NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Acoustic data optimisation for seabed mapping with visual and computational data mining

A dissertation submitted for the degree of
Doctor of Philosophy by

Kazi Ishtiak Ahmed

Maynooth, March, 2012

Primary adviser: Dr. Urška Demšar
Secondary Adviser: Professor A. Stewart Fotheringham

Examiners: Professor Kirsi Virrantaus
Dr. Alexei Pozdnoukhov

National Centre for Geocomputation
National University of Ireland, Maynooth
Ollscoil na hÉireann, Má Nuad
Co. Kildare, Ireland

StratAG
Strategic Research in Advanced Geotechnologies

NCG
National Centre for Geocomputation
An tIonad Náisiúnta Geoiríomhaireachta

sfi
science foundation ireland
fondúireacht eolaíochta éireann

Research presented in this paper was partially funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan.

GSI

INFOMAR
Integrated Mapping for the Sustainable Development of Ireland's Marine Resource

Marine Institute
Foras na Mara

লিলু, স্বাতী

এবং বাবা, আম্মু, মিতু ও বিনী

-----আমার পৃথিবী

Table of contents

List of figures

List of tables

List of acronyms

ADP	Ammonium dihydrogen phosphate
ANN	Artificial Neural Networks
CUWR	Columbia University's Division of War Research
DBI	Davies-Bouldin index
dB	Decible
DGPS	Differential Global Positioning System
DI	Dunn's index
FCM	Fuzzy c-means
FFT	Fast Fourier transforms
FFV	Full feature vector
GLCM	Grey-level co-occurrence matrix
GSI	Geological Survey of Ireland
GUI	Graphical User Interfaces
HUSL	University's Underwater Sound Laboratory
INFOMAR	Integrated Mapping for the Sustainable Development of Ireland's Marine Resource
LDA	Linear discriminant analysis
MBES	Multibeam echosounder
MI	Marine Institute
NRL	Naval Research Laboratory
PCA	Principal Component Analysis
PL	Propagation Loss
PZT	Lead Zirconate Titanate
SBES	Singlebeam echosounder
SI	Silhouette index
SOM	Self Organising Map
Sonar	Sound Navigation and Ranging
SSS	Side scan sonar
TB	Terabytes
TLR	Triangulation-Listening Ranging
TVG	Time varied gains
UCWR	California's Division of War Research
VRC	Calinski-Harabasz or Variance Ratio Criterion
WHOI	Woods Hole Oceanographic Institution

Abstract

Oceans cover 70% of Earth's surface but little is known about their waters. While the echosounders, often used for exploration of our oceans, have developed at a tremendous rate since the WWII, the methods used to analyse and interpret the data still remain the same. These methods are inefficient, time consuming, and often costly in dealing with the large data that modern echosounders produce. This PhD project will examine the complexity of the *de facto* seabed mapping technique by exploring and analysing acoustic data with a combination of data mining and visual analytic methods.

First we test the redundancy issues in multibeam echosounder (MBES) data by using the component plane visualisation of a Self Organising Map (SOM). A total of 16 visual groups were identified among the 132 statistical data descriptors. The optimised MBES dataset had 35 attributes from 16 visual groups and represented a 73% reduction in data dimensionality. A combined Principal Component Analysis (PCA) + k-means was used to cluster both the datasets. The cluster results were visually compared as well as internally validated using four different internal validation methods.

Next we tested two novel approaches in singlebeam echosounder (SBES) data processing and clustering – using visual exploration for outlier detection and direct clustering of time series echo returns. Visual exploration identified further outliers the automatic procedure was not able to find. The SBES data were then clustered directly. The internal validation indices suggested the optimal number of clusters to be three. This is consistent with the assumption that the SBES time series represented the subsurface classes of the seabed.

Next the SBES data were joined with the corresponding MBES data based on identification of the closest locations between MBES and SBES. Two algorithms, PCA + k-means and fuzzy c-means were tested and results visualised. From visual comparison, the cluster boundary appeared to have better definitions when compared to the clustered MBES data only. The results seem to indicate that adding SBES did in fact improve the boundary definitions.

Next the cluster results from the analysis chapters were validated against ground truth data using a confusion matrix and kappa coefficients. For MBES, the classes derived from optimised data yielded better accuracy compared to that of the original data. For SBES, direct clustering was able to provide a relatively reliable overview of the underlying classes in survey area. The combined MBES + SBES data provided by far the best accuracy for mapping with almost a 10% increase in overall accuracy compared to that of the original MBES data.

The results proved to be promising in optimising the acoustic data and improving the quality of seabed mapping. Furthermore, these approaches have the potential of significant time and cost saving in the seabed mapping process. Finally some future directions are recommended for the findings of this research project with the consideration that this could contribute to further development of seabed mapping problems at mapping agencies worldwide.

Preface

Acknowledgements

I want to express my deeply felt thanks to my thesis advisor, Dr. Urška Demšar; for her warm encouragement and thoughtful guidance. Her intellectual encouragement and 'disturb-at-will' pass have been instrumental in the completion of this thesis. I would also like to thank my second advisor, Professor A. Stewart Fotheringham for his valuable feedback and providing me with all the facilities I could hope for. Special thanks to Xavier Monteys from GSI for his support and feedback. I also thank the GSI for providing me the data for this research project.

I also wish to extend special thanks to Professor Kirsi Virrantaus from Aalto-yiliopisto, Finland and Dr. Alexei Pozdnoukhov for examining the finished thesis.

I would like to thank Ann-Marie Burke and Melina Lawless for taking care of all the bureaucratic matters in the most efficient manner. You are amazing! To Dr. Conor McElhinney for helping me decipher the Matlab colour management codes, Dr. Paul Harris, Dr. Paul Lewis and Rónán O'Braonáin for their help with both personal and academic matters. I would also like to thank Helen and Dr. Peter Hung for their help at different stages. Conor for the funny posts, cheered me up on numerous occasions. Lastly, thanks to Ambra, Pankaj, Maryam, Ishwari and Binbin for patiently letting me vent my frustrations at times.

On a more private level, I would like to thank Dr. Hans Hauska from KTH, Sweden. He was the first academic to encourage me to do a doctoral research. He has enlightened me through his wide knowledge of GIS and remote sensing but more so through his humility and a person I will always look up to, in all aspects of life.

Finally, I am forever indebted to my parents and family for their continued love and support. Words are inadequate to express the joy and love I feel for my daughter Lilou and my wife. Shatee, thank you for being the amazing person you are, for your love and never ending support and patience with me and my many quirks.

Publications

From doctoral work

Examining statistical segmentation of multibeam backscatter images with geovisual analytics. **Ahmed, K.I**., Demšar, U., Monteys, X. Proceedings of the 3rd ICA Workshop on Geospatial Analysis and Modeling. Gävle, Sweden, 2009.

From other work

Comparative analysis of clustering methods of Multibeam Echo Sounder (MBES) backscatter data. **Ahmed, K.I**., Hung, P., Harris, P., Caughey, H. M., McLoone, S., Demšar, U., Fotheringham, A.S., Monteys, X. Proceedings of GIScience, 2010, Zurich, Switzerland.

Developing statistical methodology to improve classification and mapping of seabed type from deep water Multi-beam Echo Sounder (MBES) data. Caughey, H.M., **Ahmed, K.I.**, Harris, P., Hung, P., Demšar, U., McLoone, S., Fotheringham, A.S., Monteys, X., O'Toole, R. Proceedings of GISRUK 2010, London, UK, 2010.

Developing a statistical methodology to improve classification and mapping of seabed type from deep water multi-beam echo sounder. **Ahmed, K.I**., Caughey, H. M., Harris, P., Hung, P., Demšar, U., Fotheringham, A.S., McLoone, S., Monteys, X., O'Toole, R. Technical report submitted to Geological Survey of Ireland (GSI), Ireland, 2011.

Chapter 1
Introduction

In this first chapter we briefly establish the context for the work developed in this thesis and give an overview of its structure.

> *It was during and after the Second World War that the great expansion [in oceanography], which is still going on, began. The realization by governments of the importance of marine problems and their readiness to make money available for research, the growth in the number of scientists at work and the increasing sophistication of scientific equipment, have made it feasible to study the ocean on a scale and to a degree of complexity never attempted and never possible before. ... As man increasingly overcrowds and exploits his tiny planet, the significance of the oceans which cover seven tenths of its surface has suddenly become apparent.*
>
> *--Margaret Deacon "Scientists and the Sea 1650-1900" 1971*

Ocean covers 70% of Earth's surface yet our understanding of its waters to date is quite minimal. Historically, the main reason for this was the unavailability of equipment for ocean exploration. Since the Second World War, with the realisation of the importance of ocean exploration, a large portion of research has focused on the technological development of underwater exploration equipment. A direct result of that research is today's sophisticated high-resolution acoustic echosounders. But not as much effort has been given on updating the methods used to process acoustic data collected by these advanced sensors. To this day, the acoustic data processing methods still stem from those used in the World War II era. These methods are not efficient in dealing with large volume of high resolution acoustic data produced by the modern sensors and as a result, creating quality seabed maps from these data is difficult, time consuming, and in some cases costly. This drawback in acoustic data processing is the motivation for the alternative approaches proposed in this thesis.

The requirement for good quality seabed maps has risen sharply in recent decades for a variety of reasons, such as environmental research, management of marine and coastal resources and oil and gas exploration. Acoustic sonar systems, which transmit and receive an acoustic pulse from a device on a survey vessel, are typically used for this purpose. Seabed survey data usually contain the travel time of the acoustic pulse to the sea floor and back and the strength of the signals. Various measurements such as the depth to the seafloor (bathymetry), depth to sub-surface sediment layers (sub-bottom), and the reflectance of the sea floor (intensity of backscattered energy) are usually derived from these data. Among all the data collected, acoustic backscatter data are directly connected to the sediment characteristics of the seabed (Brown & Blondel, 2009; Goff et al., 2004; Huges Clarke et al., 1997) and are often used for classification of seabed type.

Acoustic data are acquired using echosounders. Two types of echosounders, which are commonly used for seabed mapping, are multibeam echosounders (MBES) and side scan sonars (SSS). In the recent years, survey data from singlebeam echosounders (SBES) are also being used for seabed classification due to their high resolution and sub-surface penetration ability. There are a number of automated classification methodologies, most of them using image segmentation techniques (for MBES and SSS) or direct feature extraction techniques (for SBES) (Preston, 2009; Preston et al., 2004; Satyanarayana et al., 2007; Zimmermann & Rooper, 2008).

Ultimately, a successful classification depends heavily on the quality of the data. In the case of seabed mapping, the data generated from seabed surveys are extremely large in volume with high degree of noise. Statistical features, which act as descriptors of the backscatter data, are usually generated in an effort to reduce the data volume, thus reducing the effect of noise as well as computational intensity. There are several complexity issues with this method, which is commonly used in MBES and SSS backscatter image segmentation as well as with SBES data. According to Preston (2009), while the high number of statistical features is calculated because of the wide diversity of MBES/SSS images, many of the statistical features are highly correlated and therefore most likely redundant, i.e. they do not contribute any new information about acoustic similarity to the process. Because the resulting feature space is so highly-dimensional (i.e. 132 dimensions for MBES), it is “convenient” (Preston, 2009) to use a dimensionality reduction technique to facilitate interpretation of similarity patterns. On the other hand, SBES backscatter returns are univariate time series in nature and contain a peak when the emitted echo hits the seabed. Standard statistical descriptors (mean, standard deviation etc.), which almost always form part of the statistical feature set that is generated from SBES backscatter, may not always capture the ‘true’ mean or its variation due to the nature of the data. Other features (randomness, correlation noise etc) generated from SBES also depend on the standard statistical descriptors for their own generation and thus would also include the systematic noise. Before extraction of any statistical features, the time series data are filtered through various outlier detection algorithms. As most of these algorithms were not optimised to detect outliers in sonar time series data, there is always a possibility that some outliers can go undetected and thus affect the quality of classification. In addition to the noise issues, the traditional clustering method, a combined Principal Component Analysis (PCA) and k-means, was developed in the post-Second World War era. At that time, the computers lacked the capability of dealing with large volume of acoustic datasets and therefore their dimensionality was reduced by using PCA and selecting the first three components. The three components typically account for between 90-95% of the information. This method includes the risk of missing small but interesting areas when the survey area is large as a result of omission of 5-10% of the information that might include that area.

The noise issues stemming from data redundancy (MBES) and outliers (SBES) as well as the risk of information loss with the de facto PCA/k-means clustering methods form the core research interests in this study. The primary objective of this thesis is to provide a set of methods that enables acoustic data optimisation to reduce the noise effects arising from data redundancy and outliers as well as to test alternate clustering methods that can improve the seabed mapping quality. The following core research questions, which rise from the primary objectives of this research, are addressed in four analysis chapters:

1. Redundancy in feature data for MBES: Is it necessary for MBES classification to produce such a large number of statistical features or could the dimensionality be kept lower by avoiding redundancy? And if so, which of the features are correlated with each other and therefore redundant? Would clustering a optimal dataset (redundancy is removed) produce better, similar or worse cluster definition to that of the cluster definition generated from 132 statistical features using the de facto clustering method?
2. Clustering of SBES time series: Can visual analytics provide an efficient way of detecting outliers that are undetected using traditional outlier detection methods?
3. Would direct clustering of the SBES backscatter produce representative clusters thus eliminating the dependency of generating features as well as provide a quick overview of underlying clusters in the survey area?
4. Improving seabed mapping from MBES and SBES: Can the optimised data produce quality clusters that will ultimately result in better quality seabed classes? Seabed surfaces can have plants, shells, swimming fish etc all of which can contribute to the noise in MBES data. Can the sub-surface information of SBES be combined with MBES data to provide a better definition of the underlying seabed, thus avoiding the interference from plants and other particles lying on seabed?

A typical high-resolution sensor results in terabytes of survey data. A typical MBES feature extraction for a medium to large survey area takes days, often weeks to process. An optimisation procedure for redundancy reduction of MBES features can reduce this time to a significant level thus saving cost. A successful direct clustering of SBES would mean that surveyors can use this to get an initial idea of the underlying seabed clusters and optimise their ground truth collection points thus reducing additional survey runs for ground truth collection. This can also be much cost saving. A successful classification of combined dataset (MBES + SBES) can result in a better quality seabed map thus enabling researchers and investors to explore and analyse the seabed geology more efficiently. These factors motivate the central focus of this thesis, which is to implement Visual Analytics in improving the sonar datasets in an effort to produce better quality seabed mapping thus reducing cost and time in the long run. The core objectives of this thesis are:

- To explore the two types of datasets, MBES and SBES, using a Visual Analytical approach.
- Using Visual Analytics to detect redundancy in MBES features and to determine the optimal number of features required for clustering and subsequent classification.
- To use visual exploration to detect outliers in time series SBES data and thus optimise the dataset for clustering.
- To produce a classification map from optimised MBES and SBES datasets. Focus will also be given to the potential for combining SBES and MBES data to evaluate if subsurface information from SBES can contribute to better classification of the seabed.

This work will examine the complexity of the traditional acoustic data mining method through a Visual Analytical approach by exploring the statistical features and backscatter time series data with a combination of data mining, knowledge discovery, and visualization methods. The ultimate goal of the project is to examine the potential of Visual Analytics to help reduce the complexity of the echosounder image classification methods and thereby facilitate seabed type characterization. The research components of this thesis are framed around the central research questions, which form the basis of the analytical chapters (chapters three to six). The rest of the thesis is structured as follows:

Chapter two constitutes a review of the central theoretical and methodological issues that are built upon in the three analytical chapters. This chapter discusses the development of echosounders from a historical perspective, their working principal, as well as different aspects of seabed mapping. This chapter also briefly discusses the concept of Visual Analytics and visual data mining as well as their potential implementation in seabed mapping.

Chapter three assesses attribute redundancy of statistical feature data produced by standard automatic methods for classification of multibeam echosounder (MBES). A Self Organising Map (SOM) was used on the 132-dimensional statistical feature space (also known as Full Feature Vector or FFV). Once the SOM was trained, the result was displayed in the 132 component planes (one for each feature vector), where each plane was coloured according to the FFV that it represented. These planes were then examined for similar colour distribution patterns that defined visual groups of FFVs. These patterns indicate that FFVs in each of these visual groups are probably correlated. Correlation is further confirmed with a subsequent statistical analysis. We further evaluate classification of the original feature data with the one produced from an optimised feature set and discuss potential improvement for seabed mapping.

Chapter four explores the potential of direct clustering of SBES data. The SBES data volume is much less than that of MBES as it comprises of the return of one emitted beam. However, acoustic backscatter is a complex function of many factors (frequency, seabed slope, grain size, presence of flora & fauna etc.). One other alternative can be to apply direct clustering algorithm to the echo time series segments that include the peak curve (echoes hitting seabed surface) and sub-surface information. The central focus of this chapter is to use visual exploration technique to detect outlier that may still exist after the data have been filtered through an outlier detection algorithm. TimeSearcher© tool was used to visually explore the SBES dataset and after optimisation of the data, fuzzy c-means was used to cluster the dataset. This chapter outlines the results obtained from visual exploration and subsequent fuzzy clustering.

Chapter five focuses on the potential of the clustering of combined MBES and SBES data and compares the results from this combined dataset to that of MBES classification. The central focus of this chapter is to evaluate the potential of SBES features contributing to the improvement of seabed maps otherwise obtained using only MBES data. As the SBES contains subsurface information, it is possible that the SBES features can help to better define the MBES classes as MBES data cannot penetrate the seabed and can contain false scatters from seabed vegetation, schools of fish swimming at the bottom etc.

The results from MBES, SBES and combined MBES and SBES are validated and outlined in Chapter six. This chapter mainly focuses on comparison of results obtained in Chapters 3, 4, and 5. Examination of the quality of the clustering results obtained from MBES, SBES, and MBES + SBES clustering by labelling the clusters with appropriate class definitions and comparing the results with ground truth data collected from the survey area.

Chapter seven provides a summary of the results from each of the analysis chapters and their significance in the context of seabed classification from acoustic data. Their novelty in a wider methodological context is also discussed.

Chapter 2

Literature review

This second chapter reviews the relevant literature and puts the thesis into the context of existing research. In particular, we discuss the development of sonar systems and its applications, seabed mapping with sonars and challenges associated with sonar data, data mining, visual analytics and clustering of sonar data.

Chapter contents

Man’s concern with the depths of water began as soon as he mastered the ability to travel on the water. It was vital information for the primitive sailors to prevent their boats from running aground. ‘Lead lines’ have been used for the measurement of ocean depths for thousands of years. The earliest record of lead lines dates back to 2000BC. The boat model found in the tomb of Meketre in Thebes

(MMA), 2011) shows a sailor poised with what looks like a lead line on the bow. For about 4000 years, though improved through mechanization, the technology behind ocean's depth measurement remained the same (Mayer, 2006).

Other developments such as electromagnetic waves, successfully used in environmental monitoring and exploration, could not be used for seabed surveys. Water has strong conductivity and is highly dissipative, thus rendering the electromagnetic wave useless due to rapid attenuation (Bass, 1972; Lurton, 2002; Mayer, 2006). Sound, on the other hand, has better transmission in water. Its propagation in water is four to five times higher than in air. Sound waves can reach higher levels and with less attenuation and can propagate over large distances. Despite these favourable characteristics, water still brings a number of limitations to the acoustic waves.

Attenuation of signals due to the absorption of sound waves in water limits their ranges to some degree. The propagation speed is also very low (1,500 m/s) compared to that of electromagnetic waves in space (300,000 km/s). The variation in sound speed and reflection on seafloor and sea surface interfaces causes perturbations of the propagation. This results in inhomogeneous insonification of the propagation medium and delayed echoes also known as 'multiple paths'. The heterogeneity of the medium, reflection of the sound waves on seafloor and sea surface, and frequency changes (Doppler Effect) due to the relative movement of sonars and targets deforms the transmitted signal. In addition to the deformation of signals, noise is added to the echoes from the ambient noise in the ocean coming from the movements of the sea surface, volcanic and seismic activity, shipping, living organisms, rain as well as the self-noise characteristic of the acoustic system and its platform (surface vessel or submarine). The characteristics of water vary in space and time. These fluctuations depend on, but are not limited to, geographical and seasonal variations in temperature and salinity, seabed relief, swell, currents, tides, internal waves. All these factors give the underwater acoustic signals a mostly random fluctuating character (Lurton, 2002; Mayer, 2006).

The following sections discuss the theoretical background of underwater acoustics, how the sound propagates and how it is measured. Then the sonar systems, which use underwater acoustics for navigation, exploration and mapping, will be discussed. These sections will focus on the development of sonar systems from a historical perspective containing a brief outline of their applications and the main focus being how they are used in the seabed mapping. The sections that follow will focus on data mining and visualisation. These sections will also include a review of visual data mining and visual analytics as well as their potential as an alternative data mining technique for sonar datasets.

2.1. Propagation of sound in water

Sound faces several constraints and undergoes multiple transformations while travelling through water. The speed of sound depends heavily on the medium. For example, the speed of sound is approximately 341 m/s in air at a temperature of 18°C. In contrast, that speed increases to around 1,524 m/s in salt water at approximately the same temperature. To understand how sound travels in water, one needs to understand the fundamental notions associated with the physical nature of acoustic waves (Lurton, 2002; Waite, 2002).

The amplitude of the sound signal decreases as the wave propagates through water. This is one of the first effects of propagation and is due to both geometrical effects and absorption. The later is directly linked with the chemical properties of the sea water (Lurton, 2002; Waite, 2002). An acoustic wave originates from the oscillation of pressure (acoustic pressure) and travels through an elastic medium (solid, gas or liquid). The mechanical property (density ρ and elasticity modulus E) of the medium dictates the propagation speed or velocity. The elasticity modulus quantifies the relative variation of volume and density due to pressure variation and as water is less compressible than air, acoustic velocity is much higher in water than in air. The velocity, c of sound in water can be described by the following equation (Lurton, 2002; Urick, 1982):

$$c = \sqrt{\frac{E}{\rho}} \tag{2.1}$$

The velocity of sound in water is affected by the oceanographic variables of temperature, salinity, and pressure. The velocity increases with increasing water temperature, increasing salinity and increasing pressure or depth (Figure 2.1).

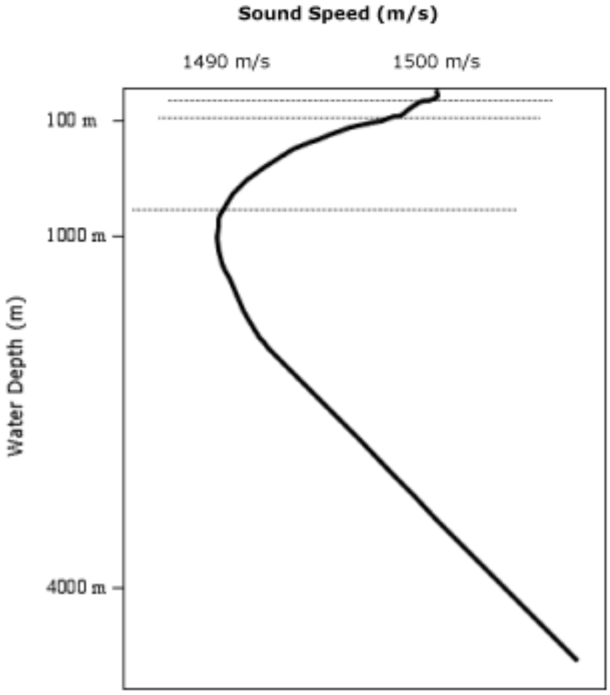


Figure 2.1: Profile of speed of sound in water (adapted from DOSITS, 2011)

Acoustic waves are also characterized by their frequency f. Frequency is the number of vibrations per second and is expressed in Hz. The pressure variation in a sound wave repeats itself in space over a specific distance. This distance is known as the wavelength of the sound, usually measured in meters and represented by λ. As the wave propagates, one full wavelength takes a certain time period to pass a specific point in space; this period, represented by T, is usually measured in fractions of a second. The relation between sound velocity and wavelength, period and frequency is given by the equation below:

$$\lambda = cT = \frac{c}{f} \tag{2.2}$$

The frequencies in underwater acoustics vary from 10 Hz to 1 MHz. The corresponding wavelengths would be around 150 meters and 0.0015 meter (DOSITS, 2011; Lurton, 2002; Urick, 1982). These diverse values of frequency and wavelength correspond to the variation of physical processes of propagation medium as well as the acoustic system itself. The selection of a frequency for a particular application directly depends on the following:

– The effect of dampening of sound wave in water. This is inversely proportional to the frequency. It limits the maximum range that is usable under the circumstances.

– The size of the acoustic sound source. For a given transmission power, it increases with lower frequencies.

– The directivity of the acoustic sources and receivers. It improves the spatial selectivity as frequency increases.

– Physical properties of the target also have a direct influence on the frequency. A smaller target will reflect less energy back to the receiver.

All these constraints are taken into account when a frequency is chosen for a particular application (Lurton, 2002).

2.1.1. Measurement of underwater acoustics: logarithmic notation

The intensity of a sound wave is the average rate of flow of energy per unit area perpendicular to the direction of propagation. Power, measured in watts, is the amount of energy per unit time and intensity is therefore measured in watts per square meter. Sound intensity is often specified as a logarithm of the ratio of a sound's intensity to reference intensity. This is often called the "Bel" in honour of Alexander Graham Bell, the inventor of the telephone. The human ear is very sensitive and can detect changes in relative intensity of as little as 1/10 of a Bel (a decibel is 1/10 of a Bel). For that reason, relative sound intensities are often reported in decibels (dB). The decibel is a relative unit, not an absolute one. The relative intensity, I, in decibels, is calculated as the ratio of the intensity of a sound wave to reference intensity:

$$I = 10\log\left(\frac{I_{sound}}{I_{reference}}\right) dB \qquad (2.3)$$

Acoustic intensity is rarely measured directly. Underwater microphones (hydrophones) measure the pressure (amplitude) of a sound wave rather than its intensity. Intensity of a sound wave is proportional to the square of its pressure p (Lurton, 2002; Urick, 1982):

$$I = \frac{p^2}{2\rho c}\left(\frac{watts}{m^2}\right) \tag{2.4}$$

The intensity in dB can be computed directly from the measured pressure:

$$I(dB) = 10\log\left(\frac{p_{sound}^2}{p_{reference}^2}\right) = 20log\left(\frac{p_{sound}}{p_{reference}}\right) \tag{2.5}$$

To be able to compare relative intensities given in dB to one another, a standard reference intensity or reference pressure is always used. For this reason, sound levels expressed in decibels include a reference pressure. It is common practice to use 1 microPascal (µPa) as the reference pressure and 1 W/m2 as the reference intensity for underwater sound (DOSITS, 2011; Urick, 1982). The logarithmic nature of the dB scale means that each 10 dB increase is a ten-fold increase in acoustic power. A 20-dB increase is then a 100-fold increase in power and so on.

2.1.2. Propagation loss in underwater acoustics

The ‘propagation loss’ or the ‘transmission loss’ is an important phenomenon for acoustic systems. Propagation loss (PL) occurs mainly in two ways: loss of intensity due to geometric spreading and absorption of energy due to the chemical characteristics of the medium itself. This is an important parameter as it constraints the amplitude of the signal and therefore, the receiver’s performance depends a great deal on signal-to-noise ratio.

Geometric spreading is the process where the acoustic intensity reduces due to the spreading of the sound wave as it propagates from the source to a larger surface. Water is regarded as a dissipative propagation medium. Part of the energy from the transmitted wave is absorbed and is dissipated through viscosity or molecular relaxation. The effect of viscosity can be observed in both fresh and salt water. Molecular relaxation is the reduction of molecules to ions induced by the local pressure variation of the sound. At high frequencies (≥500 kHz), the variation of pressure is too rapid for the relaxation mechanism to take effect (i.e. molecules to recompose themselves) and as a result the energy is not absorbed and permanently dissipates (Lurton, 2002; Waite, 2002).

Formation of air bubbles at the sea surface can affect the acoustic characteristics (velocity, attenuation etc.) of the propagation medium. Air bubbles are usually created by sea surface movements and/or by boat movement. The effect of air bubbles on acoustic attenuation is mostly local and decreases with depth. Its effect is neglected when the depth is below 10 to 20m (Lurton, 2002).

Other phenomena that contribute to underwater acoustic attenuation are refraction, scattering, and the presence of ocean boundaries. These ensure that free-field (free of any interference) conditions are practically non-existent in the real world and therefore require clear understanding in order to develop proper speed profiles for underwater surveys.

2.2. Sonar systems

Though these characteristics of the sound waves in water were discovered quite early, the actual use of underwater acoustics is fairly recent. The first breakthrough came in the early 1900s with the development of piezoelectric crystals or ceramic based transducers capable of generating and receiving sound waves (Lurton, 2002; Mayer, 2006).

Acoustic echosounders, also known as sonars (SOund Navigation And Ranging) use acoustic signals for target and obstacle detection. They do so by either receiving the echo transmitted by the system and sent back by the target (active sonar), or by receiving the acoustic noise directly radiated by the target (passive sonar).

Typical active sonar uses an emitter to transmit high-power acoustic signals. The signal is then reflected by the target and this reflected echo is received by an antenna (or an array of transducers) against a background of noise and reverberation (unwanted echoes from the sea surface and sea bed and from scatters within the volume of the sea). This signal is then processed and lastly used for measurement and identification or characterisation of the target. The range of a target can be calculated from measuring the time between transmission of a pulse and reception of an echo. Active sonars are sometimes known as echo ranging systems (Preston et al., 2001; Waite, 2002).

Passive sonars, on the other hand, detect the signal from the sound radiated from a target using a hydrophone (an underwater microphone) against a background of the ambient noise of the sea and the self-noise of the sonar platform. These systems classify the target from the analysis of the frequency spectrum of the signal and its variation in time (Waite, 2002).

Though the development of echosounders is fairly recent, the use of underwater acoustics for depth measurement and navigation dates centuries back. The following section gives a brief overview of the development of acoustics and echosounders in the monitoring of depth, navigation and seabed.

2.3. A historical overview of acoustics and echosounders

Aristotle (384–322 BC) was among the first to assert that one can hear sounds in water as well as in air. Nearly 2000 years later, Leonardo Da Vinci (1452-1519) made the observation - "If you cause your ship to stop and place the head of a long tube in the water and place the outer extremity to your ear, you will hear ships at a great distance from you." A significant advance in the physical understanding of acoustical process came with Marin Mersenne and Galileo independently discovering the laws of vibrating strings. Mersenne published his work on the nature and behaviour of sound in L'Harmonie Universelle in the late 1620's. This work and his later experimental measurements on the speed of sound in air provided the foundation for acoustics (Allaby, 2009).

The next advancement in acoustics came in 1687 when Sir Isaac Newton published the first mathematical theory of how sound moves, in his famous Philosophiae Naturalis Principia Mathematica. Although Newton focused on sound in air, the same basic mathematical theory applies to sound in water (Press, 2011).

There were, however, reservations on whether sound could travel through water. In 1743, Abbé J. A. Nollet conducted a series of experiments to settle that dispute. With his head underwater, he reported hearing a pistol shot, bell, whistle, and shouts. He also noted that an alarm clock clanging in water could be heard easily by an underwater observer, but not in air. His experiments were one of the first demonstrations of sound motion in water (DOSITS, 2011).

2.3.1. Studies of underwater acoustic in the 1800s

The first record of successful measurements of the speed of sound in water dates back to the early 1800s. Using a long tube, scientists Jean-Daniel Colladon, a physicist, and Charles-Francois Sturm, a mathematician, recorded how fast the sound of a submerged bell travelled across Lake Geneva in 1826 (Figure 2.2).

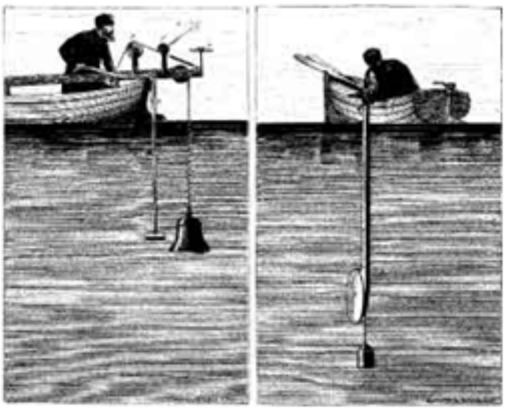


Figure 2.2: Illustration of J. D. Colladons' experiment in Lake Geneva, Souvenirs et Memoires, Albert-Schuchardt, Geneva (DOSITS, 2011)

Charles Bonnycastle performed the first documented echo sounding experiments in 1838. Lt. Matthew Fontaine Maury, commander of the U.S. Navy Depot of Charts and Instruments, attempted to use sound to measure the depth of the ocean in 1859. His experiments were unsuccessful, as he did not use an underwater receiver to listen for the echo. 1877 and 1878 can be regarded as a major breakthrough years for acoustics. The British scientist John William Strut, also known as Lord Rayleigh, published 'The Theory of Sound' in two volumes. This arguably marked the beginning of the modern study of acoustics. Lord Rayleigh was the first to formulate the wave equation that formed the basis for all work on acoustics. His ground breaking work set the stage for the development of the science and application of underwater acoustics in the twentieth century (Allaby, 2009; Press, 2011).

In the late 1800's, managing the navigation challenges of ever increasing ship traffic was a major concern. The lights and the loud sirens of the lighthouses and lightships did not travel far enough to warn ships about the dangers of shallow waters and rocks. In 1889, the American Lighthouse Board mentioned a combination of underwater bell and microphone system devised by Lucien Blake as an alternative to the traditional warning method (Allaby, 2009).

2.3.2. The 20th century advances (pre WWI)

In 1901, a group of scientists formed a company called the “Submarine Signal Company” based on a common belief that underwater sound could be used to develop a more reliable warning system to the increasing ship traffic. They developed an instrument that comprised an underwater bell located under the light ship or near lighthouses that could be detected by receivers installed on ships. To receive the bell signals, ships used a similar carbon-granule microphone developed by Thomas Edison. The microphone was put in a waterproof container, serving as a hydrophone. In doing so, the company may have, perhaps, applied the first practical use of underwater acoustics (SSC, 1907; DOSITS, 2011; Lurton, 2002).

Unfortunately, the ship-mounted hydrophones also picked up background noise, including ship machinery, splashing water, and fish, which made it difficult to hear the sounds from the bells. In mid-April 1912, Reginald A. Fessenden, a consulting engineer, was asked to redesign the hydrophones to filter out such noise. In conjunction of redesigning the hydrophones, Fessenden also suggested that the sources (bells) be improved instead. He proposed replacing the bells with louder, electric-powered sound generators designed to produce an audible tone (Allaby, 2009; DOSITS, 2011).

The development of active acoustic systems gathered pace after the unfortunate loss of the Titanic in 1912. Within a week of the ship’s tragic collision with an iceberg, L. R. Richardson filed a patent for an invention called echo ranging that used sound and its echoes off objects to determine distances in air. A month later, he filed a patent application for doing the same thing underwater. However, at this time, an appropriate acoustic source still did not exist (SSC, 1907; Lurton, 2002).

Fessenden, while working as a consultant for the Submarine Signal Company, designed an echo ranging device around the same time resembling a high-powered underwater loud speaker. It was capable of both producing and detecting sounds and was later called the “Fessenden Oscillator” (Figure 2.3).

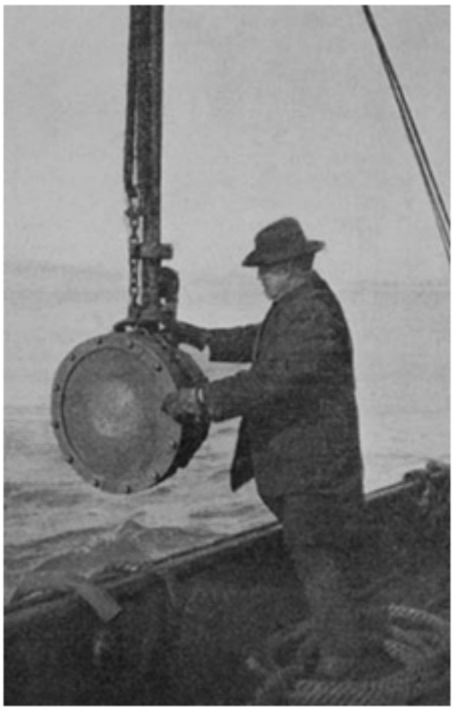


Figure 2.3: Reginald Fessenden and the Fessenden Oscillator. In "Submarine Signalling," Scientific American Supplement, No. 2071, pp. 168-170, Sept. 11, 1915. Image courtesy of NOAA Photo Library

In 1914, Fessenden conducted his first echo-ranging trials with the oscillator. The goal of his trials was detection of seafloor and icebergs. He was able to accurately detect the seafloor at a depth of 31 fathoms (57 meters approx). He also successfully used his oscillator to detect an iceberg that was approximately around 40 meters high, 137 meters long from a distance of about 3.2 kilometers. Despite these encouraging results his oscillator was not put into commercial production before 1923. Following the World War I, the company started marketing a low-frequency echosounder and called it “fathometer” as depth was in fathoms. By the mid-1930's, practically every submarine used an underwater acoustic system adapted from Fessenden Oscillator (Allaby, 2009).

2.3.3. Development during the world wars

The first efficient passive detection devices were developed during the First World War by the Allies to address the threat of German submarines. The major breakthrough came from a French physicist, Paul Langevin. Between 1915 and 1918, He experimented on river Seine and at sea to show that it was possible to transmit signals to detect submarines, giving both their angles and distance.

The use of submarines and underwater mines in WWI greatly influenced the development of underwater acoustics. German submarines targeted shipping between the United States and Europe. Explosions from contact mines suspended on underwater cables also took their toll (Allaby, 2009; DOSITS, 2011). Nearly 10 million tons of cargo was sunk in two years and had a crippling effect on the U.S. and European Allied Forces' supply lines. A total of 146 war vessels (including 40 submarines), 267 auxiliary vessels, and 586 merchant ships were sunk due to mines by both German and Allied forces were (Lasky, 1974, 1975, 1977).

The effectiveness of submarines and underwater mines in naval warfare was undisputable and the Allied forces needed an effective system to address this threat. This led to an increased interest of the military in underwater acoustics which subsequently became closely associated with military applications and research (Allaby, 2009; Lasky, 1975, 1977).

At that time, submarines were detected by listening for their engines or propellers. The sonar operator wore a two-earphone device and mechanically rotated the receiver to determine the direction of the sound (Figure 2.4). A number of different towed receivers were also developed for use by surface ships, in an attempt to reduce noise generated from the ships by putting the hydrophones further. But this approach was largely inefficient (Lasky, 1977).

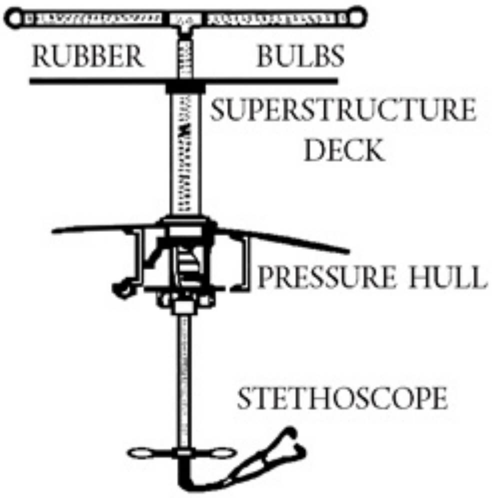


Figure 2.4: World War I Type SE-4214 (SC) sound receiver as it was installed on a U.S. submarine (Lasky, 1977)

Paul Langevin, a French physicist, experimented with the ‘piezoelectric effect’ discovered by Paul-Jacques and Pierre Curie in 1880, to build an echo-ranging system between 1915 and 1918. When a changing voltage is applied to a crystal at the desired frequency, they expand and contract. This generates a sound wave and this property of the crystal is called the ‘piezoelectric effect’. His developed a device made of quartz crystals placed between two steel plates to generate sound and tested it on River Seine and at sea to show that it was possible to transmit signals to detect submarines, giving both their angles and distance. However, his breakthrough came too late to be implemented during World War I (Lurton, 2002).

2.3.4. Between wars: Non-military development

During the time between WWI and WWII, scientists concentrated their focus on fundamental concepts of underwater sound propagation and on exploration of ocean and its inhabitants. A significant discovery during this time was by H. Lichte, a German scientist, who developed the theory on refraction of sound waves in seawater. Lichte theorized in his 1919 paper that sound waves are refracted when they encounter slight changes in temperature, salinity, and pressure. He used existing static measurements on seawater to compute the velocity of sound in terms of its determining variables. From a number of field studies in a variety of shallow sea water areas, he concluded that like sound propagation up-wind and down-wind in air, sound ranges should be better in winter than in summer (Lichte, 1919).

Following the WWI, echosounders became commercially available. They were already used extensively for helping ships avoid running aground in shallow water. Their wider availability immensely enriched our understanding of seafloor structure in the deep sea. In 1922, echosounders were used to determine a suitable underwater telegraph route between Marseilles, France, and Philippeville, Algeria. This is regarded as one of the first civilian application of echosounders (DOSITS, 2011).

Another significant discovery during this time is the ability of low frequency sound waves to penetrate into the seafloor and that sound reflected differently from individual layers in the sediment. This enabled the scientists, for the first time; to profile the sub-bottom layers beneath the seafloor. This proved to be vital in the studies of the history of the Earth and for prospecting for oil and gas under the seafloor. Pioneering work was done by Maurice Ewing and Allyn Vine, Bracket Hersey, and Sidney Knott at the Woods Hole Oceanographic Institution (WHOI) (DOSITS, 2011; Worzel, 1994). Ewing, Vine, and Joe Worzel experimented with homemade bombs by exploding them on the seafloor and recording the echoes. In 1934, they developed one of the earliest seismic recorders designed to receive sound signals on the seafloor (Hersey, 1977).

Use of acoustics in fisheries also took off during this time. The possibility of detecting echoes from schools of sardine and herring was suggested in 1924 by P. Portier in France (DOSITS, 2011). The first successful published experiment in detecting fish by acoustic means was done by Kimura, K. in 1929 in Japan. He placed a transmitter and a receiver near the two ends of a pond in such a way that the reflections were well received. He observed that each time a fish crossed the acoustic path; there was a fluctuation in the wave amplitude (Kimura, 1929).

2.3.5. Between wars: Military developments

After the end of the WWI, with the threat of Germany removed, there was a significant cutback in funding acoustic research in UK and USA. The general consensus was that the noise from the submarines could be quieted to such a degree that it would render the passive detectors useless. So, more focus was given to the development of echo ranging systems to detect submarines and measure the range and direction to them (Namorato, 2000).

During the interwar years, England actively worked on improving their transducers’ capability. But as they believed that the threat from submarine warfare was not significant, their development was not quite at the same level as that of USA. The British Anti-Submarine Detection and Investigation Committee (its acronym, ASDIC, became a name commonly applied to British SONAR systems), by 1939, developed transducers that were of quartz-steel and considered their day time sonar capabilities very satisfactory. It is largely believed that the British only signed the Anglo-German Naval agreement as they were confident with their ASDIC transducer (Lasky, 1977; Press, 2011).

Germany had a different approach to their acoustical research. Though Germany was constrained by the Treaty of Versailles from rebuilding the submarine force, the realities were somewhat different. In 1922, they set up a company in Holland called IVS Limited. It was staffed by a German naval construction group which continued their work on the U-boat. By 1930, German submarines were being built, tested, and even sold. After the Anglo-German Naval Agreement in 1935 which allowed Germany to rebuild their navy and submarine force, Germany not only stockpiled submarines but also significantly improved the passive sonar systems. The most significant was 'The Gruppen Horch Gerat (GHG)' system that was custom built to fit different type of ships and submarines. The GHG comprised of large arrays of 108 to 120 hydrophones arranged in an elliptical or semicircular pattern on either side of the hull. Average ranges for the GHG were 10 km at speeds against targets of 21 knots. By 1945, GHG was perhaps the best passive sonar in use on either side (Holt, 1947; Lasky, 1974).

But it was really in the United States that acoustics continued to flourish despite limited funds and personnel. The Sound Division of the Naval Research Laboratory (NRL), headed by Harvey Hayes, accomplished a substantial amount during 1920s and 1930s. They studied quartz-steel ultrasonics in an effort to improve the Langevin apparatus. This resulted in the development of newer ultrasonic equipment to be used on the submarine and surface vessels. But their range was considered to be too limited and the Sound Division worked on further to develop the JK transducer from Rochelle-Salt piezoelectric crystals. JK is the Navy term for passive sonars where J means that it can only be used for listening and K is merely the model. Used on submarines as a passive receiver, it was employed to find and classify targets at long ranges. Further research on the JK system led to the development of the rubber spherical window for the JK projector. This was an active sonar unit and was called the QB system. QB is also a navy term for active sonars with Q indicating that the transducer can be used for both sending and receiving signals and B indicating the model. To reduce the background noise, a streamlined dome was placed around the transducer (Holt, 1947; Lasky, 1974; Namorato, 2000).

As the research went on, NRL developed a depth finder and a variety of submarine detection equipment in the 1930s. They then proceeded to develop the QC transducer, which was continuous ping screening sonar. NRL even combined two projectors (QC and JK) and used on submarines with JK to detect and classify, QC for range. By 1939, NRL had effectively developed passive/active sonar for the Fleet (Holt, 1947; Lasky, 1974, 1975; Namorato, 2000).

Finally, the last significant development was the Bathythermograph (BT), developed by NRL and Woods Hole Oceanographic Institution. The BT gave a measurement of the temperature as a function of depth near and around the ship. The path of the sound beam could be developed from different temperature profiles. BT proved to be a quite important development for the use of underwater acoustics in warfare and was a standard installation on all US ships and submarines in WWII (Lasky, 1975, 1977).

In general, the accomplishment of the British, German and American researchers was quite remarkable during this period under severe budget constraints. More was required to be done and World War II appeared to have provided the stimulus and environment to take the development of sonar technology further (Namorato, 2000).

2.3.6. World War II (WWII)

The developments in transducers during and after WWI combined with advances in electronics and better understanding of the propagation of sound in the ocean provided the basis for development of sonar systems on the onset of WWII. In many respects, World War II saw the culmination of what had started in World War I. With new developments in warfare such as Blitzkrieg and with the appearance of the Luftwaffe and "wolf packs" and with German submarines causing serious damage to shipping off the east coast of the United States, the demands on acousticians was phenomenal (Lasky, 1974, 1975, 1977; Namorato, 2000).

Several research institutes were set up in the US at the beginning of the war: with Columbia University's Division of War Research (CUWR) at New London, Harvard University's Underwater Sound Laboratory (HUSL) in Cambridge, and the University of California's Division of War Research (UCWR) in San Diego. Their relentless efforts resulted in many American ships being equipped and continuously upgraded with echo ranging and passive listening systems as the war progressed. Other types of equipment employing transducers and underwater sound were also developed such as acoustic homing torpedoes, acoustic mines and sonobuoys (a relatively small expendable sonar system that is dropped/ejected from aircraft or ships conducting anti-submarine warfare or underwater acoustic research). A large amount of practical experience was accumulated from the use of all this equipment, and it provided a firm basis for many new developments during and after the war (NDRC, 1946).

In an effort to deal with these threats from German and Japanese acoustic mines and homing torpedoes, numerous technological improvements were made to the existing sonar systems. Sonar domes, for example, were improved so as to reduce self-noise caused by the flow near the transducer and to minimize noise through a sound baffle-sound absorber. The QC sonar was modified to reduce reverberations, improve bearing indications and deviations, and ease the operation of the system itself. Developments were also made to the JP, JT (both passive sonars), and TLR (Triangulation-Listening Ranging) to assist the submarine in its attack capabilities (Lasky, 1975).

Another significant development in acoustic research during World War II was SOFAR (Sound Fixing and Ranging). In 1943, Maurice Ewing and J.L. Worzel discovered permanent sound channels in the ocean at depths between 500 and 1300 meters. They found that sound from small TNT explosions could travel well over 1000 kilometres at such depths. Using this information, they developed SOFAR, which was extensively used by the downed pilots. By setting off a small explosive device, the signals could be picked up at receiving stations far away and rescue attempts for the pilot could be made. Finally, in 1946, the NDRC issued the famous “Red Books” in 22 volumes. It contained a collection of summary technical reports on all that was researched and accomplished in underwater acoustics in the United States during the war. These volumes are considered as benchmarks in the history of acoustics and its applications to warfare (Namorato, 2000).

In summary, World War II saw underwater acoustics develop into a highly sophisticated, multi-disciplined science supported and sustained by the Navy. As a result, a large portion of the research was kept under the veil of secrecy. It is only recently that the scale of the development that took place during the WWII and the cold war has come to light and the magnitude of the development could be researched and analysed to its entirety.

2.3.7. Post WWII and the cold war era

After the war and until the late 1950s, sonar systems were mainly used to monitor ship convoys and shipping corridors. With the development of submarines capable of launching nuclear missiles and attack nuclear submarines, development of underwater acoustics gathered pace again. Priority was given to the development of passive sonar systems, as vast areas of the ocean were required to be monitored for nuclear submarines.

With the onset of Cold War and the rapid development of nuclear submarines, underwater acoustics became a fundamental part of military research but as a science, it divided itself up into more specialized areas and contributed significantly to other evolving ancillary research fields such as military oceanography and medicine (Namorato, 2000). As the development of nuclear submarine gathered pace rendering the existing sonar systems less efficient, focus was given to fundamental acoustics and oceanographic research. This, during the 1950s, led to the development of split beam and all round scanning sonars. The transducers were also developed significantly (Lasky, 1977; Press, 2011).

With the introduction of digital signal processing in the late 1960s and the evolving computer performances, passive sonars became very sophisticated. That, however, was countered by the equal pace of sophistication of submarines by dramatically lessening acoustic noise radiating from them. So the focus had shifted again on the development of active sonars in the 1990s. The new breed of active sonars could also operate at lower frequencies and played a vital role in the conflicts of the late 20th century in detecting submarines (Falklands War) or under water mines (Gulf War) (Lurton, 2002).

2.3.8. Civilian developments

The civilian oceanographic industry and research institutes benefited directly from the military developments of acoustic echosounders during and after the WWII. The civilian industries were able to quickly adapt from the declassified information and upgrade the commercial echosounders. Singlebeam echosounders (SBES) were being extensively used for depth measurements since their development. They replaced the lead lines that were used for depth measurements 4000 years until the early 19th century. Since their first successful use in detecting fish in 1929 (L. Ding, 1997; Kimura, 1929) using forward scattering technique, government institutes and researcher started to use SBES more frequently in fisheries and biomass monitoring after the second world war (Lurton, 2002; Mayer, 2006).

Much of the work after the war was orientated towards sonar system design and development. A vast commercial aspect acted as a strong motivation in the search of improved materials for the transduction led to the development of ammonium dihydrogen phosphate (ADP), lithium sulfate and other crystals in the early 1940's (Sherman et al., 2007). A.R. von Hippel in 1944 discovered piezoelectricity in permanently polarized barium titanate ceramics (Dresselhaus, 2004). Jaffe et al. (1954; Jaffe, 1955) developed a stronger piezoelectricity in polarized lead zirconate titanate ceramics. These discoveries vastly improved the lead zirconate titanate (PZT) transducers quality and initiated the modern era of piezoelectric transducers.

A great advantage of piezoelectric ceramics and ceramic-elastomer composites is that it can be made in variety of shapes and sizes with many variations of composition to address specific properties of interest in underwater exploration. This often resulted in practical systems built for a specific customer. This flexibility has led to the development and manufacture of innovative, relatively inexpensive transducer designs which was quite unimaginable in the early days of transducers (Sherman & Butler, 2007).

2.4. Sonar applications: a brief outline

Applications of sonars can be divided into two broad categories: military and civilian application. Depending on the application area, sonars are mainly classified based on their functions.

2.4.1. Military application

Based on the function mode, sonars in military use are divided into two main categories - active and passive sonars:

Active sonars

These are sonars that are able to transmit acoustic signals and receive the reflected echoes from the target. They consist of an array of projectors to transmit acoustic pulses into the water. The time of echo arrival at the receiver is used to estimate the distance of the target as well as the angle of arrival of the echo. The echoes can be further analysed to give more details on the target (Lurton, 2002; Waite, 2002).

Each transmission of echoes from active sonar is known as a 'ping'. The term 'ping' can be rather ambiguous as it can be a single acoustic pulse transmitted from the sonar or a sequence of pulses. It can also be the total time between two transmissions, that is, the sum of the duration of both the pulses and the receive period. The meaning is usually clarified by the context of its use (Waite, 2002).

Passive sonars

These are sonars that are capable of only receiving acoustic noises (pulses) radiated by the target (for example: ships, submarines, torpedoes etc.). They can consist of a hydrophone or a series of hydrophones designed to detect radiated noise against a background of ambient and self-noise. Passive sonars have no known civilian application (Lurton, 2002; Waite, 2002).

2.4.2. Civilian applications

Use of sonars in civilian sector is highly varied and growing with time. Sonars have evolved with the growing need of advanced scientific instruments from expanding programmes of environment study and monitoring, industrial fisheries and offshore engineering. The following active sonars are commonly used in the civilian sector:

Bathymetric sounders

These sonars are mainly singlebeam echosounders (SBES) and specialised in depth measurements and replaced the 'lead line' method. A narrow beam is transmitted vertically downward and the time delay of the echo is measured. This produces a depth or a bathymetric measurement. These sonars are ubiquitous in naval navigation (Lurton, 2002; Mayer, 2006).

Fishery sounders

After the first successful detection of fish by Kimura (1929), use of sonars for detection of fish shoals took off after the Second World War. Fisheries also use singlebeam echosounders, which work in a similar way as the bathymetric echosounders. To facilitate fish detection, the echosounders also carry additional tools to process the echoes originating from the entire water column (Lurton, 2002).

Sidescan sounders

Side scan sonars are a type of sonar that is extensively used in seabed mapping and are towed by the survey ship or submarine. They are capable of producing highly accurate observations and are used in the acoustic imaging of the seabed. They transmit short pulses in the horizontal direction that sweeps the seabed signals as a function of time. This time series echoes yield an image of irregularities, obstacles, and changes in structures of the bottom surface. As they are towed, they can be used in surveys trying to detect specific objects by lowering them closer to ground thus providing a very high-resolution image of a smaller target area. These systems are mainly used in marine geology studies and shipwreck and mine detections (Lurton, 2002).

Multibeam sounders

Multibeam echosounders, also known as Swath echosounders, comprise of an array of transducers mounted directly beneath a ship's hull. Multibeam sonars emit sound waves to produce fan-shaped coverage of the seafloor. The main use of these sounders is seafloor mapping for its ability to provide an accurate topography of the seabed. Each of its transducers is capable of transmitting narrow width beams, and sweeps a large swath of the seabed in each survey run. The results are usually a highly dimensional acoustic database that can be used to produce an accurate topographic map of seabed. It can also produce acoustic images if the angular aperture is large enough. Multibeam echosounders are also commonly used for geological and oceanographic research, and since the 1990s for offshore oil and gas exploration and seafloor cable routing (Lurton, 2002; Mayer, 2006).

Sediment profilers

These are singlebeam echosounders specialising in utilising their beams' ground penetrating capability more effectively. The frequency of the sonar is kept to a minimum, which allows the beam to penetrate from 10 to hundreds of meters of the seafloor. The frequency is adjusted based on the preliminary study of seafloor type. Another method for sediment profiling is seismic systems. In this method, explosive or percussive sources coupled with long antennas are used to explore several kilometres under the seafloor. These systems are primarily used for oil and gas exploration and in geophysical studies (Lurton, 2002).

Acoustic Doppler systems

These systems are commonly used for river current velocity and discharge measurement. The acoustic Doppler profiler is mounted to a vessel that moves across the river perpendicular to the current. Water velocities are measured when the acoustic Doppler profiler transmits acoustic pulses along three or four beams at a constant frequency. The instrument processes the echoes to estimate the difference in frequency (shift) between transmitted pulses and received echoes, known as the Doppler Effect. It can be used to measure the relative current velocity. For discharge measurement, it transmits a series of acoustic pulses known as pings. Pings for measuring water velocities are known as water pings, and pings for measuring the boat velocity are known as bottom-tracking pings. These pings are normally interleaved and are referred to as an ensemble. A single ensemble may be compared to a single vertical echo from a conventional echosounder discharge measurement. This system is extensively used in hydrological studies (Lurton, 2002; Oberg & Schmidt, 1994; Yorke & Oberg, 2002).

2.5. Systems for mapping the seabed type

The sonar systems that are extensively used for seabed mapping are MBES and side scan sonars (SSS). In the recent years, SBES systems have also been used in this purpose. The following sections will discuss the mapping of seabed type using MBES and SBES only.

Mapping seabed type is important in a variety of applications, such as environmental research, management of marine and coastal resources and oil and gas exploration. Typically, tools used for this purpose are acoustic sonar systems, which transmit and receive an acoustic pulse from a device on a survey vessel. Data collected consist of the travel time of the acoustic pulse to the sea floor and back and the strength of the signals. From this, measurements such as the depth to the seafloor (bathymetry), depth to sub-surface sediment layers (sub-bottom), and the reflectance of the sea floor (intensity of backscattered energy) can be derived.

There is a strong link between acoustic backscatter and sediment characteristics of the seabed (C J Brown & Blondel, 2008; Goff et al., 2004; Huges Clarke et al., 1997), therefore such data are often used for classification of seabed type. Multibeam (MBES) and Singlebeam (SBES) echosounders are now commonly used in the accurate mapping of the seafloor and depth profiling. Before MBES was developed, only SBES systems were available for this purpose. But the major drawback of SBES systems is that it returns only one depth value per ping. The early SBES systems had broad beams (30-60 degrees). The area ensonified by these beams were large with area to water depth ratio varying from 0.5 to 1.

The problem with large ensonification is that the first echo recorded can come from any part of this vast area. The early systems lack the within beam angular discrimination capability and it was assumed that the echo came from directly beneath the surveying vessel. This potentially resulted in a defocused and somewhat inaccurate picture of the seafloor topography. Later on to address this issue, high resolution, narrow beam SBES systems were developed. But the problem with this system was that it covered a small area. The sampling was dense in the along-track direction and widely spaced ship tracks meant that only sparse sampling of seafloor was possible. This problem was addressed, later in the 1970s, with the development of MBES systems (Mayer, 2006).

2.5.1. Echo sounding: common features & measurement types

Echosounders use the echo return and corrects it for several noise sources such as the motion of the sonar platform, inherent vibration etc. to build up a representation of the seabed and targets between the source and the seabed. The operating frequencies range between 3 kHz (sub bottom profiler) and 500 kHz (shallow water mapping). The frequency is predetermined and is influenced by the requirements of range (depth) and target size. The two main components of an echosounder are its transducer and display unit (Figure 2.5). The transducer creates and receives sound waves. Before sending out another wave, the display unit pauses and waits for the echo to strike the transducer. If the echo is detected, the distance to the object is calculated and shown on the display unit, which also sounds an alarm if the distance goes below a certain value (Waite, 2002).

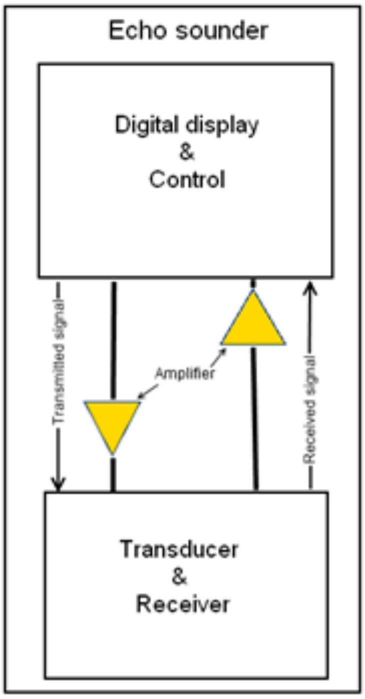


Figure 2.5: Echosounder schematics (adapted from WHSC, 2011)

2.5.2. Singlebeam echosounders (SBES)

Singlebeam bathymetry systems are generally configured with a transceiver (Figure 2.5), which is basically a system that has both a transmitter and a receiver. The system is either mounted to the hull, or side-mount to the ship. The hull-mounted transceiver transmits a high-frequency acoustic pulse in a beam directly downward into the water column. It then records the reflected echoes off the sea floor beneath the vessel (WHSC, 2011; HOMD, 2011).

This transmit-receive cycle repeats at a fast rate, on the order of milliseconds. The continuous recording of water depth below the vessel yields high-resolution depth measurements along the survey track. Additional information such as heave, pitch, roll of the vessel can be measured with a Motion Reference Unit in combination with a Differential Global Positioning System (DGPS). The system also consists of a sound velocity profiler that acquires data about the precise sound velocity in the ambient water mass. These velocity measurements are used in the depth calculation (WHSC, 2011; HOMD, 2011).

The figures below show the working principal of a typical SBES system. A transmit/receive switch (TR switch), which consists both the transmitter and the receiver, generates the pulse (Figure 2.6a) and then receives the echo, which includes the signal (backscatter information) and the noise (Figure 2.6b). The received analogue echo is filtered and then converted into a digital signal or data stream (Figure 2.6c) by the analogue to digital converter (A to D converter) in the system. A detector and low-pass digital filter is then applied to the signal to remove the carrier and higher frequency components, including the out-of-band portion of the remaining noise. The output is a smoothed signal, which is also known as ‘echo envelope’ (Figure 2.6d & e). Decimation is often used on the smoother echo to reduce the data rate. The resulting digital signal serves as the raw material for sediment classification (Preston et al., 2000).

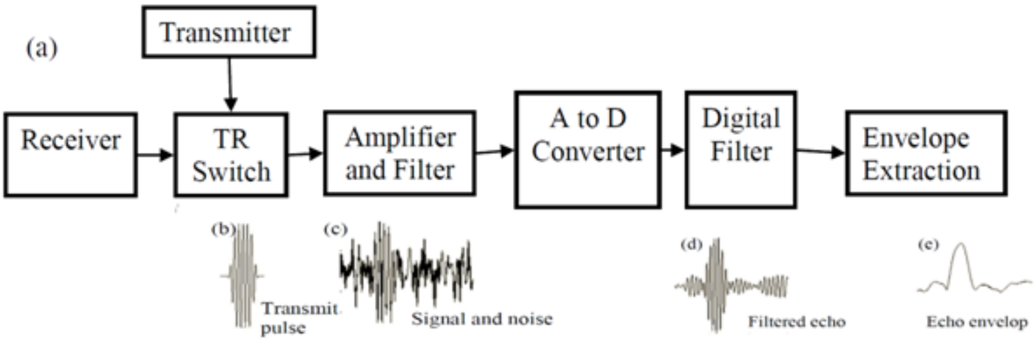


Figure 2.6: Diagram of a singlebeam echosounder's working principle. The upper row shows the important components and processes in a SBES system and the lower row shows the resulting signal at crucial steps (adapted from Brown et al., 2007)

Many present-day systems record echo data digitally. These raw signals are filtered before sampling and A/D conversion. Filtering the raw signals limit bandwidth to prevent aliasing (when the sample rate is less than twice the highest frequency in the analogue signal) in sampling and suppress noise that is outside the echo bandwidth. An echo envelope is generated from the filtered full-waveform data using the Hilbert transform (Haykin, 1994).

The Figure 2.7 below shows a typical seabed echo from a SBES system. The steep initial rise represents the instance when the transmitted pulse has reached the seabed and has returned to the transducer. The peak represents the instance of the sound wave hitting the seabed. After the signal has reached its peak, it gradually descends. This section contains the information of sub bottom properties of the seabed. It also contains higher amount of noise due to higher degree of volume scattering.

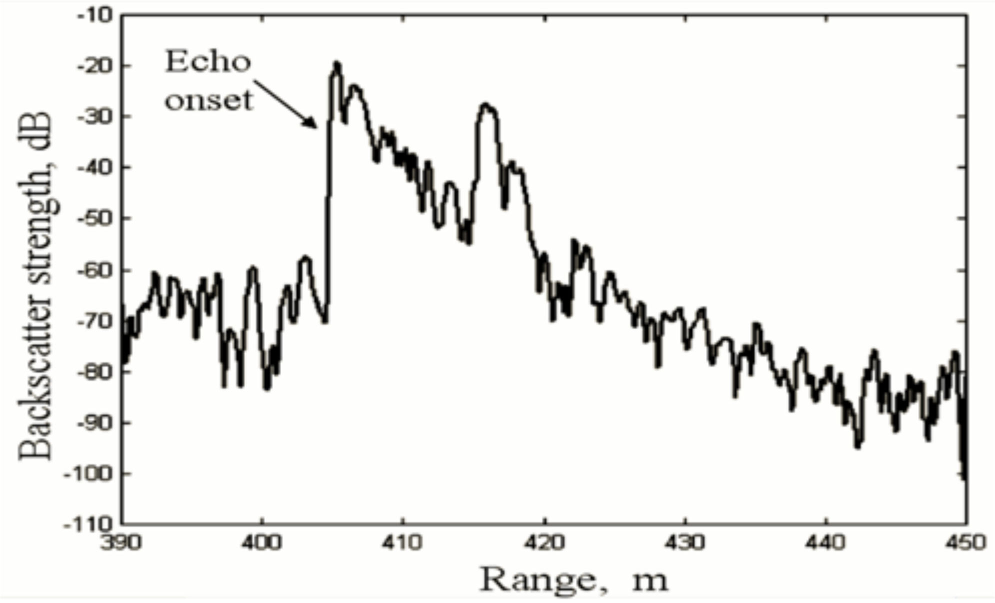


Figure 2.7: Representative seabed echo from a calibrated echosounder. The water depth is the range to echo onset; after that, echo duration depends on beam width, seabed slope and roughness, and penetration into the sediment (C.J. Brown et al., 2007)

Seabed mapping using SBES

When a set of good quality artifact free SBES echoes are collected, the next step is to extract features from them. Features in seabed mapping are a set of statistical descriptors that are generated from the backscatter data i.e. the echoes as described above. These statistical descriptors vary from common statistical measures (such as mean, standard deviation etc.) to more specific descriptors (such as textural, fractal dimensions etc.). Figure 2.8 shows the general process of feature generation the classification of SBES echoes. High-quality classes are derived from features, not from the echo amplitudes themselves. One of the objectives of feature extraction is to generate features that describe the shape of the echo i.e. features that can be used to describe the formation of echo, the nature of the curve etc. (for example: skewness, kurtosis). The number of shape features depends heavily on the analyst's imagination and experience. With the aim to capture descriptions of the echo shape and spectral character as numerical values, the procedure starts with the generation of an echo time-series as a sequence of digital samples, and the sample number that corresponds to the bottom pick. In practice, the time-period between some small fraction of the peak (for example 5%), and the peak itself, or the length at 50% of peak is measured.

Features that depend on arbitrary numbers are regarded as less suitable, but can have practical value (C.J. Brown et al., 2007).

With the presence of variability and noise, a better approach is finding shape features with statistical measures. This is achieved by dividing the samples into a number of geographic windows (mostly rectangular) with each window containing a predefined number of samples. Cumulative sums, quantiles, and histograms are calculated on each of these windows. In this way the capture of the rise time is less susceptible to noise and variability. The relative number of samples in each window expresses echo duration and decay rate and is dependent on the length of the sample window. Many other variations are possible and depend on the seabed geology, surveyor's experience with the transducers and the region (Brown et al., 2007).

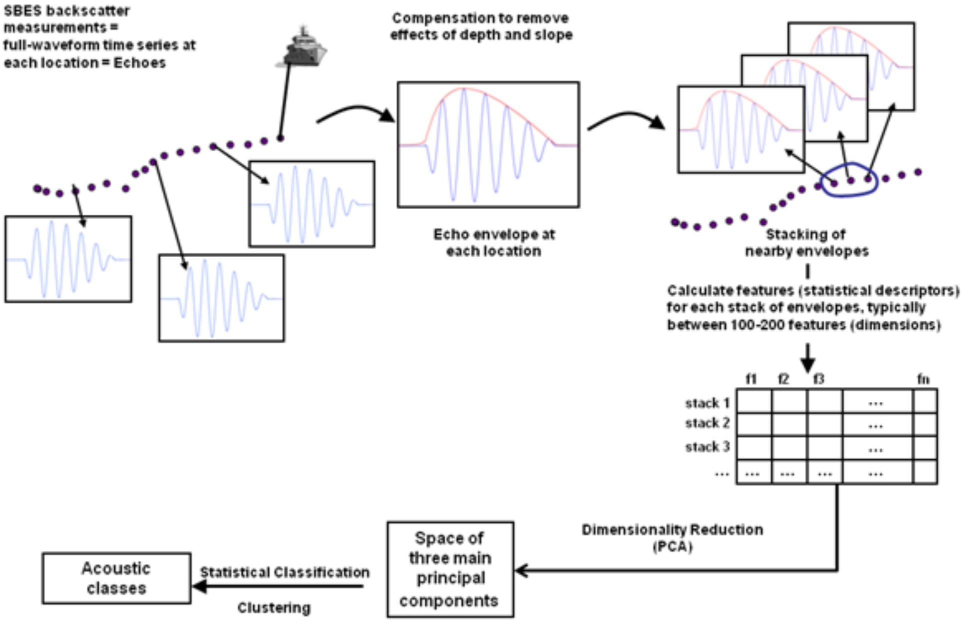

SBES backscatter measurements = full-waveform time series at each location = Echoes
Compensation to remove effects of depth and slope
Echo envelope at each location
Stacking of nearby envelopes
Calculate features (statistical descriptors) for each stack of envelopes, typically between 100-200 features (dimensions)
f1
f2
f3
fn
stack 1
stack 2
stack 3
...
Dimensionality Reduction (PCA)
Space of three main principal components
Statistical Classification
Clustering
Acoustic classes

Figure 2.8: Generic workflow of seabed classification using SBES

A second group of features, which captures the spectral character of the echoes, are also generated. Sediment roughness has a direct effect on the variability in the echo tail, so features that capture spectral content in this variability can be useful for discrimination. The Fast Fourier transform (FFT) is commonly used to provide a numerical estimate of the spectra. The power spectrum is calculated by applying FFT on the autocorrelation of the amplitudes. It expresses echo power in frequency bands. Wavelet transforms can also provide complementary information if the elementary wavelet is chosen carefully (Tegowski & Lubniewski, 2000). These methods usually operate on normalized echo time-series (with the maximum scaled to one). Therefore, the features generated are independent of echo amplitude (Hamilton et al., 1999).

Once the set of features to be derived are determined, a feature database if prepared that includes feature descriptors for every echo. The next step is to use principal component analysis (PCA) to reduce the dimensionality. In the next step, a clustering algorithm (usually k-means) is run on the space of three principal components comprising of 90-95% of the information. The clusters are then analysed and labelled based on the ground truth information available and a classified map is generated (Brown et al., 2007; Hamilton et al., 1999; Tegowski & Lubniewski, 2000).

2.5.3. Multibeam echosounders (MBES)

Multibeam echosounder works in a similar way as the singlebeam echosounder and can be considered as an extension of SBES. The development of MBES rose from the drawback of SBES, which could only take one measurement and the result is sparse sampling of the seafloor when the beam width is wider (Mayer, 2006). A multibeam echosounder transmits and receives an array of beams with individual small widths (0.50-30 degrees each) across the axis of the ship. This array of beams sweeps a large corridor around the ship's path. The intersection of the transmit pulse and the received beams results in many (100-240) simultaneous depth measurements across a wide swath, with echo measurements having excellent vertical resolution. The recent MBES systems use their larger angular width to record acoustic images. This provides the users with bathymetry (depth of water relative to sea level) and backscatter measurements (reflection of signals) at the same time (Augustin et al., 1994; Lurton, 2002; Mayer, 2006; Mcgonigle et al., 2009). Since the late 1970s, MBESs have greatly evolved and have become very varied. These systems can survey large areas rapidly and accurately. Some of the common uses of MBES are: deep water systems (12-30 kHz) for regional mapping of deeper ocean and continental shelves, shallow-water systems (100-200 kHz) for mapping continental shelves, and high-resolution systems (300-500 kHz) for local studies such as hydrology, shipwreck location and inspection of underwater structures (Lurton, 2002; Mayer, 2006).

