure 1.1 Exploitation status of marine fish stocks. (Image from UNEP GEO-4, 2007; source: Sea Around Us Project (SAUP) 2006.)

The *ecosystem-based approach* to natural resource management is a major principle underlying modern management practices (De la Mare, 2005; Garcia and Cochrane, 2005; Frid *et al.*, 2006). It was adopted by the Parties of the 1992 Convention on Biological Diversity (CBD) as a strategy for the integrated management of land, water and living resources, that promotes conservation and sustainable use in an equitable way (Gjerde, 2006). A key element of ecosystem-based management is the establishment of *Marine Protected Areas* (MPAs). The CBD defines a marine protected area as “any defined area within or adjacent to the marine environment, together with its overlaying waters and associated flora, fauna and historical and cultural features, which has been reserved by legislation or other effective means, including custom, with the effect that its marine and/or coastal biodiversity enjoys a higher level of protection than its surroundings.”

Marine science and technology are developing at a fast pace; they provide the necessary input and support for natural resource management and policy decisions. There is an urgent need to apply new scientific insights to the management of the global aquatic environments. In fact, much of the current understanding of the open ocean and deep seabed stems from explorations carried out in the last five to ten years, according to a United Nations report (Gjerde, 2006). International research projects and global cooperative efforts, such as the Census of Marine Life (CoML) (O'Dor, 2004; Yarincik and O'Dor, 2005) are helping to assess and explain the changes in past and present diversity, distribution and abundance of marine species, and to protect future ocean life. Transnational organisations and conventions such as the International Council for the Exploration of the Sea (ICES) and the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) coordinate international marine research focused on specific regions

such as the North Atlantic, the Baltic Sea and the North Sea in the case of ICES, and the Southern Ocean and the Antarctic in the case of CCAMLR.

A wide range of sensing and measuring devices and instruments is deployed in the oceans, collecting very large amounts of scientific data. Measurements are made at varying spatial and temporal scales. These measurements lead to an understanding of the systems directly, or they can be used as input to mathematical models. Quantities of interest include ocean temperatures, currents, salinity and acidity levels, seabed depth and habitats. Observations of marine life are conducted by tracking tagged individuals or estimating abundance and spatial distributions of stocks. Instruments can be mounted on buoys, ships, or underwater vehicles. A graphic impression is given in Figure 1.2.

Scientists from various disciplines are in need of tools, algorithms and systems to process and analyse these often disparate data sets, and to discover systematic patterns that can explain a system as complex as the global oceans. The work reported in this thesis represents a significant contribution to this field of research.

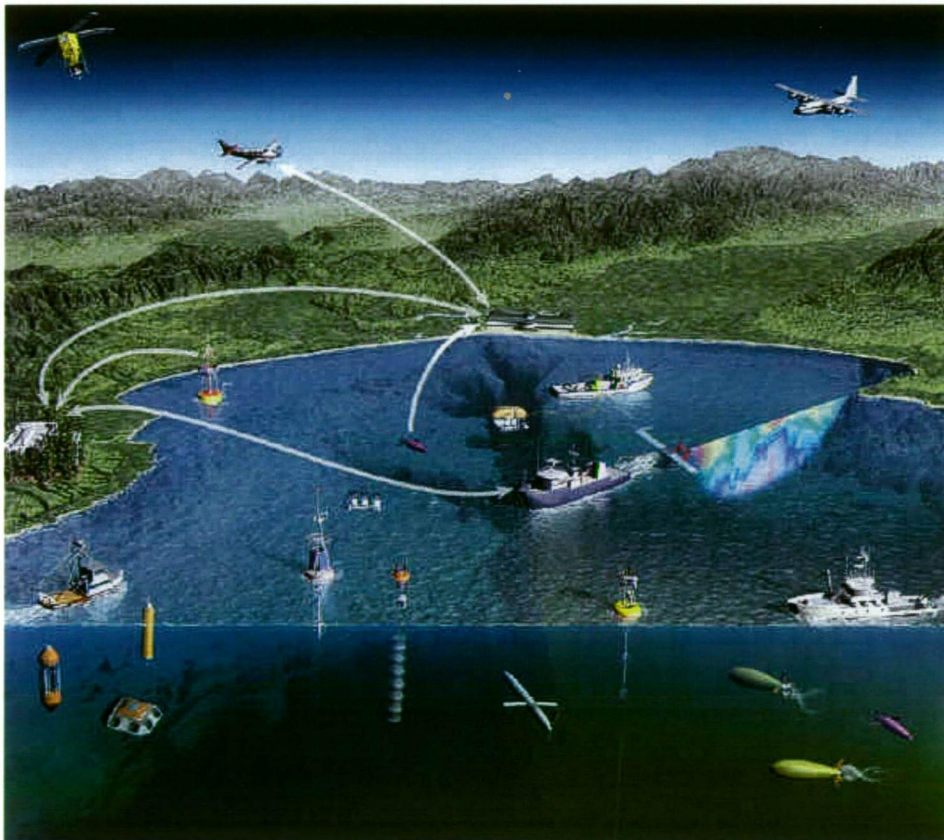


Figure 1.2 Depiction of a deployment of a multitude of sensors to observe the aquatic environment. Copyright: Monterey Ocean Observing System (MUSE Project).

1.3 PROBLEM DESCRIPTION

Acoustics is a common method used to study the underwater environment. Electromagnetic waves such as visible light are attenuated rapidly in water, whereas acoustic waves can propagate over long distances and penetrate to large depths. Acoustics is based on the examination of the characteristics of reflected sound (Urick, 1983). With respect to fisheries, acoustics is a more cost-effective and less intrusive method to conduct stock assessments than catching fish. Fisheries acoustics is the research area that studies the use of underwater sound to study fish, their behaviour, spatial distribution, and abundance (Simmonds and MacLennan, 2005).

Acoustic instruments for observing the aquatic environment are generally referred to as sonars. Echosounders, which are sonars with a single beam looking downward, are the standard acoustic devices used in routine fisheries surveys. Multibeam sonar systems have multiple beams pointing in different directions. Their use in fisheries research is relatively new. The use of multibeam sonar is well established for hydrographic applications, to measure the bathymetry or seabed depth. Sonar instruments designed for that purpose were typically not capable of collecting sound echoes from the water column, which is the body of water between the seabed and the transducer. However, most modern multibeam sonar systems designed for bathymetric applications can also collect data from the water column.

Multibeam sonars collect large amounts of data. Since the data are collected underwater using acoustic devices, the term *hydroacoustic data* is commonly used. Furthermore, the acoustic measurements are spatially and temporally referenced, hence the term *spatio-temporal hydroacoustic data*. Not only the data volumes but also the different instruments used to collect the data pose challenges in terms of data processing and analysis (Buelens *et al.*, 2006). A recent ICES report refers to this problem as the *data bottleneck* (ICES, 2007b). A need exists to reduce this bottleneck: effective, fast, automated algorithms are needed to process and analyse the data into an intelligible, informative and manageable representation. This is the problem that is addressed in this research.

1.4 RESEARCH OBJECTIVES

The central objective of this thesis is the development of a *data mining* process for the hydroacoustic data obtained by the new generation of multibeam sonar instruments capable of collecting data from the water column. In particular, the use of such data for fisheries applications is a primary point of focus.

A data mining process is a data handling and manipulation procedure leading to new insights in the data and what they represent (Cios *et al.*, 1998). In fisheries research, these insights will contribute to improved biomass estimates and stock assessments, to a better insight in schooling behaviour, and generally a better understanding of ecosystems in which fish populations are an essential component.

The data mining procedure is required to be able to handle the large amounts of data from various sonar instruments in a generic fashion. It must lead to an informative representation of the data in such a way that relevant higher level structures and concepts become available through the application of versatile and sophisticated algorithms.

1.5 THESIS SYNOPSIS

For this thesis, the two most important fields of research are underwater acoustics and data mining. General overviews of these subjects are given in chapter 2.

The data mining process that is presented in this thesis consists of two main phases: a data preprocessing phase and a pattern analysis phase, which are developed in chapters 3 and 4 respectively.

Chapter 5 contains case studies in which data are processed using the proposed data mining approach. Examples of modeled data, and real multibeam sonar and single beam echosounder data are given.

Final conclusions are drawn in chapter 6.

2 BACKGROUND

Many great advances in applied research occur when researchers from traditionally separate fields work together and combine their knowledge. One such area of research is fisheries acoustics. It has been a multidisciplinary field of research for decades. Contributors come from various disciplines including:

- physics (acoustics),
- engineering (instrumentation such as transducers),
- statistics, mathematics and computer science (data analysis), and
- biology, ecology and oceanography (users of the systems and the data).

Section 2.1 describes the field of underwater acoustics and the role multibeam sonar has started to play in recent years, particularly in fisheries research. Section 2.2 provides a general background on data mining and its role in the analysis of large data sets in general and spatio-temporal hydroacoustic data sets in particular.

2.1 UNDERWATER ACOUSTICS

2.1.1 Underwater acoustic measurements

Underwater acoustics enables the detection and location of fish, and the measurement of the bottom depth. Techniques to determine characteristics of fish

such as size, age, or species exist and are a topic of ongoing research. Similarly, the determination of bottom characteristics is possible using acoustic techniques. General overviews of fisheries acoustics are given in Simmonds and MacLennan (2005) and Misund (1997). Determination of the bathymetry (seafloor depth) is commonly achieved using multibeam sonar (de Moustier, 1988; Hughes Clarke *et al.*, 2000). A good overview of the state of the art in acoustic seabed characterization is presented in a recent ICES cooperative research report (ICES, 2007a). A general text on underwater acoustics is Urick (1983), and on acoustics Crocker (1998). This section is based on the references quoted, unless indicated otherwise.

Acoustics is the theory of sound propagating through a medium subject to scattering, reflection and absorption. Sound waves can propagate through water because of its elasticity which allows periodic compression and expansion. A sinusoidal sound wave is characterized by its frequency f , which is the number of cycles per second with which the pressure p varies relative to the ambient pressure level.

The sound speed c describes the speed with which wave fronts, or pressure peaks, move through the medium. The wavelength λ is the distance between two consecutive peaks. The following relation holds:

$$c = \lambda f. \quad (2.1)$$

The sound speed is dependent on the medium. For water, c is typically in the range 1450-1550 m/s, depending on water temperature, ambient pressure and salinity. The wavelength poses a limit on the spatial resolution of targets when observed using acoustic instruments.

Sonar instruments transmit pulses comprised of a few cycles of a sine wave that lasts for a finite time: the pulse duration. Such a pulse is also referred to as a ping. Sonars commonly transmit pulses at frequencies centred in a narrow band around a centre frequency f_0 . This centre frequency is the frequency that is quoted when discussing sonar instruments. For most purposes the signals such systems generate are treated as single frequency signals. Wide band systems transmitting signals at a range of frequencies, or chirp systems which vary the frequency during transmission are not discussed in this thesis.

The pulse duration and the wavelength of the signal determine the resolution of a sonar in the along-beam direction. The across-beam resolution is dependent on the angular beam width and the range.

A sound wave causes the water particles to vibrate. The amplitude of this particle movement is called the particle displacement, and the rate of change is the particle

velocity v . In a plane wave, the pressure and particle velocity both vary as a sine wave and are in phase. The pressure, p , is related to the velocity, v , by the formula:

$$p = \rho c v \quad (2.2)$$

where ρ is the water density. When a small source generates a wave, the wave fronts travel away from the source in all directions, in a spherical manner, and the wave is not planar. The relation between pressure and particle velocity is more complex in that case, and depends on the wavelength and the distance from the source. The far field of a source is determined by the distance at which the relation between pressure and particle velocity can be approximated by relation (2.2), a planar wave; the approximation does not apply in the so-called near field, or Fresnel zone.

A travelling wave carries energy. The flux J is the energy of the wave passing through a unit area perpendicular to the wave front. The intensity I is the energy flux per unit time. The intensity is the product of the pressure and the particle velocity:

$$I = p v = p^2 / \rho c . \quad (2.3)$$

Usually, the average intensity over one or more cycles of the wave is required, in which case the mean squared sound pressure is substituted for p . In particular, it is customary to work with root-mean-square (RMS) pressure amplitudes:

$$p_{RMS}^2 = \int_{1 \text{ cycle}} (p(t) - p_0)^2 dt . \quad (2.4)$$

In this thesis RMS pressure is assumed when using the term pressure or pressure amplitude.

The quantity:

$$Z = \rho c \quad (2.5)$$

is the acoustic impedance, which is almost constant over the sound path in typical underwater environments.

Sonar instruments measure the pressure, and convert the mechanical energy into electrical energy, effectively reporting the pressure as a voltage. Under the assumption that the impedance is constant, the squared voltage is proportional to the intensity, as is seen from eq. (2.3).

Sonars transmit pulses by means of a transducer, which converts an electric signal in an acoustic signal, thus generating a sound wave. Wave fronts travel outwards from the transducer, spreading spherically. In the far field, the intensity I varies with the inverse square of the range r :

$$I = I_0 / r^2 \quad (2.6)$$

where the range r is the distance to the transducer, and I_0 is the reference intensity, which is the intensity normalized to unit range.

Absorption is the loss of energy of a wave travelling through water. The lost energy is converted to heat. This is due to the particle movements, with higher frequencies incurring higher particle velocities. This is why low frequency waves penetrate deeper into the water, as they lose energy less quickly. The pressure, and hence the intensity of a wave decreases exponentially as:

$$I(r) = I_0 10^{-\alpha r / 10} \quad (2.7)$$

with α the absorption coefficient.

When a wave is transmitted by a transducer, it travels away from it to encounter a target such as a fish. A proportion of the energy of the incident wave is backscattered by the target. This backscattered wave travels in the opposite direction to that of the transmitted pulse, and is received some time later by the transducer. In this thesis it is assumed that the same transducer is used for transmission of a pulse and reception of its echo, or at least that the transducers are close enough to be considered the same in practical applications.

The backscattering cross section σ_{bs} is a measure of the proportion of incident energy that is backscattered by a target:

$$\sigma_{bs} = r^2 I_b / I_i \quad (2.8)$$

with I_b and I_i the backscattered and incident intensities respectively. The inverse square law for energy spreading means that σ_{bs} is a constant for a given target. The target strength TS is the logarithm of the ratio of the backscattering cross section and a reference area of 1 square meter (Clay and Medwin, 1977):

$$TS = 10 \log_{10} \sigma_{bs} . \quad (2.9)$$

The logarithmic measure TS is usually used to describe target strengths of aquatic organisms and is expressed in decibels.

For a target at a range r backscattering some of the incident energy, the time that elapses between the transmitted sound wave leaving the transducer and the backscattered signal arriving at the transducer is equal to the time needed for the sound wave to travel a distance of $2r$. Hence, when an echo is received at the transducer at a time t after transmission, the range to the target responsible for that echo is:

$$r = ct / 2 . \quad (2.10)$$

A pulse of duration τ , transmitted between times t_1 and t_2 , has a length in the range direction of:

$$ct_2 / 2 - ct_1 / 2 = c\tau / 2 . \quad (2.11)$$

Since two targets can be resolved only when each of them results in a separate echo pulse, it is clear that targets that are closer than a distance of $c\tau / 2$ cannot be observed individually.

In a situation when there are many targets close together, as is typically the case with fish schools, the targets form a combined return pulse, which does not allow for determination of individual targets of fish but which has an intensity that is still proportional to the combined target strengths of the individual scatterers (Foote, 1983). The volume backscattering coefficient s_v is defined as:

$$s_v = \frac{1}{V} \sum_{i \in V} (\sigma_{bs})_i \quad (2.12)$$

where V is the sampling volume and the sum is taken over all targets in V . The sampling volume is that volume for which targets within it are observable by the sonar. The logarithmic equivalent is commonly used, the volume backscattering strength S_V :

$$S_V = 10 \log_{10} s_v . \quad (2.13)$$

The importance of this theory is that it allows for the counting of the number of fish. When fish are not close together, the number of return pulses can simply be counted. In the other situation, considering eq. (2.12), and assuming that the distribution of the target strengths of the fish is known with expected value $\langle \sigma_{bs} \rangle$, eq. (2.12) can be written as:

$$s_v = \frac{n \langle \sigma_{bs} \rangle}{V} \quad (2.14)$$

with n the number of targets in the volume V .

In situations where the transmitted sound pulse cannot penetrate to the deeper layers of dense fish schools, the shallower scatterers are said to cause a shadowing effect. This effect can lead to underestimation of fish numbers.

The value of s_v is directly calculated from the voltage output from a calibrated sonar (section 2.1.3). Based on knowledge of, or assumptions about, the target strength of the observed fish, eq. (2.14) can be used to determine their number, n . The underlying theory is that of echo integration, which is not elaborated on in this context. A good discussion, together with references to the original literature on the subject, can be found in section 5.4 of Simmonds and MacLennan (2005).

Until now the targets have been assumed to be point targets, or fish. Similar theory applies of course to scattering from the seafloor. However, a number of important differences exist. Unlike relatively small scatterers, the seabed is fixed in that it is not displaced by the incident wave. The way in which the incident energy is backscattered is different: the seabed can absorb much of the incident energy, or let the energy penetrate to certain depths from which it is backscattered slightly later than from the seabed surface. The references provided at the beginning of this section provide a good background on the subject. In the present context only two facts are relevant:

- the return pulse from the seabed can be used to determine the depth, or bathymetry,
- the characteristics of the return pulse can be used to derive properties of the seabed surface, such as its roughness or hardness.

2.1.2 Sonar instruments

It is instructive to differentiate between sonars that transmit and receive on a single channel using a mostly narrow beam, and sonars that transmit and receive on many channels simultaneously. The former are known as single beam sonars, the latter as multibeam sonars. General references on this subject include Mitson (1983) and Medwin and Clay (1998).

Sonars generate electrical signals, which are converted to acoustic signals in the water by a transducer. The transducer contains a number of piezo-electric elements to convert electric signals to acoustic signals and vice versa. The same or possibly another transducer is used to convert the return signal, or echo, back to an electric signal. The electrical signal is digitized through sampling, and the samples are

either stored directly or transmitted on a computer network, where they can be picked up by data logging software.

Single beam sonars and echosounders have circular or elliptical transducers (Figure 2.1). The size of the transducer determines the beam width of the acoustic beam, given the frequency. Beams are typically 5-15 degrees wide. All the elements in the transducer are activated simultaneously by the same electric signal, and the received signals are summed to constitute a single output signal. There are a number of variations that offer more possibilities; two common ones used in fisheries research are:

- *Split beam echosounders* allow for separate reception on each quadrant of a circular transducer. Using the phase differences between halves (pairs of quadrants), the direction of arrival of the received signal can be determined, through which it is possible to locate targets in the beam in three dimensions: range, as usual, and additionally two angular coordinates off the vertical.
- *Dual beam echosounders* allow for separate transmission and reception on a circular subset of the circular transducer. The greater the diameter of the circle of activated elements, the narrower the beam. By using a wide and a narrow beam, the differences between the two signals can be used to determine how far off the central axis a target is located.



Figure 2.1 A 120 kHz transducer of a Simrad EK60 split beam echosounder.
Copyright: Simrad AS, Norway.

Signals received by single beam echosounders are sampled and stored to disk. The data consist of a series of received voltage signals, where each received signal is the echo from a transmitted pulse. Usually the signal is corrected for absorption and spreading losses through the variation of the gain in the sonar's amplifier with time;

this is known as applying a time-varying-gain or TVG. TVG corrected signals are shown visually by means of an echogram: a visual display, with range on the vertical axis and the transmit times on the horizontal axis (Figure 2.2). Each sample is coloured by its backscatter amplitude, where warmer colours indicate higher amplitudes (what the actual values are is not relevant in the present context).

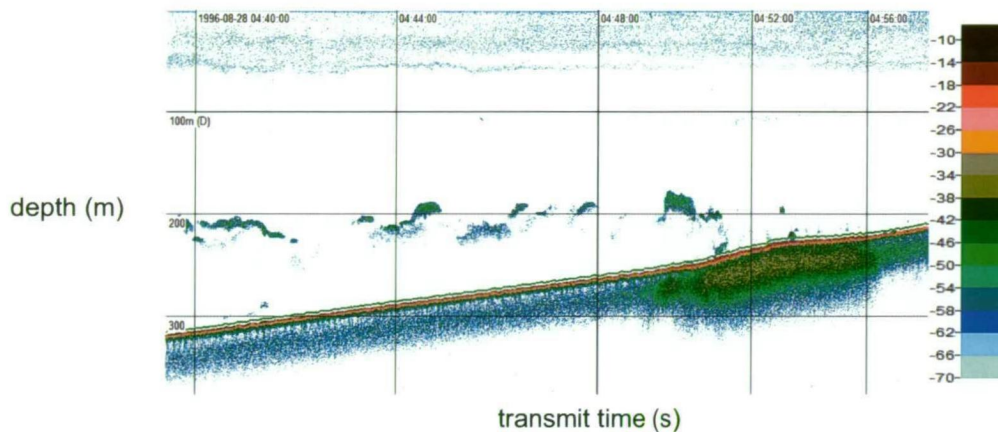


Figure 2.2 Example of a single beam echogram from a Simrad EK500 echosounder (S_V values in dB).

Multibeam sonars have transducers that consist of elements that are activated individually. The elements of such transducers are typically arranged in a linear flat or curved array. Since the beam width is very wide in the direction perpendicular to the direction of the array, it is customary to have one array for transmission, and another array perpendicular to the first for reception (Figure 2.3). This is known as a Mills-Cross array. Through this technique it is possible to attain narrow beams in both directions. The narrow beams are coplanar.



Figure 2.3 Transducer arrays of a Reson Seabat 8160. Copyright: Reson A/S, Denmark.

When transmitting, each element can be activated individually. This makes it possible to steer the beam electronically by changing the phase of the transmit pulse slightly from element to element. When this is done based on the input from a motion sensor, the beam can be stabilized for vessel motion such as pitch and roll. On reception, signals are received on the individual elements. The phases of these signals are used to form individual beams, pointing in different angular directions.

Until the early-to-mid 1990s, digital recording technology was not capable of outputting the complete signals for all elements. At the time, the signals were processed on dedicated Digital Signal Processing boards, which implemented algorithms to detect the bottom. The primary capability of such multibeam sonars was to get accurate bottom detections in each beam. The detections were output and stored to disk. Advances in technology have made it possible for complete multibeam sonar signals to be output, often together with the on-board determined bottom depths. The complete signal includes the backscatter returns from the water column.

The process of resolving the beams from the phase differences is known as beamforming, and is usually conducted prior to storing the data to disk, as is a TVG compensation. The backscatter amplitudes of the samples can be plotted in an echogram (Figure 2.4). All the data samples in such an echogram are collected from a single ping (one transmit-receive cycle). The corresponding single beam data are one vertical line of samples in Figure 2.2. In other words, a multibeam sonar collects a complete image as in Figure 2.4 for each vertical line of single beam data as in Figure 2.2.

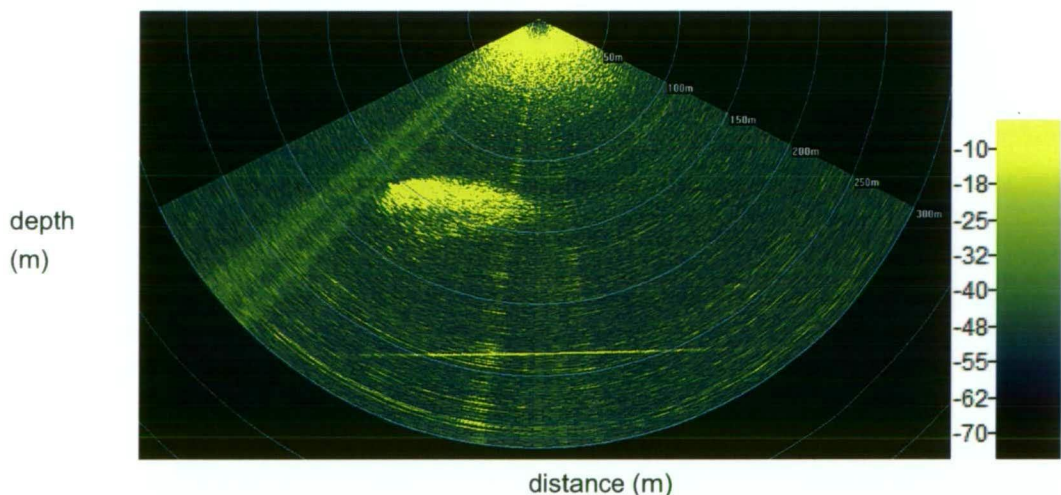


Figure 2.4 Example of a multibeam echogram from a Reson Seabat 7000 series model (S_v in dB, not calibrated).

Typical multibeam systems collect data for 100-300 beams simultaneously. As a result, multibeam data sets are typically two orders of magnitude larger than single beam data files for the same number of pings.

Another class of sonars have cylindrical or spherical transducers (Figure 2.5). Their primary purpose is finding fish at long ranges from the vessel. As opposed to the two designs discussed above, this type of sonar has beams pointing typically in a circular fashion away from the transducer, forming a conical shape. To differentiate these models from the multibeam sonar described above, they are referred to as omnidirectional sonars.



Figure 2.5 Spherical transducer head of the Furuno FSV-30 series. Copyright: Furuno Electric Co. Ltd., Japan.

2.1.3 Multibeam sonar for water-column measurements

Multibeam sonar is the best and most widespread instrument to determine bathymetry, and at the same time bathymetry is the most common use of multibeam sonar (de Moustier, 1988; de Moustier and Matsumoto, 1993; Chakraborty and Schenke, 1995; Hammerstad, 1995; Mitchell, 1996; Pratson and Edwards, 1996; Brissette, 1997). Bathymetric multibeam sonar is an active field of research, for example the processing of bottom detections (Brouns *et al.*, 2003; Calder, 2003; Canepa *et al.*, 2003) or the inclusion of the backscatter amplitudes for the bottom detections, which provides insightful information of bottom types, characteristics and seabed habitats (Clarke *et al.*, 1996; Keeton and Searle, 1996; Chakraborty *et al.*, 2001; Preston *et al.*, 2001; Fonseca *et al.*, 2002; Chakraborty *et al.*, 2003; Gallaudet and de Moustier, 2003; Hellequin *et al.*, 2003; Mayer, 2006; ICES, 2007a).

The capability for multibeam sonars to output the complete signal, including the backscatter echo returns from the water column, rather than just that from the seabed, is recognized as having great potential for fisheries research. The main advantages over using echosounders were identified as increased sampling volume without loss of resolution, and extra spatial information. The first studies using multibeam water-column data concentrated on fish behaviour, such as vessel avoidance (Misund and Aglen, 1992; Soria *et al.*, 1996), schooling behaviour (Gerlotto *et al.*, 1999), predator-prey interactions (Noettestad and Axelsen, 1999) and fish migration (Hafsteinsson and Misund, 1995). Behavioural studies have continued since, providing new insights that would not have been possible to achieve without multibeam sonar (Axelsen *et al.*, 2001; Johnson *et al.*, 2001; Benoit-Bird and Au, 2003; Gerlotto and Paramo, 2003; Gerlotto *et al.*, 2004; Brehmer *et al.*, 2006; Gerlotto *et al.*, 2006).

Behavioural studies are qualitative, in that the actual levels of the backscattered signals are not used to derive information about the observed organisms. In fisheries, an important aspect of surveys is estimating the number of fish, possibly specified to species or age group. When sonar is used for this purpose, it is said to be used in a quantitative manner. In order for quantitative work to be possible, a sonar has to be calibrated. Calibrating is the process of establishing values for parameters that are used in the calculation of target strength (TS) or volume backscattering strength (S_v) from the voltage signal from the sonar. One of the parameters in question is the on-axis sensitivity, which must be such that a target with a known target strength, placed in the centre of the acoustic beam, is observed by the sonar system to have that target strength. In a multibeam system, this has to be the case for each beam. The other parameter relates to the beam pattern of the acoustic beam, and is known as the equivalent beam angle, which is a measure of the beam width (Simmonds and MacLennan, 2005).

Calibrating single beam echosounders is common practice. The procedure to conduct a calibration is described in Foote *et al.* (1987). The calibration of multibeam sonar is more involved but not essentially different. A protocol for calibrating multibeam sonar is described in Foote *et al.* (2005). It is partially based on results from earlier preliminary multibeam calibration experiments (Chu *et al.*, 2001b; a; Cochrane *et al.*, 2003; Melvin *et al.*, 2003). There are outstanding issues related to the varying angular aspects of insonification of fish by multibeam sonar.

Since water-column data from multibeam sonar has three spatial dimensions, appropriate visualization tools are needed to present the data samples (Mayer *et al.*, 2002; Arsenault *et al.*, 2004; Wilson *et al.*, 2005). This aspect is elaborated on in section 4.2.

While water-column multibeam data have been used for more than a decade, there are a number of challenges that remain with respect to data processing and analysis (Buelens *et al.*, 2006):

- The data volumes are very large. Since one ping of multibeam data contains two orders of magnitude more samples than its single beam equivalent, storing and handling multibeam data are an issue.
- Standardization is needed. Calibration is a good step towards ensuring that data are independent of the conditions under which they were collected. However, differences in instrumentation make it difficult to compare data collected with different instruments, including other multibeam sonar models, or even single beam and omnidirectional sonars.
- Visualization is important. The spatial complexity of multibeam data means that it is difficult to represent graphically, especially when data covering large areas or long time spans must be visualized simultaneously.
- Automation is essential. Since the data volumes are large and visual inspection is less straightforward than with single beam data, automated algorithms are needed to detect noise or errors in the data.
- Segmentation or object detection algorithms capable of identifying subsets of the data as coherent structures automatically is essential to aid in data analysis.
- Classification algorithms based on previously segmented data would be very useful to reduce processing and analysis time.

These are computational challenges, with solutions to be found in the field of computer science. The next section discusses computational aspects of data analysis.

2.2 DATA MINING AND PATTERN ANALYSIS

2.2.1 The data mining process

Within the broad theme of computer science, covering areas as diverse as programming languages and operating systems, some fields of research concern the analysis of data by computers in an automated fashion. The informatization of society has instigated renewed interest in this field since the early 1990s, with the dawn of the internet and the increasing ease with which data can be collected,

stored and accessed. Large data sets are being created and collected continuously, from sales and bank transaction records to seismic sensor recordings, from surveillance camera video footage to meteorological satellite imagery. Analysis of such data sets is often not trivial. Many data sets are very large, preventing human expert investigation. Classical statistical analysis can be of use, but is often limited because of strong assumptions that are needed, such as normality of distributions, or linearity of problems or models. These assumptions are not needed in many of the computational methods arising from computer science research (Breiman, 2001; Cox *et al.*, 2001).

Data mining is the process of analysing data sets with the purpose of discovering previously unknown relationships or patterns. A number of introductory and review papers on the subject are available in the literature (Fayyad *et al.*, 1996; Cios *et al.*, 1998; Jain *et al.*, 2000; Smyth, 2000; Grossman, 2001; Ramakrishnan and Grama, 2001; Smyth, 2001; Ramakrishnan, 2003; Yao, 2003). The data mining process can be presented as a stepwise process (Figure 2.6) (Van Hulle, 2004). When the data that are the subject of the data mining processes arise from scientific experiments, measurements or models, the term *scientific data mining* is used (Fayyad *et al.*, 1996; Grossman, 2001).

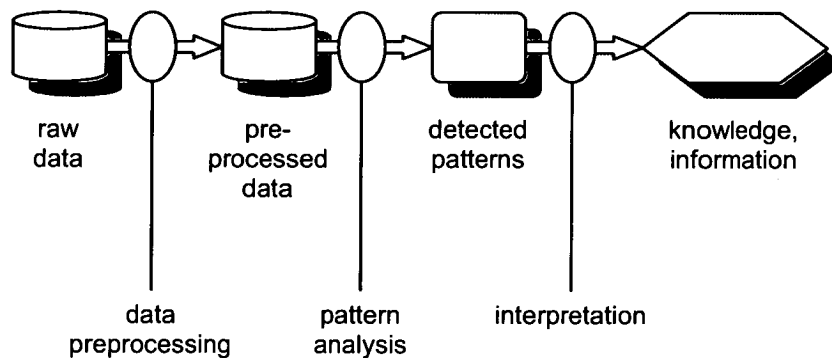


Figure 2.6 Schematic overview of the data mining process.

Prior to searching for patterns or regularities, input data sets are typically preprocessed with the purpose of representing the data of interest in a generic manner. Preprocessing is sometimes referred to as *feature extraction*, because an important aspect of data preprocessing is precisely the conversion of data into a representation by means of properties or features that best describe the data at hand.

Pattern analysis is the core component of the data mining process. Its aim is to find relationships in data sets. Many general pattern analysis algorithms have been published. The most common and important ones are found in standard reference books (Duda *et al.*, 2000; Hastie *et al.*, 2001; Bishop, 2006).

Patterns arise when elementary data units are related in such a way that the relationship conveys information. Some examples are:

- Pixels constituting an image: individual pixels convey very little, but when combined as an image, they convey information about the subject in the image. The pattern is what is shown in the image.
- Sounds combined together to form words: the order and transitions of spoken sounds gives them the meaning of words and sentences. The sequential order of sounds constitutes patterns conveying information.
- Credit card transaction data recorded by banks: individual transactions may look legitimate, but when considered together with other transactions may indicate fraudulent card use. The patterns in transaction records can indicate legitimacy of card usage.

Before an algorithm is capable of identifying patterns in data, it must be tuned to do so. This tuning is commonly referred to as *learning* or *training*. In order to train an algorithm to perform a certain task, data are needed. In *supervised learning*, a data set is available and the patterns within it are known. In *unsupervised learning*, the patterns are unknown and must be discovered by the algorithm. In the examples above, data sets and their corresponding patterns are: images and what they show, sounds and what they mean, sets of transactions and whether they are legitimate or not. Pattern analysis algorithms aim to detect such patterns in new data that were not used in training the algorithm. This is an important aspect of such algorithms: they must generalize well to unseen data (Hastie *et al.*, 2001).

The final phase of the data mining process consists of the interpretation of the detected patterns, which leads to new insights or information. This usually involves the graphical presentation of the raw or preprocessed data with an indication of the patterns. Where the data are spatial, as is the case in this research, it is customary to display the data in a spatial coordinate frame in two or three dimensions. Information visualization concerns the graphical representation of information, which is an important aspect of the final stage of the data mining process (Ware, 2004).

2.2.2 Spatio-temporal hydroacoustic data

The data sample values in spatio-temporal hydroacoustic data sets are backscatter energy amplitudes. In a typical deployment the sonar is mounted on a vessel carrying a positioning system such as a GPS, a compass, and a motion sensor collecting vessel attitude information such as pitch, roll and heave. Using the position, bearing and attitude information each data sample can be located in a

stored and accessed. Large data sets are being created and collected continuously, from sales and bank transaction records to seismic sensor recordings, from surveillance camera video footage to meteorological satellite imagery. Analysis of such data sets is often not trivial. Many data sets are very large, preventing human expert investigation. Classical statistical analysis can be of use, but is often limited because of strong assumptions that are needed, such as normality of distributions, or linearity of problems or models. These assumptions are not needed in many of the computational methods arising from computer science research (Breiman, 2001; Cox *et al.*, 2001).

Data mining is the process of analysing data sets with the purpose of discovering previously unknown relationships or patterns. A number of introductory and review papers on the subject are available in the literature (Fayyad *et al.*, 1996; Cios *et al.*, 1998; Jain *et al.*, 2000; Smyth, 2000; Grossman, 2001; Ramakrishnan and Grama, 2001; Smyth, 2001; Ramakrishnan, 2003; Yao, 2003). The data mining process can be presented as a stepwise process (Figure 2.6) (Van Hulle, 2004). When the data that are the subject of the data mining processes arise from scientific experiments, measurements or models, the term *scientific data mining* is used (Fayyad *et al.*, 1996; Grossman, 2001).

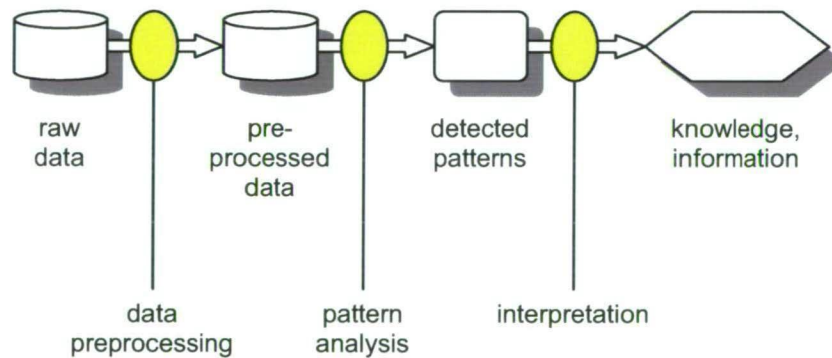


Figure 2.6 Schematic overview of the data mining process.

Prior to searching for patterns or regularities, input data sets are typically preprocessed with the purpose of representing the data of interest in a generic manner. Preprocessing is sometimes referred to as *feature extraction*, because an important aspect of data preprocessing is precisely the conversion of data into a representation by means of properties or features that best describe the data at hand.

Pattern analysis is the core component of the data mining process. Its aim is to find relationships in data sets. Many general pattern analysis algorithms have been published. The most common and important ones are found in standard reference books (Duda *et al.*, 2000; Hastie *et al.*, 2001; Bishop, 2006).

georeferenced space, by means of its longitude, latitude and depth below the water surface. In addition, the time at which each sample is collected is available.

There are large differences between the size of data files collected by single beam instruments and multibeam instruments. Some indicative values are listed in Table 2.1. In practice data rates and file sizes can vary because of a number of factors, including the disk space needed for one sample (6 to 14 bits is common), the number of samples per beam (depends on the range and sampling rate), the ping rate (depends on range and vessel speed), and the size of the meta data and non-acoustic data such as position (these are small compared to acoustic data, and are not included in Table 2.1). Processing multibeam files containing data from a multi-day survey remains a challenge even with high-end computer hardware.

	<i>Single beam</i>	<i>Multibeam</i>
Number of beams	1	120
Samples / beam	1000	1000
Ping rate	10 pings/sec	2 pings/sec
Ping size on disk	1 KB	120 KB
Data rate / minute	600 KB/min	14.4 MB/min
File size 24 hours of data	864 MB	20.7 GB

Table 2.1. Some indicative values of data rates and file sizes, assuming one byte per sample is needed.

Sonar manufacturers have not put substantial efforts in data compression, despite the fact that studies have indicated that suitable compression schemes exist, capable of compressing the data in a lossless manner to 70% of its original size (Wu *et al.*, 1997; Pitman, 2002). Data thresholding and resampling seem to be the methods of choice to reduce the number of data samples. While thresholding is aimed at removing samples containing no information from scatterers, resampling is aimed at reducing the resolution and retaining some information from all samples. More advanced methods are conceivable, such as identifying noise and side lobing artefacts and removing the relevant samples selectively. Identifying redundancies in the data, such as those caused by repeated sampling of the same volume, and selective removal of the relevant samples is another alternative. More research is needed for these advanced methods to become common practice; the methods developed in chapter 3 of this thesis deliver a contribution in this area.

To make use of multibeam data for fisheries applications, fish schools must be detected in the data. The relevant references in section 2.1.3 provide little detail about how data from fish schools was extracted from the data sets. Ad-hoc methods are used, again based on a combination of resampling either to a coarser resolution, as by Gerlotto *et al.* (2004), or onto a regular grid, as by Benoit-Bird and Au (2003). Schools detection algorithms have been developed in software, but are not widely documented. All these approaches are to some degree ad-hoc, require manual

intervention, are not suitable for large data sets, or are often not general enough to be applicable to data collected by different instruments. The methods developed in chapter 4 of this thesis offer an alternative approach.

Another important aspect that has not received much attention is that of standardization of multibeam water-column data, for easy sharing, storing and using of the data sets. There are a number of initiatives that are relevant in this context.

The ICES Working Group for Fisheries Acoustics Science and Technology (WGFAST) have edited and adopted a standard data format for fisheries acoustics raw and edited data: *HydroAcoustic data format* (HAC) (ICES, 2005). HAC is designed for single beam echosounder data, including dual and split beam, but does not currently cover multibeam data. It is unclear at this stage whether the format will be extended in the future.