Seabed classification from MBES is generally attempted by analysis of backscatter amplitudes. However, backscatter is affected by diversity of ocean floor types and lateral heterogeneity of sub bottom layers, which makes the acquired data difficult to analyse (Arescon Ltd., 2001; Xinghua & Yongqi, 2004, 2005). Due to the large volume of data acquired in MBES surveys, computer-assisted classification has become a logical choice to achieve statistically valid and objective segmentations (Cutter et al., 2003; Hellequin, 1998; Hellequin et al., 2003; Mcgonigle et al., 2009). With the recent development of side-scan sonar (SSS) and MBESs, image-based seabed classification based on the characteristics of acoustic backscatter became the focus of sustained effort to arrive at effective segmentation (Cutter et al., 2003; Mcgonigle et al., 2009; Preston et al., 2004; Preston et al., 2001).

Structure of MBES systems and working principal

A typical MBES system consists of an array of transducers (transmission and reception arrays), electronics for the transmission array, user interface, and ancillary systems (Figure 2.9). The electronics for the transmission array controls signal generation, amplification, impedance matching of the transducers as well as the properties of the transmission beam (width, inclination, level etc.). This function is influenced by the configuration parameters, which in turn depend on the seabed topology, weather, salinity etc. The principle functions of the electronics for the reception unit are digitization and demodulation of signals, and beam forming. It also performs level correction to keep the signal amplitudes within an acceptable range as constant as possible.

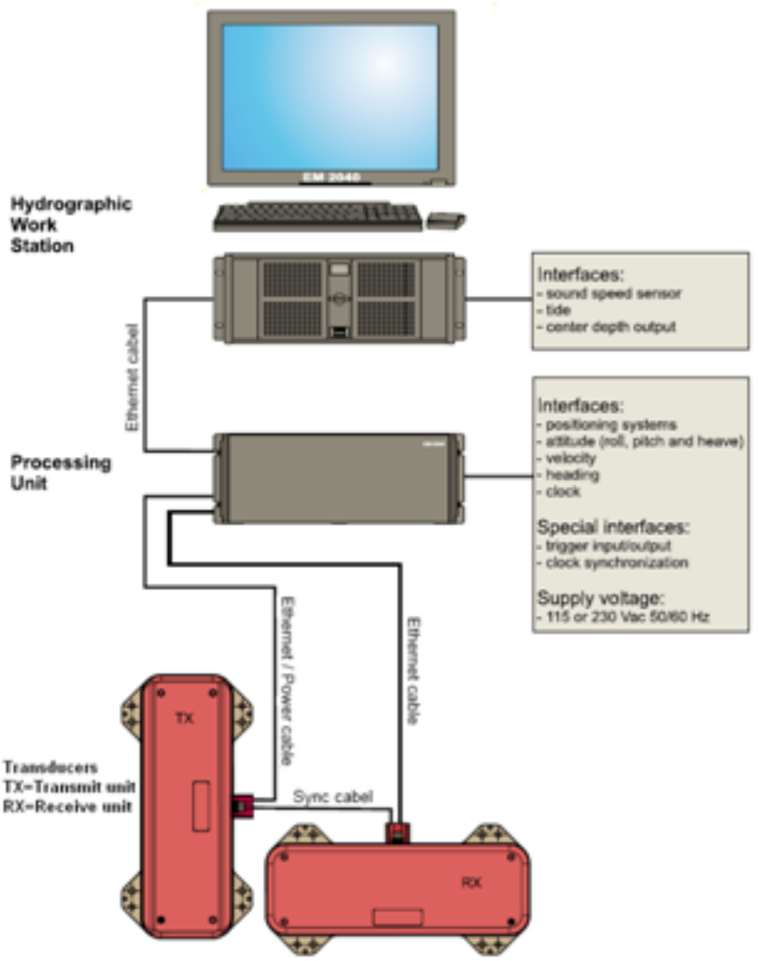


Figure 2.9: Components of a MBES system (EM 2040) (Kongsberg, 2011)

It controls the quality of preliminary bathymetry measurements and imagery signals by correcting for platform movement and acoustic paths (usually by passing filters). In the newer MBES systems, all the electronics are controlled by a dedicated personal computer or workstation on board the survey vessel (Lurton, 2002; Preston et al., 2004, 2000).

Most MBES use two arrays of transmission and reception sounders in one transducer head that can be hull-mounted (fixed) or pole-mounted (portable). The arrays are installed along the axis of the supporting platform. The setup of MBES also allows for large width or large swath (Figure 2.10). The MBES systems have superior angular discrimination. This allows the beam footprints (i.e. area ensonified by the beam) to be kept small. The along track discrimination is controlled by the transmission array while the across track discrimination is controlled by the reception array (Lurton, 2002; Preston et al., 2004, 2000).

The transmit array is in the along-ship direction and generates a pulse that is wide in the across-ship direction and very narrow in the along ship direction. The receive array is located in the across ship direction and form receive beams that are narrow in the across-ship direction and wider in the along-ship direction (Figure 0.10). These early systems formed lower number of beams (only 16), which limited their swath width to up to 45 degrees (0.75 times the water depth). Modern systems can form more beams (typically 100–240), which enable them to generate much wider swaths (typically 100–150 degrees) than the early systems. The intersection of the transmit pulse and the receive beams results in many simultaneous depth measurements across a wide swath (100 to 240). Each measurement provides excellent horizontal and vertical resolution. There are now a wide range of multibeam sonar systems available, operating at frequencies from 12 kHz (for deep-water mapping) to 455 kHz (for working in water depths less than 100 m) (Mayer, 2006).

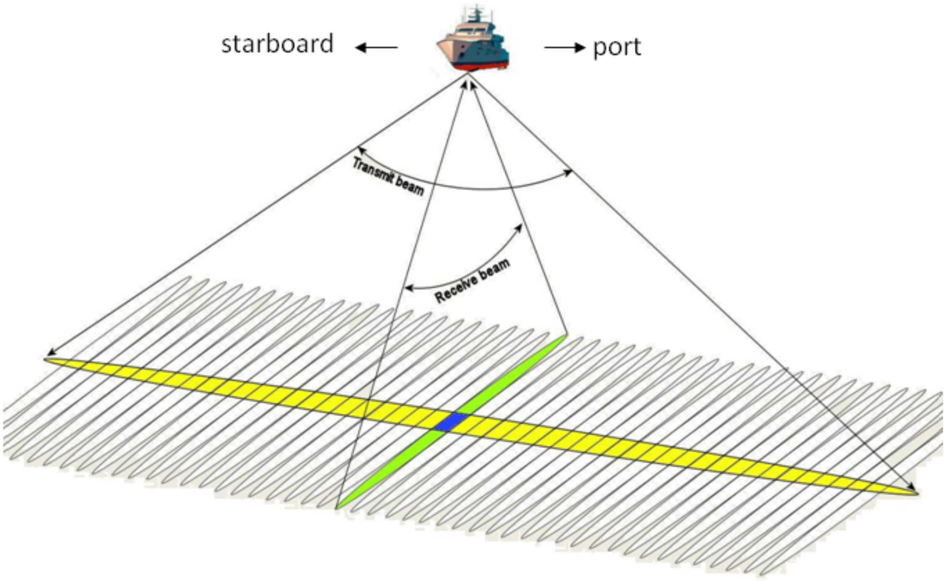


Figure 2.10: MBES working principal

A multibeam echosounder (MBES) transmits and receives an array of beams with individual narrow widths (0.50-50) in along-track direction and wide (1000-1500) across-track direction. The area ensonified by an individual beam refers to the energy or acoustic pulse from the transducer which has reached part of the seafloor (Clarke et al., 1997). The area ensonified by each beam takes the shape of an ellipse. The dimension of this ellipse varies principally as a function of depth, individual beam pointing angle, and the incidence angle. The limited dimension of each formed beam allows the estimation of the smallest feature in the seafloor that can be surveyed within the ensonification region. This array of beams sweeps a large corridor around the ship's path. The intersection of the transmit pulse and the received beams results in many (100-240) simultaneous depth measurements across a wide swath, with echo measurements having excellent vertical resolution. The recent MBES systems use their larger angular width to record acoustic images. This provides the users with bathymetry (depth of water relative to sea level) and backscatter measurements (reflection of signals) at the same time (Augustin et al., 1994; Lurton, 2002; Mayer, 2006; Mcgonigle et al., 2009). Since the late 1970s, MBESs have greatly evolved and have become very varied. These systems can survey large areas rapidly and accurately. Some of the common uses of MBES are: deep water systems (12-30 kHz) for regional mapping of deep oceans and continental shelves, shallow-water systems (100-200 kHz) for mapping continental shelves, and high-resolution systems (300-500 kHz) for local studies such as hydrography, shipwreck location and inspection of underwater structures (Lurton, 2002; Mayer, 2006).

Each MBES survey provides two types of measurement: bathymetric and backscatter data. Bathymetric data is extensively used for terrain modelling; navigational maps etc. while, the common use of backscatter data is for fisheries, geological exploration and seabed mapping.

Backscattering in seabed surveys

The echoes that are transmitted from the echosounder propagate in the ocean and are then reflected in all directions by objects like fish, plankton, bubbles, submarines and ship wreckage and the boundaries of the medium (seabed and sea surface). The portion of the echo that returns to the echosounder is called the backscatter (Figure 2.11). Clear understanding of their properties is essential to the good functioning of the sonar systems (Lurton, 2002).

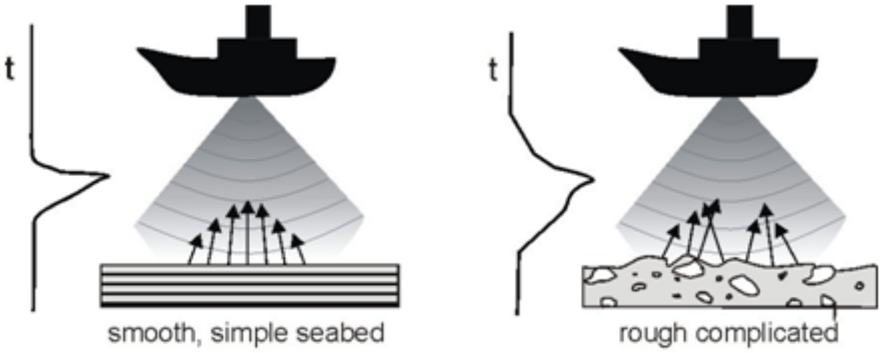


Figure 2.11: Backscatter and the associated signal profile on smooth and rough surface (MARUM, 2011)

In reality, the reflecting surface is never an ideal plane surface. The acoustic reflections are therefore much more complex. The backscatter characteristics depend on the signal strength, frequency, wavelength, incidence angle, and the local characteristics of the relief itself (MARUM, 2011). A portion of the incident wave is reflected in the specular direction. This is the coherent part of the wave. The rest of the wave is scattered over the entire space. A high roughness of the seabed will result in a relatively smaller specular reflection and a higher scattering (Waite, 2002).

During an acoustic survey, the backscatter amplitude will largely depend on the incidence angle or roughness considered prior to the survey. When the transducer is close to normal of the target, the large portion of the reflection will come from the specular echo, while the reflection at oblique incidence angles largely come from the scattering (Lurton, 2002, Waite, 2002). Other factors that influence the backscatter coefficient are the impedance (product of density and sound velocity), contrast (between the seabed and the medium) and the property of the underlying soil volume of the seabed (Buckingham, 2000; Chotiros, 1995).

Seabed mapping using MBES

Seabed classification from MBES is generally attempted by analysis of backscatter amplitudes. However, backscatter is affected by diversity of ocean floor types and lateral inhomogeneity of sub bottom layers, which makes the acquired data difficult to analyse (Arescon Ltd., 2001; Xinghua and Yongqi, 2004, 2005). Due to the large volume of data acquired in MBES surveys, computer-assisted classification has become a logical choice to achieve statistically valid and objective segmentations (Cutter et al., 2003; Hellequin, 1998; Hellequin et al., 2003; McGonigle et al., 2009).

With the recent development of side-scan sonar (SSS) and MBESs, image-based seabed classification based on the characteristics of acoustic backscatter became the focus of sustained effort to arrive at effective segmentation (Cutter et al., 2003; McGonigle et al., 2009; Preston et al., 2001; Preston et al., 2004).

Backscatter data from MBES can be used to gain insight on the spatial distribution of seabed properties (Goff et al., 2004; Hughes Clarke et al., 1997). Sonar acoustic waves within the echo levels and frequency can penetrate up to few tens of cm in soft sediments thus providing information on surface and sub-surface properties. Amplitude backscatter returns will be influenced by a combination of geological and non-geological variables. Geological factors will be a combination of surface and/or subsurface scattering processes within the sediment. Non–geological factors are generally divided in geometric and radiometric factors. The first ones are controlled by the beam incident angle and range to the system. Radiometric factors are controlled by system settings such us system power, time varied gains (TVG) and absorption coefficients.

Image classification is usually carried out on statistical features generated from the MBES backscatter image. A series of rectangular patches are distributed over the backscatter images (see Chapter 3). The placement of these patches depends on the quality of the data. It is also influenced by the grazing angle and range of the sonar as well as constraints that come with different sensor models. The influences of these are removed through the process of image compensation during the next step where the image is separated into rectangular patches, which are superimposed on the image on each side (port & starboard) of the vessel (Preston, 2009; Quester Tangent, 2007). A matrix of amplitudes from each patch is generated and fed into image feature algorithms.

Backscatter data for seabed classification can have flawed values (due to unreasonable depth picks, reflections from fish or artefacts, etc.). In addition, images can be smeared due to excessive vessel movement. Before feeding the data into feature algorithms, the backscatter data are first filtered to clean the data of these anomalies.

The goal of the feature extraction from MBES backscatter image is to capture the amplitude and texture characteristics using some of the published methods. A number of features can be generated and the type of features to be generated depends on the hydrographer who usually has prior knowledge of the area surveyed. The features are called Full Feature Vectors (FFVs) and the result is a large matrix in which each column represents the values of one feature and each row contains all the features extracted from one rectangular patch.

The most common features are the mean, standard deviation, and higher moments of the amplitudes in the rectangular patch. Other features can include quantiles, histograms, and other measures of the amplitude distribution. Texture features provide a “feel” of the image. It discriminates between uniform and irregular regions and among types of patterns (Blondel et al., 1998). The texture features are usually derived from a grey-level co-occurrence matrix (GLCM) that captures the changes in grey level between neighbouring pixels. Haralick and Shanmugam (1973) first described a number of GLCM features with names like prominence, shade, and entropy. Fractal dimension, with amplitude treated as if it were altitude, is another useful feature that captures image texture (Carmichael et al., 1996; Tegowski & Lubniewski, 2000).

The resolution of resulting classification maps is set by the size of these rectangular patches, with smaller patches giving higher resolution and larger patches resulting in lower resolution maps. The main drawback with smaller patches is that it can restrict the selection of features. For a single pixel rectangle, the only possible feature is amplitude. The classification map is, therefore, generated by a sonar mosaic organized into units by a set of amplitude thresholds (Brown et al., 2007).

Various techniques for image processing can be applied to backscatter images for image segmentation in order to provide classification maps of the seabed type. In addition to backscatter measurements, sonar geometry and the geometry of the entire multibeam system have to be taken into account in order to achieve a valid classification that depends on sediment characteristics rather than sonar artefacts (Preston et al., 2001). Another approach to segmentation relies on the calibrated backscatter levels and on their variation with grazing angle (Hughes Clarke et al., 1997). Other recent approaches to automatic classification are based on the statistical nature of the image, irrespective of absolute calibration, which uses the backscatter amplitude and statistical properties of multibeam sonar images to classify seabed sediments. A combination of PCA and k-means is currently the most common way of clustering and seabed data (McGonigle et al., 2009; Preston et al., 2001; Preston et al., 2004).

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a mathematical procedure that converts a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components using an orthogonal transformation. The number of principal components is less than or equal to the number of original variables (Ding & He, 2004; Gorban et al., 2007; Jolliffe, 2002; Pearson, 1901).

PCA is a widely used data compression technique. The PCA is computed by determining the eigenvectors and eigenvalues of the covariance matrix. Eigenvectors are a special set of vectors associated with a linear system of equations (i.e., a matrix equation) that are sometimes also known as characteristic vectors, proper vectors, or latent vectors (Marcus & Minc, 1988). The eigenvectors of a square matrix are the non-zero vectors that, after being multiplied by the matrix, remain parallel to the original vector. For each eigenvector, the corresponding eigenvalue is the factor by which the eigenvector is scaled when multiplied by the matrix. If A is a square matrix, a non-zero vector **v** is an eigenvector of A if there is a scalar λ (lambda) such that

$$A\mathbf{v} = \lambda\mathbf{v} \qquad 2.6$$

The scalar λ is said to be the eigenvalue of A corresponding to **v**. A covariance matrix or dispersion matrix is a matrix whose element in the i, j position is the covariance between the i th and j th elements of a random vector (i.e., of a vector of random variables). To perform PCA for a given dataset, $X = \{X_{m,n}\}$, the empirical mean is first calculated (Jolliffe, 2002):

$$u_m = \frac{1}{N}\sum_{n=1}^{N} X_{m,n} \qquad 2.7$$

Here, X is a $M{\times}N$ matrix with $m = 1, \dots, M \; and \; n = 1, \dots, N$. In the next step, the deviations from the mean are calculated. The objective behind mean subtraction is to find a principal component basis that minimizes the mean square error of approximating the data.

$$B = X - uh \qquad 2.8$$

Here, h is a $1 \times N$ row vector of all 1s and the dimension of B is $M \times N$. In the next step, a $M \times M$ empirical covariance matrix C is calculated from the outer product of matrix B with itself.

$$C = \frac{1}{N} \sum B.B^T \qquad 2.9$$

From the covariance matrix, its eigenvectors and eigenvalues are computed. A matrix V of eigenvectors is created which diagonalizes the covariance matrix (Jolliffe, 2002).

$$V^{-1}CV = D \qquad 2.10$$

Here, D is the diagonal matrix of covariance matrix C and has the dimension of $M \times M$. The columns of the eigenvector matrix V and eigenvalue matrix D are the sorted in the order of decreasing eigenvalue while maintaining the correct pairings between the columns in each matrix. In the next step, the cumulative energy content (g) of each eigenvector is computed. The cumulative energy content (g) for the mth eigenvector is the sum of the energy content across all of the eigenvalues from 1 to m.

$$g_m = \sum_{q=1}^{m} D_{p,q} \; for \; m = 1, \dots, M \qquad 2.11$$

A subset of eigenvectors, the first L columns of V as the $M \times L$ matrix W, is selected as basis vectors:

$$W_{p,q} = V_{p,q} \; for \; p = 1, \dots, M \; \& \; q = 1, \dots, L \qquad 2.12$$

Where, $1 \leq L \leq M$. The vector g is usually used as a guide to choose appropriate value of L. The objective is to choose a value of L as small as possible while achieving a reasonably high value of g (Jolliffe, 2002).

$$\frac{g_{[m=L]}}{\sum_{q=1}^{m} D_{p,q}} \geq 0.9 \tag{2.13}$$

2.6. Noise in acoustic surveys

Noise in acoustic surveys, stated by Waite (2002), "is the background against which sonars must detect signals from target". Lurton (2002) divided the noise present in acoustic sonar survey into four types: Ambient noise, self-noise, reverberations, and acoustic interference (Figure 2.12). It is an important component to underwater surveys and ads to the signal, decreasing its quality.

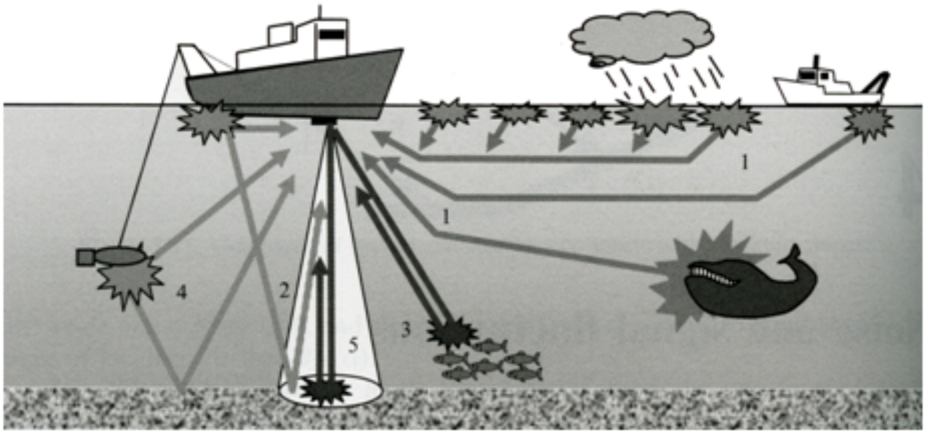


Figure 2.12: Noise affecting a hull-mounted sonar system: (1) ambient noise; (2) self-noise; (3) reverberation; (4) acoustic interference (Lurton, 2002)

Ambient noise originates from natural sources (waves, fish movement etc.) or man-made sources (shipping, drilling, submarine movements etc.). Self noise stems from the echosounder itself. Noise from radiated energy of the supporting platform, flow noise, electrical interference etc. is different forms of self-noise. Reverberations are exclusive to active sonars. It is the echo generated from the signal itself (parasite echo). It is often regarded as a limitation for sonar systems. Acoustic interference is the noise generated from other acoustic systems working in the same area. In some cases, these interferences can be from sources that are situated far away from the echosounder in question (Lurton, 2002; Waite, 2002).

Reverberation

The effects of reverberations in underwater surveys can be very significant at times and can sometimes render the survey run useless. It is induced by the propagation medium and stems from its boundaries (surface and bottom) and the heterogeneity present within this boundary (bubbles, fishes, intrinsic fluctuations etc.). The spectral characteristics of reverberations are very similar to that of backscattered target signals and are therefore very difficult to distinguish. The level of reverberation decreases with time but the rate is slower than the target's echo.

Noise reduction

Suppressing noise sources often requires a lot of attention. The vessel and ambient noise can be dominant at low frequencies, while receiver noise usually appears at higher frequencies and at longer ranges. Noise from various sources may be apparent on the echogram if the display sensitivity is increased and in areas of the water columns that are free of scatters are selected (Brown et al, 2007).

The simplest way to minimize ambient noise is to maximise the signal-to-noise ratio. This can be done by increasing the signal strength and constraining the angular spread as much as possible. Another way is the use of adaptive filtering whose properties vary with noise fluctuations. Increasing the number of receivers also helps in reducing the ambient noise levels as it can increase the level of the coherent part of the echo received (Hodges, 2010; Lurton, 2002; Waite, 2002).

The best way to address the self-noise issue is to identify the sources of self-noise. An accurate identification can allow precise measure to be taken to reduce the noise. In some cases, for example: radiated noise from the platform, it can be very difficult or impossible and very expensive to reduce the noise significantly. It is often cost effective and more common to optimize the location of the transducer in the hull and enclosing it in profiled fairings that reduces turbulence. Noise from electrical interfaces is directly related to the quality of the components chosen (Hodges, 2010; Lurton, 2002; Sherman & Butler, 2007).

A common approach to reduce reverberation is to decrease the signal strength and increasing its spectral width (Sherman & Butler, 2007). Adjusting the antennas to be spatially selective also reduces reverberations. However, it should be kept in mind that these two measures should not interfere with the formation of echo on the target (Hodges, 2010; Sherman & Butler, 2007).

Acoustic interference can be reduced by making sure that all the acoustic components on board have compatible frequencies as well as there is little or no acoustic sources near the survey area. This can be often difficult to achieve, as unlike electromagnetic signal frequencies, acoustic signal frequencies have no legal regulations. Moreover, interference can also occur from the harmonics of the main signal. Some other steps that can reduce the effects of acoustic interference are spatial filtering, optimal spacing of the transducers, and the use of acoustic barriers (Hodges, 2010; Lurton, 2002).

2.7. Effects of environmental variability on signal quality

Apart from the noise, environmental phenomena present during the acoustic survey can have significant effect on the signal quality and result in signal fluctuations.

Effects of ocean water characteristics

Oceans are unstable and heterogeneous. The characteristics of ocean water vary both spatially and temporally. Perturbations in acoustic transmissions can stem from this variability. The variation in ocean water in space and time can be divided into three main scales: small, intermediate and large.

The small scale is relative to wavelength and heterogeneity of the medium at this scale can induce scattering and fluctuations in the signal. The intermediate scale is relative to the sampling rate and at this scale; instabilities in the medium (for example: swell) can cause time delays in signal arrival and amplitude fading. At large scales, the variation in sound velocity profiles or water depths can induce permanent bias in target positioning (Hodges, 2010; Lurton, 2002; Marage & Mori, 2010; Sherman & Butler, 2007).

Spatial variability

The seafloor topography changes over space at variable scale (Hodges, 2010; Lurton, 2002; Marage & Mori, 2010; Sherman & Butler, 2007). While the small slope variation over a large scale have little effect on signals, strong topography features such as seamounts, ridges, continental slopes can refract the signal through reflections on inclined surfaces and introduce reverberations (Sherman & Butler, 2007). The type of the seabed adds to the effects from the relief. Though seabed types are often homogeneous in deep waters, they can vary rapidly in coastal areas.

These variations sometimes affect the quality of the acoustic signals (Hodges, 2010; Marage & Mori, 2010). The sound velocity profile varies spatially and can also change locally due to geographical or environmental constraints, for example: currents or gyres, estuaries etc. Seasonal change can also induce a regional variation in the signal (Lurton, 2002).

Temporal variability

The temporal nature of several environmental phenomena in the ocean can affect the acoustic signal characteristics. The temporal changes in currents can lead to amplitude instability. An increased current speed can also introduce higher scattering of the echo. This effect is known as ‘scintillation’ effect. Internal waves generated from the variation of density due to depth change can induce a sound velocity variation of up to several meters per second (Lurton, 2002). These fluctuations in sound velocity reduce the spatial coherence of the acoustic signals (Flatté, 1979). Tides, which last about half a day, can also affect the sound signal in shallow waters. The daily and seasonal temperature variation in the ocean can, sometimes, induce modification of the sound velocity profiles and can noticeably affect the sound field structure (Lurton, 2002; Marage and Mori, 2010).

2.8. Quality & compensation

The priority in any acoustic survey is to ensure a good echo registration with low signal to noise ratio and there are several approaches to achieve that. The noise from the transducer or self-noise can be kept to a minimum by frequent observation and calibration of the sonar system. These observations and verification is done by routine checks and observing the operations during the survey.

Quality control

Quality control and image compensation are another way to improve the echo quality. Quality control involves a careful watch over the survey process and filtering or keeping aside any echo that appears to be degraded. Interference from the transducer and clipping (loss of information) can be diagnosed using an echogram (amplitude vs. time or depth plots). Clipping (loss of information) can be diagnosed if the maximum possible digital value is present in a series of consecutive samples, or if some lesser digital value is frequent but never surpassed (Brown et al., 2007).

Compensation

Water depth, characteristics of the water column, bottom slope and sediment type are the major factors that affect the amplitude and shape of an echo. Compensating for depth effects can often be challenging. Shape characteristics of an echo, such as rise time and decay time, increase with depth. Compensation is done by re-sampling of the digital echo so that the echo has the length it would have had if it had come from some selected reference depth (Pouliquen, 2004). In shallow water, the spreading time of the beam front on the seabed is quite short and assumed not to be a dominant contributor to echo durations, so calculating the re-sampling rate is more complicated (Preston et al., 2003). Acoustic classes are assumed to be heavily influenced by depth if no compensation is done by re-sampling (Lubniewski & Pouliquen, 2004). Therefore, compensation is highly recommended.

2.9. Sonar data: challenges

Common sonar systems that are frequently used for seabed classification are MBES, side scan sonar, and recently, SBES. The following sections will focus of some of the challenges researchers are often faced with when dealing with MBES and SBES data in the context of seabed classification.

MBES systems have become very popular since its introduction to public domain in the late 70s for their ability to provide high-resolution backscatter information on a large area in fewer survey runs compared to SBES. But SBES systems have seen a recent interest as they can provide vertical backscatter information, which contains seabed information with minimal distortions as well as subsurface information as the echoes can penetrate the seabed several meters.

Large datasets

Due to their capability to capture high-resolution backscatter returns, surveys from current MBES and SBES systems often result in very large volume of data. The size of a backscatter and bathymetric dataset can be several terabytes (TB). Pre-processing these datasets is computationally intensive and can therefore be time consuming. Clustering of the processed data can also be challenging. Most of the popular clustering algorithms were not developed to accommodate such large datasets. Therefore, it is always a possibility that the algorithms may fail in their clustering attempt and the data need to be further reduced (for example: by sampling at equal interval, feature extraction etc.).

Feature extraction: MBES

Feature extraction is an effective way to reduce the volume of data generated from seabed surveys. Statistical features (statistical descriptors) are extracted using different techniques from segments of MBES backscatter returns as described previously. The number of features can vary from traditional statistical descriptors (mean, standard deviation) to more specific descriptors (such as: fractal dimensions, texture analysis etc.). A large number of statistical features are usually generated for MBES (around 130) (Le Gall, 1993; Hellequin, 1998; Milvang et al., 1993; Pace & Gao, 1988).

These statistical features are generated with the main objective to capture as many useful aspects of the data as possible. While extracting the features, selection of features i.e. type and number of features that are to be generated are frequently examined. This is done to see which features can provide useful discrimination and also to keep check if any algorithm is consistently producing redundant features. Each set of statistical descriptors from each segment is known as Full Feature Vectors (FFVs). There is no consensus as to the number and type of features that are generated. For example, from several studies (Preston et al., 2001; 2003), it was evident that the mean contributed least towards discrimination of seabed classes and redundancy is likely to exist when numbers of features are higher.

Feature extraction: SBES

Although SBES backscatter datasets potentially contain valuable seabed classification information, they typically receive less attention in this context. This is because SBES was not designed for that purpose and is mostly used for high resolution bathymetry determination. It is fairly recent that SBES's potential in seabed classification is explored. The advantage of SBES beams is that they are sent down as vertical beams directly beneath the survey ship and the echosounder receives the vertical reflections off the seabed. Unlike angular beams, the vertical beams contain the seabed information with minimal distortions. SBES beams can also penetrate the seabed for several meters and can provide vital sub-surface information.

The two software platforms that are most commonly used for feature extraction from SBES data are RoxAnn^{TM} and QTC Impact^{TM}. RoxAnn^{TM} systems derive feature values from the tail section of the echo time series using an echo-integration methodology and QTC Impact^{TM}, on the other hand, uses only the first part of the echo returns to extract features using principal component analysis (PCA) followed by k-means clustering. Both methods suffer from the noisy nature of sonar signals (Satyanarayana et al., 2007; Zimmermann & Rooper, 2008), contributing uncertainty to seabed classification (Peter, McLoone, & Monteys, 2010).

Addressing the challenges

In this thesis we explore alternatives to address some of the challenges outlined above. In particular, we focus on applying methods from computational data mining and Visual Analytics to improve the process of seabed classification from both MBES and SBES data and a combination of both data types.

A better application of data mining techniques can help reduce the complexities in analysing both MBES and SBES data. There are many data mining and visual exploration approaches available that can be applied to sonar data to facilitate identification of underlying patterns and reduce computational complexity. The following sections will briefly discuss the chosen techniques for both MBES and SBES data.

2.10. Data mining, Visual Analytics and applications to sonar data

In the last couple of decades, with the availability of massive storage systems and high-resolution sensors, we have been inundated with large volume of remotely sensed data. The major problem with an increasing volume of data is that people's understanding of the data decreases alarmingly. Potentially useful information that is hidden in the data can therefore go undetected or not taken advantage of. Data mining techniques, in their simplest form, require identification of a problem, along with collection of data that can lead to better understanding, and computer models to provide statistical or other means of analysis. The process of data mining is often supported by visualisation tools, that display data, or through some fundamental statistical analysis (for example: correlation analysis), all for the purpose of knowledge discovery. The ultimate objective of this process is finding patterns in data with the data stored electronically and the search for pattern is automated or at least augmented. But this is not a new idea. It has long been accepted that patterns in data can be sought automatically, identified, validated and can even be used for prediction. What is new with combined computational, visual and statistical data mining approach are increased opportunities for finding and describing patterns in extremely large datasets, a tool for helping to explain that data and make predictions from it, which is otherwise difficult if not impossible.

A recent approach to data mining is incorporating visualisation methods to aid in the process of knowledge discovery. For both MBES and SBES, such alternative combined data mining techniques can provide a way facilitate and improve classification. The focus of this research project is to incorporate this approach to help improving seabed mapping from both MBES and SBES data. A brief account of historical progress of visualisation technique is given below.

Brief history of visualisation technique

There are many historical accounts of developments within the fields of probability (Hald, 1990), statistics (Pearson, 1978; Porter, 1986), astronomy (Riddell, 1980), cartography (Wallis & Robinson, 1987), which relate to some of the important developments contributing to modern data visualization. There are other, more specialized accounts, which focus on the early history of graphic recording (Hoff & Geddes, 1959, 1962), statistical graphs (Funkhouser, 1936, 1937; Royston, 1956; Tilling, 1975), fitting equations to empirical data (Farebrother, 1999), economics and time-series graphs (Klein, 1997), cartography (Friis, 1974) and thematic mapping (Robinson, 1982), and so forth; a detailed overview of some of the important scientific, intellectual, and technical developments of the 15th–18th centuries leading to thematic cartography and statistical thinking can be found in Robinson (1982). Wainer and Velleman (2001) provide a recent account of some of the history of statistical graphics.

The earliest visualization arose in geometric diagrams, in the positions of stars and other celestial bodies, and in the making of maps to aid in navigation and exploration. The idea of coordinates was used by ancient Egyptian surveyors circa 200 BC in laying out towns as well as positions of earthly and heavenly holy subjects. The map projection of a spherical earth into latitude and longitude by Claudius Ptolemy in Alexandria would serve as reference standards until the 14th century (Friendly, 2005, 2008; Robinson, 1982).

In the 14th century, the idea of a plotting a theoretical function (as a proto bar graph), and the logical relation between tabulating values and plotting them appeared in a work by Nicole Oresme [1323–1382] Bishop of Liseus (Oresme, 1968). By the 16th century, techniques and instruments for precise observation and measurement of physical quantities, and geographic and celestial position were well developed (for example, a "wall quadrant" constructed by Tycho Brahe [1546–1601], covering an entire wall in his observatory). Particularly important developments during this period were the triangulation and other methods to determine mapping locations accurately and the first modern cartographic atlas (Teatrum Orbis Terrarum by Abraham Ortelius, 1570) (Funkhouser, 1937).

These early steps comprise the beginnings of data visualization. By the end of 17th century, the necessary elements for the development of graphical methods were at hand— some real data of significant interest, some theory to make sense of them, and a few ideas for their visual representation (Funkhouser, 1937; Robinson, 1982).

The 18th century witnessed the expansion of visualization and new graphic forms. Iso-lines and contours as well as thematic mapping of physical quantities were regularly used in cartography. The two most prominent contributors in graphical methods were also from this century. First is Johann Lambert (1728-1777) who introduced the ideas of curve fitting and interpolation from empirical data points. Another is William Playfair (1759-1823), widely considered the inventor of most of the graphical forms widely used today- the line graph in 1786, pie chart and circle graph in 1801 (Friendly, 2008). The 19th century witnessed the explosive growth in statistical graphics and thematic mapping. In statistical graphics, all of the modern forms of data display were invented: bar and pie charts, histograms, time-series plots, scatterplots, and so forth. In thematic cartography, mapping progressed from single maps to comprehensive atlases, depicting data on a wide variety of topics (economic, social, medical, physical etc.) During this period graphical analysis of natural and physical phenomena (weather, tides, etc.) began to appear regularly in scientific publications (Friendly, 2008; Friendly & Denis, 2005; Robinson, 1982).

In the late 20th century, John W. Tukey introduced a variety of new, simple and effective graphical displays under the rubric of “Exploratory Data Analysis (EDA)”- stem-leaf plots, box plots, hanging rootograms, two-way table displays etc. (Tukey, 1977). By the end of this century, with the advancement in computing technology, graphical terminals and plotters would lead a tremendous growth in new visualisation methods and techniques (Andrews, 1972; Friendly, 2005, 2008).

2.10.2. From data mining towards visual analytics

Today’s computer systems allow us to store and exchange amounts of data that until very recently were considered extraordinarily vast. Data collected are considered a potential source of valuable information, providing a competitive advantage to its holders. The data are often automatically recorded via sensors and monitoring systems and many parameters are usually recorded, resulting in data with a high dimensionality.

Spatial datasets that contain both location and attribute information are more and more commonly encountered in many areas, including environmental science. Such datasets are often extremely large, which makes the task of discovering meaningful information and knowledge in these datasets very difficult. With most of today’s data management systems, it is only possible to view quite small portions of these data. Having no possibility to adequately explore the large amounts of data that have been collected because of their potential usefulness, the data becomes useless and the databases become ’Data Dumps’ (Keim et al., 2003; Keim et al., 2008).

This unprecedented data explosion resulted in a need for an alternative data mining approach. It is estimated that about 50% of human brain’s neurons are associated with vision and human visual perception system and pattern recognition skills are considered extremely efficient (Simoff et al., 2008). One of the recent approaches to knowledge discovery in large datasets is therefore to use a combination of computational data mining and visual data exploration techniques. The discipline concerned with this approach is Visual Analytics, a recent new sub discipline of Information Visualisation, and in the case of spatial and spatio-temporal data, Geovisual Analytics (Andrienko et al., 2007). Visual Analytics looks at the integration of data mining and analytical reasoning for the purpose of exploring spatial data, where the process is supported by interactive visual interfaces and information visualisation methods. It is multidisciplinary and includes methods from information visualisation and data analysis (including statistics, data mining and mathematical modelling), but also looks at cognitive aspects of how humans perceive and use computerised visualisations (Andrienko et al., 2007). Figure 2.13 shows the overall process of visual data mining.

According to Simoff et al., (2008), data are first pre-processed. Then one or a number of visualisation techniques is selected by the user the data are mapped to visual representation.

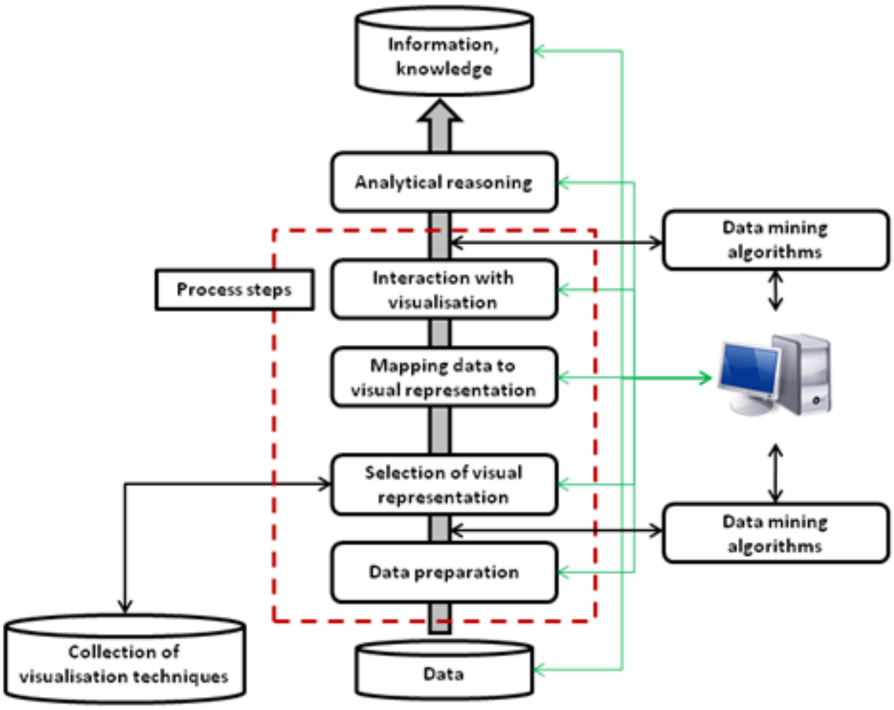

Information, knowledge
Analytical reasoning
Data mining algorithms
Process steps
Interaction with visualisation
Mapping data to visual representation
Selection of visual representation
Data mining algorithms
Data preparation
Collection of visualisation techniques
Data

Figure 2.13: Visual data mining process (adapted from Simoff et al., 2008)

The user then applies a combination of visual interaction and analytical reasoning to infer on the underlying pattern in data. The knowledge acquired from this interaction and reasoning is stored for further validation. The choice of data visualisation tool depends on the nature of the dataset and its underlying structure (Ankerst et al., 2000; Siebes, & Wilhelm, 2000).

Contemporary techniques in visualisation (Guo et al., 2005; Yan & Thill, 2007) have been brought to the force in recent years owing to advances in technology, but the essential aim remains the same: ‘to turn large heterogeneous data into information (interpreted data) and subsequently, into knowledge (understanding derived from information)’ (Hernandez, 2007). Typically, visualisation of spatial data is concerned with what might be termed ‘spatial ontology’ in that it seeks to discover the relationship between objects from a spatial perspective, or, more specifically, ‘what exists where’ (Galton, 2003; Goodchild et al., 2007; Longley et al., 2005).

Today, any visualization system must answer the these fundamental queries:

- What characteristics the visualization systems hold?
- What kinds of data should it support?
- What capabilities should it provide?

The answers to these queries almost entirely depend on the particular task and application. For some users a visualization system may be nothing more than a simple image viewer or plotting program. For others it is integrated software dedicated to their personal field of work, such as a computer algebra program or a finite-element simulation system. While in such integrated systems visualization is usually just an add-on, there are also many specialized systems whose primary focus is upon visualization itself (Longley et al., 2005).

On one hand, there are many self-contained special-purpose programs written for particular applications. Examples include flow visualization systems, finite-element post-processors, and volume rendering software for medical images. On the other hand, several general-purpose visualization systems have been developed since scientific visualization became an independent field of research in the late 1980s. These systems are not targeted to a particular application area, but provide many different modules which can be combined in numerous ways, often adhering to the data-flow principle and providing means for ‘visual programming’ (Abram & Treinish, 1995; Dyer, 1990; Foulser, 1995; Upson et al., 1989).

Visualisation tools can be used to visualise the effectiveness of the data mining model, as well as to analyse the potential deployment of the model. The gains chart, for example, provides a visual summary of the usefulness of the information provided by one or more statistical models for predicting a target event in comparison to always guessing it occurs. It can also be used to compare and contrast the performance of the models at the time they are built and once they are deployed.

Other multidimensional data visualisation tools are useful in analyzing the data mining model results, as well as comparing and contrasting multiple data mining models. These are just a few examples of how we can use data visualisation to explore the decision making and evaluation processes of data mining models. However, choosing the right kind of visualisation tools is absolutely imperative in the success of the research objective and often require substantial amount of testing before implementation (Soukup & Davidson, 2002).

For visual data mining to be effective, it is important to include the human in the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today's computers. Visual data exploration aims at integrating the human in the data exploration process, applying its perceptual abilities to the large data sets. The basic idea of visual data exploration is to present the data in some visual form, allowing the human to get insight into the data, draw conclusions, and directly interact with the data. Visual data mining techniques have proven to be very effective in exploratory data analysis and they also have a high potential for exploring large databases. Visual data exploration is especially useful when little is known about the data and the exploration goals are vague (Keim et al., 2008; Longley et al., 2005).

2.10.3. Data mining & visual analytics: a comparison

While both approaches aim to analyse and enhance the underlying properties of data, there are some differences between them. Data mining is computer-centred. A computer performs data analysis and a human analyst somehow interprets and uses the results. Visualisation may be involved in the data mining process. The main objective of using visualisation for data mining is helping the user to understand the results and sometimes enabling the user to select and prepare the data for input. It, at times, also enables the user to direct the work of the algorithm. Visual Analytics, on the other hand, is human-centred. A human solves a complex problem and the computer is there to aid in the process of problem solving. Visualisations are needed for activating the perceptual and cognitive capabilities of the human. These capabilities include perception of patterns, identification and association, abstraction and generalisation, and reasoning and insight (Andrienko & Andrienko, 2009).

Visual analytics seeks to maximise human capacity to perceive, understand, and reason about complex and dynamic data and situations and to augment this cognitive reasoning with perceptual reasoning through visual representations and interaction. It aims to transform data into a representation that is appropriate to the analytical task and to effectively convey the important contents and analytical results to various audiences in a meaningful way (Keim, 2002; Andrienko and Andrienko, 2009).

An advantage of visual analytics is that in this approach the computers and humans work synergistically. Computers can store and process great amounts of information and are very fast in searching information. They are very fast in processing data and can extend their capacities by linking with other computers. They can also efficiently render high quality graphics, both static and dynamic. Humans, on the other hand, are flexible and inventive. They can deal with new situations and problems effectively and can solve problems that are hard to formalise. Humans can reasonably act in cases of incomplete and/or inconsistent information and can simply see things that are hard to compute. They can employ their knowledge and experience in different situations without much difficulty. What visual analytics does is to combine the best of both i.e. combines the capability of heavy duty computing of modern day computers with the extremely effective visual perception system and pattern recognition skills of humans (Andrienko and Andrienko, 2009; Simoff et al., 2008; Keim, 2002).

2.10.4. Data visualisation

Visual data exploration typically follows a three-step process: Overview of data, zoom and filter, and then details on demand - which has often been regarded in the visual analytics domain as the information seeking mantra (Shneiderman, 1996). In the overview, the user identifies interesting patterns and focuses on one or more of them. For analysing the patterns, the user needs to explore the details of the data. An efficient way is to distort the overview visualisation in order to focus on the interesting subsets. To further explore the interesting subsets, the user needs a drill-down capability in order to get the details about the data (Shneiderman, 1996).

There are a number of well-known techniques for visualizing datasets in visual data mining (x-y plots, line plots, and histograms etc.). But these commonly used techniques, though useful for data exploration, are limited to relatively small and low-dimensional data sets. In the recent decades, a number of novel information visualisation techniques suitable large datasets have been developed (Card et al., 1999; Spence, 2007; Ware, 2000).

The type of visualisation to be used depends mostly on three criteria (Keim, 2001, 2002): the data to be visualized, the visualisation technique, and the interaction and distortion technique. The data can be one-dimensional (ex: time series), two-dimensional (ex: a geographical map), or multivariate (ex: relational tables) in nature (Abello & Korn, 2002; Havre et al., 2002; Kreuseler et al., 2000; Shneiderman, 1992; Stolte et al., 2002). The dataset can also comprise text and hypertext, such as news articles and web documents (Keim, 2002; Havre et al., 2002) or can have inherent hierarchy and graphs such as cell phone calls, web documents etc. (Kreuseler et al., 2000).

The data type influences heavily the choice of visualisation. Depending on the data, it can be standard 2D/3D displays, such as bar charts and x-y plots (Stolte et al., 2002) or it can be geometrically transformed displays, such as landscapes and parallel coordinates (Kreuseler et al., 2000). It can be icon-based displays, such as needle icons and star icons (Abello and Korn, 2002), dense pixel displays, such as the recursive pattern and circle segments techniques (Keim, 2000). Some other examples of visualisation techniques are graph sketches (Abello and Korn, 2002), stacked displays such as treemaps (Johnson & Shneiderman, 1991; Shneiderman, 1992) or dimensional stacking (Ward, 1994).

The third factor that influences the choice of visualisation technique is the interaction and distortion technique used. This allows users to directly interact with the visualisations and can be of several types: interactive projection (Asimov, 1985), interactive filtering (Chris Stolte, Tang, et al., 2002), interactive zooming and distortion (Kreuseler et al., 2000), and interactive linking and brushing (Stolte et al., 2002; Kreuseler et al., 2000).

All the three factors (data type to be visualized, visualisation technique, and interaction & distortion technique) that influence the choice of visualisation can be assumed to be orthogonal. This means that for any data type, any of the visualisation techniques may be used in conjunction with any of the interaction or distortion techniques. A system that is designed to deal with multiple data types may use a combination of multiple visualisation and interaction techniques (Keim, 2001; Keim, 2002).

One of the major focuses of this research is interactive visualisation system. For an effective data exploration it is necessary to use some interaction and distortion techniques. Interaction techniques allow the data analyst to directly interact with the visualisations and dynamically change the visualisations according to the exploration objectives, and they also make it possible to relate and combine multiple independent visualisations. Distortion techniques help in the data exploration process by providing means for focusing on details while preserving an overview of the data. The basic idea of distortion techniques is to show portions of the data with a high level of detail while others are shown with a lower level of detail. Below are short descriptions of some common interactive visualization techniques.

Dynamic Projections

The basic idea of dynamic projections is to dynamically change the projections in order to explore a multidimensional data set. A classic example is the Grand-Tour system (Asimov, 1985), which tries to show all interesting two-dimensional projections of a multi-dimensional data set as a series of scatter plots. The sequence of projections shown can be random, manual, pre-computed, or data driven. Systems supporting dynamic projection techniques are XGobi (Buja & Cook, 1996; Swayne et al., 1992), XLispStat (Tierney, 1991), and ExplorN (Carr et al., 1996).

Interactive filtering

The exploration of datasets is usually done by a direct selection of the desired subset (browsing) or by a specification of properties of the desired subset (querying). Browsing is very difficult for very large datasets and querying often does not produce the desired results. Therefore a number of interaction techniques have been developed to improve interactive filtering in data exploration. An example of an interactive tool that can be used for an interactive filtering is Magic Lenses (Bier et al., 1993; Fishkin & Stone, 1995). The basic idea of Magic Lenses is to use a tool like a magnifying glass to support filtering the data directly in the visualisation. The data under the magnifying glass are processed by the filter, and the result is displayed differently than the remaining data set. Other examples of interactive filtering techniques and tools are InfoCrystal (Spoerri, 1993), Dynamic Queries (Ahlberg & Shneiderman, 1994; Eick, 1994; Goldstein & Roth, 1994), Polaris (Stolte et al., 2002).

Interactive zooming

Zooming is a well-known technique, which is widely used in a number of applications. In dealing with large amounts of data, it is important to present the data in a highly compressed form to provide an overview of the data but at the same time allow a variable display of the data on different resolutions. Zooming does not only mean to display the data objects larger but it also means that the data representation automatically changes to present more details on higher zoom levels. The objects may, for example, be represented as single pixels on a low zoom level, as icons on an intermediate zoom level, and as labeled objects on a high resolution. Examples of techniques and systems, which use interactive zooming, include TableLens approach (Rao & Card, 1994), PAD++ (Bederson, 1994; Bederson & Hollan, 1994; Perlin & Fox, 1993), IVEE/Spotfire (Ahlberg & Wistrand, 1995), and DataSpace (Anupam et al., 1995).

Interactive Distortion

Interactive distortion techniques support the data exploration process by preserving an overview of the data during drill-down operations. The basic idea is to show portions of the data with a high level of detail while others are shown with a lower level of detail. Examples of distortion techniques include Bifocal Displays (Spence & Apperley, 1982), Perspective Wall (Mackinlay et al., 1991), Graphical Fisheye Views (Furnas, 1986; Sarkar & Brown, 1994), Hyperbolic Visualization (Lamping & Pirolli, 1995; Munzner & Burchard, 1995), and Hyperbox (Alpern & Carter, 1991).

Interactive Linking and Brushing

There are many possibilities to visualise multidimensional data but all of them have some strength and some weaknesses. The idea of linking and brushing is to combine different visualization methods to overcome the shortcomings of single techniques. Scatterplots of different projections, for example, may be combined by coloring and linking subsets of points in all projections. In a similar fashion, linking and brushing can be applied to visualisations generated by all visualisation techniques described above. Interactive changes made in one visualisation are automatically reflected in the other visualisations. Typical examples of visualization techniques which are combined by linking and brushing are multiple scatterplots, bar charts, parallel coordinates, pixel displays, and maps. Most interactive data exploration systems allow some form of linking and brushing. Examples are Polaris (Stolte et al., 2002) and the Scalable Framework (Kreuseler et al., 2000). Other tools and systems include S Plus (Becker et al., 1988), XGobi (Becker et al., 1996; Swayne et al., 1992), Xmdv (Ward, 1994), and DataDesk (Velleman, 1992; Wilhelm et al., 1995).

In the context of this thesis and visualising sonar data - multidimensional and time series visualisations are considered relevant and are briefly discussed in the following sections.

Multivariate data visualisation tools

These are the most commonly used data visualisation tools. These tools enable users to visually compare data variables (column values) with other data variables using a spatial coordinate system (Soukup & Davidson, 2002). Figure 2.14 shows examples of some of the most common visualisation graph types. Other common multivariate graph types not shown in Figure 2.14 include contour, histogram, error, Westinghouse, colour grid and box graphs. A detail outline on multivariate visualisation tools can be found in Harris (1999).

Most multivariate visualisations are used to compare and contrast the values of one column (data dimension) to the values of other columns (data dimensions) as well as to investigate the relationships between two or more continuous or discrete columns in the dataset. Table 2.1 lists some common multivariate graph types and their functions (Soukup and Davidson, 2002).

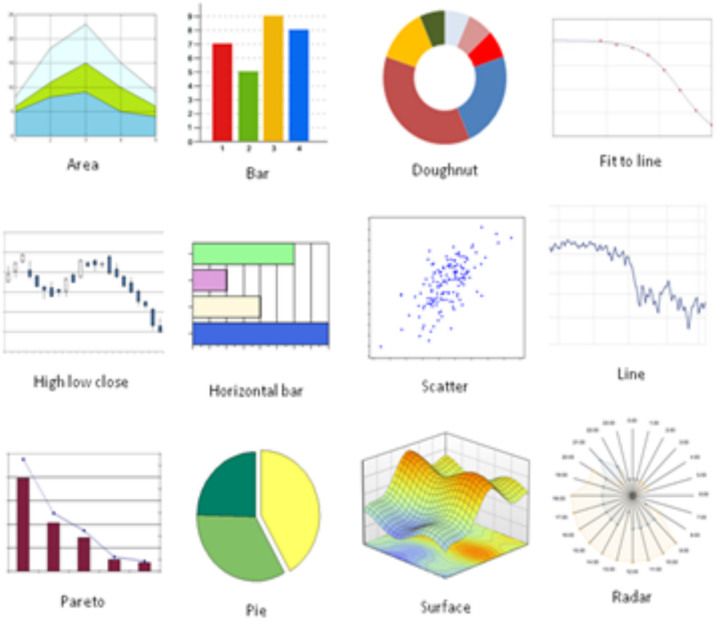


Figure 2.14: Commonly used visualisation tools for multivariate data (adapted from Harris, 1999)

2.1: Functions of common multivariate visualisation tools (Soukup and Davidson, 2002)

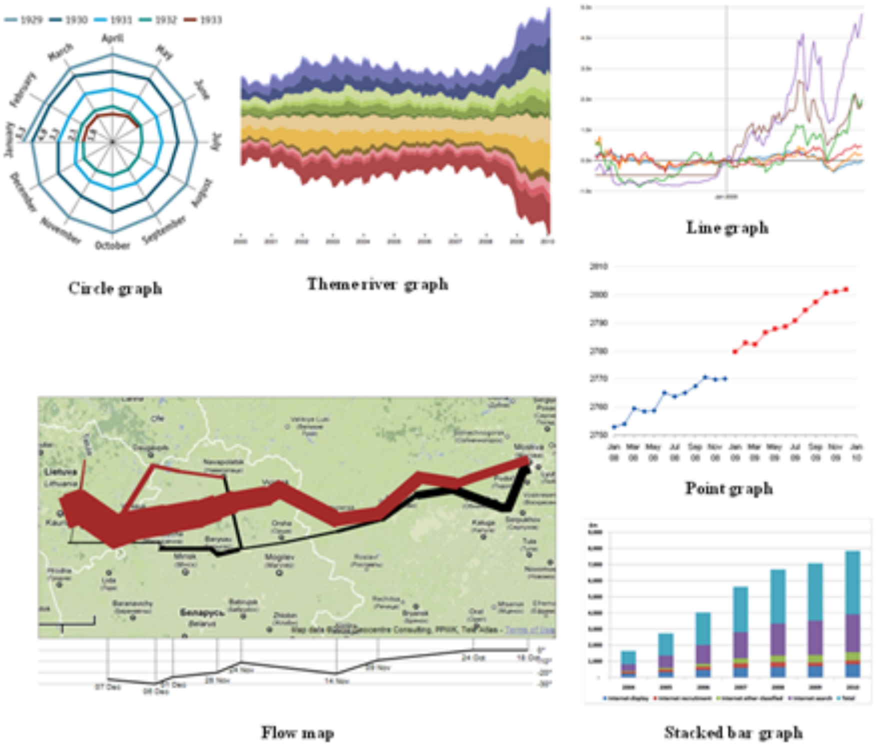
Graph type	Function
Column and bar	compare discrete (categorical) column values to continuous column values
Area, bar, line, high-low-close, and radar	compare discrete (categorical) column values over a continuous column
Pie, doughnut, histogram, distribution, box	compare the distribution of distinct values for one or more discrete columns
Scatter	investigate the relationship between two or more continuous columns

Time series visualisation

Recent advancements in sensor technology have made it possible to collect enormous amounts of time series data, often in real-time. Time series data are data elements that are a function of time. The data elements can represent different data types, for example nominal, ordinal, and quantitative data or tuples of these in the case of multivariate data (Marc et al., 2001). Appropriate visualisation of time-series data is invaluable for exploring temporal data searching for underlying patterns that were previously unknown (Hao et al., 2007; MacEachren, 1995).One of the earliest records of time series plots dates back to the 10th or 11th century from a text from a monastery school (Tufte, 2001). Lambert was one of the first scientists to re-discover the application of time series charts in the 18th century. In his book “Pyrometrie oder vom Maasse des Feuers und der Wärme” published in 1779, he displayed periodic variation in soil temperature in relation to depth under the surface using line graphs (Muller & Schumann, 2003).

The most common and frequently used visualisation techniques for time series data are sequence charts, point charts, bar charts, line graphs, and circle graphs (Marc et al., 2001). The conventional methods of visualising time dependant data (Figure 2.15) allow the conclusion of quantitative statements and facilitate the exploration of special data features and patterns as well as data values, time steps and positions according to the underlying scales without temporal limitations. As in standard cases, the choice of a technique depends on the kind of data available (i.e. point graphs for point data, line graphs for continuous data, bar graphs for cumulative data, and circle graphs for cyclic data) (Harris, 1999).

More complex graphs can be generated by mapping a static graph as an independent representation of the data element d_i for a time-step t_i (i= instance of time) onto a more general graph representation for more than one data element (Müller and Schumann, 2003). Chess plot is more appropriate if the target graph is a sequence graph (Monmonier, 1990). Change chart, stacked bar chart, parallel coordinate plots (by mapping the different time-steps to the individual axis) are some examples of visualisation technique that allows the linking of independent representations of data for each time-step to a single map (Figure 2.15) (Inselberg, 1997; Muller & Schumann, 2003). The main limitation of this type of visualisation is that in most of the cases, representations are limited to a single variable over several time steps or a limited number of time variables and time steps. (Müller and Schumann, 2003).



Circle graph

Themeriver graph

Line graph

Point graph

Flow map

Stacked bar graph

Figure 2.15: Some conventional time series plots (derived from Müller and Schumann, 2003)

A substantial body of work can be found in the development of strategies for storing and indexing time series data. Algorithmic and statistical methods for identifying patterns have also provided substantial functionality to deal with time series data in a wide variety of situations (Berndt & Clifford, 1996; Faloutsos et al., 1994).

A major drawback of algorithmic research is that it only addresses one aspect of the data mining problem. The question of query formulation i.e. selecting the questions that are worth asking is often left unanswered (Aris et al., 2005; Buono et al., 2007; Hochheiser & Shneiderman, 2001; Shmueli et al., 2006). For example, problems involving identification of outliers in a time series dataset involves detecting time series that are similar to a known series (query sequence) thus isolating the records that are not or detecting the records that are dissimilar to the rest. This involves specification of several query parameters. In addition to the input query, users must provide parameters defining the range of allowable similarity or dissimilarity. But the identification of parameters such as these requires trial-and-error processing, which is often challenging and computationally expensive. A central problem for users is that the effects of small changes on parameters such as similarity or dissimilarity tolerances may be hard to gauge without running multiple trials (Buono et al., 2007; Shmueli et al., 2006; Aris et al., 2005; Hochheiser and Schneiderman, 2001).

Visualisation of time series data introduces the various perspectives that may be suitable for interpreting these data sets and was influenced by factors such as periodicity (Brewer et al., 2000; Carlis & Konstar, 1998), multiple scales of resolution (Keim, 1996; Powsner & Tufte, 1994; van Wijk & Van Selow, 1999), and the need to display multiple variables at each time period (Carlis & Konstar, 1998; Powsner & Tufte, 1994). Tools such as QuerySketch (Wattenberg, 2001) and Spotfire (Spotfire, 2011) enable the querying of time series datasets but do not fully meet the need for interactive visualisation.

Interactive tools and visualisations are usually focused on searches for patterns involving well-specified changes over well-defined time periods. Data mining algorithms for time series, on the other hand, are more ambitious in the sense that they often address the challenge of finding patterns that occur at arbitrary times and are assumed “similar” in some general manner that can often account for variations in scale and duration, discontinuities or other features (Berndt & Clifford, 1996; Rafiei & Mendelzon, 2000; Yi et al., 1998).

Combining the interactivity of dynamic query tools with the power of these data mining approaches presents several challenges. In order to support these algorithms, a query interface must include mechanisms for specifying tolerances of approximate fits, lengths of allowable gaps, tolerances in time dilation or contraction, and other constraints. Visualisation of the query results can also be challenging. The visualisations would need to display the results with sufficient contextual information to explain why the result was a match. The requirement of combining the rapid, incremental updates of information visualisation with the computational requirements of data mining can also prove to be difficult in the implementation of interactive visualisation tools (Hochheiser & Shneiderman, 2001; Hochheiser, 2003).

2.10.5. Visual analytics and sonar data

As mentioned in section 2.8, seabed surveys result in large volumes of data from each survey. It is computationally intensive and difficult to explore such data and to produce reliable seabed maps (Preston et al., 2004). The goal of this thesis is to use visual analytical methods to improve the process of seabed classification from sonar data. In particular, the main focus of this thesis is on three issues:

- Removing redundancy in MBES feature data
- Using visual exploration and time series data mining to support seabed classification from SBES data
- Evaluating new seabed mapping methodologies for combined MBES and SBES data

Here we briefly introduce each of these problems and in the remainder of this section give an overview of relevant visualisations and data mining methods.

For MBES data, PCA is used to reduce the dimension from 132 statistical features and the first principal three components are usually used for k-means clustering. The first three components contain somewhere between 90-95% of the information (Preston et al., 2004; Preston, 2009). However, important information can be lost in the remaining 5-10%, especially if the seabed is rapidly varying. An alternative approach can be to optimise the number of statistical features so that only the numbers of statistical features that are required to describe the data completely are selected. This approach is based on the assumption that 132 statistical features have redundancy in them (Preston, 2009). Visual analytics can provide a technique to explore the underlying redundancy in MBES features and provide an alternative method for reducing dimensionality such that all relevant information is preserved (Soukup and Davidson, 2002).

We chose to use Kohonen's Self Organising Map (SOM) (Kohonen, 1989; Skupin & Agarwal, 2008) to explore the distribution of attributes and detect redundancy in MBES features dataset. SOM provides a unique feature called component planes that shows the attribute distribution in the feature space. Similar attributes will show similar distribution 10/03/2012 14:51. This feature has the potential to reduce bias as the component planes distribution can be subject to interpretation of diffident users.

SBES data consist of univariate time series measured at various spatial locations i.e. at each spatial location, SBES emits an acoustic wave and after the time delay (time required for the reflected echo to reach echousounder receiver) the receiver starts to record the backscatter at pre determined time interval. The time interval is dependent on the working frequency of the echosounder. The interval is greater at lower frequency and less at smaller working frequencies. This time interval together with the seabed depth determines the data dimension of each time series for example, the data dimension increases with increasing survey frequency and/or seabed depth and decreases with low frequency and/or shallow seabed depth. The challenge is with the high number of time series records available for the survey area. To get the data into a more manageable form, it is usually scaled down before any pre-processing. Because of the dense proximity of the time series measurements in geographical space, the standard approach using bar- and line-charts (Hao et al., 2007) is ineffective for visual analysis of these data.

Another problem is outlier detection. Although there are a number of computational algorithms that are available to detect outliers in time series data (Hung et al., 2010), with large volumes of data there is always a possibility that some outliers can go undetected. Visual exploration can be of use to identify outliers by using humans' capability of detecting anomalies in the distribution of data.

Here we use the 'TimeSearcher©' tool (Hochheiser & Shneiderman, 2001; Keogh et al., 2002), which provides us with an augmented visual query mechanism for finding patterns in time series data (Hochheiser, 2003; Keogh et al., 2002) and can be useful in detecting outliers in SBES data. The interactive distortion capability in TimeSearcher© allows data exploration by preserving the overview of data during drill-down operations i.e. shows portion of the data with emphasis while others are shown with a lower level of detail (Hochheiser & Shneiderman, 2001; Keogh et al., 2002).

The new MBES echosounders also come equipped with SBES echosounders and simultaneous measurements are acquired in the same survey run. In the case of a heavy presence of sea plants, shells, boulders etc., the MBES classification can be noisy at times. In these circumstances, information from SBES data can be particularly useful as SBES echoes can penetrate the seabed for several meters. The echo returns from the sub-surface can give us an understanding of the underlying layers where MBES does not penetrate. Optimal features from MBES (see Chapter

3) will be combined with statistical features generated from SBES backscatter. The resulting dataset will be clustered using fuzzy c-means as fuzzy logic allows overlapping of clusters, which is expected in the case of combined MBES and SBES where seabed surface echo returns from MBES can differ from sub-surface returns of SBES in the same location.

2.11. Visual and computational data mining methods used in this thesis for sonar data

In this section we discuss the visual and computational data mining methods that were used to address the research challenges presented previously. These include: A Self Organising Map and its visualisations, time series visualisation in TimeSearcher© and various clustering methods used for classification of sonar data.

2.11.1. Self Organising Maps (SOM)

Visual data mining tools can also be used to visualise the outputs of the data mining model and the selection of the visualisation tool depends on the nature of dataset and the underlying structure of the resulting model. However, not all data mining algorithms can be visualised with ease. One such example is neural networks. Neural network models simulate a large number of interconnected simple processing units segmented into three steps (input, hidden, and output layers). Visualising the entire network with its inputs, connections, weights, and outputs as a two- or three-dimensional picture can be very challenging and continues to be a matter of research (Craven & Shavlik, 1991; Lang & Witbrock, 1988; Soukup & Davidson, 2002; Uzak et al., 2008).

The Self-Organising Map (SOM) is an artificial neural network used for unsupervised classification of data. It maps multi-dimensional data onto a lower dimensional space, usually represented as a two-dimensional hexagonal lattice. The mapping function preserves the probability density and the aspatial topology of the input data, which means that the patterns in the attribute space are preserved in the result space. As an unsupervised neural network, the SOM uses similarity between input data objects as its only measure of separating data into groups (clusters) and is therefore a data-driven method (Deboeck & Kohonen, 1998).

The SOM mapping is typically from the multi-dimensional data space with n dimensions (attributes) onto a lattice of usually hexagonal cells, which represent neurons in the neural network. Each of the cells is assigned a vector of weights at the beginning of the procedure. The SOM algorithm then takes each data object in turn and finds the location of the cell that is the best match for this data object. The data object is then placed into the cell and the weights of the cell and of the neighbouring cells are recalculated to reflect this best match. The process of recalculating weights is repeated after each data object, until all weights are stable. This is the training process of the SOM. Which cells are affected by the same data object is determined through a pre-defined neighbourhood kernel function. This function has the highest value at the best match cell and monotonically decreases with distance to finally reach 0 at a certain distance from the best match cell. This means that nearby cells of the SOM all learn from the same data object and as a consequence, when this procedure is repeated over and over, it means that in the final model, similar data objects are mapped to cells in close proximity to each other (Figure 2.16). The location of the data objects in the SOM lattice therefore defines grouping of data according to their similarity (Kohonen, 2000).

Since the SOM preserves both the topology and the distribution of data objects, it is a useful knowledge discovery tool for spatial domain and has been used in a number of recent applications, for example, to list a few. An overview of SOM applications in GIScience can be found in Agarwal and Skupin (2008).

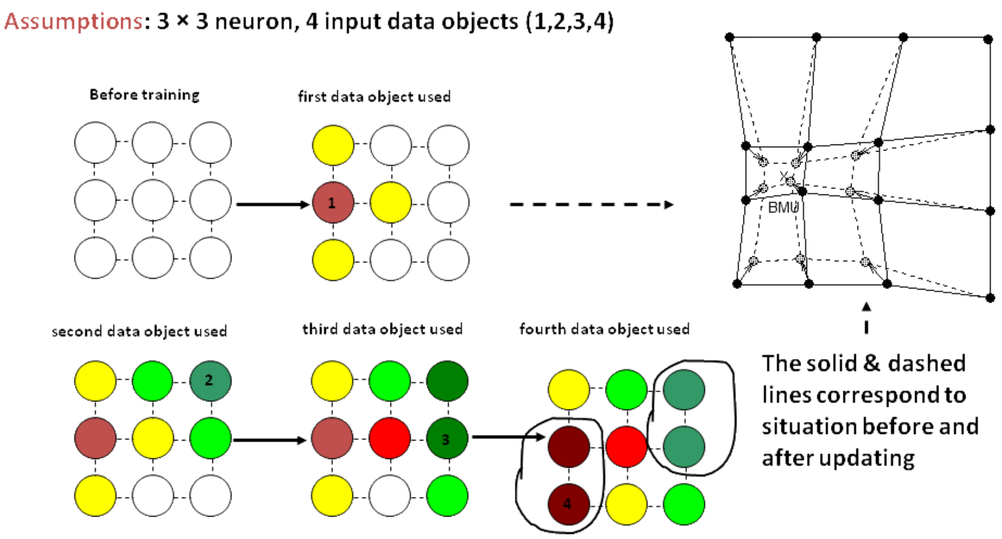

Assumptions: 3 × 3 neuron, 4 input data objects (1,2,3,4)
Before training
first data object used
1
second data object used
2
third data object used
3
fourth data object used
4
BMU
The solid & dashed lines correspond to situation before and after updating

Figure 2.16: Network training process in SOM

The result space of a SOM is a lattice and therefore two-dimensional, which makes it suitable to present graphically. Visualising a SOM means visually exploring the model that was created through network training. There are many different SOM visualisation methods (Vesanto & Ahola, 1999) and they can be used for different purposes: the most basic one is to show the similarity of data objects through their position in the lattice. The method that is of interest in this research project is the component planes visualisation, as this method allows us to explore attribute similarity (i.e. similarity between attributes and not between particular data objects), which was the purpose of our experiment in eliminating redundancy of MBES features. In the component planes visualisation, one SOM lattice is displayed for each attribute, i.e. there are as many lattices as there are attributes. Figure 2.17 shows a typical example of SOM component planes. After the data training process, neurons, which are represented by hexagonal cells are coloured according to the attribute values of the data object in the neuron cell. The distribution of data in the cells is defined by the SOM mapping and is the same in all lattices. The result is that component planes, which belong to correlated attributes, have similar colour distribution. The lattices can then be visually compared to each other to identify similar distributions of values in two or more attributes – these are revealed as similar patterns at identical locations in different lattices. These patterns are used to identify visual groups of similar attributes (Koua & Kraak, 2004; Vesanto, 1999).

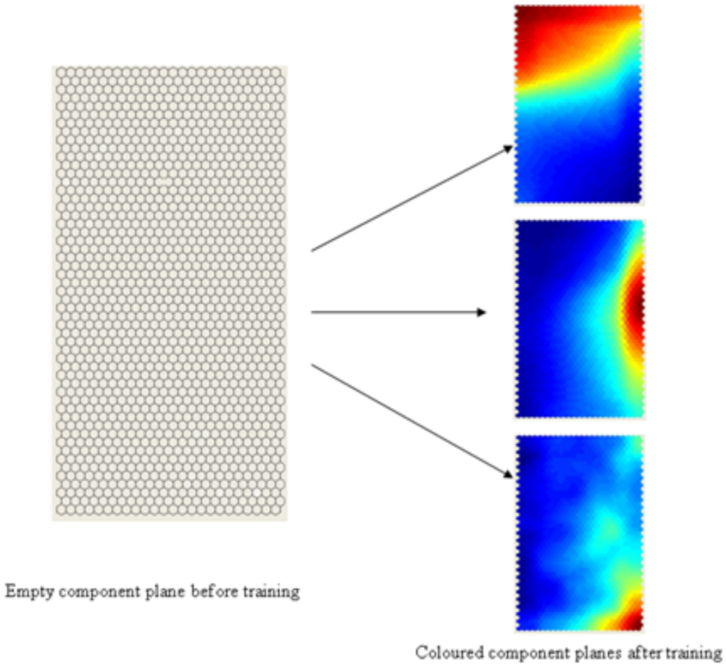

Empty component plane before training
Coloured component planes after training

Figure 2.17: SOM component planes

The 'SOM toolbox' (Vesanto, 1999) developed for Matlab environment was used for the visual exploration and evaluation of MBES feature datasets. Matlab is a high level programming language with a powerful visualisation and graphical user interface. It also has a very efficient implementation of matrix calculus. These features make Matlab a powerful tool for data mining research as they allow fast prototyping, testing and customising of the algorithms. In addition to these capabilities, there are also a large number of toolboxes intended for a variety of modeling and analysis tasks that can be added to the Matlab environment. These toolboxes are based on a wide span of methodologies from statistical methods to Bayesian networks or from mapping functions to optical/radar image analysis (Vesanto et al., 2000). The SOM Toolbox was developed to meet the on going need of an efficient, easy-to-use implementation of SOM in Matlab for research purposes. The main objective behind the development of the toolbox was to provide a simple yet powerful data mining tool with strong visualisation capabilities. The toolbox facilitates data processing, different topologies to initialise and train SOM, visualisation of SOMs in various ways, and analysis of the properties of the SOMs and data. Examples include SOM quality, clusters on the map and correlations between variables (Vesanto & Alhonierni, 2000; Vesanto, 1999; Vesanto et al., 2000).

The highlights of the SOM Toolbox include the following (J Vesanto & Alhoniemi, 2000; Juha Vesanto, 1997, 1999):

- Modular programming style: The functions in the toolbox are constructed in a modular manner. This means that the users can tailor the code to match their needs.
- Component names, masks and normalizations: Users can give different names to the input vectors and use different kinds of (reversible) preprocessing operations. The components can also be masked, or weighted.
- Batch or sequential training: Users can improve the training process considerably by using the batch version. There are also other training variants, like supervised SOM.
- Map dimension: The toolbox supports N-dimensional maps, although visualisation is limited to two dimensions.
- Advanced graphics: Matlab's strong graphics capabilities can also be implemented to generate figures.
- Graphical User Interfaces (GUIs): SOM toolbox come with some graphical user interfaces, although it is recommended to use the command line versions of the functions as they are more efficient.

2.11.2. Time series visualisation: TimeSearcher© and the time boxes

Interactive exploration of the contents of time series datasets can be a useful tool as it can enable the users to quickly construct queries, modify parameters, and examine results from the datasets. These tools also help quickly develop the understanding of the dataset as a whole, which is useful for guiding the construction of queries, thus facilitating knowledge discovery (Hochheiser and Schneiderman, 2001). The combination of graphic displays with easily customisable user-interface widgets for query formulation allows users to explore data sets easily (Ahlberg & Shneiderman, 1994). One such tool is TimeSearcher© (Hochheiser & Shneiderman, 2001; Hochheiser, 2003).

TimeSearcher© is a time series visualisation tool that allows interactive exploration of time series data (Buono et al., 2005; Hochheiser & Shneiderman, 2004). It is an information visualisation tool based on the use of Timeboxes (Hochheiser and Shneiderman, 2004) to pose queries over a set of entities with one or more time-varying attributes. TimeSearcher© is written in Java, using Piccolo for all graphics rendering and scenegraph management (Buono et al., 2005; Shmueli et al., 2006; Hochheiser and Shneiderman, 2004).

TimeSearcher© can display multiple time series representing multiple variables (e.g. precipitation, wind velocity, rainfall intensity etc.) and can also associate each item with a set of attribute data (metadata) that remain constant over time (e.g. car rating or auction start day). In this research project, TimeSearcher© version 3 is used for exploration of SBEs time series data. The user interface of this version has four major parts:

1. Overview
2. Variables View
3. Details List, and
4. Items List and Attribute Statistics.

Figure 2.18 shows the interface of TimeSearcer© (TimeSearcher, 2011).

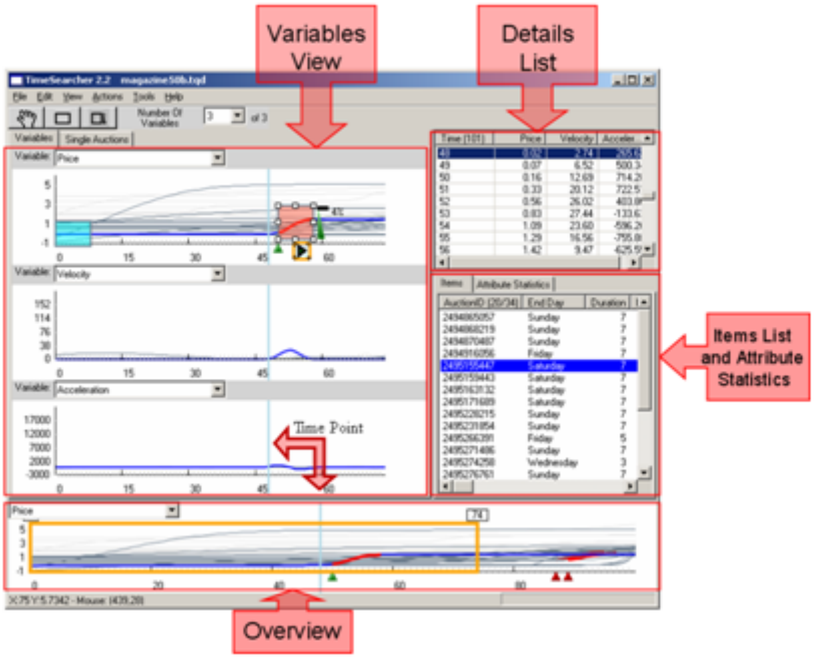


Figure 2.18: The TimeSearcher© User Interface, Main Components (TimeSearcher, 2011)

The rectangular query regions that can be drawn directly on a two-dimensional display of time series data in TimeSearcher© are known as "Timeboxes". The x-axis of Timebox represents the extent of the time period of interest and the y-axis specifies a constraint on the range of values of interest in that given time period. In general, a Timeboxe acts as a filter that accepts only those items that have values in the given range and time extent defined by the Timebox (Hochheiser and Shneiderman, 2004; Hochheiser, 2003; Keogh et al., 2002).

The application of Timebox is very simple and can be created by simply drawing rectangle in TimeSearcher©. As the box is drawn, it is constrained to occupy an integral number of time points. Items in a dataset that completely meet all of the constraints implied by the one or multiple active Timeboxes are highlighted for exploration. Figure 2.19 shows an example of application of Timeboxes in TimeSearcher© (Hochheiser and Shneiderman, 2004; Hochheiser, 2003; Keogh et al., 2002).

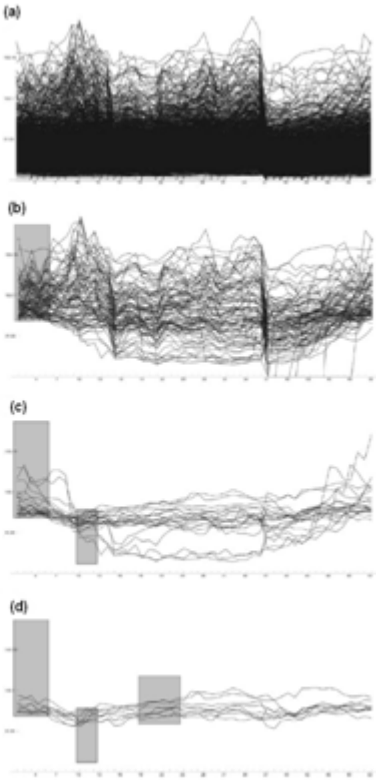


Figure 2.19: Timebox queries in TimeSearcher©: (a) Graph overview display for the entire data set. (b) Single Timebox query (c) Two Timeboxe query: refining the query in (b). (d) Three Timebox query: a further refinement of the query in (c) (Hochheiser, 2004).

The graph overview provides an ongoing display of the effects of the addition of Timeboxes and an overview of the result set. This enables the users to explore a large time series dataset interactively with minimal cognitive overhead. This enables users to quickly try a wide range of queries with ease and the modification of these queries allows users to evaluate the effects of changes in query parameters easily (Hochheiser and Shneiderman, 2004; Hochheiser, 2003; Keogh et al., 2002). This can be extremely helpful in identifying specific patterns of interest, as well as in gaining understanding of the SBES dataset as a whole.

2.11.3. Clustering of sonar data

Classifiers are often used in data mining research for identification of important classes of objects within a data repository. It is particularly useful when a dataset contains examples that can be used as the basis for future decision making. A range of different types of classification algorithms have been developed over the years and they mostly fall into the following methods: nearest neighbour methods (k-means, Fuzzy c-means), decision tree induction (Hierarchical classification), error back propagation (ANN, SOM), reinforcement learning (Markov Decision Process, MDP), and rule learning (GALE, extended classifier system (XCS)) (Bull et al., 2007; Lagoudakis & Parr, 2003; Salzberg, 1997).

When classifying sonar data, the most common approach is to use a combination of PCA and k-means (Brown et al., 2007; Mcgonigle et al., 2009; Preston et al., 2003; Sutherland et al., 2007). PCA is used for orthogonalisation of the feature dataset and dimensionality reduction (selection of three principal components). K-means is then used for clustering and the clusters are labelled using various ground truth data (Chivers et al., 1990; Orlowski, 1984; Preston et al., 2003). Artificial Neural Networks (ANN) have been used in several studies for the classification of MBES data (Blondel, 2000; Haralick, 1979; Lundblad et al., 2006; Marsh & Brown, 2009; Reed & Hussong, 1989). There have also been some classifiers that use ANN for seabed classification – for example, GENIUS developed by Danish Hydraulic Institute (DHI) and SeaClass^{TM} by Triton Elics International (Brown et al., 2007; DHI, 2011; Triton Elics International, 2004). Atallah et al. (2003, 2002) used wavelet analysis to classify and segment sonar signals scattered from underwater seabed. Lucieer (2005) used linear discriminant analysis (LDA) for generating classes from SBES data. This technique predicts classes based on how close a set of measurement variables are to the multivariate means of the levels being predicted (Hastie et al., 2009). Lucieer used 'Spatial Analyst' tool in ArcGIS to each of the formula generated from the LDA to generate the probability surfaces (Lucieer, 2005).

Fuzzy classifiers have also been used for the classification of seabed type. There have been a number of studies that focused on using fuzzy logic to achieve seabed classification. For example, Lucieer (2008) successfully tested object oriented hierarchical classification, a technique that applies fuzzy rule based membership function, for segmentation of data acquired from side scan sonars. Tamsett (1993), used fuzzy logic on power spectra features from side scan sonars and Narayanan et al. (2011) used fuzzy classifier to achieve a soft classification of mixed seabed using LIDAR bathymetric data. Lucieer and Lamarche (2011) tested unsupervised fuzzy classification on data acquired from a comprehensive 32 kHz MBES bathymetry and backscatter survey of the Cook Strait, New Zealand (survey area ~8500 km2) to map deep water substrates in the Cook Strait, New Zealand.

This thesis focuses mainly on two classifiers that are commonly used for seabed mapping: k-means and fuzzy c-means. The following sections briefly discuss these two classifiers.

K-means classifications

K-means is the most frequently used algorithm in sonar data clustering (Brown et al., 2007). This algorithm is iterative and assigns an arbitrary cluster vector as a first step. In the second step it classifies each pixel to the closest cluster. Next the new cluster mean vectors are calculated based on all the pixels in one cluster. The second and third steps are repeated until the difference between the iteration is small. The differences are commonly defined in two different ways: either by measuring the change in distance in the mean cluster vector between iterations or by the percentage of pixels that have changed between iterations (Hamerly & Elkan, 2002; Macqueen, 1967; Rekik et al., 2006).

k-means clustering

The term "k-means" was first used by James MacQueen (1967), though the idea goes back to Hugo Steinhaus (1957). The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published until 1982 (Lloyd, 1982).

For a given a set of observations (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, the principal objective of k-means clustering is to partition the n observations into k sets ($k \leq n$) $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2 \qquad 2.14$$

where, μ_i is the mean of points S_j. The most common k-means algorithm uses an iterative refinement technique. Given an initial set of k means $m_1^{(1)},\ldots,m_k^{(1)}$, the algorithm proceeds by alternating between two steps (David, 2003):

Assignment step: Assign each observation to the cluster with the closest mean (i.e. partition the observations according to the Voronoi diagram generated by the means).

$$S_i^t = \left\{x_p : \left|\left|x_p - m_i^t\right|\right| \leq \left|\left|x_p - m_j^t\right|\right| \forall 1 \leq j \leq k\right\} \qquad 2.15$$

where each x_p goes into exactly one S_i^t, even if it could go in two of them.

Update step: Calculate the new means to be the centroid of the observations in the cluster.

$$m_i^{t+1} = \frac{1}{S_i^t} \sum_{x_j \in S_i^t} x_j \qquad 2.16$$

The algorithm is deemed to have converged when the assignments no longer change. Commonly used initialization methods are Forgy and Random Partition (Hamerly and Elkan, 2002). The Forgy method randomly chooses k observations from the data set and uses these as the initial means while the Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial means to be the centroid of the cluster's randomly assigned points (Hamerly and Elkan, 2002).

From a statistical perspective, the clusters obtained by k-means can be interpreted as the Maximum Likelihood Estimates (MLE) for the cluster means if it is assumed that each cluster comes from a spherical Normal distribution with different means but identical variance (and zero covariance). This leads to some of the general disadvantages of the k-means algorithm (Bradley & Fayyad, 1998; Jain et al., 1999):

- K-means works best for datasets with clusters that are spherical and that have the same variance. This is often not true for sonar datasets.
- The result of k-means may be heavily influenced by the initial choice of random cluster centres. This effect can be alleviated by repeated clustering and picking the set of results with the minimum between-cluster errors.
- The number of clusters must be declared before the start of the algorithm. Typically this is not known a priori, leading to the common practice of "guessing" the best number. Inappropriate choice of 'k' may yield poor clustering results. Although various criteria are available to estimate the optimal number of clusters, including the elbow criterion (Aldenderfer & Blashfield, 1984), Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978), this number cannot always be found.
- Another problem is that while k-means attempts to minimise intra-cluster variance, the algorithm does not necessarily converge towards the global minimum variance - the solution can be trapped in one of the local minima due to the nature of the update algorithm. One final disadvantage of the k-means algorithm is that it does not recognise clusters that are non-convex, non-isotropic, ring-like or non-globular.

Fuzzy c-means classifications

The fuzzy c-means (FCM) clustering algorithm is an unsupervised classification algorithm which accommodates the vagueness in class definitions by allowing the class clusters to overlap. The clusters are assumed to be optimal when the multivariate within-cluster variance is minimal (Bezdek et al., 1984; Dunn, 1974).

Similarly to k-means, FCM algorithms treat each data record as an independent sample and apply an iterative procedure starting with an initial random allocation of the samples to be classified into class clusters. Based on the cluster allocation, the centre of each cluster (also known as the centroid) is calculated. The centroids are re-allocated in each iteration until the optimal locations of the cluster centres are found; at which point the algorithm has converged (V. Lucieer & Lucieer, 2009). The similarity of each sample to a cluster is expressed by a membership value ranging from 0 to 1, where 1 represents perfect similarity. The Euclidean distance measure is employed to quantify distance from samples to cluster means.

The FCM algorithm attempts to partition a finite collection of n elements X = $\{x_1,...,x_n\}$ into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centres $C = \{c_1,...,c_c\}$ and a partition matrix $U = u_{i,j} \in [0,1], i = 1, \dots, n \,\&\, j = 1, \dots, c$, where each element u_{ij} tells the degree to which element x_i belongs to cluster c_j . Like the k-means algorithm, the FCM aims to minimize an objective function (Bezdek, 1981). The standard function is:

$$u_k(x) = \frac{1}{\sum_j \left(\frac{d(\mathrm{centre}_{k,x})}{d(\mathrm{centre}_{j,x})}\right)^{2/(m-1)}} \qquad 2.17$$

which differs from the k-means objective function by the addition of the membership values u_{ij} and the fuzzifier m. The fuzzifier m determines the level of cluster fuzziness. A large m results in smaller memberships u_{ij} and hence, fuzzier clusters. In the limit m = 1%, the memberships u_{ij} converge to 0 or 1, which implies a crisp partitioning. The basic FCM Algorithm, given n data points ($x_1, \dots, x_n$) to be clustered, a number of c clusters with ($c_1, \dots, c_c$) the center of the clusters, and m the level of cluster fuzziness with,

In fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. An overview and comparison of different fuzzy clustering algorithms is available (Nock and Nielsen, 2006).

Any point x has a set of coefficients giving the degree of being in the kth cluster $w_k(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$c_k = \frac{\sum_x w_k(x)x}{\sum_x w_k(x)} \qquad 2.18$$

The degree of belonging, $w_k(x)$, is related inversely to the distance from x to the cluster center as calculated on the previous pass. It also depends on a parameterm that controls how much weight is given to the closest center (Bezdek, 1981).

The membership function (sum of the membership values) is the basic idea in fuzzy set theory; its values measure degrees to which objects satisfy imprecisely defined properties. Fuzzy memberships represent similarities of objects to imprecisely defined properties. It is calculated by determining the distance of the point to the cluster centres (also known as degree of membership value) by taking into consideration the degree of fuzziness. The fuzzy exponent (also known as fuzziness) represents the amount of overlap between classes. A value close to 1 result in clusters with a distinct boundary and a very high value of fuzziness would indicate a fully overlapping set of clusters (Lucieer & Lucieer, 2009; Ozkan & Turksen, 2007).

Clustering of natural classes has always been challenging as there is usually no distinct boundary between classes as they tend to overlap. FCM has been successfully applied in the geographical context to overcome this class overlap issue (Burrough et al., 2000). FCM is an exploratory technique that initially does not allow inference about the geographical proximity of the overlapping classes, though it is likely that the resulting membership values will be spatially correlated if such correlation exists in the source data (Burrough et al., 2000). Therefore it seems to be a promising alternative technique for seabed type classification from MBES data, since borders between geological seabed type classes are rarely crisp.

In the next subsection, we discuss the technique to analyse clusters for their within-cluster compactness and between-cluster separation without using ground truth data.

2.11.4. Internal cluster validation

There is extensive literature available on data clustering. Despite the intensive research, a critical issue in data clustering is the estimation of the number of clusters contained in the data. Most of the traditional clustering algorithms such as k-means, fuzzy c-means, EM (Expectation Maximisation), and hierarchical clustering require cluster number to be defined a priori by the users (Dempster et al., 1977; Everitt et al., 2011; Hastie et al., 2009; Jain & Dubes, 1988; Macqueen, 1967). As the actual number of clusters is generally unknown and estimation of clusters from visual exploration is prohibitive when the dataset is multidimensional and large in size, this approach of defining cluster numbers a priori is quite restrictive in practice. Common solutions to this problem are to partition with different number of clusters and then select the best result according to a specific quality criterion (Everitt et al., 2011; Jain et al., 1999). There are a number of clustering validity measures available as quantitative criteria for evaluating the quality of data partitions. They can be divided into two broad categories – External and Internal validity indices.

External validity indices are the measures of the agreement between two partitions, one of which is usually a known/golden partition, e.g. true class labels, and another is from the clustering procedure. Internal validity indices evaluate clustering results by using only features and information inherent in a dataset. They are usually used in the case that true solutions are unknown. For this project, some commonly used internal validation methods were tested (Everitt et al., 2011; Jain et al., 1999).

There are a number of functions available for internal validation of clusters. Some of these are: Silhouette index, Davies-Bouldin index, Calinski-Harabasz index, Dunn’s index, R-squared index, Hubert-Levin index (C-index), Krzanowski-Lai index, Hartigan index, Root-mean-square standard deviation (RMSSTD) index, Semi-partial R-squared (SPR) index, Distance between two clusters (CD) index, weighted inter-intra index, Homogeneity index, and Separation index (Everitt et al., 2011; Jain et al., 1999).

For this research, four indexes were tried as they are more commonly used than others. They are: Calinski-Harabasz or Variance Ratio Criterion (VRC) index (Calinski & Harabasz, 1974; Everitt et al., 2011), Davies-Bouldin index (Jain & Dubes, 1988), Dunn's index (Dunn, 1974; Halkidi et al., 2001) and Silhouette index (Everitt et al., 2011; Halkidi et al., 2001). The validation index approaches are briefly discussed below:

Calinski-Harabasz Index (VRC)

For a given data set $X = \{x(1) \dots \dots, x(N)\}$ of N data objects and a partition of these data into k mutually disjoint clusters, the Variance Ratio Criterion (VRC) is given as (Calinski and Harabasz, 1974):

$$VRC = \frac{trace\ (B)}{trace\ (W)} \times \frac{N-k}{k-1} \tag{2.19}$$

Where W and B are the within-group and between-group dispersion matrices. The trace of matrix W is the sum of the within-cluster variance (its diagonal elements). The trace of the matrix B is the sum of the between-cluster variances. Compact and separated clusters are expected to have small values of trace (W) and large values of trace (B). Therefore, a good data partition will yield a greater value of the ratio between traces of B and W. The normalisation term (N-k)/(k-1) prevents this ratio to increase monotonically with the number of clusters, making VRC an optimisation criterion with respect to k.

Davies-Bouldin Index

This index, like VRC, is also based on a ratio that involves within-group and between-group distances. The index is calculated as below (Jain & Dubes, 1988):

$$DB = \frac{1}{k} \sum_{i=1}^{k} D_i \tag{2.20}$$

Where Di = max {D$_{ij}$}, $i{\neq}j$. D$_{ij}$ is the within-to-between cluster spread for the i^{th} and j^{th} clusters and is expressed as$\frac{\overline{d_\iota}+\overline{d_J}}{d_{i,j}}$. $\overline{d_\iota}$ and $\overline{d_J}$ are the average within group distance for the i^{th} and j^{th} clusters. d$_{i,j}$ is the inter-group distance between clusters i and j. The term D$_i$ represents the worst-case within-to-between cluster spread involving i^{th} cluster and therefore, compact well-separated clusters are always distinguished by small values of DB.

Dunn's Index

Dunn's index is also based on geometric measures of cluster compactness and separation. This index can be defined as (Dunn, 1974):

$$DN = \begin{matrix} min \\ p,q \ \in \{1,\dots,k\}, p \neq q \end{matrix} \left\{ \frac{\delta_{p,q}}{\begin{matrix} max \\ l \in \{1,\dots,k\} \end{matrix} \Delta_l} \right\} \tag{2.21}$$

Where Δ_l is the diameter of the l^{th} cluster and $\delta_{p,q}$ is the set distance between clusters p and q. The set distance is defined as the minimum distance between a pair of objects across clusters p and q. The diameter of a given cluster (cluster l in this case) is defined as the maximum distance between a pair of objects within that cluster. Therefore compact and well-separated clusters are always represented with high values of DN.

Silhouette Index

Silhouette index is another well known cluster validation index and is based on, like other validation methods mentioned above, geometric considerations about compactness and separation of clusters. Before defining the criterion, let us consider that the j^{th} object of the dataset, x(j), belongs to a given cluster p∈{1,…,k}. Then, let a$_{p,j}$ be the average distance of this object to all other objects in cluster p. Finally, let b$_{p,j}$ be the minimum average of the distances between this object to all other object in another cluster q (p ≠ q and q=1,...,k). This minimum average distance represents average dissimilarity of object x(j) to its closest neighbouring cluster. Then, the Silhouette index (also known as the Silhouette Width Criterion) can be expressed as (Everitt et al., 2011; Kaufman & Rousseeuw, 2005):

$$S_{x(j)} = \frac{b_{p,j} - a_{p,j}}{\max\left(a_{p,j}, b_{p,j}\right)} \tag{2.22}$$

The denominator in the above equation is a normalising factor and a higher value of the Silhouette index represents compact and well-separated clusters.

Chapter 3

Redundancy detection and clustering of multibeam backscatter data

This chapter discusses the application of the Self Organising Map (SOM) for the detection of data redundancy in MBES data. It also compares the clustering results obtained from using both the standard MBES data as well as MBES data optimised using SOM.

Chapter contents

The use of multibeam echosounders (MBES) in seabed mapping is fairly recent (late 1970s) and at this moment is the instrument of choice for most seabed mapping projects (Mayer, 2006).

3.1. Research background and justification

The goal of this project is to examine the complexity of the classification of MBES backscatter data (Preston, 2009; Preston et al., 2001), which is commonly used for production of seabed maps. This method generates a very complex and highly dimensional new dataset of statistical features (statistical descriptors derived from segments of MBES image) from the original MBES image which is then used for clustering and classification. One of the drawbacks of the method is that many attributes in this complex dataset are very similar to each other, resulting in large redundancy on information. The main focus in this chapter will be on this particular complexity issue – redundancy of a method-generated feature dataset. The attribute similarity in this dataset is explored using a method from Visual Analytics – a Self Organising Map (SOM). This experiment introduces the idea of using a visual analytical approach and tests the feasibility of the idea of optimal feature vector determination on a set of MBES dataset to minimise the redundancy. The experiment is then further expanded by producing an alternative clustering of the backscatter data using a subset of non-redundant features and then compares that to the map generated using the traditional method. The ultimate goal of the project is to facilitate a more efficient seabed classification from MBES backscatter data by avoiding unnecessary redundancy as much as possible.

The chapter is structured as follows: the following sections briefly describe the data acquisition and processing. This is followed by discussion of the experiment that was performed. Lastly results are presented followed by discussion of the results.

3.2. Producing the feature dataset from MBES backscatter data

Various techniques for image processing can be applied to MBES backscatter images for image segmentation. In addition to backscatter measurements, sonar geometry and the geometry of the entire multibeam system have to be taken into account in order to achieve a valid classification that depends on sediment characteristics rather than sonar artefacts (Preston et al., 2001). Another approach to segmentation relies on the calibrated backscatter levels and on their variation with grazing angle (Hughes Clarke et al., 1997). Other recent approaches to automatic classification are based on the statistical nature of the image, irrespective of absolute calibration. One of the main commercial developments (specifically for MBES) based on this approach is Quester Tangent Corporation’s QTC Multiview^{TM} (Mcgonigle et al.,2009; Preston et al., 2004; Preston et al., 2001), which uses the backscatter amplitude and statistical properties of multibeam sonar images to classify seabed sediments. This is the method that is researched in this project. Figure 3.1 shows the general workflow from data acquisition to feature extraction using QTC Multiview^{TM}.

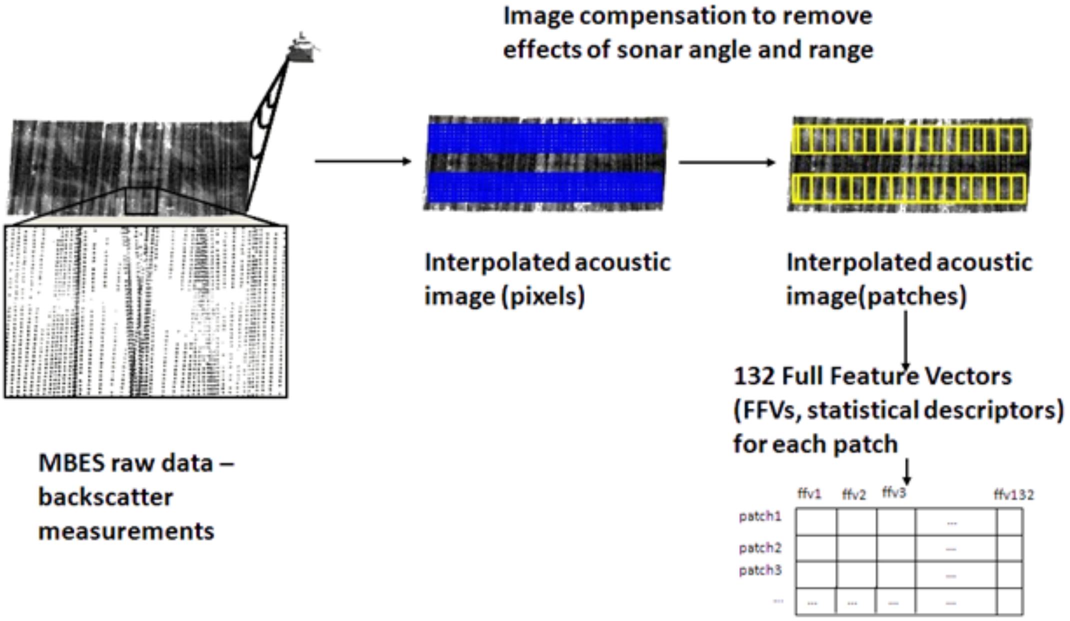


Image compensation to remove effects of sonar angle and range
Interpolated acoustic image (pixels)
Interpolated acoustic image(patches)
132 Full Feature Vectors (FFVs, statistical descriptors) for each patch
MBES raw data – backscatter measurements
ffv1
ffv2
ffv3
ffv132
patch1
patch2
patch3
...

Figure 3.1: A typical MBES data acquisition and feature extraction process (Preston, 2009)

Backscatter data for seabed classification can have flawed values (due to unreasonable depth picks, reflections from fish or artefacts, etc.). In addition, images can be smeared due to excessive vessel movement. In QTC Multiview^{TM}'s approach, the backscatter data are first filtered to clean the data of these anomalies. The image of the seabed is then divided into a number of rectangular patches. The placement of these patches depends on the quality of the data. It is also influenced by the grazing angle and range of the sonar as well as constraints that come with different sensor models. The influences of these are removed through the process of image compensation during the next step where the image is separated into rectangular patches, which are superimposed on the image on each side (port & starboard) of the vessel (Preston, 2009; Quester Tangent, 2007).

3.2.1. MBES raw data processing

In the next step, a total of 132 features are calculated for each patch from the backscatter image by applying a number of statistical algorithms on each of these rectangular patches. The features are called Full Feature Vectors (FFVs) and the result is a large matrix in which each column represents the values of one feature and each row contains all the features extracted from one rectangular patch. This matrix represents the FFV space, which is 132-dimensional, i.e. each feature can be regarded as one new dimension/attribute. Table 3.1 presents a list of the features (Preston, 2009; Preston et al., 2001) in the same order as they appear in the newly formed FFV space. They are grouped into several sets.

The first set of features includes mean, standard deviation and two high-order statistical moments (skewness & kurtosis). These features indicate interface roughness and changes in acoustic impedance. It should be noted that even though kurtosis is calculated at this step, it is assigned weighting of zero later on in the procedure and is therefore disabled.

The next set of features is extracted from histograms and quantiles. Histogram features indicate heterogeneity. Quantile features, related to the histogram features, express the distribution of backscatter intensities.

The next set of features consists of Pace power spectrum ratios, which are calculated from the ratios of log-normalized power in various frequency bands. These represent the power spectrum of backscattering strength calculated through a Fast Fourier transformation (FFT) and a median filter (Pace & Gao, 1988)

An important approach to any image analysis is quantifying the texture content of an image. In Preston (2009) classification procedure, this is done by using Gray Level Co-occurrence Matrices (GLCMs). In this procedure, a total of 63 textural features are calculated related to GLCMs in the following way: suppose a set of sound signals from MBES, called a ping, is emitted, at distinct angles, towards the sea floor and their reflectivity values recorded (Renard et al., 2005). If we consider a sequence of that reflectivity data (say their reflectivity values are m and n), the co-occurrence of reflectivity m and n is the number of pairs of samples that are in a fixed spatial relationship and have the reflectivity m and n. A GLCM for a particular patch is a square matrix where each (*n, m*) element represents the number of times that the backscatter amplitude changes from n to m for a particular direction and step. In QTC Multiview^{TM} a step size of 1, 2, and 3 pixels are used in along-track direction, across-track direction, and in the direction that is at the 45^0 angles between the two directions. This produces a total of 9 GLCMs for each patch. Seven textural features are then calculated for each GLCM – these include correlation, shade, prominence, contrast, energy, entropy, and homogeneity (Preston, 2009; Preston et al., 2004).

Table 3.1: 132 features (Full Feature Vectors, FFVs) calculated from image patches (after Preston et al., 2001; Preston, 2008).

Statistical Descriptors	**Number of features**
Mean, standard deviation, skewness, kurtosis	4
Quantiles	9
Pace features from power spectral ratios	15
GLCM correlation	9
GLCM shade	9
GLCM prominence	9
GLCM contrast	9
GLCM energy	9
GLCM entropy	9
GLCM homogeneity	9
Power spectrum	32
Histogram	8
Fractal dimension	1
	Total = 132

Finally, one of the last features to be calculated is the fractal dimension. This feature provides a measure of the structure and distribution of backscatter as well as depth variations of the seabed image (Collins & Preston, 2002; Quester Tangent, 2007; Xinghua & Yongqi, 2004). This feature however, together with kurtosis, is disabled in further modelling, even though it is calculated at this stage.

3.3. Challenges associated with MBES classification from FFV data space

In the final step of the seabed classification, image patches are clustered according to their similarity in the FFV space (Figure 3.2). The idea behind this is that patches with similar values of FFVs exhibit similar acoustic characteristics and therefore identify response from a particular type of seabed. The approach used for this step in the QTC Multiview^{TM} classification is to first reduce the dimensionality of the FFV space from 132 to three dimensions by taking the first three components from the Principal Components Analysis (PCA) performed on the FFV space. Then the so-obtained three-dimensional space is clustered using k-means clustering into unsupervised classes. These classes are finally compared to a catalogue of acoustic classes to link the computationally derived classes with particular seabed types (Preston, 2009).

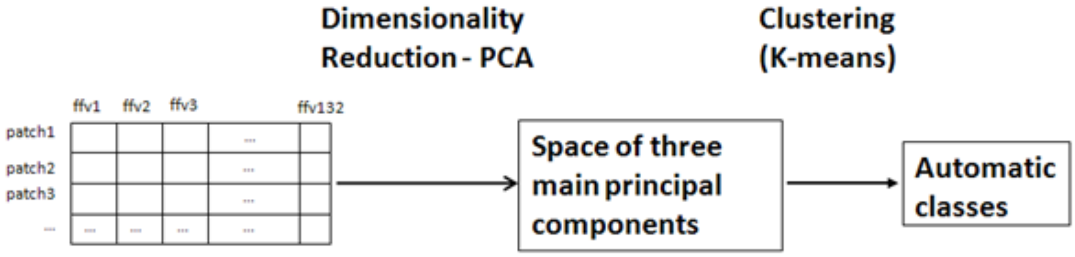


Figure 3.2: QTC Multiview^{TM} classification scheme for shallow water MBES data

There are several complexity issues with this method. According to Preston (2009), while the high number of FFVs is calculated because of the wide diversity of sonar images, many of the FFVs are highly correlated and therefore most likely redundant, i.e. they do not contribute any new information about acoustic similarity to the process. Because the resulting FFV space is so highly dimensional (132 dimensions), it is “convenient” (Preston, 2009) to use a dimensionality reduction technique to facilitate interpretation of similarity patterns. However, the dimensionality reduction method (i.e. PCA) and the number of components used are chosen arbitrarily (Preston, 2009). This brings up the main research question in this project:

> *Is it necessary to produce such a highly dimensional FFV space or could the dimensionality be kept lower by avoiding redundancy? And if so, which of the FFVs are correlated with each other and therefore redundant? It is also important to consider if redundancy is related to a particular wavelength. Since the wavelength differs depending on the survey area depth, a high frequency beam can have a higher degree of scattering while a lower frequency beam should have higher penetration and less scattering. Therefore it would be interesting to see if within feature correlation remains similar with varying wavelength.*

This project aims to investigate the similarity of FFVs in order to uncover potential redundancy in the FFV space using a Visual Analytical method that simultaneously produces dimensionality reduction and clustering – Self Organising Map (SOM).

3.4. Alternative approach: Self Organising Map (SOM)

For the experiment, Visual Analytical approach is adapted for exploring the similarity of statistical variables – FFVs – used in the classification of the multibeam echosounder images. The method chosen for this exploration was a Self Organising Map (SOM) together with its component planes visualisation, which is introduced in the rest of this section.

3.4.1. Visual analytics approach for MBES: SOM

The goal of the experiment is to determine through visual exploration which of the FFVs are sufficiently similar to each other to not bring any additional information to the classification and could therefore be potentially removed from the procedure. The Self Organising Map (SOM) is chosen for this purpose because its component planes visualisation allows for easy visual identification of attribute similarity, as described in Chapter 2. SOM will be used on the 132-dimensional FFV space (Figure 3.3). Once the SOM is trained, the result is displayed in the 132 component planes (one for each FFV), where each plane is coloured according to the FFV that it represents. These planes are then visually examined for similar colour distribution patterns that define visual groups of FFVs. These patterns indicate that FFVs in each of these visual groups are probably correlated.

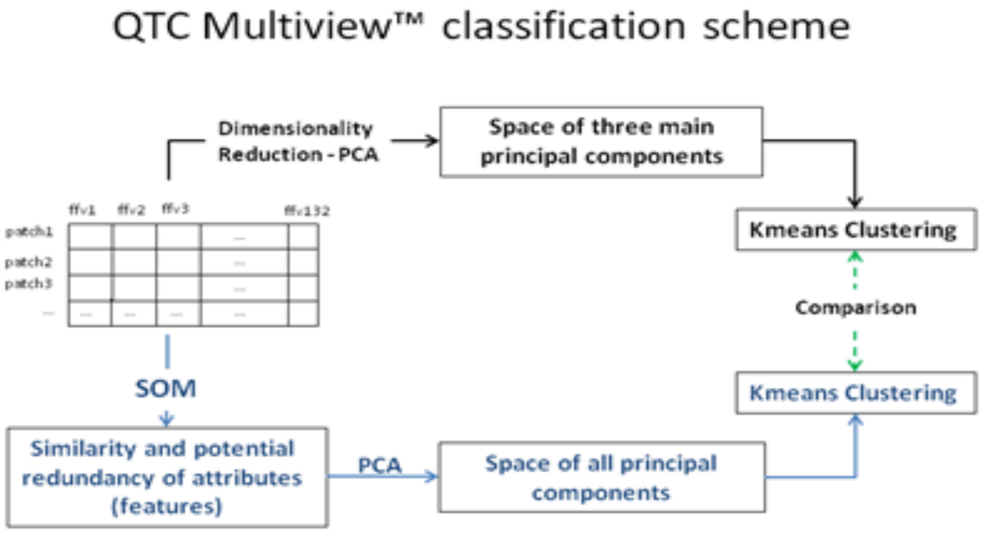

QTC Multiview™ classification scheme
Dimensionality Reduction - PCA
Space of three main principal components
ffv1
ffv2
ffv3
ffv132
patch1
patch2
patch3
Kmeans Clustering
Comparison
Kmeans Clustering
SOM
Similarity and potential redundancy of attributes (features)
PCA
Space of all principal components

Figure 3.3: SOM based exploration and subsequent clustering of MBES

Both the 132 FFVs and the reduced dimension dataset would then be clustered using a combination of PCA-k-means. The cluster results are then compared with each other and with ground truth data to establish whether reduced dimension dataset are able to produce representative clusters (Figure 3.3).

3.5. MBES data

The Geological Survey of Ireland (GSI), together with the Irish Marine Institute (MI), has been conducting sonar surveys in the Malin Sea, NW of Ireland since 2003, as part of the Irish National Survey (INSS) and INFOMAR programs (Cullen, 2003). The survey lines used in this study were acquired in the north part of the Malin Shelf, over 100 km long, between $55^0$56′N to $55^0$54.N and from $90^0$10′W to $70^0$24′W (Figure 3.4).

Bathymetric and backscatter MBES data were collected between 120-180m water depth using a Simrad EM1002 multibeam echosounder, with an operational frequency of 93-98 kHz and pulse lengths of 0.2 ms and 0.7 ms. Acoustic footprints (the width of the Earth the sensor measures- i.e., the sensor's field of view) range from 3 to 10m in diameter. The acquisition software, used in conjunction with EM1002 MBES hardware, allows the collection of time series of echo amplitudes for each beam (Simrad-Kongsberg, 1999) and applies, in real-time, a series of first order predicted corrections for propagation losses, spherical divergence and changes in the insonified area (Simrad-Kongsberg, 1999). Digital Images from backscatter data were assembled using the echo-amplitudes received. Backscatter values were filtered, compensated for angle and range and finally resample to a regular grid (Preston et al., 2001).

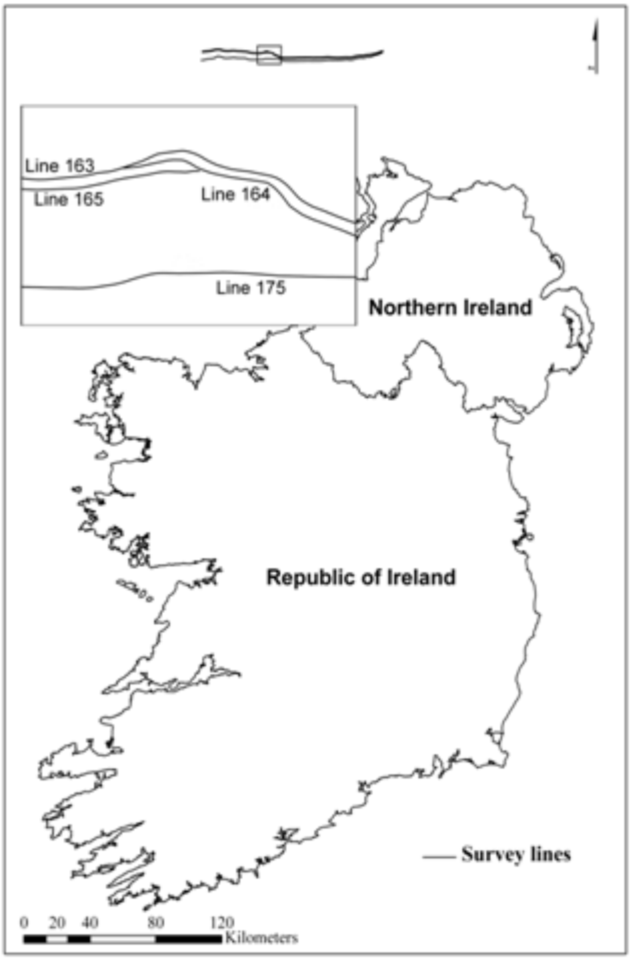


Figure 3.4: Location of the MBES survey lines used in the experiment

Raw MBES data from the survey lines (CE03_02_163, CE03_02_164, CE03_02_165, and CE03_02_175) have been processed into the FFV dataset by GSI using QTC Multiview™. This dataset contains all the FFVs for the rectangles of the survey lines 163, 164, 165, and 175 for two pulse lengths (0.2ms and 0.7ms) as well as navigation information. The 0.2ms pulse length represents data from the shallow part of the survey area whereas the 0.7ms pulse length represents data from the deeper part of the survey lines. The pulse length change is automatic and is determined by the pre-determined intensity of emitted sound echo. The data of all four survey lines consist of a total of 2,10,216 records (rectangular patches) for the 0.2ms pulse length and 1,49,552 records for the 0.7ms pulse length. The header of each data file, along with the information of the echosounder and source, provides the total number of records, number of samples in each rectangular patches and ping information. In addition to this, each record in the file comes with its own acquisition information: the date and time of the data capture, beam number, depth and geographic location of the record (central point of the patch). This information is followed by the values of 132 features (FFVs) calculated from the backscatter for each patch, which are of interest in this study and which represent the 132 dimensions of the FFV dataset.

QTC Multiview™ generated FFV files had to be converted to the standard table format before they can be used in any statistical software package (Matlab, R etc). A Matlab function was written to read the headers and features of each sample from the FFV file and create a comma separated value (CSV) file where each row stored the values of one sample. SOM toolbox for Matlab, developed in the Laboratory of Computer and Information Science at the Helsinki University of Technology (LCIS, 2011) was the function package of choice for the analysis of data using Self Organising Maps in this experiment. To be used in the SOM toolbox, these data needed to be in spreadsheet or table data format. A second Matlab function was written to convert the CSV files in to table data format and store only the feature values. The data for SOM toolbox can have any number of samples but samples must have a fixed length.

Each of the converted data files were then grouped and combined based on their pulse lengths resulting in two large dataset: one for 0.2ms pulse length and one for 0.7ms pulse length. To summarise, the dataset for visual exploration consisted two files- one for 0.2ms pulse length with 2,10,216 data records (patches) and the other for 0.7ms pulse length with 1,49,552 data records, each of which has values in 132 dimensions (FFVs).

3.6. Experiment

After the dataset was prepared as per the specification of the toolbox, it was loaded into the Matlab environment using the SOM toolbox. First the data had to be normalised as each variable in the dataset had different value ranges – without the normalisation, there was a possibility that any attributes with a larger range would visually dominate over other attributes in the component planes visualisation. The

SOM toolbox is equipped with various normalising functions (for example, histogram equalization, logarithmic scaling, variance, range, etc.). For this study the ‘variance’ function was applied to the data, in which the mean is removed from each of the columns in the data matrix and then each column is scaled by the standard deviation.

Normalised data were used as the input for the SOM, which was trained based solely on the data. The SOM algorithm in SOM toolbox automatically determines the map size (i.e. the size of the SOM lattice, which is given by the number and 2D distribution of neurons) and the training parameters based on the data and then create, initialise and train the SOM. The map size was not manually set as the automated determination of map size were fairly consistent across the dataset. The training process is completed in two phases. It starts with a large initial neighbourhood radius and large learning rate. The radius and learning rate become finer through repetitive tuning, until a stable state is reached. For this study we used the default training algorithm (linear initialization and batch training).

After the training was completed and the data distributed in the SOM cells, the results were visualised using the component planes visualisation (Figure 3.5 & Figure 3.6).

3.7. Results

The main objective of this research is to look for redundancy and relationships among the 132 features in an effort to reduce dimensionality and gain a better control for subsequent classification. This was achieved by analysing the component planes of the attribute distribution of the dataset. Each of the 132 SOM component planes shows a hexagonal SOM lattice coloured according to the values of the respective FFVs. The FFV values are in the same order as in Table 3.1.

3.7.1. Attribute similarity using SOM

Inspection of the component planes provided an idea of the distribution of the values in each component. By analysing the component planes (Figure 3.5 & Figure 3.6) it was possible to identify visual clusters consisting of features with similar visual patterns. A total of 22 visual groups were identified for the data with 0.2ms pulse length and 19 visual groups for the 0.7ms pulse length data (Table 3.2 & Table 3.3) by grouping together features that showed similar colour distribution pattern in their respective component planes.

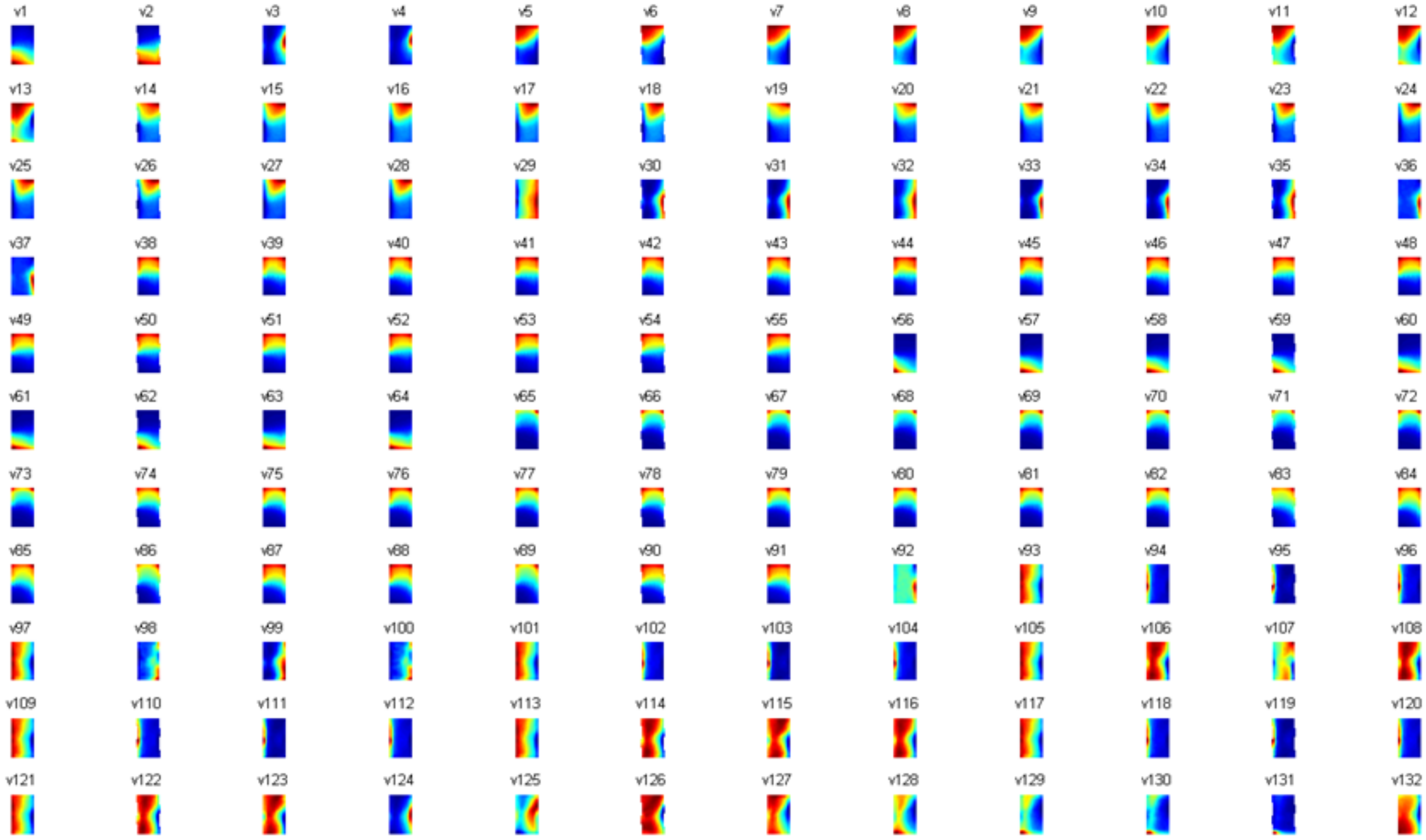


Figure 3.5: The 132 SOM component planes for 0.2ms pulse length. Panel a) shows component planes for FFVs 1-36, panel b) for FFVs 37-72, panel c) for FFVs 73-108 and panel d) FFVs 109-132. Similar colour distribution patterns in different planes indicate attribute

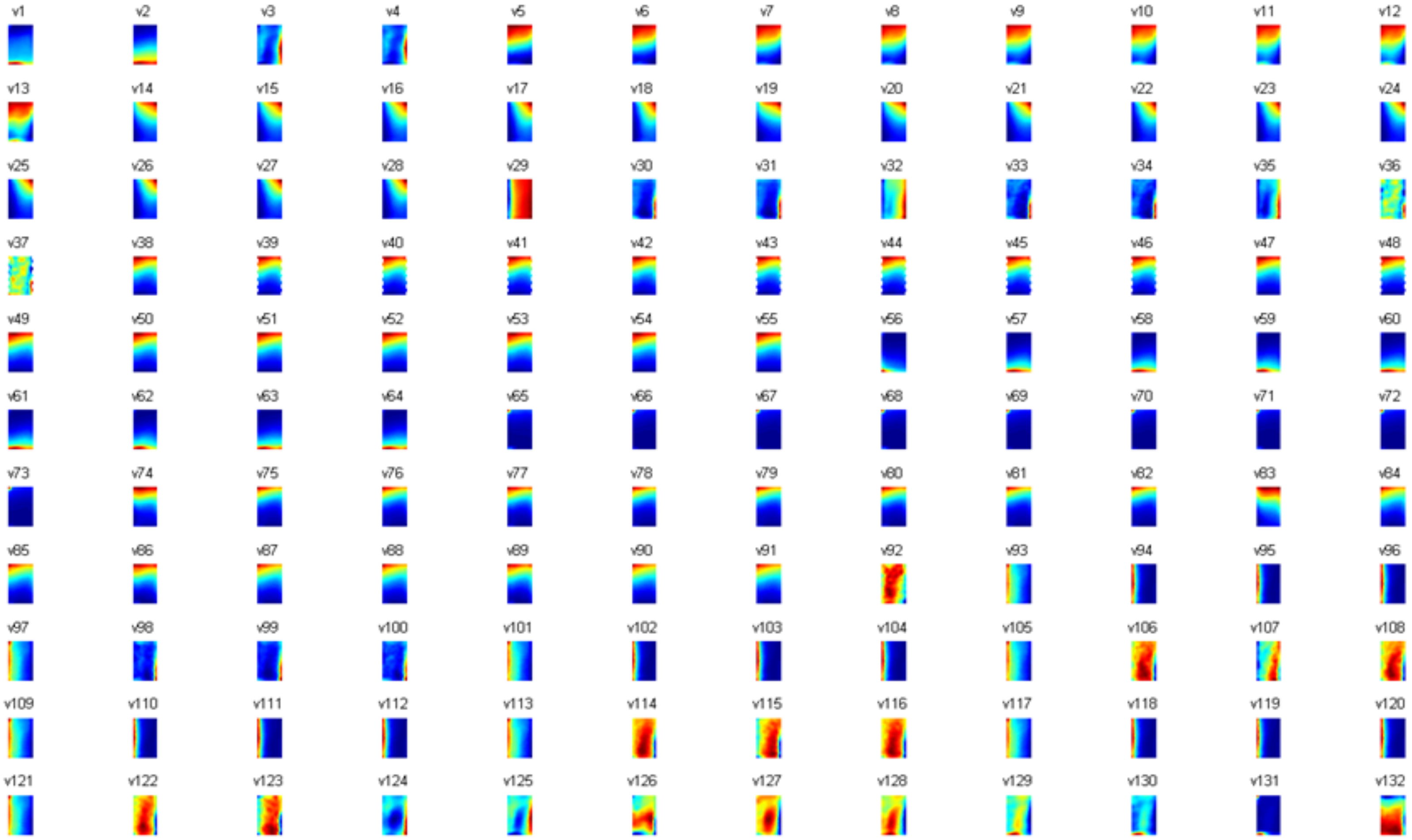


Figure 3.6: The 132 SOM component planes for 0.7ms pulse length. Panel a) shows component planes for FFVs 1-36, panel b) for FFVs 37-72, panel c) for FFVs 73-108 and panel d) FFVs 109-132. Similar colour distribution patterns in different planes indicate attribute

The example in Figure 3.7 shows the component planes for the FFVs 3 to 18. A total of three different visual groups can be identified in this selection of component planes. The first group consists of FFVs 3, and 4. These FFVs represent statistical descriptors of type ‘Skewness’ and ‘Kurtosis’. The second group includes component planes corresponding to FFVs 5 to13 and are of type ‘Quantiles’. The last group consist of component planes representing FFVs 14 to 18, which are of type ‘Pace features from power spectral ratios’.

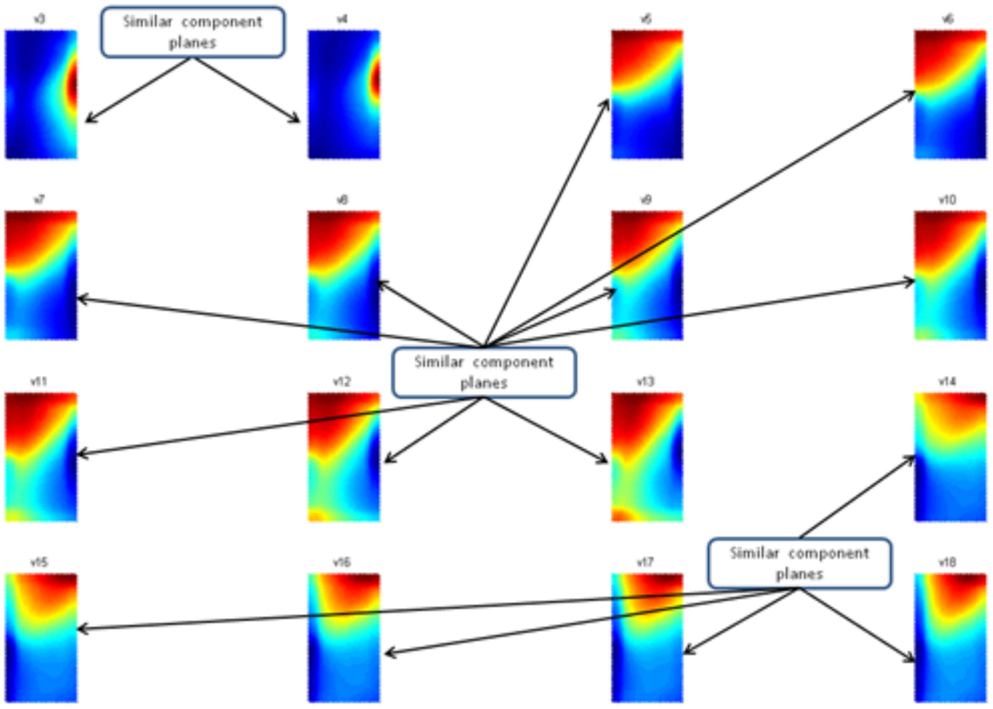


Figure 3.7: Illustration of how visual groups were identified on a smaller selection of component planes

As for these datasets (0.2ms & 0.7ms pulse lengths), a total of 16 visual groups are identified for both the pulse lengths in all 132 component planes – these are summarised in table 3.2 & 3.3. The group 16 is made of individual singletons i.e. FFVs that were visually distinct from all others. A total of twenty FFVs are placed here as they showed low visual similarity and correlation with any other FFVs. Details of visual groupings are included in Appendix 3, which also contains correlation table that are explained below.

Table 3.2: Visual groups identified in the SOM component planes (0.2ms pulse length)

Visual group No.	FFVs in the visual group	Description of the FFVs in the group	Level of visual similarity and correlation between component planes in the group
1	5-13	Quantiles	High
2	14-28	Pace features from power spectral ratios	High
3	30-31	GLCM correlation	High
4	33-34	GLCM correlation	High
5	32, 35	GLCM correlation	High
6	36-37	GLCM correlation	High
7	38-55	GLCM shade, GLCM prominence	High
8	56-64	GLCM contrast	High
9	65-73	GLCM energy	High
10	74-91	GLCM entropy, GLCM homogeneity	High
11	94:96, 102:104, 110:112, 118:120	Power spectrum	High
12	93, 97, 101, 105, 109, 113, 117, 121	Power spectrum	High
13	122:123	Power spectrum	High
14	3:4	Skewness, kurtosis	High
15	98:99	Power spectrum	High
16	1, 2, 29, 92, 100, 106, 107, 108, 114-116, 124-132	- - - - - -	Singletons

Table 3.3: Visual groups identified in the SOM component planes (0.7ms pulse length)

Visual group No.	FFVs in the visual group	Description of the FFVs in the group	Level of visual similarity and correlation between component planes in the group
1	5-13	Quantiles	High
2	14-28	Pace features from power spectral ratios	High
3	30-31	GLCM correlation	High
4	33-34	GLCM correlation	High
5	32, 35	GLCM correlation	High
6	36-37	GLCM correlation	High
7	38-55	GLCM shade, GLCM prominence	High
8	56-64	GLCM contrast	High
9	65-73	GLCM energy	High
10	74-91	GLCM entropy, GLCM homogeneity	High
11	94:96, 102:104, 110:112, 118:120	Power spectrum	High
12	93, 97, 101, 105, 109, 113, 117, 121	Power spectrum	Moderate
13	122:123	Power spectrum	High
14	3:4	Skewness, kurtosis	High
15	98:99	Power spectrum	High
16	1, 2, 29, 92, 100, 106, 107, 108, 114-116, 124-132	------	Singletons

Identified visual groups show that almost in all (except for group 16-singletons) cases statistical algorithms used to generate FFVs in a particular group generate a similar distribution of values for each patch. For example, the algorithm to generate Quantiles generates nine separate FFVs (numbers 5-13), which are all in the same visual group (no. 2) in pulse length 0.2ms and are therefore likely to be highly correlated. As another example, ‘Pace features from power spectral ratios’, shows similar distribution pattern in all component planes. The rest of the component planes were grouped in a similar fashion.

To confirm the visually derived similarity grouping, pair-wise correlation coefficient was derived between members of each group-except for group 16 as it is comprised of component planes with low visual similarity to other component planes and therefore is expected to have low correlation. Table 3.4 and Table 3.5 summarises the average, median, mode and the standard deviation of the pair-wise correlation coefficients from both set of visual groups. Full pair-wise tables for each visual group are given in Appendix 3.

Table 3.4: Summary of correlation coefficient within each visual group in 0.2ms pulse length

Visual Group	No. Of Component planes	Mean Corr. Coeff.	Std. Dev.	Variance
VG 1	9	0.9427	0.0586	0.0034
VG 2	15	0.9609	0.0342	0.0012
VG 3	2	0.9891	0	0
VG 4	2	0.9451	0	0
VG 5	2	0.9775	0	0
VG 6	2	0.8714	0	0
VG 7	18	0.9865	0.0078	6.1495e-005
VG 8	9	0.9791	0.0140	0.0002
VG 9	9	0.9792	0.0155	0.0002
VG 10	18	0.9665	0.0203	0.0004
VG 11	12	0.8903	0.0182	0.0003
VG 12	8	0.5167	0.0072	5.1207e-005
VG 13	2	0.7474	0	0
VG 14	2	0.9089	0	0
VG 15	2	0.6011	0	0
VG16(Singletons)	20	0.0555	0.3329	0.11109

Table 3.5: Summary of correlation coefficient within each visual group in 0.7ms pulse length

Visual Group	No. Of Component planes	Mean Corr. Coeff.	Std. Dev.	Variance
VG 1	9	0.9498	0.0525	0.0027
VG 2	15	0.9679	0.0340	0.0011
VG 3	2	0.9895	0	0
VG 4	2	0.9460	0	0
VG 5	2	0.9717	0	0
VG 6	2	0.8739	0	0
VG 7	18	0.9759	0.0165	0.0003
VG 8	9	0.9519	0.0461	0.0021
VG 9	9	0.9745	0.0209	0.0004
VG 10	18	0.9606	0.0235	0.0005
VG 11	12	0.9408	0.0112	0.0001
VG 12	8	0.6161	0.0034	1.1861e-005
VG 13	2	0.9076	0	0
VG 14	2	0.8491	0	0
VG 15	2	0.7432	0	0
VG16 (Singletons)	20	0.0095	0.2553	0.0652

The tables above show the average of all features in one visual group, their standard deviation and variance. Both the tables show that the average values have very low standard deviation and variance. A low standard deviation indicates that variation or "dispersion" from the average is low i.e. the data points are very close to the mean. The variance is a measure of how far a set of numbers is spread out from each other. It is one of several descriptors of a probability distribution, describing how far the random variable stray from its mean and a low value indicating that variables are not far from their mean. Therefore, it can be assumed that the component planes of similar FFVs (included in Appendix 3) can demonstrate the similarity of attributes and that one single descriptor of this type, for example an average of all or one single selected feature, could provide the same amount of information to the process as all the features together in that group.

3.7.2. Clustering

A separate dataset was then prepared based on the visual grouping. The normalised FFVs were first grouped into 16 groups based on the component plane exploration and correlation coefficient results. For groups 1-15, the mean was taken as a representative value for that group. The group 16, comprising of singletons, were added as is. This resulted in a much smaller dataset with 35 variables for each pulse length compared to the original dataset of 132 variables.

PCA + k-means was the classification algorithm of choice as the objective is to observe if lower dimension of variables produce similar results with all other steps remaining unchanged.

The Principal Component analysis (PCA) was done on both 132 variables and reduced dimension variables extracted from visual grouping. In order to not lose any information, all components were selected for k-means. In this sense, PCA was used as an orthogonalisation method to produce an input data for the k-means clustering where all dimensions are orthogonal and independent of each other. It was not used as a dimensionality reduction technique. During the k-means clustering, the number of clusters specified a priori varied from three to eight. In order to minimise the impact of having a 'poor' local minima by chance, k-means was replicated 10 times for each set of clusters with the algorithm running with a new set of initial cluster centroid positions with each replication. The cluster labels with the lowest value for within-cluster sums of point-to-centroid distances were retained as the optimum results.

Maps of the clusters obtained for each pulse length and cluster combination are presented in Figures 3.8 to 3.13. In order to facilitate visual comparison, clusters are sorted based on cardinality. That is the cluster with the highest number of elements will always be labelled as 'cluster 1'.