For data sets to be exchangeable and distributable, good metadata is vital. Metadata is the description of the actual measurement data, such as names, units, scales, and descriptions. If data is to be shared at a global scale, global initiatives are needed. One such example is the Marine Metadata Interoperability project (MMI), established to promote the exchange, integration and use of marine data through enhanced data publishing, discovery, documentation and accessibility (MMI, 2007).

Hundreds of institutes and organizations world wide are making their marine and oceanographic data available through data centres or repositories. It is essential that data can be shared across data centres, which is facilitated by using the same data formats and metadata definitions. The Intergovernmental Oceanographic Commission (IOC) of UNESCO has been running its International Oceanographic Data and Information Exchange (IODE) facility since 1962 to enhance marine research, exploitation and development by facilitating the exchange of oceanographic data and information between participating member states and by meeting the needs of users for data and information products (IODE, 2007).

ICES has a working group on marine data management. The activities of this group include the establishment of guidelines with respect to data and metadata storage and access (ICES, 2006). It can be expected that formulated advice will be in line with the corresponding IODE guidelines.

An important facet of any water-column multibeam data processing scheme is that it results in data structures which lend themselves to straightforward archiving in data centres, in such a way that the data can be obtained and analysed easily by interested parties in the future. Ecosystem-based resource management considers all available data sources in combination, with attention being paid to correlations and interactions (Garcia and Cochrane, 2005).

3 DATA PREPROCESSING

3.1 OBJECTIVES

Modern computer hardware is capable of storing the large amounts of data collected by multibeam sonars on hard disks. However, hard disk access is still relatively slow, and data processing can be computationally intensive, particularly for some advanced and complex algorithms. It is therefore desirable to reduce the data volume while retaining as much information as possible. A second and equally important aspect is normalization across instruments and instrument settings. From the point of view of postprocessing analysis algorithms it is desirable to have the acoustic measurements in a unified form and format, from which any instrument-specific details are removed.

In this chapter, a data preprocessing algorithm based on acoustic model inversion is proposed. An acoustic and sonar model is presented in section 3.2; an approach to inverting the model is proposed in section 3.3. Section 3.4 discusses the resulting data representation. The outcomes of this data preprocessing algorithm are summarized in section 3.5.

3.2 ACOUSTIC MODELING

3.2.1 Concept

Given an underwater environment including aggregations of fish and the seabed, what would this look like when observed with a multibeam sonar instrument? In this section a forward model is developed that answers this question. The ultimate question is the reverse: given the data recorded by a multibeam sonar instrument, what did the underwater environment consist of in terms of scatterers such as fish schools and the seabed? In order to answer the latter question, the forward model must be inverted. This is discussed in the next section.

Besides the primary goal of establishing an analytical description of the process that must be inverted, a secondary benefit of a forward model is that it enables the creation of arbitrarily simple or complex data sets. This will prove very useful when evaluating inverse models, due to the difficulty of obtaining real-world ground truthed data sets.

The forward model incorporates two components: an acoustic model, in this case an acoustic ray tracing model, and a model of a multibeam sonar. The input to the model consists of a description of a three-dimensional underwater scene in which the multibeam sonar will be deployed. The output consists of a sequence of complex-valued sonar data sets, commonly referred to as pings. A schematic overview is shown in Figure 3.1.

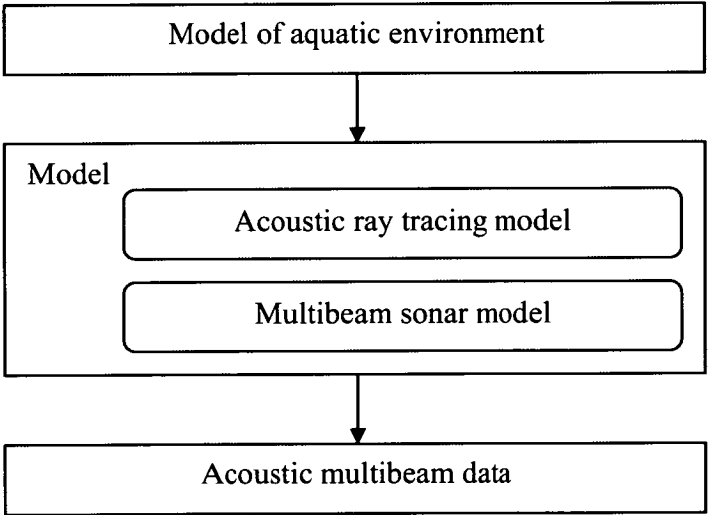


Figure 3.1 Overview of the forward model, its two components and its input and output.

3.2.2 Model input

The input to the acoustic model is a model of the aquatic environment consisting of a description of a three-dimensional underwater scene, containing a seafloor surface and volumetric objects representing aggregations of fish. It is assumed that the multibeam sonar is mounted on a vessel surveying the area of the three-dimensional scene. A trajectory for that imaginary vessel can be defined. In its simplest form, it is assumed that the acoustical characteristics of the seabed (hardness and roughness) are constant for the whole surface. Furthermore, it is assumed that the distribution of the number of fish in the fish schools is Poisson. The Poisson distribution is given by:

$$P_v(n) = \frac{v^n e^{-v}}{n!} \quad (3.1)$$

where v is the Poisson parameter, which in this case is equal to the average distance between individuals in the aggregation. The quantity $P_v(n)$ gives the probability of n individuals occurring in a unit volume within the school.

The acoustical properties of the seabed and the density of fish within a school is parameterized, as well as the target strength of the fish in the school. Both the seabed and the fish schools are modeled by individual point scatterers (Middleton, 1967; Ol'shevskii, 1967; Bell, 1997; Tillett *et al.*, 2000). In the case of the seabed, the scatterers are placed close enough to each other for the model to treat it as a solid surface. The density of such points on the seabed surface depends on the frequency of the sonar, and is such that the mean distance between two points is less than a quarter of the wavelength of the sonar system. Fish schools are modeled as point targets in an enclosing volume shell. The target strength of the point scatterers is the average target strength of the fish species being modeled.

The point model is defined as

$$\Omega = \{p_i\} \quad \text{with } 0 \leq i \leq N,$$

and N the number of point targets p_i in the three-dimensional environment. A representation of the set Ω is given in Figure 3.2.

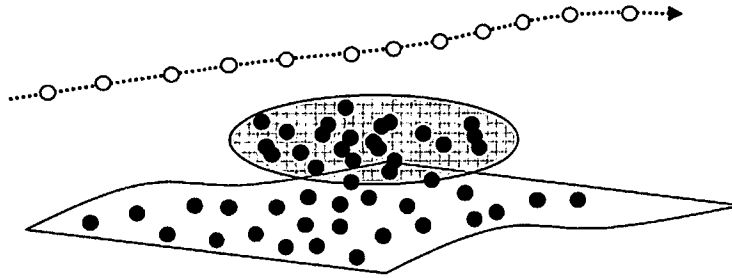


Figure 3.2 The black dots represent the p_i in Ω . The dotted line is the vessel cruise track; the white dots represent the locations where a multibeam ping will occur.

3.2.3 Acoustic ray tracing

Different standard acoustic computational models are described in the literature, including Urick (1983) and Crocker (1998). For the purpose of modeling multibeam sonar, acoustic ray tracing offers a computationally feasible and straightforward yet sufficiently sophisticated approach (Ziomek, 1989; Bell, 1997; Bell and Linnett, 1997; Etter, 2001). The ray tracing model computes the acoustic pressure at each element of the transducer face. Each pressure value is obtained by combining the responses of the scatterers in the point model. The following equation describes the ray tracing model:

$$p_{i,j}(t) = \sum_{k=1}^N \delta_{i,k} p_0 A_k W_t(t_k + t_{k,j}) 10^{-2\alpha r(k)/10} r(k)^{-4} + \eta(i, j, t) \quad (3.2)$$

with:

$p_{i,j}(t)$	the pressure received at time t in ping i by transducer element j ,
$\delta_{i,k}$	$\delta_{i,k} = 1$ if point k is in the transmit beam for ping i , $\delta_{i,k} = 0$ otherwise,
p_0	the reference pressure level (transmit pressure as measured at 1m from the transducer),
A_k	the proportion of the incident amplitude that is backscattered by point k :

$A_k = \sqrt{\sigma_{bs}(k)} / r(k)$ with $r(k)$ the distance from the array centre to point k , and $\sigma_{bs}(k)$ the backscattering cross section of point k ,

$10^{-2\alpha r(k)/10} r(k)^{-4}$ absorption and spreading loss, with $r(k)$ as above and α the absorption coefficient,

$\eta(i, j, t)$ an additive noise term, which in the model can be set to Gaussian, or zero,

$W_i(t_k + t_{k,j})$ the eikonal at time t , evaluated at time $t_k + t_{k,j}$, where t_k is the acoustic travel time from the centre of the transmit array to point target k , and $t_{k,j}$ is the travel time from point target k to element j of the receive array. W is a function of the transmitted pulse shape and pulse length:

$W_i(s) = \kappa(t - s)$ where $\kappa(\cdot)$ is the transmit pulse. For example, with $\kappa(\cdot)$ a block pulse, $\kappa(t) = e^{i\omega t}$ for $0 < t < \tau$, and 0 otherwise; τ is the pulse length and ω is the angular frequency, $\omega = 2\pi f$, with f the operating frequency of the multibeam sonar.

Calculation of t_k and $t_k + t_{k,j}$ requires knowledge of the sound speed c (provided as a parameter), and of the geometry of the multibeam transducer. Knowledge of the transducer arrays is assumed. Only first order scattering is considered as this is known to be the dominant effect (Foote, 1983).

3.2.4 Modeling multibeam sonar

A parameterized model of a generic, typical multibeam sonar is developed. The receiving transducer array is assumed to be a flat linear array. Its length and the number of individual transducer elements are parameterized, as well as its operating frequency. Known and published recommendations for transducer element sizes are adhered to: the element spacing l must be chosen such that $l \leq \lambda/2$, for a wavelength λ in order to avoid spatial aliasing (Knight *et al.*, 1981).

This sonar model allows for the simulation of any type of multibeam sonar where the beams are in the same plane, and are oriented so that they form a fan shape. Commercial instruments in this category include the Simrad Mesotech SM2000, the SM20, the Kongsberg EM series, the Reson Seabat series (6K, 7K and 8K models), and the Simrad ME70.

The acoustic ray tracing model, eq. (3.2), allows for calculation of the pressure levels at each of the transducer array elements, as a function of time. The multibeam transducer model will ‘measure’ these pressures $p_{i,j}(t)$ as voltages $V_{i,j}(t)$ for ping i , element j , at time t . A sampling rate can be chosen, and the voltages are digitally sampled accordingly. A time varied gain (TVG) is applied to the voltages to compensate for the absorption and spreading losses. TVG-compensated samples are written to disk by the model, and are referred to as the raw data. Discrete complex raw data samples are denoted by $c_{i,j,s}$, where i and j are as before, and s is the sample index for increasing ranges, $0 \dots S-1$ for S samples. With f_s the sampling frequency of the system, $c_{i,j,s} = p_{i,j}(s / f_s)$.

The raw data is subsequently beamformed. The beam former implemented in the model is the standard Fourier-based beam former (Rudnick, 1969). A beamformed complex data sample in ping i , beam j , range index s is obtained as:

$$d_{i,j,s} = \sum_{k,l} w(j,s,k,l) e^{i2\pi\phi_{k,j}} c_{i,k,l} \quad (3.3)$$

where the summation over k is over the transducer elements and over l is over the range indices of the samples, and:

$$\phi_{k,j} = d_k \sin(\alpha_j) / \lambda \quad (3.4)$$

where λ is the wavelength of the acoustic signal, d_k is the distance from the centre of the array to the centre of transducer element k and α_j is the angle of the central axis of beam j .

The function $w(j,s,k,l)$ is a windowing function. Different choices are possible, all with specific advantages and trade-offs, see for example (Curtis, 1998; Chu *et al.*, 2001b). Windowing functions not being the focus in this context, a simple windowing function is chosen:

$$\begin{aligned} w(j,s,k,l) &= 1 && \text{if } l = s, \text{ for all } k \text{ and } j, \\ w(j,s,k,l) &= 0 && \text{otherwise.} \end{aligned}$$

With this choice for w , equation (3.3) reduces to:

$$d_{i,j,s} = \sum_k e^{i2\pi\phi_{k,j}} c_{i,k,s} \quad (3.5)$$

with $\phi_{k,j}$ as in (3.4). The $d_{i,j,s}$ in (3.3) and (3.5) are the beamformed complex data samples s in ping i , beam j . It is customary to work with the amplitude or intensity

(squared amplitude), which is often the only information that is written to disk by real-world multibeam systems, $a_{i,j,s} = |d_{i,j,s}|$.

3.2.5 Model output

Some formalism is introduced. The set of points Ω is the input to the model, resulting in the beamformed data samples $d_{i,j,s}$. Defining $\Delta = \{d_{i,j,s}\}$ and denoting the analytical model with M , the modeling process is written as:

$$\Delta = M(\Omega). \quad (3.6)$$

The data generated by the model is referred to as *synthetic* data. Acoustic data are commonly represented graphically as an echogram: a colour-coded image of the signal amplitudes. An example of an echogram of beamformed synthetic data is given in Figure 3.3. This example is from the data file that is discussed in detail as a case study in section 5.1.

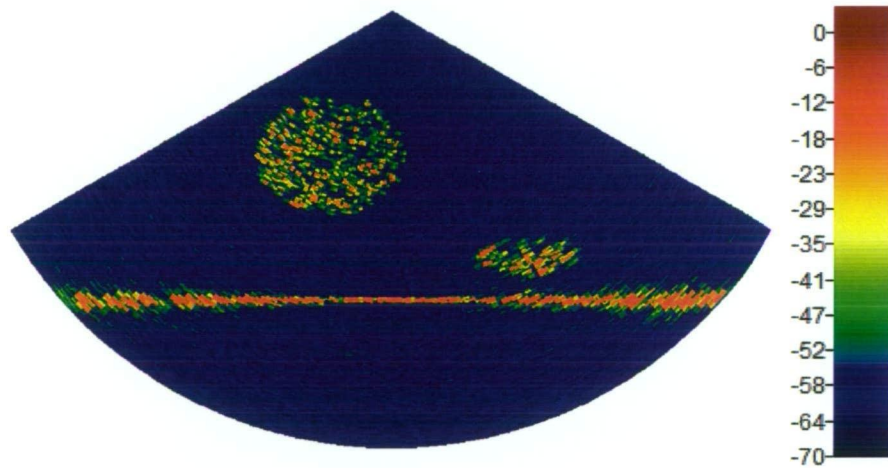


Figure 3.3 Echogram representing one ping of synthetic data, showing two aggregations of fish above a flat seabed (S_V in dB).

3.2.6 Model validation

The process of investigating whether the synthetic data generated by the model is representative of real multibeam sonar data is *model validation*. Two approaches to model validation are pursued: a statistical data analysis, and the simulation of a real-world data collection scenario.

Statistical validation

The simplest form of model validation is determining whether the synthetically generated acoustic data resemble true acoustic data. When represented as echograms, the human eye perceives the synthetic data as plausible, but an objective statistical measurement of similarity to real data must be made in order to substantiate such a claim. A criterion for similarity can be stated as (Bell and Linnett, 1997):

Definition 3.1 (Statistical similarity). Two acoustic data sets are defined to be statistically similar if their constituting amplitude values are likely to be drawn from the same probability distribution.

Theoretically, it is expected that the Probability Density Function (PDF) of full water-column multibeam amplitude data values follows the K-distribution (Di Bisceglie *et al.*, 1999; Chitroub *et al.*, 2002; Abraham and Lyons, 2004). The K-distribution is well established as a model for the amplitude statistics of scattered waves (Jakeman and Tough, 1987; Hongler, 1988; Jakeman and Tough, 1988; Lyons and Abraham, 1999), and is given by:

$$P_K(x) = \frac{4}{\Gamma(\nu)} \left(\frac{\nu}{\mu} \right)^{1/2} \left(\left(\frac{\nu}{\mu} \right)^{1/2} x \right)^\nu K_{\nu-1} \left(2 \left(\frac{\nu}{\mu} \right)^{1/2} x \right) \quad (3.7)$$

where $\Gamma(\cdot)$ is the gamma function, $K_{\nu-1}$ is a modified Bessel function of the second kind, of order $\nu-1$, and ν and μ are the parameters of the distribution. The parameter μ is the mean, and ν is the order parameter. The order parameter can be interpreted as the amount of coherent clutter in the data (Abraham and Lyons, 2002); coherent clutter arises when there are non-random aggregations of scatterers causing coherence in the echo return signal.

Given a data set, the parameters are estimated using the standard Maximum Likelihood (ML) method (Pesavento *et al.*, 1998). Whether the amplitudes of a given data set are K-distributed is assessed using the Pearson χ^2 -test, with the null hypothesis H_0 : ‘The distribution of the amplitude values follows a K-distribution’. First, a selection of multibeam pings from data sets collected by real instruments is tested, including data from a Simrad Mesotech SM2000 sonar (Figure 3.4 (a)). Structural noise must be avoided, since that can distort the amplitude sample distribution. Structural noise can be caused for example by interference with other acoustic instrumentation on board the vessel. The null hypothesis can not be rejected at the 5%-significance level on the basis of the data considered ($p = 0.0346 < 0.05$). Second, a selection of pings from the synthetic data set shown in Fig 3.3 and discussed in section 5.1 is tested (Figure 3.4 (b)) This again does not lead to a rejection of the null hypothesis at the 5%-significance level ($p = 0.0165 < 0.05$).

In Figure 3.4, histograms of real and synthetic data are shown, with the K-distribution PDF (3) overlaid, using the ML-estimates of the parameters and ν and μ . This statistical assessment shows that the synthetic data obtained through the model are similar to real data, according to definition 3.1. The target strength and spatial distribution of the scatterers do affect the parameters of the distribution, but do not affect the nature of it: amplitude values are K-distributed under sometimes very different scattering regimes.

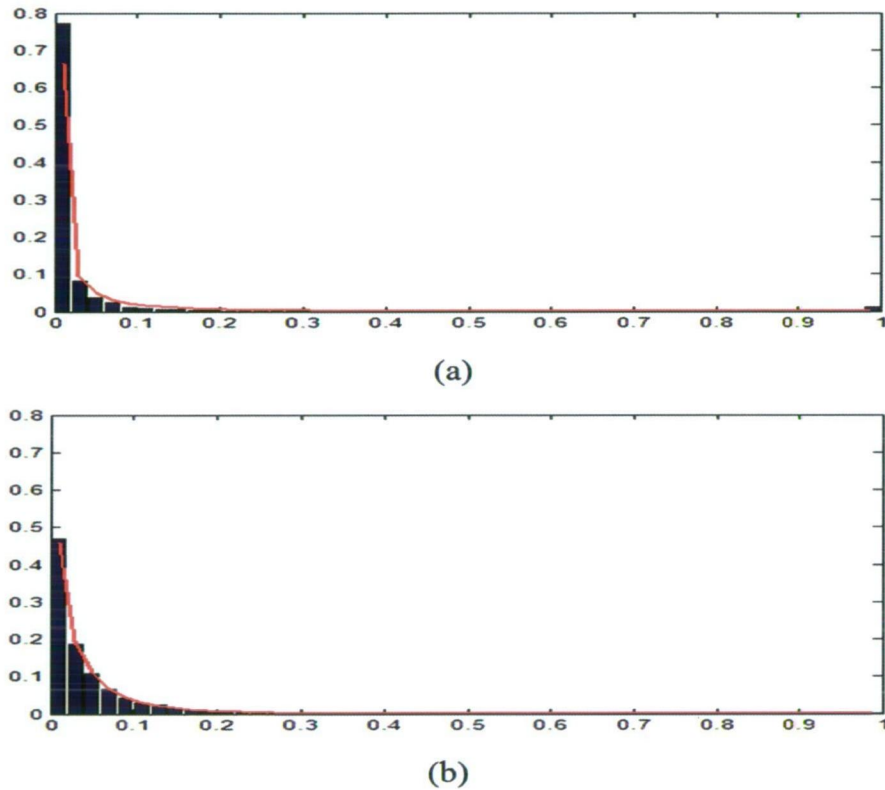


Figure 3.4 Histograms (bars) and ML estimates of the K-distribution PDF (lines) for (a) real data and (b) synthetic data. The synthetic data had more scatterers in the water column than the real data, explaining the more gradual drop-off in (b). The artefact in the bin at the value of 1.0 in (a) is due to the limited dynamic range of the instrument used (SM2000), causing saturation of the echo signal.

Simulation of a real-world data collection scenario

The statistical validation as described in the previous section is a general measure of similarity, only showing that the synthetically generated acoustic images exhibit the same statistical properties as real multibeam acoustic images. However, it does not prove validity of the model. Validity is less straightforward to define, and the following definition is used:

Definition 3.2 (*Validity*). The multibeam data collection model is valid if the resulting synthetic data resemble real data resulting from monitoring a real environment with a real multibeam sonar. The input to the model must be an accurate description of the real underwater scene.

This latter condition prevents thorough testing for validity, because, typically, an exact description of the underwater environment is not available. Further issues are that the actual multibeam sonar instrument used is likely to have peculiarities causing it to differ from the modeled system, that random noise in the real system affects the outcome, and that some simplifications and assumptions have been made in the model.

The only real-world multibeam data sets that are collected under controlled circumstances in a known environment are taken in test tanks, most commonly during calibration experiments. Such a data set was obtained (courtesy of Dr K. Foote and Dr D. Chu at Woods Hole Oceanographic Institution, USA). This data set contains full water-column data collected in a dock, in a controlled calibration experiment, with a Kongsberg Mesotech SM2000 multibeam system. A calibration sphere was moved through the beams in steps of 0.2 degrees, and kept at a constant range of 11 meters. A description of this scenario is assembled, and used as input to the simulation model. The resulting synthetically generated data set is studied in comparison with the real data set. A calibration is performed on both the real and the synthetic data sets (Cochrane *et al.*, 2003; Foote *et al.*, 2005), and the resulting calibrated sets are subsequently compared. At each angular position of the sphere, all the samples at a range of 11 meters are selected and stacked up to form a single image, representing a sweep of the calibration sphere through the beams. This is done both with the real data as well as with the synthetic data, and the resulting images are shown in Figure 3.5.

Unfortunately in Figure 3.5 (a), the ping rate and the movement of the sphere were not synchronised, resulting in an unequal number of pings per sphere position, explaining the slightly curved nature of the sphere trajectory as observed in Figure 3.5 (a).

It is found that the synthetic data resembles the real data to a satisfactory level. Furthermore, an analysis as described in the previous section shows statistical similarity as defined in definition 3.1. This section is concluded by accepting the model as valid per definition 3.2.

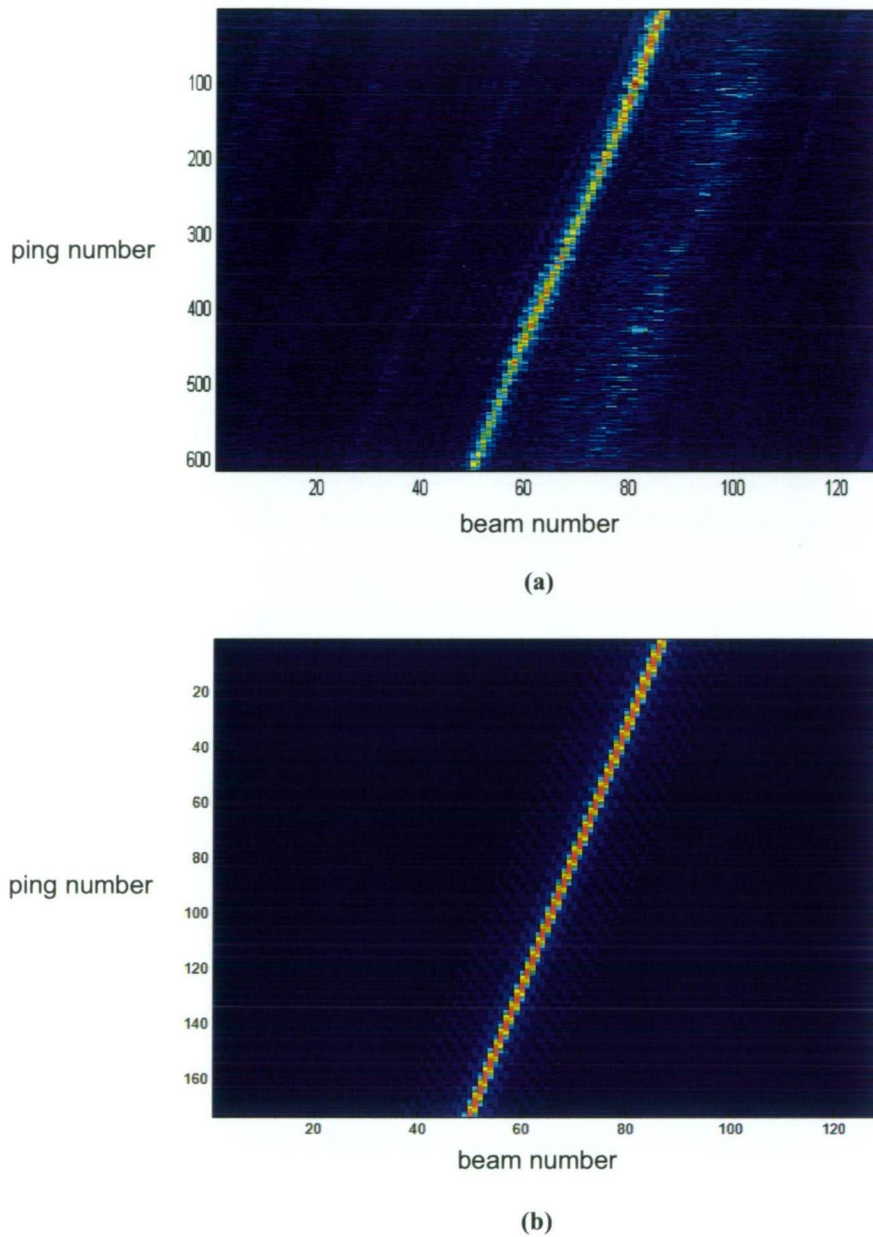


Figure 3.5 Images from (a) a real sphere, and (b) a modeled sphere. The sphere was moved from beam 50 to 86. The vertical axis indicates ping numbers. In (b), the sphere is moved 0.2 degrees from ping to ping. The structural noise observed in (a), to the right of the echo of the sphere was caused by the presence of concrete pylons in the dock.

3.3 MODEL INVERSION

3.3.1 Concept

Equation (3.6) describes a forward model. Given a description of an underwater scene, Ω , and the analytic model M , it is possible to calculate the data set Δ , which is the expected outcome of deploying a multibeam sonar measuring Ω . The real problem to be solved is the inverse. Given a data set obtained by deploying a multibeam sonar, what is an accurate description of the observed underwater scene? Inverting (3.6) formally gives:

$$\Omega = M^{-1}(\Delta). \quad (3.8)$$

Unfortunately, the model M is not analytically invertible because of the noninvertibility of the beamforming (eq. (3.3)), and the presence of random noise in the model (eq. (3.2)).

The situation where the inverse of a known model has to be determined is an inverse problem. There are various approaches to model inversion. The one that is followed here is to approximate M by an invertible function, F . If F is invertible, it is possible to calculate:

$$F^{-1}(\Delta) = \hat{\Omega} \quad (3.9)$$

with $\hat{\Omega}$ an estimate of Ω . $\hat{\Omega}$ needs to be a close approximation of Ω for F to be useful. It is essential to choose a model F which is invertible, and which approximates M closely.

3.3.2 Model approximation

A multibeam sonar is in fact a synthesis imaging instrument. Synthesis imaging is the generation (or synthesis) of an image based on signals received on multiple sensors, typically ordered in a sensor array. Various physical observation and measurement processes are forms of synthesis imaging, for example in astronomy (Starck *et al.*, 2002) and medical ultrasound imaging (Molthen *et al.*, 1995). Synthesis imaging systems are commonly modeled and described as convolutions (Rafaely, 2004), with the inverse being a deconvolution (Konstantopoulos *et al.*, 1990; Lingvall *et al.*, 2003; Lingvall, 2004). This approach has been applied to the study of fish target strength in the past (Clay, 1983), and is followed here.

The function F is chosen to be a convolution C , an approximation of the model M . The inverse problem (3.9) is now stated as:

$$\hat{\Omega} = C^{-1}(\Delta) \quad (3.10)$$

with C^{-1} representing a deconvolution. Assuming that a proper choice for C can be determined, (3.10) allows for the calculation of $\hat{\Omega}$, an estimate of the underwater environment measured by the multibeam sonar. However, deconvolution, as in (3.10), is an ill-posed problem. This can be understood intuitively by considering a convolution as a smoothing operation, filtering out high-frequency features. Two descriptions of an underwater scene that differ in the high-frequency features only, will result in the same convolved data set, hence the inverse problem has no unique solution and is ill-posed. In multibeam sonar, as in other synthesis imaging systems, this is in fact the case due to the limited resolution of the system.

A number of solutions to solve this ill-posed problem have been established in the literature (Starck *et al.*, 2002), and the problem is a topic of ongoing research (Lingvall *et al.*, 2003). Different approaches essentially enforce different forms of regularization of the problem. A standard yet powerful technique that has become commonly accepted in recent years is the so-called Lucy-Richardson algorithm, sometimes referred to as Richardson-Lucy, which is an expectation-maximization algorithm (Richardson, 1972; Lucy, 1974). This algorithm is known to be stable and does not generate artefacts unlike some other algorithms, such as the Wiener filter (Starck *et al.*, 2002). The Lucy-Richardson algorithm is used here to calculate C^{-1} . For simplicity the deconvolution is calculated on a ping-by-ping basis. A possible extension which is not pursued further here is to include across-ping deconvolutions.

Denote the observed data in a ping, obtained from the model or as the result of a multibeam deployment, by Δ_i . The assumption that the ping data Δ_i are obtained from a convolution of the latent variable which has to be estimated, $\hat{\Omega}_i$, with a point spread function (PSF) P is written as:

$$\Delta_i(x, y) = C(\hat{\Omega}_i(x, y)) = \int P(x_1 - x, y_1 - y) \hat{\Omega}_i(x_1, y_1) dx_1 dy_1 \quad (3.11)$$

with (x, y) the echogram coordinates given by (beam number, range cell). Using the conventional $*$ notation for a convolution, this can be rewritten as:

$$\Delta_i(x, y) = (P * \hat{\Omega}_i)(x, y). \quad (3.12)$$

The idea of the Lucy-Richardson algorithm is to calculate the most likely $\hat{\Omega}_i$ given the data Δ_i . This leads to an equation which can be solved iteratively:

$$\hat{\Omega}_i^{n+1}(x, y) = \frac{\Delta_i(x, y)}{(P * \hat{\Omega}_i^n)(x, y)} P^*(x, y) \hat{\Omega}_i^n(x, y) \quad (3.13)$$

where $P^*(x, y) = P(-x, -y)$. It has been shown that this iteration converges to the maximum likelihood solution (Dempster *et al.*, 1977).

For this algorithm to be applicable, the PSF must be known and must not contain any free parameters. In the application at hand, the PSF must be chosen so that $C(\Omega)$ is close to $M(\Omega)$. This can be achieved through the construction of a special input set for the model, one that consists of one point only. Let this data set be Ω_1 , see Figure 3.6 (a). The model that is used here is discussed in more detail as a case study in section 5.1.

Using the forward model, Δ_1 is calculated as $M(\Omega_1)$ (Figure 3.6 (b)). Δ_1 is a data set consisting of a single ping, which contains the acoustic image of a single scatterer as observed through the modeled multibeam system. The PSF of the convolution C is now defined in terms of Δ_1 , by choosing the local neighbourhood of the response in the output image Δ_1 . All sample values that are significantly different from zero must be included in the PSF. It follows by construction of C that $C(\Omega_1)$ will be a very good approximation of $M(\Omega_1)$. $C(\Omega_1)$ is shown in Figure 3.6 (c). Because of the additive nature of the model equation (3.2), this statement can be generalized to conclude that $C(\Omega)$ will be a good approximation of $M(\Omega)$, for any input set Ω . Finally, the PSF is used in the deconvolution algorithm to obtain $\hat{\Omega}_1$ as the most likely estimate of Ω_1 (Figure 3.6 (d)).

It must be emphasized that the acoustic image in Figure 3.6 (b) is the modeled raw multibeam data, and that the image presented in Figure 3.6 (d) is the preprocessed data obtained through the inversion method.

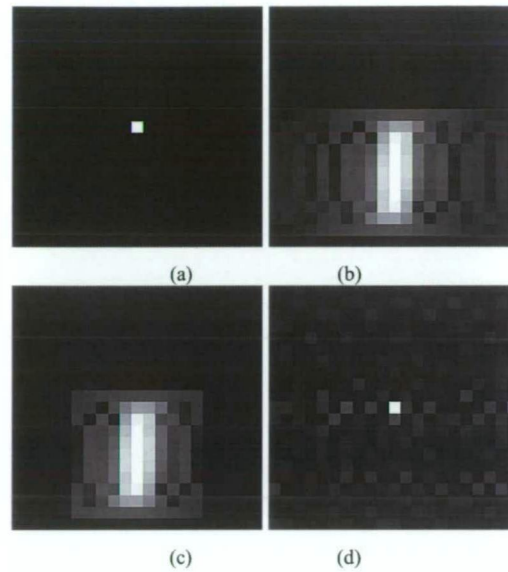


Figure 3.6 Normalized amplitudes (black = 0.0, white =1.0) for (a) an input point set consisting of a single scatterer; (b) the forward model applied to (a), from which the PSF is derived; (c) the forward convolution applied to (a) using the derived PSF; (d) deconvolution applied to (b) obtained via the Lucy-Richardson algorithm.

3.3.3 Deconvolution for real-world data

In the case of real data, rather than modeled data, the model M is not available. Information about real world sonar systems is not generally released into the public domain by instrument manufacturers, and hence it is not possible to model such systems accurately. Furthermore, the actual physical conditions of the underwater environment affecting transmission and scattering, such as the sound speed, water temperature, salinity, etc. are not always exactly known.

As explained in the previous section, finding C^{-1} is equivalent to finding an appropriate PSF. In the modeled data, the PSF is defined in terms of the output data of the model, without explicit knowledge of the model itself. For this to be possible with real data, an appropriate data set is needed. Such a data set must include the response of a single scatterer, and it must also be known where the scatterer was located in the acoustic beam at the time of the ping.

Placing a single scatterer, such as a calibration sphere, in the acoustic beam in a known location is part of the sonar calibration procedure (Cochrane *et al.*, 2003; Foote *et al.*, 2005). This means that in practice, anyone undertaking serious fisheries work with a multibeam instrument will have the required data set available to construct the PSF needed for the deconvolution C^{-1} .

In general, the response of a multibeam system is sensitive to the actual location of the point target. Calibration of a multibeam system is essentially a procedure to capture such variability, and includes the calculation of appropriate parameters to correct for this effect. It is anticipated that the variability in response is minimized in a correctly calibrated system, which means that the PSF derived from fully calibrated data will be well defined, although some angular averaging may be required.

However, in many real world situations, calibration information may not be available. This does not necessarily mean that the proposed inversion technique can not be applied to the collected data. Rather than basing the deconvolution on a known PSF, the PSF itself can be estimated during the convolution process. Applying a deconvolution without a known PSF is known as *blind deconvolution* (Tsumuraya *et al.*, 1994). This is an iterative procedure making use of a deconvolution algorithm and an additional maximum likelihood estimator for the PSF. Equation (3.12) is written in logarithmic terms as:

$$\log(\Delta_i) = \log(P) + \log(\hat{\Omega}_i). \quad (3.14)$$

This expression is used in the blind deconvolution algorithm to estimate the PSF. A variant of the Lucy-Richardson algorithm is used, where in each iteration both the estimates of $\hat{\Omega}_i$ and the PSF are updated:

$$\hat{\Omega}_i^{n+1}(x, y) = \frac{\Delta_i(x, y)}{(\hat{P}^n * \hat{\Omega}_i^n)(x, y)} \hat{P}^{n*}(x, y) \hat{\Omega}_i^n(x, y) \quad (3.15)$$

$$\log(\hat{P}^{n+1}) = \log(\hat{\Omega}_i^{n+1}) - \log(\Delta_i) \quad (3.16)$$

where \hat{P}^n is the estimate of the PSF in iteration n . A good initial estimate \hat{P}^0 of the PDF is helpful and should be provided where possible.

This technique enables the application of the inversion method to any multibeam data set.

3.3.4 Deconvolved multibeam sonar data

Applying the inverse model to either real or synthetic data results in a new data set, $\hat{\Omega}$. Since $\hat{\Omega}$ is obtained as the result of a deconvolution applied to a set of acoustic images, $\hat{\Omega}$ itself is a set of acoustic images, as in Figure 3.6 (d).

It is the set of images of the points constituting the point model that is an estimate of the underwater environment as observed by the multibeam system. In order to

obtain the points themselves, simple thresholding (above the noise level) and retaining of local maxima is applied, yielding a set of points:

$$\Theta = \{s_i\}, i = 1 \dots n. \quad (3.17)$$

The points $\{s_i\}$ are the minimal set of scatterers needed to result in the data that were observed by the multibeam sonar. It is important to note that a scatterer s_i is not necessarily a true point scatterer in the water. Rather, it is a conceptual measurement indicating the presence of a general object in the water, which could be an extended or solid object, such as a dense fish school or the seabed.

Example

The model used to obtain the acoustic image in Figure 3.6 (b) is adopted. Hence the PSF shown in Figure 3.6 (c) applies and can be used in the deconvolution. A model input data set is created, and a subset consisting of 1085 points representing a fish school and a flat seabed is selected. All these points are in the acoustic beam of the single ping under consideration.

The output of the model is represented in Figure 3.7. Polar coordinates are used, with angle (or equivalently beam number) on the abscissa and range from the transducer on the ordinate axis. This representation has the advantage that all samples are equally large in the echogram (in terms of pixels in the image).

A deconvolution is applied, using the PSF applicable to the multibeam sonar model used (that from Figure 3.6 (c)). The result is shown in Figure 3.8. In this image, peaks appear as higher amplitude pixels, and are indicative of scatterers. The actual scatterers are obtained by thresholding and selecting the local maxima: see Figure 3.9.



Figure 3.7 Acoustic image obtained by running the forward model on a point set consisting of 1085 points, representing an aggregation of fish above the seabed. The coordinate system of this image is polar: angle on the abscissa and range on the ordinate axis. The samples are normalized amplitudes (white = 0.0, black = 1.0).

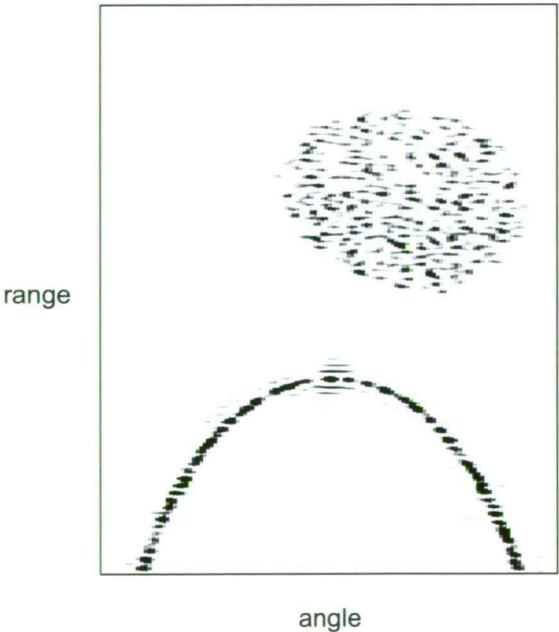


Figure 3.8 Deconvolved acoustic image obtained from applying the Lucy-Richardson algorithm to the acoustic image in Figure 3.7.