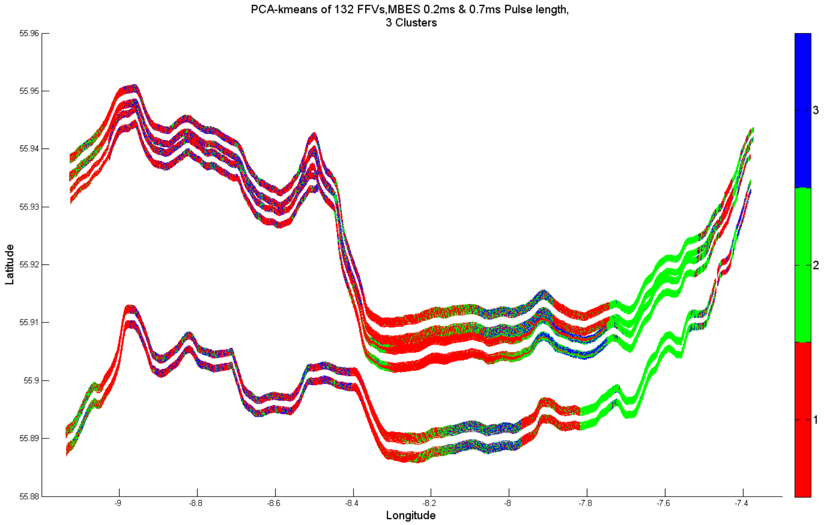


Figure 3.8a: PCA-k-means clustering of combined pulse length dataset of original 132 FFVs for 3 clusters

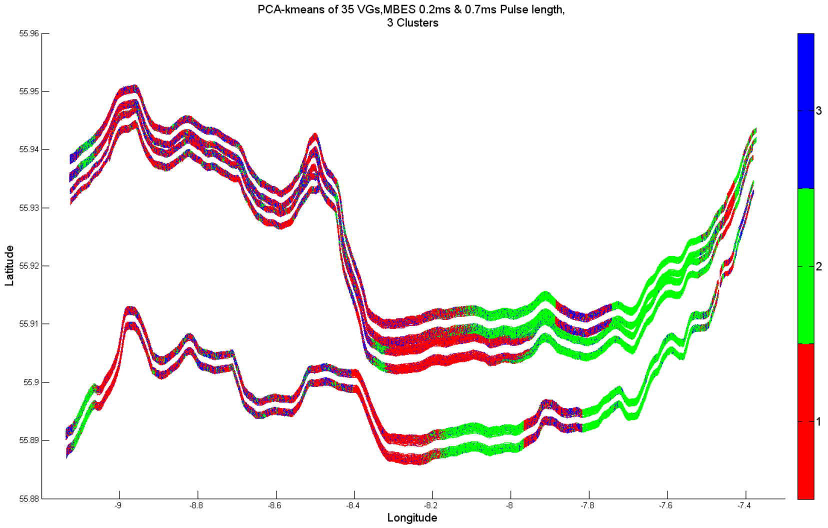


Figure 3.8b: PCA-k-means clustering of combined pulse length dataset of 35 FFVs from 16 visual groups for 3 clusters

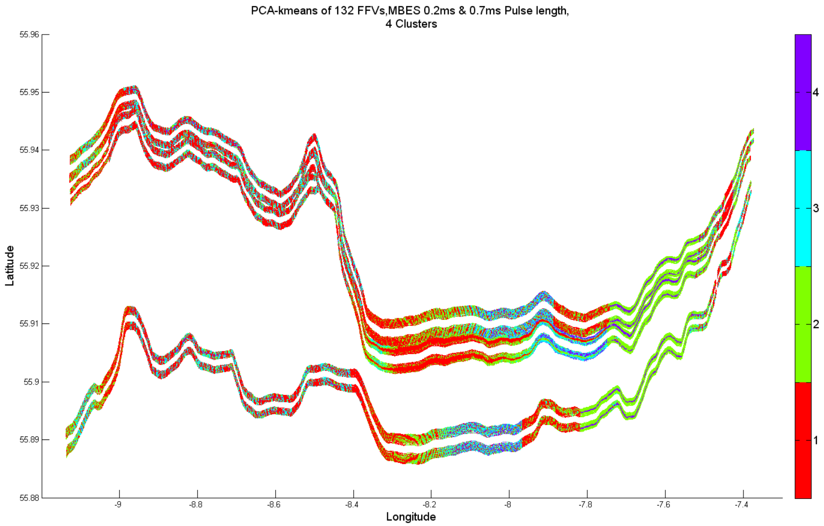


Figure 3.9a: PCA-k-means clustering of combined pulse length dataset of original 132 for 4 clusters

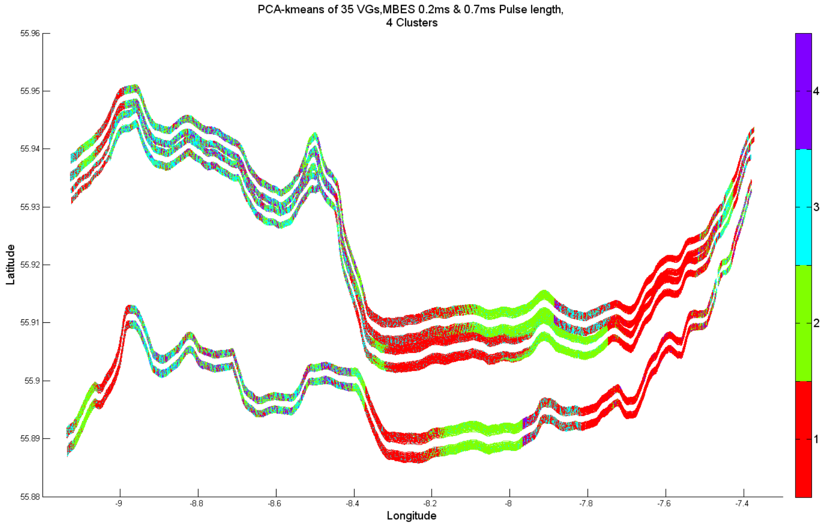


Figure 3.9b: PCA-k-means clustering of combined pulse length dataset of 35 FFVs from 16 visual groups for 4 clusters

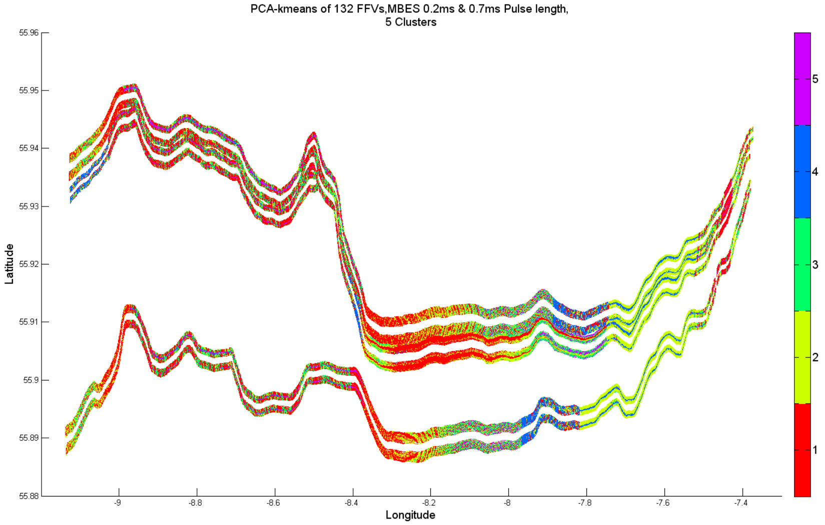


Figure 3.10a: PCA-k-means clustering of combined pulse length dataset of original 132 for 5 clusters

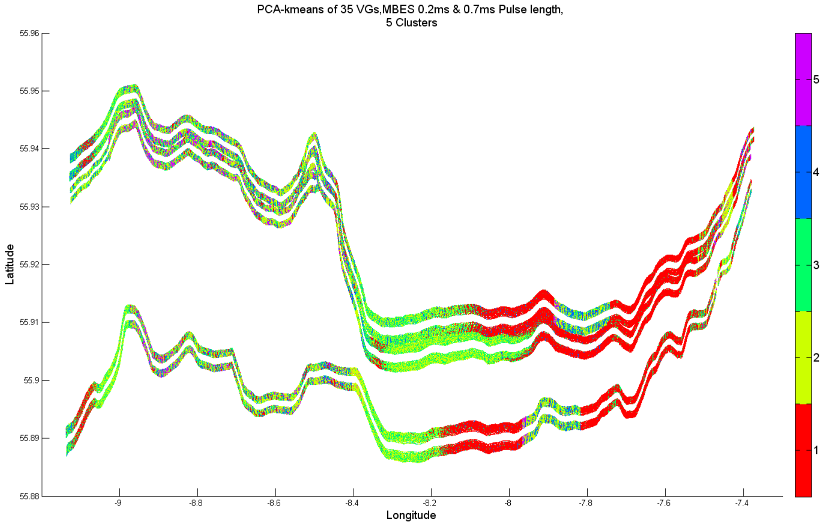


Figure 3.10b: PCA-k-means clustering of combined pulse length dataset of 35 FFVs from 16 visual groups for 5 clusters

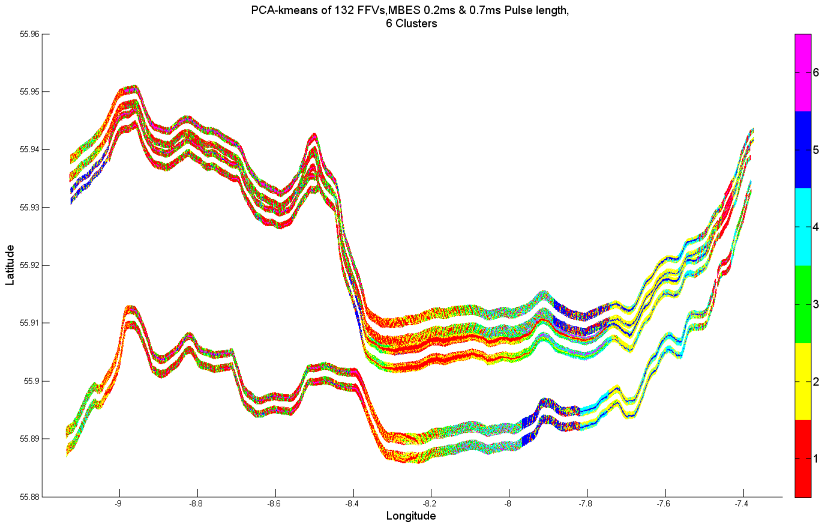


Figure 3.11a: PCA-k-means clustering of combined pulse length dataset of original 132 for 6 clusters

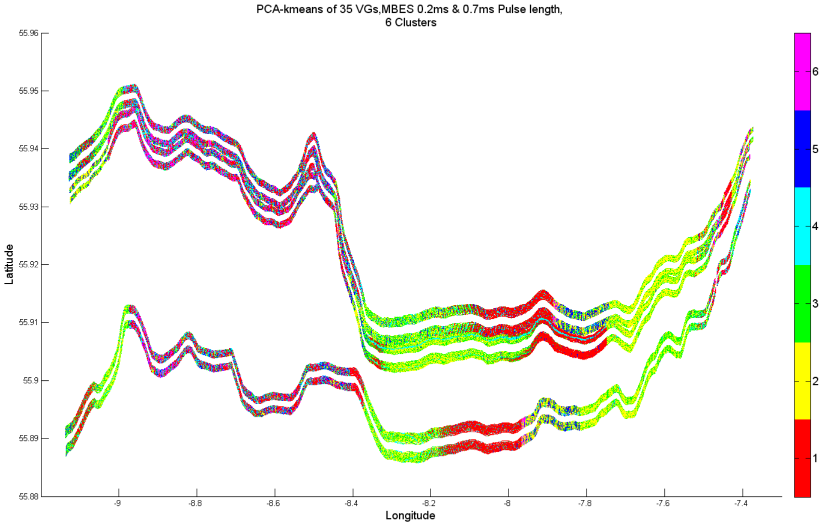


Figure 3.11b: PCA-k-means clustering of combined pulse length dataset of 35 FFVs from 16 visual groups for 6 clusters

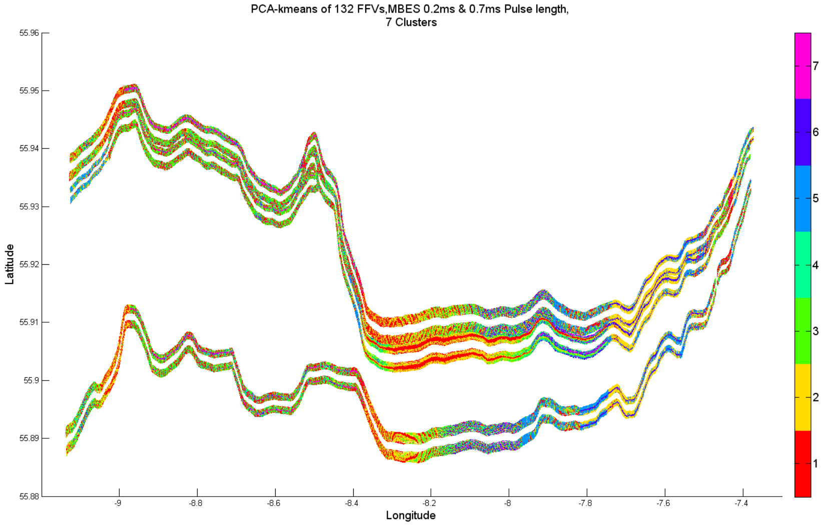


Figure 3.12a: PCA-k-means clustering of combined pulse length dataset of original 132 for 7 clusters

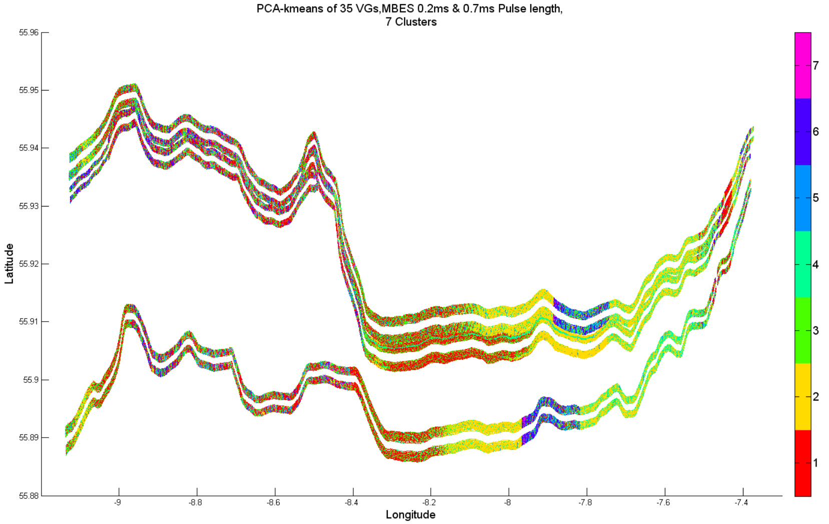


Figure 3.12b: PCA-k-means clustering of combined pulse length dataset of 35 FFVs from 16 visual groups for 7 clusters

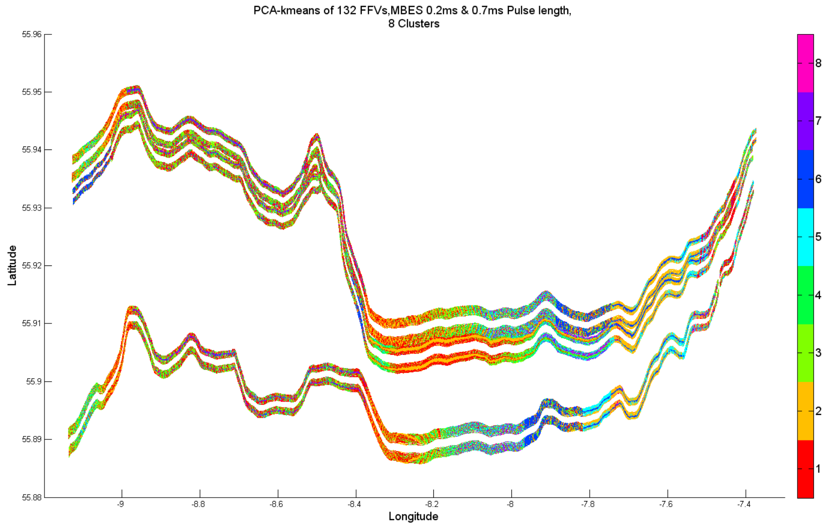


Figure 3.13a: PCA-k-means clustering of combined pulse length dataset of original 132 for 8 clusters

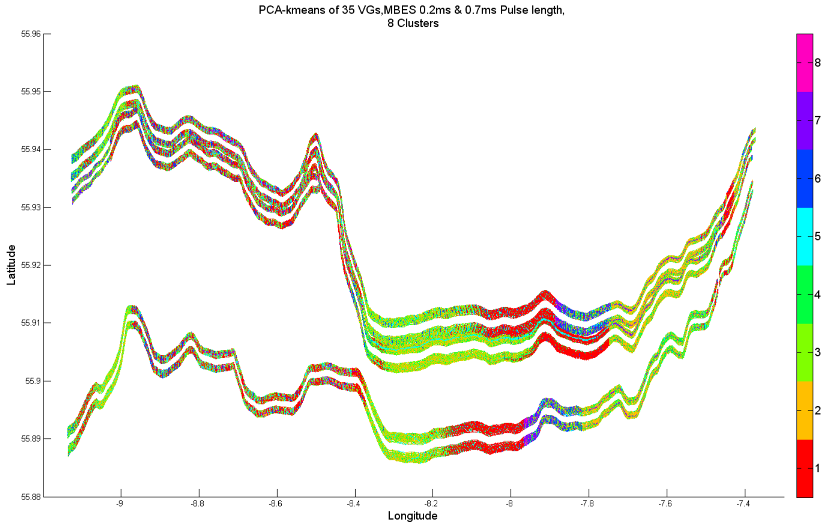


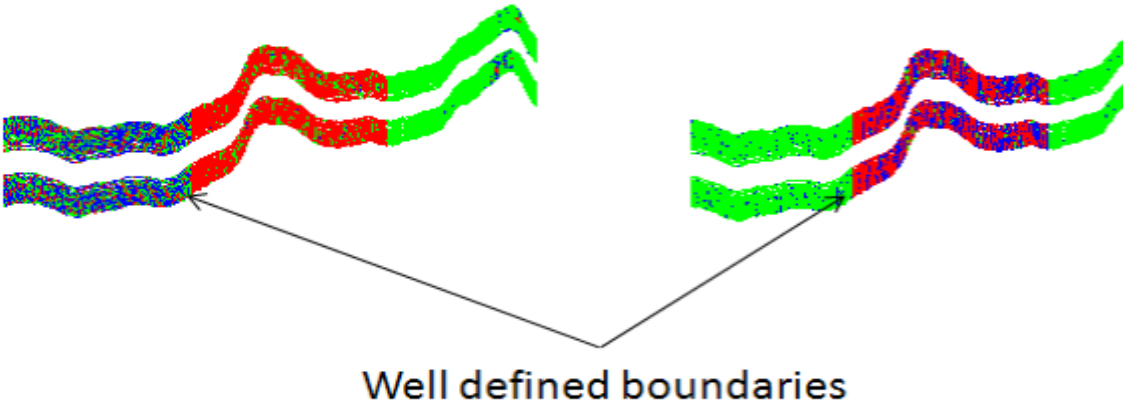
Figure 3.13b: PCA-k-means clustering of combined pulse length dataset of visual 35 FFVs from 16 groups for 8 clusters

The clustering results obtained using PCA and k-means are visualised using geographical scatter plots (each cluster label was plotted using their longitude and latitude) as shown in Figures 3.8 - 3.13. Each class label is visualised using different colours. An arbitrary number of classes (3 to 8) are chosen. In terms of the spatial distribution of clustering labels, it appears that the survey area can be divided into four to five well-defined clusters. The spatial distribution of major clusters appears to be mostly continuous, which is geologically acceptable. The other smaller clusters appear to be distributed in a pepper grain distribution, which, though acceptable for seabed geology, can represent noise or misclassification. The sixth cluster label appears to be more mixed and distributed compared to the first three (cluster three, four and five) cluster sets. For seven and eight set of clusters, the cluster boundaries become vague and the level of misclassification increases. Therefore, it can be assumed that the set of seven and eight clusters can be discarded as the results are too noisy to be considered as geologically valid clusters.

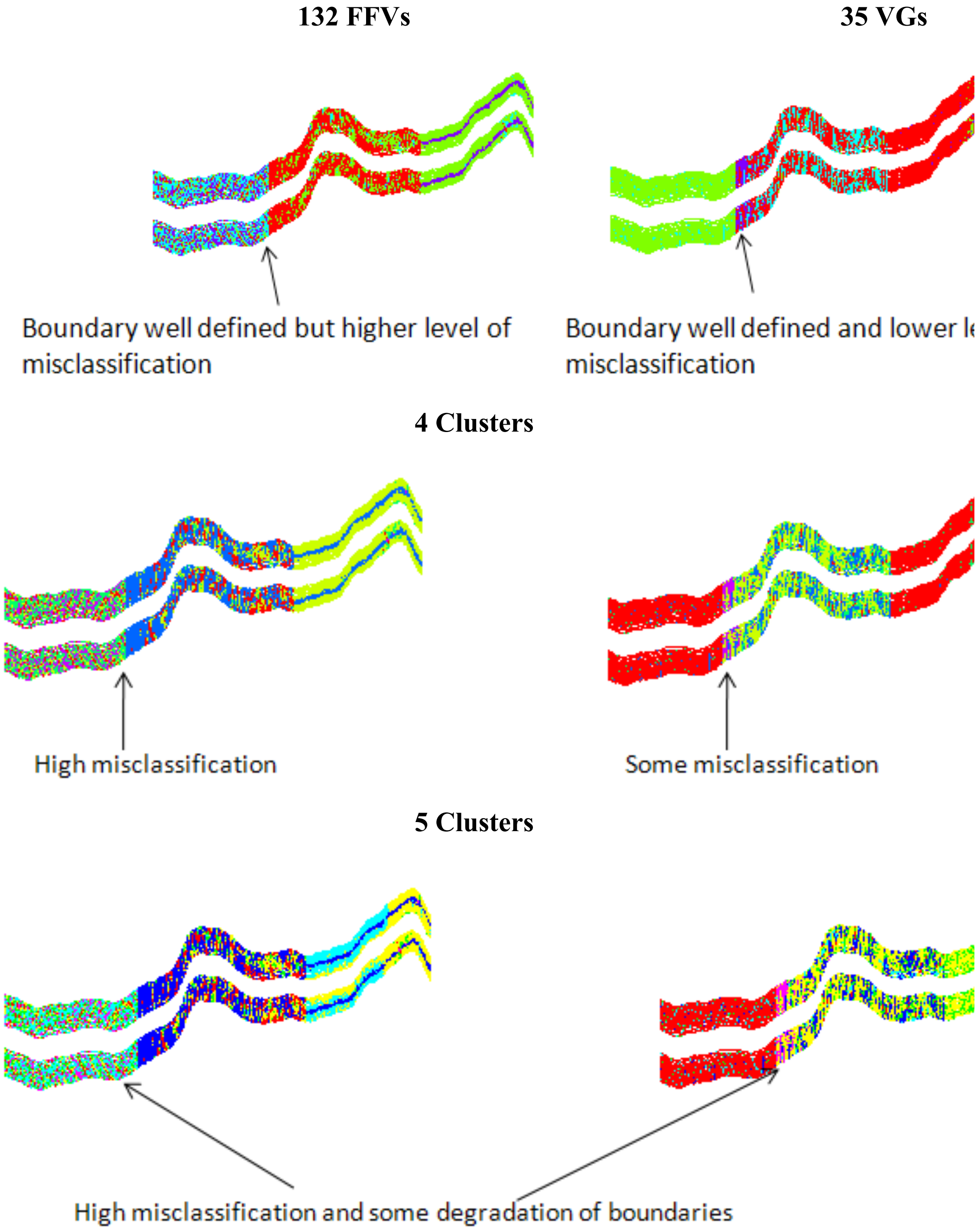
3.7.3. 132 FFVs vs. 35 VGs

To compare the effectiveness of reduced variables for clustering, a visual comparison is done with the geographical scatter plot of the clustering results from 132 FFVs (Figure 3.14).

132 FFVs 35
VGs



3 Clusters



132 FFVs

35 VGs

Boundary well defined but higher level of misclassification

Boundary well defined and lower l misclassification

4 Clusters

High misclassification

Some misclassification

5 Clusters

High misclassification and some degradation of boundaries

6 Clusters

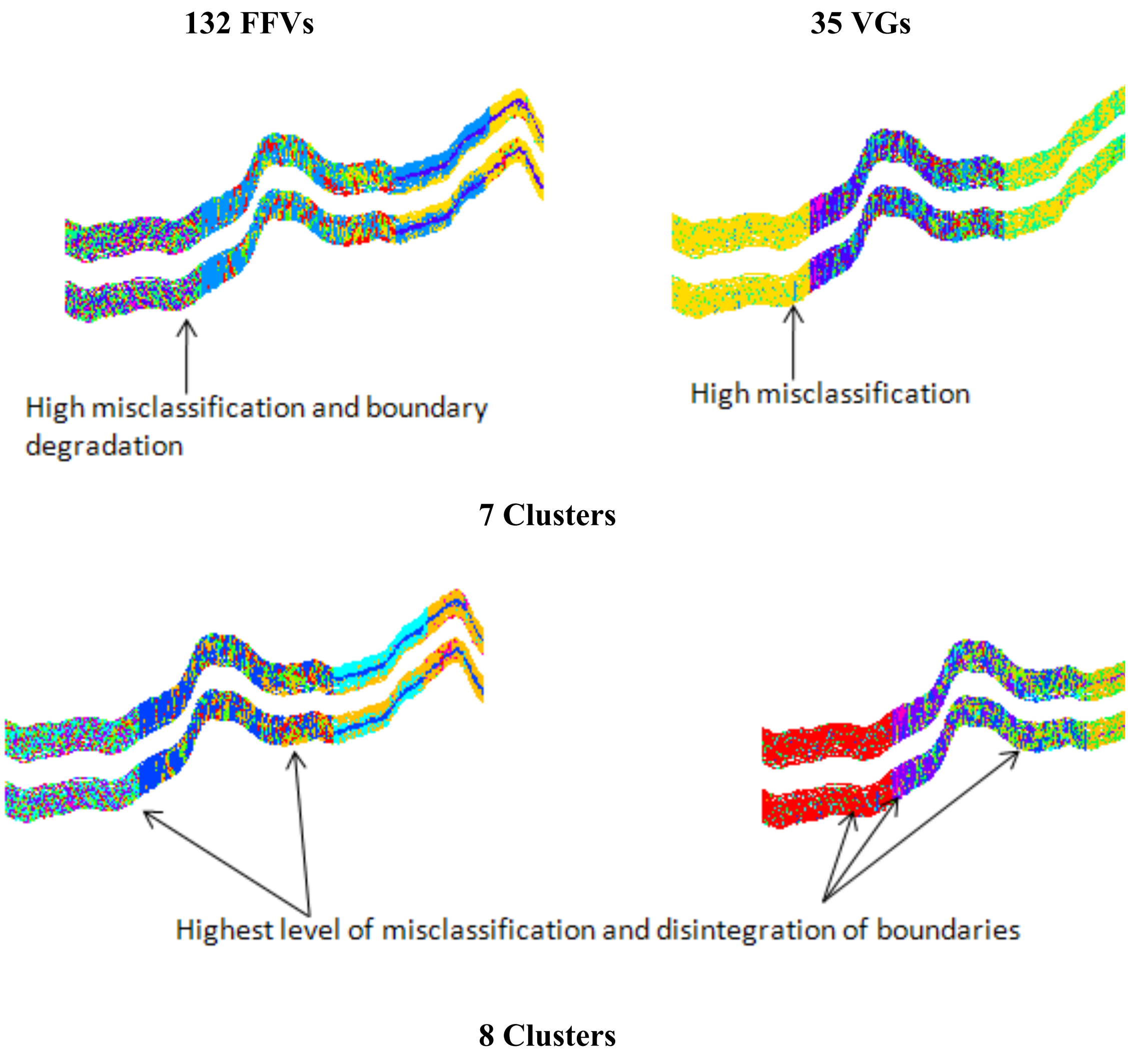

132 FFVs
35 VGs
High misclassification and boundary degradation
High misclassification
7 Clusters
Highest level of misclassification and disintegration of boundaries
8 Clusters

Figure 3.14: Comparison of clusters between dataset of 132 FFVs and 35VGs

A segment of the survey lines are taken for this comparison and the clusters and their boundaries were inspected. From initial comparison (Figure 3.14), the cluster results are very similar for three clusters. In both cases, the boundaries are well defined. Both the cluster sets have low level of misclassification. The cluster boundaries for the sets of four and five clusters appear to be better defined for reduced visual group dataset compared to the results of original dataset with 132 FFVs. A line of misclassified points can be seen for 132 FFV dataset from 4 clusters onwards. This misclassification only appears in the reduced dimensioned dataset when the cluster number is 8. While the cluster boundaries are well defined in both sets, the cluster labels from 132 FFVs showed significantly higher amount of pepper grain clusters which are likely to be a result of misclassification. From six to eight clusters, the boundaries of cluster labels from both dataset blurs with increasing amount of misclassification. A heavy presence of misclassified cluster labels can be seen throughout the survey lines and a valid identification of clusters would, therefore, be increasingly difficult.

In brief, from visual inspection, it appears that with reduced number of variables- a total of 35 visual groups instead of 132 full feature vectors- a better definition of cluster boundaries is achievable.

3.8. Discussion

With the rapid improvement of Multibeam Echosounders, seabed mapping is an emerging research area in Geographic Information Science. With the availability of ever-larger volumes of data, visual analytics can be an effective approach in the analysis of these datasets and the results of our analysis show that SOM is well suited in the exploratory analysis of large acoustic data. It also seems to be well poised for subsequent clustering and classification, which is the next step of this experiment.

Visual groups in the SOM component planes show that many FFVs are correlated on these particular MBES datasets and that there is high redundancy of information in the process. Because of this, not all 132 FFVs are necessary for further analysis - a sufficient number of attributes required (i.e. a minimal number of attributes that convey the same amount of information as all present 132 FFVs) equals the number of visual groups, which amounts to a total of 16 for both 0.2ms and 0.7ms pulse length survey lines. This number of attributes could be achieved by either selecting one representative FFV from each visual group (except for group 16) or taking an average (or some other combined measure) of all FFVs in each visual group as the representative of the respective group. As visual group 16 comprises of only singletons, they are all added resulting in a total of 35 variables. Another aspect of the result is that if we compare the similarity of colour patterns between the groups (see Appendix 3); some of the patterns are visually very different from each other. This means that probably the 35 dimensions is the optimum of variables required for conveying the information from this dataset and cannot be reduced any further.

In addition to visual similarity in the component planes, the visual groups showed a high degree of correlation within their respective groups, except for the singletons, which had a very low degree of correlation with each other as expected. This further strengthens that SOM component planes are well suited to detecting attribute similarity and a visual exploration can provide quite accurate assessment of the inherent similarity/dissimilarity among attributes in a dataset.

Different sets of cluster runs using k-means on all the principal components reveal that the lower dimension dataset appears to provide better-defined cluster boundaries. Though there were a considerable number of misclassified points in all cluster runs, this is not unexpected. The seabed, especially the shallow waters, can comprise plants, gravel, pebbles, shells etc. And the presence of these can contribute to misclassification even though the underlying ground is homogeneous in nature. Therefore it is more important to have a better cluster boundary as a starting point. The labels within those boundaries can be determined using ground truth data such as: underwater video recordings, grab samples, and grain size determination from the grab samples.

From the cluster results, the boundary definitions started to disintegrate when the clustering algorithm produced six clusters or more. The level of misclassification was significant enough to deduce that from six clusters onwards, the cluster boundaries were not defined well enough to achieve a valid classification and thus are discarded for classification and mapping. Another conclusion from the cluster results is that reduced dimensioned dataset has better cluster definition and therefore should produce better classification maps. This is to be validated by producing classification and subsequent mapping of the cluster result (Chapter 5).

In QTC Multiview^{TM} (Preston, 2009) classification methodology, the number of dimensions is reduced from 132 to three using the PCA as the next step after the FFV space is derived. When comparing this to the required and necessary attributes in our results, the question arises if three dimensions in the QTC method are really enough to capture the variability of backscatter characteristics or if some other larger number of the dimensions should be considered. Our concern is that whether only three components of PCA is optimum for dimensionality reduction for such large datasets. In a typical survey area, each survey line has around six to eight million records and all the survey lines cover a large area (a couple of hundred square kilometres). As the first three components of PCA consist of 90-95% of the information, there is a possibility that the remaining 5-10% of excluded dimensions could hold essential information due to the sheer volume of the dataset and survey area. Therefore, for this project, all the components were included when clustering the data.

In this experiment we used MBES data from four survey lines collected in an area (Malin Sea) covering several seabed types. While it may be sufficient to use 35 variables in this environment, the numbers can be further reduced in other areas with different acoustic variability. It can also happen that the number of variables may need to increase if a larger area is more spatially varied (for example: presence of ridges, cliffs etc.). Therefore we plan to apply the same approach to other MBES data lines by further analysing other areas with different seabed types (e.g Irish Sea, coastal embayment) to examine if acoustic variability has an effect on the similarity of FFVs collected from a larger area. However, this is beyond the scope of this project and is currently considered for future research.

We also did not take into account the spatial correlation of the MBES data in this approach as we only used the attribute space (i.e. FFVs of backscatter) as the input to the SOM and did not consider the geographic location of the patches. Therefore geographic space did not play any role in this analysis. However, due to the continuity in the geographic processes that formed the seabed, spatial autocorrelation is most likely present in the backscatter data when the survey area is very large and should be considered in the analysis. This could be done by using a GeoSOM (Bação et al., 2005) instead of the classic Kohonen SOM. In the GeoSOM, the algorithm is run on both the attribute and the geographic space, which means that both the geographic distance as well as the distance in the attribute space plays a role in how similarity of data objects is defined. For this project, the survey area was not large enough (just 4 survey lines) to consider spatial correlation to be a significant variable for classification.

As we examined one particular complexity issue of the QTC classification method, potential similarity and redundancy of FFVs, the particular visualisation of the SOM result - the component planes – proved to be an adequate tool to examine similarity of attributes (the FFVs). It is shown that SOM can be an effective tool for examining similarity of attributes in an acoustic database and can be used as a dimensionality reduction tool to bring an extremely large acoustic database to a more manageable size and, therefore, can make the clustering and subsequent mapping less computationally intensive and less time consuming.

Chapter 4

Visual exploration and Clustering of SBES data

This chapter presents a new approach in the processing and clustering of singlebeam echosounder (SBES) data. A visual exploration tool will be used to detect outliers that would have, otherwise, gone undetected in automatic outlier detection algorithm. A direct clustering method will also be tested on the optimised SBES dataset.

Chapter contents

Characterisation of seabed floor type can provide useful information for various applications such as seabed mapping (Biffard et al., 2005), impact assessment of human activity (Eastwood et al., 2007), nature conversation (Hamilton, 2001) and underwater plant species classification (Brown, et al., 2007). Some of the traditional method of sediment type characterisation such as coring, grab sampling and visual inspection are increasingly becoming impractical due to the time and cost involved and can often lead to misleading spatial distribution of seafloor types, especially in survey areas where sediment patterns are too complex (Ellingsen et al., 2002). The difficulty of direct seabed access, even in shallow regions, means that most seabed surveys are carried out remotely. By far the most popular and widely used technique in seabed-related remote sensing is a survey based on the measurement of sound energy using either multibeam or side scan sonar. Seabed classification is also done prior to ground truth data collection. This is usually done on a smaller set of data to provide an overview of the underlying seabed geology. Normally, either MBES of Side Scan Sonar (SSS) are used for this purpose. Singlebeam sonars are rarely used in seabed mapping and when they are used a traditional feature based approach is adopted. A direct clustering approach (clustering of raw backscatter data) of singlebeam sonar however has not been used in the context of seabed mapping, despite its ability to provide additional subsurface information and thus contribute in areas with complex sediment structure as well as saving time and cost from not having to generate any statistical features. This chapter evaluates the potential of direct clustering of singlebeam echosounder (SBES) data to aid seabed mapping, particularly in the context of ground truth data collection planning.

4.1. Research background and justification

Common applications of SBES include bathymetry or the measurement of seabed depth, based on echo return timings of a transducer that emits one sound wave (see Chapter 2). Today, most seabed type characterisations are based on classifications of digital data from multi-beam echo sounders (MBES), where multiple sounders are used simultaneously to allow a greater area of coverage during seabed surveying. Although SBES time series potentially contain valuable echo data in yielding classification information, SBES typically receives less attention because it was not primarily designed for that purpose. However, the echo time series interval directly beneath the detected seabed has a strong signal return and minimal distortion within the return beam as the return is from vertical scattering. Therefore, this part of SBES time series contains relatively undistorted information on the subsurface composition and is thus a good candidate for seabed classification. A direct clustering of the SBES echoes would eliminate the need of feature extraction for exploratory clustering prior to ground truthing. Exploratory clustering normally using MBES or SSS features (and rarely SBES features) is done to give the surveyors/geologists an initial idea of the class distribution of the survey area (Hung et al., 2010). This clustering helps the marine geologists to optimally design ground truth sample collection points as gathering and retrieving ground truth samples can often be a very expensive operation. However, feature extraction on large datasets, as MBES, SSS or SBES data are, is usually a very time consuming procedure. A successful direct clustering of SBES data would, therefore, save time, reduce cost by eliminating unnecessary ground truthing and thus further optimise the survey process. This project discusses the viability of such an approach.

This data processing exercise assumes that the rate of seabed type variation is much smaller than the spatial extent of survey, while the characteristics of different seabed types are captured only in a limited section of the echo return signal. The proposed approach of clustering the SBES aims at clustering only the segments of backscatter returns that capture the seabed surface information as well as some subsurface information.

The chapter is structured as follows: the following section discusses SBES data acquisition and processing. This is followed by a description of the experiment that was performed. Lastly results are presented followed by discussion of the results.

4.2. SBES data acquisition

Data were collected by Geological Survey Ireland (GSI) and the Marine Institute (MI) as part of the INFOMAR programme during a series of Malin Sea survey in the year 2003. The main equipment used were the Kongsberg™ EA600 singlebeam echosounder (Kongsberg & Ea, n.d.), which simultaneously emits and records echo returned at sonar frequencies 12, 38 and 200 kHz, respectively. Malin Sea (Chapter 3, Figure 3.4) is located to the north of the Republic of Ireland and has been chosen as a case study due to its known seabed type variations. Both MBES and SBES returns were acquired from a single survey vessel during the same survey runs simultaneously. Chapter 3 analysed the data acquired from MBES. In this chapter an example survey track, line 163, is used throughout the project for analysis. The track runs East-West at latitude of 56N and covers a survey distance of about 110 km.

4.3. Challenges with the SBES dataset

In the conventional method of classifying seabed type, statistical features are extracted from SBES datasets for clustering and classification. RoxAnn^{TM} and QTC Impact^{TM} are the two most commonly used software platforms for this purpose.

Principal component analysis (PCA) is performed on the extracted features for orthogonalisation and dimensionality reduction (by selecting first three components), followed by k-means clustering. The noisy nature of the SBES datasets affects both these methods thus introducing a relatively high level of uncertainty in the classified map (Hung et al., 2010; Satyanarayana et al., 2007; Zimmermann & Rooper, 2008).

This research project focuses on an alternative approach: a direct clustering of SBES data. Data used in this study consist of echo segments from both above and below the seabed surface. For this method to be effective, it is important that the data are first cleaned of any outliers that could contribute to the overall noise and subsequent misclassification. Therefore a visual exploration of the dataset is setup after it was smoothed using second order Butterworth low-pass filtering and 'cleaned' of the bad samples by an automatic outlier detection algorithm.

Low-pass filters are electronic filters that enhance low frequency signals but attenuate signals with frequencies higher than a cut-off value. The actual amount of attenuation for each frequency varies from filter to filter. Designed by Steven Butterworth (1930), the frequency response or gain (the ability of a circuit, often an amplifier, to increase the power or amplitude of a signal from the input to the output) in this filter is given by:

$$H(j\omega) = \frac{1}{\sqrt{1+\varepsilon^2\left(\frac{\omega}{\omega_p}\right)^{2n}}} \tag{4.1}$$

Here, n represents the filter order, omega ω, the radian frequency, is equal to $2\pi f$ (f=frequency) and epsilon ε is the maximum pass band gain (Electronic-Tutorials, 2011). The smoothed echoes are then passed through an outlier detection algorithm (Hung et al., 2010), which is described in the next section.

Our particular SBES dataset contains a segment from the raw echo that is 5m above and below seabed surface (Figure 4.1). In Figure 4.1, the value '0' on x axis represents the seabed surface. As the echo hits the seabed surface, it should reach its peak amplitude. However, from the figure we can see that the data contain considerable measurement error in bathymetry determination and contain a good amount of vertical displacement. Because of this, the peak amplitudes are distributed as much as 3m above seabed surface. This error raises the risk of similar classes (time series with a similar shaped peak) being classified as different classes and therefore the resulting clusters are likely to be noisy in nature with not-so-well defined geographic boundaries. However, the subsurface information should result in representative clusters, as it is the echo time series interval immediately beneath the detected seabed that provides the clearest information for seabed classification, due to the strong signal return and ease of detection. Because of the noisy nature of the data, to achieve reliable and consistent classification results, we introduced an additional step in the outlier detection procedure. The data already come processed with Butterworth filter and automatic outlier detection. In this chapter we add visual exploration using a time series visualisation too as the final step in outlier detection procedure to ensure the best possible optimisation of the dataset prior to direct clustering.

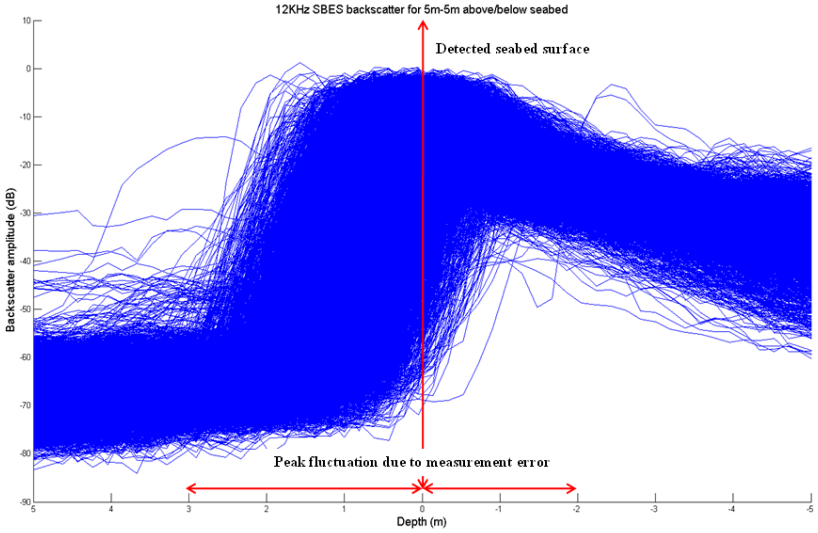


Figure 4.1: SBES backscatter at 12 KHz

Our dataset was collected at three different frequencies to accommodate the varying degree of softness of the seabed. Lower frequencies (12 & 38 KHz) provide good penetration of the seabed when the surface is hard while the higher frequency (200 KHz) yields better surface information. The study area is expected to be sandy in nature with burrows and shells. As a result, the 200 KHz data showed significant fluctuations (Figure 4.2) even after filtering. This frequency was, therefore, not expected to yield good clustering results. This assumption was later confirmed in our analysis.

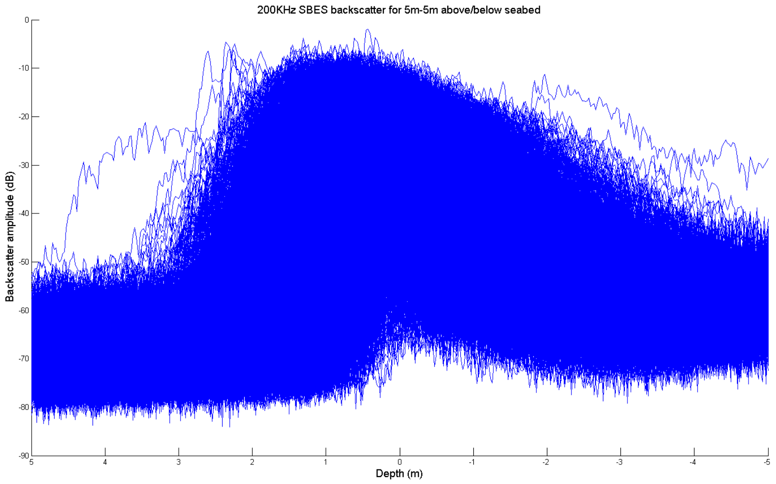


Figure 4.2: SBES backscatter at 200 KHz

4.4. SBES data processing

A typical SBES dataset consists of high dimensional (data records at varying depths) time series data containing heteroscedastic (random in nature) noise. A large number of samples are usually collected in each survey. Since SBES data contain less geological information and redundancy for quality assurance compared to MBES, additional measures need to be taken to ensure the quality of the raw data before further processing. The pre-processing of SBES data involves inspection for data integrity (usually done by an expert geologist on the survey boat), initial 'cleanup' to mitigate the effects of systematic errors of sonar measurement, including tidal movement, decibel normalisation and conversion to industry-standard format for compatibility and storage purposes. In the next step, spatial sub-sampling is carried out to reduce data size, followed by seabed depth determination targeted for feature extraction. These data processing steps are carried out by GSI during and after survey before supplying the data. The SBES dataset (line 163) used in this study was pre-processed by Hung et al. (2010) as explained in the following section.

4.4.1. SBES dataset cleaning

For this project, the aim of identifying seabed depth is to facilitate the capture of complete echo envelope by segmenting the sonar data into two sections, above and below the seabed. Before our experiment, described in sections 4.5 and 4.6, data were further cleaned by Hung et al. (2010) as follows. With the assumption that the seabed is located somewhere at the peak amplitude of each echo return from the lowest frequency sonar, each time series was smoothed using a second-order Butterworth low-pass filter. The aim is to smooth the backscatter to some extent so that outliers in the time series dataset are removed and ‘good quality’ data are kept for further analysis (Bianchi & Sorrentino, 2007; Butterworth, 1930). However, upon initial inspection of the data (Figure 4.1), it was observed that there were small but regular bathymetric fluctuations (~5m). These are probably caused by the inherent measurement noise and are unlikely to be due to geological variations.

Some corrupted samples with abnormal changes in bathymetry still remained in the dataset after its initial pre-processing, which could potentially cause data inconsistency. The dataset was further processed by an automatic procedure to identify such bad samples. The procedure that was implemented was as follows (Hung et al., 2010):

We consider each unfiltered SBES time series as column vector **z** where

$$z_i = \{z_{ik}\} \quad i = \{1,2,\dots,m\}\; k = \{1,2,\dots,n\} \tag{4.2}$$

Here, i is the number of time series collected, while k is the number of temporal samples in each time series. The output matrix can therefore be written as:

$$Z = \{z_i\} \quad i = \{1,2,\dots m\} \tag{4.3}$$

$$\tilde{Z} = \{z_i\} \quad i = \{1,2,\dots,m\} \tag{4.4}$$

Here, ~ denotes data derived from the filtered SBES data. Next, a ten-fold stacking was performed on the dataset to improve the signal-to-noise ratio:

$$Y = \{y_j\} = \left\{\left[\frac{1}{10}\sum_{i=1}^{10} z_{i+j}\right]_j\right\}, \qquad j = \left\{1, 10, 20, \dots, \frac{m}{10}\right\} \tag{4.5}$$

The filtered version $\tilde{Y}$ was obtained in a similar fashion. To indicate the segment of echo return used, the convention ‘A_B_’ was employed. As such, A*a*B*b*, which represents the echo segment ‘a’ meters above and ‘b’ meters below the seabed from each time series. For example, A5B25 represents 5 m above and 25 m below the seabed. Then, **Y** matrices are grouped together to form the corresponding **X** family matrices.

$$X = \{x_i\} = Y_{AaBb}, \qquad where \;\; X \subset Y, i = \{1, 2, \dots, m\} \tag{4.6}$$

$$\tilde{X} = \{\tilde{x}_i\} = \tilde{Y}_{AaBb} \tag{4.7}$$

$$\tilde{X} = \{\tilde{x}_i\} = \tilde{Y}_{AaBb} \tag{4.8}$$

Next we define the variation in mean spatially filtered echoes, v, as

$$v = \{v_i\} = \{E(\tilde{x}_i) - E(x_i)\} \qquad i = \{1, 2, \dots, m\} \tag{4.9}$$

Here E is the expectation operator. In the next step, the following automatic outlier decision rule was employed:

$$b_i = \begin{cases} 1 & 'bad' \;\; if \; |v_i| > \alpha s(v) \\ 0 & 'good' \quad otherwise \end{cases} \qquad i = \{1, 2, \dots, m\} \tag{4.10}$$

Here $s(v)$ is the standard deviation of all v on the same survey track and α is a threshold based on the size of the ‘bad’ data region. In other words, any time series sample vector found outside α times the standard deviation of v will be labelled as ‘bad’. The ‘bad’ data detection is performed for all three frequencies. This resulted in a cleaned dataset that was provided to us for subsequent direct clustering. However, before proceeding with the direct clustering we used visual exploration to evaluate if the ‘cleaned’ dataset using the above automatic method was adequate in detecting all the bad samples or outliers.

4.5. Experiment, step 1: Visual exploration of SBES for outlier detection

Visual exploration in general facilitates data visualisation so that human cognitive skills can be implemented for pattern recognition in large datasets. The objective of using visual exploration on time series SBES dataset is to identify and detect any echo time series that appears to be a bad sample and still remains in the dataset after it has been filtered and cleaned using the automatic algorithm described in the previous section.

The tool of choice for visual exploration is TimeSearcher©, which was developed at the Human–Computer Interaction Laboratory, Department of Computer Science, University of Maryland (Hochheiser and Shneiderman, 2001). This software facilitates interactive visual exploration of time series data.

4.5.1. SBES exploration with Time Searcher

Once the data are loaded into TimeSearcher©, the overview panel at the bottom of the screen displays the time series for one of the variables. This panel can be used to specify a specific region of the time series to be focused on. Furthermore it is possible to explore the attributes in the time series and to select specific time series based on selected attributes.

The Timebox query within TimeSearcher© was used for visual exploration of the SBES dataset. Timeboxes are rectangular query regions drawn directly on a two-dimensional display of time series data. Time boxes in TimeSearcher© can be created by clicking on the desired starting point of the Timebox and drag the pointer to the desired location of the opposite corner. The extent of the Timebox on the x-axis specifies the time period that it constraints, while the extent of the y-axis specifies the constraint the Timebox puts on the range of values of interest in the given time period. Given a set of time series datasets, a Timebox acts as a filter that accepts only those items that have values in the given range during the interval spanned by the box.

To be specific, a Timebox has four tuples (Figure 4.3): b=(t_{min}, t_{max}, V_{min}, V_{max}). Suppose, n_i is an item in a time series dataset (N), where $n_i \in$ N and n_i(j) is the value of n_i at time j. The item n_i will only be selected by the Timebox if it satisfies the following:

$$t_{min} \leq j \leq t_{max} \tag{4.11}$$

$$V_{min} \leq n_i(j) \leq V_{max} \tag{4.12}$$

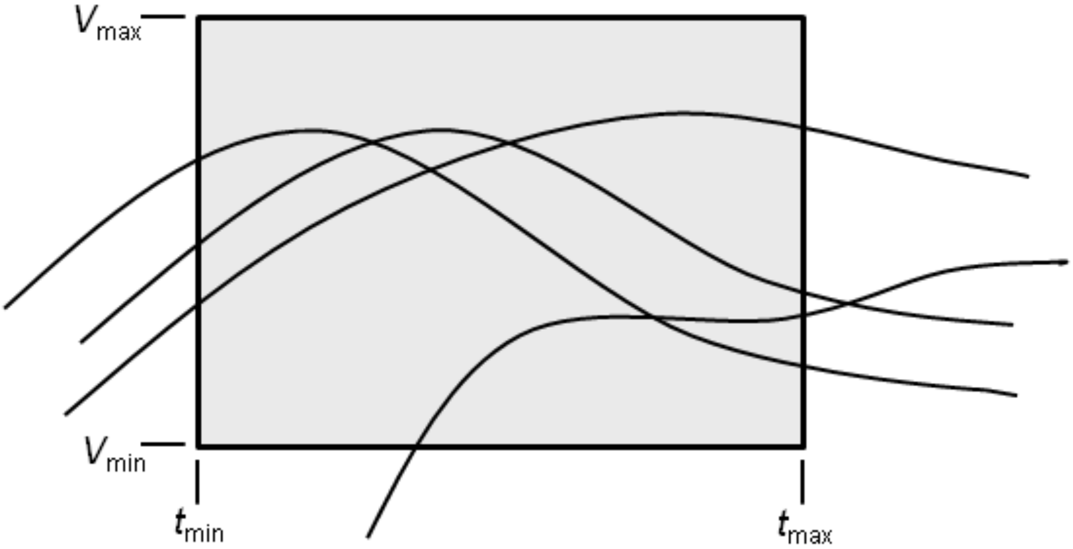


Figure 4.3: A Timebox query expresses constraints in time and value

A Timebox can be dragged to a new location or resized via appropriate resize handles on the corners in the TimeSearcher© environment. Every time a Timebox is updated, the query is re-processed and the visualisation updated accordingly. Multiple Timeboxes can be drawn to specify conjunctive queries. Time series data segments in any subsequent Timeboxes must meet all of the constraints implied by the previous Timebox in order to be included in the result set.

In our experiment, first the filtered SBES dataset was loaded into TimeSearcher© for visual exploration. Timebox query within the software environment was then used to interactively detect any anomaly that may still have remained in the dataset. Once the approximate locations of outliers were detected, the Timeboxes were further optimised in size and location to isolate and enhance the outliers. Once the outliers were identified, they were removed from the dataset using Matlab and the data was then ready to be used for direct clustering.

4.5.2. Results

The first objective of this research was to use visual exploration to detect outliers in acoustic time series data that may still be present after the data is filtered. Figures 4.4, 4.5, and 4.6 show the results of outlier detection using TimeSearcher©.

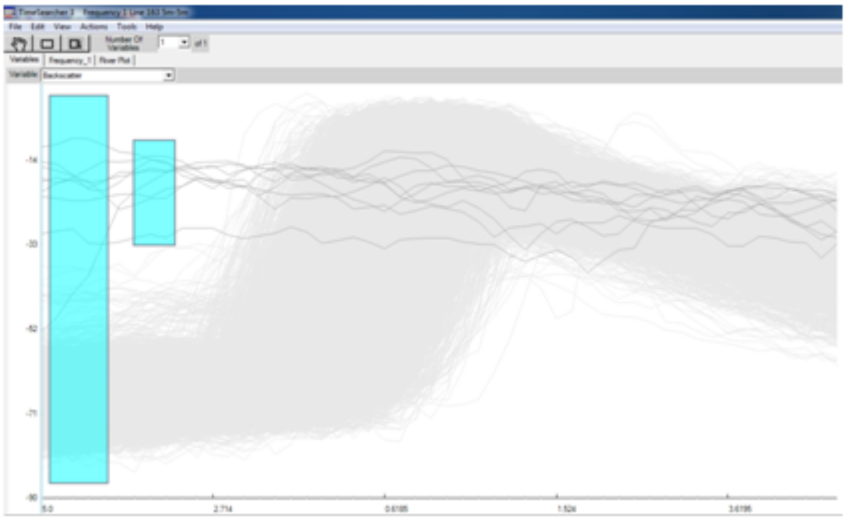


Figure 4.4: Detected outliers in SBES echo returns (12 KHz)

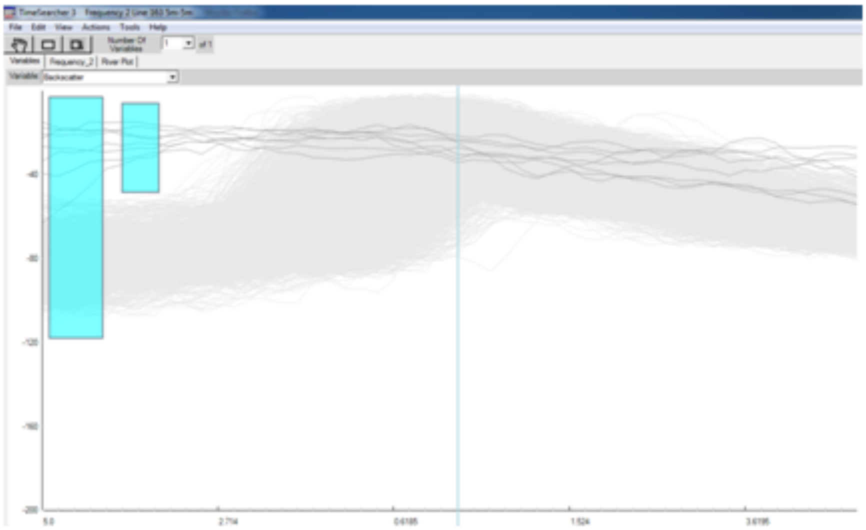


Figure 4.5: Detected outliers in SBES echo returns (38 KHz)

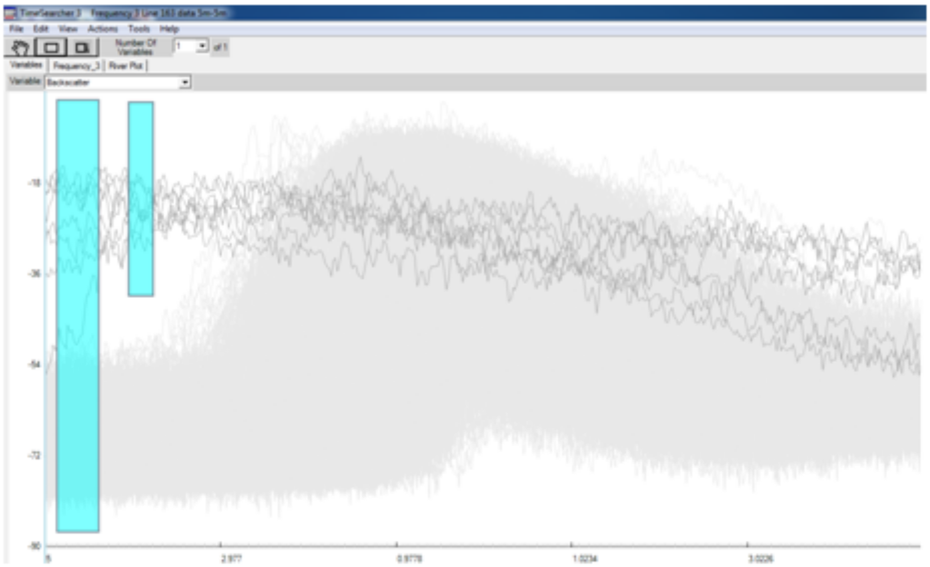


Figure 4.6: Detected outliers in SBES echo returns (200 KHz)

In TimeSearcher©, each time series is assigned a unique identifier. In our case the echo returns were assigned the identifiers T1, T2,…, Tn. Based on their identification numbers, the identified outliers in all three frequencies appear to be located in the same geographic locations and surveyed in sequential order. The locations of the outliers are at the very right end of the survey line (Figure 4.7). This indicates that these echo returns were recorded at the beginning or at the end of this particular survey run. Therefore, it is probable that they represent a gross error rather than a systematic error. Gross error is attributed to the surveyor, while systematic errors occur due to inherent limitations of the instruments and can be calibrated or their effect quantified.

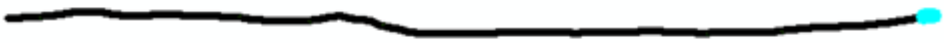


Figure 4.7: location (highlighted) for SBES echo returns (12, 38, and 200 KHz)

These outliers or bad samples were very obvious in the visual exploration due to the fact that none of them contained any peaks. Peak occurs when the echo hits the seabed surface and as the echo penetrates the seabed, the peak value dissipates. The absence of a peak in the echo return indicates that the echoes have not hit the seabed or that the surveyor was in the process of re-calibrating the instruments as is often the case at the beginning or the end of any survey run. As there was no peak present in these echo returns, no further tests were necessary and the echoes were discarded from the dataset.

It is, however, important to point out that in spite of the obvious effect (the absence of a peak) these echoes were not detected as outliers in the automatic outlier detection process described in the previous section. Without additional visual exploration they would have remained in the data and influenced the results of subsequent classification.

4.6. Experiment, step 2: Clustering of SBES data

Two clustering algorithms were selected for direct clustering of SBES backscatter echoes. The de facto PCA and k-means clustering (QTC, 1997) was first tested on the dataset. PCA was used to orthogonalise the dataset and all components were used for k-means clustering. Fuzzy clustering was initially selected for its ability to accommodate overlapping of clusters. However, as the survey dataset for this study is relatively small and as only one survey line was available for testing, we decided not to use different levels of fuzziness. This decision was based on the fact that the seabed in this area is mostly sandy in nature and therefore just one narrow survey line would not be adequate in capturing the overlapping of different clusters. Therefore, a fuzziness of 1% was used to treat the clusters as discrete clusters. Each of the algorithms were run several different number of clusters (3 to 8) and in order to minimise the impact of hitting ‘poor’ local minima by chance, a total of 10 Monte Carlo simulations were performed for each cluster runs. The labels with the least mean of errors were retained as the optimum results and presented.

After the clustering, the optimal number of clusters was determined using cluster validation indices, which measure the quality of clustering through level of separation between clusters. There are several available and for this experiment four indices were used: Calinski-Harabasz (VRC) index, Davis-Bouldin index, Dunn index and Silhouette index (see Chapter 2). For the VRC index, the cluster number with the highest index value represents the optimal number of clusters. For the other three, it is the lowest index number that indicates the optimal cluster number. The results of clustering and cluster validation are described in the following section.

4.6.1. SBES data clustering

One of the objectives of this research is to test if direct clustering of SBES instead of commonly used feature based clustering can produce representative clusters from a dataset of SBES echoes. Here we present results from the two methods as described above: PCA + k-means & Fuzzy c-means. In order to facilitate visual comparison, results are presented based on cardinality sorting (ascending order) of the clusters obtained i.e. the cluster with highest number of points are always labelled as cluster 1. Due to the compactness of data points and relatively small length of the survey line, the visualisation of the clustered point data was not possible. At full extent, the points overlapped significantly and as a result the survey line appeared very dark, even after the points were enhanced (see Figures 4.8-4.10). A different approach was taken to visualise the coherence of obtained cluster boundaries. The survey line was converted to raster with spatial resolution set to three meters. This resolution was selected after several trials and the raster with three meters resolution appeared to produce the best geologically viable output. Once the cluster areas in the raster representation were extracted, the clusters were labelled and coloured based on the largest amount of clustered points within each area. The clustered points belonging to each cluster area were counted and if one cluster label in any given cluster section had 50% or more of the total points then that cluster was coloured with the colour of the dominant cluster points. The purpose of generating a rasterised representation of results was to provide a clearer view of the location of cluster boundaries (i.e. spatial distribution of demarcation lines between cluster areas). Therefore, when examining cluster areas in the raster, only the location of their boundaries should be examined and cluster labels (i.e colours) should be ignored.

4.6.2. K-means clustering results

Figures 4.8-4.10 show the results from SBES clustering using PCA + k-means (three clusters).

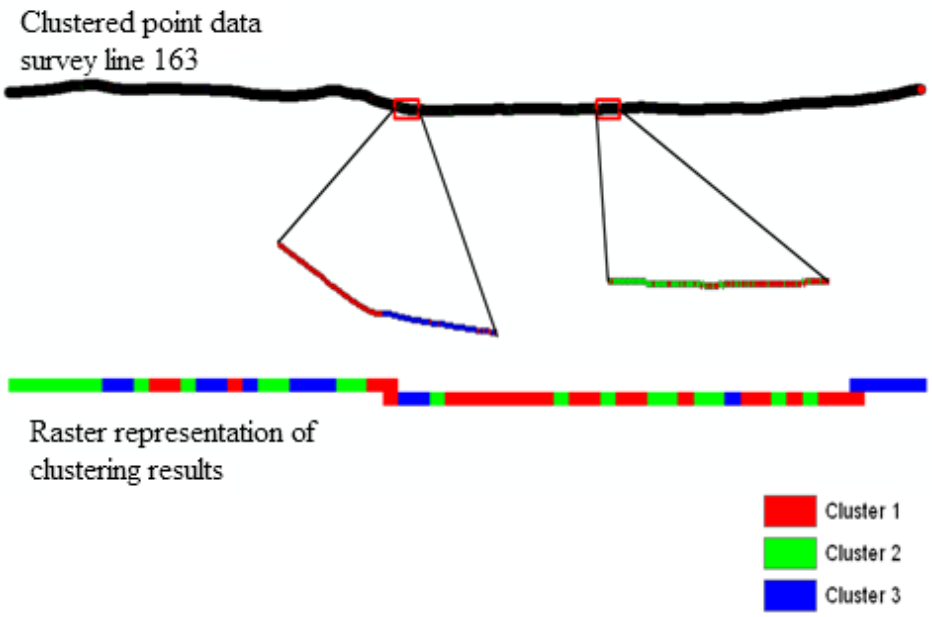


Figure 4.8: PCA + k-means clustering of SBES dataset (12 KHz)

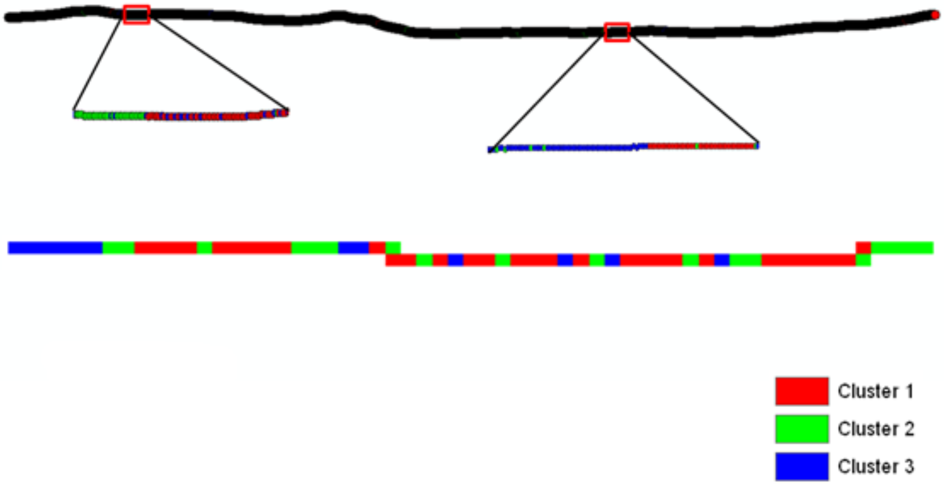


Figure 4.9: PCA + k-means clustering of SBES dataset (38 KHz)

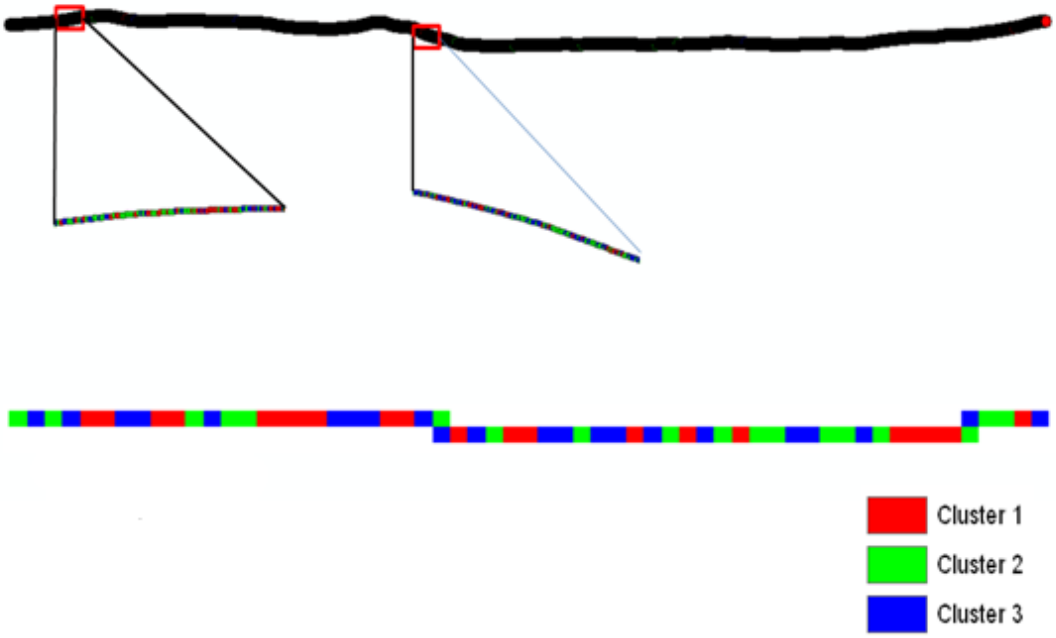


Figure 4.10: PCA + k-means clustering of SBES dataset (200 KHz)

Working frequency for SBES clustering

Visual exploration of results allowed us to get an idea of the suitability of acquisition frequencies of sonar for direct clustering. Low cluster numbers were chosen for this exploration in order to minimize the effect of misclassification.

While direct clustering of backscatter echoes is expected to produce some misclassification due to the noisy nature of the data, the level of misclassified points was higher compared to that of MBES classification (Chapter 3). From the figures, it was evident that while the cluster boundaries were to some extent similar in distribution for frequencies 12 and 38 KHz. Clustering of the 200 KHz dataset did not seem to provide a reliable result. Comparing the raster boundaries and zoomed sections of the clustered 200 KHz dataset, it seems that the effectiveness of k-means was severely affected due to the nature of the data. The heavy fluctuation in the data, present even after smoothing (section 4.2) contributed to the heavy presence of misclassification. This probably indicates that the seabed in this area is likely not soft enough for this frequency to be effective. It is therefore unlikely that any valid cluster boundaries could be obtained from direct clustering of 200 KHz dataset when the cluster number is higher than three.

The main objective of this part was to extract boundaries to provide a guideline for optimizing ground truth. Clustering of both 12 KHz and 38 KHz frequencies seems promising for this purpose (Figures 4.8-4.10). However, results from the 200 KHz dataset were not considered for further analysis due to the reasons explained above. Consequently, both the 12 and 38 KHz frequencies were accepted as working frequencies for this study and clustering results from both frequencies will be validated using ground truth data in Chapter 6.

Cluster boundary extraction from the SBES dataset

Due to the narrow geographic extent of the dataset and the close proximity of the data points, cluster boundary visualization was not possible when the whole dataset was plotted. Enhancement of points (by increasing their size for visualisation) was not effective due to overlapping. Therefore rasterisation of the point data was implemented to get the overview of cluster boundaries as explained previously. Figures 4.11 and 4.13 show the results for PCA + k-means on both SBES frequencies (12 and 38 KHz) in raster format. As mentioned above, here it is the geographic distribution of boundaries between clusters that is of interest, rather than the cluster labels (i.e. colours) of each cluster area.

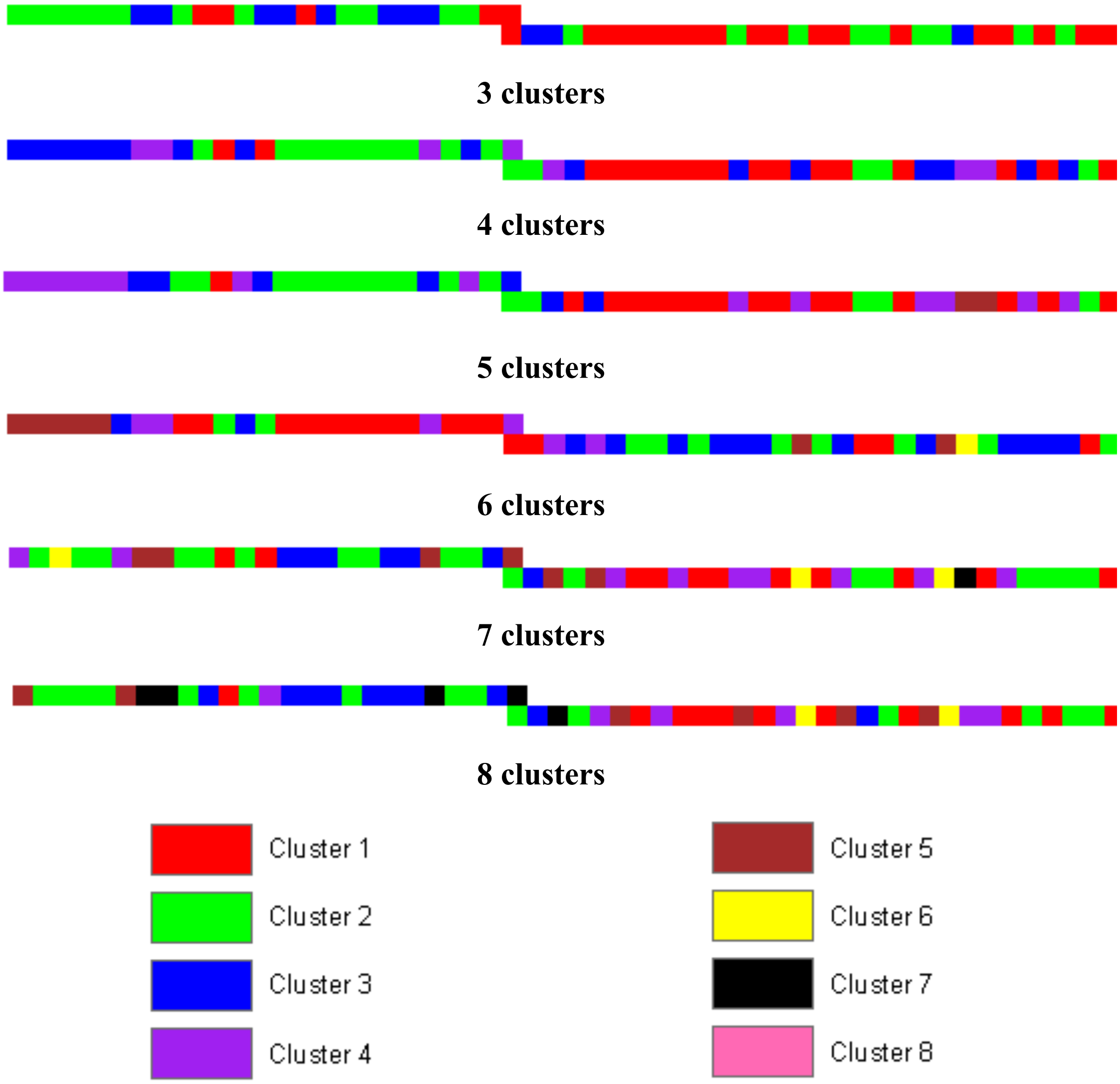
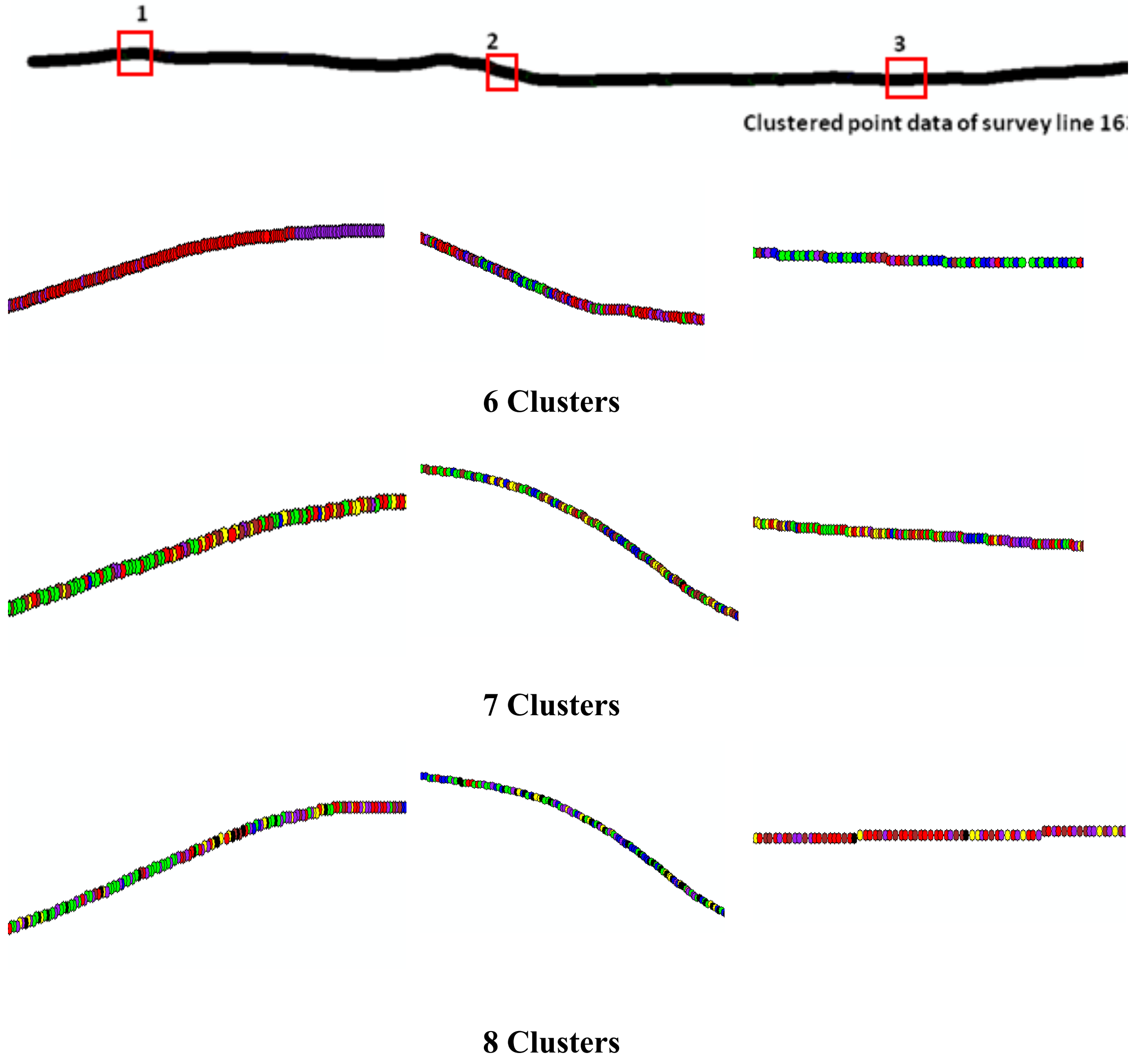


Figure 4.11: PCA + k-means cluster boundaries extracted from rasterised survey line (12 KHz)



6 Clusters

7 Clusters

8 Clusters

Figure 4.12: Misclassification in three selected locations (1, 2, and 3) for PCA + k-means with cluster numbers 6, 7, 8 (SBES 12 KHz)

The rasters obtained from the 12 KHz dataset clustering show a heavy presence of misclassification when higher numbers of clusters are used. Upon close examination of the point dataset (i.e. in locations 1, 2, and 3 in Figure 4.12) and the rasters (Figure 4.11), results from seven and eight clusters in the 12 KHz dataset failed to display clusters with well defined boundaries. Misclassification (i.e. a very mixed pattern) was present throughout the survey line and therefore it was concluded that results with both seven and eight clusters should be discarded from further analysis. Cluster result with six clusters seemed to give somewhat of a mixed picture. PCA + k-means clustering produced clusters with relatively well defined boundaries from the extreme left to the centre of the survey line (Figure 4.11), while the area from the centre to the extreme right end of survey line appeared heavily mixed and therefore misclassified. This could be due to the fact that the right portion of the survey line contains areas with mixed geological features, something that should be further investigated. However, because of the misclassification present in one half of the area, results with six clusters were also discarded from further analysis.

From figure 4.13, which represents the PCA + k-means results from SBES dataset obtained with 38 KHz, we can see that for 6 clusters and upwards the whole region is divided into very small raster areas. This indicates a high probability of misclassification. A close examination of selected areas (1, 2, and 3) is displayed in figure 4.14. There, clustered points are heavily mixed and a boundary is not identifiable. Therefore, as with 12 KHz data, cluster boundaries will not be extractable for cluster numbers greater or equal to 6. Cluster results with 3, 4, and 5 clusters however do contain clearer cluster boundaries and can be further used for subsequent cluster validation and labelling (Chapter 6).

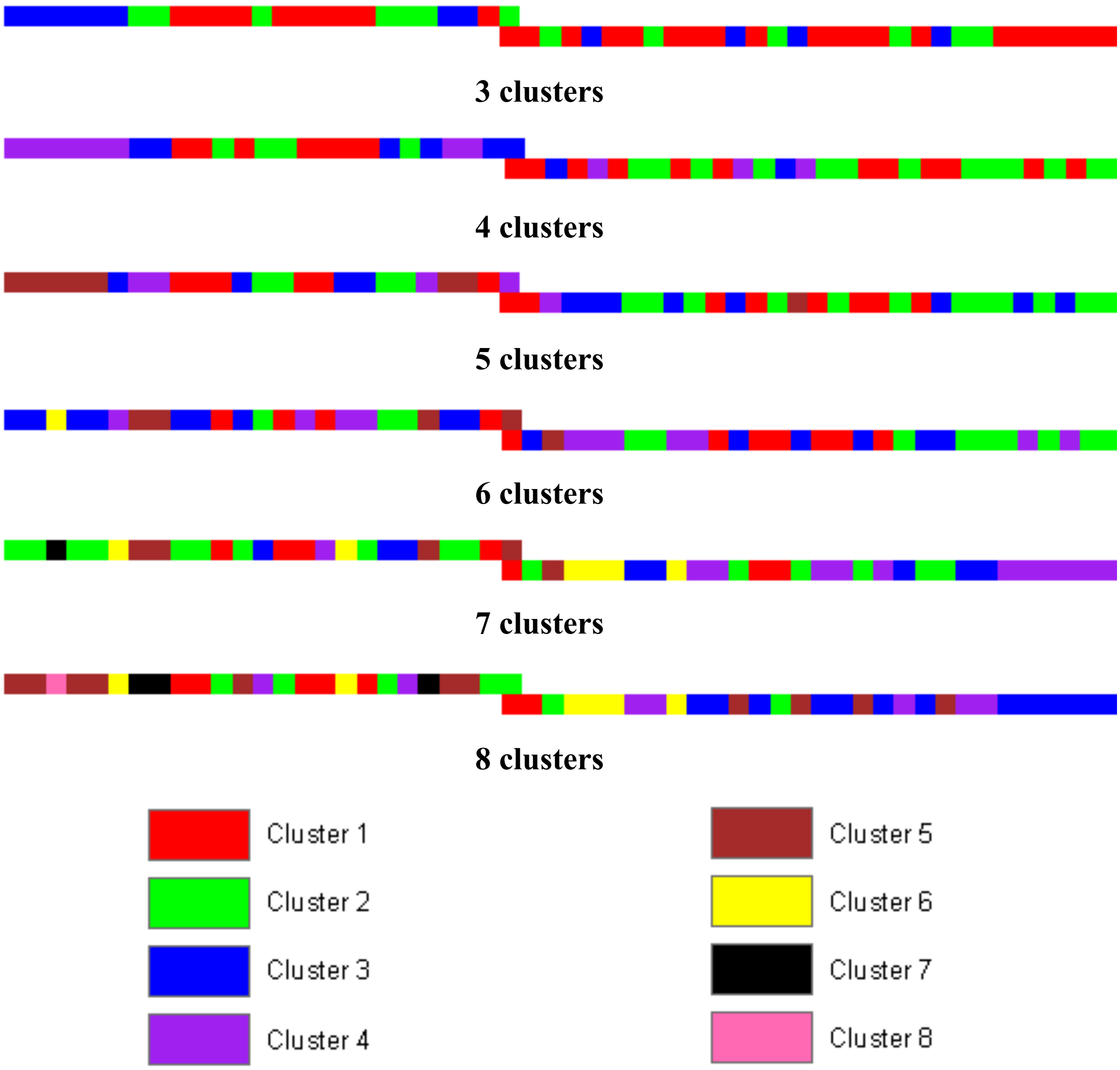
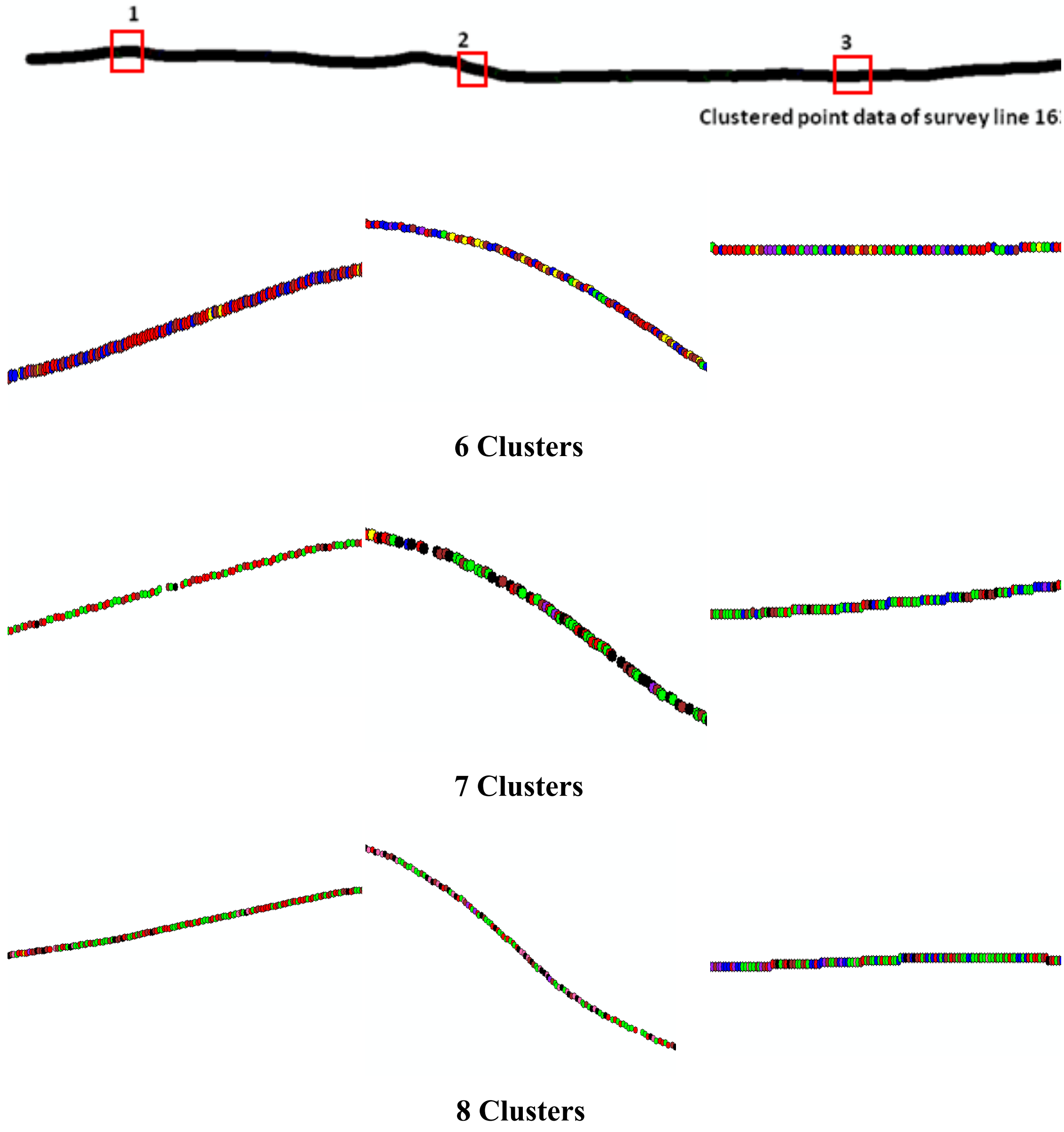


Figure 4.13: PCA + k-means cluster boundaries extracted from rasterised survey line (38 KHz)



6 Clusters

7 Clusters

8 Clusters

Figure 4.14: Misclassification in three selected locations (1, 2, and 3) for PCA + k-means results with cluster numbers 6, 7, and 8 (SBES 38 KHz)

4.6.3. Fuzzy c-means clustering results

Fuzzy c-means (Bezdek et al., 1984; Dunn, 1974), also known as soft k-means, was developed as a generalisation of k-means clustering, hence the two algorithms share a lot of similarities, including the need to specify the number of clusters a priori and the potential for getting trapped at local minima or saddle point of the objective cost function. The main differences between the two related clustering techniques are that k-means assigns a single label to each sample after completion without exception, whereas FCM returns a vector of label probabilities for each sample. In effect, FCM allows one sample the possibility of belonging to two or more clusters. This makes it particularly suitable for clustering of geological data where overlap between classes is almost always present. While Euclidean distance measure is normally used with k-means, inverse distance weighting is employed in FCM.

In FCM, each sample point has either a strong or weak association, namely membership function, to each cluster, determined by the inverse distances to centres of clusters and influenced by the degree of fuzziness. The iterative optimisation involves fuzzy partitioning by consecutively updating every membership function (U=$\{u_{ij}\}$) and sample (c_j) until the termination criteria is satisfied or the maximum number of iterations is reached. As in the case with k-means, this procedure only converges to a local minimum. To obtain the clustering labels, the membership function with the highest probability is required.

To test if the discrete fuzzy boundary was effective, a plot of the difference between the largest (U_{max}) and second-largest membership function value (U_{max-1}) for the selected fuzziness was considered (Figure 4.15 and 4.16) for all clusters. A data point is assigned to a cluster based on the U_{max} value of the membership function, while Umax-1 represents the second highest probability value for that data point to belong to another cluster. If the boundaries are discrete, then there should be a sharp change in the difference in probability values i.e the probabilities of a sample belonging to one cluster or the other should have a large difference as they can only belong to one class. The difference between the highest membership function to the next starts to decrease when the fuzziness is increased, indicating the increased probability of one point belonging to two or more clusters. With higher fuzziness, the line in the plot would start to get straighter as the difference got lower. From the figures, it can be seen that the change in difference in the membership function is quite sharp for all the clusters, indicating that the 1% fuzziness was in fact treating the samples as discrete.

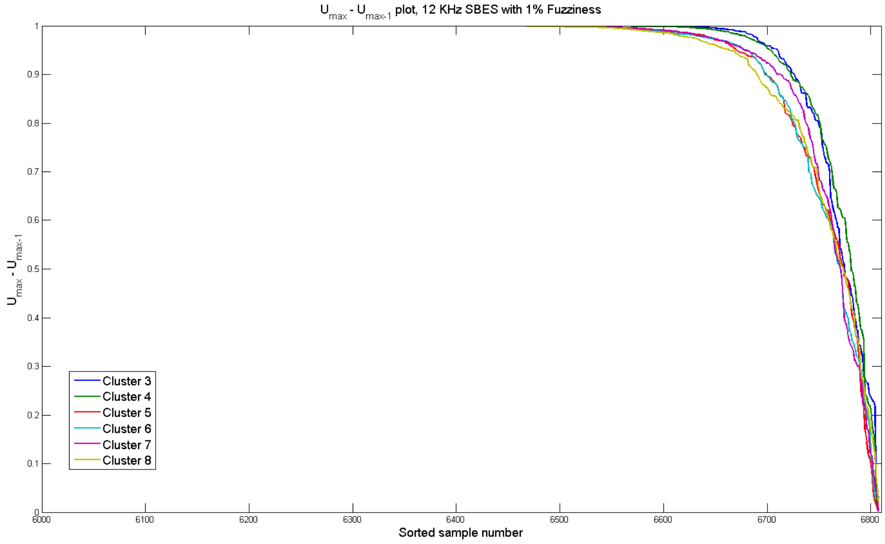


Figure 4.15: U_{max}-U_{max-1} plot for 1% fuzziness (12 KHz dataset)

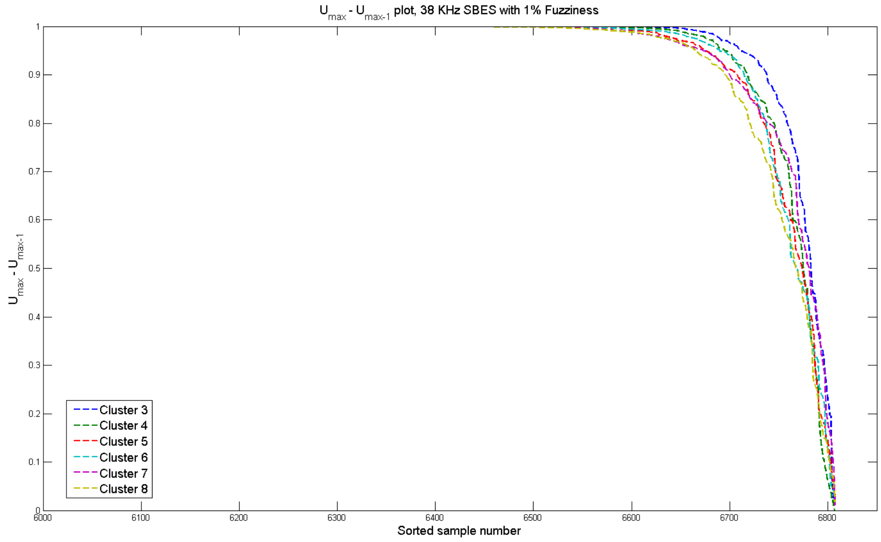


Figure 4.16: Umax-Umax-1 plot for 1% fuzziness (38 KHz dataset)

During the clustering computation, the number of clusters (p) specified a priori was varied from 3 to 8, while the degree of fuzziness (m) was set to 1% (m=1.01) for each value of p. In order to minimise the impact of hitting ‘poor’ local minima by chance, a total of 10 Monte Carlo simulations were performed for each case of p and m, and the labels with the least mean of errors were retained as the optimum results and presented. Example maps of the clusters obtained from fuzzy c-means are presented in Figures 4.17 to 4.18. Raster representation in these maps was created in the same way as for PCA + k-means.

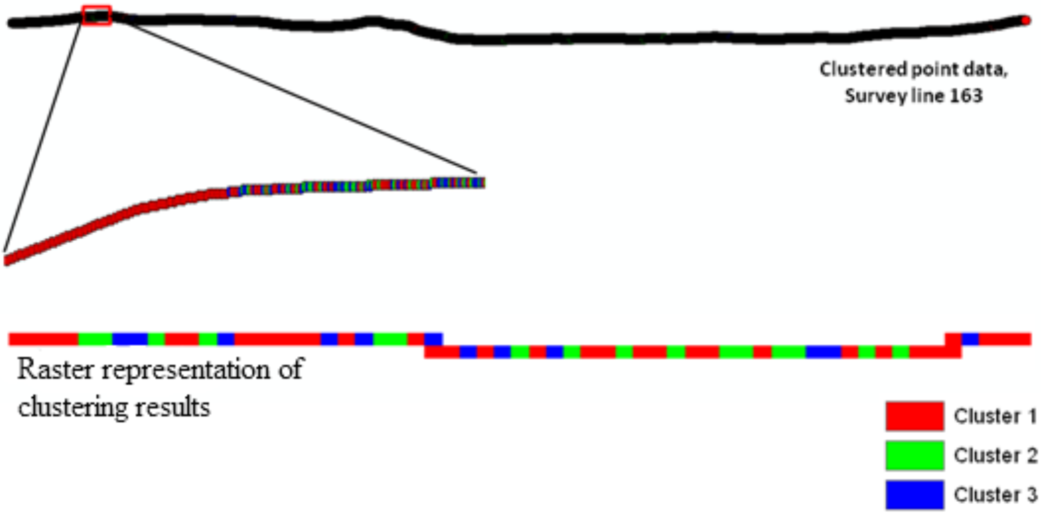


Figure 4.17: Fuzzy c-means clustering of SBES dataset (12 KHz)

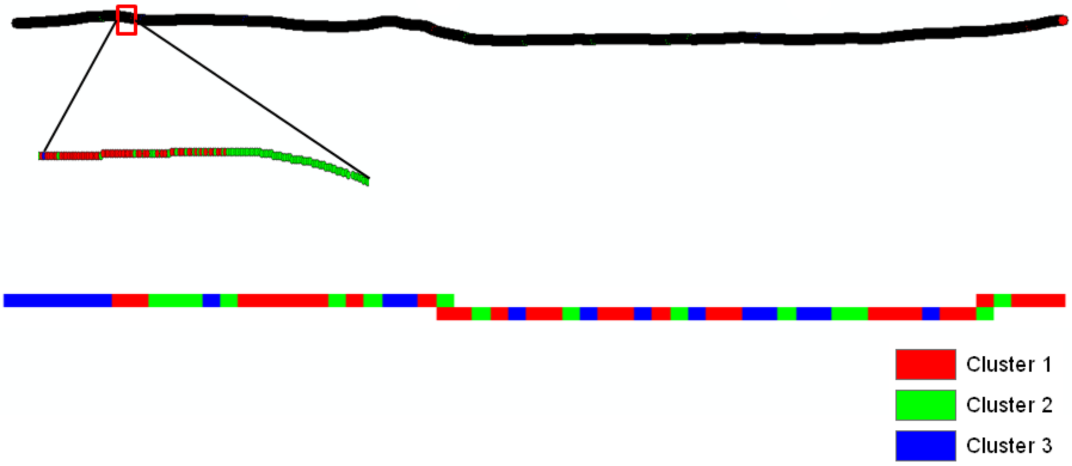


Figure 4.18: Fuzzy c-means clustering of SBES dataset (38 KHz)

As before, the results obtained from fuzzy c-means were closely examined and as the result cannot be displayed to its entirety due to the compactness of the points, a few selected points were selected and displayed in the above figures. As before the aim was to identify boundaries between cluster areas. We can see from one selected zone in Figures 4.17 and 4.18 that cluster boundaries are relatively well identifiable. This was consistent throughout the survey line. As with PCA + k-means we also compared geographical distribution of cluster boundaries for all values of ‘p’ (cluster number). Figure 4.19 and 4.20 shows the fuzzy c-means clustering results for 12 and 38 KHz datasets in raster form for p=3 to 8. As before, it is the location of the boundaries between areas that is of interest and not the specific cluster labels (colours) of each cluster area.

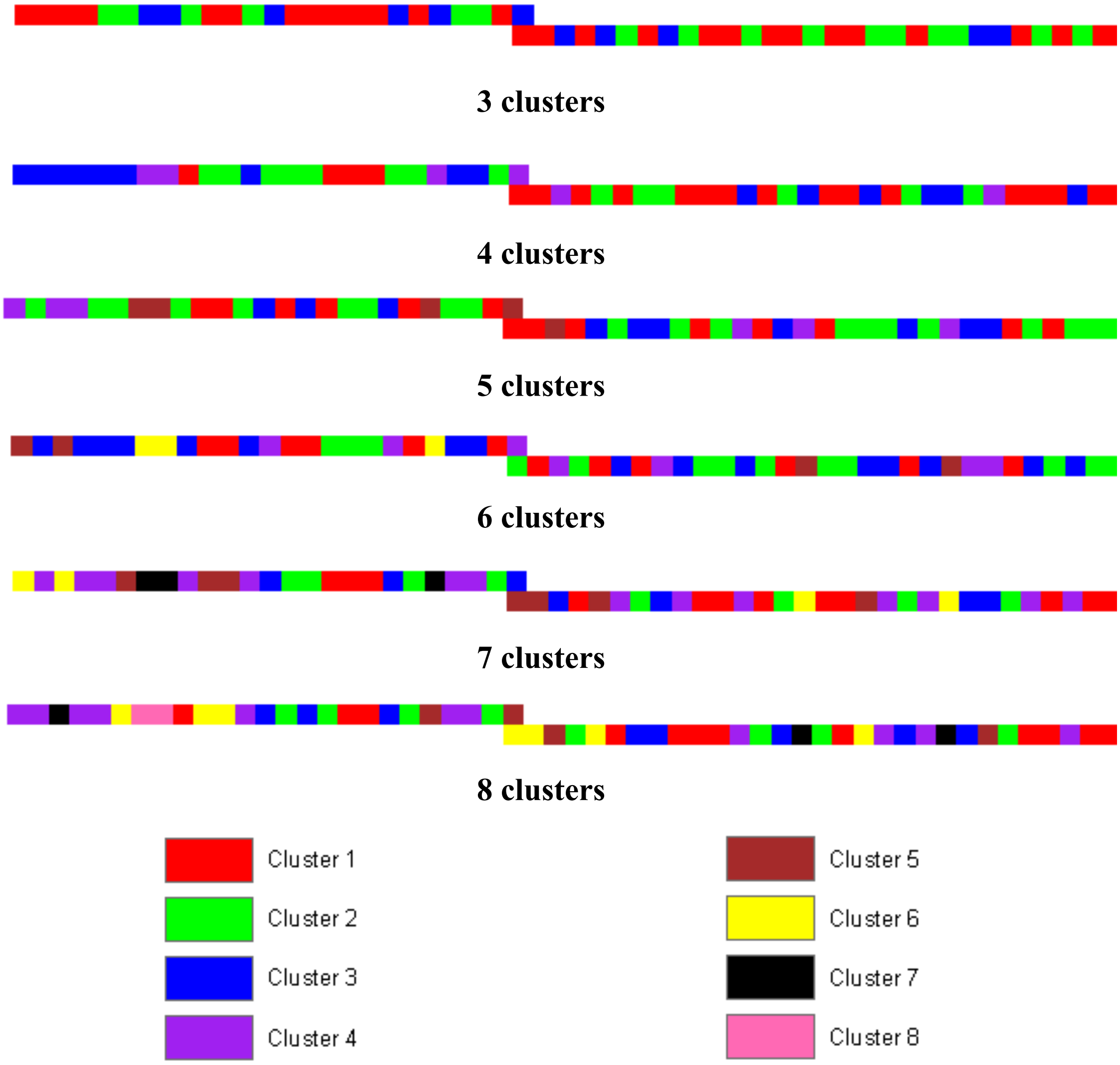


Figure 4.19: Fuzzy c-means cluster boundaries extracted from rasterised survey line (12 KHz)

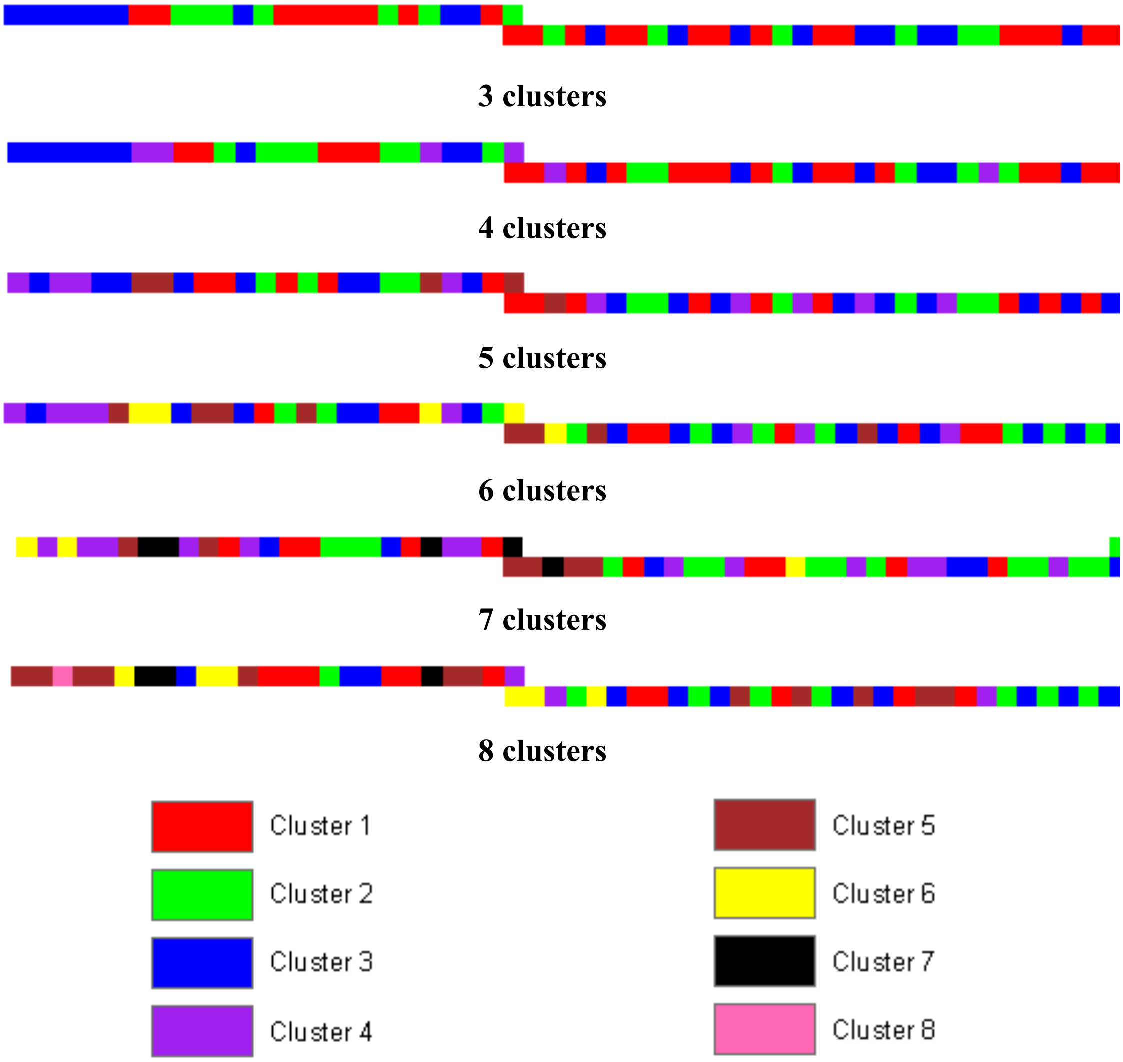


Figure 4.20: Fuzzy c-means cluster boundaries extracted from rasterised survey line (38 KHz)

From the above figures, it appears that representative boundaries only exist for cluster numbers three and four. From five clusters and onwards, the level of misclassification increases significantly as the distribution of raster boundaries appears random in nature and cluster areas in the raster become smaller with frequent alteration of labels (colours). To further investigate this, the clustered point datasets (12 and 38 KHz) were visually explored. Figures 4.21 and 4.22 show some example locations.

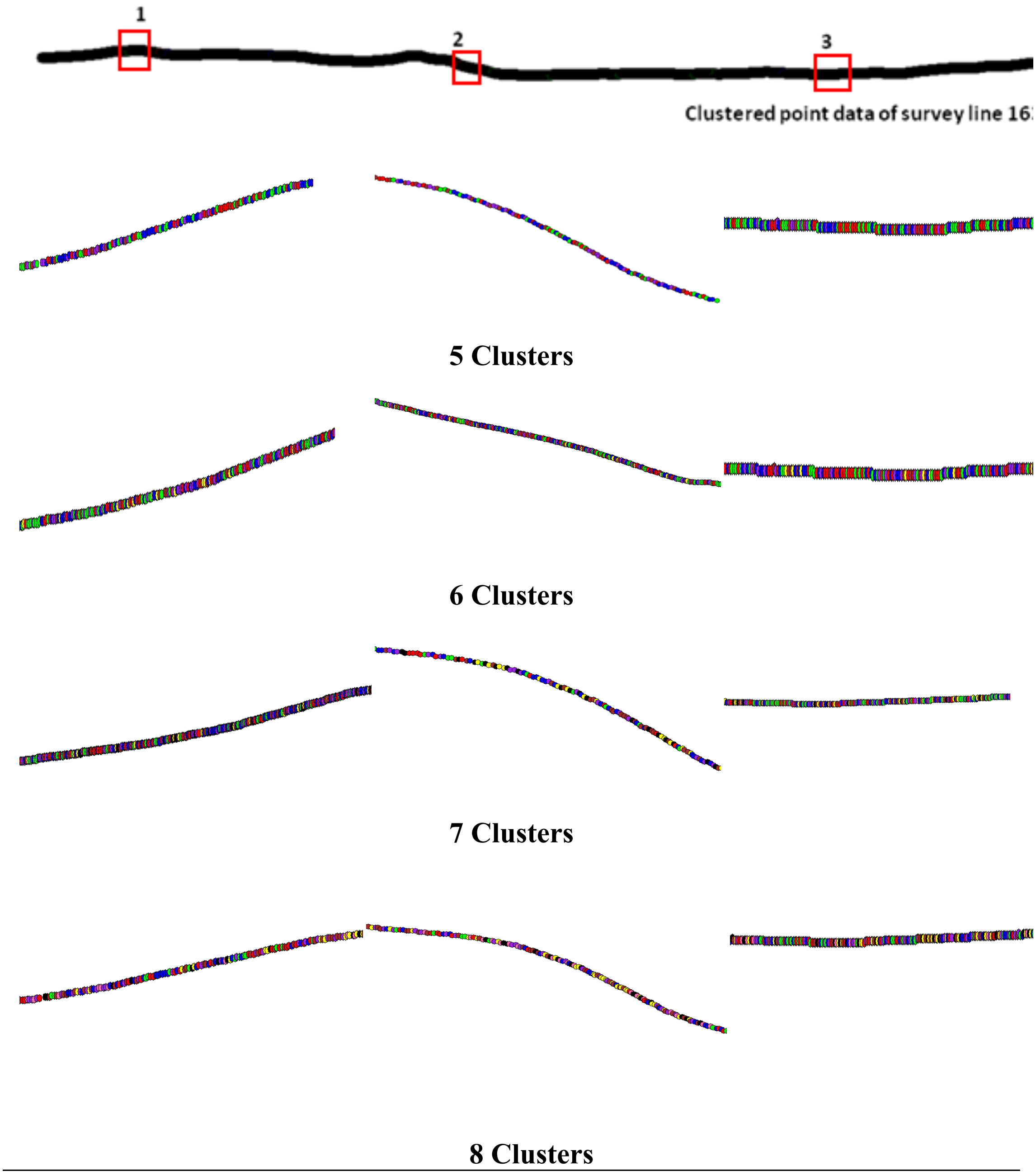
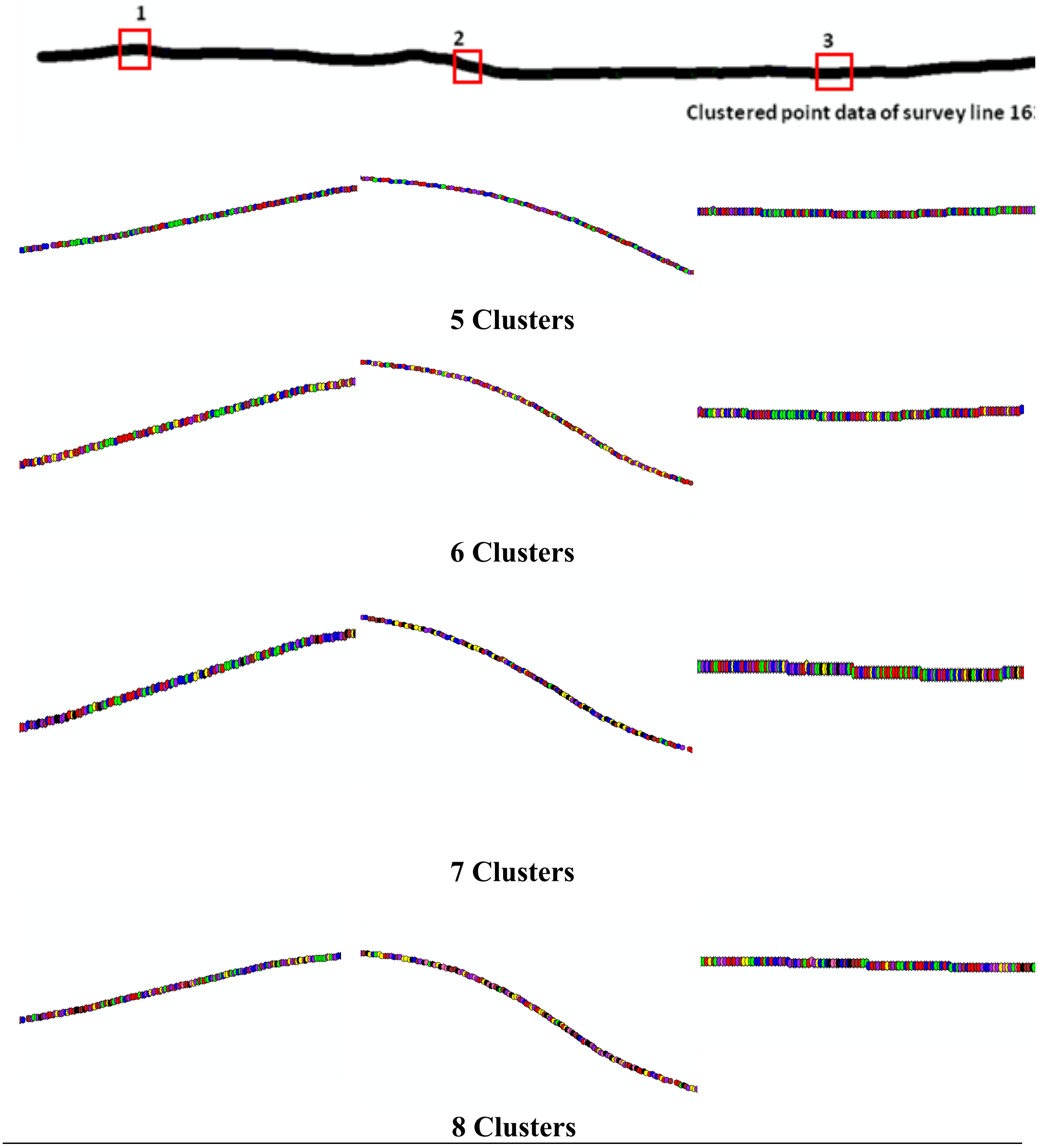


Figure 4.21: Misclassification in three selected locations (1, 2, and 3) for fuzzy c-means results with cluster numbers 5, 6, 7, 8 (SBES 12 KHz)



5 Clusters

6 Clusters

7 Clusters

8 Clusters

Figure 4.22: Misclassification in three selected locations (1, 2, and 3) for fuzzy c-means results with cluster numbers 5, 6, 7, 8 (SBES 38 KHz)

Consequently, for both frequencies (12 KHz and 38 KHz), fuzzy c-means clustering results seem to contain a high level of misclassification when the cluster number is greater or equal to five. In comparison to PCA + k-means, fuzzy c-means appears to be more prone to misclassification. Therefore, based on this finding, clusters 5, 6, 7, and 8 will not be included in further analysis and validation in Chapter 6.

4.6.4. Internal cluster validation

For this experiment, we selected four commonly used indices for cluster quality assessment and optimal cluster number estimation: Calinski-Harabasz or Variance Ratio Criterion (VRC) index (Calinski & Harabasz, 1974; Everitt et al., 2011), Davies-Bouldin index (Davies & Bouldin, 1979; Jain & Dubes, 1988), Dunn's index (Dunn, 1974; Halkidi et al., 2001) and Silhouette index (Everitt et al., 2011; Halkidi et al., 2001). The working principal of the indices are described in Chapter 2. We calculated these four validity indices were extracted from the cluster results obtained from SBES echo returns (12 and 38 KHz) for PCA + k-means and fuzzy c-means clustering and for cluster numbers 3 to 8. Tables 4.1-4.4 lists the results of these indices.

Table 4.1: Values of validity Indices for k-means clustering from SBES dataset (12 KHz)

	3 Clusters	4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Calinski-Harabasz index	2037.73	1832.52	1667.37	1500.77	1375.42	1354.30
Davies-Bouldin index	1.42	1.69	1.80	1.59	1.75	1.59
Dunn index	1.09	0.63	0.54	0.78	0.56	0.67
Silhouette index	0.14	0.13	0.13	0.10	0.09	0.08

Table 4.2: Values of validity Indices for k-means clustering from SBES dataset (38 KHz)

	3 Clusters	4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Calinski-Harabasz index	2952.61	2270.06	1879.08	1676.46	1676.00	1589.05
Davies-Bouldin index	1.11	1.28	1.44	1.39	1.50	1.57
Dunn index	1.37	1.14	0.83	0.89	0.84	0.75
Silhouette index	0.21	0.15	0.12	0.11	0.12	0.11

Table 4.3: Values of validity Indices for fuzzy c-means clustering from SBES dataset (12 KHz)

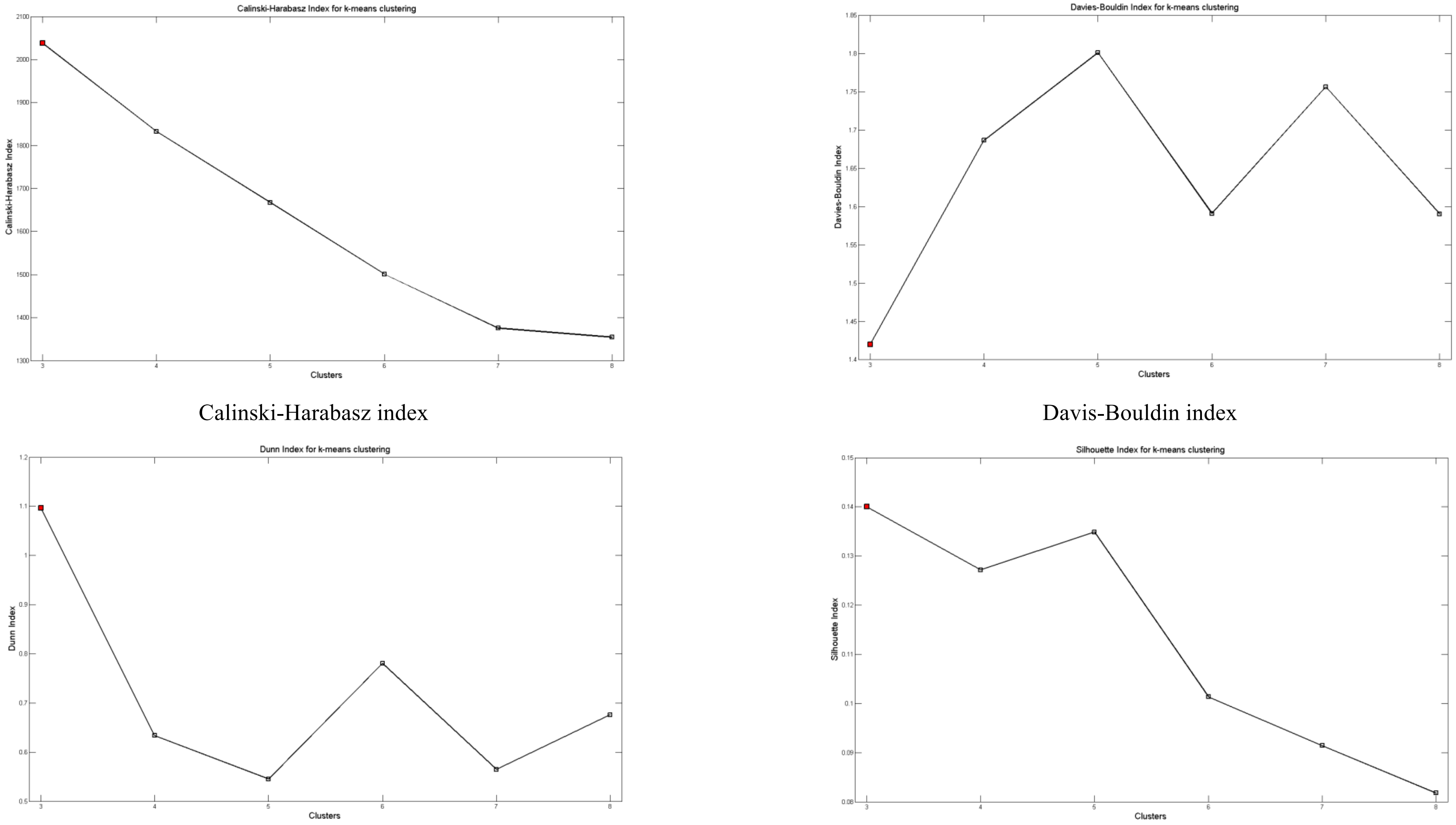
	3 Clusters	4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Calinski-Harabasz index	3995.89	3456.04	3024.01	2670.44	2457.06	2246.47
Davies-Bouldin index	0.973	1.08	1.18	1.32	1.23	1.30
Dunn index	1.56	1.30	1.14	1.01	1.00	0.98
Silhouette index	0.27	0.22	0.19	0.16	0.17	0.16

Table 4.4: Values of validity Indices for fuzzy c-means clustering from SBES dataset (38 KHz)

	3 Clusters	4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Calinski-Harabasz index	3789.63	3265.20	2833.96	2523.17	2319.61	2124.66
Davies-Bouldin index	1.00	1.11	1.24	1.30	1.26	1.31
Dunn index	1.50	1.24	1.08	0.92	0.95	0.95
Silhouette index	0.26	0.22	0.18	0.18	0.17	0.15

Figures 4.23 to 4.26 show these results as graphs. The aim of this is to identify the optimal number of clusters, i.e. at which number does each algorithm provide the best separation of data into the most compact clusters. As explained in Chapter 2, this happens when Calinski-Harabasz, Dunn, and Silhouette indices reach their maximums and Davies-Bouldin index reaches the minimum.

In our case, all four validity indices indicate that for both frequencies and clustering algorithms, three is the number of clusters where the best between-cluster separation and within-cluster compactness are reached. This is less than the optimal number of clusters identified in MBES clustering (see Chapter 3). However, this is expected as SBES contains subsurface information and if the seabed surface has a heavy presence of shells, burrows or plants, the subsurface information can actually provide a more accurate account of what lies beneath those depositions. This is likely the case with this particular dataset, since the study area (Malin Head) is known to have a heavy deposition of shells. Thus, it is acceptable that the number of clusters for SBES clustering can be less than MBES clustering. This finding will be further analysed in Chapter 6 where we evaluate both MBES and SBES results versus a ground truth database.



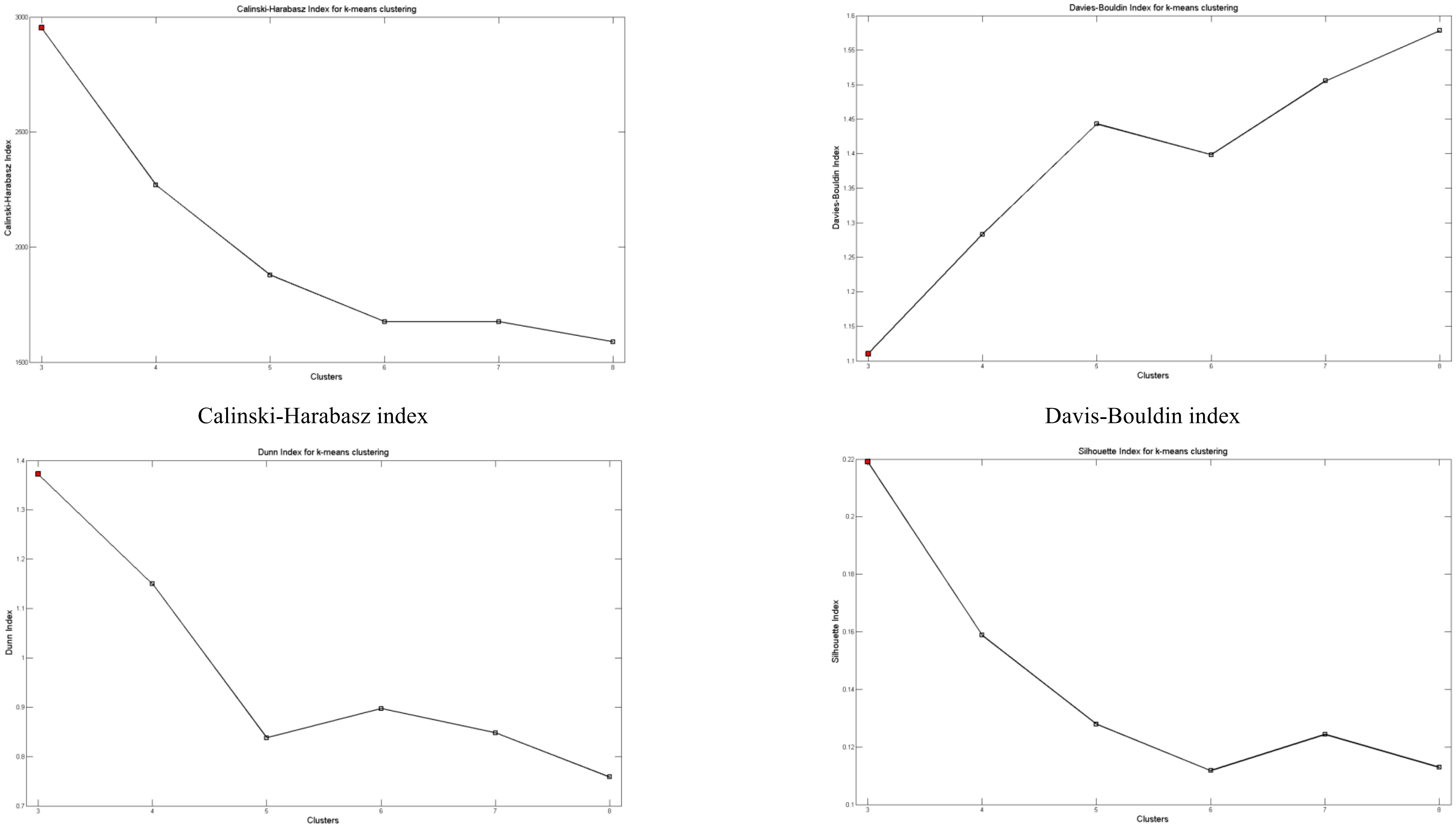
Calinski-Harabasz index

Davis-Bouldin index

Dunn's index

Silhouette index

Figure 4.23: Validity indices for PCA + k-means clustering from SBES backscatter dataset (12 KHz)



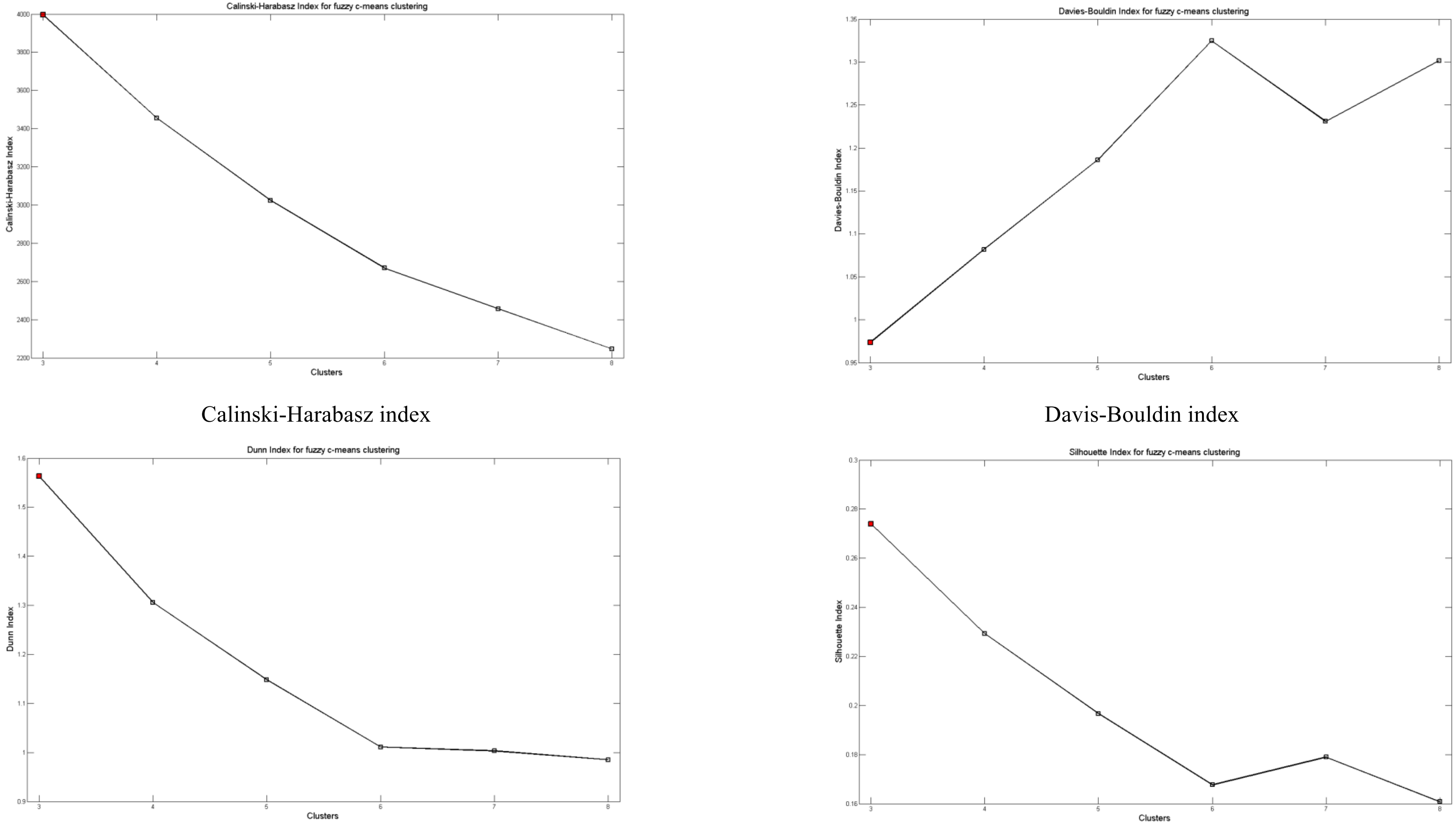
Calinski-Harabasz index

Davis-Bouldin index

Dunn's index

Silhouette index

Figure 4.24: Validity indices for PCA + k-means clustering from SBES backscatter dataset (38 KHz)



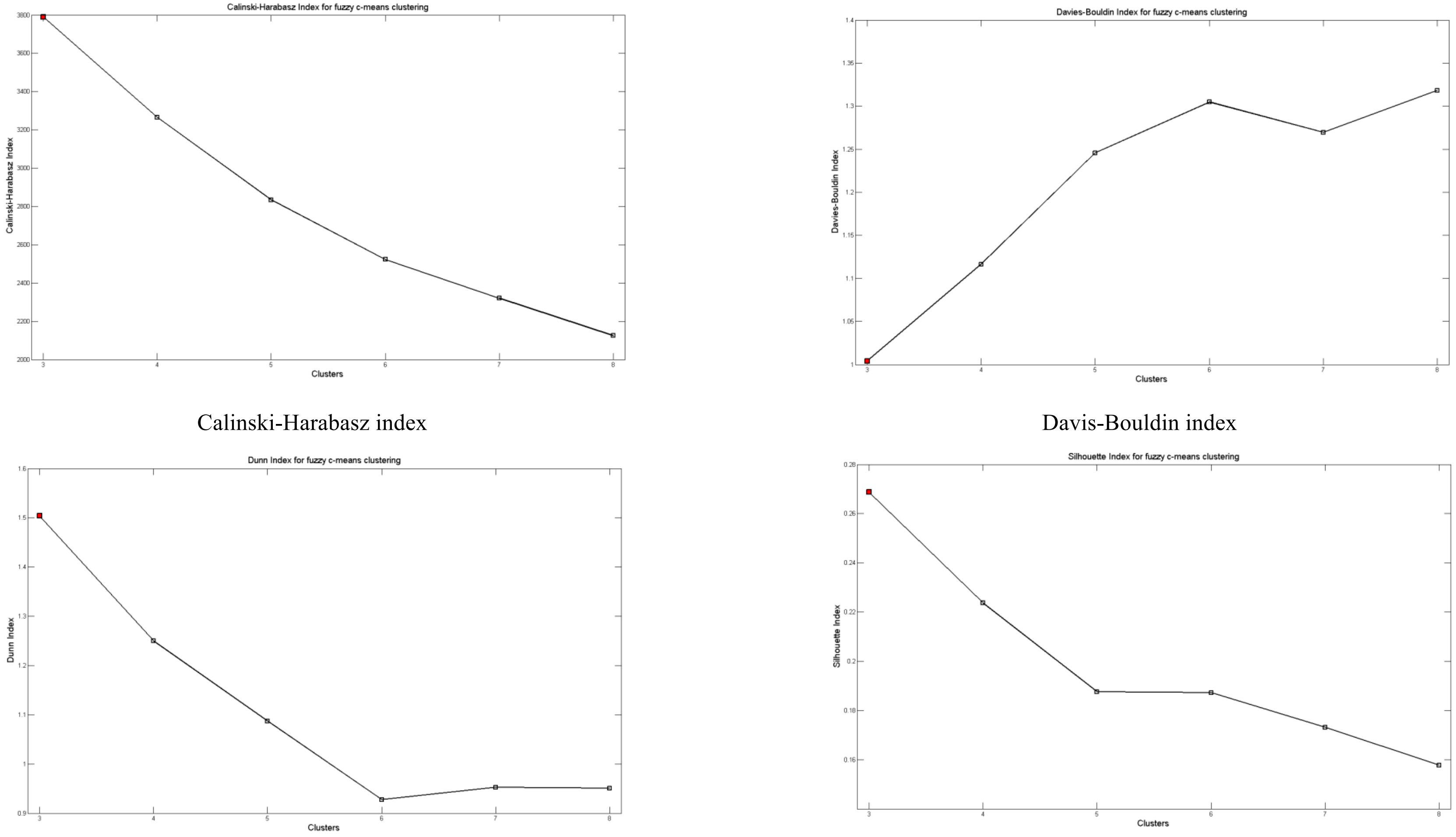
Calinski-Harabasz index

Davis-Bouldin index

Dunn’s index

Silhouette index

Figure 4.25: Validity indices for fuzzy c-means clustering from SBES backscatter dataset (12 KHz)



Calinski-Harabasz index

Davis-Bouldin index

Dunn's index

Silhouette index

Figure 4.26: Validity indices for fuzzy c-means clustering from SBES backscatter dataset (38 KHz)

4.7. Discussion

The recent decade brought a rapid development in SBES echosounders. Their resolution has improved to the point where they can produce very accurate information of seabed and seabed subsurface. They are normally fitted to the survey vessel in conjunction with MBES echosounders and thus each MBES survey run also yields high resolution SBES data for the same location. However, in most cases only MBES data are used for clustering and subsequent classification of seabed type. The potential of SBES to provide high resolution subsurface information can be of interest as the resulting clusters could provide a more accurate description of the seabed than that of MBES.

SBES data in general contain a higher degree of noise compared to MBES data. This is due to the high resolution, sensitivity to suspended particles in water, presence of schools of swimming fish and other factors. Therefore, the inherent systematic noise in SBES data needs to be carefully explored and processed by experienced geologists and surveyors. Good quality raw SBES backscatter data are fundamental to any classification, especially when data are prone to the inevitable noise. This was the first drawback of the dataset that was available for this study and the reason for the design of our experiment: first using visual exploration for outlier detection, then comparing the clustering methods and finally validating cluster results using validity indices.

The SBES dataset available for this study also contained a significant vertical displacement. This means that the optimal depth detected from the data fluctuated by ±5m (approximately 3m above seabed and 2m below seabed from the derived average depth over the survey line). This fact probably had a direct effect on the clustering as the algorithms would not only have to consider the shape of the echo envelope but also the location of the peak (above, below or on seabed). This can lead to misclassification as peaks similar in shape can be placed at a different location within two time series due to this fluctuation and thus such samples will have a different cluster label assigned to them instead of being labelled with the same cluster label. After discussion with the collaborators from the Geological Survey of Ireland (GSI), this error was contributed to surveyor’s failure to optimally calibrate the echosounder configuration for seabed depth and thus is regarded as a gross error.

This is considered as one of the contributing factors of misclassification in the results.

Another drawback this study faced is the fact that only one dataset was available for clustering. This meant that area that was available for clustering was long and very narrow as SBES emits only one vertical beam producing a narrow high resolution backscatter per ping. A higher number of datasets would have enabled us to explore and analyse the overlap or mixture of seabed classes more effectively and expand cluster boundary detection in the direction perpendicular to that of the survey line. This can be another of the reasons why the clustering results had a heavy presence of misclassification.

Visual exploration of the dataset after it was ‘cleaned’ using a Butterworth low-pass filter and an automatic algorithm to detect ‘bad’ samples yielded promising results. A total of 81 records were found in the visual exploration which did not have any peaks and were thus deemed outliers, but were not detected in the automatic procedure. The absence of peaks means that at any point those echoes did not hit the seabed and therefore did not contain any information on the seabed. These echoes could be discarded without any further testing. The reason why the automatic algorithm failed to detect these 81 ‘bad’ samples requires further investigation and is recommended for future studies. The fact that a visual exploration tool was successfully used to incorporate human cognitive ability in detecting additional outliers is an indication that this approach can be used as an added step for data processing of acoustic backscatters obtained from SBES. This can improve data quality as each survey schedule usually involves hundreds of runs resulting in a large quantity of SBES data.

Two clustering approaches were tested on the dataset- the *de facto* PCA + k-means and fuzzy c-means. The application of fuzzy c-means was somewhat limited as different combinations of fuzziness could not be tested due to the narrowness of the dataset and the fact that only one dataset was available thus making the study area for this study quite small for different fuzziness to be effectively tested. The fuzziness was set to 1% so that the algorithm treats the cluster boundaries as discrete. The visualisation of clustering results was challenging as well. As the survey line was narrow, the clustered point overlapped and as a result could not be visualised clearly. Therefore, a different approach had to be taken for visualisation. The cluster areas were extracted by rasterising the data with working resolution set to 3 meters. This enabled us to visualise the boundaries between areas. The cluster areas were coloured based on the summary of the clustered point present within that boundary – the area was coloured as per the dominant cluster points. The drawback of rasterisation was that it took away the sense of boundary mixture and gave a very discrete and somewhat 'unrealistic' (Figures 4.11, 4.13, 4.20, 4.21) cluster boundary picture. However, this does not affect further analysis as rasterised boundaries were generated only to provide a visual overview of the clusters present and the raster images were not considered in further analysis.

Both PCA + k-means and fuzzy c-means algorithms performed similarly- with identifiable boundary definitions when the number of clusters was low and showing a high level of misclassification at higher cluster numbers. From five clusters and upwards, both algorithms failed to provide any representative clusters. This could be due to the fact that the survey area does not contain many discrete classes. Subsurface information contained in the dataset may also have contributed to higher degree of misclassification. The study area is known to be more or less sandy with a heavy deposition of shells. The subsurface information can therefore enhance the relatively homogeneous nature of the subsurface layer and thus reduce the number of discrete clusters in the dataset. This may be the reason why both algorithms were effective when the cluster numbers were low, but not for higher cluster numbers.

The results were validated using four different internal validation measures- Calinski-Harabasz (VRC) index, Davis-Bouldin index, Dunn's index, and Silhouette index. In all four occasions the optimal numbers of clusters was suggested to be three - the lowest number of classes in the testing. Internal validation indices provide an estimation of the inherent number of clusters in a dataset based on compactness and separation in the absence of ground truth data. This however serves only as a guideline and further tests and validation based on ground truth data are required to establish the actual number of classes. Therefore, all three result sets (for 3, 4, and 5 cluster numbers respectively) obtained for two frequencies (12 KHz and 38 KHz) were kept for further validation against ground truth data (Chapter 6).

Results from this chapter show that direct clustering of SBES backscatter dataset can, to some extent, provide information on inherent seabed type clusters. This could help in the process of how ground truthing is designed. The traditional approach to ground truthing is to do a feature based clustering of MBES backscatter data to give the surveyor an idea of the clusters present in a survey area so that s/he can determine optimal locations for ground truth collection points. However, feature based clustering of MBES data is a very time consuming procedure as well as being reliant on domain knowledge of the surveyor. Without the presence of any ground truth data, the selection of features to be extracted from MBES almost always relies on the experience of geologists involved in the survey. The selection of features is then redefined once enough ground truth samples are collected and the geologist has some idea of the nature of the seabed to be surveyed. An effective direct clustering of SBES could however provide an immediate approximation of clusters and eliminate the need of feature extraction and sampling, which would result in saving time for both feature computation and actual acquisition of ground truth samples. In this regard the results from direct clustering of SBES echoes are promising.

If direct SBES clustering could provide a representative classification when data are large enough to cover a good area for the algorithms to work effectively i.e. when more than one survey line is taken into consideration, then it could potentially be integrated into the actual seabed mapping process instead of just for estimation of optimal ground truthing locations. Tests are required to confirm this, but if successful, then the need to extract features at any point in SBES classification could be eliminated. This could lead to significant time and cost saving as feature extraction from a whole SBES database (usually terabytes of data) requires weeks of intensive computing.

In the next chapter we attempt to join seabed surface information from MBES data with subsurface information provided by SBES. The main objective is to test if a combination of SBES and MBES datasets leads to an improved classification of seabed type.

Chapter 5

Data fusion: Combining SBES with MBES

This chapter evaluates the potential of improving seabed type mapping quality by combining SBES features with that of MBES.

Chapter contents

The research focus in this chapter is to evaluate clustering and subsequent mapping of the seabed using a new approach by combining the MBES and SBES datasets with the aim of improving seabed mapping quality obtained otherwise using MBES data only.

5.1. Research background and justification

MBES has been the most commonly used echosounder for seabed mapping since its arrival in the civilian domain in the 1970s (Mayer, 2006). The emitted acoustic beams in an MBES are fan shaped and the returns include a large number of angular backscatter responses (see Chapter 3). The drawback of having a large amount of angular backscatter is that it introduces noise to the datasets. Ideally, the preferred backscatter would be the one that is returned from the beam that was vertical as it would contain the least amount of distortion. However, this would mean that for a large area, numerous survey runs would be required which would not be economically viable. As a result, despite the noise, angular beams are included (hence the name multibeam).

The limiting factor of using only one scanner type (MBES) for seabed classification is that it cannot provide reliable classes on all occasions. Some seabed types are more easily discriminated by one sensor than the other. Different seabed types that have a heavy deposition of lose objects of similar size (for example: shell and fine gravel) can give similar backscatter responses due to MBES's inability to penetrate beyond a few centimetres. It is therefore an improvement to use data from more than one sensor type for the same purpose. SBES is a good candidate for fusion with MBES due to its unique ability to penetrate seabed surface for several meters and therefore can bring additional information that can be potentially useful for better discrimination between seabed classes.

In the current surveys, both MBES and SBES are fitted to the hull of the survey vessel recording echo returns from both the sensors simultaneously. This results in two separate acoustic databases – one for the MBES and one for the SBES in a single survey run (Figure 5.1). In addition to combined collection of data, SBES backscatter is the only backscatter collected in the survey that has high resolution vertical echo returns and therefore has the potential to contribute to the classification process.

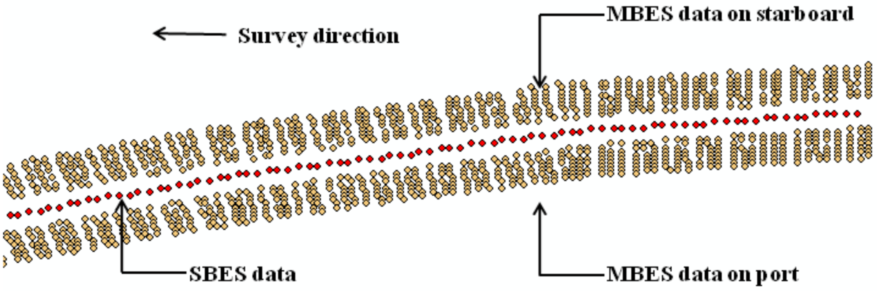


Figure 5.1: MBES and SBES data collected from a single survey run

This forms the central research objective of this chapter. The goal is to use SBES's low distortion returns in conjunction with MBES to better define the class boundaries. There are only a handful of studies that focus on fusion techniques in acoustic classification. Kerneis & Zerr (2005) combined a DEM image from SBES with images from Side Scan Sonar (SSS) for improvement of classification results. Motao et al. (2002) used data fusion techniques within MBES data and between MBES and SBES data to improve the swath accuracy.

This research project is aimed at incorporating both MBES and SBES in seabed classification at the same time. A feature based classification approach is adopted for this study in an effort to reduce computational intensity. The MBES statistical features will include features that resulted from dimensionality reduction using SOM (see Chapter 3), while for SBES a set of statistical features will be generated from the time series backscatter dataset. The features generated from SBES will be briefly discussed later in this chapter.

5.2. Study area

Data were collected by Geological Survey Ireland (GSI) and the Marine Institute (MI) as part of the INFOMAR programme during a series of Malin Sea surveys in the year 2003. The main equipment used were the Kongsberg™ EA600 singlebeam echosounder (Kongsberg & Ea, n.d.) and the Kongsberg™ Simrad EM1002 multibeam echosounder (Simrad-Kongsberg, 1999). The details of the data frequency are outlined in Chapters 3 and 4. Malin Sea is located to the north of the Republic of Ireland and has been chosen as a case study due to its known seabed type variations. The survey acquired both MBES and SBES returns in the same survey runs simultaneously. Chapter 3 analysed the data acquired from MBES while the SBES dataset was analysed in Chapter 4. As SBES echo returns from only one survey line, line 163, are available, the MBES data from the same survey line are used throughout the project for analysis. The track runs East-West at latitude of 56N and covers a survey distance of about 110 km (Chapter 4 pp. To be updated on final version).

5.3. Producing the feature dataset from SBES backscatter data

The MBES data come with 132 statistical features or FFVs and these FFVs were further reduced using a visual analytics technique – a Self Organising Map (SOM). The final result was an optimized reduced MBES dataset that contained 35 FFVs (Chapter 3). The SBES dataset consisted of raw backscatter time series. To combine the SBES dataset with that of MBES, they needed to be of similar type. Therefore, statistical features were generated from the SBES dataset. Various statistical techniques can be applied to SBES time series backscatter data for feature generation. For the SBES time series dataset, a total of four statistical features were used to capture the quantitative echo characteristics. These are described below.

Mean and standard deviation

The first two statistical features that were generated are mean and standard deviation. The temporal mean ($\bar{x}$) and standard deviation (s) are regarded as one of the most important statistical measures in acoustic feature extraction (Lurton, 2002). The mean or average echo return value reflects the influence of different types of sediments on echo time series. In geology, it is related to the impedance contrast of the seafloor. Standard deviation of sonar time series also contains information that has direct relationship with the seabed geology and tends to relate to the seabed roughness. The differences in roughness help to define cluster boundaries and thus make standard deviation an important feature in seabed classification.

Measure of randomness

While being regarded as useful features, the applicability of mean and standard deviation is generally restricted to the statistical properties of one single echo time series. A separate feature is required to convey information about the relationship between adjacent time series. To address this, a measure of randomness ($r$and) is proposed using the standard deviation of the autocorrelation of the echo difference data. This measure of randomness (rand) is defined as (Hung et al., 2010):

$$rand(X) = \left\{ s\left(\frac{a_i}{||a_i||_\infty} \right) \right\} \tag{5.1}$$

Where, $$a_i = \{u_{1-m}, \ldots, u_{-1}, u_1, \ldots, u_{m-1}\} \tag{5.2}$$

$$u_{1-m} = \sum_{\lambda=1-m}^{m-1} d_i d_{i-\lambda} \tag{5.3}$$

$$d_i = \frac{X_i}{s(X_i)} - \frac{X_{i-1}}{s(X_{i-1})} \tag{5.4}$$

Here X_i is the extracted segment from original echo time series for a specific depth (in this case 5m above and below the seabed), m is the number of spatial data samples (i.e. number of time series collected) and i is the position of temporal sample ($0 < i \leq n$) with n being the total number of temporal samples in each time series. d_i is the row vector of the difference of two adjacent echo time series, σ is the standard deviation, λ is the time series delay during convolution u, and a_i is the 'modified' autocorrelation row vector of d. To ensure that the autocorrelation is unaffected by the magnitude of the data, each correlation sequence is normalised by the largest value in the sequence as returned by the infinity norm. It has been found that the randomness measure will be more representative if the largest value of a conventional version of a_i is removed before calculating the standard deviations of the autocorrelation sequence. In theory, a value of unity can only be found at the no lag point (u_0) in the autocorrelation unless the sequence is perfectly autocorrelated to itself. The randomness measure is applied to adjacent samples only.

Measure of correlation noise

The measure of correlation noise is a measure of relationship between neighbouring echo time series. It is a measure based on the signal-to-noise ratio of the mean correlation coefficient noise and is denoted by c (Hung et al., 2010):

$$c(X) = log\left|\frac{\bar{r}_{10}(\tilde{X}_i) - \bar{r}_{10}(X_i)}{\bar{r}_{10}(X_i)}\right| \tag{5.5}$$

Where, $$\bar{r}_{10}(X_i) = \tfrac{1}{10}\textstyle\sum_{j=1}^{10} r(X_i\,, X_{i+j}) \tag{5.6}$$

Here r is the correlation coefficient between two time series and $\bar{r}_{10}$ is the average correlation from 10 adjacent time series records. $\bar{X}_i$ is same as X_i except that it is extracted for the same segment from the echo time series. Correlation noise is defined as the difference between the r from the original and temporally filtered version of X_i. The measure of correlation noise quantifies the amount of differences between consecutive samples based on correlation. It is expected to convey more information about the data in a spatial context.

5.4. Experiment design

This study will test the potential of clustering a combined MBES and SBES dataset. The main objective of this research is to test a novel technique for classification of seabed type using datasets that were optimised using visual analytics techniques, which involves combining SBES with MBES and subsequent clustering of this combined dataset.

Once the dataset is prepared, it will be clustered using two different clustering methods: PCA+ k-means and fuzzy c-means. A ground truth database will then be developed from the grab samples obtained from the study area. Three different ground truth databases are to be developed- for MBES, SBES, and combined MBES-SBES. Validation of the clustering results against these ground truth databases is discussed in Chapter 6. Figure 5.2 shows the overall process involved in this chapter.