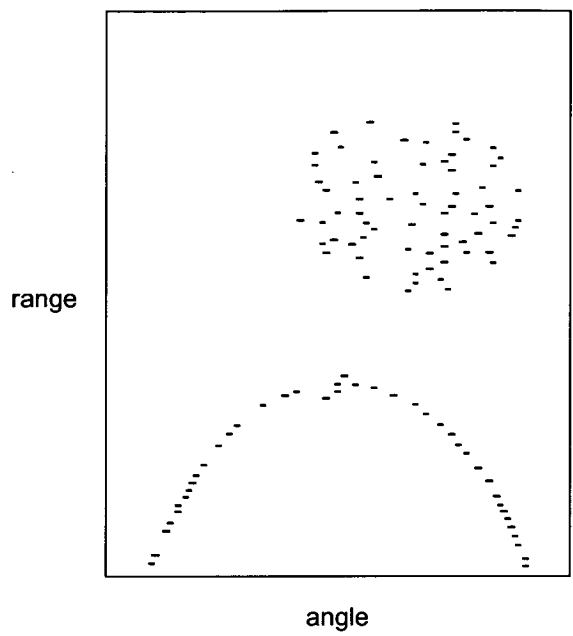


Figure 3.9 Scatterers obtained from thresholding the acoustic image from Figure 3.8 and retaining local maxima.

There are 116 scatterers. They provide an approximation of the underwater environment that was observed. It is irrelevant to compare this number with the number of input data points, as the input data points are independent of the model and are generally closer together than the resolution of the modeled system. That is in particular the case for modeling contiguous seabeds and dense fish aggregations.

The achievement of the inversion method is the conversion of the raw data in Figure 3.7 to the deconvolved data in Figure 3.9. The raw data in this case consists of 128 beams x 800 range cells, which equals 102,400 data samples. This data set is reduced to a representation by 116 scatterers only, a massive reduction of data size.

In fact, the ping-based data discussed so far in this example are one of a sequence of 35 pings. Considering all pings together and placing the input points as well as the resulting scatterers in a three-dimensional environment, a three-dimensional picture emerges (Figure 3.10).

This example demonstrates that the representation by means of scatterers (Figure 3.10 (b)) is a close approximation of the true scattering regime (Figure 3.10 (a)).

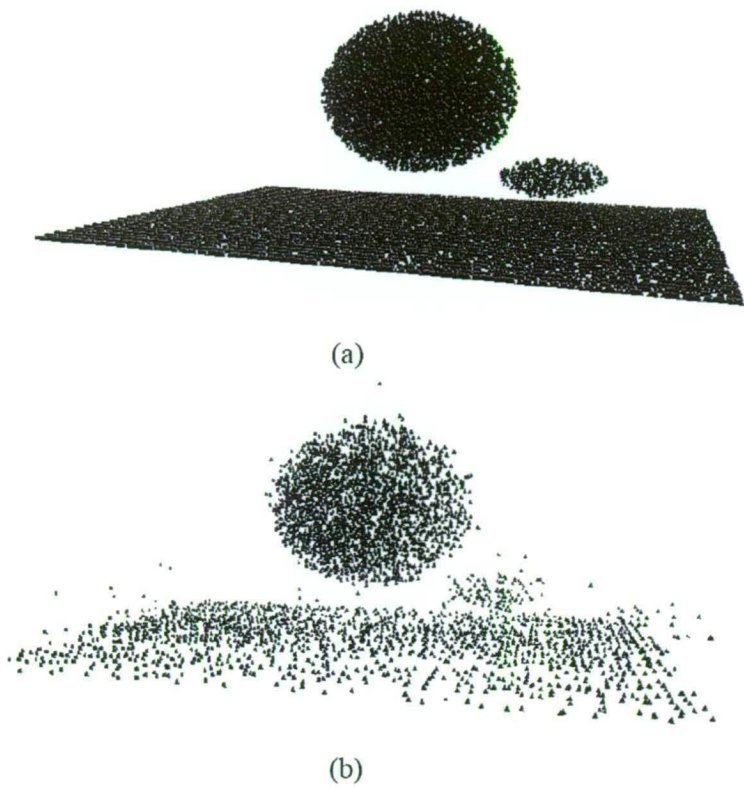


Figure 3.10 (a) An input point set, and (b) the corresponding representation by scatterers.

3.4 SCATTER NODES

3.4.1 Definition

Scatterers as presented in eq. (3.17) can be written in terms of their components as:

$$s = (\mathbf{x}, t, b) \quad (3.18)$$

with \mathbf{x} the spatial three-dimensional coordinates, t the time stamp, and b the backscatter energy for scatterer s . The components \mathbf{x} and t are derived in a straightforward fashion from the geometry and timing information in the original raw multibeam data set. The backscatter value b is the value from the sample at location \mathbf{x} and time t from the original data set; it is the backscatter amplitude from an assumed point scatterer which is located on-axis for that particular beam.

In order to retain information from the raw data set for use by subsequent postprocessing algorithms, additional features are extracted from the raw data set and are associated with the scatterers; the concept of *scatter nodes* is introduced.

Definition 3.3 (*Scatter nodes*) Scatter nodes are feature-rich spatio-temporal hydroacoustic data points.

Applying the deconvolution algorithm to multibeam sonar data results in scatterers. Such scatterers enriched with additional features are scatter nodes.

Quantities of interest are added to the scatter nodes as properties or features, to render the nodes as informative as required. These features are combined into a single feature vector \mathbf{v} of the scatter node s :

$$s = (\mathbf{x}, t, \mathbf{v}). \quad (3.19)$$

By means of scatter node features, information extracted from the raw multibeam data can be passed on to subsequent processing algorithms. Which features should be carried through may depend on the analysis that one wishes to perform. A number of features that can be extracted from the raw measurements are presented below.

3.4.2 Feature extraction

Backscatter energy

A key aspect in calibrated sonar for fisheries is the relation between backscatter energy and fish biomass. The proportion of transmitted energy that is received at the transducer is indicative of the fish density in the acoustic beam. This is well established for single beam sonar (MacLennan, 1990; Simmonds and MacLennan, 2005):

$$N \sigma_{bs} = s_v V, \quad (3.20)$$

with N the number of fish in a volume V , σ_{bs} backscattering cross section of an individual fish, and s_v the mean volume backscatter. The value of s_v is obtained from the raw data measurements through a process known as echo integration (Foote, 1987; MacLennan, 1990; Foote, 1991; Foote and Steffanson, 1993). Given estimates for σ_{bs} and V , and the measured s_v , relation (3.20) allows for the calculation of an (estimated) N .

At the time of writing, echo integration standards are not established yet for multibeam sonar. Promising results on calibration of multibeam systems are available (Chu *et al.*, 2001b; Melvin *et al.*, 2003; Foote *et al.*, 2005). Calculation of volumes to be used in (3.20) is discussed in Tang *et al.* (2006). The first experimental multibeam echo integration routines are made available in the Echoview software (Myriax, 2008). Research on these topics is ongoing and no text-book method is available yet.

Typically, raw data measurements are used and needed in the calculation of s_v . Stepping away from the raw data, and continuing further analysis based on scatter nodes must not compromise the options for echo integration or use of backscatter energy. Therefore the necessary information is extracted as features of the scatter nodes. Within a ping of raw multibeam sonar data for which scatter nodes are determined, let A_i be the set of raw samples which are nearest to scatter node i . A feature of the scatter nodes representing the integrated backscatter energy is then defined as:

$$\beta_i = \sum_{j \in A_i} w_j d_j \quad (3.21)$$

with d_j the raw data samples, properly calibrated, and w_j sample weights. The algorithms currently implemented in Echoview (Myriax, 2008) use the sample volumes as the weights w_j .

This way, the energy of each raw sample contributes to a scatter node exactly once: hence the total energy content of the ping is maintained:

$$\sum_{\text{samples } j} w_j d_j = \sum_{\text{nodes } i} \beta_i. \quad (3.22)$$

Such a scatter node feature will allow the scatter nodes to be used for biomass estimation. It is anticipated that future research outcomes on this subject will appear in the scientific literature. Such results will be transferable to the scatter node concept, through their encoding as scatter node features.

At this stage the concern is to capture as much information as possible into features of scatter nodes. From the set A_i of raw samples that contribute energy to a node i , other characteristics can be calculated that may be of use in further processing. Statistical moments measure the distribution of the raw backscatter sample values of the samples around scatter nodes. The (weighted) mean is given in eq. (3.21). Higher moments include standard deviation, skewness and kurtosis, which can all be incorporated as features into the feature vector \mathbf{v} of the scatter nodes.

Temporal information

In situations where diurnal or seasonal effects are expected to be important, one may consider deriving additional temporal information from the time stamp t of the scatter nodes, such as:

- time of day (day/night) for diurnal effects,
- time of year (month or season) for seasonal effects.

Echogram textures

In single beam fisheries acoustics, texture measures have been used to classify fish schools, and to differentiate between plankton and fish. For example Kieser *et al.* (2006) make use of a class of texture measures known as Gray Level Co-occurrence Matrices (GLCMs). These texture measures capture variability in texture between regions on the echogram, and may be indicative of certain species or classes of species or scatterers.

Such texture measures can be calculated directly off the echogram in two dimensions (a ping based approach). Care must be taken in using such measures, as they are really image processing tools and consider the echogram as an image. Consequently, the texture measure values do depend on the resolution of the system.

Hence they are not instrument independent, and in a way violate the fundamental idea of scatter nodes being a normalized data representation across instruments and across instrument settings. However, if one uses a single instrument and collects all data with the same instrument settings (beamforming options, pulse lengths and rates, etc.), texture measures can prove very useful.

Return echo pulse shape

Properties of the shape of the echo return pulse are commonly used in seabed classification and habitat mapping. The idea is that different types of seabed reflect the impacting acoustic pulse differently. Analysing the return pulse could therefore allow for identification of the seabed type, see for example Kloser *et al.* (2001) and Preston *et al.* (2001), and references therein. There is much debate about which features of the return pulse are indicative of the seabed type. Opinions range from one or two features, to many (more than hundred). In the latter case an analysis of the relative merit of each feature is typically conducted. Any features in which one may be interested, in the seabed characterization context, can be transferred directly to scatter nodes. In the context of this thesis no contributions to this ongoing discussion on seabed return echo features are made, and no specific features are described. However, it is important to note that such features can be attributed to scatter nodes. This will enable the use of the algorithms discussed in the next chapter to be applied to seabed classification problems, where relatively little attention is paid to classification algorithms as such.

Non-acoustic information

Information that is available but not directly derived from the backscatter measurements can be attributed to the scatter nodes so that it can be taken into account in the subsequent analysis. An example of such information is the distance to the seabed. Many multibeam sonars provide a bathymetry output, which is a measurement of depth for each sonar beam. Using the bathymetry output, the height above the seabed can be estimated for each node. This information can be useful in differentiating between seabed nodes and mid-water nodes, or between nodes indicative of pelagic and benthic fish species.

Features obtained from other data sets

Any information that is available about the conditions of the data and the environment it was collected in can be of value. Examples include the temperature of the water at the location, time and depth of the scatter node. Equivalently water salinity can be used as a feature, or chemical composition of the water. If water

currents are known to affect certain fish species it may be helpful to include measurements from instruments like acoustic Doppler current profilers (ADCPs) into the feature vector of the nodes. In the context of this thesis no use has been made of such additional data sources. It is expected that doing so would be valuable in certain circumstances when there is an obvious or expected correlation between the subject of the multibeam data analysis and auxiliary data one may have access to. Examples of relevant auxiliary data include depth profiles of temperature and acidity levels.

Other features

There is no limit to which features, and how many, can be attached to scatter nodes. It is anticipated that further research will lead to features being found or established that have great potential to support further processing.

3.4.3 Bathymetric soundings as scatter nodes

Scatter nodes are somewhat reminiscent of *soundings* as they are known in bathymetric surveying and mapping (de Moustier, 1988; 1993). Bathymetric soundings are derived from the raw multibeam data and are estimates of the seabed depth at the map locations of the soundings. There is at most one sounding per beam. Various processing algorithms exist to create bathymetric charts of the seabed (Calder and Mayer, 2003; Canepa *et al.*, 2003). The only property that is used is the location of each point. From the measured range from the transducer and the location of the surveying vessel, a geo-referenced coordinate is calculated for each sounding.

This method has been extended to not only map the bathymetry of the seabed but also the seabed type (Reut *et al.*, 1985; de Moustier and Matsumoto, 1993). In order to do so a number of other features of soundings are used, primarily their backscatter values (the point backscatter energy at the location of the sounding). Other studies investigate the use of the echo return pulse for classification purposes (Preston *et al.*, 2001).

This line of research does not consider midwater measurements. Everything is based on a good ping-based seabed detection, leading to soundings, from which some features are calculated. This approach can be brought into the full water-column picture by considering bathymetric soundings as scatter nodes. While definition 3.3 is introduced primarily to cover feature-rich scatterers obtained by applying a deconvolution to the raw full water-column sonar measurements, scatter

nodes can also be obtained by other means, such as standard multibeam seabed detection algorithms.

There is value in analysing the seabed habitat together with the mid-water ecosystem to provide a better understanding. The inclusion of soundings as scatter nodes allows for that. When some scatter nodes are obtained through the application of the inversion method and others through routine bathymetric processing, it is useful to include this knowledge into the nodes as a categorical feature (0 = deconvolution result, 1 = bathymetric sounding).

3.4.4 Scatter nodes from single beam sonar data

Deconvolution can be applied to single beam sonar data echograms, and scatter nodes can be derived. The deconvolution can be applied either in one dimension, on a per ping basis, or in two dimensions across pings. The latter will lead to better results as it takes correlations between pings into account.

An alternative way of deriving scatter nodes from single beam sonar data is to regard scatter nodes as lower resolution data samples. Down sampling single beam sonar data is achieved through the averaging of backscatter values of samples over a particular range, and over a particular number of pings or distance or time interval. The samples obtained by down sampling the original data can be considered scatter nodes, and features can be assigned to them similarly to how this is done with scatter nodes derived from multibeam sonar data. An example of this type of scatter node is presented as a case study in section 5.4.

In both cases, deconvolving or down sampling, the scatter nodes are a representation of a larger number of underlying original data samples. An important difference is that the nodes obtained through down sampling are regularly spaced, while the nodes obtained through the deconvolution are located where relevant echo signals were present. Hence, the spatial information carried by the nodes obtained through down sampling is less useful than it is in the other case. Nevertheless, defining nodes through down sampling of data does provide an efficient mechanism of transforming single beam sonar data into a scatter node representation, for use by subsequent pattern analysis algorithms.

3.5 OUTCOMES

The deconvolution method that is derived as a model inversion technique is used as a data preprocessing algorithm for multibeam sonar data. Additional features are extracted from the raw data, leading to scatter nodes: feature-rich spatio-temporal hydroacoustic data points.

Preprocessing is the first step in the scientific data mining process, and results in an alternative representation of the data (Figure 3.11).

The ultimate result of the model inversion technique is the transformation of any given multibeam data set into its corresponding set of scatter nodes. Scatter nodes are a fundamental concept that will be used throughout the remainder of this thesis. Their derivation is summarized in Algorithm 3.1.

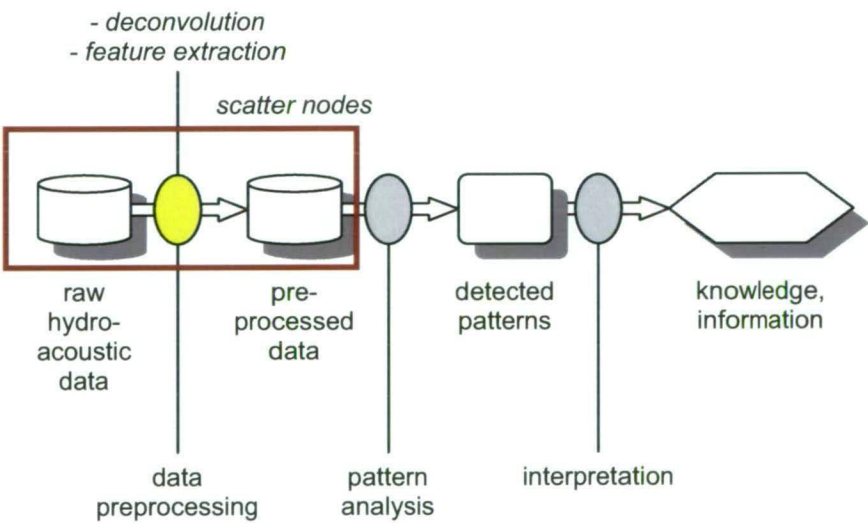


Figure 3.11 Data preprocessing is achieved through deconvolution and feature extraction, leading to preprocessed data in the form of scatter nodes.

Algorithm 3.1: *Derivation of scatter nodes from multibeam sonar data*

1. Obtain a data set containing the return echo from a single scatterer, preferably a calibration sphere. The settings on the instrument must be those that will be used during further data collection.
2. In the beamformed amplitude data containing the acoustic image from the calibration sphere, identify a region of samples around the sphere's location such that all samples that are significantly different from zero are within that region. The amplitudes of the samples in this region define the PSF.
3. Collect any kind of data with the instrument, using exactly the same settings as those used to find the PSF.
4. Apply the Lucy-Richardson deconvolution algorithm to that data set, using the PSF determined in step 2. If no calibration information is available, blind deconvolution can be applied instead.
5. Apply a suitable threshold (above the noise level), and define the scatter nodes as the local maxima in the deconvolved echogram.
6. Extend the scatter nodes with features obtained from the raw data set.

Two important aspects of the scatter node representation are now discussed: its compactness, and its usability compared to raw multibeam measurements.

3.5.1 Data compactness

The size of a raw multibeam data set Δ is affected by the particular details of the multibeam instrument, such as how many beams there are, what the sampling rate is, and how data are stored on disk. In addition, operational settings will affect the overall size: ping rate, pulse length, selected range, etc. Clearly these aspects will be reflected in the number of scatter nodes that will be obtained when applying the inversion method. Therefore it is most instructive to assess the ratio of the number of samples in the raw data to the number of scatter nodes. This ratio is affected by two components. Firstly, the particular details of the instruments and settings: given a single scatterer, how many samples end up being non-zero? This can be assessed from the PSF. In common multibeam systems such as those by Simrad, Kongsberg and Reson, signal transmit frequencies range from 200-450 kHz with pulse lengths of 0.05-0.20 ms and sampling rates in the order of 40-90 kHz. This results in echo return pulses of typically 8 samples long. The beam widths of the main lobes of the beam patterns typically result in widths of echo return pulses of approximately 5

beams, ignoring lower level side lobe effects. This alone would result in a non-zero PSF for 5 times 8 samples: a ratio of 1 scatter node to 40 samples. The second factor affecting this ratio is the underwater environment being observed. The whole sampling volume is represented through samples, but only the zones containing scatterers will result in scatter nodes, with no scatter nodes representing empty water. Experimentally evaluating the ratio of scatterers to samples in areas with fish present leads to typical values of 1/100. Note that this ratio includes the first effect mentioned. Typically a small number of scatter nodes are determined in each beam containing several hundreds of raw samples. Examples are given in chapter 5.

In summary, for a real-world data file collected during a survey, the amount of data will be reduced by a factor of hundreds when applying the inversion method. To what extent such a difference is reflected in file sizes depends on the data storage format of both the raw multibeam data and the scatter node data.

3.5.2 Usability

The other advantage of transforming multibeam data into scatter nodes is that scatter nodes are a more convenient data representation for visualization and further analysis by post processing algorithms. No off-the-shelf software packages other than Echoview (Myriax, 2008) are known to support multiple water-column multibeam sonar data formats, while many analytical and visualization software packages support data that are of the form of expression (3.19): essentially a set of multidimensional points.

Furthermore, the scatter node representation enables straightforward comparison of data sets collected using different instrument settings, or even by different instruments. In this sense, the inversion method acts as a means of normalizing the raw sonar data. Scatter nodes approximate the true scattering regime in an instrument-independent manner, since the deconvolution removes instrument-specific effects.

While data from various instruments obtained with various settings can all be transformed in a scatter node representation, those differences will carry through to the scatter node features. For example shorter pulse lengths will result in a higher range resolution, as will higher frequencies, and instruments with a higher dynamic range or regimes with higher signal to noise ratios will result in more accurate backscatter values of the original samples and hence of the scatter nodes. If such instrument related information is expected to be useful in future analyses, it can be attributed as additional features to the scatter nodes.

4 PATTERN ANALYSIS

4.1 OBJECTIVES

The next phase in the data mining process is concerned with the processing and analysis of scatter node data. Suitable pattern analysis algorithms must be applied to data sets containing scatter nodes in order to derive useful information. The aim is to identify groups of scatter nodes that belong together, and that are likely to be indicative of the same larger-scale object or concept. From a fisheries perspective, larger scale objects of interest that are observed by multibeam sonar are aggregations of fish, or fish schools. The seabed too is a larger scale object, used for bathymetry and habitat mapping.

Section 4.2 presents some possibilities for analysing and inspecting scatter nodes visually. New computational pattern analysis algorithms for scatter node data are developed in sections 4.3 and 4.4. Implications and properties of patterns detected using these novel methods are discussed in section 4.5. The outcomes of this chapter are summarized in section 4.6.

4.2 EXPLORATORY DATA ANALYSIS

4.2.1 Concept

Given a data set containing scatter nodes, a straightforward approach is to study a graphical representation of the scatter nodes, and determine visually whether any useful information can be observed. The field that is concerned with constructing such graphic representations of scientific data is known as *scientific visualization*

(Hansen and Johnson, 2004). The aim of scientific visualization is to aid in the understanding of the data. Creating and exploring data visualizations interactively is called *exploratory data analysis*.

In fisheries acoustics, it is customary to have visual representations of the echosounder or sonar recordings. Historically, the first analogue electronic echosounder recordings were plotted using an electronic plotter. Subsequent data analysis was based on such plots, by manual expert investigation. In modern echosounder and sonar systems, a visual display remains the most important primary tool for initial data analysis and assessment. Obvious anomalies in the system or its setup become apparent immediately. Scrutinizing a visual representation of the data is an essential part of quality control.

The visualization of multibeam water-column data is an important aspect of its usability (Mayer *et al.*, 2002; Wilson *et al.*, 2005). Typical displays of multibeam water-column data are either simply two-dimensional images or three-dimensional representations of original samples. Because of the compactness of the scatter node representation, there are much fewer scatter nodes than there are raw samples. This allows for a 3-dimensional graphical representation of much larger data sets, spanning longer time frames, and covering larger spatial areas. An appropriate visualization system should represent as many of the components of the scatter node components as possible, desirable or feasible.

4.2.2 Visualizing scatter nodes

Scatter nodes have different components, all of which can be represented or made accessible through appropriate visualization techniques. In this section some general aspects of visualization systems are discussed (Foley *et al.*, 1995; Hansen and Johnson, 2004).

Coordinate space

The fact that scatter nodes have a spatial component, usually coordinates in a georeferenced space (longitude, latitude, and depth below the water surface), suggests that the visualization be based around a georeferenced coordinate system in which the vector data of the scatter nodes will be displayed. Since this space is three-dimensional, a visualization on a computer display needs to provide an interface to manipulate the point of view and the zoom level interactively.

Apart from the spatial coordinates, the time coordinate is important. The combination of a three dimensional coordinate space with time as an additional

dimension is sometimes called four-dimensional visualization (Wilson *et al.*, 2005). Time can be introduced as a dimension through the concept of a time line for the three dimensional spatial coordinate system. A precise moment or limited time period for which the corresponding data are displayed can be selected. Progressing the visualization through time leads to a sequence, series, or animation.

Symbols or glyphs

Simply considering the spatial and temporal components of scatter nodes, they are points in a three or four dimensional space. The symbol with which they are represented graphically is sometimes referred to as a *glyph* in the literature (Schroeder *et al.*, 2006). Symbols in common use are points, spheres, crosses, etc.

Additional information can be encoded through symbols. Categorical features are most suitable for this purpose. For a categorical feature with possible values in a limited set $\{c_1, \dots, c_m\}$, a corresponding set of m symbols $\{s_1, \dots, s_m\}$ can be chosen so that points with feature value c_i are plotted with symbol s_i .

Colour

The colour with which the symbols are plotted can be used to encode additional information. An appropriate colour scale is used to map the numerical value of a feature to a particular colour.

Size

The size at which the symbols are plotted can fulfil the same role as colour. Encoding a feature through the size of the symbols works best for features that are of the same order of magnitude for all points, otherwise the small values will not stand out in the display. A solution in such cases is to use a logarithmic scale for that particular feature, and have the size of the symbols proportional to the logarithm of the feature rather than to the feature itself.

Transparency

Symbols can be drawn at varying levels of transparency, from 0% (opaque) to 100% (completely transparent and hence invisible). The transparency level can be used to encode feature values. Typical features that are encoded through transparency are related to accuracy, reliability or uncertainty: points with low

accuracy or reliability, or high uncertainty, are plotted in a more transparent way than accurate or reliable points.

Another use of transparency is in visualizing time. When plotting points that occur at a certain point in time, points of the recent past can be added to the same visual display, with a transparency level increasing with increasing time gap between that data and the present moment. The result is a fading away of data points as new data is added through progressing time.

4.2.3 Echoview

An advanced software tool for acoustic data is the Echoview package, which has extensive visualization capabilities (Wilson *et al.*, 2005; SonarData, 2007; Myriax, 2008). This is the only known off-the-shelf software currently available that can handle multiple water-column multibeam sonar data formats, from a range of instruments and manufacturers. This is very useful for visualizing raw multibeam data sets in conjunction with the scatter nodes derived from it. At the time of writing, the support for spatio-temporal vector data in Echoview is limited. After they are derived externally, scatter nodes can be imported into Echoview as objects. As a result, it is not possible to interactively change the scatter node features encoded by size or colour after they have been imported into Echoview. Nevertheless, the combined display of scatter nodes with the underlying raw data is invaluable.

The typical visualization of raw multibeam sonar data is a two-dimensional *echogram* (Figure 4.1): it represents the data constituting one *ping* (one transmission/reception cycle).

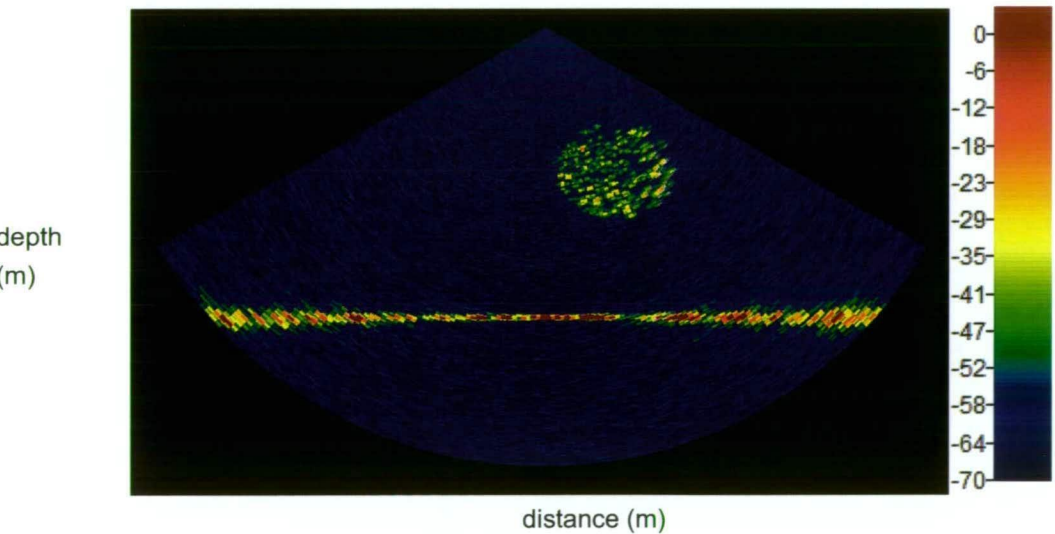


Figure 4.1 Standard two-dimensional echogram of raw multibeam data (S_v in dB).

Echoview includes a four-dimensional data viewer (Wilson *et al.*, 2005). A set of consecutive two-dimensional echograms can be shown as a sequence of images in a three-dimensional space with time as a fourth dimension. Time is controlled through a time slider (Figure 4.2). It is necessary to impose a data threshold, otherwise the many samples in empty water block the samples of interest from view.

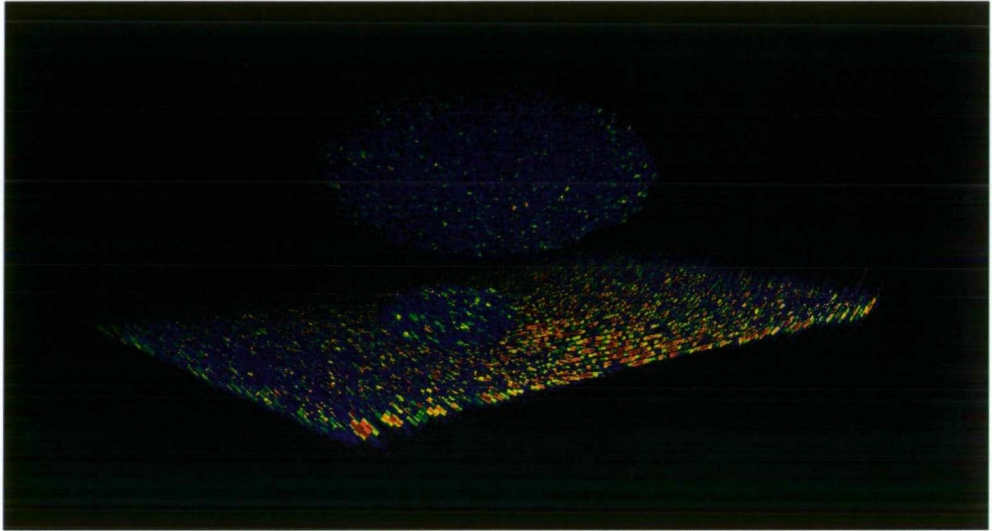


Figure 4.2 Three-dimensional view with additional time dimension. Past data is fading away.

Scatter nodes, as introduced in this research, can be visualized together with the raw data. They can be shown in the two-dimensional or in the three-dimensional view (Figure 4.3 and Figure 4.4 respectively).

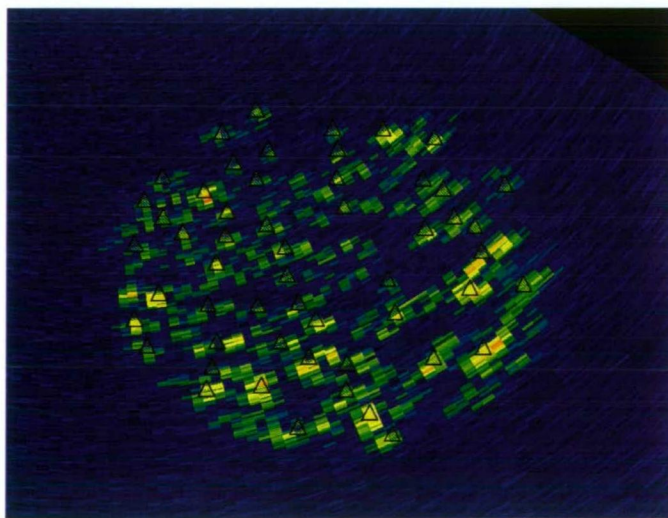


Figure 4.3 Close-up of part of the echogram in Figure 4.1, with derived scatter nodes plotted as triangles.

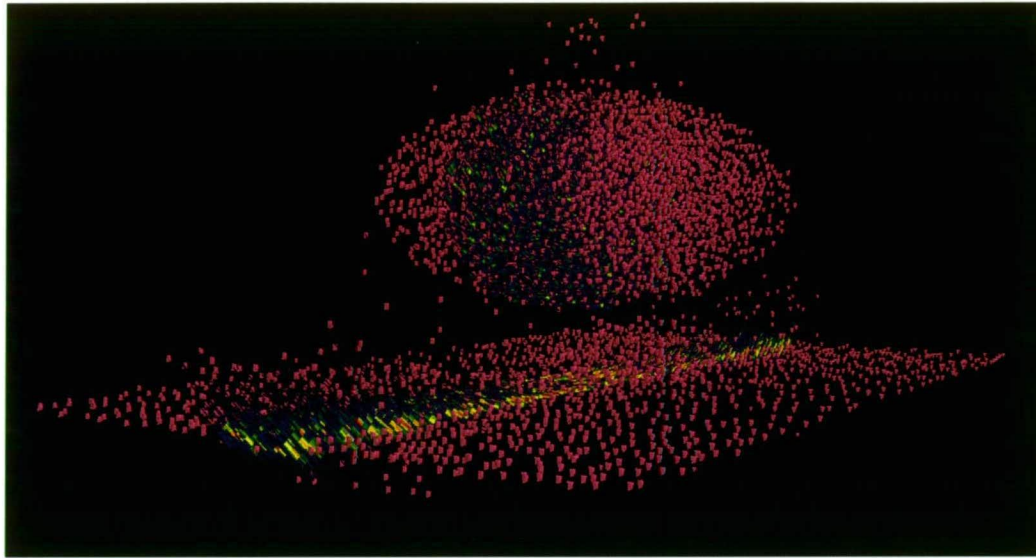


Figure 4.4 Three-dimensional view, showing derived scatter nodes in red, together with part of the underlying raw data.

These visualizations place the scatter nodes at their correct position in space. In two dimensions they are represented by triangles, in three dimensions by tetrahedra. A tetrahedron is a volumetric object formed by four points, interconnected by lines of equal length. In Echoview it is not convenient to vary the size or colour of these symbols by a particular property or feature of the nodes.

The combined visualization of the raw multibeam data and the scatter nodes derived from it gives immediate visual feedback about the data preprocessing algorithm. The visualizations can be scrutinized to investigate whether enough scatter nodes are constructed, whether no parts of the data are lost, and whether noise is not distorting the results. In general, one wants a good coverage of the raw above-threshold backscatter samples by scatter nodes. An inspection of the three-dimensional visualization may reveal how many aggregations of fish are present, and what their spatial extent is. Such information can be of use in tuning subsequent pattern analysis algorithms.

The data set that was used to create the images in this section is discussed in detail in section 5.1.

4.2.4 Eonfusion

A beta version of the software package Eonfusion (by Myriax Pty Ltd) is used to visualize scatter nodes. Eonfusion is a data analysis package for environmental data and includes a four-dimensional visualization engine. The user of the software can interact with the data and the graphical environment, making the visualization dynamic and interactive. At the time of writing the software supports generic spatio-temporal vector data such as scatter nodes, but it has no support for raw multibeam data. It is expected that support for such data types will become available in the future.

Figure 4.5 shows an example, created with Eonfusion, in which the size and colour with which scatter nodes are plotted is varied with their backscatter energy levels. The same nodes as those of Figure 4.4 are shown, but now large red points are scatter nodes with high backscatter energy sample values, while smaller yellow nodes have lower energy levels.

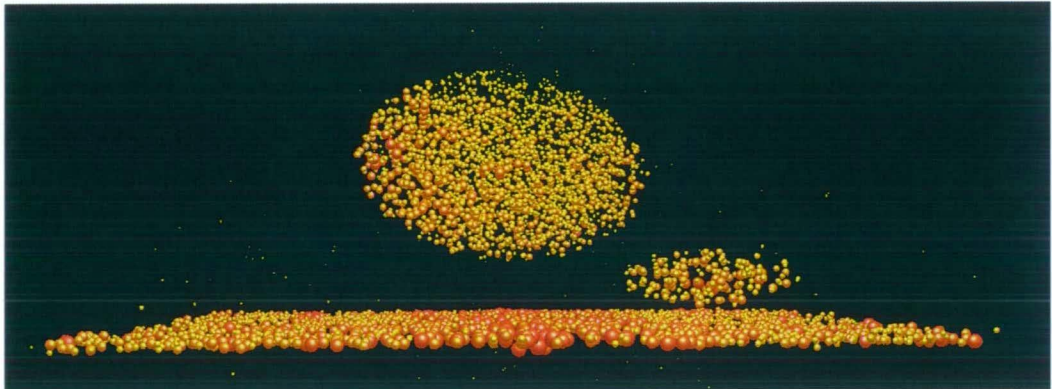


Figure 4.5 The size and colour of scatter nodes is varied with their backscatter energy levels. Large red points indicate high levels; small yellow ones indicate low levels.

4.2.5 Other packages

Scatter nodes are essentially points in space and time with an additional set of features. Scatter nodes can be seen as point data, as vector data, or as records in a spatial database. The terminology used to describe this kind of data varies between software packages. Relevant software packages supporting this kind of data include

- analytical packages such as Matlab (® The Mathworks Inc.), S-Plus (® Insightful Corp.), and R (open source),

- spatial visualization software such as Fledermaus (® IVS 3D), and
- spatial information analysis software such as Voxler (® Golden Software).

This is not an exhaustive list of all software available that supports scatter node data, but it demonstrates that the scatter node representation of multibeam sonar data is sufficiently general and generic to be supported by many existing packages. Furthermore, most of the packages mentioned here include data export routines to export data into other file formats for use by different software tools.

4.3 SPATIAL CLUSTERING

4.3.1 Concept

The aim of clustering algorithms is finding groups or clusters of similar scatter nodes automatically. Scatter nodes as defined in definition 3.3 are similar if all or some of their components are similar. Considering the spatial components \mathbf{x} , scatter nodes are expected to be indicative of the same object if they are close to each other, in a geometrical sense. Similarity in this case is measured as spatial proximity. In terms of the temporal component t , scatter nodes are similar if they were collected not long apart in time. Similarity with respect to the feature vectors \mathbf{v} is less straightforward, as similarity is measured in a feature space, in which a variety of metrics could be applied.

It is clear that all three components, space, time and other features must be considered simultaneously in order to assess similarity of scatter nodes. This is not trivial, and will be discussed in section 4.4, where appropriate methods are developed.

Prior to that, it is worth investigating whether there are standard text-book clustering algorithms that can be used to cluster scatter node data. The benefits of such investigation are twofold. Firstly, the existence of applicable algorithms delivers evidence of the value of the scatter node representation. If standard algorithms can be applied, then the preprocessing routine enables the use of such algorithms for multibeam sonar data. Secondly, a working text-book algorithm provides a base line approach to clustering scatter nodes. It can be used in benchmarking tests with new algorithms. It is customary to compare results of new algorithms with results of established ones, to assess their usefulness and value.

Algorithms to jointly consider spatial, temporal and feature components of data points are not found in the reference literature on clustering and pattern analysis (Duda *et al.*, 2000; Hastie *et al.*, 2001; Bishop, 2006). Therefore, the scatter nodes are reduced to their spatial components in the remainder of this section. From studying the scatter node data sets visually (section 4.2), it is expected that the spatial components alone can be sufficient to yield reasonably good clustering results for some purposes, for example for the detection of fish schools.

Ignoring the temporal and feature components of scatter nodes, scatter nodes are represented as column vectors x_i for $i = 1, \dots, n$, and n the number of scatter nodes in the data set Θ .

Assuming k clusters are to be found, a clustering algorithm is a function f which maps each data point to an integer value $1, \dots, k$ indicating the cluster number:

$$f: x_i \in \Theta \mapsto \{1, \dots, k\}. \quad (4.1)$$

Different clustering methods allow different function classes for f , and differ in the way f is established.

In the next section an overview of clustering methods is given, with an assessment of whether they can be of value in clustering the spatial components of the scatter nodes.

4.3.2 Overview of clustering methods

Several standard clustering methods are considered. A number of general texts discuss various methods (Duda *et al.*, 2000; Hastie *et al.*, 2001; Bishop, 2006). Good review papers on clustering include Wei *et al.* (2003) and Xu and Wunsch (2005). References to specific publications are given below, where appropriate. The list of algorithms presented below is not an exhaustive listing of all methods. Rather, it is a summary of the most common algorithms or families of algorithms that are considered within the scope of this thesis. The methods discussed are simple, ad-hoc algorithms that have proven valuable in at least some application domains.

Each of the standard texts referenced above presents clustering algorithms in a different taxonomy. Given the multitude of algorithms and the diverse ways they operate, no taxonomy is perfect. Roughly following Xu and Wunsch (2005), the following families of clustering algorithms are distinguished.

- **Criterion or optimization methods** find an optimal cluster assignment for the data points under consideration, based on some well defined decision or optimization criterion.

Examples: k-means, Gaussian mixtures, CLARANS.

- **Hierarchical or agglomerative methods** group data points together in clusters, starting from having n clusters each containing a single data point. These are *bottom-up* methods.

Examples: DBSCAN, CURE, Chameleon.

- **Partitioning or divisive algorithms** partition the data into clusters, starting from a single cluster containing all the data points. These are *top-down* methods.

Examples: BIRCH

- **Advanced clustering methods** are more involved and less ad-hoc, and do not typically fit one of the first three groups mentioned. Algorithms in this family have roots in fuzzy logic, neural networks, kernel methods and other areas related to machine learning.

The example algorithms listed for each of the first three families are considered here. Advanced clustering algorithms are discussed further in section 4.4.

K-means

The most widely used clustering algorithm is the k-means algorithm, originally proposed in Lloyd (1982) and discussed in many text books on pattern analysis (Duda *et al.*, 2000; Hastie *et al.*, 2001; Bishop, 2006). It is a simple and popular method, often used as a benchmark for comparison with other algorithms. Because of its significance, and the fact that extensions to this algorithm are considered in section 4.4.6, a detailed description is given.

Let $\{x_1, \dots, x_n\}$ be the data set to be clustered. In the current context the x_i are the spatial components of the scatter nodes, but the algorithm applies to any kind of feature vectors for which a (Euclidian) metric is defined.

The goal is to partition this data set into k clusters. The k clusters are assumed to be represented by their cluster centres, μ_1, \dots, μ_k . The k-means algorithm is an iterative procedure to find these cluster centres. The algorithm can be derived and

defined in different ways; here, the general line of thought of Bishop (2006) is followed.

For each \mathbf{x}_i in the data set, binary indicator variables $r_{ij} \in \{0, 1\}$ are defined, describing which of the k clusters the data point i belongs to: r_{ij} is 1 if data point \mathbf{x}_i belongs to cluster j , and 0 otherwise. An objective function, sometimes called a distortion measure or loss function is defined as:

$$J = \sum_{i=1}^n \sum_{j=1}^k r_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \quad (4.2)$$

with $\|\cdot\|$ the Euclidian distance metric. The objective function J represents the sum of squares of the distances between each data point and the centre of the cluster it belongs to. The goal is to find values for all r_{ij} and $\boldsymbol{\mu}_j$ so as to minimize J . This is achieved through an iterative procedure consisting of two steps or phases, corresponding to successive optimizations with respect to r_{ij} and $\boldsymbol{\mu}_j$. The algorithm is initialized by selecting initial values for the $\boldsymbol{\mu}_j$. This is commonly done by selecting k data points at random.

Then, in the first phase J is minimized with respect to r_{ij} , keeping the $\boldsymbol{\mu}_j$ fixed. The terms involving different i in eq. (4.2) are independent, so J can be minimized for each i separately by choosing r_{ij} to be equal to 1 for whichever value of j gives the minimum value of $\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$. This is assigning the data points \mathbf{x}_i to their nearest cluster centre:

$$\begin{aligned} r_{ij} &= 1 && \text{if } j = \arg \min_m \|\mathbf{x}_i - \boldsymbol{\mu}_m\|^2, \\ r_{ij} &= 0 && \text{otherwise.} \end{aligned} \quad (4.3)$$

In the second phase, J is minimized with respect to the $\boldsymbol{\mu}_j$ keeping the r_{ij} fixed. The function J is quadratic in $\boldsymbol{\mu}_j$ and can be minimized by setting its derivative with respect to $\boldsymbol{\mu}_j$ equal to zero:

$$2 \sum_{i=1}^n r_{ij} (\mathbf{x}_i - \boldsymbol{\mu}_j) = 0 \quad (4.4)$$

which can be solved for $\boldsymbol{\mu}_j$, giving:

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^n r_{ij} \mathbf{x}_i}{\sum_{i=1}^n r_{ij}}. \quad (4.5)$$

The denominator is equal to the number of data points in cluster j . This expression for μ_j has a simple interpretation: it is the mean of all the data points assigned to cluster j .

These two phases of re-assigning the data points to clusters and re-calculating the cluster centres are repeated until there is no further change in the assignments. Since each phase reduces the value of J , convergence is assured. However, the algorithm may converge to a local rather than a global minimum of J .

As a consequence of utilizing the criterion of eq. (4.2), cluster shapes are hyper-spherical. This is a known limitation, and is one preventing the application of the k-means algorithm to scatter node data sets. Fish schools can take on various non-spherical shapes, and obviously the seabed is non-spherical. To illustrate this, a simple data set is created using the model developed in section 3.2, and scatter nodes are derived (Figure 4.6).

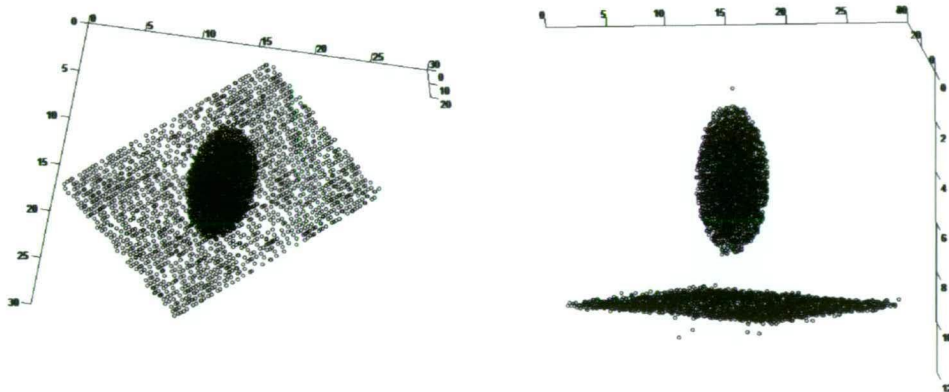


Figure 4.6 Two views of the same set of scatter nodes that is to be clustered. An ellipsoidal aggregation of fish and a flat seabed are distinguished.

Clearly, from a visual inspection of this data set, two clusters are present: an ellipsoidal fish school and the seabed. K-means is run with $k = 2$. The result is presented in Figure 4.7.

K-means is unable to correctly identify the two clusters, which are obviously distinguished by visual inspection. The reason is that k-means cannot detect cluster shapes that are not spherical. In this case the flat rectangular cluster representing the seabed is causing k-means to fail.

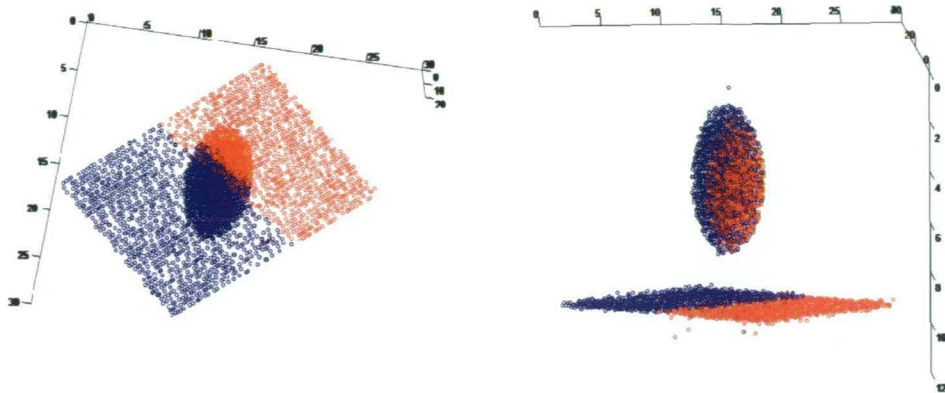


Figure 4.7 The data set from Figure 4.6, clustered into 2 clusters using k-means. Red and blue indicate cluster membership. K-means fails to identify the obvious clusters.

Gaussian mixtures

The underlying assumption of Gaussian mixture models is that the probability density from which the data set $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ was obtained is a linear superposition of k Gaussian distributions:

$$P(\mathbf{x}) = \sum_{j=1}^k \pi_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad (4.6)$$

with mixing coefficients π_j , $N(.,.)$ the normal distribution with mean $\boldsymbol{\mu}_j$ and variance/covariance vector $\boldsymbol{\Sigma}_j$.

The goal is to find values for π_j , $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ using the maximum likelihood criterion, which aims at selecting those values of the free parameters that maximize the joint likelihoods of the data, given the parameters:

$$\sum_i P(\mathbf{x}_i | \pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \quad (4.7)$$

The procedure through which this is commonly achieved is *Expectation Maximization*. It is an iterative procedure consisting of two steps, an expectation step and a maximization step. Details are found in the text books referenced above. After convergence, a point \mathbf{x}_i is assigned to that cluster m for which the probability is maximal:

$$m = \arg \max_j P(\mathbf{x}_i | \pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \quad (4.8)$$

In fact, this algorithm is a generalization of the k-means algorithm. The difference is that with Gaussian mixtures, cluster shapes can be hyper-ellipsoidal rather than hyper-spherical. From the example given in Figure 4.7 it is clear that this is still too limiting in the case of clustering scatter nodes, as clusters can only be convex ellipsoids, and the rectangular cluster representing the seabed cannot be expected to be identified correctly. Furthermore, while the example school in the data set is ellipsoidal, real fish schools can have varying shapes, depending on the fish species and their schooling behaviour (Lawson *et al.*, 2001; Gerlotto and Paramo, 2003; Gerlotto *et al.*, 2004).

CLARANS

The algorithm *Clustering Large Applications based on RANdomized Search* (CLARANS) was proposed by Ng and Han (2002), and is a generalization of CLARA, a random sampling approach (Xu and Wunsch, 2005). The CLARANS algorithm defines optimization criteria, such as eq. (4.2) for k-means or eq. (4.7) for Gaussian mixtures. Rather than implementing an iterative process, it proceeds by randomly selecting subsets of data points that are considered for cluster re-assignment. The advantage over k-means and Gaussian mixtures is that it is less sensitive to initialisation, and hence less likely to converge to a local extremum of the criterion function. A disadvantage is its computational complexity and running time.

As the resulting cluster shapes are not different from those found by k-means or Gaussian mixtures, CLARANS is not considered further in the current context.

DBSCAN

Density Based Spatial Clustering of Applications with Noise (DBSCAN) was first proposed in Ester *et al.* (1996) and extended in Sander *et al.* (1998). The intuitive idea is to assign dense aggregations of points to the same cluster. Zones of low density separate different clusters. Clusters are grown from an initial data point and new clusters are created as needed, based on a density criterion. DBSCAN can handle noise or outliers: points that are not in sufficiently high-density regions are regarded as noise.

This approach meets the needs for clustering scatter nodes very well. DBSCAN is specifically aimed at spatial clustering, it deals with noise very well, and it matches the intuitive concept of clusters in scatter node data. DBSCAN is described and studied in more detail in section 4.3.3. As a preliminary test, DBSCAN is run on the data set from Figure 4.6. The resulting DBSCAN-based clusters are shown in

Figure 4.8. DBSCAN is able to correctly identify the two clusters that are expected from visual inspection.

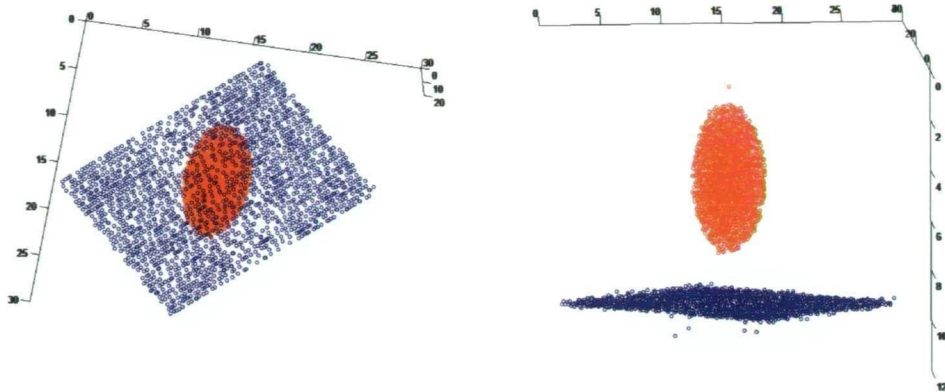


Figure 4.8 The data set from Figure 4.6 clustered using DBSCAN. The two clusters that are expected intuitively are determined correctly.

CURE

Clustering Using REpresentatives (CURE) was proposed in Guha *et al.* (2001). Where k-means, Gaussian mixtures and CLARANS are all based on a mean cluster centre, CURE uses multiple points to represent a cluster. This introduces some flexibility in terms of cluster shapes, and does not limit them to be hyper-spherical or ellipsoidal in shape.

CURE is reported to have difficulty in finding clusters of hugely different sizes (Xu and Wunsch, 2005). Since in the application at hand, the clustering of scatter nodes, it is expected that clusters will vary greatly in size, CURE is not considered further.

Chameleon

Chameleon is a clustering algorithm based on graph theory (Karypis *et al.*, 1999). It uses a k-nearest-neighbour graph, where each data point is a node. Nodes are connected to their k nearest neighbours for some value of k and some distance metric. The algorithm then proceeds by finding connected clusters in this graph. Conceptually this is not very different from the DBSCAN approach. However, an important difference is the fact that Chameleon connects points that are near to each other using a ranking system, while DBSCAN uses absolute distances. Chameleon could give rise to clusters consisting of points that are sparsely distributed relative to points in other clusters. This is not desirable in the case of scatter nodes, where the spatial features of scatter nodes are considered explicitly, and absolute proximity in geometric terms is relevant.

BIRCH

In Zhang *et al.* (1996), an algorithm called *Balanced Iterative Reducing and Clustering using Hierarchies* (BIRCH) is proposed. BIRCH is a tree-based method, where each node can be seen as a cluster summary at that level. Clustering proceeds by splitting nodes according to some criterion, so introducing smaller clusters. The advantage of BIRCH is that it is relatively efficient and can handle very large data sets. However, it has problems also: it is unclear how to establish optimal splitting criteria, and, again, it uses cluster centres as representatives, leading to hyper-spherical clusters. The latter prevents the use of BIRCH in the clustering of scatter nodes.

In summary

The requirements that are important for a clustering algorithm for scatter nodes are given in Table 4.1, with an assessment of each requirement against the considered algorithms.

	<i>Arbitrary cluster shape</i>	<i>Different cluster sizes</i>	<i>Handles noise or outliers</i>	<i>Uses spatial aspect of data</i>	<i>Acceptable complexity</i>
K-means	-	+	-	+	+
Gauss. mixt.	-	+	-	+	+
CLARANS	-	+	-	+	-
DBSCAN	+	+	+	+	+
CURE	+	-	-	+	+
Chameleon	+	+	+	-	+
BIRCH	-	-	+	+	-

Table 4.1 Overview of clustering algorithms and their characteristics
(+ : requirement met, - : requirement not met).

In conclusion it is decided to use DBSCAN as the base line algorithm for clustering scatter nodes as it scores well against all requirements. DBSCAN allows for arbitrary cluster shapes of potentially hugely different sizes, and can handle noise in the data. The concept of clusters being regions of high density fits well with the intuitive idea of clusters in scatter node data. Furthermore, the algorithm is designed for spatial data specifically, and does take into account the fact that the features are coordinates in a geometric space.

4.3.3 Spatial clustering with DBSCAN

DBSCAN basics

The original version of DBSCAN (Ester *et al.*, 1996) has been generalised (GDBSCAN), extending the notion of neighbourhoods and metrics (Sander *et al.*, 1998). An optimization procedure that allows for sequential applications of DBSCAN with different parameters was proposed in Ankerst *et al.* (1999) under the name OPTICS (*Ordering Points To Identify Clustering Structure*). In the context of this thesis, only the spatial components of scatter nodes are considered. Furthermore, spatial densities of nodes in clusters are not found to vary greatly and hence no extensive iterative searching is done. For these reasons no further attention is paid to GDBSCAN and OPTICS, and the discussion concentrates on the fundamental DBSCAN algorithm as it was originally proposed (Ester *et al.*, 1996).

In investigating DBSCAN for scatter nodes, a serious shortcoming of the algorithm was identified. The clustering resulting from DBSCAN is not unique. The outcome depends on the order in which the data points are presented to the algorithm. A permutation of the input data set can result in a different clustering being obtained. This weakness is recognized to some extent by the original authors (Ester *et al.*, 1996), but no solution is provided. Furthermore, this aspect of the algorithm is often ignored in the literature when considering DBSCAN (Xu and Wunsch, 2005). The DBSCAN-based extensions GDBSCAN and OPTICS make a brief mention of this problem but do not investigate its consequences (Sander *et al.*, 1998; Ankerst *et al.*, 1999).

In the following paragraphs it is shown that arbitrary clustering can occur for points near cluster boundaries. Since these points are crucial in determining subsequent decision boundaries in classification problems, the DBSCAN indeterminism is unacceptable. Because the randomness is most manifest in boundary regions between clusters, DBSCAN is known to perform poorly in such cases (Yip *et al.*, 2006). An improved version of DBSCAN, resulting in a truly unique clustering, is proposed in this thesis; it will be referred to as Unique-DBSCAN or UDBSCAN, as it results in a unique solution to the clustering problem.

The DBSCAN algorithm is now presented, followed by the extensions and modifications leading to UDBSCAN. Since the context of this section is self-contained, definitions and lemmas are numbered with the prefix ‘D-’ (from DBSCAN). This prevents confusion with the general discussion of the thesis.

DBSCAN according to Ester *et al.* (1996)

The presentation of the DBSCAN algorithm given here follows Ester *et al.* (1996) roughly, although the mathematical notation utilized here is more formalized.

Let S be some N -dimensional space with a distance metric $\|.,.\|$. Let X be a finite subset of S containing the points to be clustered, $X = \{x_1, \dots, x_n\}$.

Definition D-1. The ε -neighbourhood of a point x in X is

$$N_\varepsilon(x) = \{y \in X \mid \|x, y\| \leq \varepsilon\}.$$

The fundamental idea of DBSCAN is to create clusters of points in X such that locally dense distributions of points give rise to clusters. Intuitively, points in clusters are expected to have a minimum number of other points in their local neighbourhood. Points on the boundaries of clusters, towards low-density areas, will generally have fewer points in their neighbourhoods while still belonging to the cluster. This idea is formalized as follows. Let m be the minimum number of points expected in a ε -neighbourhood of a point inside a cluster.

Definition D-2. A point x in X is a core point with respect to ε and m , if and only if $|N_\varepsilon(x)| \geq m$.

$|N_\varepsilon(x)|$ denotes the cardinality of the set $N_\varepsilon(x)$.

Definition D-3. A point x is directly density-reachable from a point y with respect to ε and m , if

- 1) $x \in N_\varepsilon(y)$ and
- 2) y is a core point.

Note that if x is directly density-reachable from y , the reverse is not necessarily true. It is only true if x is also a core point.

Lemma D-1. If x is a core point and x is directly density-reachable from y , then y is directly density-reachable from x .

Proof

Since x is directly density-reachable from y , $x \in N_\varepsilon(y)$, and $\|x, y\| \leq \varepsilon$. The distance metric $\|.,.\|$ is symmetric by definition, hence $y \in N_\varepsilon(x)$. Since x is also a core point, y is directly density reachable from x .

The concept of density-reachable is now defined.

Definition D-4. A point x is density-reachable from a point y with respect to ε and m , if and only if there is a chain of points x_1, x_2, \dots, x_k , with $x_1 = y$ and $x_k = x$, such that x_{i+1} is directly density reachable from x_i , $\forall i \in \{1, 2, \dots, k-1\}$.

This relation is again not symmetrical in general, and lemma D-1 can be generalized.

Lemma D-2. If x is a core point and x is density-reachable from y , then y is density-reachable from x .

Proof

By induction from the proof of lemma D-1.

The notion of border points can now be formalized. This is not done explicitly in the original publications (Ester *et al.*, 1996; Sander *et al.*, 1998).

Definition D-5. A point x in X is a border point, if and only if x is density reachable from a point y , and x is not a core point.

Two border points may not be density reachable from each other, while they are both density reachable from a common core point. Such a relationship is now defined.

Definition D-6. A point x_1 is density-connected to a point x_2 with respect to ε and m , if and only if there is a point y such that both x_1 and x_2 are density reachable from y .

Lemma D-3. The relation ‘density-connected’ is symmetrical.

Proof

Follows immediately from definition 6.

The density-based notion of a cluster can now be defined.

Definition D-7. A cluster C of points in a set X , with respect to ε and m , is a non-empty subset of X satisfying the following conditions:

(maximality) If $x \in C$ and y is density reachable from x with respect to ε and m , then $y \in C$,

(connectivity) $\forall x, y \in C$, x is density-connected to y with respect to ε and m .

A clustering of X is simply the set of all possible clusters.

Definition D-8. The clustering, with respect to ε and m , of D , is the set $\Phi_{\varepsilon, m} = \{ C \mid C \text{ is a cluster of } X \text{ with respect to } \varepsilon \text{ and } m \}$, and if C is a cluster of X with respect to ε and m , then $C \in \Phi_{\varepsilon, m}$.

There may be points that do not belong to any cluster. Such points are said to be noise.

Definition D-9. Let $\Phi_{\varepsilon, m} = \{C_1, \dots, C_k\}$ be the clustering of the set X with respect to ε and m . The noise of the set X with respect to ε and m is the set $N_{\varepsilon, m} = \{x \in X \mid \forall i, x \notin C_i\}$.

In order to find the clusters of a set X , an algorithmic process must be outlined. The following two lemmas are helpful in doing so.

Lemma D-4. Let $x \in X$ be a core point. Then the set $C = \{y \in X \mid y \text{ is density-reachable from } x\}$ is a cluster as per definition D-7.

Proof

Since x is a core point, x is density-reachable from itself, or $x \in C$, hence C is not empty. Maximality: Let $y \in C$ and r density-reachable from y . y is density-reachable from x by definition of C . Therefore r is density-reachable from x , and $r \in C$. Connectivity: Let $y, r \in C$. Both are density-reachable from x by definition of C . Then y is density-connected to r by definition D-6.

Lemma D-5. Let C be a cluster of X and let x be a core point, $x \in C$. Then C equals the set $A = \{y \in X \mid y \text{ is density-reachable from } x\}$.

Proof

The set A is a cluster because of Lemma D-4. Assume that there is a point $y \in C$, $y \notin A$. Since $y \in C$, y is density-connected with x . Since x is a core point, y is density-reachable from x , and hence $y \in A$. Therefore $C \subseteq A$. Similarly, assuming a point $y \notin C$, $y \in A$ leads to $A \subseteq C$. In conclusion $A = C$.

This concludes the formal basis for the DBSCAN algorithm as in Ester *et al.* (1996), where it is not stated that it is possible for clusters to overlap. The authors of Sander *et al.* (1998) do recognize the fact that cluster overlap is possible.

Lemma D-6. Let $C_1 \in \Phi_{\epsilon, m}$ and $C_2 \in \Phi_{\epsilon, m}$ be two clusters of points in X , and $C_1 \neq C_2$. Then for all $x \in C_1 \cap C_2$, $|N_\epsilon(x)| < m$ (i.e. x is not a core point).

Proof

Assume x a core point. Since x is in both C_1 and C_2 , lemma D-5 states that $C_1 = C_2$, which is a contradiction.

While stating this lemma, Sander *et al.* (1998) do not attribute any further significance to it. This lemma is in fact the basis for the indeterminism arising in DBSCAN. For completeness the following lemma is formulated.

Lemma D-7. A core point p can only belong to a single cluster C_i with respect to ϵ and m .

Proof

Follows immediately from lemma D-6.

It is important to note that, based on lemmas D-6 and D-7, it is clear that when a clustering $\Phi_{\epsilon, m}$ of points in X is achieved, not every point in X will belong to a single cluster.

Indeterminism in DBSCAN

Lemmas D-4 and D-5 allow for an algorithmic formulation of a clustering procedure. In essence, the lemmas say that one can use any core point as a seed, and grow a cluster by adding every point that is density-connected to a point already in the cluster. When the cluster-growing process is complete for the first cluster, the process is repeated, each time starting with a core point that is not in a cluster yet. The clustering finishes when all core points are in a cluster. Pseudo-code for this algorithm is given in Ester *et al.* (1996).

Lemma D-6 clearly indicates that, in principle, there may be points that belong to more than one cluster. DBSCAN simply proceeds by assigning such points to clusters on a first-come-first-serve basis: once a point is included in a cluster it cannot be included in a second cluster. As such, the actual clusters resulting from the algorithm are disjoint sets, while the clusters according to the underpinning formalism outlined above are not necessarily disjoint. This is the essence of the indeterminism in DBSCAN. Altering the order in which data points are presented to the algorithm can affect the cluster to which they will be assigned. This is clearly undesirable. The DBSCAN clustering algorithm will assign some points arbitrarily to one cluster or another. Lemma D-6 states that this is the case only for points that are not core points.

The following example illustrates the importance of such non-core points. Once clusters are identified, they can be used for subsequent classification of new points. The classification boundary will be chosen to be the cluster boundaries as resulting from the clustering. The key achievement of such clustering scenarios is not only the clustering of the given points, but rather the determination of decision boundaries between classes. The data to be clustered are represented in Figure 4.9.

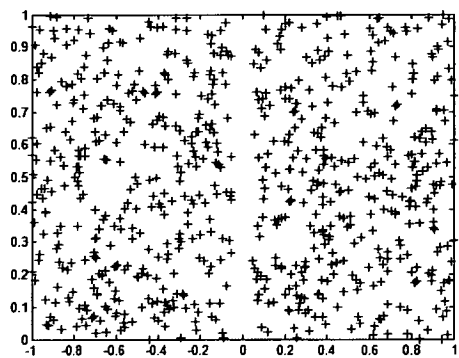


Figure 4.9 An example data set to be clustered.

This is a simple example just to illustrate the indeterminism in DBSCAN. The points were clustered twice, with identical parameters, but the order of points was changed between runs. The results are presented in Figure 4.10 (a)-(d). Cluster

assignments are represented by the symbols \circ and \blacklozenge . Observe that in Figure 4.10 (a)-(b) there is one point belonging to the \circ -cluster that was captured by growing the \blacklozenge -cluster (marked by the red ellipse). Starting with growing the \circ -cluster, as in Figure 4.10 (c), leads to this point being captured by the \circ -cluster. However, now another point is incorrectly captured by the \circ -cluster (marked by the red ellipse). Given that the only difference between the two runs is the order in which points were processed, this is an unacceptable indeterminism in the algorithm, and clearly leads to arbitrary distortions in the decision boundaries between classes (Figure 4.10 (b) and (d)).

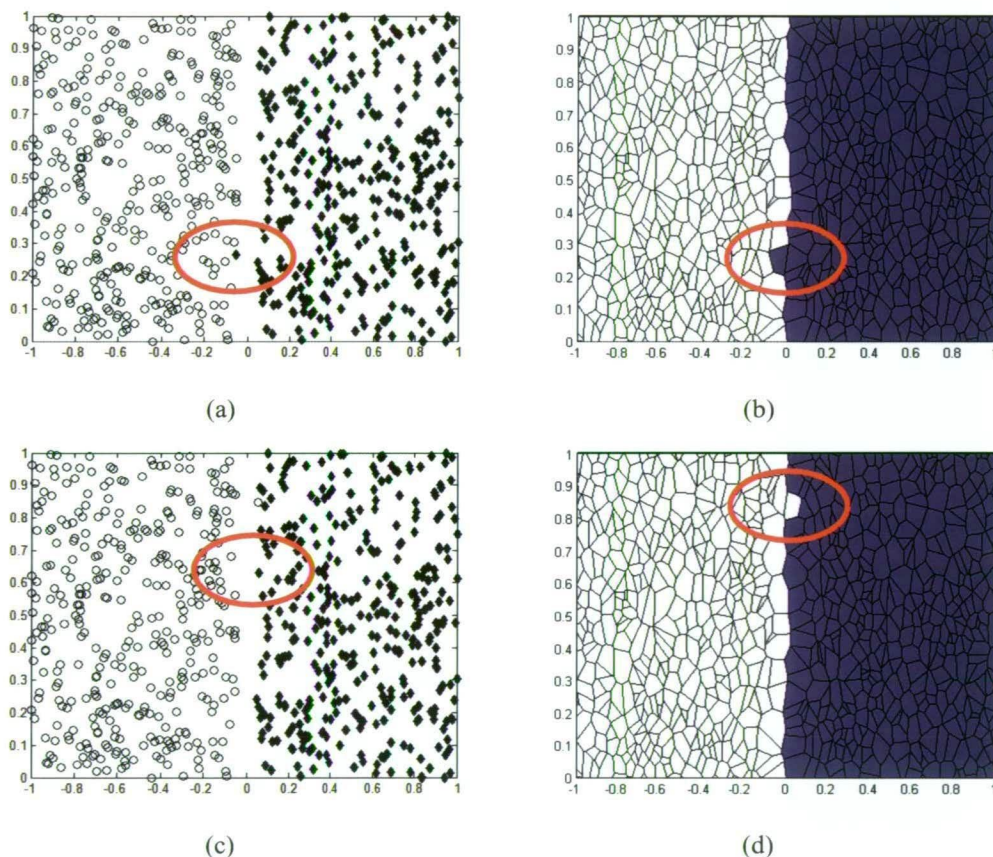


Figure 4.10 (a) and (b) are the clusters and classification zones resulting from running DBSCAN a first time. (c) and (d) are equivalent, obtained by running DBSCAN a second time, with the only difference that the list of input data points was permuted after the first run of DBSCAN. It can be seen that the two runs of DBSCAN do not result in equal clusterings. Different arbitrary distortions can be seen (marked by red ellipses).

It is tempting to ignore the fact that some points are arbitrarily assigned to one cluster or another (Ester *et al.*, 1996; Sander *et al.*, 1998) by contending that it will generally be only a small number of points that suffer from this inconsistency, and that the majority of points will be clustered identically, independent of the order in

which the points are presented to the algorithm. This is true, but it is argued here that it is precisely those critical points that are the important ones in subsequent analysis.

Unique-DBSCAN

It is shown above that points belonging to multiple clusters result in the indeterministic behaviour of DBSCAN. A definition of such points is now introduced formally.

Definition D-10. A critical point x is a point for which there are clusters $C_i \in \Phi_{\varepsilon, m}$ and $C_j \in \Phi_{\varepsilon, m}$, such that $C_i \neq C_j$, and $x \in C_i \cap C_j$.

The critical points are points in the boundary zones between clusters, and play a crucial role in establishing cluster boundaries to be used as decision boundaries in subsequent classification algorithms. Critical points cannot be core points according to lemma D-7. It can be seen from definition D-10 that critical points are points that are border points for more than one cluster. The following definition will allow the formulation of an extended algorithm leading to unique clustering.

Definition D-11. The cluster membership function for points x and clusters C_i is

$\mu_i(x) = \mu(x, C_i) = 1$	if x is a core point of C_i ,
$\mu_i(x) = \mu(x, C_i) = (\varepsilon m)^{-1} \sum_j (\varepsilon - \ x, y_j\)$	if x is a border point of C_i , with y_j the core points of C_i in the ε -neighbourhood of x ,
$\mu_i(x) = \mu(x, C_i) = 0$	otherwise.

The membership functions μ_i take on values in the range $[0, 1]$. These functions immediately provide a fuzzy clustering. This is not further explored in the present context, as DBSCAN is a crisp clustering algorithm. The concept of membership functions will be used for resolving conflicts that arise with regard to critical points in DBSCAN. The original DBSCAN algorithm can now be formulated in terms of membership functions as algorithm 4.1.

Algorithm 4.1: *Ordinary DBSCAN.*

1. Construct a clustering $\Phi_{\varepsilon, m}$ (see Definition D-8).
2. Calculate membership functions (see Definition D-11).
3. Assign points to clusters as follows:
 - (a) if $\mu_i(x) \neq 0$ and $\mu_j(x) = 0$ for $j \neq i$, then $ca(x) = C_i$,
 - (b) if $\mu_i(x) = 0$ for all i , then $ca(x) = N_{\varepsilon, m}$,
 - (c) otherwise, $ca(x) = \text{rand}(\{C_i \mid \mu_i(x) \neq 0\})$.

The function $ca(.)$ is the cluster assignment function which assigns points to clusters. The function $rand(.)$ takes a set as its argument and selects one element from the set at random. The cluster $N_{\epsilon,m}$ is the noise set as in definition D-9. Step 3(c) in Algorithm 4.1 covers cluster assignment for critical points. Assigning such points to clusters at random is clearly undesirable. The developed formalism, in particular the introduction of the concept of membership functions, suggests an alternative algorithm, Algorithm 4.2, which will be called Unique-DBSCAN, or UDBSCAN in short.

Algorithm 4.2: *Unique-DBSCAN (UDBSCAN)*.

1. Construct a clustering $\Phi_{\epsilon,m}$ (see Definition D-8).
2. Calculate membership functions (see Definition D-11).
3. Assign points to clusters as follows:
 - (a) $ca(\mathbf{x}) = C_i$, with $i = \arg \max_j \mu_j(\mathbf{x})$ if i is unique,
 - (b) $ca(\mathbf{x}) = N_{\epsilon,m}$ otherwise.

UDBSCAN assigns all points to clusters based on the membership functions. This is in fact the classical approach of turning a fuzzy clustering into a crisp clustering: points are assigned to that cluster for which the membership function is maximal. The difference between Algorithms 4.1 and 4.2 is in the handling of critical points, points which have non-zero memberships for more than one cluster. While Algorithm 4.1 assigns them randomly to a cluster, Algorithm 4.2 assigns them maximizing the membership function. The ultimate achievement of the newly proposed UDBSCAN algorithm can now be proven.

Lemma D-9. UDBSCAN provides a unique clustering.

Proof

Core points and border points that are not critical points belong to precisely one cluster (lemma D-7 and definition D-10). Hence there is a unique membership function that is non-zero for such points (definition D-11) and step 3(a) in Algorithm 4.2 becomes trivial.

For critical points: assume the critical point \mathbf{x} is assigned to two different non-noise clusters C_i and C_j , $C_i \neq C_j$ in two different runs of the UDBSCAN algorithm. Since \mathbf{x} is assigned to C_i , $\mu_i(\mathbf{x}) > \mu_k(\mathbf{x})$ for all $C_k \neq C_i$, in particular $\mu_i(\mathbf{x}) > \mu_j(\mathbf{x})$. Equivalently, since \mathbf{x} was assigned to C_j in the second run, the following must hold also: $\mu_j(\mathbf{x}) > \mu_l(\mathbf{x})$ for all $C_l \neq C_j$, in particular $\mu_j(\mathbf{x}) > \mu_i(\mathbf{x})$. This is a contradiction. In the case where \mathbf{x} was assigned to a non-noise cluster C_i in one run, and the noise set in another run, $\mu_i(\mathbf{x}) > \mu_k(\mathbf{x})$ for all $C_k \neq C_i$. On the other hand, it being assigned to the noise set means that there were two different non-noise clusters C_i and some C_j for which $\mu_i(\mathbf{x}) = \mu_j(\mathbf{x})$. This is a contradiction and concludes the proof.

The same data set used in the example in Figure 4.9 and Figure 4.10 is now subjected to the UDBSCAN algorithm. The unique classification resulting from this experiment is represented in Figure 4.11. While the main achievement is the uniqueness of the clustering, it can be seen that the UDBSCAN-clustering matches our intuitive understanding more closely than some of the various clusterings that can arise from the ordinary DBSCAN algorithm.

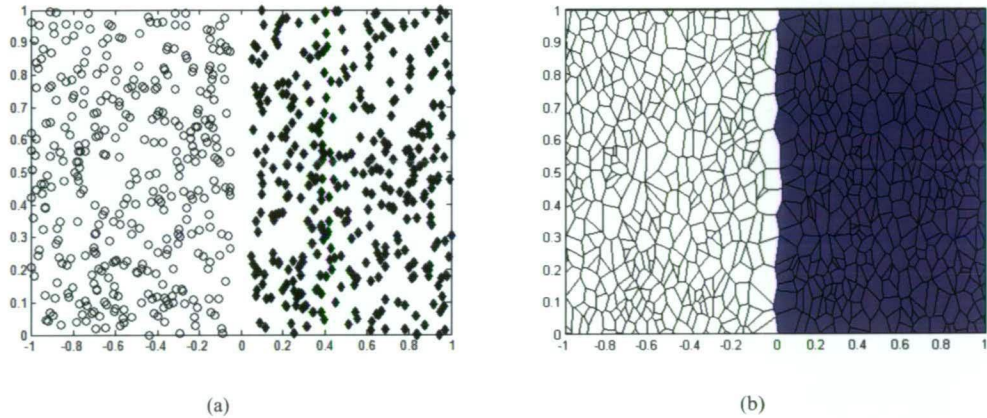


Figure 4.11 (a) and (b) The example data as clustered by UDBSCAN. The UDBSCAN clustering is unique.

The precise extent of the difference in results between DBSCAN and UDBSCAN depends on the data at hand. The number of border points for a cluster depends on the ratio of the length of its border to its area, and on the density. The number of border points that are critical points depends on the spatial proximity of other clusters. In areas where clusters are close together, separating them is a difficult problem, and is not one which DBSCAN is known to handle well (Yip *et al.*, 2006). The data points in such zones are crucial in determining boundaries between clusters, hence a consistent non-random cluster assignment is essential. UDBSCAN accommodates this need.

Benefits of (U)DBSCAN

DBSCAN was selected from a range of standard algorithms as one that is considered promising with respect to clustering scatter nodes based on their spatial components. It scored best against a range of requirements that were put forward (see Table 4.1). Modifications are made, leading to UDBSCAN. A preliminary, simple example has been given already (Figure 4.8). Further examples are given in chapter 5, where a number of case studies are presented.

It is in any case convenient to have access to a standard algorithm, which can be applied easily, and which can be expected to give reasonable first results in terms of clustering scatter nodes. UDBSCAN will fulfil this role, and be utilized for

- obtaining a first clustering quickly and easily,
- providing a reference against which other algorithms can be compared.

If, in a given situation, the UDBSCAN clustering is satisfactory (see section 4.5.5), there is of course no need pursue the application of alternative clustering algorithms. However, it is expected that using more information from the scatter nodes than only their spatial components –which cannot be done with UDBSCAN– will lead to improved results.

Limitations of (U)DBSCAN

In the example data set of Figure 4.6, DBSCAN is able to find the expected clusters, as shown in Figure 4.8. In fact, UDBSCAN was used to find these clusters. This data set is rather clean and does not pose any problems.

A still simple yet slightly more complex data set is presented in Figure 4.12. It is the same data set as Figure 4.6, but now with a second, smaller aggregation of nodes added not far above to the seabed (indicated by a red ellipse). Some random noise is also introduced.

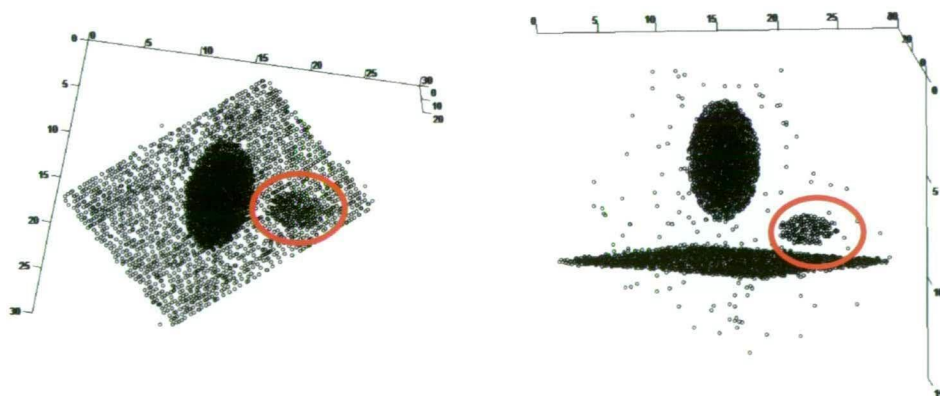


Figure 4.12 The same data set as in Figure 4.6, but now with an additional fish school (indicated by the red ellipse) and some random noise added.