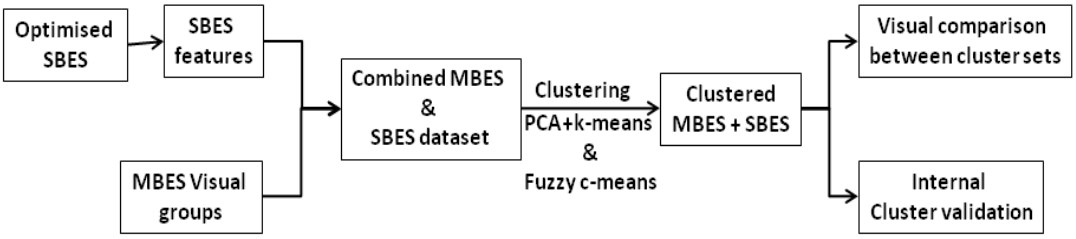

Optimised SBES
SBES features
MBES Visual groups
Combined MBES & SBES dataset
Clustering
PCA+k-means & Fuzzy c-means
Clustered MBES + SBES
Visual comparison between cluster sets
Internal Cluster validation

Figure 5.2: Overview of experiment process

Cluster results will be visually compared with the results obtained from MBES clustering. The performance of the algorithms in terms of boundary definitions will also be subject to visual comparison. Lastly, four internal cluster validation indices will be calculated for estimation of the optimal number of classes, as previously with separate MBES and SBES cluster results (Chapters 3 & 4).

5.5. Data processing

Before the extraction of SBES features, the optimal seabed depth was calculated from the SBES backscatter dataset. Once the depth was determined, 5 meters above and below that depth (estimated location of the seabed surface) were used to segment the dataset so that it would represent echo returns from the seabed surface as well as some the subsurface (see Chapter 4). Figure 5.3 shows the process of combining SBES with MBES dataset.

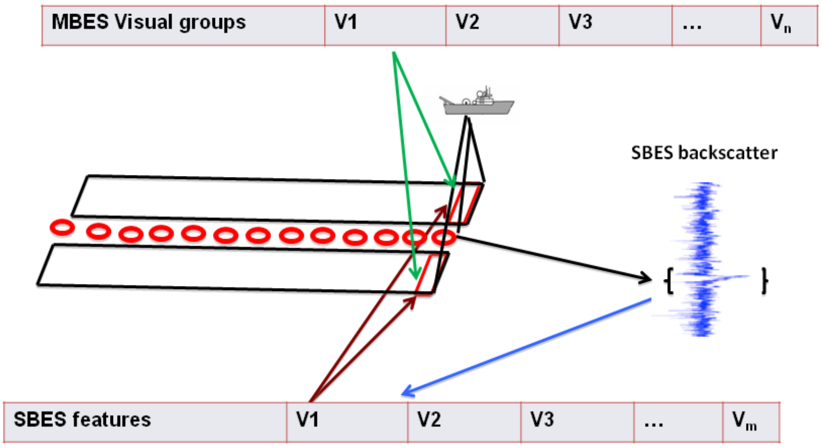


Figure 5.3: Combining MBES and SBES features

Once the data from the selected segment (5m-5m above and below seabed) is isolated, a number of statistical features are generated (described in section 5.3). In the next step, the statistical features (mean, standard deviation, randomness, and correlation noise) were combined with the MBES data.

5.5.1. Combining MBES and SBES

Once the SBES feature dataset was created, the nearest SBES data location to each MBES was determined. This had to be done as the number of SBES echo returns available was far more than that of MBES sweeps (though in each sweep, MBES had more returns (Figure 5.1) due to the presence of multiple beams). Here a sweep is regarded as each set of rectangular patch records generated from each ping of MBES (separated by dotted lines in Figure 5.4). Therefore, the number of SBES returns had to be reduced to match the number of MBES sweep in order to be combined. To do so, using the most central MBES patches as references (Figure 5.4), the nearest SBES echo returns to those references were selected.

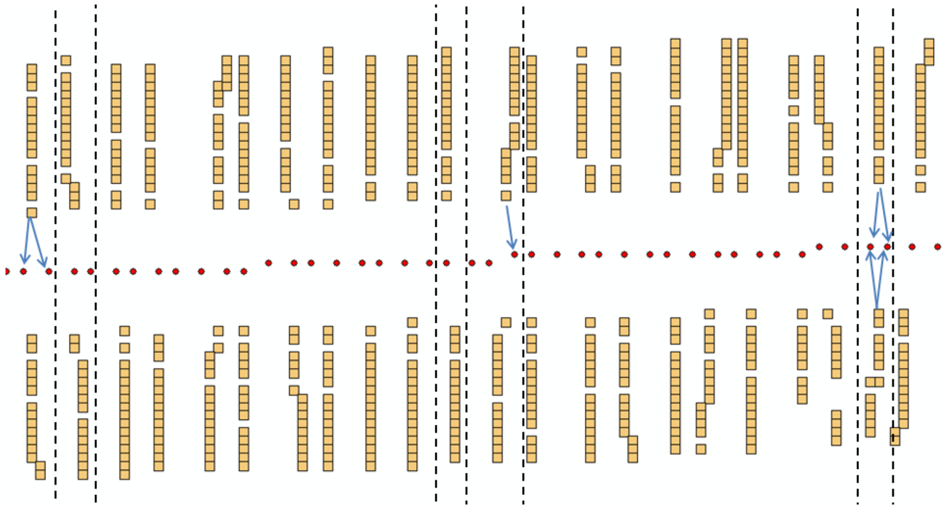


Figure 5.4: Selection of nearest SBES locations

Once the SBES locations were isolated, the statistical features of those locations were added to the MBES patches belonging to the same sweep (Figure 5.5). The patches belonging to the same MBES sweep contain the same set of SBES statistical features as it is expected that the geological variation in the vertical (North) direction to be minimal due to the narrow SBES survey line and can be ignored.

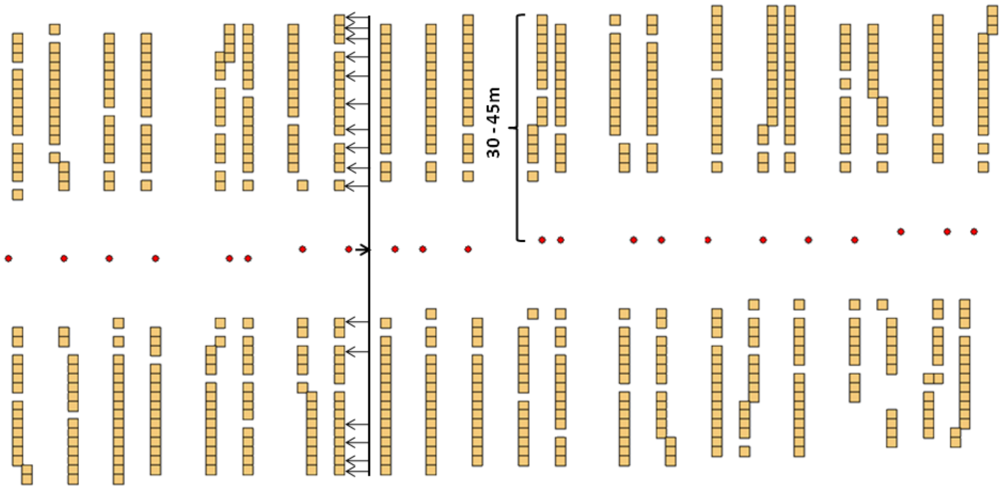


Figure 5.5: Combining SBES statistical features with MBES features

If there are any significant variations in the vertical (along seabed depth) direction, these should be captured by the MBES features. The combined dataset was clustered using two different clustering algorithms, as described in the next section.

5.6. Clustering of combined MBES and SBES dataset

PCA + k-means and fuzzy c-means were the classification algorithms of choice. The Principal Component Analysis (PCA) was used as an orthogonalisation method to produce input data for the k-means clustering, where all dimensions are orthogonal and independent of each other. In order not to lose any information, all components were selected for k-means and PCA was not used as a dimensionality reduction technique. During the k-means clustering, the number of clusters specified a priori varied from three to six. As in other two chapters, k-means was replicated 10 times to minimise the probability of hitting a 'poor' local minima by chance. For each replication the algorithm ran with a new set of initial cluster centroid positions and the cluster labels with the lowest value for within cluster sums of point-to-centroid distances were retained as the optimum results.

Though fuzzy clustering can accommodate overlapping of clusters, here it was run with 1% fuzziness so that the algorithm would treat the clusters as discrete. This was decided based on the fact that the seabed in this area is mostly sandy in nature and just one survey line (line 163) would surely not be adequate in capturing the overlapping of different clusters. Like k-means, this was also replicated 10 times to minimise the chance of being stuck in the local minima.

5.6.1. Clustering results

The main objective of this research is to test if the inclusion of SBES statistical features results improves boundary definition in seabed clustering. In the following section, clustering results obtained from the combined MBES and SBES dataset (Line-163) are visualised and compared.

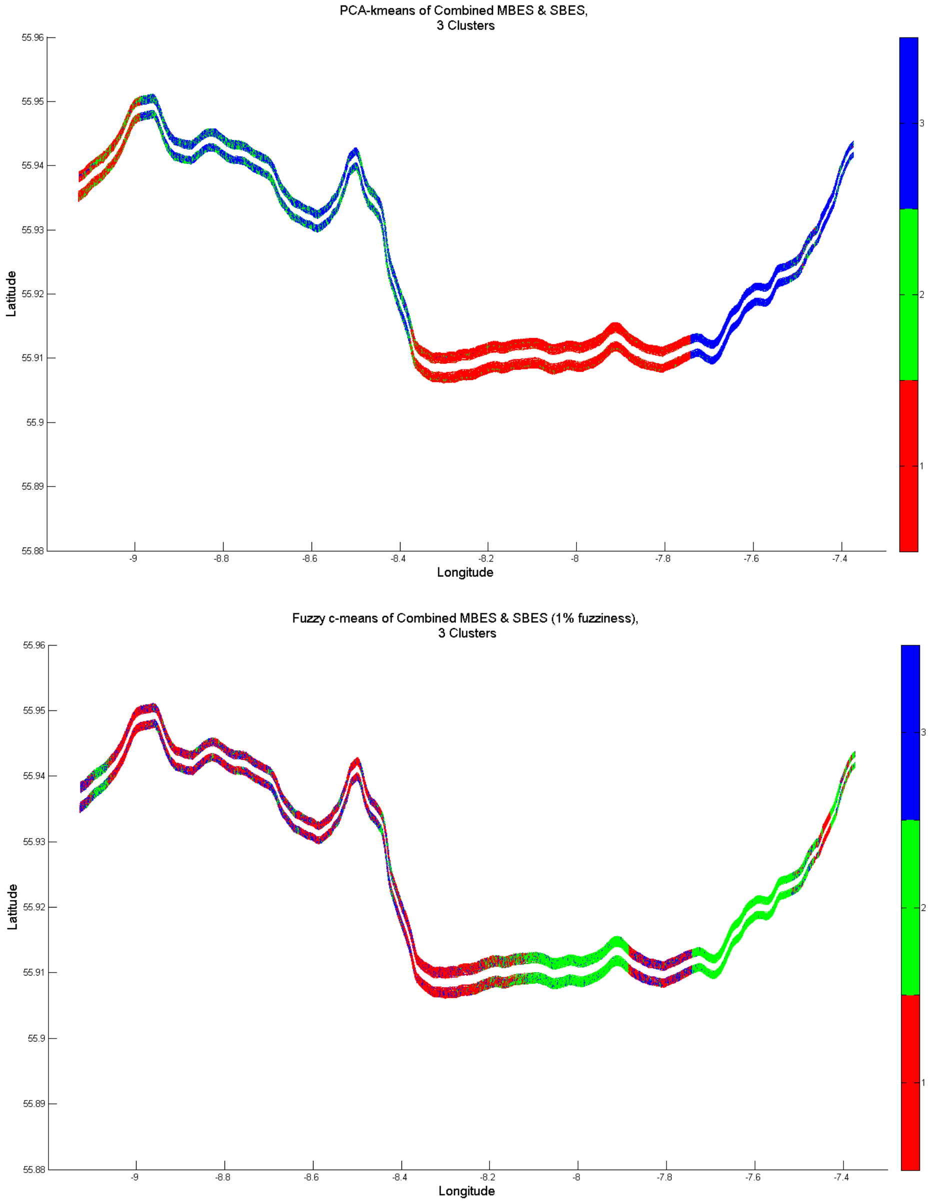


Figure 5.6: PCA-k-means (top) and fuzzy c-means (bottom) cluster results of combined MBES & SBES dataset (3 clusters)

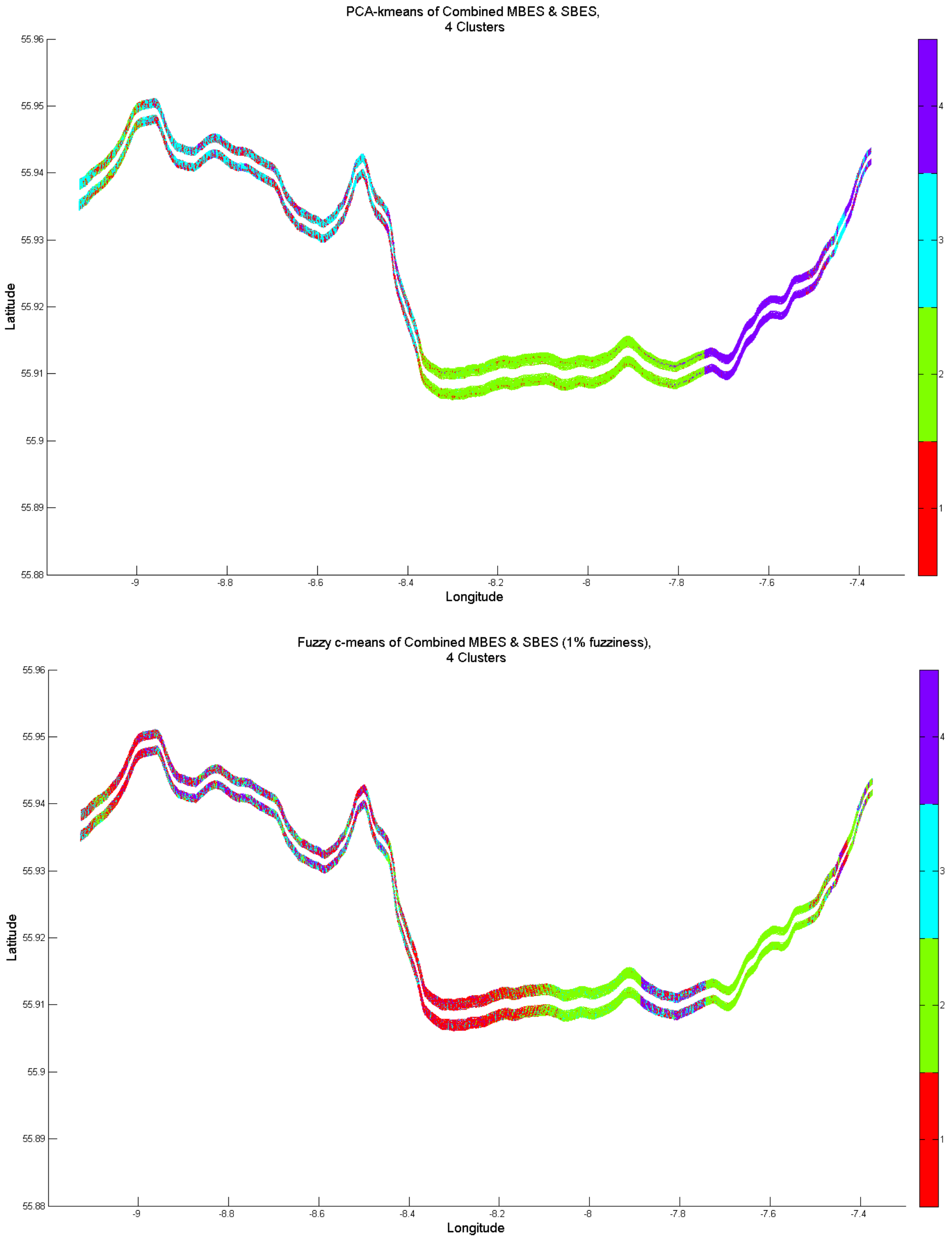


Figure 5.7: PCA-k-means (top) and fuzzy c-means (bottom) cluster results of combined MBES & SBES dataset (4 clusters)

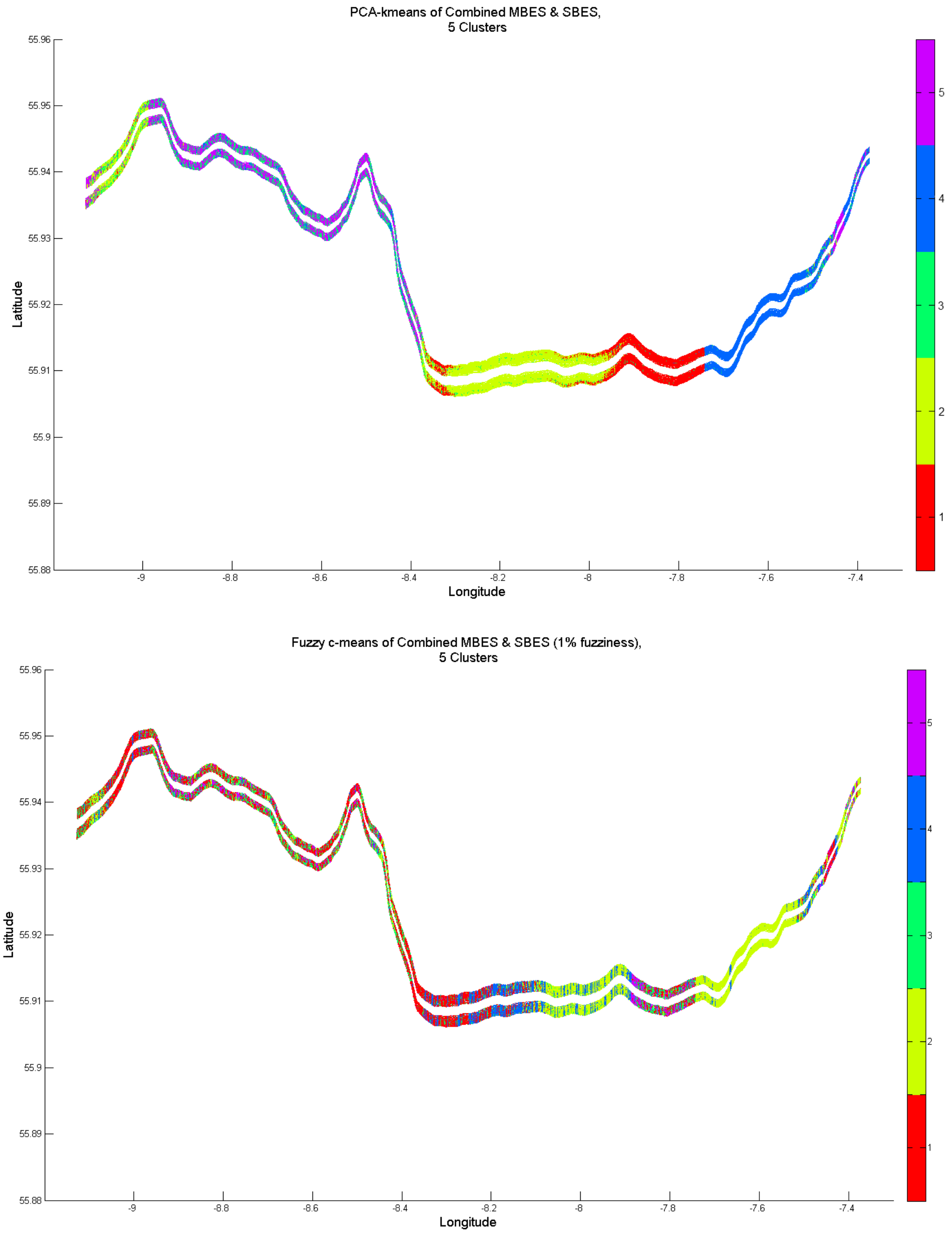
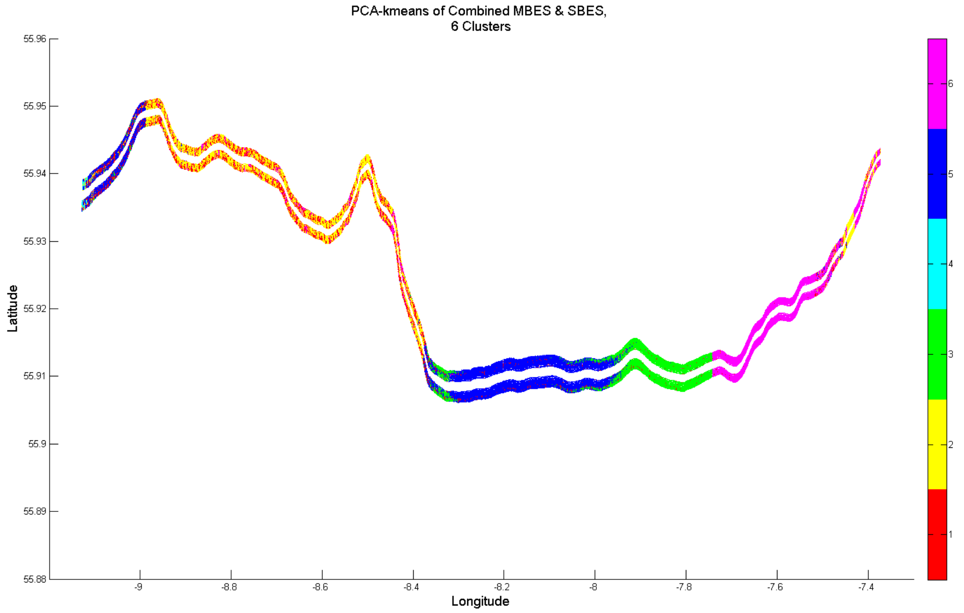


Figure 5.8: PCA-k-means (top) and fuzzy c-means (bottom) cluster results of combined MBES & SBES dataset (5 clusters)



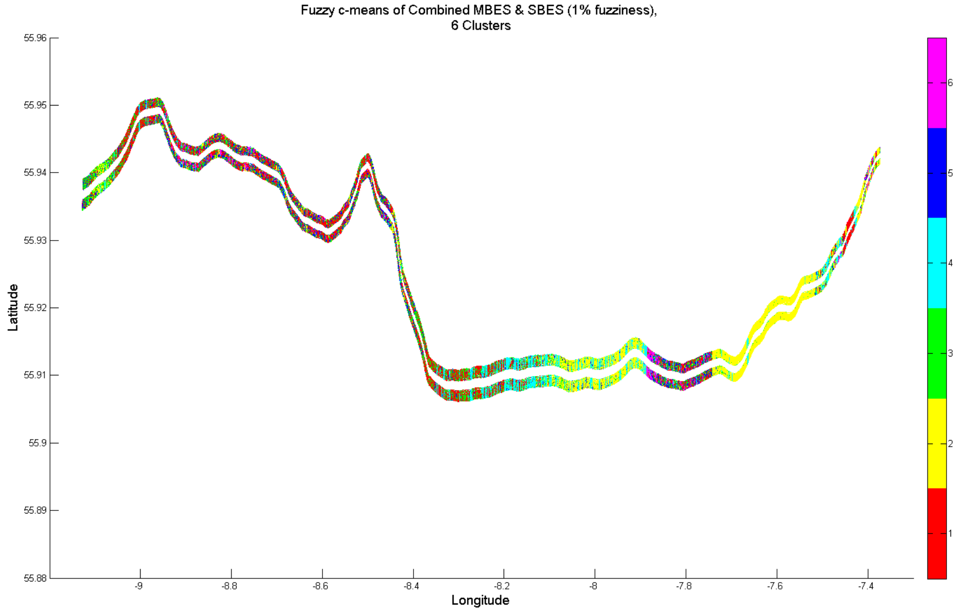


Figure 5.9: PCA-k-means (top) and fuzzy c-means (bottom) cluster results of combined MBES & SBES dataset (6 clusters)

From the results, two things are readily noticeable when compared with the results obtained MBES clustering only (Chapter 3). First, for both classifications, boundaries appear to be better defined (i.e. less mixed). Second, when the cluster outputs from two algorithms are compared visually, the combined PCA and k-means appear to have yielded better defined cluster boundaries compared to that of fuzzy c-means, while fuzzy c-means appear to have identified more segmentation in the survey line than k-means.

The fact that the cluster definition for the combined dataset clustering visually appears better than cluster results from just MBES datasets is an indication that the statistical features from SBES may have contributed to the better definition of boundaries. This observation will be validated by comparison with ground truth (Chapter 6). When visually comparing the performance of the algorithms, the results obtained from fuzzy c-means appear to have more misclassification than that of k-means. The level of misclassification increases with the number of clusters. Additionally, at higher cluster numbers the misclassification becomes spatially random ('pepper grain' shape) in nature, which is expected when the classes are more or less homogeneous (sandy) with different levels of shell presence.

For three clusters (Figure 5.6), both algorithms produce relatively well defined boundaries. However, this may not be representative as, from the figures, it is apparent that cluster one and three (for k-means) and clusters one and two (for fuzzy c-means) are the dominant clusters with clusters two and three appearing as 'peppered' throughout the survey line respectively. For four clusters, the cluster number three appears to be heavily misclassified in both algorithms and lacks proper boundary definition. For five and six clusters, k-means appears to perform better than fuzzy c-means with clearer boundaries between clusters. In the k-means clustering result, clusters four and five appear to be heavily mixed in the centre-left cluster (Figure 5.8). This could be due to the fact that this region has two classes that share a lot of geological properties. For fuzzy c-means, clusters one, two and three appear to be mixed in the centre location of the map (Figure 5.8). The same pattern was also noticeable with MBES clustering (Chapter 3). This can be due to the fact that the centre part of the study area comprises of classes that are similar in nature. The similar nature of misclassification can be observed for cluster six for both algorithms with heavier mixed classification concentration in the map centre.

5.6.2. Internal cluster validation

For this experiment, as with MBES and SBES, four commonly used internal cluster validation indices were generated and compared. As before, they are: Calinski-Harabasz or Variance Ratio Criterion (VRC) index (Calinski and Harabasz, 1974; Everitt et al., 2011), Davies-Bouldin index (Davis and Bouldin, 1979; Jain and Dubes, 1988), Dunn's index (Dunn, 1974; Halkidi et al., 2001) and Silhouette index

(Everitt et al., 2011; Halkidi et al., 2001). Their working principals are described in Chapter 2. Tables 5.1-5.2 and Figures 5.10-5.11 show the internal validation index values calculated for both k-means and fuzzy c-means.

Table 5.1: Internal validation indexes for k-means clustering on combined MBES, SBES dataset

	Cluster 3	**Cluster 4**	**Cluster 5**	**Cluster 6**
Calinski-Harabasz Index	1021.2557	1142.7402	1332.6618	1457.0885
Davies-Bouldin Index	2.1824	1.9945	2.2286	1.9013
Dunn Index	0.75586	0.62445	0.32581	0.80723
Silhouette Index	0.1185	0.11895	0.12303	0.12919

Table 5.2: Internal validation indexes for fuzzy c-means clustering on combined MBES, SBES dataset

	Cluster 3	**Cluster 4**	**Cluster 5**	**Cluster 6**
Calinski-Harabasz Index	1121.1156	1423.624	1152.955	1487.1102
Davies-Bouldin Index	12.7785	12.5364	12.9354	10.5949
Dunn Index	0.098838	0.09871	0.063141	0.11708
Silhouette Index	0.41408	0.27422	0.12534	0.82308

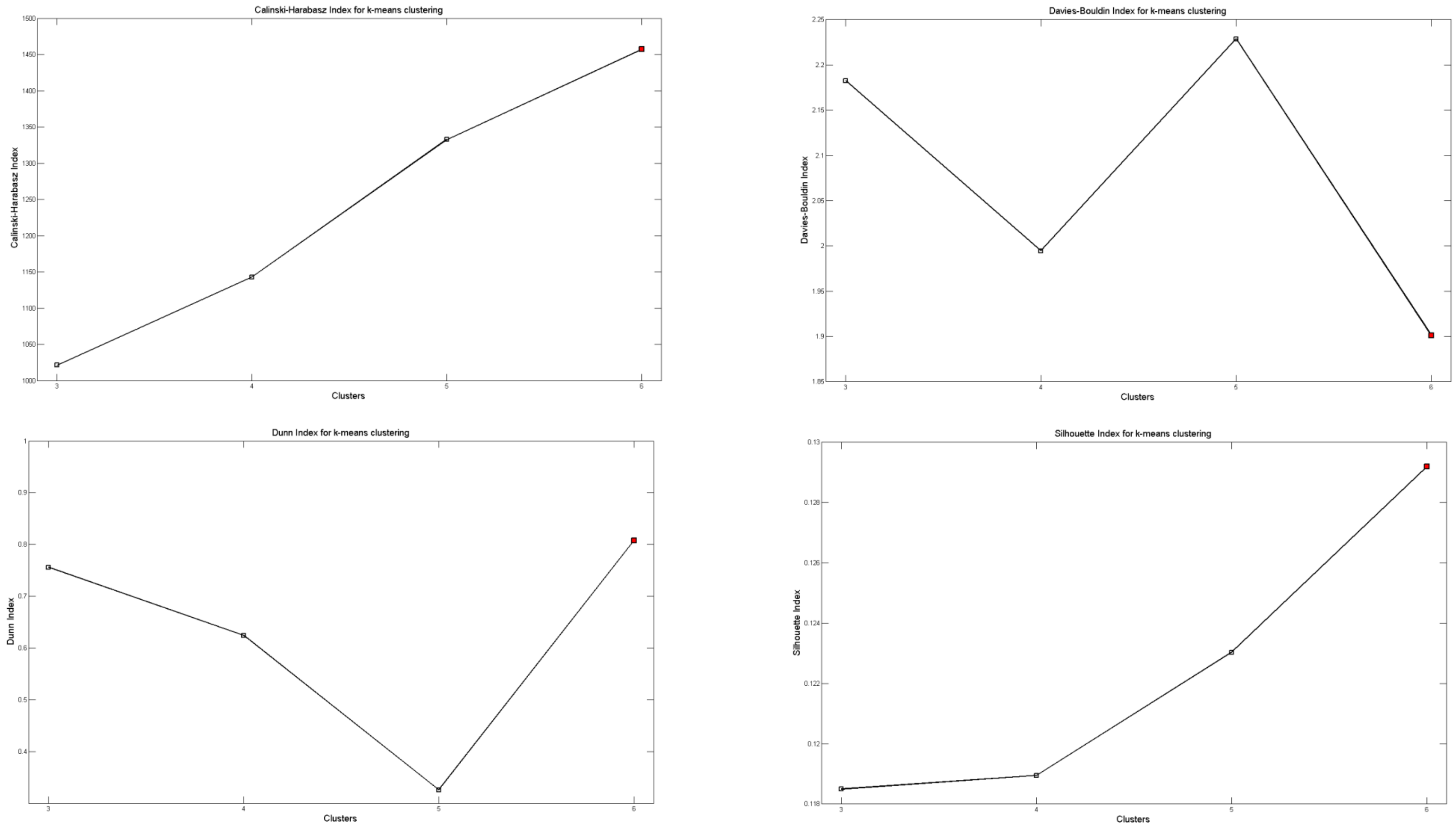


Figure 5.10: Internal validation index for k-means clustering of combined MBES, SBES dataset (red marker shows the estimated optimal number of clusters)

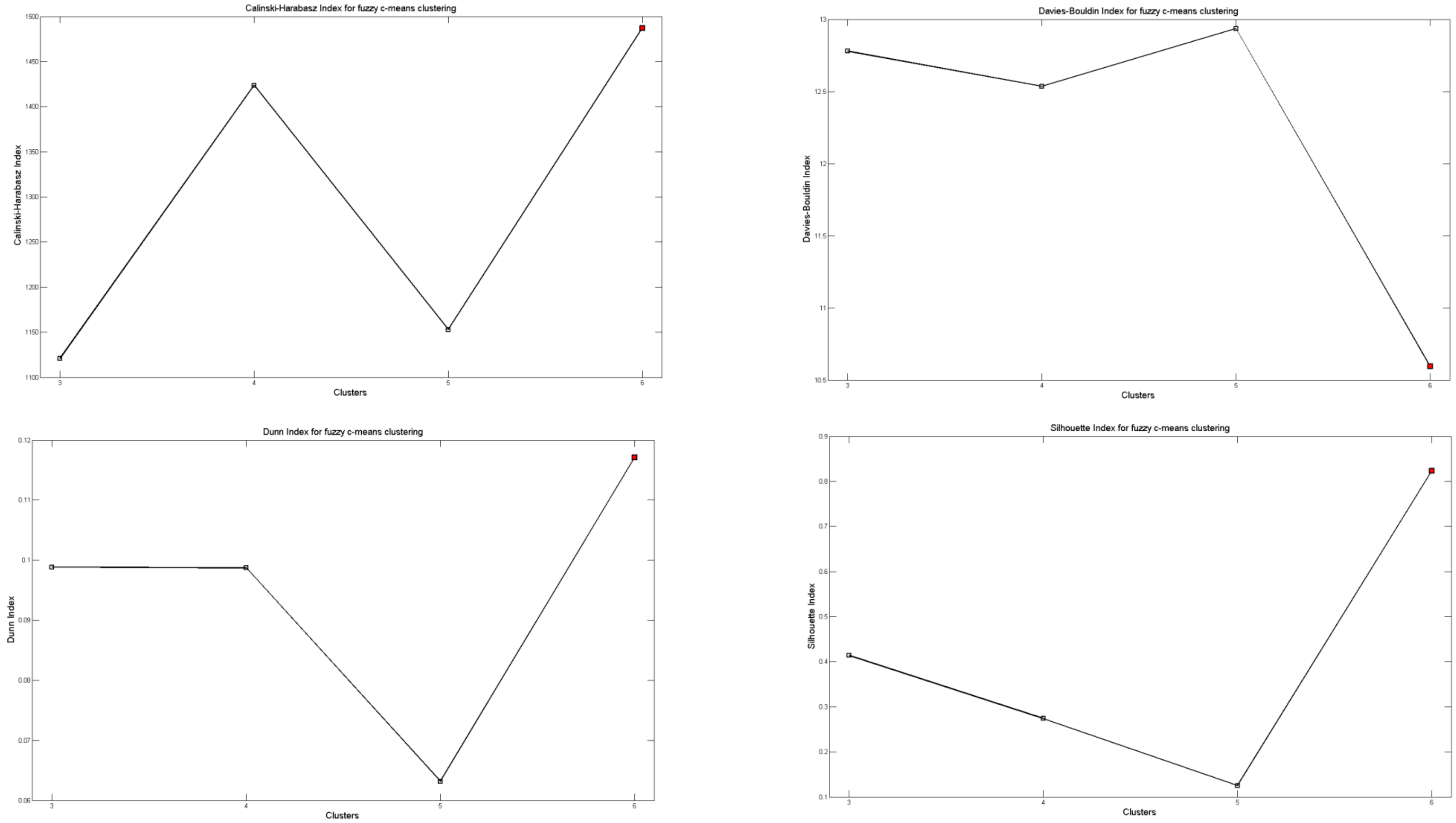


Figure 5.11: Internal validation index for fuzzy c-means clustering of combined MBES, SBES dataset (red marker shows the estimated optimal number of clusters)

These indices suggest the optimal number of inherent clusters in the data based on compactness and separation (see Chapter 2). In figures 5.10 and 5.11, the suggested optimal number of clusters is shown by the ‘red’ mark. The table shows the corresponding index values for each index. All four validity indices appear to indicate that for both clustering algorithms, six clusters are the optimum number of clusters which give good separation and compactness. This is similar to that of ‘MBES only’ clustering shown in Chapter 3. The study area (Malin Head) is known to be sandy in nature with varying concentration of shells. Thus, it is quite possible that the clusters are comprised of a mixture of different grain size sands with varying level of shells, gravel, burrows etc. This will be further analysed in Chapter 6 where we develop a ground truth database for both MBES and SBES data.

5.7. Discussion

Rapid development in sensor technology has seen a significant improvement in the quality and accuracy of both MBES and SBES echosounders over the last decade. These sensors are now capable of providing high resolution seabed backscatter data and thus produce better seabed maps. Though both MBES and SBES are usually fitted to the hull of the same survey vessel and produce simultaneous echo returns from the same survey locations, only MBES data have been extensively used for seabed classification. The use of SBES is normally limited to fish shoal detection and digital elevation model generation. However, the potential of SBES to provide high resolution subsurface information can be of interest in seabed mapping as the resulting clusters could provide a more accurate description of the seabed that that of MBES.

Because of their high resolution and sensitivity to suspended particles in water, good quality raw backscatter SBES data are fundamental to any classification, especially when the data are prone to inevitable noise. Given a good quality dataset, SBES should be able to provide a reliable seabed classification as it is the only echosounder that records vertical echo returns with minimal angular distortions. The main objective of this study was to test if SBES can provide complementary information to MBES classification i.e. if the MBES and SBES datasets are combined, would they yield better classification results than that of MBES only classification?

For this study, the visual groups (16 groups or 35 variables) obtained from SOM implementations on MBES were used as the main dataset. We only used line 163 of MBES data as SBES data were available for this survey line only. Four types of statistical features were generated from the 'cleaned' SBES dataset (Chapter 4). The SBES dataset had a high volume of echo returns and as a result the data were combined using a nearest distance approach i.e. only the closest SBES echo locations to the MBES locations were selected. Once the combined dataset was complete the dataset was clustered using both PCA-k-means and fuzzy c-means and the results visualised.

After visual exploration of the results, the combined dataset appears to have produced better defined cluster boundaries compared to MBES clustering (Chapter 3). Among the clustering algorithms k-means had better cluster definition than fuzzy c-means while fuzzy c-means appeared to have performed better when only MBES data were used. As with any clustering of geological features that are homogeneous to some extent, the level of misclassification increased with the number of clusters. The misclassification was more concentrated in the centre to centre-left part of the map. This becomes more prominent in the dataset with 5 and 6 clusters and indicates that the centre region of the map may have classes that are more similar in nature. This observation will be further validated with ground truth data comparison (Chapter 6).

The visual comparison of cluster results in this chapter with those of MBES (Chapter 3) indicates that SBES may have contributed to a better definition of cluster boundaries than the ones achieved with MBES data only. However, this is subject to ground truth validation, which we discuss in the next chapter. The main objective behind the combination of data from different sensors was to see if high resolution vertical echo returns from SBES could improve the cluster boundaries obtained from MBES clustering. The results, when visually compared, appear to show that SBES did in fact improve the boundary definitions to some extent.

The major drawback of this study was that the SBES dataset available had significant vertical displacement and thus was not optimal for classification. This means that the optimal depth detected from the data fluctuated by ±5m (approximately 3m above seabed and 2m below seabed from the derived average depth over the survey line).This could eventually compromise the quality of the classification outcome to some degree as the statistical features generated from this dataset included not only the echo envelope but also some portion of above and below seabed data segment and therefore would contain a significant bit of noise. However, it was still expected that SBES would, to some degree, contribute to the classification quality.

Another limitation was that the study area was limited to one survey line (line 163). Though the length of the survey line was approximately 100 km, the width varied from 45m to 60m. This is quite narrow and limits the clustering algorithm to capture the geological variation in East-West direction only. The narrow study area was the main reason behind limiting the fuzzy c-means to 1% fuzziness only as higher degree of fuzziness would not be logical for such a narrow study area.

In the next chapter, we develop a ground truth database for the validation of cluster results obtained in Chapters 3, 4, and 5. In particular, cluster results using the estimated optimal number of clusters for MBES, SBES and combined MBES/SBES will be cross-validated against the ground truth data and their accuracy analysed, which should indicate how representative the computational clusters are of actual seabed. Their accuracy will also be compared with that of the MBES datasets clustered with all features. This will provide a quantifiable measure of how effective our approach was compared with the de facto standard mapping procedure (Preston, 2009).

Chapter 6

Cluster validation

This chapter provides a quantifiable measure of the cluster analysis undertaken in the previous chapters through cluster validation using ground truth data.

Chapter contents

The ultimate conclusion of any seabed clustering task is the generation of maps validated by ground truth. Good quality seabed maps have a number of environmental and economic uses. This chapter includes the validation analysis of clusters obtained in the previous studies described in Chapters 3, 4, and 5. The cluster results obtained in the previous chapters will be cross validated against the ground truth results and their accuracy analysed. This will indicate the ‘goodness’ of the clusters in comparison to the class labels from the ground truth dataset, thus ultimately providing a quantifiable measure of how effective our approach is in optimising the datasets.

6.1. Research background and justification

The objective of clustering is the partitioning of a given dataset for discovering groups and identifying distributions and patterns in the underlying data. Clustering objective is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters. In order to establish the final data partitioning from any clustering results, the clusters need to be validated against known class labels from the same locations. Cluster validation is concerned with the quality of clusters generated by clustering algorithms. Any validation technique for a give partitioned dataset usually attempts to answer questions such as:

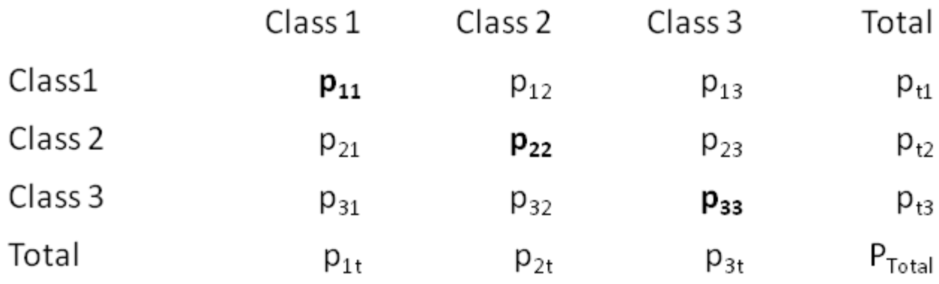
- For any number of clusters: how pronounced is the cluster structure that has been identified?
- When more than one algorithm is used: how do clustering solutions from different algorithms compare?
- For varying cluster numbers: how do clustering solutions for different values of algorithmic parameters (e.g. the number of clusters) compare?

When information about true class membership is available (i.e. ground truth), external cluster validation techniques are usually used. These provide an objective way of assessing algorithm performance. When such external knowledge is not available, internal measures need to be used which attempt to measure the quality of the clusters based on the intrinsic properties of the data.

The ultimate goal of this PhD research is to evaluate the potential of visual analytic methods in improving the quality of seabed mapping. Previous chapters concentrated on visual comparison of cluster results obtained from optimised MBES and SBES datasets as well as on data-driven internal cluster validation. This chapter will focus on quantifying those clustering results through cross correlation with external ground truth data. This is done by using both a confusion matrix and a kappa coefficient, two common techniques used in remote sensing for validating the cluster results obtained in segmentation of remotely sensed images of natural objects or features (Campbell, 2008; Lillesand et al., 2004). In the following, we briefly introduce these two methods.

Confusion matrix

In a confusion matrix, cluster results are compared with known class labels from the same locations. The confusion matrix provides a measure of how the algorithm performed. An example is shown in Figure 6.1, the rows (1,2,...,i) correspond to the known class of the data and the columns (1,2,...,j) (here, i = j) correspond to the predictions made by the algorithm.



	Class 1	Class 2	Class 3	Total
Class1	$\mathbf{p_{11}}$	p_{12}	p_{13}	p_{t1}
Class 2	p_{21}	$\mathbf{p_{22}}$	p_{23}	p_{t2}
Class 3	p_{31}	p_{32}	$\mathbf{p_{33}}$	p_{t3}
Total	p_{1t}	p_{2t}	p_{3t}	P_{Total}

Figure 6.1: An example of a confusion matrix

The value of each of element in the matrix represents the number of predictions made with the class corresponding to the column. The diagonal elements show the number of correct classifications made for each class, while the off-diagonal elements represent the number of misclassifications (Campbell, 2008; Mather & Koch, 2010).

There are two types of error measures that can be calculated from the confusion matrix: error of commission and error of omission (Mather and Koch, 2010; Campbell, 2008; Lillesand et al., 2004). The commission error, which occurs when one incorrectly identifies data points associated with a class as other classes, or when one improperly separates a single class into two or more classes, can be calculated as:

$$Commission\ Error = \frac{Total\ of\ off - diagonal\ elements\ in\ column\ j}{Total\ of\ row\ i} \qquad (6.1)$$

For example,

$$\text{Commission Error of 'Class 1' (Figure 6.1)} = \frac{\mathrm{p}_{12} + \mathrm{p}_{13}}{\mathrm{p}_{t1}} \qquad (6.2)$$

Errors of omission occur whenever the algorithm fails to recognize data points that should have been identified as belonging to a particular class; that is data belonging to a particular class are classified into different classes. This is calculated as:

$$\text{Omission Error} = \frac{\text{Total of off} - \text{diagonal elements in row i}}{\text{Total of row i}} \tag{6.3}$$

To give an example, the omission error of class 1 in the confusion matrix in Figure 6.1 is given as:

$$\text{Omission Error of class 1 (Figure 6.1)} = \frac{\mathrm{p}_{12} + \mathrm{p}_{13}}{\mathrm{p}_{\mathrm{t}1}} \tag{6.4}$$

Another measure of the quality of clustering is the mapping accuracy, which provides a measure of correct classification of individual classes are calculated using both correct and misclassified samples for each class as follows:

$$\text{Mapping accuracy of class 1 (Figure 6.1)} = \frac{\mathrm{p}_{11}}{\mathrm{p}_{11} + \mathrm{p}_{12} + \mathrm{p}_{13} + \mathrm{p}_{21} + \mathrm{p}_{31}} \tag{6.5}$$

Finally, the overall accuracy of the clustering algorithm can be calculated from the ratio of the sum of the diagonal elements of the confusion matrix versus the number of elements in the confusion matrix:

$$\text{Overall accuracy (Figure 6.1)} = \frac{\mathrm{p}_{11} + \mathrm{p}_{22} + \mathrm{p}_{33}}{\mathrm{P}_{\text{total}}} \tag{6.6}$$

Kappa coefficient or Cohen's kappa

When two classification algorithms applied to the same data are to be compared, the kappa coefficient (k) is often used to summarise the information provided by the confusion matrix. Kappa is calculated from the observed and expected frequencies on the diagonal of confusion matrix. This is a common evaluation approach in remote sensing where experimental results are compared with those that are known (i.e. ground truth) (Campbell, 2008). The calculation of kappa is based on the difference between how much agreement is actually present ("observed" agreement) compared to how much agreement would be expected to be present by chance alone ("expected" agreement) and is calculates as (Bishop et al., 2007; Mather & Koch, 2010):

$$\mathrm{Kappa(k)} = \frac{\text{observed agreement} - \text{expected agreement}}{1 - \text{expected agreement}} \tag{6.7}$$

The observed agreement is calculated as follows (Figure 6.1):

$$\text{Observed agreement} = \frac{\mathrm{P_{11} + P_{22} + P_{33}}}{\mathrm{P_{Total}}} \tag{6.8}$$

And the expected agreement is calculated from expected cell frequencies, which is calculated as below (Figure 6.1):

$$\text{Expected cell frequency,} \qquad \mathrm{P_i} = \frac{\mathrm{P_{it} \times P_{ti}}}{\mathrm{P_{Total}}} \tag{6.9}$$

where, *i*=range of classes (1-3 in this case)

Finally, expected agreement is calculated using the expected cell frequencies as follows:

$$\text{Expected agreement} = \frac{\mathrm{P_1 + P_2 + P_3}}{\mathrm{P_{Total}}} \tag{6.10}$$

The value of kappa is presented as a percentage. A value of zero indicates that there is no agreement whereas a value of 1 indicates complete agreement between the classification output and the reference data. Kappa statistic measures the proportion of agreement between two rates with correction for chance and the higher the value of kappa, the stronger the agreement. Kappa scores ranging from 0.4-0.6 are considered to be fair, 0.6-0.75 are good, and scores greater than 0.75 are excellent. Though these guidelines are frequently used in different research fields they are not universally accepted as no evidence was ever given to support this guideline. Despite this the kappa coefficient remains a frequently used method in quantifying the classification accuracy (Fleiss, 1981; Landis & Koch, 1977; Mather & Koch, 2010).

In the next section, we describe how we used these measures (confusion matrix, kappa coefficient) to validate our clustering results against a ground truth database generated from grab samples collected and provided by GSI and MI.

6.2. Data for validation

Three different sets of clustered data sets were produced in Chapters 3, 4, and 5: from MBES clustering, from SBES clustering, and from combined MBES and SBES clustering. For clarity, we give a short description of each of these datasets before introducing the ground truth database against which our results are compared:

Clustered MBES data

The 'original' form of MBES data comprised 132 Full Feature Vectors (FFVs) (see Chapter 3). The dimensionalities of the datasets were optimised using a self organising map (SOM), which resulted in 35 variables in the form of 16 visual groups. The datasets (both the original and optimised) were clustered using the commonly used combined PCA and k-means. For the original datasets, the first three principal components were used which comprises of around 95% of the information. These three components were then clustered using k-means. Therefore, for the original datasets, PCA functioned as both dimensionality reduction and orthogonalisation tool. For the datasets optimised with visual groups, all 35 components were used in the k-means clustering thus including 100% of the information. Here, PCA functioned only as an orthogonalisation tool. Though different cluster numbers were tested, the internal validation indices estimated that six was the optimal number of clusters for the data. Therefore we only used results with six clusters for validation. Thus for the MBES clustering, two datasets were selected for validation:

- 6 clusters from k-means clustering of MBES datasets (132 FFVs)
- 6 clusters from k-means clustering of MBES datasets (35 FFVs)

Clustered SBES dataset

The original SBES dataset came from only one survey line (line 163) and comprised a time series of backscatter amplitude in three frequencies (see Chapter 4). The original dataset was filtered using a Butterworth filter and an automatic outlier detection algorithm was run to get rid of the 'bad' samples. The dataset was further evaluated using a visual time series exploration tool TimeSearcher© and further outliers were detected. Once the data were cleaned, they were clustered using PCA-k-means and fuzzy c-means with 1% fuzziness. Different cluster numbers (3 to 8 clusters) were tested. The results were then evaluated using four different internal validation indices and it was suggested that the optimal number of clusters for the dataset were three. A frequency of 200 KHz in this dataset was deemed too noisy to produce any valid clusters. Thus for the SBES clustering, four datasets were selected for validation:

- 3 clusters from k-means clustering of SBES datasets (12 KHz)
- 3 clusters from fuzzy c-means clustering of SBES datasets (12 KHz)
- 3 clusters from k-means clustering of SBES datasets (38 KHz)
- 3 clusters from fuzzy c-means clustering of SBES datasets (38 KHz)

Clustered MBES+SBES dataset

As the SBES data were only available for one survey line (line 163), MBES records from that line were selected to produce the combined dataset (see Chapter 5). The MBES dataset comprised 35 variables from 16 visual groups, while four different statistical features were generated from the SBES dataset. They were combined using the nearest neighbour method (Euclidean distance) i.e. for each MBES ‘sweep’, the closest SBES was located and combined (Chapter 5). The combined dataset was clustered using both PCA-k-means and fuzzy c-means with different cluster numbers (3 to 6 clusters). From the internal validation tests, it was estimated that the optimal number of clusters was six. Therefore, two datasets were selected from the cluster results of the combined dataset:

- 6 clusters from k-means clustering of MBES+SBES dataset
- 6 clusters from fuzzy c-means clustering of MBES+SBES dataset

The next section describes the process of the generation of ground truth databases for validation of above mentioned clustered datasets.

6.3. Ground truth database

For any mapping project, a good ground truth is important for proper cluster validation. Mapping of the seabed is not any different when it comes to cluster validation and labelling. However, collection of ground truth for seabed mapping poses a number of challenges unlike traditional Earth surface mapping.

In the shallow waters, ground truthing is done by collecting grab samples, running video lines and/or sending a diver to collect samples to establish a ground truth database for the region. For deep water surveys, collection of ground truth data gets more complicated as the survey area increases with the depth. The operation cost for instruments used for collecting samples becomes very high and sending divers beyond a certain depth becomes impossible. Unlike high resolution satellite images in visible spectrum (ex: Quick Bird) or photographs taken using aerial photography, there are no means available to obtain high resolution images for the entire seabed survey area to be used as a reference ground truth image for cluster validation. Therefore, a combination of sample collection, videos and still images serves as the ground truth for seabed mapping (MESH, 2011).

Sample collection

There is a variety of sampling devices available for collecting samples. These are usually limited to grabs and corers which sample sediments and their infauna (animals living within the seabed) as well as trawls and dredges. Human observers play a part through the use of video and still cameras mounted on towed sledges, drop-frames or Remote Operated Vehicles (ROVs), but these only provide a view of the surface substrata. A combination of these sampling and observational methods is required to provide all the information needed to classify the seabed (Brown et al., 2002).

It is important that sampling should be representative rather than exhaustive as the collection of samples is both costly and time-consuming to process, analyse and interpret. To be representative, appropriate sampling techniques should be used on each seabed type. A minimum sampling requirement should be set, considering the level of classification accuracy and confidence required in the final map. Single sampling of each ground type forces the assumption that it is homogeneous. Replicate sampling allows some assessment of variability within and between different clusters. The number of replicates taken may be determined by a rule of thumb and expert judgement based on the assessment of the information on the heterogeneity/homogeneity of the ground as it appears to the remote sensing instrument (Brown et al., 2002).

In many cases it will be necessary to set aside a proportion of the ground truth samples to test the accuracy of the map once it has been produced. The need for these 'validation' samples should be factored into the survey design.

Optimising the ground-truth survey design is often an iterative process, and is kept flexible to a certain degree. A draft plan of the ground truthing survey is usually handed to those who will conduct the survey to check operational feasibility (access to sites, navigation hazards, Health & Safety matters, etc). However, the fine detail of the design frequently depends on the outcome of the seabed survey and the prevailing conditions at the time of sampling.

For the GSI and Marine Institute (MI) surveys, which provided ground truth data used in this chapter, three types of equipments were used for the collection of grab samples: Hamon grab, Van Veen grab and Day grab.

Hamon grab

The Hamon Grab (Figure 6.2) comprises of a sampling scoop that is box shaped and is mounted in a triangular frame. On reaching the seabed, tension wires are released which activates the grab. This causes the sampling bucket to pivot through 90º thus driving the sample bucket through the sediment and pushing seabed sediment into the bucket (Figure 5.6). The open end of the bucket then comes against a rubber sealed steel plate which stops the sediment escaping during recovery. Weights are attached to the grab to minimize the lateral movement of the supporting frame during sample collection. A major drawback of the Hamon grab is that the sediment sample is mixed during the process of collection and retrieval, thereby precluding the examination or sub-sampling of an undisturbed sediment surface.

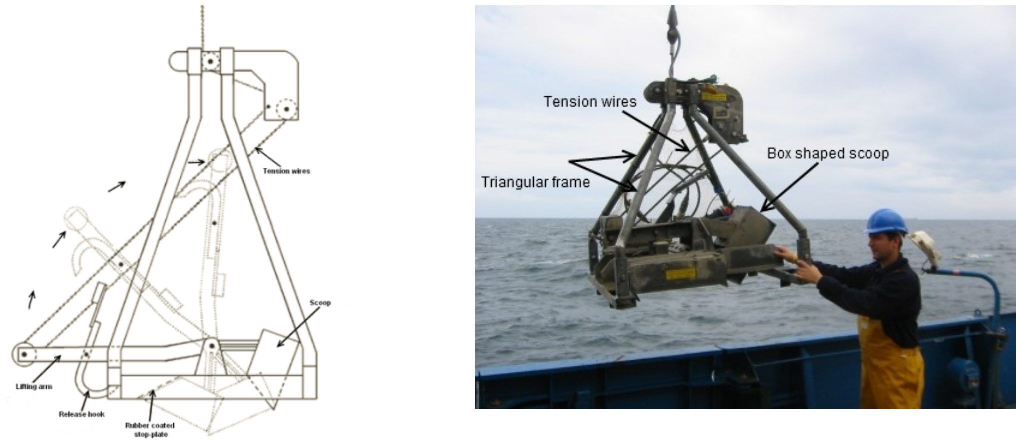


Figure 6.2: A Hamon grab (Brown et al, 2002)

Van Veen grab

Van Veen Grabs are very simple sampling devices based on two hinged bucket sections connected to extended lever arms (Figure 6.3). A simple locking device attached to a single lift line uses the weight of the grab to hold the jaws of the buckets open during descent to the seabed. On contact with the seabed the weight of the grab is taken off the locking mechanism which falls away. During recovery the weight of the grab acts on the ends of the lever arms applying a substantial closing force on the grab buckets. The grab is a simple robust mechanism only requiring single wire operation and is effective in any water depth. But it also shares the major drawback of Hamon grab as the sediments are significantly disturbed during sampling.

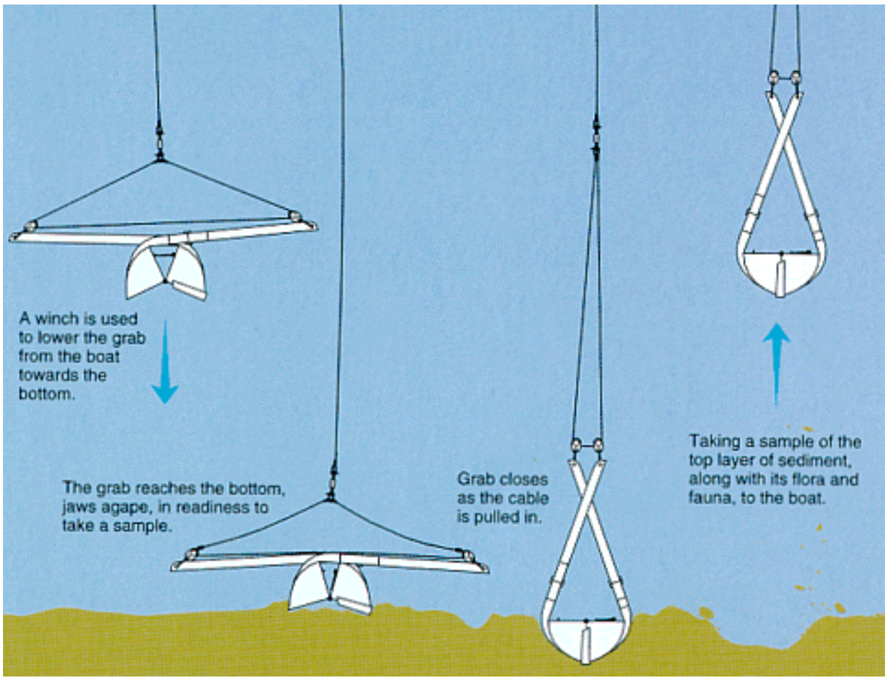

A winch is used to lower the grab from the boat towards the bottom.
The grab reaches the bottom, jaws agape, in readiness to take a sample.
Grab closes as the cable is pulled in.
Taking a sample of the top layer of sediment, along with its flora and fauna, to the boat.

Figure 6.3: Working principal of Van Veen grab (GeoSI, 2011)

Day Grab

Day Grabs consist of two hinged bucket sections slung within a pyramidal frame (Figure 6.4). Within the frame tensioned stainless steel warp wires retain the sample buckets in the open position. The instrument is triggered by contact of pad feet with the seabed. Once triggered, the weight of the instrument is transferred along the warp wires, forcing shut the grab jaws. Sample recovery is thus not reliant on the momentum of the grab during impact reducing sediment disturbance to a minimum.

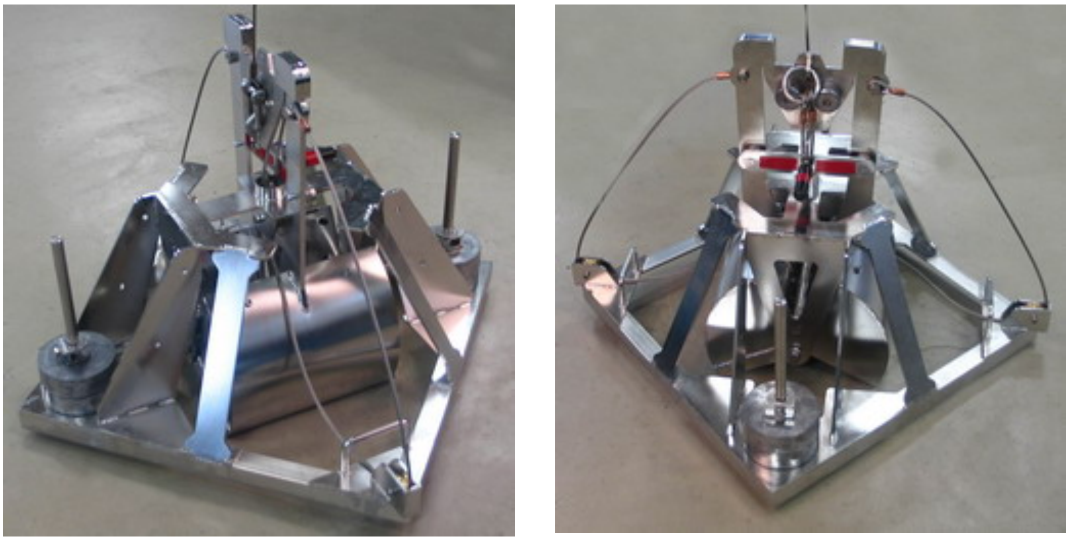


Figure 6.4: A Day grab (KC Denmark, 2011)

A Day grab is mainly designed for sampling soft sediments i.e. sands, muds etc. Its efficiency somewhat diminishes when the sediments are coarse in nature due to the tendency of larger particles to prevent closure of the buckets, causing loss of sample. However, where there is a high percentage of soft sediment (sands or muddy sands) associated with a gravelly component a Day grab is used for its capability to sample finer particles, albeit with the likelihood of a relatively high failure rate.

Ground truth samples at Malin Head

The survey location used for this research represents a small section of the larger survey area at Malin Head. A total of 42 grab samples were taken at and around the survey location (Figure 6.5). Though GSI routinely use all three of the above mentioned grabs for sample collection, only a Hamon grab was used for sample collection in this study area for its simplicity, smaller size and effectiveness in shallow waters. Once the grab samples were collected, they were analysed by geologists and labelled. Almost all the grab samples were a mixture of different types of sand, shells and burrows. Spatial locations of these grab samples along with their descriptions were prepared by GSI and supplied to us to be used in the generation of ground truth databases.

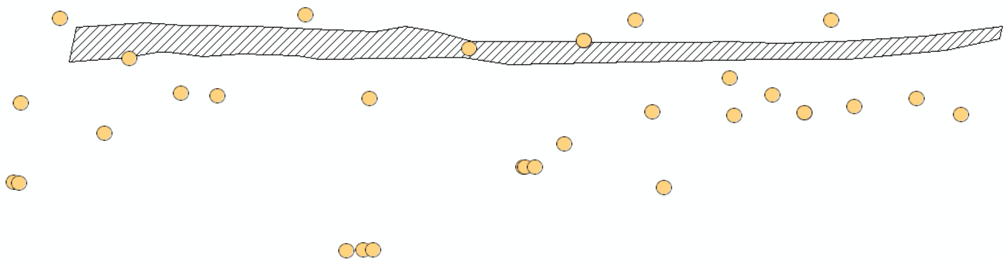


Figure 6.5: Grab sample locations around the study area

From the set of grab samples, different classification zones were created based on the composition of seabed materials in each grab samples. This was done to generalise the properties of the seabed from the grab samples for easier interpretation. A total of seven classification zones were created (Figure 6.6). Table 6.1 summarises the seabed properties within these zones.

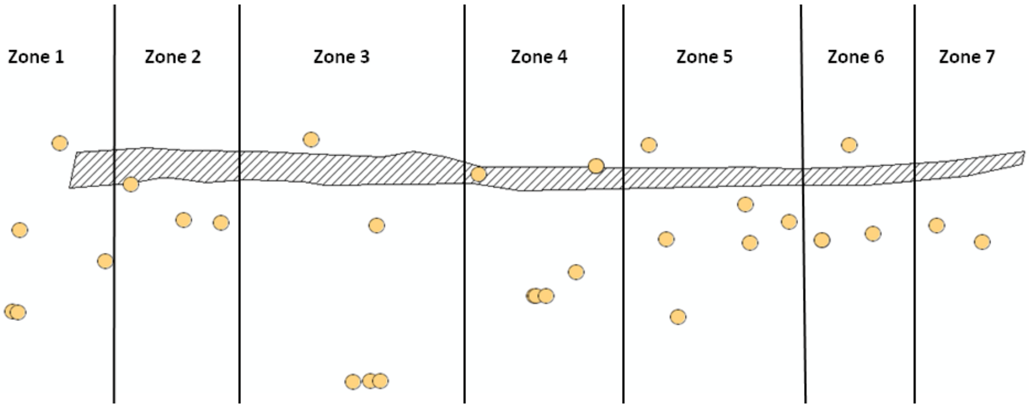


Figure 6.6: Different classification zones for the survey area based on the grab samples collected

These zones are created by grouping together the grab samples that showed similar seabed properties i.e. each zone is basically a generalisation of the seabed properties obtained from the grab samples that were located in close proximity and shared most common geological features. The zones are assumed to be vertical due to the small width of the study area and there are not enough grab samples available from the study area to establish a more representative boundary of class zones. Though the length of the study area is about 100 kms, the effect of vertical zone boundary on classification accuracy is assumed to be minimal as the grab sample indicate no abrupt change of boundaries. The effectes of vertical zoning on accuracy can only be quantified for a wider study area that includes a higher number of grab samples.

Table 6.1: Summary of classification zones in Malin Head

Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Zone 7
Silty sand with organic material. Green brown in colour.	Medium grain sand with small presence of shell, gravel and cobble. Green brown in colour.	Silty clay with gravel size clasts. Moderate presence of sand.	Silty sand. Presence of small shells and gravel size clasts.	Fine grain sand. Heavy presence of shells (20% of sample).	Medium grain sand. Heavy presence of shells and burrows.	Coarse grain sand. Presence of fine to coarse pebbles.

From Table 6.1, it is apparent that the there is not much fundamental variation in seabed in the survey area. The area mainly comprises of different mixtures of sands with varying level of shell, gravel, clast, and pebble presence. Shells in this area are mainly comprised of clams, scallops, and dentalia.

Class labels for MBES and combined MBES-SBES classification

The next step is to define cluster labels from the zones. Clustering labels should be based upon the underlying composition of seabed. That is, the grain size of the soil/sand should be the main divisor of clusters. In that regard there are five types of soil present in the area – silty clay, silty sand, fine, medium and coarse grain sand. However, the presence of different level of shells, gravel and other materials will contribute to the determination of cluster labels. Based on this, we have aggregated classification zones from Table 6.1 into six cluster labes as follows: ‘Zone 1’ (table 6.1) represents a unique area that is not recurrent in any other zones. This area is silty sand with a presence of organic material. Therefore, this area is labelled as the first cluster or ‘Class 1’.

‘Zone 2’ and ‘Zone 6’ both consist of medium grain sand. However, ‘Zone 2’ has a small presence of shell, gravel and cobble while ‘Zone 6’ has a heavy presence of shells and burrows. This difference is likely to separate these two zones into different classes when classified using MBES data. Therefore, ‘Zone 2’ was labelled as ‘Class 2’ and ‘Zone 6’ would be labelled later on.

The main components of ‘Zone 3’ and ‘Zone 4’ are silty clay and silty sand respectively. ‘Zone 3’ has some presence of clasts that are of gravel size. There is also some presence of sand. ‘Zone 4’ has some shells and gravel size clasts. The presence of sand and similar gravel sized clasts makes these two zones very similar in nature and the backscatter response from these two zones would be very similar as well. Therefore, both these zones were labelled as ‘Class 3’. Zone 5 and 7 are very different from each other as well as compared to other zones. Therefore, zone 5, 6, and 7 were labelled ‘Class 4, 5, and 6’. Table 6.2 lists the classification class labels and their generalised description.

Table 6.2: Class description for MBES classification

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class value	1	2	3	4	5	6
Zones in class	1	2	3 and 4	5	6	7
Class title	Silty sand	Medium grain sand with light shell, gravel, and cobble	Silty sand with clay, gravel and clasts	Fine grain sand with heavy shells	Medium grain sand with heavy shells and burrows	Coarse grain sand with pebbles

Cluster labels for SBES classification

In order to differentiate between homogeneous features such as different types of sand, extensive ground truthing is required for proper establishment of cluster boundaries. The ground truth samples should be adequate enough to develop a database based on grain size and their corresponding backscatter amplitudes. This is to help the analyst distinguish between fine differences that are present in the backscatter returns. Apart from extensive ground truth availability, a combination of unsupervised clustering and supervised classification is almost always recommended in the classification of geological features that returns identical backscatters. In this process, enough clusters would be generated first using unsupervised clustering. A post-classification aggregation would be applied on the clustered database- where the analyst with the help of the extensive ground truth information aggregates the clusters into pre-determined class labels. Unfortunately, this was not possible in the case of SBES classification and a precise differentiation of the geological features (e.g. fine sand and silty sand) would not be possible with the available ground truth data.

The number of clusters present in SBES is expected to be lower than that of MBES (Chapter 4) as the subsurface content of SBES was taken into account when clustering the dataset. For the determination of cluster labels for SBES classification, the main five soil types in the area were taken into account. These types are: silty clay, silty sand, fine, medium, and coarse grain sand. The rationale is that the subsurface content of the SBES dataset contains information of the underlying seabed soil properties and is not affected by the shells, gravels and clasts present on the surface.

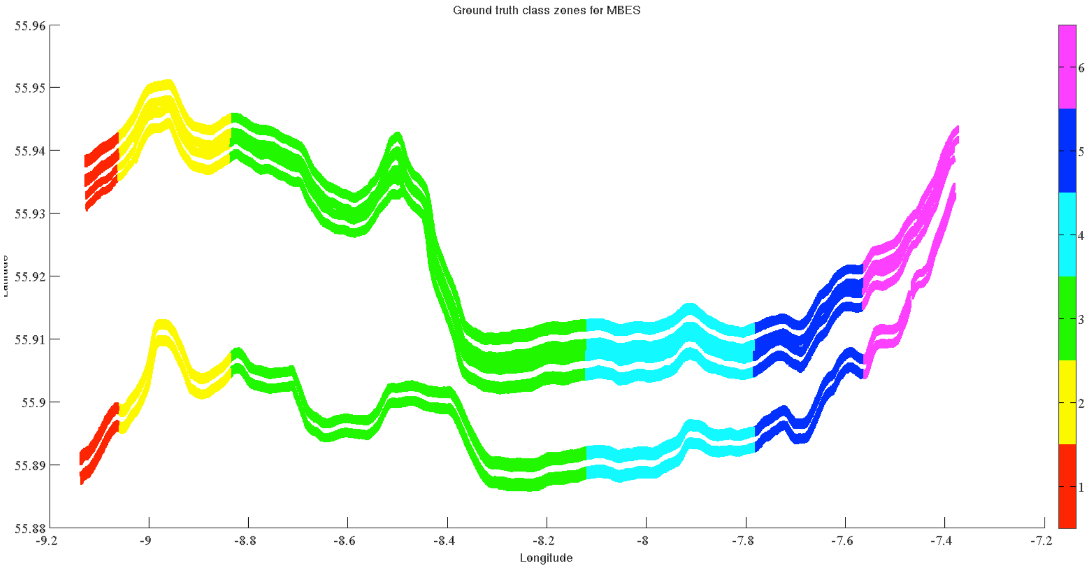
After considering the underlying properties of different zones in Table 6.1, the decision was made that SBES responses to the silty sand from zones 1 and 4, to the silty clay with sand mixture from ‘Zone 3’, and to the fine grain sand from ‘Zone 5’ are likely to be similar. Therefore, these three zones were labelled as ‘Class 1’ for SBES classification. The main argument here is the rationale behind treating of silty sand (zone 4) and fine grain sand (zone 5) similarly for cluster labelling. But without the presense of detailed grain size analysis, the backscatter responses from these two zones would be too close to distinguish.