Running UDBSCAN with the same settings as those used to obtain the clusters in Figure 4.8 results in the clusters shown in Figure 4.13. Again, only two clusters are

found, indicated by blue and red. The newly added aggregation is **not** detected as a separate cluster, rather, it is determined to belong to the same cluster as the seabed (both in blue). The large aggregation is correctly identified (in red).

The parameters are subsequently adjusted in order to find finer detail in the data set. More clusters are indeed found, with an example with 7 clusters shown in Figure 4.14. However, the new aggregation is still in the same cluster as the seabed (both in blue), with additional smaller, spurious clusters found within the seabed cluster (indicated by colours other than blue). The large aggregation is still isolated as a cluster (in red).

Also worth observing is the removal of noise, which was successful. The noise (data points not belonging to any cluster) in Figure 4.12 is no longer present in Figure 4.13 and Figure 4.14. UDBSCAN has identified those points correctly as noise; they are not plotted in these figures.

UDBSCAN has difficulty detecting the smaller aggregation near the seabed as a cluster because that aggregation is not separated enough from the nodes representing the seabed. This is indicative of a more general limitation of DBSCAN and UDBSCAN: the zones of high density must be well separated, otherwise they cannot be identified as separate clusters. The presence of noise in those zones can also disrupt the expected working of UDBSCAN, as the noise samples can act as a bridge between two high density zones.

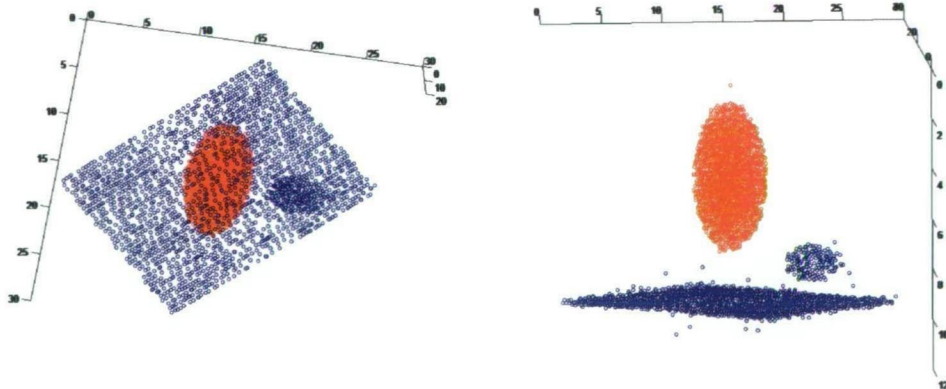


Figure 4.13 The results of UDBSCAN when applied to the data set of Figure 4.12, with the same settings as those used in obtaining the results presented in Figure 4.8. UDBSCAN is unable to identify the newly added cluster.

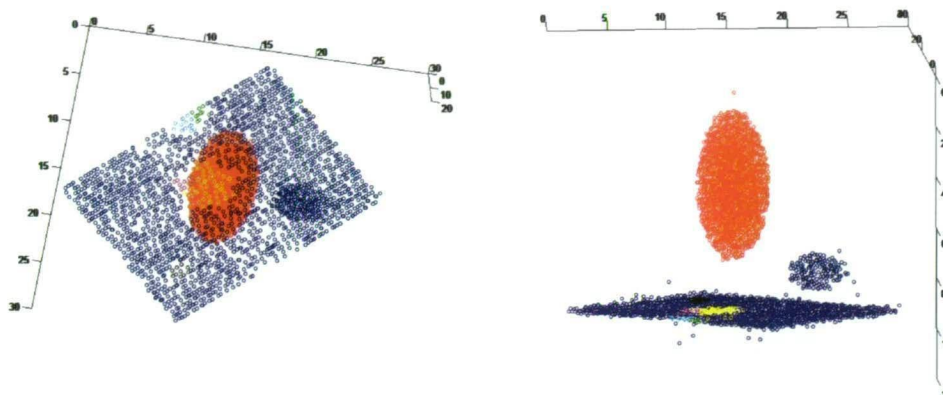


Figure 4.14 Changing the settings of the UDBSCAN algorithm for it to become more sensitive results in more clusters being found. In this case 7 clusters are identified, but rather than detecting the smaller school near the seabed, spurious, meaningless clusters are introduced.

In the examples presented here, only the spatial coordinates of scatter nodes are used. However, scatter nodes include, besides spatial properties, also a temporal and a general feature vector as components. So far these non-spatial components have been ignored when clustering scatter nodes. Bringing these into play is not a trivial matter, but it can be expected that there is value in doing so. For example, the scatter nodes in the small aggregation near the seabed may have backscatter energy levels that are a lot less than those of the scatter nodes indicative of the seabed. The next section will develop a method that uses such differences in non-spatial components to improve the clustering.

4.4 KERNEL METHODS FOR CLUSTERING

4.4.1 Concept

Scatter nodes have three distinct components: a spatial, a temporal and a feature component. In applying processing algorithms, there is benefit in utilising this knowledge for fine-tuning the clustering algorithms. No obvious way to do this could be seen in relation to any of the clustering techniques discussed in section 4.3.2 and listed in Table 4.1. These methods are all relatively simple. They provide different sensible and well understood approaches to clustering data points. However, they are rather restricted in terms of clusters they are able to determine, or are lacking theoretical foundations upon which could be built to extend them to handle spatio-temporal feature vectors. K-means, CLARANS and BIRCH are linear

methods, in the sense that boundaries between clusters are hyper-planes. This is relaxed slightly with Gaussian mixtures, in which clusters are hyper-ellipsoidal in shape with boundaries that can be quadratic. Chameleon and DBSCAN are non-linear in that any shape of cluster could arise. However, they are ad-hoc methods, based on rather intuitive concepts. Chameleon is based on proximity rankings rather than proximity metrics, which renders it less useful for spatial clustering. DBSCAN on the other hand is specifically a spatial clustering algorithm, which in turn is a limiting factor as no obvious possible extensions to include non-spatial components of data points are available. In this section, a number of alternative methods are considered.

Neural networks

The most commonly used neural network based clustering algorithm is the Self-Organising Map (SOM) (Kohonen, 2001). The objective of SOM is to represent high-dimensional input data with prototype vectors that can be visualised in a usually two-dimensional lattice structure. Each unit in the lattice is a neuron, and adjacent neurons are connected to each other, providing a topological representation of how the lattice fits itself in the input space. Input data points are connected to all neurons via adaptable weights. During the training phase these weights are adjusted to project input points to adjacent neurons.

The benefit of this approach to clustering scatter nodes is that the neurons in the lattice provide an approximation to the density in a non-parametric and non-linear fashion. However, the SOM algorithm has a number of problems (Xu and Wunsch, 2005). The number of neurons (lattice size) must be chosen in advance, which is difficult. Another problem is how SOM copes with areas of varying densities: low density areas may end up over-represented, while high-density areas may end up under-represented. Finally, the training algorithm makes use of a metric in input space. It is not clear how this could be generalized to include temporal and feature components of spatial scatter nodes.

Association rules

Association rules (Agrawal *et al.*, 1995) attempt to find regions of higher probability in the input space by determining a number of prototype vectors which have a relatively high probability of occurring. This problem is severely plagued by the so-called *curse of dimensionality* (Bellman, 1961): for typical dimensions in feature spaces (five and higher), the number of data points that would be needed to find these prototypes reliably is generally unrealistically high.

In order to obtain a tractable problem, both the goals of the analysis and the types of data to which it is applied are greatly simplified. The goal is simplified by seeking a subset of the data points that have a higher joint probability of occurring, rather than seeking a single representative. In terms of data types, association rules are commonly applied to categorical data, which is much easier than continuous data (Hastie *et al.*, 2001).

In scatter nodes, the spatial and temporal components are continuous, as are all but some of the features proposed in section 3.4.2. Continuous variables can be categorized by defining a limited set of intervals and quantizing the continuous data into this limited set of categories. Doing this with scatter node data is not expected to provide satisfactory results as either too much detail is lost in the quantization, or too many levels are defined, preventing the association rules algorithm converging.

Kernel methods

Kernel methods arise from the theory of statistical learning (Vapnik, 1995; Shawe-Taylor and Cristianini, 2004). Kernel methods are a mathematically sound way of applying known linear methods in a non-linear fashion, through an implicit mapping of data points in a high-dimensional space. These methods are known to have good convergence properties and their working is well understood, since there is an underpinning mathematical formalism based on probability, linear algebra and functional analysis (Scholkopf and Smola, 2001).

As with the other methods, it is not immediately clear how kernel methods can be applied to spatio-temporal vector data. However, their well defined properties enable the development of a mathematical framework for using kernel methods for spatio-temporal vectors, as is presented below.

In summary

Since kernel methods compare favourably with the other methods considered (Table 4.2), kernel methods are investigated further. One drawback is that kernel methods are computationally intensive and of higher computational complexity than the other methods.

Kernel methods originate from the theory of statistical learning, introduced in section 4.4.2, and are presented in section 4.4.3. In section 4.4.4 the necessary mathematical foundations are developed to extend kernel methods to handle spatio-temporal feature vectors, which are discussed further in section 4.4.5. Section 4.4.6 introduces kernel based clustering methods, which are applied to scatter nodes in section 4.4.7.

	<i>Mathematically and statistically founded</i>	<i>Capability to handle non-linear problems well</i>	<i>Understandable behaviour</i>	<i>Convergence to global optimum</i>	<i>Runtime simplicity</i>
Neural networks (SOM)	-	+	-	-	+
Association rules	+	-	+	-	+
Kernel methods	+	+	+	+	-

Table 4.2 Overview of advanced approaches to clustering, and their characteristics (+ : property applies, - : property does not apply)

4.4.2 Statistical learning theory

Statistical learning theory describes the framework in which learning algorithms are formulated. The theory of Vapnik and Chervonenkis (VC) describes such a framework for learning methods, with the main objective of controlling the generalisation capabilities of learning algorithms (Vapnik, 1995). Generalisation is an important aspect of learning algorithms: an algorithm generalizes well if it performs the task of classifying unseen data well. VC theory forms the basis of support vector machines (Scholkopf *et al.*, 1999; Cristianini and Shawe-Taylor 2000) and kernel methods (Muller *et al.*, 2001; Shawe-Taylor and Cristianini, 2004). A brief overview of VC theory is given.

Typically, in a pattern analysis or machine learning problem, the available data set is split in two parts: a training set and a test set. The training set is used to train the algorithm, the test set is used to evaluate the trained algorithm. Intuitively, a good algorithm must perform its task well on both sets. Algorithms which perform well on the training set but not on the test set are said to be over-fitting. Forcing such algorithms to perform well on the test set also comes at the cost of a reduced performance on the training set. A balance needs to be found between over-fitting and over-generalizing. VC theory provides a mathematical framework to describe these intuitive ideas in a formal manner.

A loss function L is defined, which describes the performance of a learning algorithm f on a data set D :

$$L(f, D) \in \mathbb{R} . \tag{4.9}$$

L is the error rate of algorithm f on D and takes on values in \mathfrak{R} , the real numbers. The error on the training set is known as *training error*, or *empirical risk*:

$$R_{emp}(f) = L(f, D_{tr}). \quad (4.10)$$

with D_{tr} the training data set. Minimizing empirical risk only does not guarantee a small error on the test set D_{te} . This error is given by:

$$R(f) = L(f, D_{te}). \quad (4.11)$$

In VC theory it is imperative to restrict the class of functions from which f is chosen to one which has suitable *capacity* for the amount of test data that is available. Capacity is a concept of complexity. Within VC theory, the so-called Vapnik-Chervonenkis dimension is used for this purpose (Vapnik, 1995). The VC dimension of a function f is equal to the number of points that can be *shattered* by f . A function f can *shatter* m points if it can be tuned so as to assign any m points of D_{tr} to one class or cluster, and all the other points to another one.

Given the training set, the empirical risk and VC dimension can be calculated for any f . The great achievement of VC theory is that it can provide bounds on the test error, based only on the empirical risk and VC dimension:

$$R(f) \leq \varepsilon(R_{emp}, VCdim(f)), \quad (4.12)$$

with $VCdim(f)$ the VC dimension of f .

The minimization of this bound on the test error leads to the principle that is known as *structural risk minimization*. In this sense, structural risk is the maximum expected test error.

The *support vector machine* arises naturally from VC theory: when f is restricted to the class of linear functions, the resulting structural risk minimization algorithm is the support vector machine (Vapnik, 1995; Scholkopf *et al.*, 1999). In this context, the concept of mapping the input space to a higher-dimensional feature space was introduced and has formed an integral part of support vector machines since their inception. This concept has been generalized, leading to the more general class of learning algorithms known as *kernel methods*, discussed in the next section.

Support vector machines in particular, and kernel methods in general, are commonly described without explicit reference to the underpinning statistical learning theory of Vapnik and Chervonenkis. However, this underlying theory provides a powerful framework and sound mathematical foundation on which these methods are based.

4.4.3 Kernel methods

In this section an overview of kernel methods is given. It is based on a number of review papers (Muller *et al.*, 2001; Campbell, 2002), an introductory book chapter (Scholkopf and Smola, 2001), and a recent text book on kernel methods (Shawe-Taylor and Cristianini, 2004).

Kernel methods are a generalization of the support vector machine. They are non-linear versions of known linear methods, where the non-linearity is introduced by means of a mapping of the input space to a high and possibly infinite dimensional feature space.

Data points x_i in the space X are embedded in what is generally called a feature space, F , by means of a mapping φ :

$$\varphi: X \mapsto F : x \in X \mapsto \varphi(x) \in F . \quad (4.13)$$

To avoid confusion it is worth stating explicitly that the components of the data points x_i are commonly thought of as features also. However, for clarity the term feature space is in this context reserved for the space F into which the points x_i are mapped.

The purpose of this mapping is that it is hoped that in the space F , different classes or clusters of points can be separated linearly, while that may not be possible in the original space X . Since the space F can be of higher dimension than X , it is expected that this will be possible in general. This concept is illustrated in Figure 4.15.

The algorithms that seek linear relationships in the feature space F are such that the actual coordinates of the vectors $\varphi(x_i)$ are not needed: rather, algorithms are formulated in terms of *inner products* only.

Prior to introducing kernels, some concepts from functional analysis are needed for defining and describing spaces for which inner products exist.

Definition 4.1 (Inner product space). A vector space F is an inner product space if there exists a real-valued symmetric bilinear map $\langle \cdot, \cdot \rangle$ that satisfies $\langle f, f \rangle \geq 0$ for all $f \in F$.

Bilinear means that $\langle \cdot, \cdot \rangle$ is linear in each of its arguments. The map is known as the inner product. Furthermore, the inner product is strict if $\langle f, f \rangle = 0$ if and only if $f = 0$.

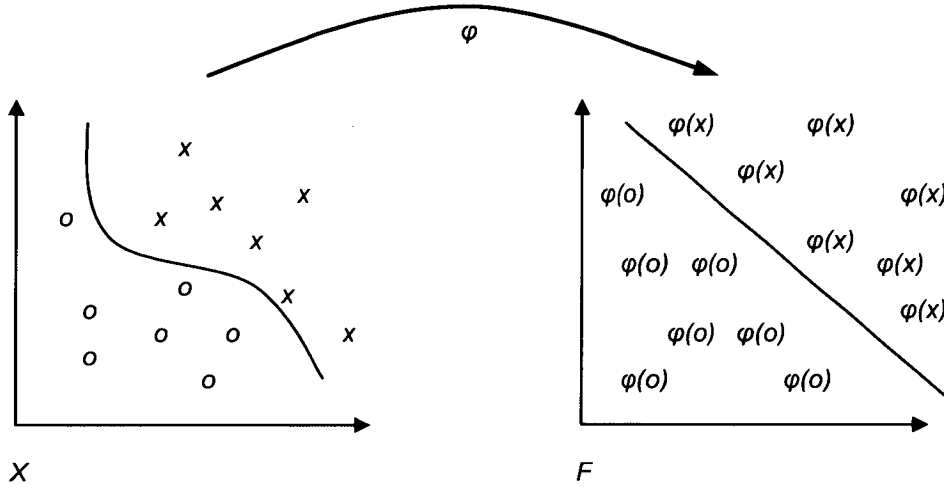


Figure 4.15 Illustration of the concept of mapping into feature space. The separation boundary between the groups of points represented by 'o' and those represented by 'x' is non-linear in X while it is linear in F . Image based on Fig. 2.1 of (Shawe-Taylor and Cristianini, 2004).

The norm induced by the inner product of a strict inner product space is defined as:

$$\|f\| = \langle f, f \rangle^{1/2}. \quad (4.14)$$

The metric d_f associated with this norm is given by:

$$d_f(f_i, f_j) = \|f_j - f_i\|. \quad (4.15)$$

Using symmetry and bilinearity of the inner product, this can be written as:

$$\begin{aligned} d_f(f_i, f_j)^2 &= \|f_j - f_i\|^2 = \langle f_j - f_i, f_j - f_i \rangle \\ &= \langle f_j, f_j \rangle - \langle f_i, f_j \rangle - \langle f_j, f_i \rangle + \langle f_i, f_i \rangle \\ &= \langle f_i, f_i \rangle - 2\langle f_i, f_j \rangle + \langle f_j, f_j \rangle. \end{aligned} \quad (4.16)$$

A vector space with a metric is referred to as a *metric space*. For the further development of kernel methods it is convenient to require the space F to have some additional properties, leading to the concept of *Hilbert space*.

Definition 4.2 (Hilbert space). A Hilbert space F is an inner product space with the additional properties that it is *complete* and *separable*.

Completeness refers to the property that every Cauchy sequence $\{h_n\}_{n \geq 1}$ of elements of F converges to an element $h \in F$, where a Cauchy sequence is one satisfying the property that $\sup_{m > n} \|h_n - h_m\| \rightarrow 0$ as $n \rightarrow \infty$. Separability refers to the property that for any $\varepsilon > 0$ there is a countable set of elements $h_1, h_2, \dots \in F$ such that for all $h \in F$, $\min_i \|h_i - h\| < \varepsilon$. Intuitively these properties describe that the limit of a converging sequence exists in the space F , and that the space F contains a countable dense subset (Moore, 1985).

Provided with the Hilbert space F and an inner product $\langle \cdot, \cdot \rangle$ defined over F , kernel methods are now described. Kernel methods are algorithms that are linear in F and can be written in terms of pair wise inner products of the embedded data points only, $\langle \varphi(x_i), \varphi(x_j) \rangle$. This latter expression can be seen as a function on the original space X , defined in terms of the mapping φ and the inner product. This special kind of function is a *kernel*, κ :

$$\kappa : X \mapsto \mathbb{R} : \kappa(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle. \quad (4.17)$$

Through this mechanism, kernels enable the calculation of distances in the feature space F . Using the expansion of the metric d_f given in expression (4.16):

$$\begin{aligned} d_f(\varphi(x_i), \varphi(x_j))^2 &= \langle \varphi(x_i), \varphi(x_i) \rangle - 2\langle \varphi(x_i), \varphi(x_j) \rangle + \langle \varphi(x_j), \varphi(x_j) \rangle \\ &= \kappa(x_i, x_i) - 2\kappa(x_i, x_j) + \kappa(x_j, x_j). \end{aligned} \quad (4.18)$$

Using this formula, distances between mapped points can be obtained from the original data points directly, by means of the kernel function κ .

Analytical expressions for kernels are stated in terms of x_i and x_j only, without considering the embedding function φ . This bypassing of the function φ , and of the feature space F , has become known as *the kernel trick*. When algorithms applied to the points in F can be written in terms of distances only, they can make use of the kernel trick.

The fact that kernels must arise as inner products of mappings into a Hilbert space poses some restrictions on the class of functions that can fulfil the role of a kernel. Valid kernel functions that are in common use include:

- the Gaussian kernel, sometimes called Radial Basis Function kernel or RBF-kernel:

$$\kappa(x_i, x_j) = e^{-(x_i - x_j)^2 / 2\sigma^2}$$

with σ^2 the width of the kernel (the variance parameter),

- the polynomial kernel of degree d with parameter a :

$$\kappa(x_i, x_j) = (x_i \cdot x_j + a)^d,$$

- the sigmoidal kernel with parameter b :

$$\kappa(x_i, x_j) = \tanh(x_i \cdot x_j + b),$$

- the linear kernel:

$$\kappa(x_i, x_j) = (x_i \cdot x_j),$$

which is the kernel obtained by choosing the identity function for the map φ .

Expression (4.17) defines a kernel function κ . As stated before, kernel methods are algorithms that do not need access to the embedded data points $\varphi(x_i)$ explicitly, but only to the pair-wise inner products. Because of the linear nature of the algorithms, it is often possible and convenient to express them in algebraic terms. For that purpose the so-called *kernel matrix* \mathbf{K} is introduced. The matrix \mathbf{K} contains all pairs $\kappa(x_i, x_j)$ as its elements:

$$\mathbf{K}_{ij} = \kappa(x_i, x_j) \tag{4.19}$$

with \mathbf{K}_{ij} the matrix element in row i , column j . The matrix \mathbf{K} is sometimes referred to as the *Gram matrix*.

An important property of a kernel matrix is that it is positive semi-definite: it is symmetric and has positive eigenvalues. The reverse also holds: any semi-definite matrix defines a kernel over the input space X . This is a direct result of the fact that kernels are defined in terms of inner products; a formal proof can be found in the literature (Cristianini and Shawe-Taylor 2000).

4.4.4 The Hahn-Banach theorem

From expressions (4.17) and (4.18) it is clear that kernels are closely related to the norm on the space F , as they are defined in terms of an inner product, which induces the norm (4.14) and associated metric (4.15).

It is exactly this general concept of a metric that is exploited in this thesis, to extend kernel methods to spatio-temporal spaces in the next section (4.4.5). Some results from the theory of functional analysis are used, in particular a very important theorem, known as the *Hahn-Banach theorem*. The theorem is named after Hans Hahn and Stefan Banach who independently proved it in the 1920s (Hahn, 1927; Banach, 1929). Narici and Beckenstein (1997) provide a comprehensive historical and mathematical overview of this important theorem.

Hilbert spaces are introduced in the previous section. This concept is relaxed, leading to *Banach spaces*.

Definition 4.3 (*Banach space*). A Banach space is a complete normed vector space.

In general, in Banach spaces, the norm does not arise from an inner product, as is the case with Hilbert spaces. A Hilbert space is a Banach space but the converse is not necessarily true. Banach spaces are introduced here as they are the spaces for which the Hahn-Banach (HB) theorem holds.

The HB-theorem is an important theorem in applied functional analysis, more specifically in constructing linear functionals with certain properties. This is usually done in two steps. First, a linear functional is constructed on a subspace of the Banach space where it is easy to verify the desired properties. Second, one appeals to a general theorem which says that any such functional can be extended to the whole space while retaining the desired properties. The basic tool of the second step is the HB-theorem.

First, *linear functionals* are introduced. A functional is a real-valued mapping defined on a Banach space. If the mapping is linear, it is called a linear functional.

In particular, inner products give rise to linear functionals in the following manner. If H is an inner product space and h is any element in H , then:

$$p_h(x) = \langle x, h \rangle \tag{4.20}$$

is a linear functional on H .

The norm of a functional is defined through the norm on the Banach space B :

$$f : B \mapsto \mathbb{R}, \|f\| = \sup_{b, \|b\|=1} |f(b)|. \quad (4.21)$$

In other words, the norm of a functional f is its supremum on the unit sphere in B .

The HB theorem is now formulated:

Theorem 4.1 (Hahn-Banach). For a Banach space B , consider a continuous linear functional f defined on a proper linear subspace M of B . Then f can be extended as a continuous linear functional f_0 such that $\|f_0\|$ on B is equal to $\|f_0\|$ on M and $f(b) = f_0(b)$ for $b \in M$.

A proper linear subspace M of a Banach space B is a Banach space itself, which is not empty, contains fewer elements than B , and contains all linear combinations of its elements.

Various proofs of the HB-theorem are found in the literature (Reed and Simon, 1980; Narici and Beckenstein, 1997; Giles, 2000), as well as in the original publications (Hahn, 1927; Banach, 1929).

The HB-theorem states that any linear functional on a subspace of a Banach space can be extended to the whole space. Since each Hilbert space is a Banach space and inner products are linear functionals, the HB-theorem applies to inner products defined on Hilbert spaces. Inner products give rise to kernels (section 4.4.3).

4.4.5 Kernels for spatio-temporal feature vectors

In this section, the HB-theorem is used to build kernels for spatio-temporal feature vectors. This will allow for the application of kernel methods to scatter nodes. The development of this enabling formalism is an important contribution of the research presented in this thesis.

For clarity in the current context the notation from expression (3.19), defining scatter nodes, is changed to:

$$x = (x_s, x_t, x_v) \quad (4.22)$$

where now x are the scatter nodes with spatial, temporal and feature components x_s , x_t and x_v respectively. Let S , T and V be the coordinate spaces of these components:

$$x_s \in S, x_t \in T, x_v \in V. \quad (4.23)$$

Let the joint coordinate space be X , so that $x \in X$. The space X consists of the subspaces S , T , and V , formally:

$$X = S \oplus T \oplus V. \quad (4.24)$$

Kernel methods can be applied to each of the spaces S , T and V separately. There are advantages to analysing each of the spaces S , T and V separately and then combining the findings to form a unified approach over the space X . These advantages and the flexibility they offer are discussed in section 4.4.6. Here, the underpinning mathematical theory needed to do this is developed.

First, kernels are defined on each of the subspaces separately. An implicit mapping to a higher dimensional feature space corresponds with each kernel. Denote the kernels with κ_s , κ_t , and κ_v . They are defined in terms of inner products on the Hilbert spaces they induce, denoted by H_s , H_t , and H_v :

$$\begin{aligned} \kappa_s(x_{s1}, x_{s2}) &= \langle \varphi_s(x_{s1}), \varphi_s(x_{s2}) \rangle \\ \kappa_t(x_{t1}, x_{t2}) &= \langle \varphi_t(x_{t1}), \varphi_t(x_{t2}) \rangle \\ \kappa_v(x_{v1}, x_{v2}) &= \langle \varphi_v(x_{v1}), \varphi_v(x_{v2}) \rangle \end{aligned} \quad (4.25)$$

with:

$$\begin{aligned} \varphi_s : x_s \in S &\rightarrow \varphi_s(x_s) \in H_s \\ \varphi_t : x_t \in T &\rightarrow \varphi_t(x_t) \in H_t \\ \varphi_v : x_v \in V &\rightarrow \varphi_v(x_v) \in H_v. \end{aligned} \quad (4.26)$$

This approach offers the flexibility and freedom to define different kernels on each of the spaces S , T and V separately.

Next, the Hahn-Banach theorem is used to extend each of these sub-space kernels to the whole space X . The extended kernels can be combined to form a single kernel function that will be used by subsequent kernel algorithms.

Using the Hahn-Banach theorem, the kernels κ_s , κ_t , and κ_v are extended to the whole space X . This is valid since kernels are linear functionals on the Hilbert spaces H_s , H_t , and H_v . These Hilbert spaces constitute the larger Hilbert space H_x , formed by the subspaces H_s , H_t , and H_v :

$$H_x = H_s \oplus H_t \oplus H_v. \quad (4.27)$$

The subspace kernels are extended in the following manner, for $x_1, x_2 \in X$:

$$\begin{aligned}
\kappa'_s(x_1, x_2) &= \kappa_s(s(x_1), s(x_2)) \\
\kappa'_t(x_1, x_2) &= \kappa_t(t(x_1), t(x_2)) \\
\kappa'_v(x_1, x_2) &= \kappa_v(v(x_1), v(x_2))
\end{aligned} \tag{4.28}$$

with $s(\cdot)$, $t(\cdot)$ and $v(\cdot)$ canonical projections from X on the subspaces S , T and V respectively.

The three kernels κ'_s , κ'_t , and κ'_v can now be combined to form a single kernel defined over the space X . There are a number of ways to achieve valid combinations.

Given two kernels κ_1 and κ_2 , the following functions are kernels also:

$$\begin{aligned}
\kappa(x_1, x_2) &= \kappa_1(x_1, x_2) + \kappa_2(x_1, x_2), \\
\kappa(x_1, x_2) &= \kappa_1(x_1, x_2) \kappa_2(x_1, x_2), \\
\kappa(x_1, x_2) &= a \kappa_1(x_1, x_2),
\end{aligned} \tag{4.29}$$

with a any positive real number. For a proof see Shawe-Taylor and Cristianini (2004).

The significance of extending sub-space kernels and combining them together lies in the fact that the resulting single kernel can be used by any kernel algorithm, to cluster scatter nodes in the joint spatial, temporal and feature spaces. This provides a powerful framework to handle disparate spaces in a unified fashion. The benefits offered by this approach will become clear in the next section, where kernel-based clustering algorithms are discussed in detail.

4.4.6 Clustering with kernels

An attractive aspect of kernel methods is that they offer a mathematically sound approach to developing non-linear variants of well established linear algorithms. Doing this has become known as *kernelizing* a method.

Since kernels arise as inner products in a feature space, the kernel function can be seen as a measure of similarity. Two input data points that are similar are mapped close together in the feature space, resulting in relatively high inner product values. The kernel matrix (4.19) is sometimes called the *similarity matrix*. This is an intuitive representation of the data, which has deep theoretical foundations and which is both simple and convenient.

Many kernel methods are formulated as algorithms that take such a similarity matrix as input. The kernel matrix acts as an interface between the data and the kernel-based pattern analysis methods.

In spatial clustering, spatial proximity based on Euclidian distances is used to define similarity between data points. Data points are similar simply when they are close to each other. The clustering process aims to group similar data points. This is the essential idea of clustering algorithms in general. Data points are characterized by a number of features. Clustering algorithms break the data points up in groups so that within each group or cluster, data points are similar, and they are dissimilar between groups. Kernel clustering methods use the kernel matrix as the reference for similarity.

Kernel k-means

The first reported kernel clustering method was named *support vector clustering* (Ben-Hur *et al.*, 2000), after its famous classification variant the support vector machine (Vapnik, 1995; Scholkopf *et al.*, 1999; Cristianini and Shawe-Taylor 2000). Support vector clustering is based on finding a number of data points in feature space which define the boundaries between clusters. These key points are called support vectors, similar to the role they have in support vector machines. This is different from standard clustering methods, where there is typically some reference point for each cluster, with cluster membership defined as the distance to the reference points. The support vector clustering procedure is rather cumbersome to implement and has not emerged as the best approach to clustering using kernels.

An alternative to support vector clustering is the kernelization of the famous k-means algorithm. This was first mentioned as a possibility by Scholkopf *et al.* (1998) and further developed by Girolami (2002). Kernel k-means has become the standard base-line kernel based clustering method, similar to the role fulfilled by the usual k-means algorithm in linear clustering.

Let $\{x_1, \dots, x_n\}$ be the data set to be clustered, $x_i \in X$. The x_i consist of spatial, temporal and feature components, as in expression (4.22). For each of these components, suitable kernel functions are chosen. These kernel functions are combined to form the kernel function κ . The $n \times n$ kernel matrix is constructed:

$$K_{ij} = \kappa(x_i, x_j), \text{ for } i, j = 1, \dots, n. \quad (4.30)$$

Implicitly, the kernel κ defines a mapping φ from X into the Hilbert space F .

Similar to the objective function (4.2) for regular k-means, an objective function is now formulated based on the metric in the feature space F :

$$J_\varphi = \sum_{i=1}^n \sum_{j=1}^k r_{ij} \|\varphi(x_i) - \mu_j^\varphi\|^2 \quad (4.31)$$

with, as before, k the number of clusters, r_{ij} the indicator variables, $r_{ij} = 1$ if data point i belongs to cluster j , and 0 otherwise. The norm used in eq. (4.31) is the norm arising from the inner product, as in expression (4.14). The cluster centres μ_j^φ are points in the feature space F :

$$\mu_j^\varphi = \frac{1}{n_j} \sum_{i=1}^n r_{ij} \varphi(x_i) \quad (4.32)$$

with $n_j = \sum_{i=1}^n r_{ij}$ the number of elements assigned to cluster j .

Essentially, the regular k-means algorithm is applied in the space F , using the data points $\{\varphi(x_1), \dots, \varphi(x_n)\}$, cluster centres $\{\mu_j^\varphi\}_{1 \leq j \leq k}$, and objective function J_φ . However, since the mapping φ is known only implicitly, that cannot be done by simply proceeding as in section 4.3.2. The problem is therefore reformulated in terms of the kernel matrix. This is applying the *kernel-trick* to the k-means algorithm.

The calculation of the norm in eq. (4.31) can be rewritten in terms of the kernel function:

$$\begin{aligned} \|\varphi(x_i) - \mu_j^\varphi\|^2 &= \langle \varphi(x_i) - \mu_j^\varphi, \varphi(x_i) - \mu_j^\varphi \rangle \\ &= \langle \varphi(x_i), \varphi(x_i) \rangle - 2\langle \varphi(x_i), \mu_j^\varphi \rangle + \langle \mu_j^\varphi, \mu_j^\varphi \rangle \\ &= \langle \varphi(x_i), \varphi(x_i) \rangle - \frac{2}{n_j} \sum_{s=1}^n r_{sj} \langle \varphi(x_i), \varphi(x_s) \rangle + \frac{1}{n_j^2} \sum_{s=1}^n \sum_{t=1}^n r_{sj} r_{tj} \langle \varphi(x_s), \varphi(x_t) \rangle \\ &= \kappa(x_i, x_i) - \frac{2}{n_j} \sum_{s=1}^n r_{sj} \kappa(x_i, x_s) + \frac{1}{n_j^2} \sum_{s=1}^n \sum_{t=1}^n r_{sj} r_{tj} \kappa(x_s, x_t). \end{aligned} \quad (4.33)$$

Using this expression, the objective function is rewritten as:

$$J_\varphi = \sum_{i=1}^n \sum_{j=1}^k r_{ij} \left(\kappa(x_i, x_i) - \frac{2}{n_j} \sum_{s=1}^n r_{sj} \kappa(x_i, x_s) + \frac{1}{n_j^2} \sum_{s=1}^n \sum_{t=1}^n r_{sj} r_{tj} \kappa(x_s, x_t) \right). \quad (4.34)$$

The kernel k-means algorithm proceeds by updating clusters until there is no improvement in J_φ , which is now written in terms of the indicator variables and kernel function only.

The regular k-means iterations consist of two phases, a reassignment of points to clusters, and an update of the cluster centres. The kernel k-means algorithm lacks the step in which the cluster centres are updated. This is caused by the implicit mapping into feature space via the kernel function.

Kernel k-means is an important improvement over regular k-means, as in kernel k-means cluster shapes are not restricted to being hyper-spherical. They are hyper-spherical in the feature space F , but they could be of any shape in the input space X since the mapping ϕ can be any non-linear mapping, and F can be of arbitrary large dimension. A comparative study of clustering algorithms including k-means shows that the kernelized versions of the algorithms perform better than their standard counterparts (Kim *et al.*, 2005).

Kernel k-means is as popular a kernel method as regular k-means is a standard linear clustering method. However, kernel k-means does suffer from the same problem as regular k-means in that it is not guaranteed to converge to a global optimum of the respective objective function. Convergence is guaranteed, but the optimum that the algorithms converge to could be local.

The next section discusses solutions to this problem and suggests ways to relax the initial problem so that a global optimum can be found.

Spectral clustering

Spectral clustering methods are a recent, new approach to clustering. A first review paper discussing these new developments gives an overview of some common methods (Verma and Meila, 2003). One such method concentrates on providing a so-called *spectral relaxation* for the regular k-means clustering algorithm (Zha *et al.*, 2001), without referring to the kernel k-means algorithm. It is only recently that the relation between kernel k-means and spectral methods has been discovered (Dhillon *et al.*, 2004; 2005). An assessment of this unified view is presented in von Luxburg (2006).

Spectral clustering methods are clustering algorithms that are based on the *spectrum* of some matrix. The spectrum of a matrix is the set of its eigenvalues. The matrix that is typically used by spectral clustering methods is a matrix that is derived from the data. Since one such matrix is the kernel matrix, it is not surprising that there is a connection between spectral methods and kernel clustering methods such as k-means.

The spectral clustering method that has emerged as the reference algorithm was first published by Ng *et al.* (2002). It has since become known in the literature as the

Ng-Jordan-Weiss (NJW) algorithm, after the authors of the paper. There are fundamental similarities between the NJW algorithm and kernel k-means clustering (Dhillon *et al.*, 2004).

Spectral methods do not make use of cluster means, as the k-means algorithms do, nor do they make use of any other kind of cluster representatives. Rather, clusters arise as subsets of the data set to be clustered, $\{x_1, \dots, x_n\}$. Let C_1, \dots, C_k be k clusters, then:

$$\begin{aligned} C_i &\subseteq \{x_1, \dots, x_n\} \\ \bigcup_i C_i &= \{x_1, \dots, x_n\} \\ C_i \cap C_j &= \{\}, \forall i, j. \end{aligned} \tag{4.35}$$

While spectral methods use the general concept of similarity between data points, in the current context the kernel matrix is used for this purpose, in accordance with discussions of the relation between kernel methods and spectral methods in the literature (Cristianini *et al.*, 2001; Dhillon *et al.*, 2004; 2005; von Luxburg, 2006).

Since no cluster means are available, objective functions like those that form the basis of the k-means and kernel k-means algorithms cannot be used. An alternative criterion is used: the *cut cost*. The concept of cut cost is derived from graph theory. If each data point x_i is considered a node in a graph, and the value of the kernel function $\kappa(x_i, x_j)$ is the weight of the edge in the graph linking nodes x_i and x_j , then the cut cost of a clustering is the sum of the weights of the edges that need to be cut to form the clusters.

The motivation is that $\kappa(x_i, x_j)$ will be low for dissimilar data points and high for similar data points. Hence, minimizing the cut cost will lead to clusters containing similar points, and different clusters containing dissimilar points.

Let y_i be the cluster membership label of point i :

$$y_i = m \Leftrightarrow x_i \in C_m \tag{4.36}$$

for $m \in \{1, \dots, k\}$. The cut cost can then be written as:

$$CK = \sum_{y_i \neq y_j} \kappa(x_i, x_j). \tag{4.37}$$

It is customary to normalize the kernel matrix. This can be done in a number of ways. Following the NJW algorithm, define a diagonal matrix \mathbf{D} , with the elements on the diagonal equal to the sum of the elements on the rows of the kernel matrix:

$$D_{ii} = \sum_j \kappa(x_i, x_j). \quad (4.38)$$

Then define the matrix L as:

$$L = D^{-1/2} K D^{-1/2}. \quad (4.39)$$

The matrix L is sometimes called the Laplacian, a term originating from the graph-theoretical approach. The cut cost for this normalized version of the kernel matrix is:

$$CL = \sum_{y_i \neq y_j} L_{ij}. \quad (4.40)$$

It is shown (Ng *et al.*, 2002) that the following procedure finds a clustering that minimizes the cut cost CL for a given matrix L .

- Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ be the n eigenvalues of L , with v_1, \dots, v_n the corresponding eigenvectors. Retain the k eigenvectors v_1, \dots, v_k corresponding to the k largest eigenvalues, and form the $n \times k$ matrix V by stacking the k eigenvectors in columns.
- Form the matrix W by normalizing the rows of V to have unit length,
$$W_{ij} = V_{ij} / \left(\sum_j V_{ij}^2 \right)^{1/2}.$$
- Consider each row of the $n \times k$ matrix W as a point w_i in a k -dimensional space. Cluster these points into k clusters using the regular k-means algorithm.
- Finally, assign the original data point to cluster m if and only if w_i is assigned to cluster m .

The connection between this approach and kernel k-means is discussed in the literature (Dhillon *et al.*, 2004; 2005). It is shown that both the NJW and kernel k-means algorithms can be written as a so-called *trace maximization* problem. The trace of a matrix expression which includes a matrix with indicator variables must be maximized to find the optimal clusters according to the respective criteria.

The important difference between the NJW and kernel k-means algorithm is in the sorting of the eigenvalues and retaining the k largest ones. If the sorting is not performed, and a random set of k eigenvalues is selected to proceed with, the NJW algorithm reduces to kernel k-means. Since in the NJW algorithm the actual

clustering is performed in a lower dimensional space (of dimension k), NJW is referred to as a spectral relaxation of kernel k-means. Furthermore, it is more stable, as the random selection of eigenvalues corresponds to the random initializations of the cluster centres normally performed in the kernel k-means algorithm.

4.4.7 Clustering scatter nodes using kernel methods

In this section the theory developed in the previous sections is recapitulated with the application of clustering scatter nodes in mind.

Given a data set of scatter nodes x_i with spatial, temporal and feature components $(x_s, x_t, x_v)_i$, the following procedure describes a stepwise method for determining patterns, or clusters, in the data set.

- Decide on appropriate kernels to use for each of the three components separately. For example, a Gaussian kernel can be used for the spatial components, while the linear kernel may be sufficient for the feature components. Let the chosen kernels be κ_s , κ_t and κ_v respectively.
- Combine the three kernels on the subspaces to one kernel on the combined space. Practically, this can be achieved easily by first calculating three kernel matrices:

$$\begin{aligned} K_{ij}^s &= \kappa_s((x_s)_i, (x_s)_j) \\ K_{ij}^t &= \kappa_t((x_t)_i, (x_t)_j) \\ K_{ij}^v &= \kappa_v((x_v)_i, (x_v)_j) \end{aligned} \tag{4.41}$$

and summing them, to give

$$K_{ij} = \frac{1}{3} (K_{ij}^s + K_{ij}^t + K_{ij}^v) = \kappa(x_i, x_j). \tag{4.42}$$

The factor $1/3$ is a normalization factor.

- This combined kernel matrix K is used in the NJW algorithm.
- The result of the algorithm is a clustering, represented by a labelling y of the scatter nodes x_i :

$$y_i = j \Leftrightarrow x_i \in C_j \tag{4.43}$$

for $j = 1, \dots, k$. These labels define k clusters C_j , where each cluster is a subset of scatter nodes.

The resulting clusters are expected to be groups of scatter nodes that are similar. Similarity is a combined similarity in spatial, temporal and feature components of the nodes. Clusters of similar nodes are indicative of the same object, such as a fish school. Examples are given in chapter 5, where a number of case studies are presented.

From expression (4.42) it can be seen that it is straightforward to exclude any of the components if desired. For example, if only features relating to backscatter values are of interest, one can simply set the spatial and temporal kernel matrices to zero, excluding them from the clustering process. This idea is generalised by introducing weights w_s , w_t and w_v as follows:

$$\mathbf{K}_{ij} = \frac{1}{w_s + w_t + w_v} (w_s \mathbf{K}_{ij}^s + w_t \mathbf{K}_{ij}^t + w_v \mathbf{K}_{ij}^v) = \kappa(x_i, x_j). \quad (4.44)$$

Gaussian kernels are widely used. In the case studies presented in chapter 5, two versions of spatio-temporal Gaussian kernels are used, the additive Gaussian kernel:

$$\mathbf{K}_{ij} = \left(\sum_c w_c \right)^{-1} \sum_c w_c e^{-(x_i^{(c)} - x_j^{(c)})^2 / 2\sigma_c^2} \quad (4.45)$$

and the multiplicative Gaussian kernel:

$$\mathbf{K}_{ij} = \left(\prod_c w_c \right)^{-1} \prod_c w_c e^{-(x_i^{(c)} - x_j^{(c)})^2 / 2\sigma_c^2} \quad (4.46)$$

where the sum and product are over all components c of the nodes, spatial, temporal and feature components. Components have weights w_c and parameters σ_c .

The use of such kernels allows for combining the spatial, temporal and other features of scatter nodes into a single kernel matrix for use in kernel clustering algorithms such as NJW.

4.5 UNDERSTANDING SCATTER NODE PATTERNS

4.5.1 Segmentation

Applying the clustering methods presented in sections 4.3 and 4.4 to scatter node data sets results in a labelling l of the nodes, $i = 1, \dots, n$:

$$l(x_i) = m \quad (4.47)$$

with $m \in \{1, \dots, k\}$ for k clusters.

This yields a segmentation of the data into segments or groups of nodes belonging together.

Through expert knowledge it may be possible to assign each segment a known label. For example, one label may correspond to fish of species A while another label may correspond to species B, and yet another to the seabed. When labels can be named, segments are generally referred to as *classes*, and the pattern analysis methods are said to yield a *classification*.

4.5.2 Classification

Classification, that is directly assigning scatter nodes to named classes, can only be achieved after a clustering method has been applied and names are assigned to segments or clusters through expert intervention. In fact, if expert knowledge is not available or is insufficient to assign true class names, one can proceed by using the labels as nominal class names, which is the approach adopted in this section.

Classification in this sense applies to unseen data: scatter nodes that were not used during the clustering process. Having clustered a set of scatter nodes, the clustering or segmentation can be used to classify new data.

Classification with UDBSCAN

Given a set of scatter nodes and their segmentation obtained through the application of the UDBSCAN algorithm, and a new scatter node y , the classification of y consists of an assignment of y to one of the established clusters.

One proceeds by assigning this new node to a cluster according to the cluster assignment rules of the UDBSCAN algorithm, namely steps (2) and (3) of Algorithm 4.2:

- if y is a core point of a cluster it is assigned to that cluster,
- if y is a border point of a single cluster it is assigned to that cluster,
- if y is a border point of multiple clusters (a critical point), the membership function is used to determine to which cluster y gets assigned,
- in all other cases y is assigned to the noise cluster.

Since unseen scatter nodes typically arise from new or repeat surveys, there will only be value in comparing spatial components of past and new scatter nodes for objects or structures that were approximately stationary between surveys. Hence the use of the method described here is less useful for fisheries applications. It can be valuable though for seabed habitat mapping applications.

Since kernel methods make use of more than just the spatial components of the scatter nodes it is recommended that kernel methods be used for the classification of unseen data.

Classification with Ng-Jordan-Weiss

The kernel clustering method of choice is Ng-Jordan-Weiss (NJW). Given a set of scatter nodes that have been segmented using the NJW algorithm, it is now described how new scatter nodes can be classified based on the NJW segmentation.

Since the NJW algorithm is based on the concept of a cost criterion, the cut cost, defined in eq. (4.40), assigning an unseen scatter node y to a cluster is straightforward. One assigns node y to that cluster so that the increase in cut cost is minimal.

This procedure is particularly powerful when segmentation is mostly or exclusively based on the non-spatial and non-temporal features. In that case, features typically relating to backscatter energy levels of the scatter nodes are used to identify coherent clusters. Similarity is based on similarities in backscatter features, in which case it is useful to classify unseen data nodes accordingly.

4.5.3 Visualizing patterns

The patterns that result from the clustering algorithms presented in sections 4.3 and 4.4 are clusters, segments, or groups, of scatter nodes. The primary means of conveying this information is visual: by graphically representing the detected segments.

In addition to the techniques discussed in section 4.2.2, scatter node clusters can be analysed visually or graphically using the following methods.

Isosurfaces are two-dimensional surfaces in a three-dimensional space. They are defined as surfaces at which a particular quantity is constant. In the case of scatter nodes, all the nodes representing a fish school can be used to derive a bounding school volume by means of an isosurface based on the backscatter energy levels of the nodes.

Alternatively, a volumetric object can be determined by triangulating the nodes representing the fish school. Triangulating point sets is common practice in computer graphics (Foley *et al.*, 1995), and is known as Delaunay-triangulation after the author who originally proposed this in the 1930s (Delaunay, 1934). The volume is built up from a series of tetrahedra, where each tetrahedron connects four nodes together and consists of three triangular facets. Additional conditions can be imposed on a shape obtained through this mechanism, imposing some regularization conditions with respect to the smoothness of the volume being generated. Alpha-shapes provide such a method (Edelsbrunner and Mücke, 1994). When the cluster of nodes representing the seabed is considered, a two-dimensional surface can be derived in the same way. Approaches to triangulating classical bathymetric multibeam soundings are available in the literature (Canepa *et al.*, 1999; Brouns *et al.*, 2003; Canepa *et al.*, 2003). Gridding is another option, where a regular grid is defined and values for the grid nodes are derived from values of the irregularly spaced scatter nodes or soundings (Calder and Mayer, 2001; Calder, 2003; Paton *et al.*, 2003).

Projecting scatter nodes, or two or three-dimensional objects derived from them, onto a two-dimensional plane can be useful in creating two-dimensional visualizations. In particular, projecting onto a horizontal plane is effective to show an areal view of the data.

4.5.4 Measuring patterns

Apart from the information that is conveyed visually by graphical representations of the data, calculating quantitative numerical metrics of the detected segments or derived objects is useful. Two types of metrics can be distinguished: geometrical or morphological metrics, and energy-based measures. The former relate to the spatial aspects of the clusters of nodes, the latter to the energy content or energy related features of the nodes. Many metrics can be calculated in the Echoview software (Myriax, 2008). The most prominent and relevant ones are listed in Table 4.3. Some are only relevant to volumetric objects such as fish schools, others only to two-dimensional surfaces such as the seabed.

	<i>Volume (school)</i>	<i>Surface (seabed)</i>
<i>Energetic</i>		
Mean backscatter level	+	+
Variance of levels	+	+
Min. and max. level	+	+
Sum of levels	+	-
<i>Geometric</i>		
Number of nodes	+	+
Volume	+	-
Area	-	+
Size (length, width)	+	+
Size (height)	+	-
Min. and max. depth	+	+
Georeferenced position	+	+

Table 4.3 Metrics that can be calculated from scatter node segments or objects derived thereof. Some metrics are only relevant to volumetric objects, others only to two-dimensional surfaces. Where a metric is relevant this is indicated using a + symbol, where not a – sign is used.

When features from other data sources are available, they can be used to derive additional metrics. For example, if water temperatures are available for each node it is clear that statistical measures of the temperatures of all nodes in a particular cluster may convey valuable information.

These metrics provide essential information to scientists. It is important to note that the metrics are obtained from the scatter nodes only, without making use of the original multibeam measurements.

4.5.5 Assessing pattern quality

Two clustering methods were discussed in previous sections, UDBSCAN in section 4.3, and kernel clustering methods in section 4.4. Both algorithms allow for some adjustment by setting parameters. In UDBSCAN, the parameters ϵ , the size of the neighbourhood, and m , the minimum number of points in the ϵ - neighbourhood, must be set prior to running the algorithm. With kernel methods, the kernel functions to be used must be chosen, as well as any free parameters in the functions.

In this section some approaches are given to assist in selecting the best set of parameters or function classes, and to determine the effectiveness of various selections chosen.

Finding a suitable clustering is a pattern analysis problem. A trained clustering algorithm is often referred to as a model, and the process of validation and evaluation is known as *model validation* and *model evaluation*.

Expert assessment

When using multibeam sonar data and the derived scatter nodes, it is often the case that the expert human eye perceives coherent sets of nodes as clusters very clearly. Such clusters can be indicative of aggregations of fish, or of the seabed. Acousticians or marine biologists familiar with the instruments and their use will have firm opinions on which patterns should be found. Presenting the outcomes of automated algorithms for inspection to experts is a common and sensible approach, though it can be argued that such a validation is not objective.

Visual inspection can reveal some obvious errors in the clustering results, even to the less experienced scientist. Examples include Figure 4.7, Figure 4.13, and Figure 4.14. In these examples, coherent structures are standing out as obvious regions of aggregated scatter nodes, yet the clustering algorithm applied is not capable of identifying those high density regions as clusters.

Analytical measures

Analytical measures of clustering quality attempt to quantify the quality of the clustering according to some criterion. In some algorithms such criteria arise naturally. For example, the objective functions used in the k-means algorithms are direct measures of quality. Expressions (4.2) and (4.31) measure the proximity of each point to its cluster centre in an Euclidean and kernel-induced feature space respectively. These criteria will yield lower values with increasing values of k , the number of clusters, as the sum of the distances to the cluster centres will decrease

with the number of centres available. Therefore, using the objective functions as a criterion is only valid for a given value of k , not to establish the optimal value for k .

Other clustering methods, such as the spectral methods, can make use of criteria that do not require a cluster centre or template. The cut costs defined in expression (4.40) is an example. In this case, the cut cost can be expected to increase with increasing values of k and so again it cannot be used to establish an optimal value for k .

A probabilistic approach is possible where the model or trained clustering algorithm corresponds to an estimated probability density function. This is appropriate for the Gaussian mixtures model. This approach can be followed for regular k-means clustering as well. The cluster centres that are estimated can be regarded as the means of unit-variance Gaussians. In that case the likelihood can be used as a measure of quality:

$$P(\{x_1, \dots, x_n\} | \mu_1, \dots, \mu_k). \quad (4.48)$$

The likelihood is the posterior probability of the data: it is the probability of obtaining the data given the model. To avoid over-fitting, it is customary to include a term involving the complexity of the model (Hastie *et al.*, 2001). Since this approach cannot be applied in a straightforward manner to the main clustering algorithms considered in this thesis, UDBSCAN and NJW, this route is not explored any further.

Cross validation and predictive accuracy

Cross validation is the most versatile approach, and is generally the preferred method for assessing pattern quality. The general idea of cross validation is to set some of the data apart, run the clustering algorithm on the remaining data samples, and assess how well the data that were set apart are clustered by the trained algorithm. Cross validation measures the predictive accuracy of the trained clustering algorithm. In assessing the predictive capabilities of the trained algorithm, one can use the objective functions for k-means, and the cut cost for spectral methods. The difference with the approach outlined in the previous section is that now these criteria are calculated on data items that were not used during the clustering process. In fact, these criteria are used twice:

- first, to cluster part of the data. The criteria are used to make sure that the algorithm performs well on this data set.

- second, to assess the clustering of unseen data. The criteria are used to assess how well the clustering generalizes to unseen data. This prevents over-fitting.

Various scenarios are possible in conducting cross validation (Hastie *et al.*, 2001). A popular approach known as the *jack-knife* or *leave-one-out* approach repeats the clustering of n scatter nodes n times, each time leaving one scatter node out so that every node is left out precisely once. Each time the cost criterion is calculated and the n obtained values are averaged. The resulting single value can be used to compare results for different values of k , the number of clusters.

A variation on this procedure is *bootstrapping*. In bootstrapping, a subset of m randomly selected nodes is left out each time during a series of runs of the clustering algorithm. The cost criteria are calculated each time on the m data nodes that were left out. It is generally unclear what the best choice of m is for a given data set size n , or how many times the clustering should be repeated. For these reasons the jack-knife approach is recommended.

4.6 OUTCOMES

The pattern analysis phase and the interpretation of its results complete the scientific data mining process (Figure 4.16).

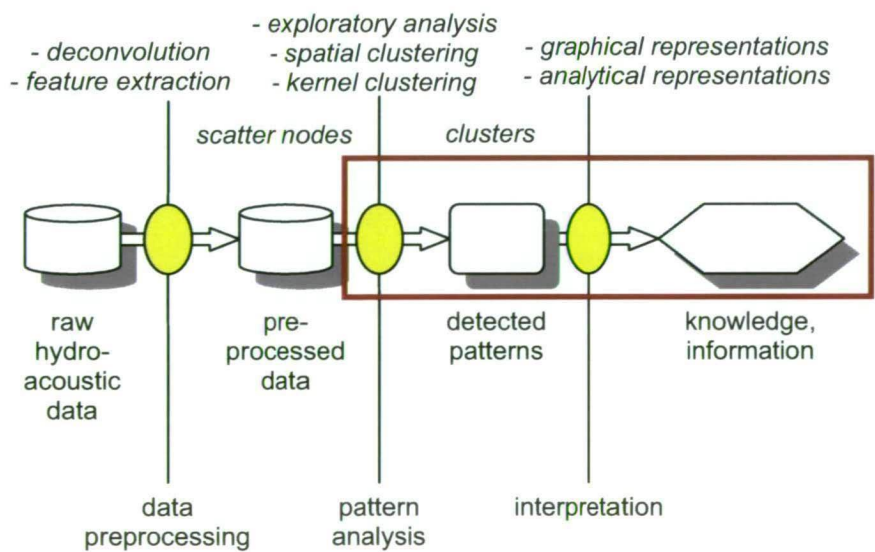


Figure 4.16 Pattern analysis and the interpretation of detected patterns concludes the data mining process.

Three approaches to analysing scatter nodes are proposed, with the aim of identifying coherent clusters. The first two, exploratory data analysis and spatial clustering, apply existing techniques and methods to scatter node data sets. A new, modified version of the spatial clustering algorithm DBSCAN is proposed, UDBSCAN, to guarantee uniqueness of the resulting segmentation.

The third approach, using kernel clustering methods, has been made possible by the development of a mathematical framework to establish kernels for spatio-temporal data. The Hahn-Banach theorem is used to extend kernels to the combined space of spatial, temporal and feature components of the scatter nodes. The formalism developed in this thesis allows for the application of kernel methods in such a manner that all components of scatter nodes are considered simultaneously. The Ng-Jordan-Weiss (NJW) clustering algorithm is identified as a suitable kernel clustering algorithm.

The pattern analysis methods of choice, UDBSCAN and NJW, result in a segmentation of the scatter nodes into coherent segments or clusters. Scatter nodes that are assigned to the same cluster have similar spatial properties in the case of DBSCAN, and similar combined spatial, temporal and other features in the case of NJW.

Clusters of scatter nodes are useful entities in interpreting hydroacoustic data. Since they contain similar scatter nodes, the constituting nodes are likely to be indicative of a larger scale structure, object or class. Clusters can be represented visually, either directly based on the scatter nodes, or otherwise by means of derived spatial objects. The features of the nodes are useful in calculating metrics of the clusters.

Case studies demonstrating the developed data mining process are presented in the next chapter.

5 CASE STUDIES

In this chapter a number of data sets are analysed using the methodology developed for deriving scatter nodes from the raw data set, and using clustering methods to identify patterns. A range of data sets collected by different sonar instruments is studied. The examples presented in this chapter illustrate the methods and serve as a guide to applying these methods in practice.

5.1 MODELED DATA

5.1.1 Description of the data set

A data set obtained from the model presented in chapter 3 was used a number of times in this research. This data set is described in detail below.

The point set used as input to the model, in order to produce this data set, consists of a flat seabed and two ellipsoidal fish schools, and contains 31,688 points. This point set is presented graphically in Figure 3.10 (a). The seabed depth is 100 m. The centre of the larger fish school is at a depth of 50 m. Its longer horizontal axis is 120 m and its shorter horizontal axis is 60 m. The height of the school is 60 m. The smaller school has horizontal axes of 40 and 50 m respectively, and is 16 m high.

The survey transect line is straight and 165 m in length. Along this transect, data for 35 pings are collected at equally spaced intervals. The sound speed is assumed fixed at 1500 m/s.

The modeled transducer array is linear and consists of 80 equally spaced elements, with a total array length of 0.3 m. The operating frequency is 200 kHz, with a pulse duration of 0.2 ms. The modeled instrument is set to collect data at ranges of up to 150 m, with 800 samples per element. The data are beamformed into 128 beams covering a 120 degree sector.

Since the absorption coefficient is assumed to be known exactly and has been corrected for, its numeric value is irrelevant and is set to unity for ease of computation. The reference pressure level is set to unity also as the ratio of the received pressure to the transmitted pressure is the quantity of interest. The backscattering cross section of all points in the point set is chosen equal, with a value of 0.01 m^2 .

The model is run, creating the corresponding synthetic data set. By design, the data set contains beamformed data for 35 pings, each consisting of 128 beams of 800 samples. One ping of data is presented graphically in Figure 3.3. Ignoring phase and storing amplitudes only results in a total of 358,400 backscatter sample values.

5.1.2 Analysis

Scatter nodes are obtained by the application of a deconvolution to the raw sample data. In this case, with a known model, a PSF can be established for use in the Lucy-Richardson deconvolution algorithm (section 3.3.2). In order to establish the PSF of the modeled system, a point set consisting of a single scatterer is used as input, and the corresponding output is calculated and used to establish the PSF. The obtained PSF is presented graphically in Figure 3.6 (c). This PSF is used in the deconvolution, which results in the scatter nodes presented in Figure 3.9 and Figure 3.10 (b).

The total number of scatter nodes is 5,026, while the input model consisted of 31,688 point scatterers. Indeed, observe the sparser density of scatter nodes compared to the initial point set in Figure 3.10 (a). This is due to the resolution of the system, and is captured in the method by the deconvolution, and the low-pass filtering effect it has.

The proportion of scatter nodes to raw data samples is 1:71. In other words, the size of the set of scatter nodes is only 1.4% of the size of the set of raw samples in this data set, if only a single feature is stored as is done in this example. The point amplitudes are the only features used.

Exploratory data analysis of the raw data in conjunction with the derived scatter nodes is instructive. Screenshots of such explorations are given in sections 4.2.3 and 4.2.4.

The full data set as described here was subjected to UDBSCAN with two choices of parameter settings (Figure 4.13 and Figure 4.14). UDBSCAN is not successful in separating the smaller school from the seabed. However, UDBSCAN is useful in the detection of noise scatter nodes. Nodes that are not spatially dense in a cluster of nodes are assigned to the noise cluster. This cluster is retained, and the remaining nodes are further analysed using kernel methods. The black points in Figure 5.1 are the nodes of the noise cluster.

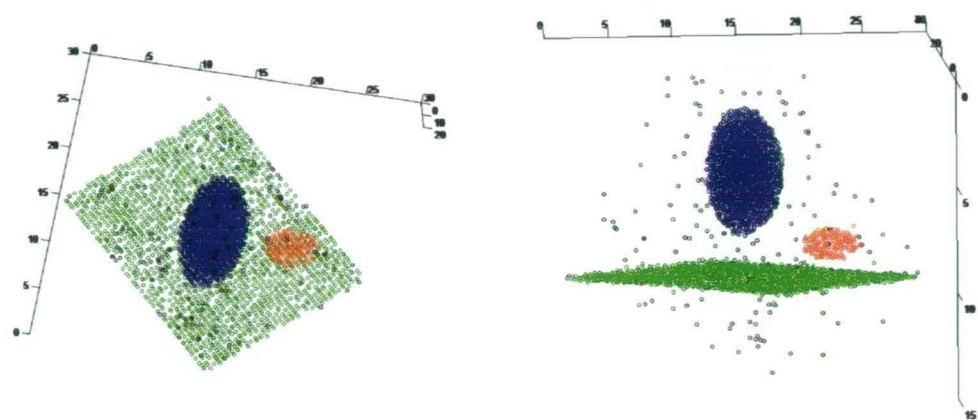


Figure 5.1 Scatter nodes clustered using the NJW kernel clustering method. The black nodes are the ones that were identified as noise by the UDBSCAN algorithm. Three clusters are identified: the seabed (green), the large fish school (blue) and the small one (red).

The NJW kernel clustering algorithm is applied to the set of scatter nodes that are not assigned to the noise cluster by UDBSCAN. In addition to the spatial components of the nodes, one feature is used: the backscatter amplitude of the nodes; all components are weighted equally. The additive Gaussian spatio-temporal kernel is used, eq. (4.45), with parameters as given in Table 5.1. The NJW clustering algorithm is run to detect three clusters (Figure 5.1): a large school (blue), a small school (red), and the seabed (green).

Components	σ	weight
X (longitude)	0.8	1.0
Y (latitude)	0.8	1.0
Z (depth)	0.05	1.0
Backscatter amplitude	0.5	1.0

Table 5.1 Parameters used in the calculation of the kernel matrix for the modeled data set.

5.1.3 Results

The clusters that are found are indeed representative of the three objects that are present in the data: the seabed and two fish schools. This is shown in Figure 5.1. Where UDBSCAN is not able to differentiate between the smaller school and the seabed, the kernel clustering method is. This example illustrates the power of the kernel clustering method over the spatial clustering algorithm UDBSCAN.

5.2 SALMON BANKS

5.2.1 Description of the data set

A data set was collected off San Juan Island, Washington, using a Simrad Kongsberg Mesotech SM20 sonar (data set courtesy of John Horne, School of Fisheries, University of Washington, Seattle, WA, USA). The survey consisted of ten transects covering about 50% of a 3 square nautical mile area known as *Salmon Banks*. The data set and a first analysis of it is discussed in Buelens et al. (2007). The SM20 operates at a frequency of 200 kHz, and covers a 120 degree swath with 128 beams.

For the purpose of the present example a section of data from the third transect is selected. Coincidentally, in this transect, a fish school is observed near the typical angular bottom side lobing effects present in the backscatter data. This example will show how the kernel clustering algorithm is capable of differentiating between the echoes from this aggregation of fish and the artefacts.

The raw multibeam data are scrutinized in Echoview. The school of interest in this example is seen on two-dimensional echograms of multibeam ping data (Figure 5.2). The school is at a depth of approximately 100 meters. The bottom sidelobing artefacts happen to occur in the vicinity of this school.

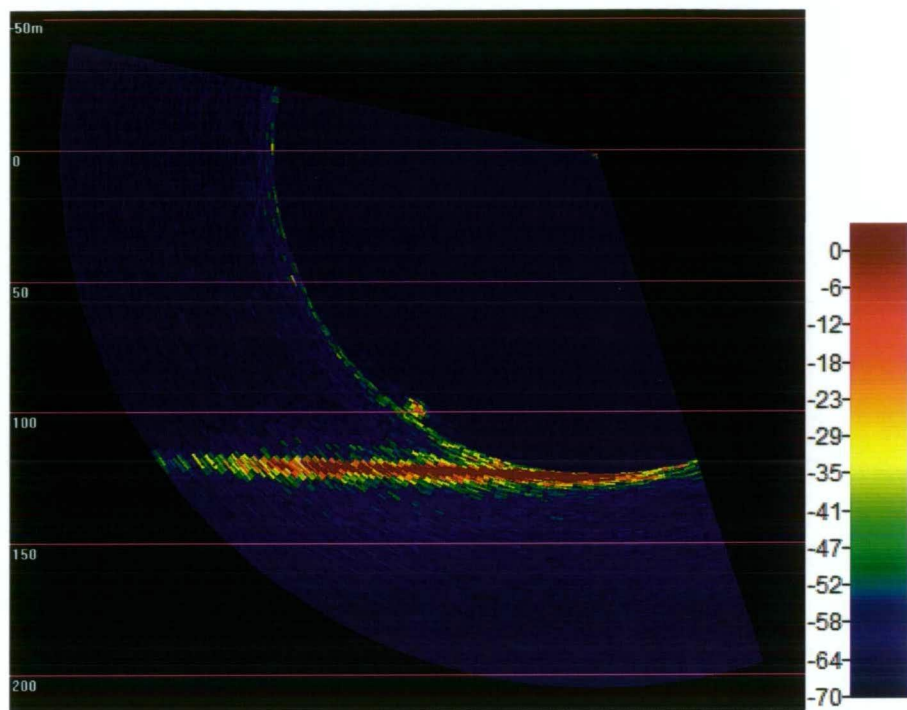


Figure 5.2 One ping from the Salmon Banks data set (S_v in dB). A school is present at a depth of about 100 meters, very close to the angular bottom sidelobe artefacts.

5.2.2 Analysis

The selected section of data consists of 150 pings, each having 798 backscatter data samples in each of the 128 beams. These 15,321,600 data samples are transformed into a set of 22,341 scatter nodes using the blind deconvolution technique as no calibration data are available. Seven features are associated with each scatter node (Table 5.2). The total number of features for all nodes together is only just over 1% of the total number of raw backscatter samples, a massive reduction. Scatter nodes are visualized in two dimensions in Figure 5.3 and in three dimensions in Figure 5.4.

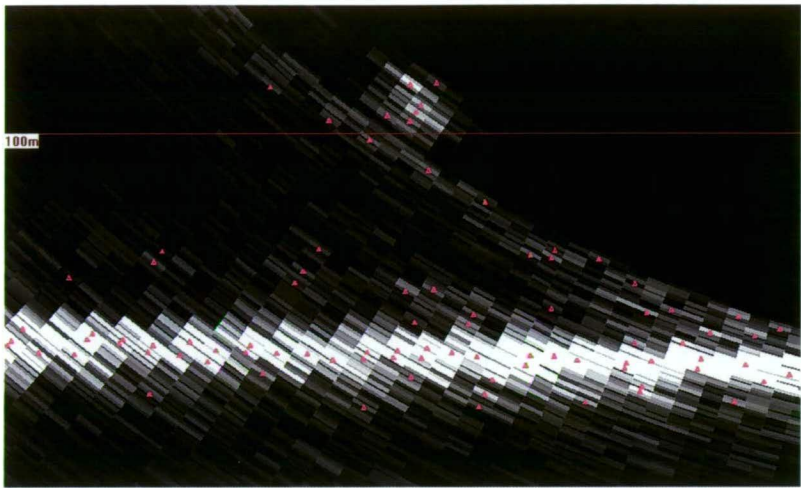


Figure 5.3 Close-up of the relevant part of the same ping, now with scatter nodes added (plotted as small pink triangles). The colour scheme for the data is set to gray scale so the scatter nodes are more clearly visible.

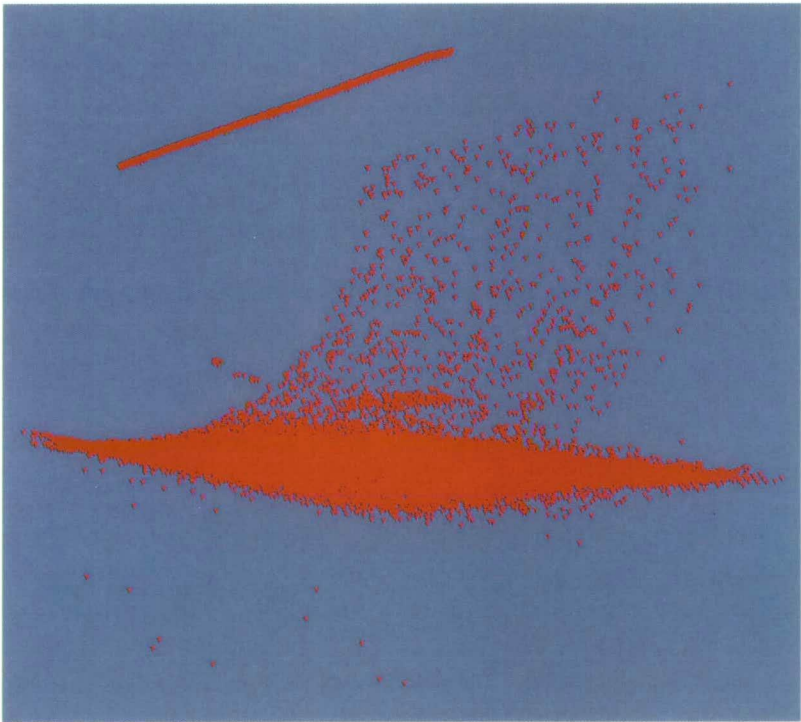


Figure 5.4 The scatter nodes represented in three dimensions.

From visual inspection it is clear that there are some distinct structures in this data set: the seabed, the sidelobing artefacts, the fish school, and noise. The aim is now to cluster the scatter nodes into groups, each representing one of these objects or structures in the data.

First the UDBSCAN algorithm is run. The parameter ε is set manually. As recommended in Ester et al. (1996), the value for m is set to 8. Large values of ε lead to few clusters being found while small values lead to more clusters. This example appears to be difficult for the UDBSCAN algorithm, as no good value for ε could be determined. Shrinking ε so that more clusters are being detected does not result in the school being identified as a cluster. For example, with a value of $\varepsilon = 10.0$, UDBSCAN identifies 14 clusters (Figure 5.5). However, none of the clusters bear any significance, other than the cluster identifying the noise band at short range, in purple in Figure 5.5. This short range noise is most likely caused by transducer ringdown or air bubbles under the vessel. Lowering the value of ε further will ultimately result in the school being identified, however, too many spurious clusters exist by then. For example for $\varepsilon = 4.0$, UDBSCAN identifies 46 clusters, which is no longer informative.

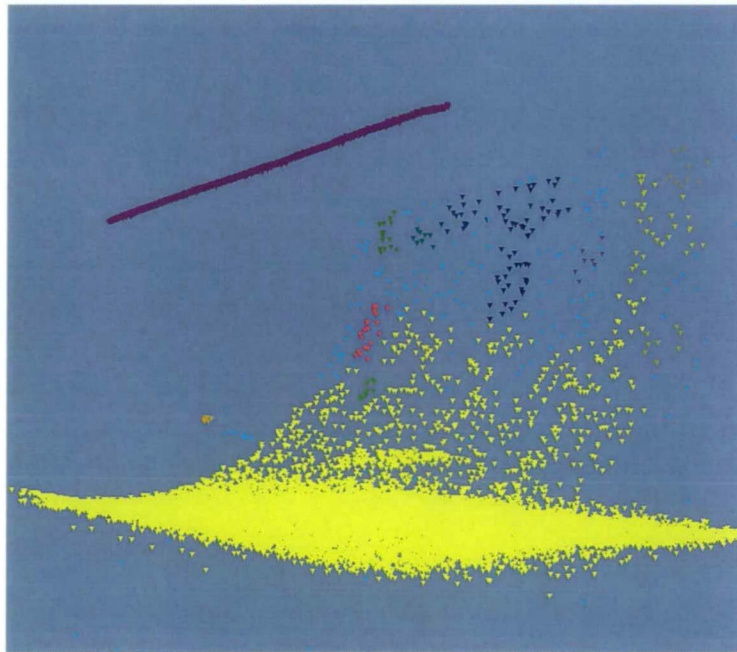


Figure 5.5 The scatter nodes clustered into 14 clusters using UDBSCAN. The denser aggregation of nodes indicative of a fish school, is not isolated as a cluster.

Next, the kernel clustering algorithm NJW is run. Since it requires the calculation and handling of an $n \times n$ matrix, with n the number of scatter nodes, running the algorithm can become slow for large n , in particular if no specific attention is paid to efficient implementations. In this case $n = 22,341$. Running NJW on 10% of that is very fast (in the order of tens of seconds), so the following approach is taken: first, a randomly selected subset of 10% of the N scatter nodes is subjected to the NJW algorithm; then, the remaining 90% of nodes are labelled using the nearest-neighbour criterion. In this first run of the algorithm the aim is to separate the

scatter nodes in the water column from the ones indicative of the seabed. The latter group is much larger, as can be seen in Figure 5.4.

The multiplicative Gaussian kernel is used, eq. (4.46). The temporal component is not used. A set of parameter values is established experimentally. The values that are used are presented in Table 5.2. Weights are either 1 or 0, indicating whether that component is used (1) or not (0). The parameter σ is the parameter of the Gaussian kernel. The non-spatial components were normalized prior to running the clustering algorithm. The results of the first run are presented in Figure 5.6. A number of different clusters cover the seabed scatter nodes, while the nodes in the water column representing the aggregation of fish as well as the angular noise pattern are isolated as a single cluster, drawn in green in Figure 5.6.

<i>Components</i>	<i>σ run 1</i>	<i>weight run 1</i>	<i>σ run 2</i>	<i>weight run 2</i>
Longitude (X)	-	0	50.0	1
Latitude (Y)	-	0	50.0	1
Depth (Z)	10.0	1	15.0	1
Point backscatter	-	0	2.0	1
Mean backscatter	0.5	1	2.0	1
Median	-	0	2.0	1
Standard deviation	-	0	1.0	1
Skewness	-	0	1.0	1
Kurtosis	0.8	1	1.0	1
Number of samples	0.8	1	2.0	1

Table 5.2 Parameters used in applying the NJW algorithm to the scatter nodes from the Salmon Banks data set.

Next, the scatter nodes of this cluster are considered separately, and subjected again to the NJW algorithm. The parameters used in this second run are listed in Table 5.2. The result is shown in Figure 5.7. Two clusters are detected, one is clearly indicative of the aggregation of fish while the other represents the angular noise pattern caused by sidelobing effects from the seabed.

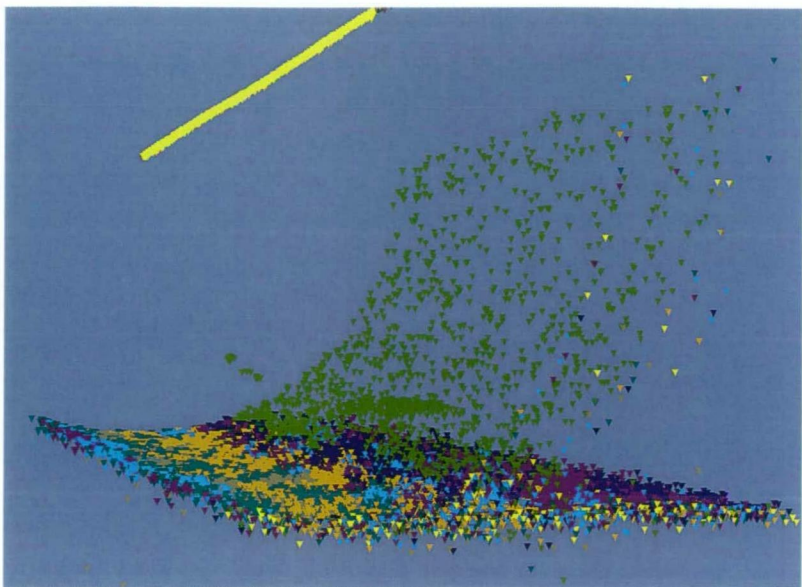


Figure 5.6 The water-column scatter nodes are isolated in the green cluster. This is the result of the first run of the NJW algorithm.

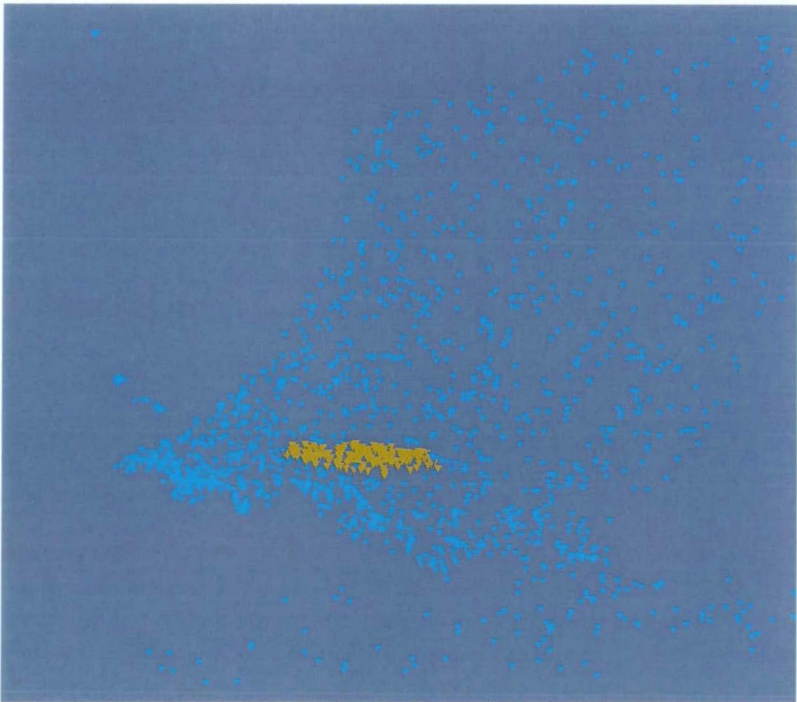


Figure 5.7 The water-column nodes identified in the first run are now clustered in a second run of NJW. The nodes indicative of the aggregation of fish are separated from the nodes resulting from the angular seabed sidelobing noise pattern.

It is interesting to note how UDBSCAN and NJW detect the short range noise band differently. UDBSCAN isolates it perfectly from the other scatter nodes, because it is spatially well separated (the purple cluster in Figure 5.5). NJW on the other hand does not identify it perfectly (the yellow cluster in Figure 5.6). This is because non-spatial components come into play as well, and there are indeed non-spatial similarities between scatter nodes in the short range noise band and other nodes. This observation suggests that there is value in utilizing the UDBSCAN results for what they are useful: identifying spatially isolated clusters of samples. Under such a regime, the short range noise nodes would be removed from the set of scatter nodes prior to running the kernel algorithm NJW.