Both zones 2 and 6 have medium grain sand and are therefore labelled as ‘Class 2’ and lastly zone 7 is the only zone with coarse grain sand. This zone is labelled as ‘Class 3’ for SBES classification. Table 6.3 lists the classification class labels and their generalised description for SBES clustering.

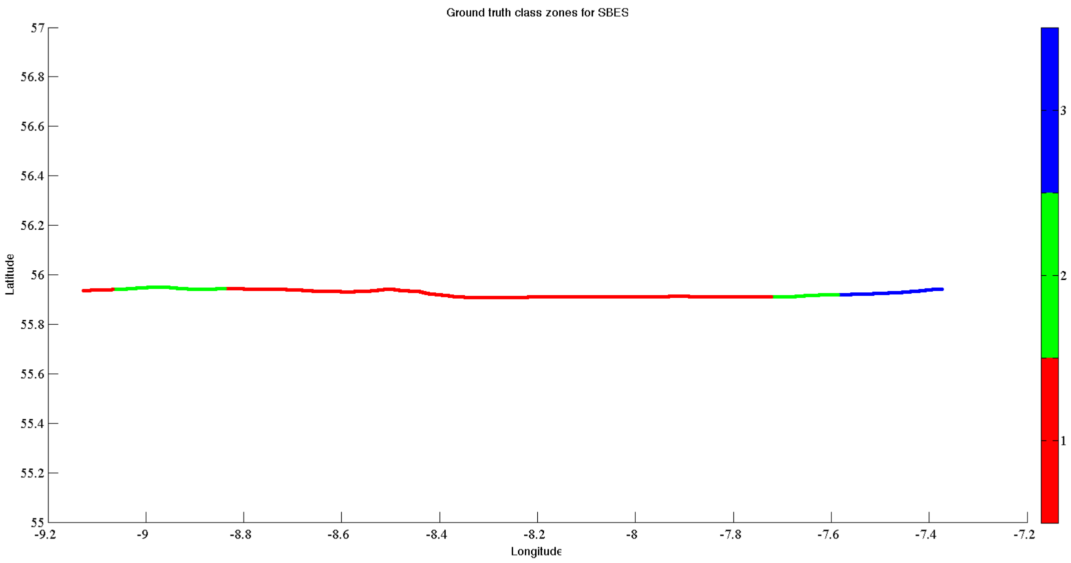
Table 6.3: Class description for SBES classification

	Class 1	Class 2	Class 3
Class value	1	2	3
Zones in class	1, 3, 4, and 5	2 and 6	7
Class title	Silty & fine grain sand	Medium grain sand	Coarse grain sand

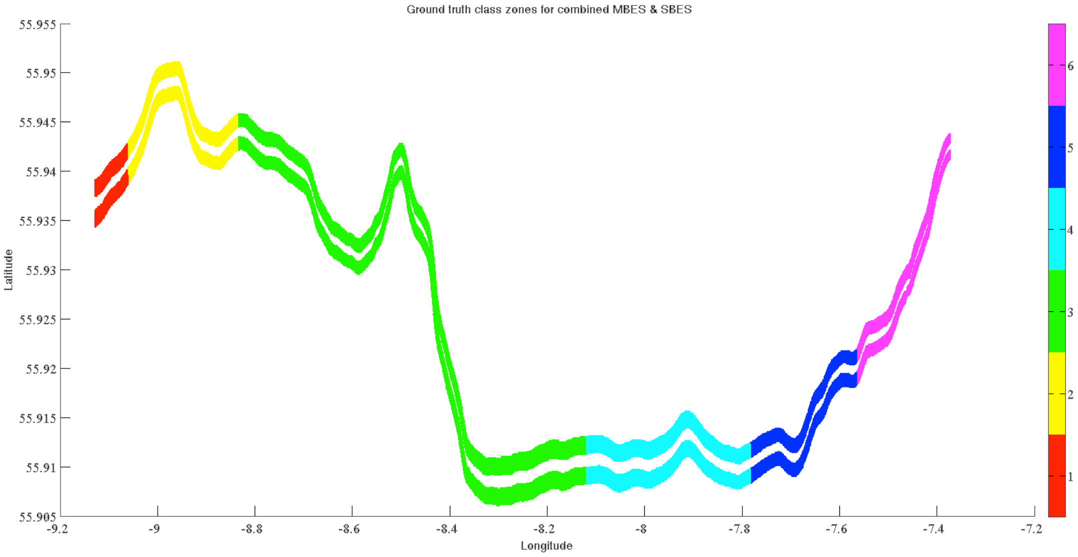
Using the above schemes (Table 6.2 and 6.3), two separated cluster validation databases were developed. One (with 6 classes) will be used to validate the MBES clusters and combined MBES and SBES clusters and the other (with 3 classes) will be used to validate the SBES clusters. Figure 6.7 shows the ground truth dataset for cluster validation.



(a)



(b)



(c)

Figure 6.7: Ground truth datasets for MBES, SBES and combined MBES + SBES cluster validation

In the following section, we show the results of combined clustering as well as he validation results.

6.4. Results and discussion

This section will include the validation results using confusion matrices and kappa coefficients for clustering results obtained in the analysis chapters 3, 4, and 5.

Validation of MBES clusters – 132 FFVs vs 35 FFVs (obtained from 16 VGs)

MBES with 132 FFVs was clustered using the defacto method- using PCA to orthogonalise the datasets and then picking the first three components which account for around 90-95% of information. Tables 6.4 and 6.5 show the validation results obtained from clustered MBES datasets (Chapter 3):

Table 6.4: Confusion matrix and accuracy of k-means clustering for 6 clusters of MBES datasets (132 FFVs)

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Row total
Class 1	**79.43%**	16.54%	1.41%	1.24%	1.16%	0.22%	17893
Class 2	2.83%	**50.49%**	42.10%	2.23%	0.38%	1.97%	48703
Class 3	4.17%	4.30%	**73.97%**	10.52%	3.82%	3.25%	137291
Class 4	3.51%	6.10%	4.75%	**82.35%**	1.68%	1.64%	70101
Class 5	2.27%	11.57%	1.77%	31.87%	**50.97%**	1.55%	46073
Class 6	1.33%	4.92%	2.53%	3.97%	40.21%	**47.04%**	39607
Column total	25350	44959	127451	89730	46225	25953	**359668**
						Total of diagonal =	**240197**
						Overall accuracy =	66.78%

Table 6.5: Confusion matrix and accuracy of k-means clustering for 6 clusters of MBES datasets (35 FFVs)

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Row total
Class 1	**86.09%**	13.44%	0.47%	0.00%	0.00%	0.00%	17893
Class 2	5.67%	**66.91%**	19.04%	7.39%	0.43%	0.57%	48803
Class 3	2.31%	2.64%	**76.76%**	9.07%	8.58%	0.64%	137291
Class 4	4.12%	1.97%	14.81%	**78.81%**	0.30%	0.00%	70101
Class 5	1.53%	11.52%	6.12%	6.82%	**74.00%**	0.00%	46073
Class 6	3.17%	9.98%	7.04%	5.73%	32.01%	**42.07%**	39607
Column total	26187	49317	130740	76722	58983	17819	**359768**
						Total of diagonal =	**259438**
						Overall accuracy =	72.11%

Classification with MBES visual groups (Table 6.5) has a higher overall accuracy than that using all 132 FFVs (Table 6.4). In the classification of MBES datasets with 132 FFVs, ‘Class 2’ which is medium grain sand (Table 6.2) appears to be poorly classified with a significant portion of it being misclassified into ‘Class 3’, which is silty sand with gravel and clasts. Mapping accuracy of this class (Table 6.6) is just 35.6% which makes it a very unreliable class in the classification map. Class accuracy is much improved when the datasets classified include visual groups. This indicates that optimised datasets produced more reliable map than the original datasets.

Table 6.6: Accuracy assessment and corresponding kappa for MBES classification

	MBES (132 FFVs)				MBES (16 VGs)			
	Omission error (%)	Comm. error (%)	Mapping Accuracy	Kappa	Omission error (%)	Comm. error (%)	Mapping Accuracy	Kappa
Class 1	20.57	62.24	48.96%		13.91	60.26	53.72%	
Class 2	49.51	41.82	35.60%	0.55	33.09	34.15	49.87%	0.63
Class 3	26.03	18.86	62.23%		23.24	18.47	64.79%	
Class 4	17.65	45.65	56.54%		21.19	30.64	60.32%	
Class 5	49.03	49.36	51.80%		25.99	54.01	48.05%	
Class 6	52.96	18.49	39.70%		57.93	2.92	40.87%	

Overall, the mapping accuracy of each class obtained from MBES data with 132 FFVs vary from 35%-62%. This indicates that the classes include a good amount of misclassification in the form of omission and commission. Omission means a portion of a class being misclassified into other classes, while commission means other classes being misclassified into that particular class. Both the omission and commission error varied from 20% to as high as 60%. This can be an indication that loss of information by selecting the first three principal components is actually contributing to the misclassification. The highest commission error can be observed for the smallest class (Class 1). Class 1 has a total of 17,893 data locations while its adjacent larger classes (Class 2 and 3) have a total of 185,994 data locations. A relatively small portion of those larger classes, a total of 7102 data locations (2.8% of Class 2 & 4% of Class 3), were misclassified as ‘Class 1’. However, this equated to almost 50% of the total data locations correctly classified as ‘Class 1’. Therefore, this high commission error does not fully indicate that the mapping quality of ‘Class 1’ is poor but shows the influence of larger classes on relatively smaller classes. ‘Class 6’, also a small class, was the most poorly classified with the highest amount of omission error. Almost half of its data locations were classified as ‘Class 5’. ‘Class 6’ contains coarse grain san with some presence of fine to coarse pebbles, while ‘Class 5’ contains medium grain san with a heavy presence of shells and burrows. The nature of particles present in these two zones makes them quite identical and could be the principal reason for heavy misclassification.

MBES with 16 visual groups were also clustered using a combination of PCA and k-means, but instead of picking the first three components, all principal components were selected for clustering thus no information was lost. The overall accuracy of the map had increased by more than 5% with kappa coefficient being 0.6383. The mapping accuracy of individual classes also increased and ranged between 40% - 64%. As in MBES with 132 FFVs, the classes 1 and 6 had the highest amount of commission and omission error respectively. The definition of ‘Class 1’ was significantly better than previous with almost all of its misclassification concentrated on the adjacent ‘Class 2’. Classes 5 and 6 were defined in a similar fashion like previous i.e. almost half of the 6th class’s data locations were misclassified as ‘Class 5’.

In comparison, it can be said that the reduced data seemed to have produced a better quality map with better class definitions than that obtained using the traditional method. It also indicates that SOM was able to optimise the MBES dataset thus contributing to the improvements.

Validation of SBES clusters – 12 KHz vs. 38 KHz and k-means vs. Fuzzy c-means

Before discussing the validation results, it should be noted that The SBES dataset available for this study contained a significant vertical displacement. As the dataset was clustered directly instead of clustering the features generated from the backscatter, this should have a direct effect on the clustering quality as similar peaks could have a different location in the respective time series due to this fluctuation and thus may have had different cluster labels instead of being labelled as members of one class. In addition to this the fact that only one survey line was available for clustering severely limits the effectiveness of the algorithms as the area that was available for clustering was long and very narrow. A higher number of survey lines would have enabled us to explore and analyse the overlap or mixture of seabed classes effectively but as this was not the case, these limitations may have contributed to the higher level of misclassification present in the results.

Tables 6.7-6.12 list validation results for SBES clustering analysis. Two frequencies of SBES dataset (12 & 38 KHz) were clustered using both PCA-k-means and fuzzy c-means (Chapter 4).

Table 6.7: Confusion matrix and accuracy of k-means clustering for 3 clusters of 12 KHz SBES dataset

	Class 1	Class 2	Class 3	Row total
Class 1	**66.28%**	28.11%	5.61%	4597
Class 2	20.70%	**63.03%**	16.27%	1420
Class 3	21.52%	20.76%	**57.72%**	790
Column total	3511	2351	945	**6807**
			Total of diagonal=	**4398**
			Overall accuracy=	64.61%

Table 6.8: Confusion matrix and accuracy of k-means clustering for 3 clusters of 38 KHz SBES dataset

	Class 1	Class 2	Class 3	Row total
Class 1	**68.91%**	21.97%	9.11%	4597
Class 2	25.07%	**63.52%**	11.41%	1420
Class 3	13.04%	26.84%	**60.13%**	790
Column total	3627	2124	1056	**6807**
			Total of diagonal =	**4545**
			Overall accuracy =	66.77%

Table 6.9: Accuracy assessment and corresponding kappa for SBES classification (k-means, 12 & 38 KHz)

	SBES (12 KHz k-means)				SBES (38 KHz k-means)			
	Omission error (%)	Comm. error (%)	Mapping Accuracy	Kappa	Omission error (%)	Comm. error (%)	Mapping Accuracy	Kappa
Class 1	33.72	10.09	60.20%	0.47	31.08	9.98	62.66%	0.50
Class 2	36.97	102.53	31.12%		36.48	86.06	34.14%	
Class 3	42.28	61.90	35.65%		39.87	73.54	34.65%	

As the optimal number of clusters from internal cluster validation was estimated to be three (Chapter 4), the cluster results were validated against the ground truth database of three classes. From the validation results, the first thing that is noticeable is that the largest class overshadows the other two small classes. As the study area comprised of mostly fine grained sand with some silty areas, it resulted in a large geographic area under a single class ('Class 1'). The largest class or 'Class 1' accounted for almost 70% of the data location while classes 2 and 3 accounted for about 20 and 10% of the data locations respectively.

In this case, the k-means performed slightly better on 38 KHz dataset than on the 12 KHz SBES dataset. However, the difference is not substantial enough to be conclusive. 'Class 2' has a significant amount of commission error (Table 6.9). This is due to the fact 'Class 1' is a very large class with almost 70% of the total dataset (4597 points) and one zone (zone 2) of 'Class 2', a very small zone in comparison, is located in between zones included in 'Class 1' (Figure 6.6 & Table 6.3). Though the overall accuracy for both datasets (12 KHz & 38 KHz) is around 65%, the mapping accuracy for the smaller classes is significantly lower. This is mainly due to the misclassification of the larger class out numbering the correctly classified smaller classes as mentioned before.

From Tables 6.7 & 6.8, we can see that the mapping accuracy of classes 2 and 3 was heavily constrained by the commission error. 'Class 2' which is surrounded by the largest class was mostly affected with 102% and 86% of commission error in 12 and 38 KHz datasets. A third of the points (1292 & 1010 data points for 12 & 38 KHz respectively) of 'Class 1' was misclassified into 'Class 2' and was larger than the number of points accurately classified (895 & 902 data points respectively) as 'Class 2'. The omission error of all three classes remains fairly similar across the different frequencies. The higher degree of commission error can be improved by recalibrating the boundary definition of the ground truth database as well as through supervised classification. Tables 6.10-6.12 display the validation results from fuzzy c-means classification of the SBES dataset.

Table 6.10: Confusion matrix and accuracy of fuzzy c-means clustering for 3 clusters of 12 KHz SBES dataset

	Class 1	Class 2	Class 3	Row total
Class 1	**61.45%**	30.73%	7.81%	4597
Class 2	24.93%	**68.87%**	6.20%	1420
Class 3	10.51%	27.09%	**62.40%**	790
Column total	3262	2605	840	**6807**
			Total of diagonal =	**4296**
			Overall accuracy=	63.11%

Table 6.11: Confusion matrix and accuracy of fuzzy c-means clustering for 3 clusters of 38 KHz SBES dataset

	Class 1	Class 2	Class 3	Row total
Class 1	**65.32%**	28.17%	6.50%	4597
Class 2	29.58%	**65.56%**	4.86%	1420
Class 3	15.57%	17.72%	**66.71%**	790
Column total	3546	2366	895	**6807**
			Total of diagonal =	**4461**
			Overall accuracy =	65.53%

Table 6.12: Accuracy assessment and corresponding kappa for SBES classification (fuzzy c-means, 12 & 38 KHz)

	SBES (12 KHz k-means)				SBES (38 KHz k-means)			
	Omission error (%)	Comm. error (%)	Mapping Accuracy	Kappa	Omission error (%)	Comm. error (%)	Mapping Accuracy	Kappa
Class 1	38.55	9.51	56.12%	0.48	34.67	11.81	58.42%	0.51
Class 2	31.13	114.58	32.10%		34.44	101.06	32.61%	
Class 3	37.59	37.59	39.85%		33.29	46.58	45.51%	

The performance of fuzzy c-means was similar to that of k-means with kappa coefficients for 12 and 38 KHz datasets being 0.4802 and 0.5086 and overall accuracies 63.11% and 65.53%. Like in k-means, the smaller classes here were also affected but the commission error contributed by ‘Class 1’. The omission error remained fairly identical across the frequencies indicating that both algorithms performed similarly on both frequencies.

Validation of combined MBES and SBES clusters

Finally, statistical features extracted from SBES were combined with the optimised MBES dataset with 35 FFVs based on nearest distance i.e. for each MBES sweep, the closest SBES location was determined and features from that location combined. The objective was to test if high resolution vertical returns from SBES could improve the classification results of MBES. Tables 6.13-6.16 display validation results obtained from clustering the combined MBES and SBES dataset.

In this case the overall mapping quality is improved from that of MBES classification (Tables 6.4-6.6). The combined PCA and k-means appear to have performed better than fuzzy c-means. Both methods have comparatively lower omission errors than that achieved with the classification of MBES with all 132 features. This means that the classes are well defined within themselves with lower misclassification.

Table 6.13: Confusion matrix and accuracy of k-means clustering for 6 clusters of combined MBES & SBES dataset

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Class 1	**2727**	25	33	17	7	14	2823
Class 2	1352	**9494**	1711	1135	1030	622	15344
Class 3	533	1725	**40454**	1632	2164	760	47268
Class 4	58	47	4601	**16016**	776	0	21498
Class 5	7	53	147	1021	**10444**	1146	12818
Class 6	109	175	6	1451	3721	**6305**	11767
Total	4786	11519	46952	21272	18142	8847	**111518**
						Total of diagonal =	**85440**
						Overall accuracy =	76.62%

Table 6.14: Confusion matrix and accuracy of fuzzy c-means clustering for 6 clusters of combined MBES & SBES dataset

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Class 1	**2760**	21	14	7	10	11	2823
Class 2	1114	**11158**	2172	257	432	211	15344
Class 3	1107	1112	**40033**	1499	1418	2099	47268
Class 4	26	1308	5754	**14093**	197	120	21498
Class 5	671	1333	475	1688	**8519**	132	12818
Class 6	261	553	284	493	5437	**4739**	11767
Total	5939	15485	48732	18037	16013	7312	**111518**
						Total of diagonal =	**81302**
						Overall accuracy =	72.90%

Table 6.15: Accuracy assessment and corresponding kappa for combined MBES & SBES classification (k-means and fuzzy c-means)

	MBES+SBES (k-means)				MBES+SBES (fuzzy c-means)			
	Omission error (%)	Comm. error (%)	Mapping Accuracy	Kappa	Omission error (%)	Comm. error (%)	Mapping Accuracy	Kappa
Class 1	3.40	72.93	57.13%		2.23	113.97	45.98%	
Class 2	38.12	13.19	54.66%		27.28	28.19	56.72%	
Class 3	14.41	13.74	75.24%	0.68	15.30	18.40	71.53%	0.63
Class 4	25.50	24.44	59.86%		34.44	18.34	55.39%	
Class 5	18.52	60.05	50.91%		33.53	58.46	41.94%	
Class 6	46.41	21.60	44.06%		59.72	21.86	33.05%	

From the validation results, we can see that the map had increased accuracy for both k-means and fuzzy c-means (76.62% & 72.90%). Performance of k-means was better than that of fuzzy classification. Apart from 'Class 6', all the other classes have reasonable omission errors indicating the points within the class are quite homogeneous. Classes 1 and 6, both smaller classes, had the highest amount of commission errors. A good portion of 'Class 2' was classified as 'Class 1', while 'Class 6' contributed heavily to the commission error of 'Class 5'. Composition of classes 5 and 6 made them quite similar hence the misclassification. 'Class 1' is the smallest class with 2823 data locations accounting for only 2.5% of the whole dataset. Though the class itself has low omission error (3.4% & 2.23% for k-means and fuzzy), misclassification in other larger classes outnumbered the total class size and reduced the map accuracy significantly. Therefore, despite 57% (k-means) and 46% (fuzzy c-means) mapping accuracy, the class would have well defined class boundary due to low omission error. The commission error can be improved through calibration of ground truth boundaries and supervised classification.

Finally, to compare the combined MBEs + SBES approach with MBES only clustering, we decided to look at the overall accuracy of MBES + SBES vs. The two MBES accuracies (132 FFVs & 35 FFVs). These overall accuracies and kappa coefficients are displayed in Figure 6.8. For the Combined data results, only k-means accuracy was selected for comparison as MBES only clustering was done only with k-means.

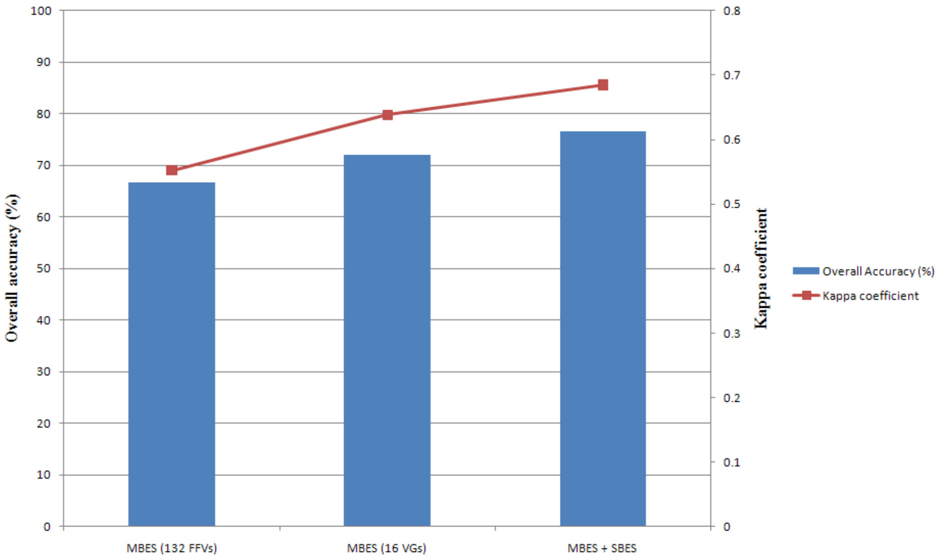


Figure 6.8: Comparison of overall accuracy and kappa coefficient for MBES and Combined MBES + SBES dataset using k-means classification

From the figure we can observe a steady improvement of mapping quality from using traditional method of seabed classification to classification of optimized datasets and finally to classification of a combined MBES and SBES dataset. The results could be further improved by using datasets covering a larger area and through optimization of ground truth databases. With a large enough dataset, different degrees of fuzziness could also be tested. Fuzzy classification with different degrees of fuzziness can provide interesting results given the homogeneous nature of the survey area. A better quality SBES dataset could also improve the overall mapping quality of both SBES and combined MBES, SBES classification.

The next chapter will summarise the studies conducted in this PhD research, based on the analysis and interpretations from chapters 3, 4, 5, and 6.

Chapter 7

Conclusion

In this last chapter we summarise the contributions of this thesis and suggest possible future research directions.

Chapter contents

7.1. Summary

The core objectives of this thesis were to provide a set of methods for acoustic data optimisation that could reduce noise by eliminating redundancy in MBES data and detect outliers in SBES data as well as to provide alternative clustering methods for seabed mapping from acoustic data. These objectives were achieved by addressing a series of four research questions outlined in the first chapter. These involved:

1. MBES data optimisation through redundancy elimination by using Self Organising Maps (SOM).
2. Detecting outliers in SBES data using TimeSearcher©- a visual exploration tool.
3. Direct clustering and classification of optimised SBES data, and
4. Classification of combined MBES and SBES data to improve mapping quality.

In this section we summarise the contributions from each chapter and comment on the results.

Chapter 2 reviewed literature from related fields of research: the development of sonar system technologies, underwater acoustics, acoustic data analysis, seabed mapping, visual analytics, data mining and clustering.

The third chapter was concerned with the redundancy issues associated with MBES data. Self Organising Maps (SOM), a well established visual analytical tool for multivariate data, was used for attribute clustering on a dataset containing statistical features of MBES backscatter from two different frequencies (0.2ms & 0.7ms). The change in frequencies is directly related to the amplitude of the emitted beam. Consequently, with all other variables remaining unchanged, the frequencies are automatically adjusted during the survey as the depth of the seabed changed in order to maintain the emitted beam amplitude a near constant. For both frequencies a total of 16 visual groups were identified among the 132 features. The groups were identical across frequencies. Each of the first 15 groups comprised of a different set of statistical features that had similar colour distribution in the SOM component planes, while the 16th group included features with dissimilar distribution from one another and to all other features. A total of 20 FFVs out of 132 were placed into the last group. As the rationale was to produce an optimal database that accounted for the majority of the information while reducing the number of variables as much as possible, only the FFVs that showed a high degree of visual similarity were placed in groups 1-15 while FFVs with any noticeable dissimilarity in colour distribution were placed in group 16. The level of similarity within each of the visual groups of FFVs was further estimated by calculating pair-wise correlation coefficients for each group. Most of the groups that exhibited a very high level of visual similarity had high correlation coefficients (0.87-0.98). Three groups showed moderate similarity with correlation coefficients around the regions of 0.5, 0.6, and 0.7 respectively. However, standard deviations of these values of these three groups were extremely low (0.0072-0). This indicated that the relationships between features in those groups were relatively constant.

Based on the visual and statistical analysis of attribute similarity, we developed an optimised MBES dataset to be used for seabed classification. This optimised MBES dataset had 35 attributes - 1 feature from each of the 15 visual groups of similar features and all 20 features from group 16, which represented a 73% reduction in data dimensionality while preserving most of the information.

Both the original data and the optimised data were subsequently clustered using the PCA + k-means approach. For the original data, the first three principal components were selected to emulate the de facto standard procedure, while for the optimised data all the components were selected for k-means clustering. The cluster results were visually compared as well as internally validated using four different internal validation methods. The optimal numbers of clusters estimated by the internal validation indices were six. Consequently the datasets (both with 132FFVs and 35FFVs) with six clusters were validated using ground truth in Chapter 6.

The following chapter, Chapter 4, discussed two novel approaches in SBES data processing and clustering – using visual exploration for outlier detection and direct clustering of time series echo returns. TimeSearcher©, a visual exploration tool of time series data was used for visually exploring the SBES dataset with the aim of outlier detection. This was done after the data were already cleaned using an automatic outlier detection procedure. Visual exploration identified further outliers the automatic procedure was not able to find. After the outliers were removed, the SBES data were clustered directly. The rationale behind direct clustering (as opposed to feature based clustering) was to provide the surveyor/geologist onboard the survey vessel a rough estimation of underlying clusters during the actual survey thus enabling them to optimise the ground truth collection locations. A good portion of the subsurface information was included in the clustering (5m below estimated seabed surface). Two algorithms, PCA + k-means and fuzzy c-means with different set of clusters were tested. Due to the noisy nature of the data, performances of both algorithms were not optimal from five clusters and up. The internal validation indices estimated the optimal number of clusters to be three. This is consistent with the assumption that the SBES time series represented the subsurface classes of the seabed. The subsurface information may have enhanced the homogeneous nature of the seabed underneath layer and thus reduced the number of inherent groupings in the dataset. The cluster results were further validated in Chapter 6.

The potential of improving seabed mapping using a combination of MBES and SBES data was discussed in Chapter 5. First, statistical features were generated from the SBES data. These were then joined with the corresponding MBES data based on identification of the closest locations between MBES and SBES. As before, two algorithms, PCA + k-means and fuzzy c-means were tested and results visualised. From visual comparison, the clusters appeared to provide better boundary definitions than when compared to the results obtained from clustering MBES data only. Between the algorithms, k-means appeared to have yielded better results than fuzzy c-means. The results seem to indicate that adding SBES did in fact improve the boundary definitions. However, two limiting factors acting here were first that the SBES data quality was less than optimal since it had a high degree of vertical displacement. Second, SBES data were available from one survey line only thus limiting the fuzzy c-means to exploit its ability to account for class overlaps, which is quite common in homogeneous geological classes. Despite these limitations, the results were encouraging and the resulting clusters were further validated in the next chapter, as with separate MBES and SBES clustering.

Cluster results from the analysis chapters were validated against ground truth data in Chapter 6. A dataset containing the details of a number of grab samples from GSI/MI survey at Malin was used as the basis for the development of ground truth database. An overview of the study area was built from the information obtained from the grab samples. The study area was divided into 7 zones based on the mixture of underlying seabed type and different types of surface objects such as shells, clasts, burrows, gravel etc. From the 7 zones, two different ground truth databases were developed: one with 6 classes for validation of MBES and combined MBES/SBES clustering and the second with 3 classes based solely on the underlying soil types (fine, medium and coarse grain) for validation of SBES classes. Clustering results were then compared to those two ground truth databases using a confusion matrix, a well-known method of cluster validation in remote sensing with corresponding accuracies and kappa coefficients. For MBES, the classes derived from optimised reduced dimension data (35 FFVs) yielded better accuracy compared to classes derived from original data with all 132 FFVs. This indicates that in spite of the higher complexity of the standard method, there is a loss of information compared to our alternative approach on using the optimised data for seabed classification. In addition, statistical analysis of the similarity of attributes confirms that SOM component planes are a suitable visual approach to identify the redundancy in acoustic features.

For SBES, the accuracy of mapping was somewhat limited due to data quality as well as the impact of one extremely large class (Class 1) on its neighbouring smaller classes (Class 2 and 3). The quality of each class (based on the omission errors) remained fairly constant across algorithms i.e. both k-means and fuzzy c-means performed similarly on both frequencies (12 & 38 KHz). Though, in comparison, the dataset from 38 KHz yielded a slightly better result, while k-means was the better performer (by a slight margin) of the two algorithms. Based on the accuracies, direct clustering of SBES data was able to provide a relatively reliable overview of the underlying classes in survey area.

The combined MBES + SBES data provided by far the best accuracy for mapping. There was almost a 10% increase in overall accuracy when compared to the results from the original MBES data (with 132 FFVs), with an increase of 4.5% in accuracy compared to the results obtained from optimised MBES data (with 35 FFVs). This demonstrates the potential of combining the high resolution SBES data with MBES.

To summarise the novel findings of this research, we gather our results into the answers to four core research questions raised in the first chapter as follows:

Is it necessary for MBES classification to produce such a large number of statistical features or could the dimensionality be kept lower by avoiding redundancy? And if so, which of the features are correlated with each other and therefore redundant? Would clustering a optimal dataset (redundancy is removed) produce better, similar or worse cluster definition to that of the cluster definition generated from 132 statistical features using the de facto clustering method?

Typical MBES data have a large amount of redundancy that can reduce the overall map quality. SOM's component planes are capable of identifying those redundancies and can help the analyst in producing optimised MBES datasets free of the redundant information. Most of the redundancies were detected in the textural and power spectrum features. This optimised MBES data increases the accuracy of mapping. This finding (35 FFVs instead of 132 FFVs) has the potential to significantly reduce the MBES data processing time which normally takes several computers weeks to process and therefore cost effective as less energy and human input would be required for data processing. With better mapping accuracies, the optimisation of MBES data also contributes to the overall output of the seabed mapping project.

Can visual analytics provide an efficient way of detecting outliers that are undetected using traditional outlier detection methods?

TimeSearcher© was able to identify 81 outliers which the automatic outlier detection algorithm failed to detect. The outliers were detected by visual exploration of the shapes of echo time series. Therefore it can be said that with large volumes of time series data, visual exploration can be an easy and effective way of detecting obvious outliers that may have slipped through the detection process of an automatic outlier detection algorithm. In doing so, visual exploration can identify potential weaknesses of the automatic algorithm used and help improve the effectiveness by addressing those weaker areas of the algorithm. This finding identifies two potential improvements in SBES data processing steps: using visual exploration as an added step to include the potential of human cognition abilities and to use as an added measure to assess the performance of automatic outlier detection algorithms.

Would direct clustering of the SBES backscatter produce representative clusters thus eliminating the dependency of generating features as well as provide a quick overview of underlying clusters in the survey area?

Direct clustering of the cleaned SBES dataset, which included subsurface information, was able to provide an overview of the underlying classes in the study area. This was a novel approach as oppose to the traditional feature based classification. This finding has the potential of having a significant impact on the use of SBES data in seabed mapping process. A quick classification of the SBES time series can enable the surveyor to better plan the ground truth data collection locations thus saving time and cost over the whole survey operation. It also has the potential of further being used in actual seabed classification provided that the data quality is optimal. If successful, SBES raw echoes can be a valuable addition to seabed mapping processes as it contains the least distorted echoes in high resolution.

Improving seabed mapping from MBES and SBES: Can the optimised data produce quality clusters that will ultimately result in better quality seabed classes? Can the sub-surface information of SBES be combined with MBES data to provide a better definition of the underlying seabed, thus avoiding the interference from plants and other particles lying on seabed?

Optimised MBES dataset was able to provide a better classification than that achieved with the original dataset (approx. 5.5% increases in overall accuracies). When the optimised MBES was combined with features from SBES, the increase in accuracy was almost 10%. This shows that adding SBES data has the potential of improving the classification results. This was a novel approach in seabed mapping as commonly data from the same sensor family is used in fusion (MBES & SSS). The rise in overall mapping accuracy was quite significant (10%) and underscores the combined potential of MBES data optimisation and subsurface information from SBES. This approach has the potential to significantly increase the seabed mapping quality without much effort as both MBES and SBES data are usually collected simultaneously in seabed surveys.

The operational cost of Celtic Explorer (the survey vessel used by the the MI & GSI) was estimated at €17,000/day in 2010 (MI, 2010). With a typical survey exploration extending up to anywhere between 3-4 weeks, the basic cost (cost of fuel, crew salary, equipment rent) can be between €357,00-€476,00. With efficient ground truth planning through direct clustering of SBES, each day less amount to a significant financial saving. If this is combined with the reduced dimension of MBES data (132 FFVs vs 35 VGs, almost 75% reduction in data size), the total saving in terms of money and manpower can be quite significant.

7.2. Future directions

In closing, we would like to suggest a few potential directions stemming from the present work. First and foremost, to firmly establish the findings in this research, the methods should be replicated on data covering a larger area with more seabed variety. The SOM component planes should be tested on a variety of seabed classes and survey areas to test if the visual groups remain unchanged (i.e. the same 16 visual groups from the same 132 FFVs) or if they need to be adjusted for each different area. Visual interpretation should also include several analysts to minimize bias. While the similarities within each visual group were confirmed by correlation analysis, the inter-group difference should also be tested to estimate the level of their separation. The geographic location of the feature vectors could also be taken into account in future studies through some measure of spatial autocorrelation measures.

Data quality, especially for SBES, is another pertinent issue for this work. The quality of the SBES data has a direct impact on the quality of the clustering results and is therefore of utmost importance. Raw SBES echo returns are noisy by nature due to their high sensitivity to suspended particles, plants, fauna, etc and care should be given while collecting these data. Avoidable errors present in the study dataset like ‘vertical displacement’, which is regarded as a gross error and is due to miscalculation of seabed depth, should not occur. In current practice, it is up to the surveyor to look out for anomalies in depth estimation. The automatic algorithm estimates the depth from time gap of echo returns. This algorithm could be further optimised with the use of a high resolution Digital Elevation Model to establish a cut off depth range. Any estimated depth outside this cut off range would be marked for further evaluation.

Clustering, combining MBES and SBES datasets produced promising results and an increase in mapping accuracy. This result could be further improved using supervised classification for which, however a good ground truth database is an important requirement. The classification should also be tested on varied seabed conditions to further establish and refine the method.

The procedure for combining MBES and SBES data used in this study was a simple one, where the closest SBES data location was added to the corresponding MBES data. This could sometimes be misleading as ‘closest’ does not necessarily mean ‘similar’. This is due to the fact that seabed classes are homogeneous and it can well be the fact that the detected closet SBES data point to an MBES data point is in fact similar to the neighbouring MBES. To counter this effect, perhaps weighted distance measures could be incorporated in the procedure in the following manner: If a seabed type map generated from MBES and corresponding ground truth data already existed for the target area, we could use this map and superimposed SBES data locations to generate a weighted cost distance surface. This cost distance surface could be used to select the ‘nearest’ and ‘similar’ SBES data location for each MBES sweep.

A successful generation of seabed maps, like any other mapping process, depends largely on the quality of collected data and available ground truth information. Given the amount and quality of data available for this study as well as limited ground truth information (i.e. sparse ground truth sampling and no grain size analysis available), the results proved to be promising in optimising the noisy acoustic data and improving the accuracy and quality of seabed mapping. Furthermore, these approaches can lead to a significant saving of time and money as surveying the sea floor is, to date, a quite expensive and time-consuming procedure, but one that is of increasing importance. With the rapid improvement of sensor quality and increasing interest of researchers, governments, explorers and investors in our oceans, the importance and demands for new improved approaches to interpret the high volume of seabed data is on the rise and the proposed approaches have the potential of delivering three major requirement of any research or exploration projects: cost effective, time saving and better quality results.

We know less about the ocean's bottom than about the moon's backside.

--Roger Revelle "Physics Today" February, 1992

Bibliography

Abello, J., & Korn, J. L. (2002). MGV: A System for Visualizing Massive Multidigraphs. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 21-38.

Abram, G., & Treinish, L. A. (1995). An extended data-flow architecture for data analysis and visualization. *Visualization'95: IEEE Conference on Visualization* (pp. 263-270).

Ahlberg, C., & Shneiderman, B. (1994). Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In B. Adelson, S. Dumais, & J. S. Olson (Eds.), *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence* (pp. 313–317). Boston, MA, USA: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=191775

Ahlberg, C., & Wistrand, E. (1995). Ivee: An information visualization and exploration environment. *International Symposium On Information Visualization* (pp. 66-73). Atalanta, GA.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. doi:10.1109/TAC.1974.1100705

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster Analysis*. Beverly Hills, CA: Sage Publications.

Allaby, M. (2009). *Oceans : a scientific history of oceans and marine life*. New York, USA: Facts on File Inc.

Alpern, B., & Carter, L. (1991). Hyperbox. *Visualization'91: IEEE Conference on Visualization* (pp. 133-139). San Diego, CA.

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics*, *28*, 125–136. JSTOR. Retrieved from http://www.jstor.org/stable/10.2307/2528964

Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M.-J., MacEachren, A., & Wrobel, S. (2007). Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science*, *21*(8), 839–857. Taylor & Francis. doi:10.1080/13658810701349011

Andrienko, Gennady, & Andrienko, N. (2009). A Practical Introduction to Geospatial Visual Analytics. Bonn, Germany. Retrieved from http://geoanalytics.net/and/

Ankerst, M., Ester, M., & Kreigel, H. P. (2000). Towards an effective cooperation of the computer and the user for classification. *KDD'2000: The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston, MA, USA.

Anupam, V., Dar, S., Leibfried, T., & Petajan, E. (1995). Dataspace: 3D visualization of large databases. *International Symposium On Information Visualization* (pp. 82-88). Atalanta, GA.

Arescon Ltd. (2001). *An Approach to seabed classification from multi-beam bathymetric sonar data*. BC, Canada. Retrieved from http://arescon.com/pics/mbeam.pdf

Aris, A, Shneiderman, B., Plaisant, C., Shmueli, G., & Jank, W. (2005). Representing Unevenly-Spaced Time Series Data for Visualization and Interactive Exploration. In M. Costabile & F. Paternò (Eds.), *Human-Computer Interaction - INTERACT 2005* (Vol. 3585, pp. 835-846). Springer Berlin / Heidelberg. doi:10.1007/11555261_66

Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing*, *6*, 128–143. doi:10.1137/0906011

Atallah, L., & Smith, P. J. P. (2003). Using wavelet analysis to classify and segment sonar signals scattered from underwater sea beds. *International Journal of Remote Sensing*, *24*(21), 4113–4128. London: Taylor & Francis, c1980-. doi:10.1080/0143116021000035012

Atallah, L., Probert Smith, P., & Bates, C. (2002). Wavelet analysis of bathymetric sidescan sonar data for the classification of seafloor sediments in Hopvågen Bay-Norway. *Marine Geophysical Researches*, *23*(5), 431–442. Springer. Retrieved from http://www.springerlink.com/index/L9024584G3R4211T.pdf

Augustin, J. M., Edy, C., Savoye, B., & Le Drezen, E. (1994). Sonar mosaic computation from multibeam echo sounder. *OCEANS "94. 'Oceans Engineering for Today's Technology and Tomorrow"s Preservation'. Proceedings*, *2*, 433-438. Brest, France.

Bass, G. F. (1972). *A history of seafaring based on underwater archeology* (p. 320). New York, USA: Walker & Company.

Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing maps as substitutes for k-means clustering. In V. Sunderam, G. Albada, P. Sloot, & J. Dongarra (Eds.), *Computational Science–ICCS 2005* (Vol. 3516, pp. 9-28). Springer Berlin Heidelberg. doi:10.1007/11428862_65

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The New S Language: A Programming Environmnent for Data Analysis and Graphics*. Pacific Grove, California, USA: Wadsworth & Brooks/Cole Advanced Books & Software.

Becker, R. A., Cleveland, W. S., & Shyu, M.-J. (1996). The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*, *5*(2), 123–155. JSTOR. Retrieved from http://www.jstor.org/stable/10.2307/1390777

Bederson, B. (1994). Pad++: Advances in multiscale interfaces. *CHI '94 : Human Factors in Computing Systems*. Boston, MA, USA.

Bederson, B. B., & Hollan, J. D. (1994). Pad++: A zooming graphical interface for exploring alternate interface physics. *UIST'94: the Seventh Annual Symposium on User Interface Software and Technology* (pp. 17-26). Marina del Ray, CA.

Berndt, D. ., & Clifford, J. (1996). Finding patterns in time series:A dynamic programming approach. In U.M Fayyad, G. Piatesky-Shapiro, P. Smyth, & R. Uthursamy (Eds.), *Advances in Knolwedge Discovery and Data Mining* (pp. 229-248). American Association for Artificial Intelligence Menlo Park, CA, USA.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: the fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191-203. doi:10.1016/0098-3004

Bianchi, G., & Sorrentino, R. (2007). *Electronic filter simulation & design* (pp. 17-20). McGraw-Hill.

Bier, E. A., Stone, M. C., Pier, K., Buxton, W., & DeRose, T. (1993). Toolglass and magic lenses: The see-through interface. *SIGGRAPH '93: Computer Graphics and Interactive Techniques* (pp. 73-80). Anaheim, CA.

Biffard, B., Bloomer, S., Chapman, R., & Preston, J. (2005). Single beam seabed classification: direct methods of classification and the problem of slope. In N.G. Pace & P. Blondel (Eds.), *Boundary Influences in High Frequency Shallow Water Acoustics* (pp. 227–232). University of Bath Press. Retrieved from http://www.questertangent.com/seabed-classification/marine-rd/papers-articles/publications/upload/docs/Singlebeamclassificationslope.pdf

Bishop, Y. M., Fienberg, S. E., Holland, P. W., Light, R. J., & Mosteller, F. (2007). *Discrete Multivariate Analysis: Theory and Practice*. New York, USA: Springer Science + Business Media.

Blondel, P. H., Parson, L. M., & Robigou, V. (1998). TexAn: textural analysis of sidescan sonar imagery and generic seafloor characterization. *OCEANS '98 Conference Proceedings* (pp. 419-23).

Blondel, Philippe. (2000). Automatic mine detection by textural analysis of COTS sidescan sonar imagery. *International Journal of Remote Sensing*, *21*(16), 3115-3128. doi:10.1080/01431160050144983

Bradley, P. S., & Fayyad, U. M. (1998). Refining Initial Points for K-Means Clustering. *ICML'98: Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 91-99). Madison, Wisconsin, USA.

Brewer, I., MacEachren, A. ., Abdo, H., Gundrum, J., & Otto, G. (2000). Collaborative geographic visualization: enabling shared understanding of environmental processes. *InfoVis'2000: IEEE Symposium on Information Visualization* (pp. 137-141). Salt Lake City, UT, USA. doi:10.1109/INFVIS.2000.885102

Brown, C J, & Blondel, P. (2008). Developments in the application of multibeam sonar backscatter for seafloor habitat mapping. *Applied Acoustics. In press, corrected proof.* doi:10.1016/j.apacoust.

Brown, C.J., & Blondel, P. (2009). Developments in the application of multibeam sonar backscatter for seafloor habitat mapping. *Applied Acoustics*, *70*(10), 1242–1247. Elsevier. doi:10.1016/j.apacoust.2008.08.004

Brown, C., Limpenny, D. S., & Meadows, W. (2002). *Guidelines for the conduct of benthic studies at aggregate dredging sites* (No. 02DPL001). Essex, UK.

Brown, C.J., Chapman, R., Coggan, R., Kieser, R., Michaels, W. L., Orlowski, A., Preston, J., et al. (2007). *Acoustic Seabed Classification of Marine Physical and Biological Landscapes*. Denmark.

Buckingham, M. J. (2000). Wave propagation, stress relaxation, and grain-to-grain shearing in saturated, unconsolidated marine sediments. *Journal of Acoustic Society of America*, *108*(6), 2796-2815. doi:10.1121/1.1322018

Buja, A., & Cook, D. (1996). Interactive high-dimensional data visualization. *Journal of Computational and Graphical*, *5*(1), 78-99. Retrieved from http://www.jstor.org/stable/10.2307/1390754

Bull, L., Studley, M., Bagnall, A., & Whittley, I. (2007). Learning classifier system ensembles with rule-sharing. *IEEE Transactions on Evolutionary Computation*, *11*(4), 496-502. doi:10.1109/TEVC.2006.885163

Buono, P, Plaisant, C., Simeone, A., Aris, A., Shneiderman, B., Shmueli, G., & Jank, W. (2007). Similarity-Based Forecasting with Simultaneous Previews: A River Plot Interface for Time Series Forecasting. *IV'07: 11th International Conference on Information Visualisation*. Zürich, Switzerland.

Buono, Paolo, Aris, A., Plaisant, C., Khella, A., & Shneiderman, B. (2005). Interactive Pattern Search in Time Series. *VDA'05: Conference on Visualization and Data Analysis, SPIE*. California, USA.

Burrough, P. A., van Gaans, P. F. M., & MacMillan, R. A. (2000). High -resolution landform classification using fuzzy k-means. *Fuzzy Sets and Systems*, *113*, 37-52.

Butterworth, S. (1930). On the theory of filter amplifiers. *Experimental wireless & the wireless engineer*, *7*, 536-541.

Calinski, R. B., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*, 1-27.

Campbell, J. B. (2008). *Introduction to Remote Sensing* (4th ed.). New York, USA: The Guilford Press.

Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann.

Carlis, J. ., & Konstar, J. A. (1998). Interactive visualization of serial periodic data. *ACM Symposium on User Interface Software and Technology* (pp. 29-38). San Francisco, CA, USA: ACM Press.

Carmichael, D., Linnett, L., Clarke, S., & Calder, B. (1996). Seabed classification through multifractal analysis of sidescan sonar imagery. *IEE PROCEEDINGS RADAR SONAR AND NAVIGATION*, *143*(3), 140–148. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.54.6257&amp;rep=rep1&amp;type=pdf

Carr, D., Wegman, E., & Luo, Q. (1996). *Explorn: Design considerations past and present*. *Center for Computational Statistics, George Mason University*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Explorn:+Design+considerations+past+and+present#0

Centre(WHSC), W. H. S. (2011). WHSC bathymetry systems: single beam bathymetry systems. Retrieved from

Chivers, R. C., Emerson, N., & Burns, D. R. (1990). New acoustic processing for underway surveying. *Hydrological Journal*, *56*, 9-17.

Chotiros, N. P. (1995). Biot Model of Sound Propagation in Water Saturated Sand. *Journal of Acoustic Society of America*, *97*(1), 199-214. doi:10.1121/1.412304

Collins, W.T., & Preston, J. M. (2002). Multibeam seabed classification. *International Ocea*, *6*(4), 12-15. Retrieved from http://www.questertangent.com/seabed-classification/marine-rd/papers-articles/publications/upload/docs/iosd02.pdf

Company(SSC), S. S. (1907). Submarine Signals. *Cornell University Library Archives*. Retrieved from

Craven, M. W., & Shavlik, J. W. (1991). Visualizing learning and computation in artificial neural networks. *International Journal on Artificial Intelligence Tools*, *1*, 399-425.

Cutter, J. G. R., Rzhanov, Y., & Mayer, L. A. (2003). Automated segmentation of seafloor bathymetry from multibeam echosounder data using local Fourier histogram texture features . *Journal of Experimental Marine Biology and*

Ecology 285-286, 355-370., *285-286*, 355-370. doi:10.1016/S0022-0981(02)00537-3

Danish hydraulic Institute (DHI). (n.d.). No Title. Retrieved August 29, 2011, from . http://www.dhi.dk/

Davies, D. L., & Bouldin, D. W. (1979). A cluster seperation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*, 224-227.

Deboeck, G., & Kohonen, T. (Eds.). (1998). *Visual Explorations in Finance: with Self-Organizing Maps*. Springer London.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society*, *39*, 1-38.

Ding, C., & He, X. (2004). K-means Clustering via Principal Component Analysis. *ICML'04: International Conference on Machine Learning* (pp. 225-232).

Ding, L. (1997). Direct laboratory measurement of forward scattering by individual fish. *The Journal of the Acoustical Society of America*, *101*(6), 3398. doi:10.1121/1.419375

Discovery of sound in the sea (DOSITS). (2011). History of underwater acoustics. Retrieved from

Dresselhaus, M. S. (2004). Arthur Robert von Hippel: Obituary. *Physics Today*, *57*(9), 76-77.

Dunn, J. C. (1974). Well seperated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*, 95-104.

Dyer, D. S. (1990). A dataflow toolkit for visualization. *IEEE computer graphics and applications*, *10*(4), 60-69.

EM 2040 Multibeam echo sounder True wide band high resolution multibeam echo sounder. (n.d.).

Eastwood, P. D., Mills, C. M., Aldridge, J. N., Houghton, C. A., & Rogers, S. I. (2007). Human activities in UK offshore waters: an assessment of direct, physical pressure on the seabed. *ICES Journal of Marine Science*, *64*, 453-463.

Eick, S. (1994). Data visualization sliders. *Proceedings of the 7th annual ACM symposium on User Interface Software and Technology* (pp. 119-120). Retrieved from http://dl.acm.org/citation.cfm?id=192472

Ellingsen, K. E., Gray, J. S., & Bjørnbom, E. (2002). Acoustic classification of seabed habitats using the QTC VIEW^{TM} system. *ICES Journal of Marine Science*, *59*(4), 825-835. doi:10.1006/jmsc.2002.1198

Everitt, B. S., Landau, S., Leese, M., & Daniel, S. (2011). *Cluster analysis* (5th ed.). John Wiley & Sons.

Faloutsos, C., Ranganathan, M., & Manolopoulos, Y. (1994). Fast Subsequence Matching in Time-Series Databases. *SIGMOD '94: ACM International Conference on Management of Data* (pp. 419-429). Minneapolis, USA: ACM Press.

Farebrother, R. W. (1999). *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900*. New York, USA: Springer.

Fishkin, K., & Stone, M. C. (1995). Enhanced dynamic queries via movable filters. *CHI '95: Human Factors in Computing Systems* (pp. 415-420). Denver, CO.

Flatté, S. (1979). *Sound in a fluctuating ocean*. Cambridge: Cambridge University Press.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York, USA: John Wiley & Sons.

Foulser, D. (1995). IRIS Explorer: A framework for investigation. *SIGGRAPH'95: ACM Computer Graphics*, *29*(2), 13–16. ACM. Retrieved from http://dl.acm.org/citation.cfm?id=204365

Friendly, M. (2005). Milestones in the history of data visualization: A case study in statistical historiography. In C. Weihs & W. Gaul (Eds.), *Handbook of Computational Statistics: Data Visualization* (pp. 34–52). Springer. Retrieved from http://www.springerlink.com/index/u2q17246n4584868.pdf

Friendly, M. (2008). A brief history of data visualization. *Handbook of data visualization*, 15–56. Heidelberg, Germany: Springer. Retrieved from http://www.springerlink.com/index/n08061427483537n.pdf

Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, *41*(2), 103–130. Wiley Online Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/jhbs.20078/abstract

Friis, H. R. (1974). Statistical cartography in the United States prior to 1870 and the role of Joseph CG Kennedy and the US Census Office. *Cartography and Geographic Information Science*, *1*(2), 131–157. Cartography and Geographic Information Society. Retrieved from http://www.ingentaconnect.com/content/cagis/cagis/1974/00000001/00000002/art00004

Funkhouser, H. G. (1936). A note on a tenth century graph. *Osiris*, *1*, 260–262. JSTOR. Retrieved from http://www.jstor.org/stable/10.2307/301609

Funkhouser, H. G. (1937). Historical development of the graphical representation of statistical data. *Osiris*, *3*(1), 269–404. JSTOR. Retrieved from http://www.jstor.org/stable/10.2307/301591

Furnas, G. (1986). Generalized fisheye views. *CHI '86 : Human Factors in Computing Systems* (pp. 18-23). Boston, MA, USA.

Galton, A. P. (2003). Desiderata for a spatio-temporal geo-ontology. In W. Kuhn, M. F. Worboys, & S. Timpf (Eds.), *Spatial information theory: Foundations of geographic information science* (pp. 1-12).

Geo Seabed Instruments AS(GeoSI). (2011). Van veen grab. Retrieved from

Goff, J. A., Kraft, B. J., Mayers, L. A., Schock, S. G., Sommerfield, C. K., Olson, H. C., Gulick, S. P. S., et al. (2004). Seabed characterization on the New Jersey middle and outer shelf: correlatability and spatial variability of seafloor sediment properties. *Marine Geology*, *209*(1-4), 147-172.

Goldstein, J., & Roth, S. F. (1994). Using aggregation and dynamic queries for exploring large data sets. *CHI '94 : Human Factors in Computing Systems* (pp. 23-29). Boston, MA, USA.

Goodchild, M., Yuan, M., & Cova, T. (2007). Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, *21*(3), 239-260. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/13658810600965271

Gorban, A. N., Kegl, B., & Wunsch, D. C. (2007). *Principal manifolds for data visualization and dimension reduction*. (A. Gorban, B. Kegl, D. Wunsch, & Zinovyev, Eds.) (Vol. 58). Springer Verlag. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Principal+Manifolds+for+Data+Visualisation+and+Dimension+Reduction#0

Guo, D., Gahegan, M., MacEachren, A., & Zhou, B. (2005). Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science*, *2*, 113-132. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2786224/

Hald, A. (1990). *A History of Probability and Statistics and their Application before 1750*. New York, USA: John Wiley & Sons.

Halkidi, M., Bastistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information System*, *17*, 107-145.

Hamerly, G., & Elkan, C. (2002). Alternatives to the k-means algorithm that find better clusterings. *CIKM'02: Proceedings of the eleventh international conference on Information and knowledge management*. New York, USA. doi:10.1145/584792.584890

Hamilton, L. (2001). *Acoustic seabed classification systems* (No. DSTO-TN-0401). Victoria, Australia. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&amp;metadataPrefix=html&amp;identifier=ADA398417

Hamilton, L. J., Mulhearn, P. J., & Poeckert, R. (1999). Comparison of RoxAnn and QTC VIEW acoustic bottom classification system performance for the Cairns area, Great Barrier Reef, Australia. *Continental Shelf Research*, *16*, 1577-1597.

Hao, M., Dayal, U., Keim, D., & Schreck, T. (2007). Multi-Resolution Techniques for Visual Exploration of Large Time-Series Data. In T. MÖller, A. Ynnerman, & K. Museth (Eds.), *Eurovis 2007: Eurogrpahics/ IEEE-VGTC Symposium on Visualization* (pp. 1-8). Norrköping, Sweden.

Haralick, R. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, *67*(5), 786-804. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01455597

Haralick, R., & Shanmugam, K. (1973). Textural features for image classification. *Systems, Man and*, *3*(6), 610-621. doi:10.1109/TSMC.1973.4309314

Harris, R. L. (1999). *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press: New York.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data mining, Inference and Prediction* (2nd ed.). Springer.

Havre, S., Hetzler, E., Whitney, P., & Nowell, L. (2002). Themeriver: Visualizing thematic changes in large document collections. *Transactions on Visualization and Computer Graphics*, *8*(1), 9-20.

Haykin, S. (1994). *Communication Systems*. New York, USA: John Wiley and Sons.

Hellequin, L. (1998). Statistical characterization of multibeam echosounder data. *OCEANS'98 Conference Proceedings* (Vol. 1, pp. 228–233). Nice, France: IEEE. doi:10.1109/OCEANS.1998.725742

Hellequin, L., Boucher, J. M., & Lurton, X. (2003). Processing of high-frequency multibeam echosounder data for seafloor characterization, *28*, 78-89.

Hernandez, T. (2007). Enhancing retail location decision support: The development and application of geovisualization. *Journal of Retailing and Consumer Services*, *14*, 249-258. Retrieved from http://www.sciencedirect.com/science/article/pii/S0969698906000452

Hersey, J. B. (1977). A chronicle of man's use of ocean acoustics. *Oceans*, *20*, 8-21.

Hochheiser, H, & Shneiderman, B. (2004). Dynamic Query Tools for Time Series Data Sets, Timebox Widgets for Interactive Exploration. *Information Visualization*, *3*(1), 1-18.

Hochheiser, H, & Shneiderman, B. (2001). Visual Specification of Queries for Finding Patterns in Time-Series Data. *Proceedings of Discovery Science* (pp. 441-446). Washington, DC, USA.

Hochheiser, Harry. (2003). *Interactive Graphical Querying of Time Series and Linear Sequence Data Sets*. University of Maryland.

Hodges, R. P. (2010). *Underwater Acoustics: Analysis, Design and Performance of Sonar*. Weiley-Blackwell, UK.

Hoff, H. E, & Geddes, L. A. (1959). Graphic recording before Carl Ludwig: An historical summary. *Archives Internationales d'Histoire des Sciences*, *12*, 3-25.

Hoff, H.E., & Geddes, L. A. (1962). The beginnings of graphic recording. *Isis*, *53*(3), 287–324. JSTOR. Retrieved from http://www.jstor.org/stable/10.2307/227784

Holt, L. E. (1947). The German use of sonic listening. *Journal of the Acoustic Society of America*, *19*(4), 678-681. doi:10.1121/1.1916537

Huges Clarke, J. E., Danforth, B. W., & Valentine, P. (1997). No Title. *High Frequency Acoustics in Shallow Water* (pp. 243-250). Lerici, Italy. Retrieved from http://www.omg.unb.ca/omg/papers/HFSW_1997_Lerici.pdf

Hughes Clarke, J. E., Danforth, B. W., & Valentine, P. (1997). Areal Seabed Classification using Backscatter Angular Response at 95kHz Areal Seabed Classification using Backscatter Angular Response at. *Response*, (1).

Hydrographic Office Marine Department (HOMD). (2011). Single-beam echo-sounding principal. Retrieved from

Inselberg, A. (1997). Multidimensional Detective. *InfoVis'1997: IEEE Symposium on Information Visualization 1997* (pp. 100-107).

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice Hall.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*, 264-323.

Johnson, B., & Shneiderman, B. (1991). Treemaps: A space-filling approach to the visualization of hierarchical information. *Proceedings of the Visualization '91 Conference* (pp. 284–291).

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York, USA: Springer.

KC Denmark. (2011). KC Day grab. Retrieved from

Kaufman, L., & Rousseeuw, P. (2005). *Finding groups in data: an introduction to cluster analysis*. Wiley.

Keim, Daniel A, Panse, C., & Sips, M. (2003). Visual Data Mining of Large Spatial Data Sets, 201-215.

Keim, D. a. (1996). Pixel-Oriented Visualization Techniques for Exploring Very Large Data Bases. *Journal of Computational and Graphical Statistics*, *5*(1), 58. doi:10.2307/1390753

Keim, D. a. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 1-8. doi:10.1109/2945.981847

Keim, Daniel, Andrienko, G., Fekete, J.-daniel, & Carsten, G. (2008). Visual Analytics : Definition , Process , and Challenges. In A. Kerren, J. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization* (Vol. 4950, pp. 154-175). doi:10.1007/978-3-540-70956-5_7

Keim, D.A. (2001). Visual exploration of large data sets. *Communications of the ACM*, *44*(8), 38–44. ACM. Retrieved from http://dl.acm.org/citation.cfm?id=381656

Keogh, E., Hochheiser, H., & Shneiderman, B. (2002). An Augmented Visual Query Mechanism for Finding Patterns in Time Series Data. In J. Carbonell, J. Siekmann, T. Andreasen, H. Christiansen, A. Motro, & H. Legind Larsen (Eds.), *Flexible Query Answering Systems* (Vol. 2522, pp. 240-250). Springer-Verlag London, UK. doi:10.1007/3-540-36109-X_19

Kerneis, D., & Zerr, B. (2005). Multisensor fusion for seabed classification. *OCEANS, 2005. Proceedings of MTS/IEEE* (pp. 815–820). IEEE. doi:10.1109/OCEANS.2005.1639853

Kimura, K. (1929). On the detection of fish-groups by an acoustic method. *Journal of the Imperial Fisheries Insitute*, *24*, 41-45.

Klein, J. L. (1997). *Statistical Visions in Time: A History of Time Series Analysis, 1662-1938*. Cambridge, UK: Cambridge University Press.

Kohonen, T. (1989). *Self-Organization and Associative Memory* (3rd ed.). Springer Verlag: Berlin.

Kohonen, T. (2000). *Self-Organizing Maps* (3rd ed.). Berlin, Germany.

Kongsberg, T., & Ea, M. (n.d.). EA 400 / 600 Seabed Classification Software Realtime Seabed Sediment Classification made easy. *Geographical*.

Koua, E. L., & Kraak, M.-J. (2004). Alternative visualization of large geospatial datasets. *Cartographic Journal, The*, *41*(3), 217–228. Maney Publishing. doi:10.1179/000870404X13283

Kreuseler, M., Lopez, N., & Schumann, H. (2000). A scalable framework for information visualization. *Information Visualization, 2000. InfoVis 2000. IEEE*

Symposium on (pp. 27–36). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=885088

Kreuseler, Matthias, López, N., & Schumann, H. (2000). A scalable framework for information visualization. *InfoVis'2000: IEEE Symposium on Information Visualization* (pp. 27-36). doi:10.1109/INFVIS.2000.885088

Laboratory of Computer and Information Science (LCIS). (2011). SOM Toolbox documentation. Retrieved from

Lagoudakis, M. G., & Parr, R. (2003). Reinforced learning as classification: Leveraging modern classifiers. *ICML'03: 20th International Conference on Machine Learning*. Washington DC, USA.

Lamping, J., R., R., & Pirolli, P. (1995). A focus + context technique based on hyperbolic geometry for visualizing large hierarchies. *CHI '95: Human Factors in Computing Systems* (pp. 401-408).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. doi:10.2307/2529310

Lang, K. J., & Witbrock, M. J. (1988). Learning to tell two spirals apart. *Proceedings of the 1988 Connectionist Models Summer School* (pp. 52-59).

Lasky, M. (1974). A historical review of underwater acoustic technology 1916-1939 with emphasis on undersea warfare. *US Navy Journal of Undersea Acoustics*, *24*(4), 559-601.

Lasky, M. (1975). Historical review of underwater acoustic technology 1939-1945 with emphasis on undersea warfare. *US Navy Journal of Undersea Acoustics*, *25*(4), 885-918.

Lasky, M. (1977). Review of undersea acoustics to 1950. *The Journal of the Acoustical Society of America*, *61*(2), 283. doi:10.1121/1.381321

Le Gall, J. (1993). Analysis and simulation of side scan sonar image texture. *The Institute of Acoustics Conference on Acoustic Classification and Mapping of the Seabed* (pp. 75-79). Bath, UK.

Lichte, H. (1919). Uber den Einfluß horizontaler Temperaturschichtung des Seewassers auf die Reichweite von Unterwasserschallsignalen.Wittenborn, A.F. "On the influence of horizontal temperature layers in sea water on the range of underwater sound signals." *Physikalische Zeitschrift*, *17*, 385-389.

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. (2004). *Remote Sensing and Image Interpretation* (5th ed.). New Jersey, USA: John Wiley & Sons.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographical information systems and science* (2nd ed.). Chichester: Wiley.

Lubniewski, Z., & Pouliquen, E. (2004). Sensitivity of echo parameters to seafloor properties and depth variability. *ECUA'04:Seventh European Conference on Underwater Acoustics*. Delft, The Netherlands.

Lucieer, VL. (2005). APPLYING DISCRIMINATE ANALYSIS TO CHARACTERISE SHALLOW ROCKY REEF HABITAT. *Proceedings of International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion*. Beijing, China. Retrieved from http://www.isprs.org/proceedings/XXXVI/2-W25/source/APPLYING DISCRIMINATE ANALYSIS TO CHARACTERISE SHALLOW ROCKY HABITAT.pdf

Lucieer, VL. (2008). Object-oriented classification of sidescan sonar data for mapping benthic marine habitats. *International Journal of Remote Sensing*, *29*(3), 905–921. Taylor & Francis, Inc. doi:10.1080/01431160701311309

Lucieer, V., & Lucieer, A. (2009). Fuzzy clustering for seafloor classification. *Marine Geology*, *264*(3-4), 230-241. doi:10.1016/j.margeo.2009.06.006.

Lucieer, Vanessa, & Lamarche, G. (2011). Unsupervised fuzzy classification and object-based image analysis of multibeam data to map deep water substrates, Cook Strait, New Zealand. *Continental Shelf Research*, *31*(11), 1236-1247. Elsevier. doi:10.1016/j.csr.2011.04.016

Lundblad, E., Wright, D., Miller, J., Larkin, E., Rinehart, R., & Naar, D. (2006). A benthic terrain classification scheme for American Samoa. *Marine Geodesy*, *29*, 89–111.

Lurton, Xavier. (2002). *An introduction to underwater acoustics: Principles and applications*. Berlin-Heidelberg, Germany: Springer Verlag.

MacEachren, Alan M. (1995). *How Maps Work: Representation, Visualization, and Design*. New York: The Guilford Press.

Mackinlay, J. D., Robertson, G. G., & Card, S. K. (1991). The perspective wall: Detail and context smoothly integrated. *CHI '91 : Human Factors in Computing Systems* (pp. 173-179). New Orleans, LA.

Macqueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. *5th Berkley Symposium on Mathematical Statistics and Probability* (pp. 281-297).

Mapping European Seabed Habitats: MESH. (2011). Data collection: Ground truthing factors. Retrieved October 12, 2011, from http://www.searchmesh.net/default.aspx?page=1877

Marage, J.-P., & Mori, Y. (2010). *Sonars and underwater acoustics*. UK: Weiley-Blackwell.

Marc, W., Marc, A., & Wolfgang, M. (2001). Visualizing time-series on spirals. *INFOVIS'01: IEEE Symposium on Information Visualization* (pp. 7-13). doi:10.1109/INFVIS.2001.963273

Marcus, M., & Minc, H. (1988). *Introduction to Linear Algebra*. New York, USA: Dover Publications.

Marsh, I., & Brown, C. (2009). Neural network classification of multibeam backscatter and bathymetry data from Stanton Bank (Area IV). *Applied Acoustics*, *70*(10), 1269-1276. doi:10.1016/j.apacoust.2008.07.012

Mather, P., & Koch, M. (2010). *Computer Processing of Remotely-Sensed Images: An Introduction* (4th ed.). Oxford, UK: John Wiley and Sons.

Mayer, L. A. (2006). Frontiers in seafloor mapping and visualization. *Marine Geophysical Researches*, *27*, 7-17.

Mcgonigle, C., Brown, C., Quinn, R., & Grabowski, J. (2009). Evaluation of image-based multibeam sonar backscatter classification for benthic habitat discrimination and mapping at Stanton Banks, UK. *Estuarine, Coastal and Shelf Science*, *81*(3), 423-437. Elsevier Ltd. doi:10.1016/j.ecss.2008.11.017

Metropolitan Museum of Art (MMA). (2011). Tomb of Merketre, Thebes. Retrieved from

Milvang, O., Huseby, R. B., Weisteen, K., & Solberg, A. (1993). Feature extraction from backscatter sonar data. *The Institute of Acoustics Conference on Acoustic Classification and Mapping of the Seabed* (pp. 157-163). Bath, UK.

Monmonier, M. (1990). Strategies for the visualization of geographic time-series data. *cartographica*, *27*(1), 30-45.

Motao, H., Guojun, Z., Yongzhong, O., & Yanchun, L. (2002). Data fusion technique for multibeam echosoundings. *Geo-spatial Information Science*, *5*(3), 11-18. doi:10.1007/BF02826383

Muller, W., & Schumann, H. (2003). Visualization methods for time-dependent data-an overview. *Simulation Conference, 2003. Proceedings of the 2003 Winter* (Vol. 1, pp. 737–745). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1261490

Munzner, T., & Burchard, P. (1995). Visualizing the structure of the world wide web in 3D hyperbolic space. *VRML'95: Symposium on Virtual Reality Modeling Language* (pp. 33-8). San Diego, CA.

Namorato, M. (2000). A concise history of acoustics in warfare. *Applied Acoustics*, *59*(2), 101-135. doi:10.1016/S0003-682X(99)00021-3

Narayanan, R., Sohn, G., Kim, H. B., & Miller, J. R. (2011). Soft classification of mixed seabed objects based on fuzzy clustering analysis using airborne LIDAR bathymetry data. *Journal of Applied Remote Sensing*, *5*. doi:10.1117/1.3595267

National Defense Research Committee (NDRC). (1946). *A survey of subsurface warfare in World War II.*

Oberg, K. A., & Schmidt, A. R. (1994). *Measurements of leakage from Lake Michigan through three control structures near Chicago, Illinois, April–October 1993* (No. 94-4112) (p. 48).

Oresme, N. (1968). *Nicole Oresme and the Medieval Geometry of Qualities and Motions: A Treatise on the Uniformity and Difformity Known as Tractatus de Configrationibus Qualitatum et Motuum*. Madison, WI, USA: University of Wisconsin Press.

Orlowski, A. (1984). Application of multiple echoes energy measurements for evaluation of sea bottom type. *Oceanologia*, *19*, 61-78.

Ozkan, I., & Turksen, I. B. (2007). Upper and Lower Valued for the Level of Fuzziness in FCM. In P. P. Wang, D. Ruan, & E. E. Kerre (Eds.), *in Fuzzy Logic, A Spectrum of Theoretical and Practical Issues Series: Studies in Fuzziness and Soft Computing* (Vol. 215, pp. 85-105). Springer.

Pace, N G, & Gao, H. (1988). Swathe seabed classification. *IEEE Journal of Oceanic Engineering*, *13*(2), 83-90.

Pearson, E. S. (1978). *The History of Statistics in the 17th and 18th Centuries Against the Changing Background of Intellectual, Scientific and Religeous Thought*. London, UK: Griffin & Co. Ltd.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, *2*(6), 559–572.

Perlin,] K., & Fox, D. (1993). Pad: An alternative approach to the computer interface. *SIGGRAPH '93: Computer Graphics and Interactive Techniques* (pp. 57-64). Anaheim, CA.

Peter, H., McLoone, S., & Monteys, X. (2010). Seabed type clustering using single-beam echo sounder time series data. *6th WSEAS International Conference on REMOTE SENSING* (pp. 308-315).

Porter, T. M. (1986). *The Rise of Statistical Thinking 1820–1900*. Princeton, NJ, USA: Princeton University Press.