5.2.3 Results

The fish school represented by the cluster plotted in orange in Figure 5.7 is studied in further detail. Marine biologists familiar with the particular area where the data were collected have identified the school as Pacific Herring (*Clupea pallasii*). The scatter nodes of the school are used to render a volume, as described in section 4.5.3. This school volume is shown graphically together with the original multibeam data in Figure 5.8.

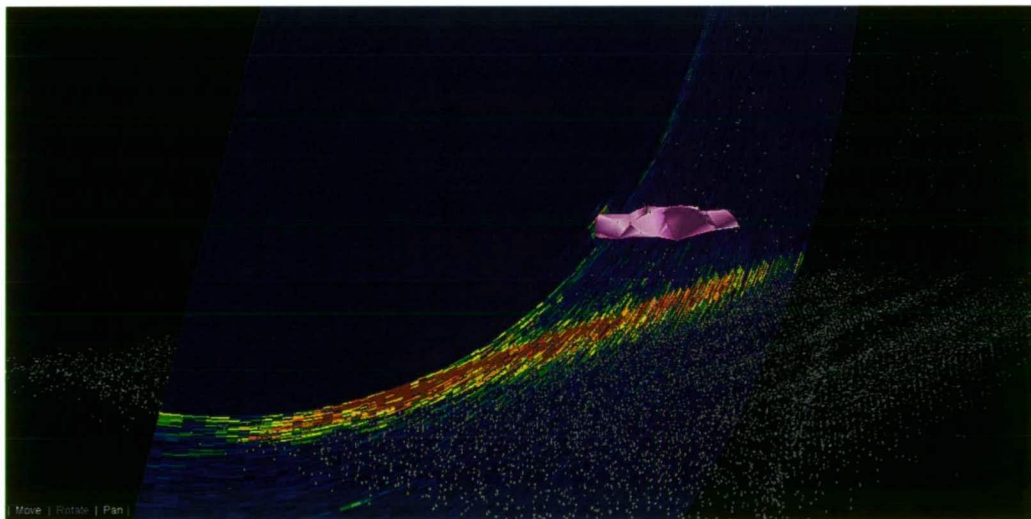


Figure 5.8 In pink, the school of Pacific Herring as detected by the NJW clustering algorithm, shown together with one ping of raw multibeam data and all scatter nodes from the whole data set (in gray).

For comparison, the same data was used for schools detection in the Echoview software. The schools detection algorithm in Echoview is based on thresholding of raw multibeam samples (Myriax, 2008). The school determined in this way is plotted in yellow in Figure 5.9, together with the earlier determined school volume

in pink. It is seen from this figure that the Echoview school in this case is more fractal in nature, which is caused by the fact that it is built up from volume elements corresponding to individual multibeam samples, while the scatter node cluster is smoother.

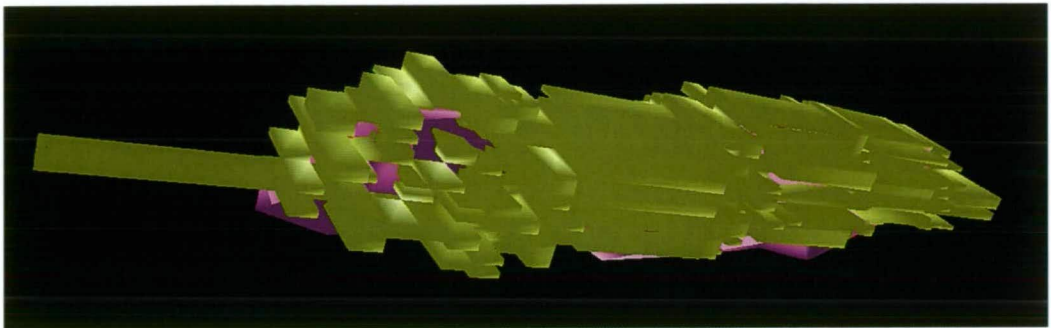


Figure 5.9 The pink school as in the previous figure, together with the corresponding school as determined by the Echoview sample thresholding based detection algorithm.

An analytical comparison is given in Table 5.3. It is difficult to tune settings for both algorithms so that comparable school objects are detected. The school object detected by sample thresholding is slightly bigger, which is reflected in the length and depth measures. The fact that the scatter node based school is a lot smoother is observed in the complexity of the describing volume, in particular the number of vertices and triangles making up the volume shell. The energetic properties, however, are very similar for both school objects. Since the sample thresholding based school contains more raw samples, it must be concluded that these additional samples are of low backscatter values and in reality may not be an integral part of the school.

	<i>Scatter node cluster school</i>	<i>Sample thresholding school</i>
Number of vertices	156	2741
Number of triangles	2460	6120
Length (North-South)	13.8 m	14.4 m
Length (East-West)	40.5 m	54.4 m
Depth minimum	93.8 m	92.0 m
Depth maximum	101.5 m	104.1 m
Centre depth	99.0 m	97.3 m
Sample mean	-39.7 dB	-40.2 dB
Sample minimum	-52.7 dB	-52.7 dB
Sample maximum	-23.0 dB	-23.2 dB
Number of raw samples	700	2271

Table 5.3 Comparison of analytical measures of schools detected by clustering scatter nodes versus by sample thresholding

Finally, the scatter nodes representing the school being considered are used to obtain a mean sample value directly, without using the raw multibeam data. This is possible because the mean of all samples contributing to each node were used to construct a feature of the node.

Calculating the mean of the relevant feature and expressing the obtained value logarithmically gives a value of -39.5 dB. This is very close to -39.7 dB, the value obtained using all the raw sample data (Table 5.3).

This case study demonstrates that the nodes, their features, and the volumetric objects derived from them may be sufficient for fisheries applications. Once the nodes and their features are derived, there is no immediate need to reconsider the raw multibeam data again, other than for comparative analyses.

5.3 LAKE OPEONGO

5.3.1 Description of the data set

During a joint project conducted by Myriax Pty Ltd, Kongsberg Mesotech and the Ontario Ministry of Natural Resources and Scientific Assessment Technology Laboratory in 1999, a Kongsberg Mesotech SM2000 multibeam sonar was deployed on a vessel on Lake Opeongo, Canada. The sonar operational frequency was 200 kHz, collecting 128 beams of data over a 120 degree swath.

The data file that is used in this case study contains recordings of a school of *Coregonus artedii* (Lake herring or Cisco). It is the same data file that is shown in plate 3.5 of Simmonds and MacLennan (2005). One ping of data in which the school of fish is clearly visible is shown in Figure 5.10.

5.3.2 Analysis

As no calibration data are available, blind deconvolution is used to derive scatter nodes. The original data consist of 91 pings of 128 beams of 245 samples each, a total of 2,853,760 samples. The point backscatter sample values and the mean of the samples nearest to each node were attributed to the scatter nodes as features. With two features per node and only 5,146 nodes in total, the original data set is reduced in volume to under 0.4% of its original size. The scatter nodes derived from the ping of data shown in Figure 5.10 are shown in Figure 5.11. In Figure 5.12 all nodes for the whole data set are shown.

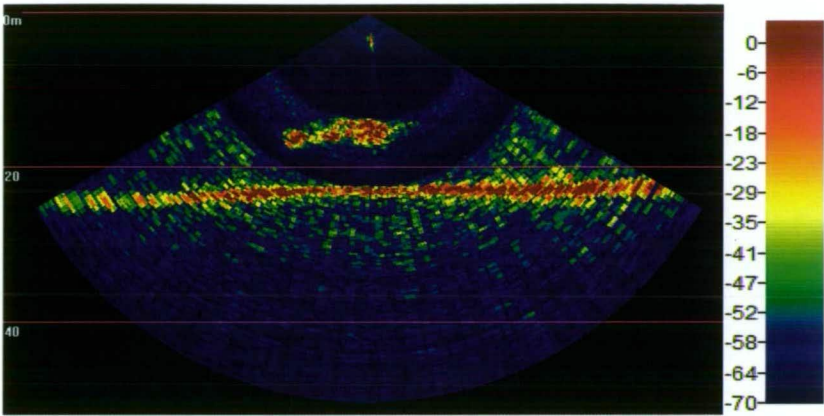


Figure 5.10 One ping of data from the Lake Opeongo data set (S_V in dB). The school of *Coregonus artedii* (Lake herring) is clearly visible above the 20m depth line, with the seabed beneath the line.

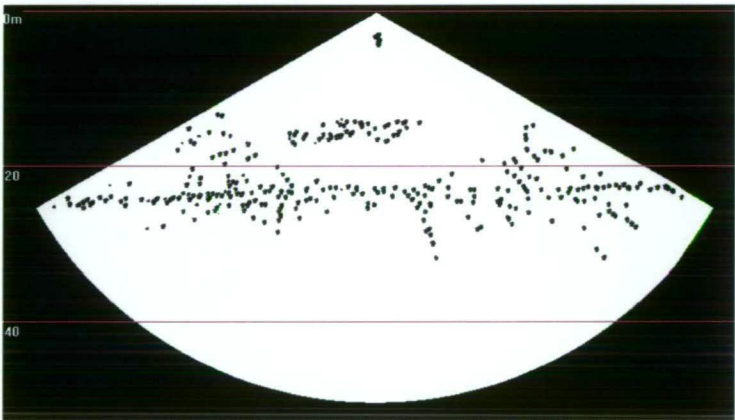


Figure 5.11 The scatter nodes derived from the ping of data shown in Figure 5.10.

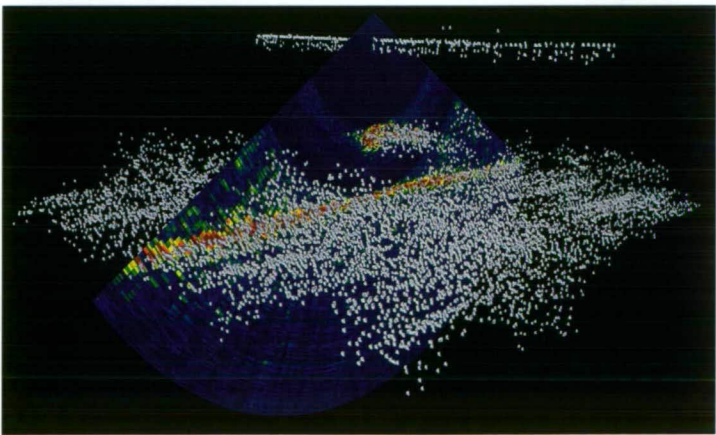


Figure 5.12 A three-dimensional view of the ping shown in Figure 5.10 together with the scatter nodes derived for the whole data set.

It is seen from Figure 5.11 and Figure 5.12 that there are many scatter nodes in dense regions close to the seabed which arise from noise rather than from true

seabed echoes. Therefore UDBSCAN is unlikely to be successful in this case since it only considers spatial aspects of the scatter nodes. The kernel clustering algorithm NJW is applied, using a multiplicative Gaussian kernel, eq. (4.46), with the settings listed in Table 5.4.

Components	σ	weight
X (longitude)	10.0	1.0
Y (latitude)	10.0	1.0
Z (depth)	4.0	1.0
Point backscatter amplitude	1.0	1.0
Mean backscatter amplitude	2.0	1.0

Table 5.4 Parameters used in the calculation of the kernel matrix for the Lake Opeongo data set.

Intuitively one would expect four clusters to be found: the fish school, the seabed, the seabed noise, and the nodes at the top, close to the transducer. The latter are due to the presence of another instrument that was mounted in the acoustic beam of the sonar. Experimental runs of the NJW algorithm were found to optimally detect the school at a value of $k = 7$. Lower values of k resulted in the school being assigned to the same cluster as parts of the seabed and the seabed noise, while higher values caused the school to be partitioned in several clusters. Of the seven clusters, four were in fact due to the bottom noise and were afterwards manually merged to a single clusters resulting in a total of four clusters. These clusters are shown graphically in Figure 5.13.



Figure 5.13 The scatter nodes from the Lake Opeongo data set, segmented into 4 clusters. The colors indicate cluster membership; the size of the spheres is representative of the mean backscatter energy level of the original samples contributing to each node.

In Figure 5.13, colour indicates cluster membership: green is the seabed, orange the seabed noise, blue the noise due to the other instrument, and purple is the fish school. The size of the nodes is representative of the mean backscatter energy levels

of the original samples contributing to that node, with bigger nodes representing higher backscatter energy levels.

5.3.3 Results

The nodes indicative of the fish school are now triangulated to form a volumetric object. Properties of this object can be calculated: the school is 32.4m long, 17.0m wide and 4.9m high. The depth of its centre is 13.1m. The nodes indicative of the seabed are triangulated to form a surface. The seabed surface area is 8,414m², ranging in depth from 22.0m to 24.9m. A graphical representation of the volumetric school object and of the surface is presented in Figure 5.14.

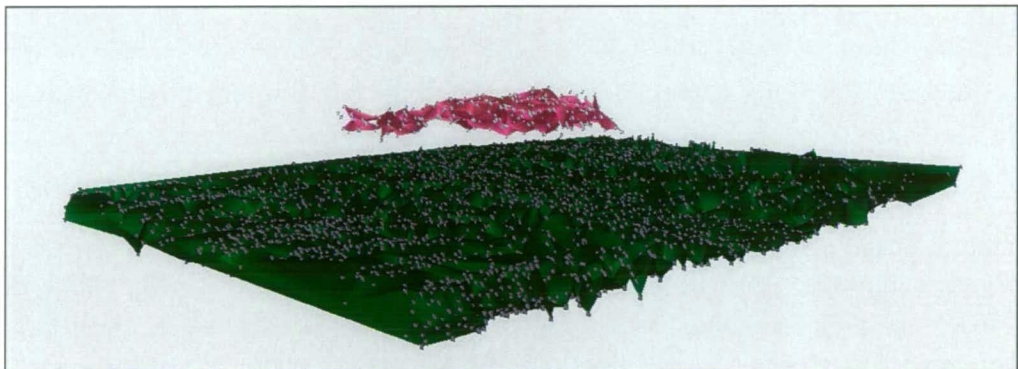


Figure 5.14 The scatter nodes of the school are constructed to build a volumetric object (in purple) and the seabed nodes are used to build a surface (in green). All nodes are drawn in gray.

5.4 SOUTHERN OCEAN

5.4.1 Description of the data set

The preprocessing and clustering methods that have been developed in this research are designed primarily for analysing multibeam sonar data. However, they are applicable to single beam echosounder data too, as is demonstrated in this case study. For standardization purposes it is valuable to have a unified approach that can be applied to both multibeam and single beam sonar data sets. Furthermore, it will facilitate the combined usage of both kinds of data, for joint analyses.

The data set was collected in 2004 by the Australian Antarctic Division (AAD) in the vicinity of Heard Island and the McDonald Islands, a subantarctic island group in the Southern Ocean, about 4,000 km south west of Australia (data set courtesy of Toby Jarvis, AAD, Kingston, Tasmania, Australia). A Simrad EK60 single beam echosounder was used with three transducers, to collect data at three acoustic frequencies: 38kHz, 120kHz and 200kHz. Data were collected for 55 minutes along a transect approximately 10 nautical miles long.

The AAD have established their own data processing standard (Jarvis, 2006). Data cleaning components of this data processing routine include the removal of the bottom echo, noise spikes and background noise. The cleaned data are then down-sampled to a lower resolution, and classified using the acoustic responses at multiple frequencies (Korneliussen and Ona, 2003). Four classes are distinguished: resonant scatterers, fluid-like scatterers and (more specifically) large and small fluid-like scatterers.

The method of deriving scatter nodes and clustering them is applied to this data set, and the results compared with those obtained by the AAD.

The segmentation of the water column into different classes has attracted some attention in the past. Recently, Anderson *et al.* (2007) used a simplified version of Gaussian mixtures to classify multifrequency echosounder data. They did not include spatial information or correlations between the responses at different frequencies. Kieser *et al.* (2006) classified single frequency single beam echosounder data using texture features of the echograms. The clustering algorithm used is k-means. An earlier attempt investigated the use of artificial neural networks (Haralabous and Georgakarakos, 1996). Along the lines of the AAD method is the method based on differences between backscatter levels at different frequencies as discussed in Kang *et al.* (2006). A systematic comparison of these various methods would be interesting and useful but is beyond the scope of the present research.

5.4.2 Analysis

The method, introduced in this thesis, to derive scatter nodes through the application of a deconvolution can be altered to correspond more closely to the customary method of reducing data in single beam echosounder applications. Rather than deconvolving the data, raw data samples are aggregated to form a lower resolution data set. Each of the lower resolution samples can be seen as a scatter node (section 3.4.4). In the AAD scheme, this down-sampling is conducted because it reduces the variance of sample-to-sample backscatter comparisons (Jarvis, 2006). The lower resolution data bins are 2 meters high and contain samples from 25 pings. Following this approach, a scatter node is defined for each lower resolution sample. Approximately 250 raw data samples are reduced to a single scatter node.

The major difference with scatter nodes obtained from a deconvolution is that the scatter nodes obtained through down-sampling are positioned regularly in space and time also in locations where no significant backscatter energy is present. Scatter nodes resulting from a deconvolution are distributed irregularly, and only at locations where some backscatter energy above the noise level was received.

Since the AAD classification scheme is based on differences between the acoustic backscatter energy levels at three frequencies, these levels are defined as features for the scatter nodes. The levels for the lower resolution samples, and hence for the scatter nodes, are obtained through averaging the contributing raw samples. In that way, each scatter node gets three features. In Figure 5.15, full resolution raw data samples are shown together with the corresponding lower resolution representation; the backscatter intensities in this image are those of the responses at 120 kHz.

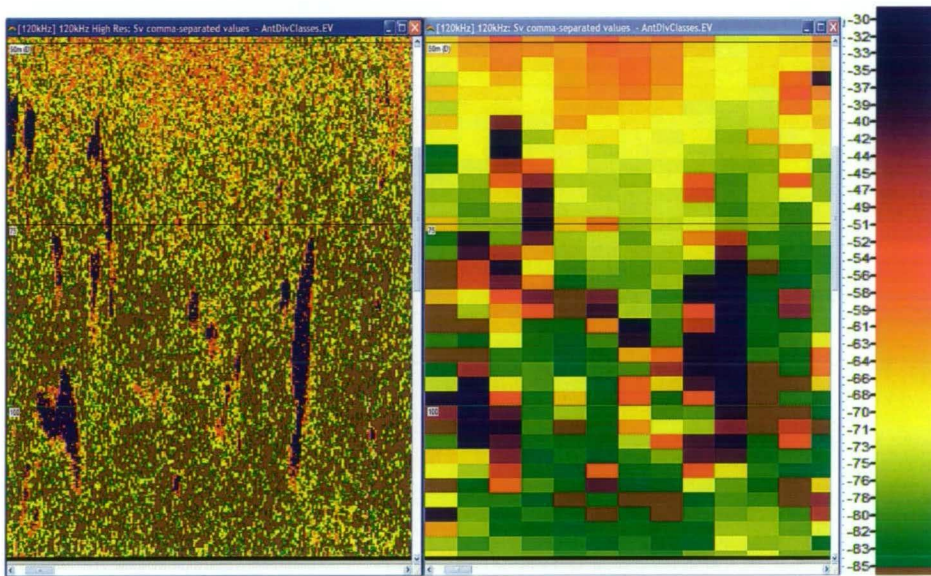


Figure 5.15 Full resolution EK60 data (S_v in dB), 120kHz (left) and the same data down-sampled to a lower resolution, with samples corresponding to scatter nodes (right).

In the AAD approach, the lower resolution samples are classified to one of four classes based on the differences between the backscatter levels at each of the three frequencies (Korneliussen and Ona, 2003; Jarvis, 2006). These samples are now considered as scatter nodes.

Since the scatter nodes arise as the result of a down-sampling operation on the raw samples, density based spatial clustering is pointless, as scatter nodes are equally dense in the entire space covered by the data set. While there is value in taking

spatial proximity into account, that alone is clearly not sufficient to cluster the scatter nodes.

The scatter nodes are subjected to the NJW kernel clustering algorithm, making use of the three backscatter energy levels of each node (38 kHz, 120kHz and 200kHz). These levels are each first normalized to the range [0, 1]. Since the data are basically two-dimensional in space, only two spatial coordinates are used: the distance along the cruise track, and the depth in the water. Additive Gaussian kernels are used, eq. (4.45), with parameters as presented in Table 5.5. From the parameters it can be seen that the backscatter levels are the features primarily used, with smaller contributions from the spatial components of the nodes.

<i>Components</i>	σ	<i>weight</i>
Distance cruise track	0.01	0.1
Depth	0.001	0.3
Backscatter at 38 kHz	0.8	1.0
Backscatter at 120 kHz	0.8	1.0
Backscatter at 200 kHz	0.8	1.0

Table 5.5 Parameters used in applying the NJW algorithm to the scatter nodes from the AAD data set.

The AAD scheme aims at identifying four classes of scatterers: resonant scatterers and fluid-like scatterers; at depths shallower than 100 meters the fluid-like scatterers are divided into two classes: small and large.

The result of the clustering procedure is an assignment of scatter nodes to one of four clusters. This is an unsupervised method, in that it is not capable of identifying the relevant classes, but merely of differentiating between them. Expert knowledge is needed to label the identified clusters as resonant scatterers, or fluid-like scatterers, large or small. In this case the AAD processing of this data set is used to achieve this expert class assignment, or classification.

5.4.3 Results

The scatter nodes are clustered into four clusters using the NJW kernel clustering algorithm. The initial results are shown in Figure 5.16. The colour codes for the classes are as follows:

- green: fluid like scatterers,

- orange: small fluid like scatterers,
- red: large fluid like scatterers,
- blue: resonant scatterers.

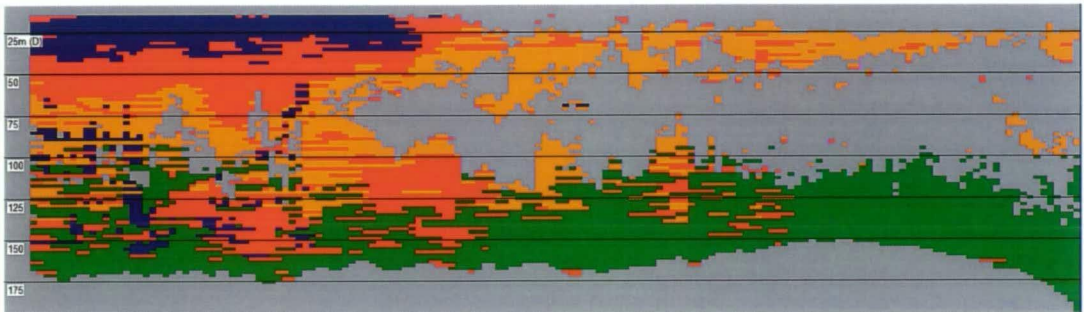


Figure 5.16 The initial result of the kernel clustering method. The small and large fluid-like scatterers, in orange and red respectively, are grouped with the overall class of fluid-like scatterers in green at depths below 100 m, resulting in the classes of Figure 5.17 (e). Resonant scatterers are in blue.

In the AAD classification, samples at depths greater than 100 meters are classified as either resonant (blue) or fluid-like (green). In the kernel-based classification this 100 meter line is not used. Therefore, samples classified into either of these two classes are set to the general class of fluid-like scatterers (green).

In Figure 5.17, the lower resolution samples corresponding to the scatter nodes are shown for each of the three frequencies, together with the classes as identified by the AAD scheme and by the kernel clustering method.

Some observations are made:

- From Figure 5.16, it can be seen that the kernel clustering method assigns hardly any samples at depths shallower than 100 meters to the green cluster of fluid-like scatterers. This is remarkable, as the 100 meter line is not taken into account in any way in the clustering process. This finding provides objective support for the explicit use of this line in the AAD processing scheme.
- The only resonant scatterers identified by the AAD method are the thin layer at shallow depths of less than 25 meters. The kernel method does find resonant scatterers (in blue) at greater depths, including at depths lower than

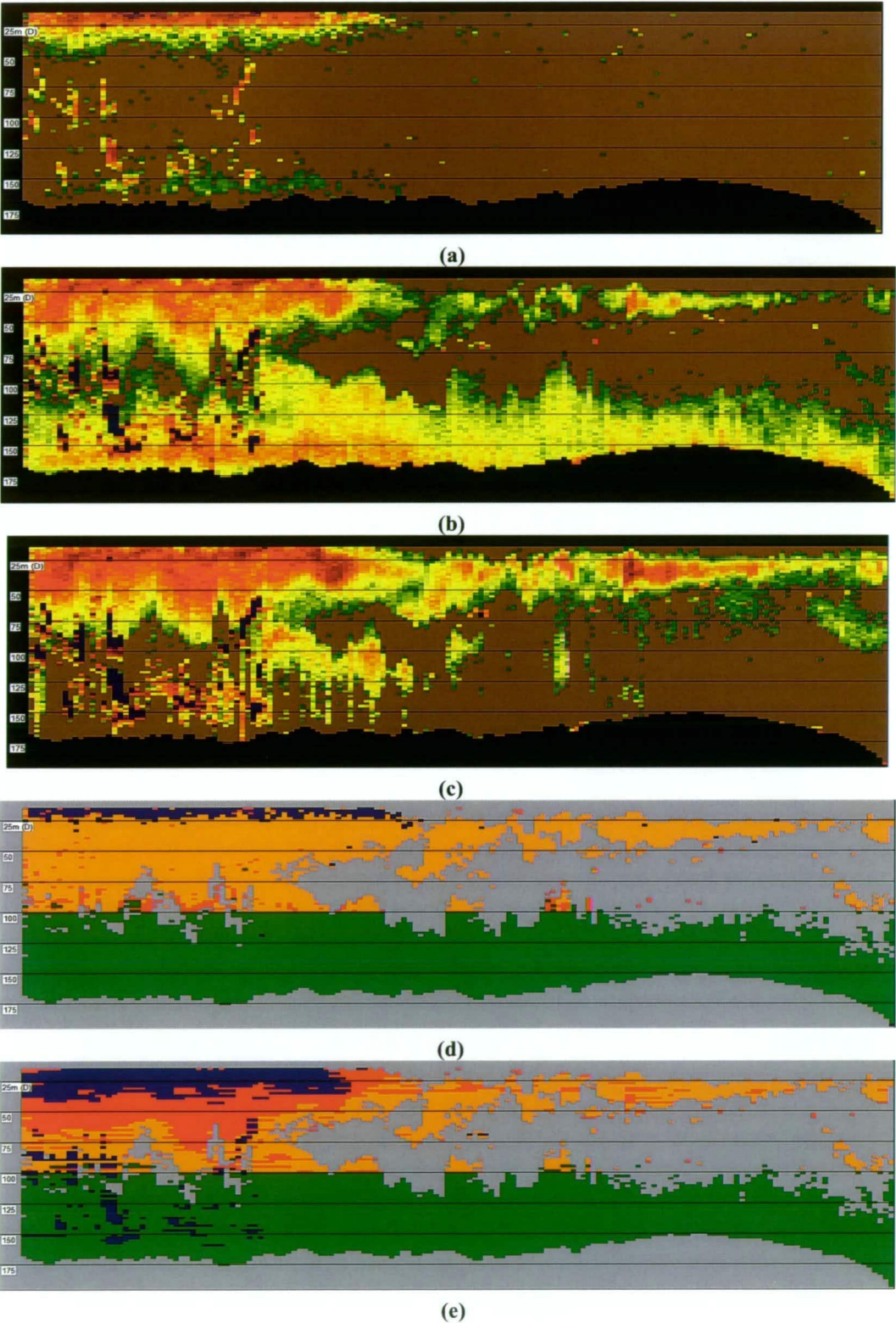


Figure 5.17 (a) 38 kHz backscatter, (b) 120 kHz backscatter, (c) 200 kHz backscatter, (d) classes according to AAD scheme, (e) classes according to the kernel method. The colour scale of (a)-(c) is the same as that of Fig. 5.15, the colour codes of the classes in (d) and (e) is as in Fig. 5.16.

100 meters, whereas the AAD method does not. While it is possible that the scatterers deeper in the water column are of a different nature than those closer to the surface, it seems to be an anomaly that the stronger scatterers deeper in the water column are not identified by the AAD method.

- The layer of resonant scatterers at shallow depths is thinner in the AAD approach. The kernel method appears to be including all above-threshold samples present in the 38 kHz data. If this is judged as an error in the kernel clustering, an approach would be to alter threshold levels prior to creating scatter node features. Using kernel methods, it is possible to include scatter node backscatter values that are obtained from different threshold levels.
- In the depth range 0 to 100 meters, the AAD method labels most scatterers that are fluid-like as small, while a substantial number are labelled as large by the kernel method. From studying the backscatter levels at the different frequencies (Figure 5.17 (a) – (c)) it can be seen that the scatterers identified as large by the kernel method have stronger backscatter levels at 120 kHz and 200 kHz. The AAD method identifies only a few scatterers as large fluid-like, and it does so in a way that is not easily explained by visual inspection of the echograms. This can be interpreted as the kernel method identifying one cluster too many. If the orange and red clusters were merged into one, Figure 5.17 (e) would become very similar to Figure 5.17 (d).

The kernel based classification is plausibly similar to the AAD approach, with some differences that are in need of expert assessment. The kernel method offers a number of benefits over the AAD approach, including the incorporation of spatial aspects of the data, and the possibility to include more features than just the backscatter returns at the three acoustic frequencies. Features that can be included can be of a totally different nature, for example echogram texture measures could prove valuable to take into account (Kieser *et al.*, 2006). Explicit formulations of hard decisions in the AAD scheme, such as the 100 meter boundary, can be incorporated into the kernel method, as demonstrated above.

The kernel clustering method is an unsupervised method to detect clusters in unclustered data. Supervised methods, on the other hand, use examples with a known classification to learn from (section 4.5.2). In the context of the AAD processing scheme it would be feasible to establish a supervised scheme, where a classifier learns from correctly classified data. Rather than unsupervised kernel methods, supervised ones could be used. In particular, the support vector machine would be a good candidate as the most common supervised kernel classification algorithm (Scholkopf *et al.*, 1999; Cristianini and Shawe-Taylor 2000).

In summary, one would proceed as follows:

- cluster data using an unsupervised method, as in the case study presented here,
- apply this unsupervised method to a range of representative data sets,
- use expert knowledge to assign class labels to the clusters obtained,
- use these data sets and the assigned class labels as the training and test sets for training a supervised classification method such as the support vector machine,
- use this trained supervised classifier to classify future unseen data.

This is a useful approach for organisations doing repeat surveys and utilizing standard data processing routines.

6 CONCLUSIONS

6.1 SPATIO-TEMPORAL HYDROACOUSTIC DATA MINING

The widespread use of multibeam sonar for fisheries applications is hampered by a number of factors, an important one of which is the data processing and analysis required. Multibeam sonar data sets are very large and deriving useful information from them is a challenging task.

In this thesis, the problem of handling multibeam water-column data is placed in a scientific data mining context, and a solution is formulated. Algorithms are developed to derive useful information from the raw spatio-temporal hydroacoustic multibeam sonar measurements. A schematic overview of the complete scientific data mining process is given in Figure 6.1.

The first phase in analysing multibeam data is a preprocessing step, to transform the multibeam sonar acoustic backscatter samples to a compact and generic data representation. The elementary units of this generic representation are points in space and time, each having one or more additional features or properties mostly relating to their backscatter energy levels. The spatio-temporal vectors, enriched with additional features, are called scatter nodes in this thesis. Scatter nodes are obtained from raw multibeam sonar data by the application of a deconvolution, which acts as a model inversion technique. They can be regarded as the minimum configuration of scatterers needed to produce the multibeam sonar data that are recorded. The additional features are extracted from the sonar data files or from other sources of information that may be available. Not only is this representation more concise in terms of data volumes, but it is a normalization at the same time:

data collected under different instrument data collection and sampling regimes, and even data from different instruments, are represented in a standard manner.

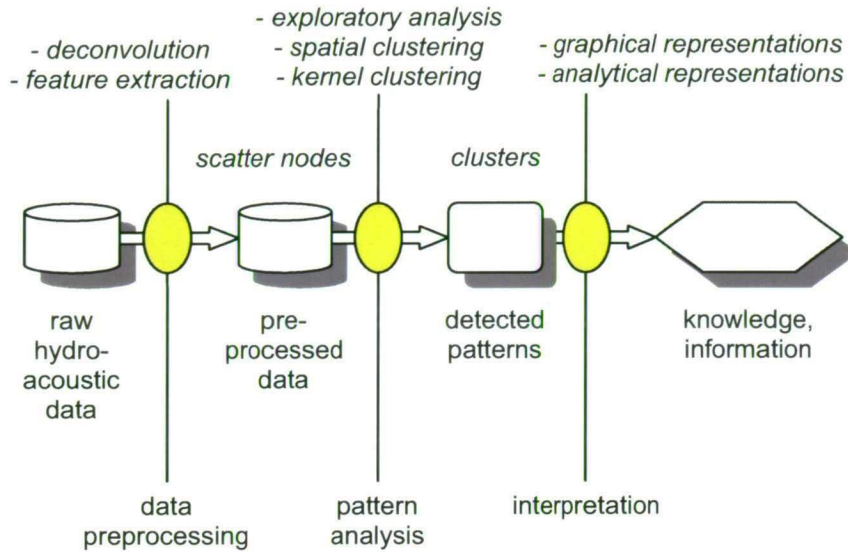


Figure 6.1 The complete scientific data mining process for spatio-temporal hydroacoustic data from multibeam sonar instruments.

Scatter nodes are a sufficiently general representation to capture other aggregated or derived data representations, such as bathymetric soundings, or down-sampled single beam echosounder data. This mechanism allows for the representation of data from various instruments in a unified form. It is also useful for the storage and archiving of backscatter measurements from sonar instruments.

The second phase of the data mining process consists of applying pattern analysis algorithms to scatter node data. Pattern analysis algorithms are used to identify groups or clusters of scatter nodes that belong together because they are representatives of the same higher level object or structure. Such coherent clusters arise for example when fish schools are present in the data, or the sea floor. Scatter nodes indicative of the same structure are expected to have attributes that are more similar than attributes of nodes from different structures. Pattern analysis algorithms aimed at finding these groups or clusters are clustering algorithms. They result in a segmentation of the data set into coherent segments or clusters. Exploratory data analysis is useful to obtain an initial insight into the data and to identify apparent structures.

In this thesis two classes of clustering algorithms are considered. Firstly, spatial clustering algorithms, only taking the spatial attributes of the scatter nodes into account, are considered. These algorithms are simple yet they can provide good

results. A density based spatial clustering algorithm, DBSCAN, is identified as particularly suitable to clustering scatter nodes. A modified version of this algorithm is proposed: Unique-DBSCAN or UDBSCAN. UDBSCAN results in a clustering that is unique for a given data set, whereas the clustering resulting from the standard DBSCAN algorithm can be different for different permutations of the scatter nodes.

Secondly, more sophisticated algorithms are considered, with the aim of including all attributes of scatter nodes, not only the spatial ones. Kernel methods are identified as relevant. They have strong foundations in functional analysis and algebra, and offer a convenient way to make known established linear statistical methods non-linear. A long standing theorem from functional analysis, the Hahn-Banach theorem, is used in this research to develop the necessary mathematical foundations to extend kernel methods to include the spatial, temporal and other feature components of scatter nodes simultaneously. Using these extended kernels, kernel clustering methods are applied to scatter nodes. The kernel k-means clustering algorithm is considered, in particular the variant known as the Ng-Jordan-Weiss (NJW) algorithm, a well established spectral relaxation of the standard kernel k-means algorithm.

The result of applying these pattern analysis algorithms to scatter node data is a segmentation of the nodes. Each segment or cluster is indicative of some underwater object such as a fish school or the seabed. These clusters can be analysed further. For example, they can be used to create volumetric objects representative of fish schools. Backscatter energy measures of fish schools are obtained through the backscatter energy related features of the constituting scatter nodes. Scatter node clusters and derived objects allow for convenient graphical representations of the data for information visualization purposes in presenting analysis results.

The segmentation routines are extendible to perform classification. In classification, unseen data are labelled to belong to a particular group or class. After a data set of scatter nodes has been clustered using any of the aforementioned methods, and clusters are labelled using expert knowledge, the algorithms can be used directly to assign unseen scatter nodes to one of the determined classes.

A number of case studies are presented. Data from different instruments are analysed using the proposed methods. These examples illustrate the usefulness of the methods and are a guide to bringing them into practice.

The proposed methods will facilitate present and future studies employing multibeam sonar technology, and are anticipated to be general enough to prove useful in supporting future developments such as the use of multibeam sonar for fish stock assessments.

6.2 AN EXTENSIBLE FRAMEWORK

The proposed data mining process can serve as a framework for the mining of spatio-temporal data in general. It is a framework in the sense that both the data and the algorithms can be generalized, extended or altered. Several options are discussed.

Instruments

In this thesis the primary focus is on multibeam sonar instruments that are capable of collecting data samples from the complete water column. In section 3.4.4 the possibility of deriving scatter nodes from single beam echosounder data is introduced, with an example given in section 5.4. This can be investigated in further detail. Of particular interest is the possibility of using the deconvolution technique, rather than the resampling technique that is currently used in single beam sonar data analysis.

The multibeam systems discussed in this thesis are limited to instruments with coplanar beam configurations. The method of deriving scatter nodes can be extended to cover instruments with other transducers and beam configurations, such as omnidirectional or scanning sonars (section 2.1.2), which have been used successfully in fisheries studies (Brehmer *et al.*, 2006). Models of this kind include the Furuno FSV30(R) and the Simrad SP and SH series models.

Furthermore, it is anticipated that new instruments specifically designed for fisheries work will become available, and that the technique of transforming the raw data into scatter nodes by means of a deconvolution will prove useful there as well. A new generation of multibeam systems for fisheries research is presented in Andersen *et al.* (2006).

Data types

The data type of interest in this thesis is spatio-temporal hydroacoustic data: acoustic measurements acquired by underwater sonar. Pattern analysis methods are applied to these data sets. As far as the pattern analysis methods are concerned, the fact that the data they are applied to are hydroacoustic is not important. Any kind of quantitative spatio-temporal measurement can be analysed in exactly the same manner. Many oceanographic data sets are of this kind: a measured or modeled quantity such as salinity, temperature, acidity or current velocity is available at a number of points in space and time. If the aim is to cluster these spatio-temporal data points into similar groups, the pattern analysis methods presented in this thesis

can be used; in particular the kernel method which allows for the combination of spatial and non-spatial components of the points in the analysis.

Algorithms

Various algorithms are used in the data mining process presented in this thesis. At each of the yellow nodes in the overview in Figure 6.1, one or more algorithms are or can be used.

One possible extension of the preprocessing phase is to apply the proposed deconvolution algorithm across pings. In the present research, only within-ping deconvolution is considered. Across-ping deconvolution is a complex matter due to the generally irregular movements of the transducer arrays from ping to ping. Expected benefits include further data reduction and higher precision, as ping to ping correlations are then taken into account.

Alternative pattern analysis algorithms other than UDBSCAN and kernel methods are possible, using the same preprocessed multibeam data in the form of scatter nodes as inputs. One possibility is to establish spatial gridding methods for scatter nodes in a similar manner as they are used for bathymetric multibeam soundings (Calder and Mayer, 2003).

Data fusion

The ecosystem-based approach to managing marine resources consists of combining many sources of information to achieve a holistic insight into the complete ecosystem (De la Mare, 2005; Garcia and Cochrane, 2005; Frid *et al.*, 2006). How to incorporate water-column multibeam sonar data in such analyses is an outstanding challenge because of the instrument and manufacturer dependent custom storage techniques for raw multibeam sonar data. Since clusters of segmented scatter nodes are simply sets of generic spatio-temporal vectors, a format well suited for import in many software packages, it is anticipated that they will facilitate the use of multibeam sonar data in the ecosystem-based approach to marine resource management.

6.3 SUMMARY

A scientific data mining approach for the processing and analysis of water-column multibeam sonar data is developed utilizing concepts and algorithms from the research areas of underwater acoustics and pattern analysis. The following is a summary of the contributions delivered by the research presented in this thesis:

- the development of a multibeam sonar model, employing an existing acoustic ray-tracing model and known multibeam sonar instrument designs,
- the application of a deconvolution as a model inversion technique in order to obtain the minimal set of scatterers that would lead to the observed data,
- the extension of this minimal set of scatterers with additional features and defining these feature-rich spatio-temporal vectors as scatter nodes; scatter nodes are a concise, generic representation of spatio-temporal hydroacoustic data sets,
- the identification of the algorithm *Density Based Spatial Clustering for Applications with Noise* (DBSCAN) as an appropriate spatial clustering algorithm for scatter nodes,
- the modification of DBSCAN to overcome an issue with uniqueness of its results; the modified version is named Unique-DBSCAN (UDBSCAN),
- the development of a mathematical foundation to enable the application of kernel methods to spatio-temporal data; the Hahn-Banach theorem plays a fundamental role in this theory, which provides a method to use spatial, temporal and other features simultaneously,
- the application of kernel clustering methods to scatter nodes using this mathematical foundation; in particular the kernel k-means method and its variant known as the Ng-Jordan-Weiss (NJW) algorithm are used to segment scatter nodes into clusters indicative of coherent structures in the data,
- demonstration of the effectiveness of the data mining process by means of a number of case studies.

These developments are capable of facilitating the routine use of water-column multibeam sonar data for fisheries applications.

APPENDIX: ABSTRACTS OF PUBLICATIONS

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2003). "Midwater acoustic modeling for multibeam sonar simulation," 146th ASA Meeting, Austin, Texas, The Journal of the Acoustical Society of America 114, p. 2308.

Simulation and modeling software has been developed to generate synthetic midwater multibeam data. Essentially, the simulator can be considered as a virtual test tank. In order to develop multibeam data analysis methods for fisheries research, it is essential to have a variety of test data sets available, which are ground truthed, georeferenced and corrected for vessel motion. Since equipment and ship time are expensive and data quality not always guaranteed, the simulator provides an effective alternative. The seabed and any objects in the water column such as fish and fish schools can be defined in a 3-dimensional space. A specification for a generic linear array multibeam sonar and its position in space and time can be chosen. The acoustic model implements the technique of acoustic ray-tracing to obtain the pressure at the transducer face, which is converted to individual samples by modeling the working of a digital multibeam system. Beamforming is performed on the fly, and both raw and beamformed complex data sets are generated. Statistical validation of the generated data has been conducted successfully.

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2004). "A framework for scientific data mining in hydroacoustic data sets," 2nd International Conference on Artificial Intelligence in Science and Technology (AISAT) (Hobart, Tasmania, Australia), pp. 104-108.

A data mining framework for handling large volumes of scientific hydroacoustic backscatter data is proposed. The method is applicable to data collected by the new generation of multibeam echosounders, capable of logging acoustic backscatter data for the full water column. Such instruments are increasingly used for fisheries applications. The data mining technique is based on an inverse modeling of the underlying physics and electronics of a generic multibeam sonar system. A set of tagged soundings is obtained, which serves as a base for further advanced analysis techniques. It is anticipated that the proposed framework will serve as a tool for scientific fisheries research.

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2005). "Model inversion for midwater multibeam backscatter data analysis," IEEE Oceans '05 Europe (Brest, France), pp. 431-435.

A model of the multibeam echosounding process was developed. This model has now been used as the basis for the application of a model inversion technique, with the aim of analysing midwater multibeam echosounder data, for fisheries applications. Research on midwater multibeam echosounding for fisheries is in its infancy. Some results have been published, announcing promising progress at the level of multibeam transducer design, beamforming algorithms and calibration procedures, but no standard postprocessing technique has emerged yet. In this paper, the postprocessing of midwater multibeam backscatter data is placed in a scientific data mining framework. Data mining aims at automatically extracting useful information and knowledge from large volumes of data which do not reveal this knowledge in a trivial manner. Multibeam acoustic data has an additional dimension compared to single beam data, and multibeam echosounding results in large data logging rates, typically several gigabytes per hour, making it suitable for applying data mining algorithms in order to analyse the data in postprocessing. A data mining technique to handle multibeam data sets is presented. The technique is based on inverse modeling. A model of the multibeam echosounding process was developed, including a physical underwater acoustics model, as well as a model of a generic multibeam transducer and its digital signal processor. This model has now been approximated by an invertible function, leading to an inverse model. Applying the inverse model to midwater multibeam backscatter data results in a set of soundings. A multibeam midwater sounding is the equivalent of a standard multibeam sounding as obtained from hydrographic multibeam instruments. In the midwater multibeam echosounding context, a sounding can represent anything in the water column, not just the seabed. These soundings can be visualized directly, allowing for exploratory data analysis in a 3d or 4d interactive environment. Furthermore, various features can be tagged to each sounding, such as the backscatter energy value and some statistical parameters of the multibeam ping from which the sounding was obtained. The term data node is used to describe the sounding and its associated feature vector. The set of data nodes serves as the basis for further advanced spatio-temporal data mining techniques. Soundings can be clustered into coherent groups, each cluster representing an object in the water column, such as a fish school. Cluster features are obtained from the feature tags of their contained data nodes, giving rise to feature vectors for each cluster. Clusters can be classified into classes of different types, using each cluster's feature vector. When a cluster is thought of as a fish school, it can be classified according to fish species or age group, for example.

The concept of a set of data nodes is a versatile concept that can be extended further, enabling the application of more advanced clustering and classification algorithms.

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2005). "A scientific data mining approach to midwater multibeam echosounding for fisheries applications," 1st International Conference on Underwater Acoustic Measurements: Technologies & Results (UAM) (Heraklion, Crete, Greece).

Midwater acoustic backscatter measurements collected by multibeam sonar offer new opportunities and challenges for fisheries applications. A scientific data mining technique to handle midwater multibeam backscatter data is presented. Most of the earlier research on multibeam echosounding for fisheries has focused on the core basic technologies of multibeam transducers, the associated signal processing, and calibration. Some work has been done with postprocessed data, but no systematic methodology for postprocessing of midwater multibeam backscatter data has emerged. In this paper, the problem is placed in a data mining framework. A model inversion technique is utilized, by applying the inverse of an approximation to the multibeam echosounding model. The proposed approach leads to a data product consisting of a collection of midwater soundings. A multibeam midwater sounding is the equivalent of the standard multibeam soundings as obtained from hydrographic multibeam instruments. These soundings can be visualized directly, allowing for exploratory data analysis in a 3d or 4d interactive environment. A sounding is a measurement in space and time, and has associated attributes or features, such as the backscatter value. Other features can be tagged to the soundings, forming generalised data nodes. Advanced spatio-temporal data mining techniques can now be applied to this set of nodes. Some further clustering techniques are presented, clustering the soundings into groups representing coherent objects in the water column, or, more specifically, fish schools. Global properties of clusters can be derived from the individual feature tags of the soundings, thus allowing for classification of schools into classes of similar types. The latest developments of this research are presented.

Buelens, B., Williams, R., Sale, A., and Pauly, T. (2006). "Computational challenges in processing and analysis of full water-column multibeam sonar data," 8th European Conference on Underwater Acoustics, edited by S. M. Jesus, and O. C. Rodríguez (Carvoeiro, Portugal), pp. 799-804.

Several multibeam sonar systems are now capable of collecting and recording data samples covering the full water column, not just the seabed. Such systems, while still facing hardware challenges such as limited dynamic range and bandwidth, collect vast quantities of data, generally an order of magnitude more than conventional hydrographic multibeam or scientific single beam sonar systems. In this paper, the challenges faced by data processing systems for analysis of full water-column multibeam sonar data are explored. Full water-column multibeam data sets are valuable to scientists from traditionally diverse fields, providing

simultaneous information about bathymetry, seabed type and habitats, and biomass in the water column. Aspects of the data processing pipeline that are considered in this paper include raw data storage, data preprocessing, visualization and exploratory data analysis, statistical data analysis and postprocessing, and presentation and interpretation of results. A general framework is outlined, and specific aspects applicable to the kind of data and problems at hand are emphasized. Proposed solutions to some of the challenges are reviewed and placed within an overall framework of multibeam sonar water-column data analysis. It will become clear that successful contributions to the field have been made, but that a general analysis method has yet to emerge.

Buelens, B., Pauly, T., Williams, R., and Sale, A. (in press). "Kernel methods for detection and classification of fish schools in single beam and multibeam acoustic data," in *ICES Journal of Marine Science, Special Issue on the Ecosystem Approach with Fisheries Acoustics and Complementary Technologies*.

A kernel method for clustering acoustic data from single-beam echosounder and multibeam sonar is presented. The algorithm is used to detect fish schools and to classify acoustic data into clusters of similar acoustic properties. In a preprocessing routine, data from single-beam echosounder and multibeam sonar are transformed into an abstracted representation by multidimensional nodes, which are data points with spatial, temporal, and acoustic features as components. Kernel methods combine these components together to determine clusters based on joint spatial, temporal and acoustic similarities. The resulting clusters yield a classification of the data in groups of similar nodes. Including the spatial components results in clusters for each school and effectively detects fish schools, while ignoring the spatial components yields a classification according to acoustic similarities, corresponding to classes of different species or age groups. The method is described and two case studies are presented.

REFERENCES

- Abraham, D. A., and Lyons, A. P. (2002). "Novel physical interpretations of K-distributed reverberation," *Ieee Journal of Oceanic Engineering* **27**, 800-813.
- Abraham, D. A., and Lyons, A. P. (2004). "Reverberation envelope statistics and their dependence on sonar bandwidth and scattering patch size," *Ieee Journal of Oceanic Engineering* **29**, 126-137.
- Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1995). *Fast discovery of association rules* (AAAI/MIT Press, Cambridge, MA).
- Anderson, C. I. H., Horne, J. K., and Boyle, J. (2007). "Classifying multi-frequency fisheries acoustic data using a robust probabilistic classification technique," *Journal of the Acoustical Society of America* **121**.
- Ankerst, M., Breunig, M., Kriegel, H.-P., and Sander, J. (1999). "OPTICS: Ordering Points to Identify the Clustering Structure," in *ACM-SIGMOD International Conference on Management of Data*, pp. 49-60.
- Arsenault, R., Ware, C., Plumlee, M., Martin, S., Whitcombe, L. L., Wiley, D., Gross, T., and Bilgili, A. (2004). "A system for visualizing time-varying oceanographic 3D data," in *Oceans 2004* (Kyoto).
- Axelsen, B. E., Anker-Nilssen, T., Fossum, P., Kvamme, C., and Noettestad, L. (2001). "Pretty patterns but a simple strategy: predator-prey interactions between juvenile herring and Atlantic puffins observed with multibeam sonar," *Canadian Journal of Zoology* **79**, 1586-1596.
- Banach, S. (1929). "Sur les fonctionelles lineaires," *Studia Mathematica* **1**, 211-216, 223-229.
- Bell, J. M. (1997). "Application of optical ray tracing techniques to the simulation of sonar images," *Optical Engineering* **36**, 1806-1813.
- Bell, J. M., and Linnett, L. M. (1997). "Simulation and analysis of synthetic sidescan sonar images," *Iee Proceedings-Radar Sonar and Navigation* **144**, 219-226.
- Bellman, R. E. (1961). *Adaptive control processes* (Princeton University Press, Princeton, NJ).
- Ben-Hur, A., Horn, D., Siegelmann, H. T., and Vapnik, V. (2000). "A Support Vector Method for Clustering," in *NIPS*, pp. 367-373.
- Benoit-Bird, K., and Au, W. (2003). "Hawaiian spinner dolphins aggregate midwater food resources through cooperative foraging," *Journal of the Acoustical Society of America* **114**, 2300.
- Bishop, C. M. (2006). *Pattern recognition and machine learning* (Springer, New York).
- Brehmer, P., Lafont, T., Georgakarakos, S., Josse, E., Gerlotto, F., and Collet, C. (2006). "Omnidirectional multibeam sonar monitoring: applications in fisheries science," *Fish and Fisheries* **7**, 165-179.
- Breiman, L. (2001). "Statistical modeling: The two cultures," *Statistical Science* **16**, 199-215.

- Brissette, M. B. (1997). "The application of multibeam sonars in route survey," in *Ocean Mapping Group* (University of New Brunswick, New Brunswick).
- Brouns, G., De Wulf, A., and Constales, D. (2003). "Delaunay triangulation algorithms useful for multibeam echosounding," *Journal of Surveying Engineering-Asce* **129**, 79-84.
- Buelens, B., Williams, R., Sale, A., and Pauly, T. (2006). "Computational challenges in processing and analysis of full-watercolumn multibeam sonar data," in *8th European Conference on Underwater Acoustics*, edited by S. M. Jesus, and O. C. Rodriguez (Carvoeiro, Portugal), pp. 799-804.
- Buelens, B., Wilson, M., and Horne, J. K. (2007). "Multibeam water column data analysis for fisheries research: a worked example in Echoview," in *Underwater Acoustic Measurements: Technologies & Results* (Crete, Greece).
- Calder, B. R. (2003). "Automatic Statistical Processing of Multibeam Echosounder Data," *International Hydrographic Review* **4**, 1-16.
- Calder, B. R., and Mayer, L. A. (2001). "Robust automatic multi-beam bathymetric processing," (University of New Hampshire, Durham).
- Calder, B. R., and Mayer, L. A. (2003). "Automatic processing of high-rate, high-density multibeam echosounder data," *Geochemistry Geophysics Geosystems* **4**.
- Campbell, C. (2002). "Kernel methods: a survey of current techniques," *Neurocomputing* **48**, 63-84.
- Canepa, G., Bergem, O., and Pace, N. G. (1999). "Trismap: A fast way to deal with new multibeam sonar data," *Sea Technology* **40**, 49, 44 pgs.
- Canepa, G., Bergem, O., and Pace, N. G. (2003). "A new algorithm for automatic processing of bathymetric data," *Ieee Journal of Oceanic Engineering* **28**, 62-77.
- Chakraborty, B., Kodagali, V., and Baracho, J. (2003). "Sea-floor classification using multibeam echo-sounding angular backscatter data: A real-time approach employing hybrid neural network architecture," *Ieee Journal of Oceanic Engineering* **28**, 121-128.
- Chakraborty, B., and Schenke, H. W. (1995). "Arc arrays: Studies of high resolution techniques for multibeam bathymetric applications," *Ultrasonics* **33**, 457-461.
- Chakraborty, B., Schenke, H. W., Kodagali, V. N., and Hagen, R. (2001). "Analysis of multibeam-hydrosweep echo peaks for seabed characterisation," *Geo-Mar. Lett.* **20**, 174-181.
- Chitroub, S., Houacine, A., and Sansal, B. (2002). "Statistical characterisation and modelling of SAR images," *Signal Processing* **82**, 69-92.
- Chu, D., Baldwin, K. C., Foote, K. G., Yanchao, L., Mayer, L. A., and Melvin, G. D. (2001a). "Multibeam sonar calibration: removal of static surface reverberation by coherent echo subtraction," *MTS/IEEE Oceans 2001. An Ocean Odyssey. Conference Proceedings (IEEE Cat. No.01CH37295)*, 2498-2502.
- Chu, D., Baldwin, K. C., Foote, K. G., Yanchao, L., Mayer, L. A., and Melvin, G. D. (2001b). "Multibeam sonar calibration: target localization in azimuth," *MTS/IEEE Oceans 2001. An Ocean Odyssey. Conference Proceedings (IEEE Cat. No.01CH37295)*, 2506-2510.

- Cios, K. J., Pedrycz, W., and Swiniarski, R. W. (1998). *Data Mining Methods for Knowledge Discovery* (Springer).
- Clarke, J. E. H., Mayer, L. A., and Wells, D. E. (1996). "Shallow-water imaging multibeam sonars: A new tool for investigating seafloor processes in the coastal zone and on the continental shelf," *Marine Geophysical Researches* **18**, 607-629.
- Clay, C. S. (1983). "Deconvolution of the fish scattering PDF from the echo PDF for a single transducer sonar," *Journal of the Acoustical Society of America* **73**, 1989-1994.
- Clay, C. S., and Medwin, H. (1977). *Acoustical Oceanography: Principles and Applications* (John Wiley and Sons, New York).
- Cochrane, N. A., Li, Y., and Melvin, G. D. (2003). "Quantification of a multibeam sonar for fisheries assessment applications," *Journal of the Acoustical Society of America* **114**, 745-758.
- Cox, D. R., Efron, B., Hoadley, B., Parzen, E., and Breiman, L. (2001). "Statistical modeling: The two cultures - Comments and rejoinders," *Statistical Science* **16**, 216-231.
- Cristianini, N., and Shawe-Taylor, J. (2000). *Support vector machines* (Cambridge University Press, Cambridge, UK).
- Cristianini, N., Shawe-Taylor, J., and Kandola, J. (2001). "Spectral kernel methods for clustering," in *Neural Information Processing Systems (NIPS)*, pp. 649-655.
- Crocker, M. J. (1998). *Handbook of acoustics* (John Wiley & Sons).
- Curtis, T. E. (1998). "Principles of sonar beamforming," (Curtis Technology Ltd., UK).
- De la Mare, W. K. (2005). "Marine ecosystem-based management as a hierarchical control system," *Marine Policy* **29**, 57-68.
- de Moustier, C. (1988). "State of the Art in Swath Bathymetry Survey-Systems," *International Hydrographic Review* **65**, 25-54.
- de Moustier, C. (1993). "Signal processing for swath bathymetry and concurrent seafloor acoustic imaging," in *Acoustic signal processing for ocean exploration*, edited by J. M. F. Moura, and I. M. G. Lourtie (Kluwer Academic Publishers, Dordrecht), pp. 329-354.
- de Moustier, C., and Matsumoto, H. (1993). "Sea-Floor Acoustic Remote-Sensing with Multibeam Echo-Sounders and Bathymetric Sidescan Sonar Systems," *Marine Geophysical Researches* **15**, 27-42.
- Delaunay, B. (1934). "Sur la sphere vide," *Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* **7**, 793-800.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society (Series B)* **39**, 1-38.
- Dhillon, I. S., Guan, Y., and Kulis, B. (2004). "Kernel k-means, spectral clustering and normalized cuts," in *KDD'04* (ACM, Seattle, Washington, USA).
- Dhillon, I. S., Guan, Y., and Kulis, B. (2005). "A unified view of kernel k-means, spectral clustering and graph cuts," (Department of Computer Science, University of Texas at Austin, Austin, TX, USA), p. 20.

- Di Bisceglie, M., Galdi, C., and Griffiths, H. D. (1999). "Statistical scattering model for high-resolution sonar images: characterisation and parameter estimation," *Iee Proceedings-Radar Sonar and Navigation* **146**, 264-272.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (2nd Edition)* (Wiley-Interscience, New York).
- Edelsbrunner, H., and Mucke, E. P. (1994). "Three-dimensional alpha shapes," *Acm Transactions on Graphics* **13**, 43-72.
- Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)* (AAAI Press, Portland, Oregon, USA), pp. 226-231.
- Etter, P. C. (2001). "Recent advances in underwater acoustic modelling and simulation," *Journal of Sound and Vibration* **240**, 351-383.
- FAO (2006). *State of the World Fisheries and Aquaculture (SOFIA) 2004* (Food and Agriculture Organization of the United Nations, Rome).
- Fayyad, U., Haussler, D., and Stolorz, P. (1996). "Mining scientific data," *Communications of the Acm* **39**, 51-57.
- Foley, J. D., van Dam, A., Feiner, S. K., and Hughes, J. F. (1995). *Computer graphics: principles and practice in C (2nd edition)* (Addison-Wesley Professional).
- Fonseca, L., Mayer, L., Orange, D., and Driscoll, N. (2002). "The high-frequency backscattering angular response of gassy sediments: Model/data comparison from the Eel River Margin, California," *Journal of the Acoustical Society of America* **111**, 2621-2631.
- Foote, K. G. (1983). "Linearity of fisheries acoustics, with addition theorems," *Journal of the Acoustical Society of America* **73**, 1932-1940.
- Foote, K. G. (1987). "Fish target strengths for use in echo integrator surveys," *Journal of the Acoustical Society of America* **82**, 981-987.
- Foote, K. G. (1991). "Summary of methods for determining fish target strength at ultrasonic frequencies," *Ices Journal of Marine Science* **48**, 211-217.
- Foote, K. G., Chu, D. Z., Hammar, T. R., Baldwin, K. C., Mayer, L. A., Hufnagle, L. C., and Jech, J. M. (2005). "Protocols for calibrating multibeam sonar," *Journal of the Acoustical Society of America* **117**, 2013-2027.
- Foote, K. G., Knudsen, H. P., Vestnes, G., MacIennan, D. N., and Simmonds, E. J. (1987). "Calibration of acoustic instruments for fish density estimation: a practical guide," *ICES Cooperative Research Report* **144**.
- Foote, K. G., and Steffanson, G. (1993). "Defintion of the problem of estimating fish abundance over an area from acoustic line-transect measurements of density," *Ices Journal of Marine Science* **50**, 369-381.
- Frid, C. L. J., Paramor, O. A. L., and Scott, C. L. (2006). "Ecosystem-based management of fisheries: is science limiting?," *Ices Journal of Marine Science* **63**, 1567-1572.
- Gallaudet, T. C., and de Moustier, C. P. (2003). "High-frequency volume and boundary acoustic backscatter fluctuations in shallow water," *Journal of the Acoustical Society of America* **114**, 707-725.

- Garcia, S. M., and Cochrane, K. L. (2005). "Ecosystem approach to fisheries: a review of implementation guidelines," *Ices Journal of Marine Science* **62**, 311-318.
- Gerlotto, F., Bertrand, S., Bez, N., and Gutierrez, M. (2006). "Waves of agitation inside anchovy schools observed with multibeam sonar: a way to transmit information in response to predation," *Ices Journal of Marine Science* **63**, 1405-1417.
- Gerlotto, F., Castillo, J., Saavedra, A., Barbieri, M. A., Espejo, M., and Cotel, P. (2004). "Three-dimensional structure and avoidance behaviour of anchovy and common sardine schools in central southern Chile," *Ices Journal of Marine Science* **61**, 1120-1126.
- Gerlotto, F., and Paramo, J. (2003). "The three-dimensional morphology and internal structure of clupeid schools as observed using vertical scanning multibeam sonar," *Aquatic Living Resources* **16**, 113-122.
- Gerlotto, F., Soria, M., and Freon, P. (1999). "From two dimensions to three: the use of multibeam sonar for a new approach in fisheries acoustics," *Canadian Journal of Fisheries and Aquatic Sciences* **56**, 6-12.
- Giles, J. R. (2000). *Introduction to the analysis of normed linear spaces* (Cambridge University Press).
- Girolami, M. (2002). "Mercer kernel-based clustering in feature space," *Ieee Transactions on Neural Networks* **13**, 780-784.
- Gjerde, K. (2006). "Ecosystems and biodiversity in deep waters and high seas," in *UNEP Regional Seas Report and Studies* (United Nations Environment Programme).
- Grossman, R. (2001). *Data mining for scientific and engineering applications* (Kluwer Academic Publishers).
- Guha, S., Rastogi, R., and Shim, K. (2001). "Cure: An efficient clustering algorithm for large databases," *Information Systems* **26**, 35-58.
- Hafsteinsson, M. T., and Misund, O. A. (1995). "Recording the migration behaviour of fish schools by multibeam sonar during conventional acoustic surveys," *Ices Journal of Marine Science* **52**, 915-924.
- Hahn, H. (1927). "Uber lineare Gleichungssysteme in linearen Raumen," *Journal fur die reine und angewandte Mathematik* **157**, 214-229.
- Hammerstad, E. (1995). "Advanced Multibeam Echosounder Technology," *Sea Technology* **36**, 67-69.
- Hansen, C. D., and Johnson, C. R. (2004). *Visualization handbook* (Academic Press).
- Haralabous, J., and Georgakarakos, S. (1996). "Artificial neural networks as a tool for species identification of fish schools," *Ices Journal of Marine Science* **53**, 173-180.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning* (Springer, New York).
- Hellequin, L., Boucher, J. M., and Lurton, X. (2003). "Processing of high-frequency multibeam echo sounder data for seafloor characterization," *Ieee Journal of Oceanic Engineering* **28**, 78-89.
- Hongler, M. O. (1988). "Dynamic Derivation of the Weak-Scattering K-Density," *Journal of the Optical Society of America a-Optics Image Science and Vision* **5**, 1649-1651.

- Hughes Clarke, J. E., De Moustier, C., Mayer, L., and Wells, D. (eds). (2000). *Fourth Asia-Pacific coastal multibeam sonar training course* (Course notes, Cairns, Australia).
- ICES (2005). "Description of the ICES HAC Standard Data Exchange Format, Version 1.60.," in *ICES Cooperative Research Report* (International Council for the Exploration of the Sea, Copenhagen), p. 86.
- ICES (2006). "Report of the Working Group on Marine Data Management (WGMDM)," (ICES Headquarters, Copenhagen), p. 67.
- ICES (2007a). "Acoustic seabed classification of marine physical and biological landscapes," in *ICES Cooperative Research Report* (International Council for the Exploration of the Sea, Copenhagen), p. 183.
- ICES (2007b). "Report of the Joint Workshop of the ICES-FAO Working Group on Fishing Technology and Fish Behaviour [WGFTFB] and the Working Group on Fisheries Acoustics Science and Technology [WGFAST] (JFATB)," (International Council for the Exploration of the Sea, Dublin), p. 35.
- IODE (2007). "International Oceanographic Data and Information Exchange," (<http://www.iode.org/>).
- Jain, A. K., Duin, R. P. W., and Mao, J. C. (2000). "Statistical pattern recognition: A review," *Ieee Transactions on Pattern Analysis and Machine Intelligence* **22**, 4-37.
- Jakeman, E., and Tough, R. J. A. (1987). "Generalized K-Distribution - a Statistical-Model for Weak Scattering," *Journal of the Optical Society of America a-Optics Image Science and Vision* **4**, 1764-1772.
- Jakeman, E., and Tough, R. J. A. (1988). "Non-Gaussian Models for the Statistics of Scattered Waves," *Advances in Physics* **37**, 471-529.
- Jarvis, T. (2006). "Hydroacoustic data post-processing using SonarData Echoview v3.50," (Australian Antarctic Division, Kingston).
- Johnson, R. L., Simmons, M. A., Simmons, C. S., and Blanton, S. L. (2001). "A New Multibeam Sonar Technique for Evaluating Fine-Scale Fish Behavior Near Hydroelectric Dam Guidance Structures," in *American Fisheries Society Symposium*, pp. 161-170.
- Kang, M., Honda, S., and Oshima, T. (2006). "Age characteristics of walleye pollock school echoes," *Ices Journal of Marine Science* **63**, 1465-1476.
- Karypis, G., Han, E. H., and Kumar, V. (1999). "Chameleon: Hierarchical clustering using dynamic modeling," *Computer* **32**, 68-+.
- Keeton, J. A., and Searle, R. C. (1996). "Analysis of Simrad EM12 multibeam bathymetry and acoustic backscatter data for seafloor mapping, exemplified at the Mid-Atlantic Ridge at 45 degrees N," *Marine Geophysical Researches* **18**, 663-688.
- Kieser, R., Tesler, W., Buelens, B., and Wilson, M. (2006). "Implementation of seabed classification procedure for echogram and fish species classification," in *ICES. 2006. Report of the Working Group on Fisheries Acoustics Science and Technology (WGFAST)* (Hobart, Tasmania).
- Kim, D. W., Lee, K. Y., Lee, D., and Lee, K. H. (2005). "Evaluation of the performance of clustering algorithms in kernel-induced feature space," *Pattern Recognition* **38**, 607-611.

- Kloser, R. J., Bax, N. J., Ryan, T., Williams, A., and Barker, B. A. (2001). "Remote sensing of seabed types in the Australian South East Fishery; development and application of normal incident acoustic techniques and associated 'ground truthing'," *Marine and Freshwater Research* **52**, 475-489.
- Knight, W. C., Pridham, R. G., and Kay, S. M. (1981). "Digital Signal Processing for Sonar," *Proceedings of the Ieee* **69**, 1451-1506.
- Kohonen, T. (2001). *Self-organizing maps, 3rd edition* (Springer-Verlag, New York).
- Konstantopoulos, C., Mittag, L., Sandri, G., and Beland, R. (1990). "Deconvolution of Gaussian Filters and Antidiffusion," *Journal of Applied Physics* **68**, 1415-1420.
- Korneliussen, R. J., and Ona, E. (2003). "Synthetic echograms generated from the relative frequency response," *Ices Journal of Marine Science* **60**, 636-640.
- Lawson, G. L., Barange, M., and Freon, P. (2001). "Species identification of pelagic fish schools on the South African continental shelf using acoustic descriptors and ancillary information," *Ices Journal of Marine Science* **58**, 275-287.
- Lingvall, F. (2004). "A method of improving overall resolution in ultrasonic array imaging using spatio-temporal deconvolution," *Ultrasonics* **42**, 961-968.
- Lingvall, F., Olofsson, T., and Stepinski, T. (2003). "Synthetic aperture imaging using sources with finite aperture: Deconvolution of the spatial impulse response," *Journal of the Acoustical Society of America* **114**, 225-234.
- Lloyd, S. P. (1982). "Least squares quantization in PCM," *IEEE Transactions on Information Theory* **28**, 129-137.
- Lucy, L. B. (1974). "Iterative techniques for rectification of observed distributions," *Astronomical Journal* **79**, 745-754.
- Lyons, A. P., and Abraham, D. A. (1999). "Statistical characterization of high-frequency shallow-water seafloor backscatter," *Journal of the Acoustical Society of America* **106**, 1307-1315.
- MacLennan, D. N. (1990). "Acoustical Measurement of Fish Abundance," *Journal of the Acoustical Society of America* **87**, 1-15.
- Mayer, L., Li, Y. C., and Melvin, G. (2002). "3D visualization for pelagic fisheries research and assessment," *Ices Journal of Marine Science* **59**, 216-225.
- Mayer, L. A. (2006). "Frontiers in seafloor mapping and visualization," *Marine Geophysical Researches* **27**, 7-17.
- Medwin, H., and Clay, C. S. (1998). *Fundamentals of Acoustical Oceanography* (Academic Press, Boston).
- Melvin, G. D., Cochrane, N. A., and Li, Y. C. (2003). "Extraction and comparison of acoustic backscatter from a calibrated multi- and single-beam sonar," *Ices Journal of Marine Science* **60**, 669-677.
- Middleton, D. (1967). "A statistical theory of reverberation and similar first-order scattered fields, Part I: Waveforms and the general process," *IEEE Transactions on Information Theory* **13**, 372-392.
- Misund, O. A. (1997). "Underwater acoustics in marine fisheries and fisheries research," *Reviews in Fish Biology and Fisheries* **7**, 1-34.
- Misund, O. A., and Aglen, A. (1992). "Swimming behaviour of fish schools in the North Sea during acoustic surveying and pelagic trawl sampling," *Ices Journal of Marine Science* **49**, 325-334.

- Mitchell, N. C. (1996). "Processing and analysis of Simrad multibeam sonar data," *Marine Geophysical Researches* **18**, 729-739.
- Mitson, R. B. (1983). *Fisheries sonar* (Fishing News Books, Farnham, Surrey, UK).
- MMI (2007). "Marine Metadata Interoperability," (<http://marinemetadata.org/>).
- Molthen, R. C., Shankar, P. M., and Reid, J. M. (1995). "Characterization of Ultrasonic B-Scans Using Non-Rayleigh Statistics," *Ultrasound in Medicine and Biology* **21**, 161-170.
- Moore, R. E. (1985). *Computational functional analysis* (Ellis Horwood Ltd, Chichester, England).
- Muller, K. R., Mika, S., Ratsch, G., Tsuda, K., and Scholkopf, B. (2001). "An introduction to kernel-based learning algorithms," *Ieee Transactions on Neural Networks* **12**, 181-201.
- Myriax (2008). "Echoview on-line documentation," (Myriax Software Pty Ltd, Australia), p. <http://www.echoview.com/WebHelp/Echoview.htm>.
- Narici, L., and Beckenstein, E. (1997). "The Hahn-Banach theorem: the life and times," *Topology and its Applications* **77**, 193-211.
- Ng, A. Y., Jordan, M., and Weiss, Y. (2002). "On Spectral Clustering: Analysis and an algorithm," in *Neural Information Processing Systems (NIPS)*.
- Ng, R. T., and Han, J. W. (2002). "CLARANS: A method for clustering objects for spatial data mining," *Ieee Transactions on Knowledge and Data Engineering* **14**, 1003-1016.
- Noettestad, L., and Axelsen, B. E. (1999). "Herring schooling manoeuvres in response to killer whale attacks," *Canadian Journal of Zoology* **77**, 1540-1546.
- O'Dor, R. (2004). "A Census of Marine Life," *BioScience* **54**, 92-93.
- Ol'shevskii, V. V. (1967). *Characteristics of sea reverberation* (Consultant Bureau, New York).
- Paton, M., Neville, D., Calder, B., Smith, S., Reed, B., and Depner, J. (2003). "Area Based Processing and Visualization for Efficient Seafloor Mapping," in *U.S. Hydrographic Conference*.
- Pesavento, A., Ermert, H., Brol Zeitvogel, E., and Grifka, J. (1998). "High resolution imaging of generalized K-distribution parameters using maximum likelihood estimation for ultrasonic diagnosis of muscle after back surgery," in *IEEE Ultrasonics Symposium* (Sendai), pp. 1353-1356.
- Pitman, T. (2002). "Development of a multibeam sonar data logging system," in *Faculty of Engineering* (University of Tasmania, Hobart), p. 71.
- Pratson, L. F., and Edwards, M. H. (1996). "Introduction to advances in seafloor mapping using sidescan sonar and multibeam bathymetry data," *Marine Geophysical Researches* **18**, 601-605.
- Preston, J. M., Christney, A. C., Bloomer, S. F., and Beaudet, I. L. (2001). "Seabed classification of multibeam sonar images," in *MTS/IEEE Oceans '01* (Honolulu, USA), pp. 2616-2623.
- Rafaely, B. (2004). "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *Journal of the Acoustical Society of America* **116**, 2149-2157.
- Ramakrishnan, N., and Grama, A. Y. (2001). "Mining scientific data," in *Advances in Computers*, pp. 119-169.

- Ramakrishnan, R. (2003). "Data mining: Fast algorithms vs. fast results," in *Foundations of Intelligent Systems*, pp. 12-13.
- Reed, M., and Simon, B. (1980). *Functional analysis* (Academic Press).
- Reut, Z., Pace, N. G., and Heaton, M. J. P. (1985). "Computer classification of sea beds by sonar," *Nature* **314**, 426-428.
- Richardson, W. H. (1972). "Bayesian-based iterative method of image restoration," *Journal of the Optical Society of America* **62**, 55.
- Rudnick, P. (1969). "Digital beamforming in frequency domain," *Journal of the Acoustical Society of America* **46**, 1089.
- Sander, J., Ester, M., Kriegel, H. P., and Xu, X. W. (1998). "Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications," *Data Mining and Knowledge Discovery* **2**, 169-194.
- Scholkopf, B., Burges, C. J. C., and Smola, A. J. (1999). *Advances in kernel methods - support vector learning* (MIT Press, Cambridge, MA).
- Scholkopf, B., and Smola, A. J. (2001). "Learning with kernels, a tutorial introduction," in *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (The MIT Press).
- Scholkopf, B., Smola, A. J., and Muller, K. R. (1998). "Nonlinear component analysis as a kernel eigenvalue problem " *Neural Computation* **10**, 1299-1319.
- Schroeder, W., Martin, K., and Lorensen, B. (2006). *Visualization Toolkit: and object-oriented approach to 3D graphics* (Kitware).
- Shawe-Taylor, J., and Cristianini, N. (2004). *Kernel methods for pattern analysis* (Cambridge University Press).
- Simmonds, J., and MacLennan, D. N. (2005). *Fisheries acoustics: theory and practice* (Blackwell Science Ltd).
- Smyth, P. (2000). "Data mining: data analysis on a grand scale?."
- Smyth, P. (2001). "Data mining at the interface of computer science and statistics," in *Data Mining for Scientific and Engineering Applications*, edited by R. Grossman (Kluwer Academic Publishers).
- SonarData (2007). "Echoview on-line documentation," (SonarData Pty Ltd, Australia), p. <http://www.sonardata.com/WebHelp/Echoview.htm>.
- Soria, M., Freon, P., and Gerlotto, F. (1996). "Analysis of vessel influence on spatial behaviour of fish schools using a multi-beam sonar and consequences for biomass estimates by echo-sounders," *Ices Journal of Marine Science* **53**, 453-458.
- Starck, J. L., Pantin, E., and Murtagh, F. (2002). "Deconvolution in astronomy: A review," *Publications of the Astronomical Society of the Pacific* **114**, 1051-1069.
- Tang, Y., Iida, K., Mukai, T., and Nishimori, Y. (2006). "Estimation of fish school volume using omnidirectional multi-beam sonar: Scanning modes and algorithms," *Japanese Journal of Applied Physics Part 1-Regular Papers Brief Communications & Review Papers* **45**, 4868-4874.
- Tillett, R., McFarlane, N., and Lines, J. (2000). "Estimating dimensions of free-swimming fish using 3D point distribution models," *Computer Vision and Image Understanding* **79**, 123-141.
- Tsumuraya, F., Miura, N., and Baba, N. (1994). "Iterative blind deconvolution method using Lucy's algorithm," *Astronomy and Astrophysics* **282**, 699-708.

- UNEP (2007). *Global Environment Outlook 4 (GEO-4): Environment for Development* (United Nations Environment Programme).
- Urick, R. J. (1983). *Principles of underwater sound* (McGraw Hill, New York, NY).
- Van Hulle, M. (2004). "Data mining: numerical methods (course notes)," (Catholic University of Leuven, Leuven).
- Vapnik, V. (1995). *The nature of statistical learning theory* (Springer, New York).
- Verma, D., and Meila, M. (2003). "A comparison of spectral clustering algorithms," (University of Washington).
- von Luxburg, U. (2006). "A tutorial on spectral clustering," (Max Planck Institute for Biological Cybernetics, Tuebingen).
- Ware, C. (2004). *Information visualization: perception for design (2nd edition)* (Elsevier, San Francisco).
- Wei, C. P., Lee, Y. H., and Hsu, C. M. (2003). "Empirical comparison of fast partitioning-based clustering algorithms for large data sets," *Expert Systems with Applications* **24**, 351-363.
- Wilson, M. P., Higginbottom, I. R., and Buelens, B. (2005). "Four-dimensional visualization and analysis of water column data from multibeam echosounders and scanning sonars using Sonardata Echoview for fisheries applications," in *1st International Conference on Underwater Acoustic Measurements: Technologies & Results (UAM)* (Heraklion, Crete, Greece).
- Wu, L. X., Zielinski, A., and Bird, J. S. (1997). "Lossless compression of hydroacoustic image data," *Ieee Journal of Oceanic Engineering* **22**, 93-101.
- Xu, R., and Wunsch, D. (2005). "Survey of clustering algorithms," *Ieee Transactions on Neural Networks* **16**, 645-678.
- Yao, X. (2003). "Research Issues in Spatio-temporal Data Mining," (Department of Geography, University of Georgia).
- Yarincik, K., and O'Dor, R. (2005). "The Census of Marine Life: Goals, strategy & scope," *Scienta Marina* **69 (Suppl. 1)**, 201-208.
- Yip, A. M., Ding, C., and Chan, T. F. (2006). "Dynamic cluster formation using level set methods," *Ieee Transactions on Pattern Analysis and Machine Intelligence* **28**, 877-889.
- Zha, H., Ding, C., Gu, M., He, X., and Simon, H. (2001). "Spectral relaxation for k-means clustering " in *Neural Information Processing Systems (NIPS)*.
- Zhang, T., Ramakrishnan, R., and Livny, M. (1996). "BIRCH: An efficient data clustering method for very large databases," in *ACM SIGMOD Conference on Management of Data*, pp. 103-114.
- Ziomek, L. J. (1989). "3-Dimensional Ray Acoustics - New Expressions for the Amplitude, Eikonal, and Phase Functions," *Ieee Journal of Oceanic Engineering* **14**, 396-399.