Pouliquen, E. (2004). Depth dependence correction for normal incidence echosounding. *ECUA'04:Seventh European Conference on Underwater Acoustics*. Delft, The Netherlands.

Powsner, S., & Tufte, E. (1994). Graphical summary of patient status. *The Lancet*, *344*, 386-389.

Press, C. (2011). Sonar Research and Naval Warfare 1914-1954 : A Case Study of a Twentieth-Century Establishment Science Author (s): Willem D . Hackmann Source : Historical Studies in the Physical and Biological Sciences , Vol . 16 , No . 1 (1986), pp . 83- Published b. *Historical Studies*, *16*(1), 83- 110.

Preston, J. (2009). Automated acoustic seabed classification of multibeam images of Stanton Banks. *Applied Acoustics*, *70*(10), 1277–1287. Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/S0003682X08001801

Preston, J M, Christney, A. C., & Collins, W. T. (2004). Automated acoustic classification of sidescan images. *In Proc. Oceans '04 MTS/IEEE, Kobe, Japan*, -.

Preston, J M, Christney, A. C., Beran, L. S., & Collins, W. T. (2004). Statistical seabed segmentation - from images and echoes to objective clustering. *Proc. 7 European Conf. On Underwater Acoustics, pp*, 813-816.

Preston, J M, Christney, A. C., Collins, W. T., Corporation, Q. T., & Road, W. S. (2003). Comparisons of acoustic classifications with multibeam, sidescan, and single-beam sonars. *Proc. Shallow Survey Conf* (pp. 8-10). Retrieved from http://www.questertangent.com/upload/docs/comparison.pdf

Preston, J M, Parrott, D. R., Collins, W. T., & John, S. (2003). Sediment classification based on repetitive multibeam bathymetry surveys of an offshore disposal site. *Oceans 2003. Celebrating the Past ... Teaming Toward the Future (IEEE Cat. No.03CH37492)*, 69-75 Vol.1. Ieee. doi:10.1109/OCEANS.2003.178523

Preston, J M, Rosenberger, A., & Collins, W. T. (2000). Bottom classification in very shallow water. *Oceans '00*.

Preston, JM, Christney, A., Bloomer, S., & Beaudet, I. (2001). Seabed classification of multibeam sonar images. *Oceans, 2001. MTS/IEEE Conference and Exhibition* (Vol. 4, pp. 2616–2623). Honolulu, USA: IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=968411

Quester Tangent Corporation(QTC). (1997). *Operator's Manual & Reference: Seabed Classification*. BC, Canada.

Rafiei, D., & Mendelzon, A. (2000). Querying time series data based on similarity. *IEEE Transactions on Knowledge and Data Engineering*, *12*(5), 675-693.

Rao, R., & Card, S. K. (1994). The table lens: Merging graphical and symbolic representation in an interactive focus+context visualization for tabular information. *CHI '94 : Human Factors in Computing Systems* (pp. 318-322). Boston, MA, USA.

Reed, T., & Hussong, D. (1989). Digital image processing techniques for enhancement and classification of SeaMARC II side scan sonar imagery. *Journal of Geophysical Research*, *94*(B6), 7469-7490.

Rekik, A., Zribi, M., Benjelloun, M., & Hamida, A. B. (2006). A k-Means Clustering Algorithm Initialization for Unsupervised Statistical Satellite Image Segmentation. *2006 1ST IEEE International Conference on E-Learning in Industrial Electronics*, 11-16. Ieee. doi:10.1109/ICELIE.2006.347204

Riddell, R. C. (1980). Parameter disposition in pre-Newtonain planetary theories. *Archives Hist. Exact Science*, *23*, 87-157.

Robinson, A. H. (1982). *Early Thematic Mapping in the History of Cartography*. Chicago, USA: University of Chicago Press.

Royston, E. (1956). Studies in the History of Probability and Statistics: III. A Note on the History of the Graphical Presentation of Data. *Biometrika*, *43*(3/4), 241–247. JSTOR. Retrieved from http://www.jstor.org/stable/10.2307/2332903

S, H. H., Siebes, A., & Wilhelm, A. (2000). Visualizing association rules with interactive mosaic plots. *KDD'2000: The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston, MA, USA.

Salzberg, S. L. (1997). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, *1*(3), 317-327.

Sarkar, M., & Brown, M. H. (1994). Graphical fisheye views. *Communications of the ACM*, *37*(12), 73–84. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Graphical+fisheye+views#0

Satyanarayana, Y., Naithani, S., & Anu, R. (2007). Seafloor sediment classification from single beam echo sounder data using LVQ network. *Marine Geophysical Researches*, *28*(2), 95-99. doi:10.1007/s11001-007-9016-7

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461-464. doi:10.1214/aos/1176344136

Sherman, C. H., & Butler, J. L. (2007). *Transducers and arrays for underwater sounds* (1st ed.). New York, USA: Springer + Business Media, LLC.

Shmueli, G., Jank, W., Aris, A., Plaisant, C., & Shneiderman, B. (2006). Exploring Auction Databases Through Interactive Visualization. *Decision Support Systems*, *42*(3), 1521-1538. doi:10.1016/j.dss.2006.01.001

Shneiderman, B. (1992). Tree visualization with treemaps: A 2D spacefilling approach. *ACM Transactions on Graphics*, *11*(1), 92-99.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, 336-343. IEEE Comput. Soc. Press. doi:10.1109/VL.1996.545307

Simoff, S. J., Böhlen, M. H., & Mazeika, A. (2008). Visual data mining: an introduction and overview. In S. J. Simoff, M. H. Böhlen, & A. Mazeika (Eds.), *Visual data mining: theory, techniques, and tools for visual analytics* (Vol. 4404, pp. 1-12). Springer Verlag Berlin / Heidelberg. doi:10.1007/978-3-540-71080-6

Simrad-Kongsberg. (1999). EM1002 Multibeam Echosounder, Operator Manual.

Skupin, A., & Agarwal, P. (2008). Introduction: What is Self-Organizing Map? In P. Agarwal & A. Skupin (Eds.), *Self-Organising Maps: Applications in Geographic Information Science* (pp. 1-20). John Wiley & Sons: Sussex, UK.

Soukup, T., & Davidson, I. (2002). *Visual data mining: techniques and tools for data visualization and mining*. John Wiley & Sons: Canada.

Spence, Robert. (2007). *Information Visualization: Design for Interaction*. Pearson Education Limited: Essex, UK.

Spence, R., & Apperley, M. (1982). Data base navigation: an office environment for the professional. *Behaviour & Information Technology*, *1*(1), 43–54. Taylor & Francis. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01449298208914435

Spoerri, A. (1993). Infocrystal: A visual tool for information retrieval. *Visualization '93* (pp. 150-157). San Jose, CA.

Stolte, Chris, Tang, D., & Hanrahan, P. (2002). Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *Visualization and Computer Graphics, IEEE Transactions on*, *8*(1), 52–65. IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=981851

Sutherland, T., Galloway, J., Loschiavo, R., Levings, C., & Hare, R. (2007). Calibration techniques and sampling resolution requirements for groundtruthing multibeam acoustic backscatter (EM3000) and QTC VIEW^{TM} classification technology. *Estuarine, Coastal and Shelf Science*, *75*(4), 447-458. doi:10.1016/j.ecss.2007.05.045

Swayne, D. F., Cook, D., & Buja, A. (1992). *User's Manual for XGobi: A Dynamic Graphics Program for Data Analysis*. *Journal of Computational and Graphical* (Vol. 5, pp. 78-99). Retrieved from http://www.jstor.org/stable/10.2307/1390754

Tamsett, D. (1993). Sea-bed characterisation and classification from the power spectra of side-scan sonar data. *Marine Geophysical Researches*, *15*(1), 43-64. doi:10.1007/BF01204151

Tegowski, J., & Lubniewski, Z. (2000). Acoustical classification of bottom sediments in southern Baltic using fractal dimension. *Fifth European Conference on Underwater Acoustics* (pp. 313-318).

Tierney, L. (1991). *Lispstat: An object-orientated environment for statistical computing and dynamic graphics*. New York, USA: Wiley.

Tilling, L. (1975). Early experimental graphs. *British Journal for the History of Science*, *8*(30), 193–213. Cambridge Univ Press. Retrieved from http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=2914288

TimeSearcher. (2011). Visual Exploration of Time-Series Data. Retrieved from

Triton Elics International. (2004). *Using SeaClass: User manual*. Retrieved from http://www.tritonelics.com

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Connecticut, USA: Graphics Press.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Upson, C., T. A. Faulhaber, J., Kamins, D., Laidlaw, D., Schlegel, D., Vroom, J., Gurwitz, R., et al. (1989). The Application Visualization System: a computational environment for scientific visualization. *IEEE computer graphics and applications*, *9*(4), 30-42.

Urick, R. J. (1982). *Sound propagation in the sea*. California, USA: Peninsula Publishing.

User, Q. T. C. M., & Tangent, Q. (2007). Quester Tangent. *Strategy*.

Uzak, M., Jaksa, R., & Sincak, P. (2008). Reduction of visual information in neural network learning process visualization. *2008 6th International Symposium on Applied Machine Intelligence and Informatics*, 279-284. Ieee. doi:10.1109/SAMI.2008.4469183

Velleman, P. (1992). *Data Desk 4.2: Data Description. Data Desk, Ithaca, NY* (Vol. 1992). New York, USA: Data Desk, Ithaca, NY. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Data+Desk+4.2:+Data+Description#0

Vesanto, J, & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, *11*(3), 586-600. doi:10.1109/72.846731

Vesanto, J, & Alhonierni, E. (2000). Clustering of the self-organizing map. *IEEE Trans. Neural Networks*, *11*(3), 586-600.

Vesanto, J., & Ahola, J. (1999). Hunting for correlations in data using the self-organizing map. *International ICSC Congress on Computational Intelligence Methods and Applications (CIMA'99)*. New York, USA. Retrieved from http://lib.tkk.fi/Diss/2002/isbn9512258978/article5.pdf

Vesanto, Juha. (1997). *Data mining techniques based on the self-organizing map*. Helsinki University of Technology. Retrieved from http://www.cis.hut.fi/projects/ide/publications/fulldetails.shtml#vesanto97master

Vesanto, Juha. (1999). SOM-based data visualization methods. *Intelligent data analysis*, *3*(2), 111–126. Elsevier. doi:10.1016/S1088-467X(99)00013-X

Vesanto, Juha, Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). *SOM toolbox for Matlab 5*. helsinki. Retrieved from http://www.cis.hut.fi/projects/somtoolbox/package/papers/techrep.pdf

Wainer, H., & Velleman, P. F. (2001). Statistical graphics: Mapping the pathways of science. *Annual Review of Psychology*, *52*(1), 305–335. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.52.1.305

Waite, A. D. (2002). *Sonar for practising engineers* (3rd ed.). Chichester, UK: John Wiley & Sons Ltd.

Wallis, H. ., & Robinson, A. H. (1987). *Cartographical Innovations: An International Handbook of Mapping Terms to 1900*. Tring, Herts: Map Collector Publications.

Ward, M. O. (1994). Xmdvtool: Integrating multiple methods for visualizing multivariate data. *Proceedings of the Conference on Visualization'94* (pp. 326 - 333). Washington, DC , USA. doi:10.1109/VISUAL.1994.346302

Ware, C. (2000). *Information Visualization: Perception for Design*. Morgan Kaufmann.

Wilhelm, A., Unwin, A. R., & Theus, M. (1995). Software for interactive statistical graphics - a review. *SoftStat'95: 8th Conference on the Scientific Use of Statistical Software*. Heidelberg, Germany.

Worzel, J. L. (1994). Allyn C. Vine: Obituary. *Physics Today*, *47*(11), 105-106.

Xinghua, Z., & Yongqi, C. (2004). Seafloor sediment classification based on multibeam sonar data. *Geo-spatial Information Science*, *7*(4), 290–296. Springer. Retrieved from http://www.springerlink.com/index/U578HG5631542166.pdf

Xinghua, Z., & Yongqi, C. (2005). Seafloor classification of multibeam sonar data using neural network approach. *Marine Geodesy*, *28*, 201-206.

Yan, J., & Thill, J.-C. (2007). Visual exploration in US airline market structures. *Papers of Applied Geography Conferences*, *30*, 10-19.

Yi, B., Jagadish, H. V., & Faloutsos, C. (1998). Efficient retrieval of similar time sequences under time warping. *14th International Conference on Data Engineering* (pp. 201-208). Orlando FL, USA: IEEE Comput. Soc. Press.

Yorke, T. H., & Oberg, K. A. (2002). Measuring river velocity and discharge with acoustic doppler profilers. *Flow Measurement and Instrumentation*, *13*, 191-195.

Zentrum für Marine Umweltwissenschaften(MARUM). (2011). Center for marine environment sciences. Retrieved from

Zimmermann, M., & Rooper, C. N. (2008). Comparison of echogram measurements against data expectations and assumptions for distinguishing seafloor substrates. *Fishery Bulletin*, *106*(3), 293-304.

van Wijk, J. J., & Van Selow, E. R. (1999). Cluster and calendar based visualization of time series data. *InforVis'99: IEEE Symposium on Information Visualization* (pp. 4-9). San Francisco, CA, USA.

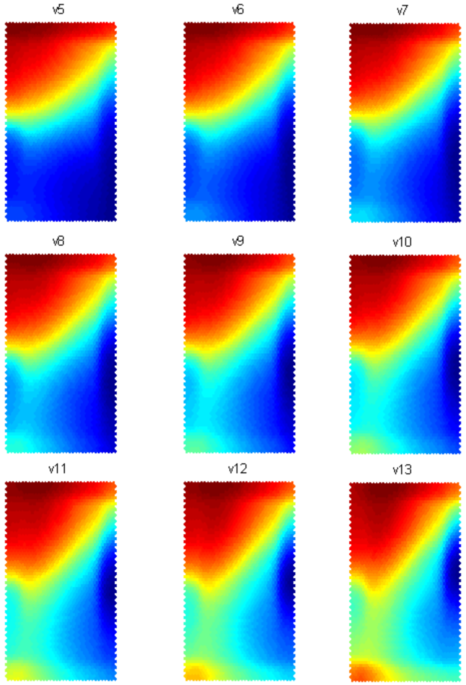


Figure A 3.1: Visual Group 1 for 0.2ms

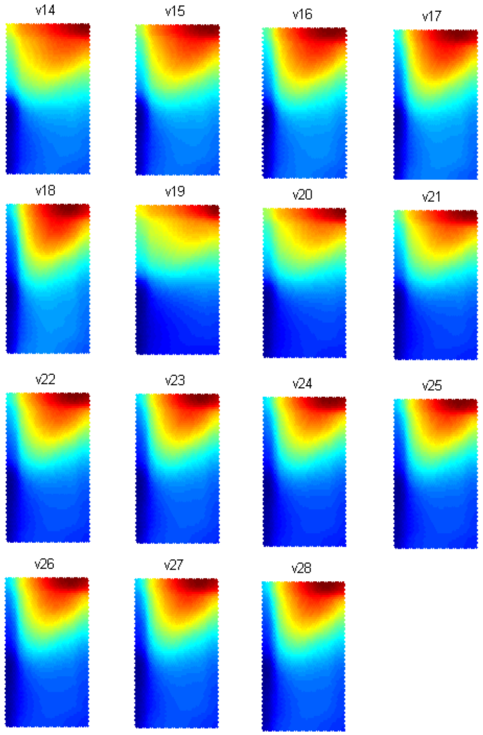


Figure A 3.2: Visual Group 2 for 0.2ms

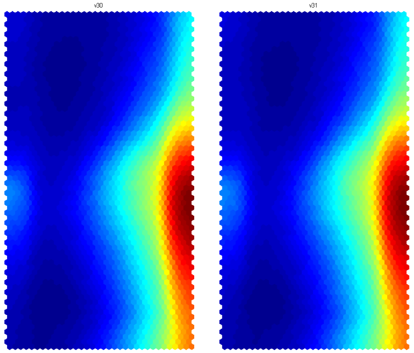


Figure A 3.3: Visual Group 3 for 0.2ms

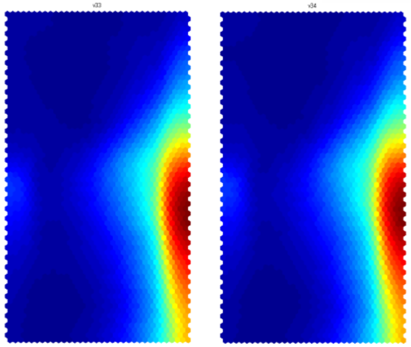


Figure A 3.4: Visual Group 4 for 0.2ms

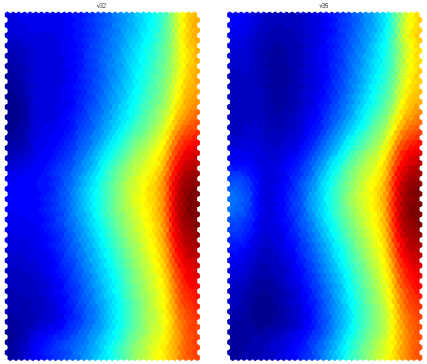


Figure A 3.5: Visual Group 5 for 0.2ms

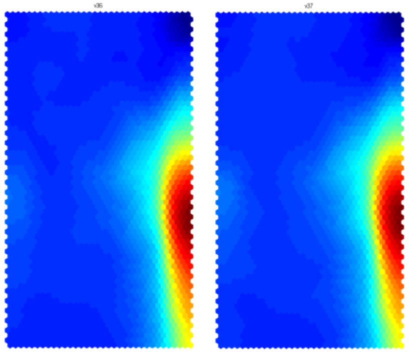


Figure A 3.6: Visual Group 6 for 0.2ms

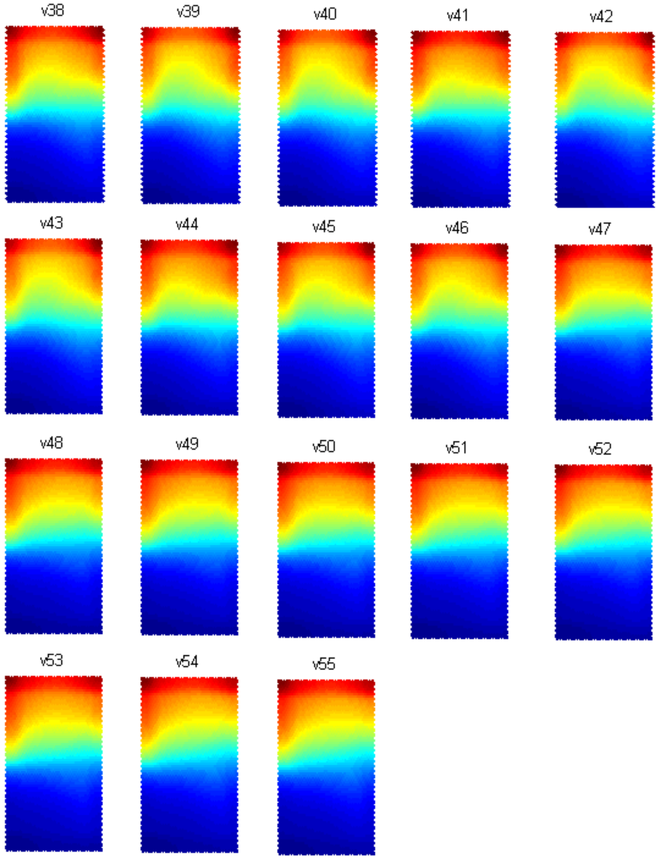


Figure A 3.7: Visual Group 7 for 0.2ms

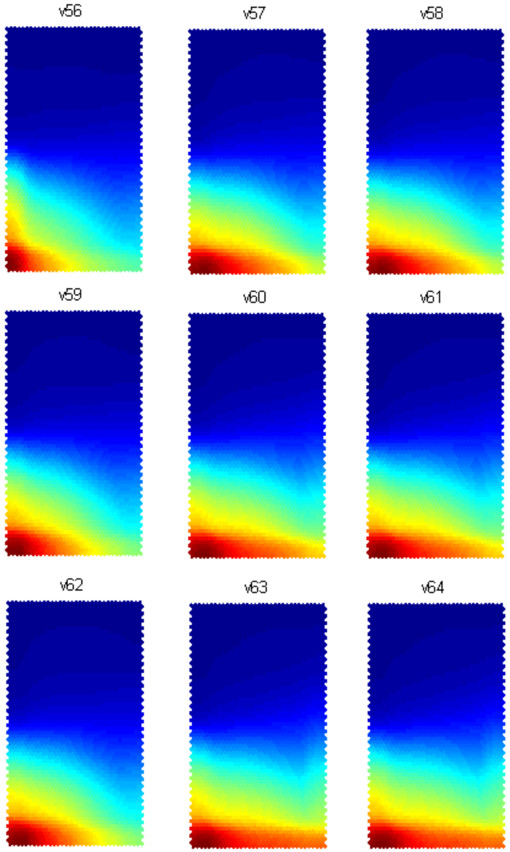


Figure A 3.8: Visual Group 8 for 0.2ms

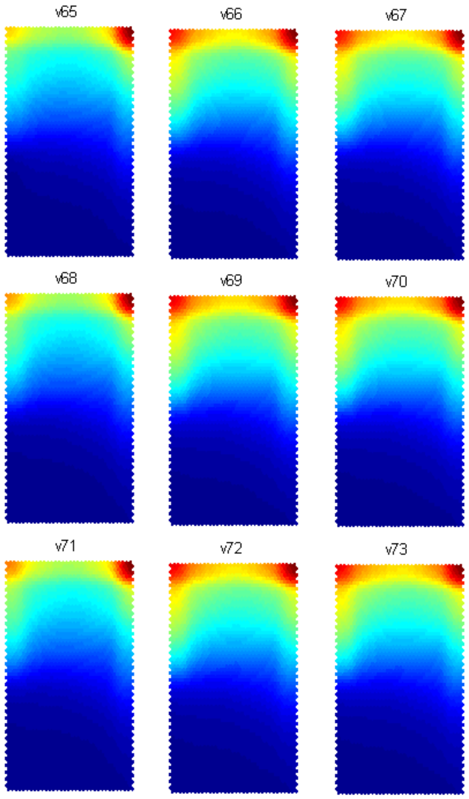


Figure A 3.9: Visual Group 9 for 0.2ms

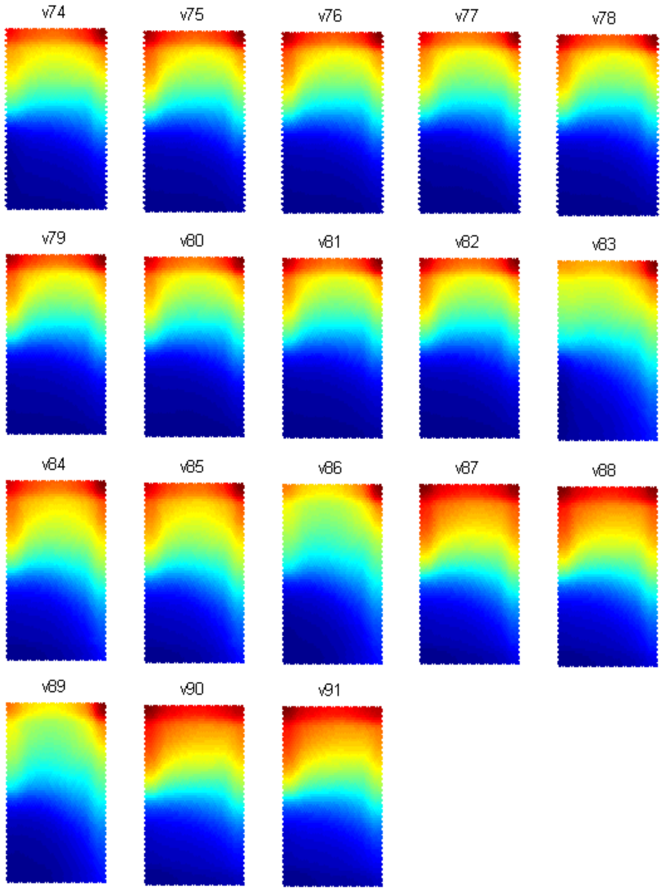


Figure A 3.10: Visual Group 10 for 0.2ms

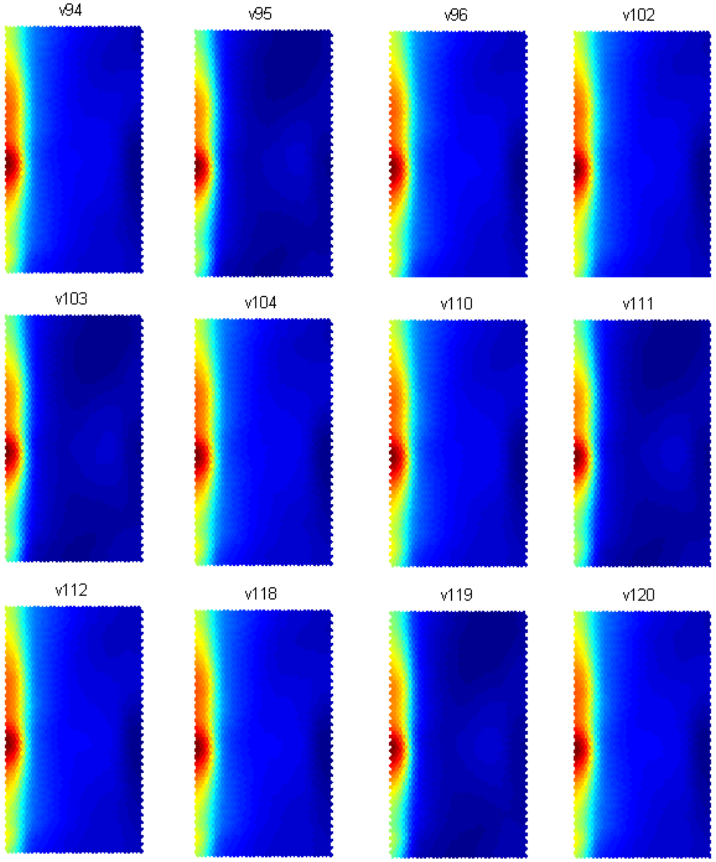


Figure A 3.11: Visual Group 11 for 0.2ms

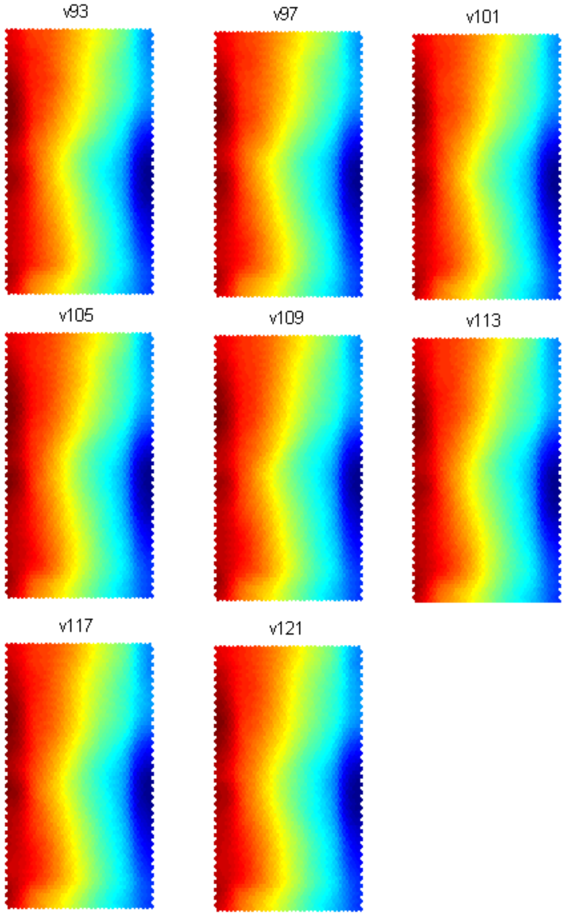


Figure A 3.12: Visual Group 12 for 0.2ms

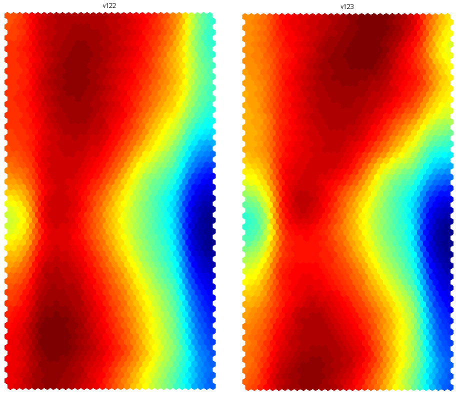


Figure A 3.13: Visual Group 13 for 0.2ms

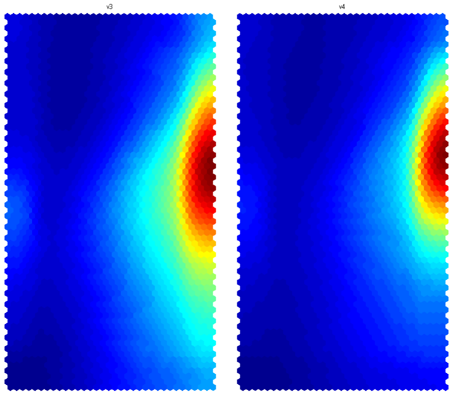


Figure A 3.14: Visual Group 14 for 0.2ms

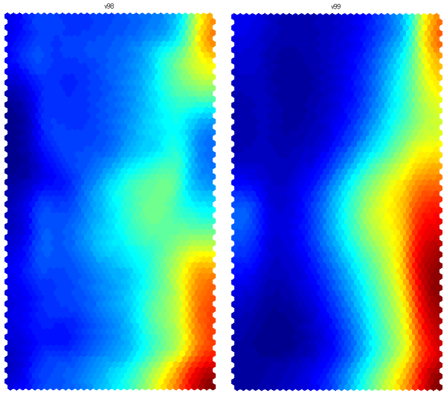


Figure A 3.15: Visual Group 15 for 0.2ms

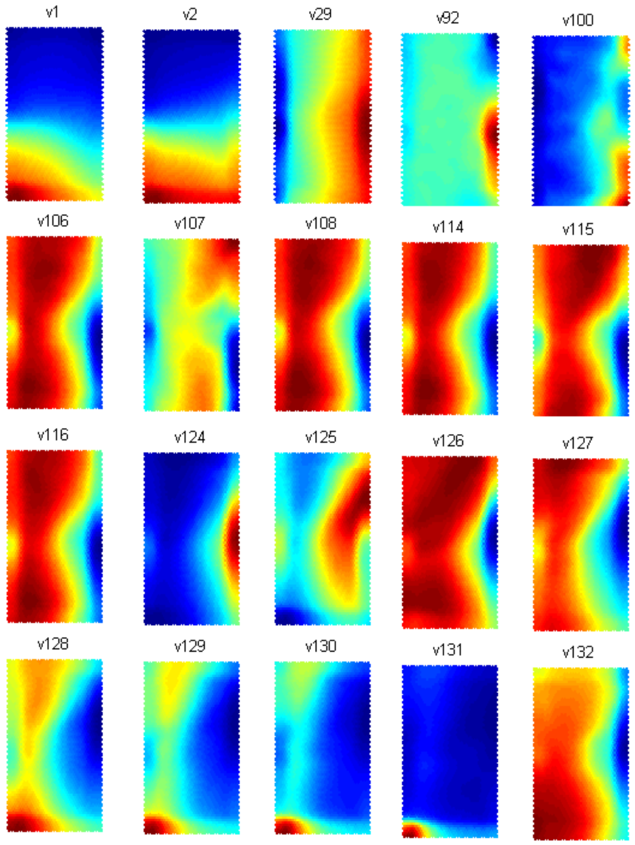


Figure A 3.16: Visual Group 16 (singletons) for 0.2ms

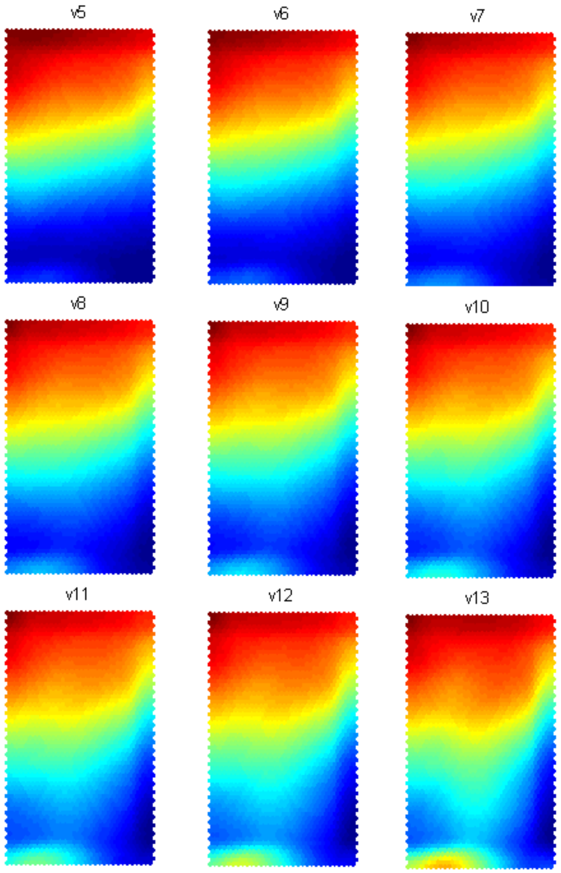


Figure A 3.17: Visual Group 1 for 0.7ms

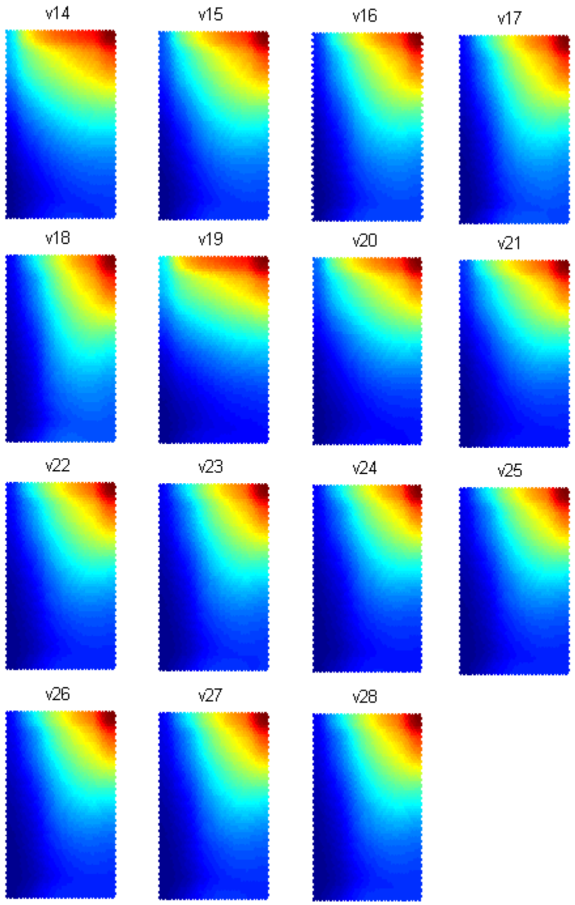


Figure A 3.18: Visual Group 2 for 0.7ms

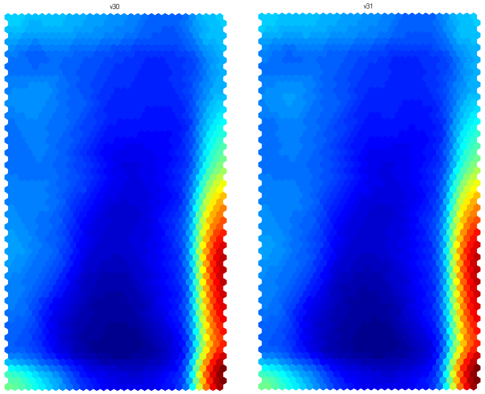


Figure A 3.19: Visual Group 3 for 0.7ms

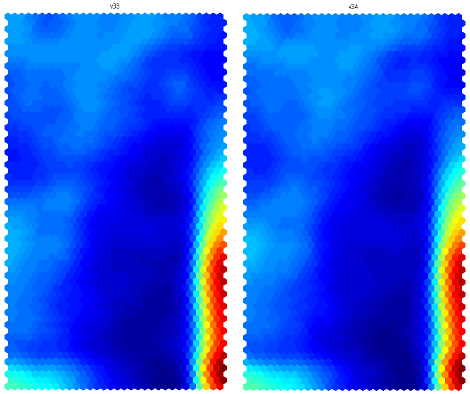


Figure A 3.20: Visual Group 4 for 0.7ms

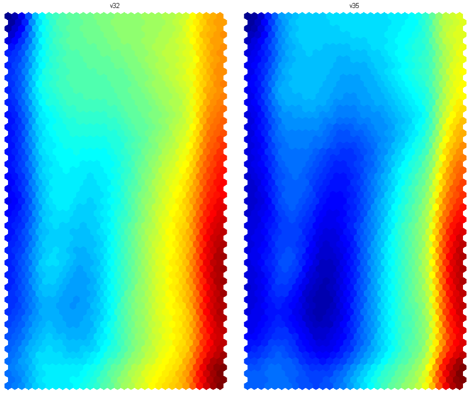


Figure A 3.21: Visual Group 5 for 0.7ms

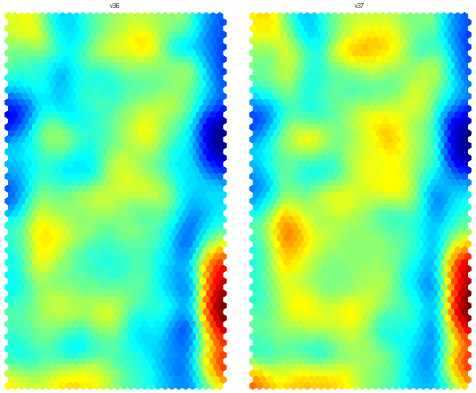


Figure A 3.22: Visual Group 6 for 0.7ms

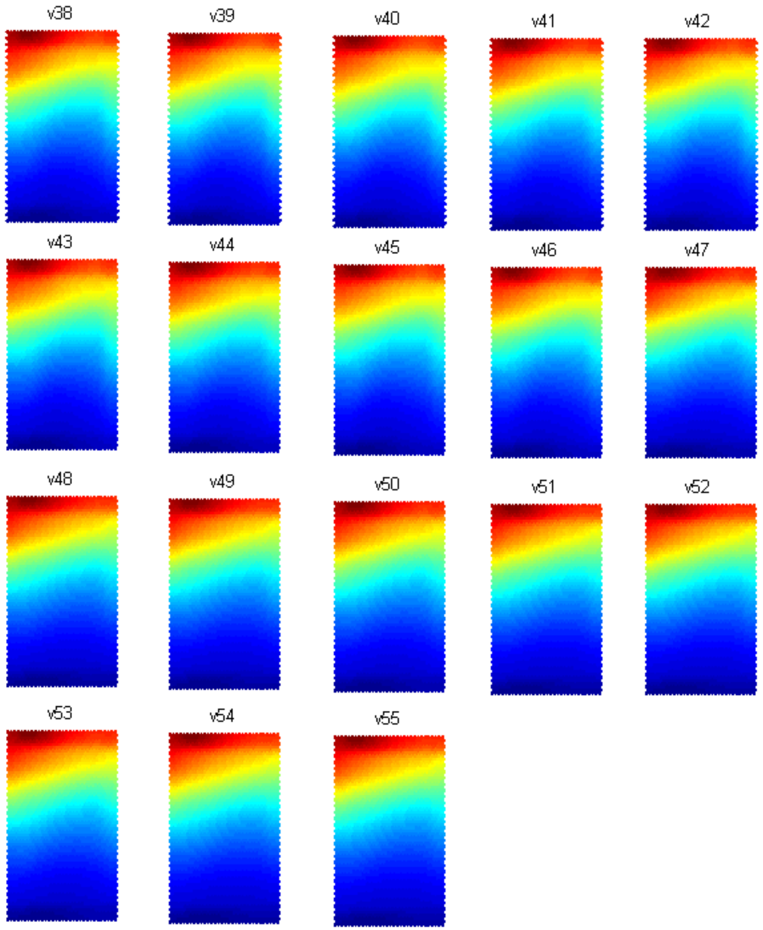


Figure A 3.23: Visual Group 7 for 0.7ms

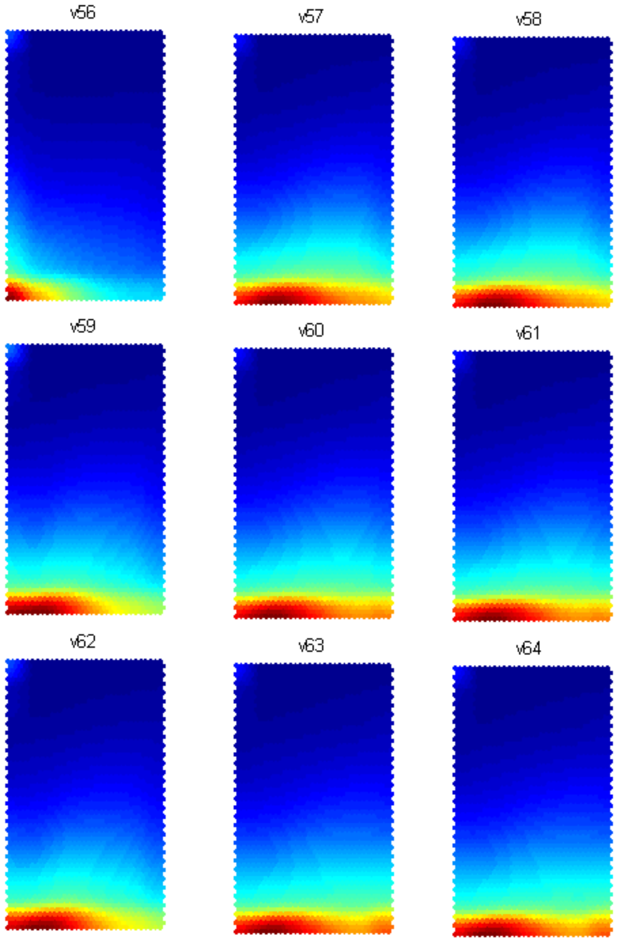


Figure A 3.24: Visual Group 8 for 0.7ms

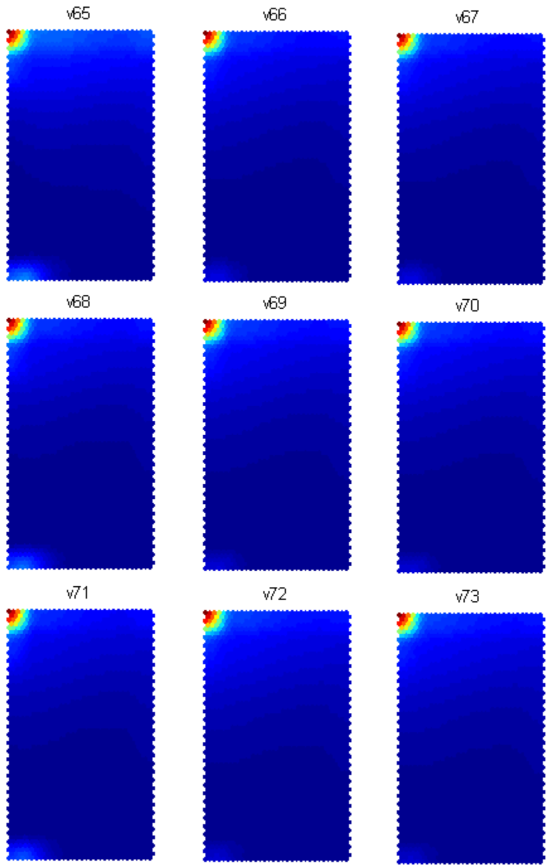


Figure A 3.25: Visual Group 9 for 0.7ms

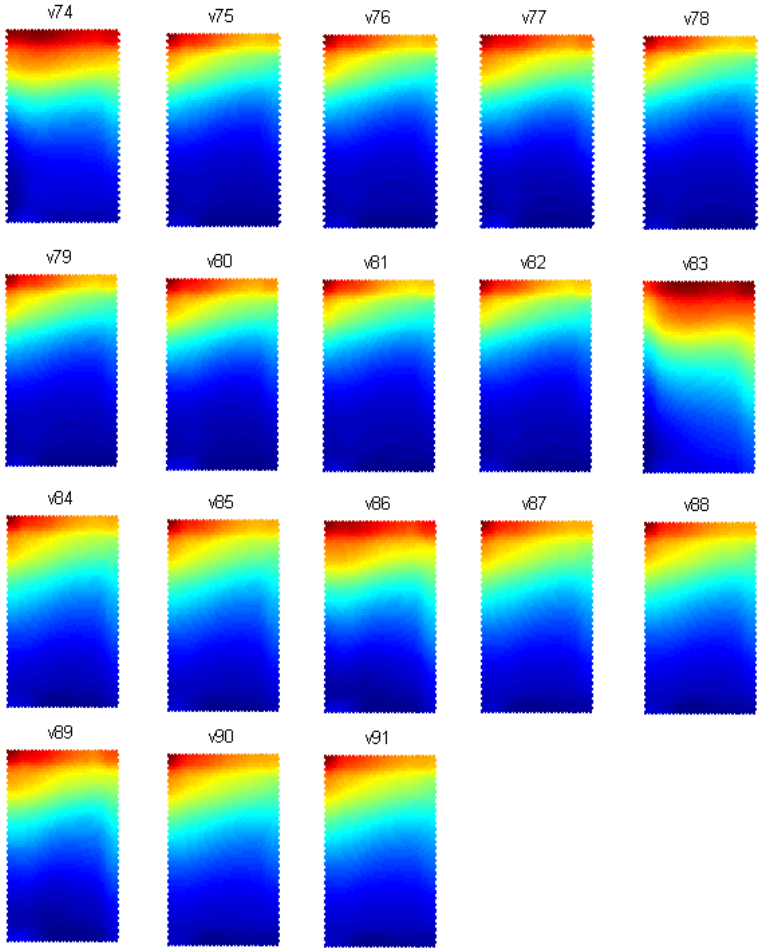


Figure A 3.26: Visual Group 10 for 0.7ms

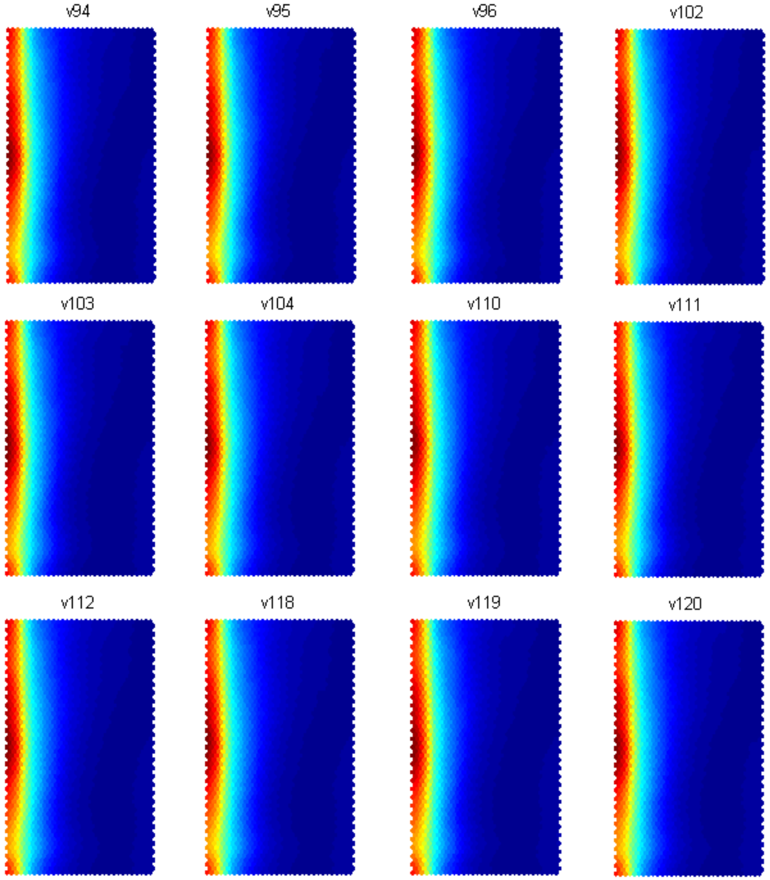


Figure A 3.27: Visual Group 11 for 0.7ms

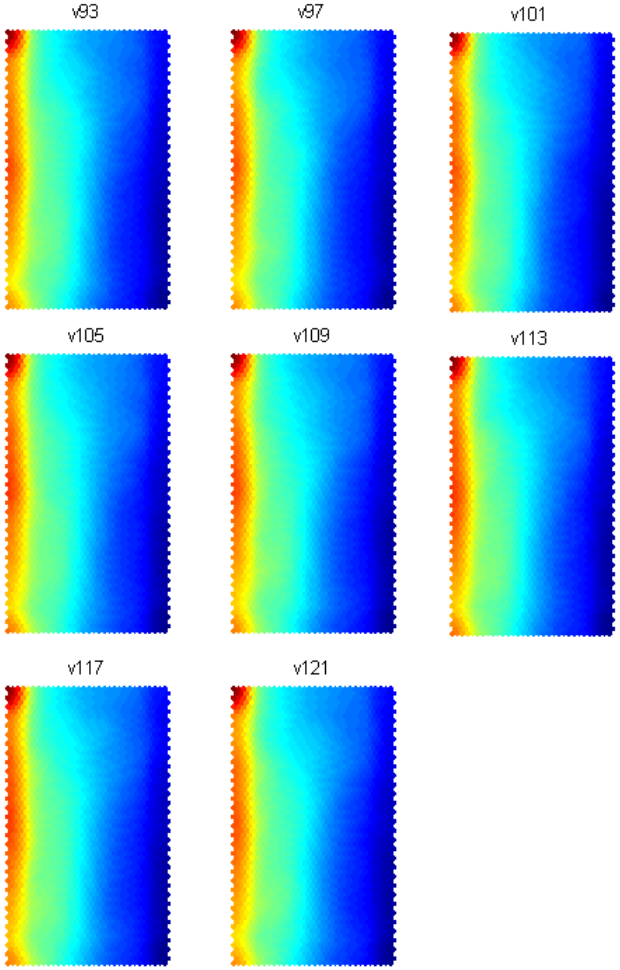


Figure A 3.28: Visual Group 12 for 0.7ms

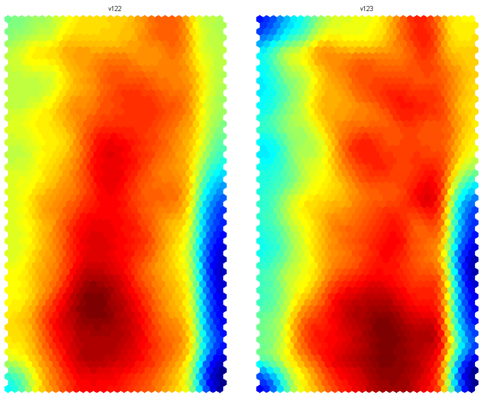


Figure A 3.29: Visual Group 13 for 0.7ms

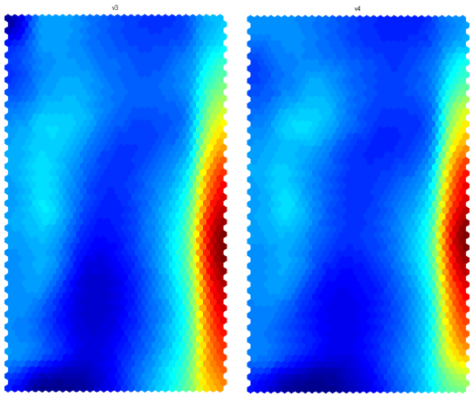


Figure A 3.30: Visual Group 14 for 0.7ms

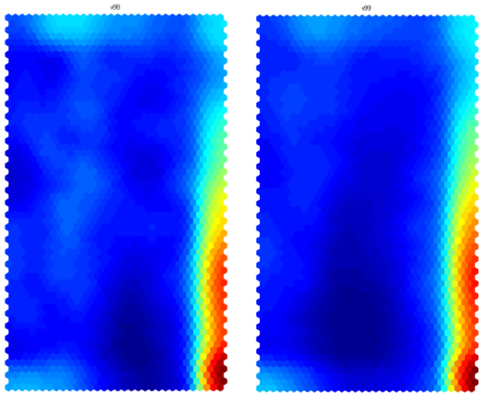


Figure A 3.31: Visual Group 15 for 0.7ms

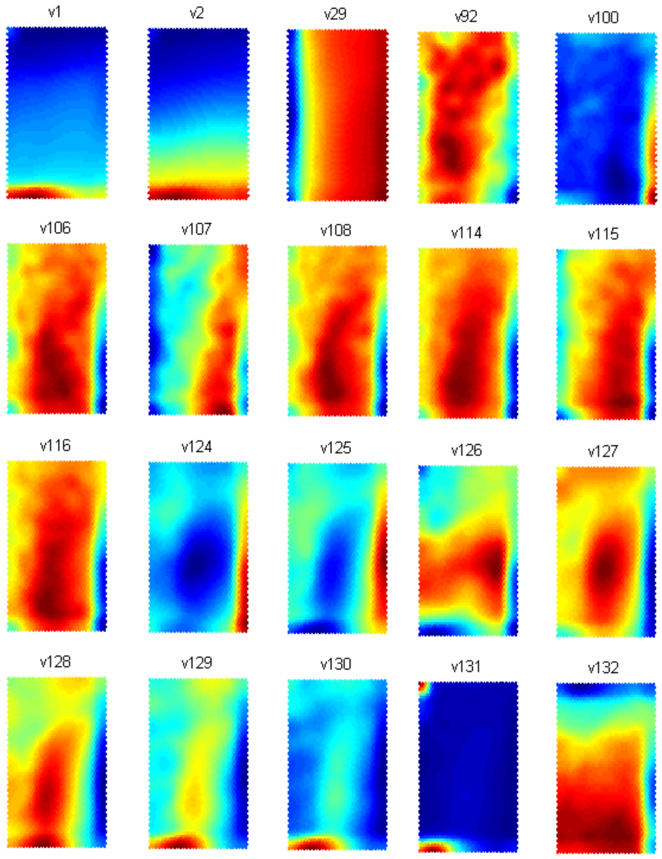


Figure A 3.32: Visual Group 16 for 0.7ms

Table A 3.1: Correlation coefficient of VG 1, 0.2ms

	FFV 5	FFV 6	FFV 7	FFV 8	FFV 9	FFV 10	FFV 11	FFV 12	FFV 13
FFV 5	1	0.9955	0.9865	0.9734	0.9562	0.9334	0.9003	0.8491	0.7579
FFV 6	0.9955	1	0.9962	0.9884	0.9766	0.9587	0.9304	0.8846	0.8002
FFV 7	0.9865	0.9962	1	0.9965	0.9887	0.9757	0.9534	0.9137	0.8345
FFV 8	0.9734	0.9884	0.9965	1	0.9963	0.9872	0.9698	0.9369	0.8654
FFV 9	0.9562	0.9766	0.9887	0.9963	1	0.9955	0.9828	0.9559	0.8940
FFV 10	0.9334	0.9587	0.9757	0.9872	0.9955	1	0.9937	0.9734	0.9210
FFV 11	0.9003	0.9304	0.9534	0.9698	0.9828	0.9937	1	0.9895	0.9480
FFV 12	0.8491	0.8846	0.9137	0.9369	0.9559	0.97336	0.9895	1	0.9762
FFV 13	0.7579	0.8002	0.8345	0.8654	0.8940	0.9210	0.9480	0.9762	1

Table A 3.2: Correlation coefficient of VG 2, 0.2ms

	FFV 14	FFV 15	FFV 16	FFV 17	FFV 18	FFV 19	FFV 20	FFV 21	FFV 22	FFV 23	FFV 24	FFV 25	FFV 26	FFV 27	FFV 28
FFV 14	1	0.9798	0.95547	0.92957	0.90035	0.98252	0.98438	0.97104	0.95157	0.92813	0.92763	0.9245	0.92305	0.92043	0.91925
FFV 15	0.9798	1	0.98339	0.96384	0.94065	0.95301	0.98807	0.98508	0.97451	0.95823	0.9552	0.95409	0.95343	0.9521	0.95144
FFV 16	0.9555	0.98339	1	0.98589	0.96818	0.9209	0.9642	0.98979	0.9869	0.97704	0.97152	0.97234	0.97244	0.97233	0.9722
FFV 17	0.9296	0.96384	0.98589	1	0.98636	0.88823	0.93864	0.96963	0.99228	0.98736	0.97928	0.98187	0.98266	0.9837	0.98404
FFV 18	0.9003	0.94065	0.96818	0.98636	1	0.8526	0.90947	0.94601	0.97334	0.99287	0.98173	0.98613	0.98765	0.98986	0.9907
FFV 19	0.9825	0.95301	0.9209	0.88823	0.8526	1	0.98278	0.958	0.9271	0.89556	0.90664	0.89753	0.89377	0.8874	0.88467
FFV 20	0.9844	0.98807	0.9642	0.93864	0.90947	0.98278	1	0.98666	0.966	0.94224	0.94735	0.94287	0.94018	0.93554	0.93351
FFV 21	0.9710	0.98508	0.98979	0.96963	0.94601	0.958	0.98666	1	0.98735	0.97025	0.9718	0.96934	0.96813	0.96488	0.96344
FFV 22	0.9516	0.97451	0.9869	0.99228	0.97334	0.9271	0.966	0.98735	1	0.98864	0.98596	0.98565	0.98529	0.98441	0.98359
FFV 23	0.9281	0.95823	0.97704	0.98736	0.99287	0.89556	0.94224	0.97025	0.98864	1	0.99378	0.99533	0.99572	0.99607	0.9961
FFV 24	0.9276	0.9552	0.97152	0.97928	0.98173	0.90664	0.94735	0.9718	0.98596	0.99378	1	0.9989	0.99826	0.99697	0.99633
FFV 25	0.9245	0.95409	0.97234	0.98187	0.98613	0.89753	0.94287	0.96934	0.98565	0.99533	0.9989	1	0.99968	0.99889	0.99846
FFV 26	0.9230	0.95343	0.97244	0.98266	0.98765	0.89377	0.94018	0.96813	0.98529	0.99572	0.99826	0.99968	1	0.99941	0.99908
FFV 27	0.9204	0.9521	0.97233	0.9837	0.98986	0.8874	0.93554	0.96488	0.98441	0.99607	0.99697	0.99889	0.99941	1	0.99981
FFV 28	0.9192	0.95144	0.9722	0.98404	0.9907	0.88467	0.93351	0.96344	0.98359	0.9961	0.99633	0.99846	0.99908	0.99981	1

Table A 3.3: Correlation coefficient of VG 3, 0.2ms

	FFV 30	FFV 31
FFV 30	1	0.98911
FFV 31	0.98911	1

Table A 3.4: Correlation coefficient of VG 4, 0.2ms

	FFV 33	FFV 34
FFV 33	1	0.94515
FFV 34	0.94515	1

Table A 3.5: Correlation coefficient of VG 5, 0.2ms

	FFV 32	FFV 35
FFV 32	1	0.97749
FFV 35	0.97749	1

Table A 3.6: Correlation coefficient of VG 6, 0.2ms

	FFV 36	FFV 37
FFV 36	1	0.87139
FFV 37	0.87139	1

Table A 3.7: Correlation coefficient of VG 7, 0.2ms

	FFV 38	FFV 39	FFV 40	FFV 41	FFV 42	FFV 43	FFV 44	FFV 45	FFV 46	FFV 47	FFV 48	FFV 49	FFV 50	FFV 51	FFV 52	FFV 53	FFV 54	FFV 55
FFV 38	1	0.9886	0.9887	0.9993	0.9875	0.9877	0.9963	0.9860	0.9865	0.9964	0.9925	0.9924	0.9958	0.9910	0.9910	0.9946	0.9888	0.9887
FFV 39	0.9886	1	0.9985	0.9876	0.9829	0.9830	0.9865	0.9773	0.9776	0.9806	0.9832	0.9827	0.9794	0.9759	0.9757	0.9780	0.9711	0.9708
FFV 40	0.9887	0.9985	1	0.9877	0.9829	0.9831	0.9866	0.9773	0.9777	0.9808	0.9829	0.9833	0.9796	0.9760	0.9759	0.9781	0.9712	0.9710
FFV 41	0.9993	0.9876	0.9877	1	0.9859	0.9861	0.9984	0.9844	0.9848	0.9953	0.9911	0.9911	0.9953	0.9894	0.9894	0.9947	0.9870	0.9869
FFV 42	0.9875	0.9829	0.9829	0.9859	1	0.9942	0.9836	0.9797	0.9799	0.9809	0.9786	0.9785	0.9797	0.9835	0.9815	0.9782	0.9744	0.9741
FFV 43	0.9877	0.9830	0.9831	0.9861	0.9942	1	0.9837	0.9798	0.9802	0.9812	0.9788	0.9787	0.9800	0.9819	0.9835	0.9784	0.9746	0.9744
FFV 44	0.9963	0.9865	0.9866	0.9984	0.9836	0.9837	1	0.9812	0.9815	0.9900	0.9852	0.9851	0.9904	0.9829	0.9828	0.9904	0.9798	0.9797
FFV 45	0.9860	0.9773	0.9773	0.9844	0.9797	0.9798	0.9812	1	0.9882	0.9814	0.9782	0.9781	0.9804	0.9785	0.9783	0.9790	0.9833	0.9794
FFV 46	0.9865	0.9776	0.9777	0.9848	0.9799	0.9802	0.9815	0.9882	1	0.9821	0.9789	0.9789	0.9811	0.9791	0.9791	0.9796	0.9804	0.9836
FFV 47	0.9964	0.9806	0.9808	0.9953	0.9809	0.9812	0.9900	0.9814	0.9821	1	0.9971	0.9971	0.9998	0.9963	0.9963	0.9991	0.9952	0.9952
FFV 48	0.9925	0.9832	0.9829	0.9911	0.9786	0.9788	0.9852	0.9782	0.9789	0.9971	1	0.9997	0.9967	0.9964	0.9964	0.9963	0.9946	0.9946
FFV 49	0.9924	0.9827	0.9833	0.9911	0.9785	0.9787	0.9851	0.9781	0.9789	0.9971	0.9997	1	0.9967	0.9964	0.9964	0.9963	0.9946	0.9946
FFV 50	0.9958	0.9794	0.9796	0.9953	0.9797	0.9800	0.9904	0.9804	0.9811	0.9998	0.9967	0.9967	1	0.9959	0.9959	0.9997	0.9949	0.9949
FFV 51	0.9910	0.9759	0.9760	0.9894	0.9835	0.9819	0.9829	0.9785	0.9791	0.9963	0.9964	0.9964	0.9959	1	0.9990	0.9954	0.9961	0.9960
FFV 52	0.9910	0.9757	0.9759	0.9894	0.9815	0.9835	0.9828	0.9783	0.9791	0.9963	0.9964	0.9964	0.9959	0.9990	1	0.9954	0.9960	0.9961
FFV 53	0.9946	0.9780	0.9781	0.9947	0.9782	0.9784	0.9904	0.9790	0.9796	0.9991	0.9963	0.9963	0.9997	0.9954	0.9954	1	0.9944	0.9944
FFV 54	0.9888	0.9711	0.9712	0.9870	0.9744	0.9746	0.9798	0.9833	0.9804	0.9952	0.9946	0.9946	0.9949	0.9961	0.9960	0.9944	1	0.9981
FFV 55	0.9887	0.9708	0.9710	0.9869	0.9741	0.9744	0.9797	0.9794	0.9836	0.9952	0.9946	0.9946	0.9949	0.9960	0.9961	0.9944	0.9981	1

Table A 3.8: Correlation coefficient of VG 8, 0.2ms

	FFV 56	FFV 57	FFV 58	FFV 59	FFV 60	FFV 61	FFV 62	FFV 63	FFV 64
FFV 56	1	0.96457	0.96463	0.98727	0.95569	0.95589	0.97876	0.94378	0.94402
FFV 57	0.96457	1	0.99949	0.98673	0.98969	0.98977	0.9892	0.97827	0.97844
FFV 58	0.96463	0.99949	1	0.98678	0.98965	0.98978	0.98924	0.97824	0.97847
FFV 59	0.98727	0.98673	0.98678	1	0.97744	0.97762	0.99862	0.96477	0.96506
FFV 60	0.95569	0.98969	0.98965	0.97744	1	0.99824	0.97986	0.98887	0.98891
FFV 61	0.95589	0.98977	0.98978	0.97762	0.99824	1	0.98002	0.98871	0.98895
FFV 62	0.97876	0.9892	0.98924	0.99862	0.97986	0.98002	1	0.96706	0.96733
FFV 63	0.94378	0.97827	0.97824	0.96477	0.98887	0.98871	0.96706	1	0.99665
FFV 64	0.94402	0.97844	0.97847	0.96506	0.98891	0.98895	0.96733	0.99665	1

Table A 3.9: Correlation coefficient of VG 9, 0.2ms

	FFV 65	FFV 66	FFV 67	FFV 68	FFV 69	FFV 70	FFV 71	FFV 72	FFV 73
FFV 65	1	0.96941	0.96954	0.99438	0.95909	0.95916	0.99016	0.95962	0.95962
FFV 66	0.96941	1	0.99772	0.97235	0.9932	0.99314	0.97294	0.9902	0.99029
FFV 67	0.96954	0.99772	1	0.97242	0.99318	0.99318	0.97303	0.9902	0.99033
FFV 68	0.99438	0.97235	0.97242	1	0.96068	0.96068	0.99751	0.96029	0.9602
FFV 69	0.95909	0.9932	0.99318	0.96068	1	0.99675	0.96115	0.99334	0.99341
FFV 70	0.95916	0.99314	0.99318	0.96068	0.99675	1	0.96109	0.99333	0.99353
FFV 71	0.99016	0.97294	0.97303	0.99751	0.96115	0.96109	1	0.96018	0.96008
FFV 72	0.95962	0.9902	0.9902	0.96029	0.99334	0.99333	0.96018	1	0.9954
FFV 73	0.95962	0.99029	0.99033	0.9602	0.99341	0.99353	0.96008	0.9954	1

Table A 3.10: Correlation coefficient of VG 10, 0.2ms

	FFV 74	FFV 75	FFV 76	FFV 77	FFV 78	FFV 79	FFV 80	FFV 81	FFV 82	FFV 83	FFV 84	FFV 85	FFV 86	FFV 87	FFV 88	FFV 89	FFV 90	FFV 91
FFV 74	1	0.9920	0.9921	0.9966	0.9900	0.9901	0.9950	0.9885	0.9888	0.9705	0.9685	0.9683	0.9662	0.9667	0.9668	0.9603	0.9628	0.9630
FFV 75	0.9920	1	0.9983	0.9954	0.9969	0.9969	0.9958	0.9950	0.9952	0.9491	0.9713	0.9708	0.9519	0.9719	0.9720	0.9490	0.9680	0.9683
FFV 76	0.9921	0.9983	1	0.9954	0.9968	0.9969	0.9959	0.9949	0.9952	0.9492	0.9712	0.9711	0.9519	0.9719	0.9721	0.9490	0.9681	0.9683
FFV 77	0.9966	0.9954	0.9954	1	0.9937	0.9937	0.9984	0.9927	0.9929	0.9623	0.9692	0.9690	0.9653	0.9664	0.9665	0.9620	0.9619	0.9622
FFV 78	0.9900	0.9969	0.9968	0.9937	1	0.9976	0.9943	0.9958	0.9959	0.9458	0.9675	0.9672	0.9488	0.9707	0.9705	0.9460	0.9669	0.9671
FFV 79	0.9901	0.9969	0.9969	0.9937	0.9976	1	0.9943	0.9958	0.9960	0.9459	0.9676	0.9673	0.9489	0.9706	0.9710	0.9460	0.9671	0.9674
FFV 80	0.9950	0.9958	0.9959	0.9984	0.9943	0.9943	1	0.9934	0.9935	0.9579	0.9683	0.9680	0.9622	0.9659	0.9660	0.9600	0.9617	0.9620
FFV 81	0.9885	0.9950	0.9949	0.9927	0.9958	0.9958	0.9934	1	0.9966	0.9451	0.9647	0.9643	0.9489	0.9660	0.9661	0.9464	0.9637	0.9635
FFV 82	0.9888	0.9952	0.9952	0.9929	0.9959	0.9960	0.9935	0.9966	1	0.9454	0.9650	0.9646	0.9491	0.9664	0.9666	0.9465	0.9637	0.9643
FFV 83	0.9705	0.9491	0.9492	0.9623	0.9458	0.9459	0.9579	0.9451	0.9454	1	0.9504	0.9503	0.9825	0.9330	0.9331	0.9748	0.9208	0.9212
FFV 84	0.9685	0.9713	0.9712	0.9692	0.9675	0.9676	0.9683	0.9647	0.9650	0.9504	1	0.9833	0.9524	0.9663	0.9664	0.9498	0.9536	0.9539
FFV 85	0.9683	0.9708	0.9711	0.9690	0.9672	0.9673	0.9680	0.9643	0.9646	0.9503	0.9833	1	0.9522	0.9660	0.9661	0.9496	0.9534	0.9535
FFV 86	0.9662	0.9519	0.9519	0.9653	0.9488	0.9489	0.9622	0.9489	0.9491	0.9825	0.9524	0.9522	1	0.9311	0.9312	0.9849	0.9167	0.9171
FFV 87	0.9667	0.9719	0.9719	0.9664	0.9707	0.9706	0.9659	0.9660	0.9664	0.9330	0.9663	0.9660	0.9311	1	0.9786	0.9268	0.9675	0.9678
FFV 88	0.9668	0.9720	0.9721	0.9665	0.9705	0.9710	0.9660	0.9661	0.9666	0.9331	0.9664	0.9661	0.9312	0.9786	1	0.9269	0.9677	0.9680
FFV 89	0.9603	0.9490	0.9490	0.9620	0.9460	0.9460	0.9600	0.9464	0.9465	0.9748	0.9498	0.9496	0.9849	0.9268	0.9269	1	0.9114	0.9119
FFV 90	0.9628	0.9680	0.9681	0.9619	0.9669	0.9671	0.9617	0.9637	0.9637	0.9208	0.9536	0.9534	0.9167	0.9675	0.9677	0.9114	1	0.9737
FFV 91	0.9630	0.9683	0.9683	0.9622	0.9671	0.9674	0.9620	0.9635	0.9643	0.9212	0.9539	0.9535	0.9171	0.9678	0.9680	0.9119	0.9737	1

Table A 3.11: Correlation coefficient of VG 11, 0.2ms

	FFV 94	FFV 95	FFV 96	FFV 102	FFV 103	FFV 104	FFV 110	FFV 111	FFV 112	FFV 118	FFV 119	FFV 120
FFV 94	1	0.91803	0.88043	0.87867	0.89112	0.87523	0.87477	0.8883	0.8751	0.87467	0.88681	0.87394
FFV 95	0.91803	1	0.91782	0.88713	0.93049	0.88804	0.88686	0.92621	0.88794	0.88796	0.92327	0.88649
FFV 96	0.88043	0.91782	1	0.87518	0.89057	0.87896	0.87351	0.8882	0.87524	0.8745	0.88589	0.87329
FFV 102	0.87867	0.88713	0.87518	1	0.9173	0.87617	0.87789	0.88925	0.87462	0.87469	0.88584	0.87324
FFV 103	0.89112	0.93049	0.89057	0.9173	1	0.91857	0.88961	0.93192	0.89199	0.88854	0.9252	0.88825
FFV 104	0.87523	0.88804	0.87896	0.87617	0.91857	1	0.87395	0.89045	0.88023	0.87414	0.88686	0.87483
FFV 110	0.87477	0.88686	0.87351	0.87789	0.88961	0.87395	1	0.91728	0.8742	0.87905	0.88843	0.87436
FFV 111	0.8883	0.92621	0.8882	0.88925	0.93192	0.89045	0.91728	1	0.91793	0.88961	0.92959	0.88986
FFV 112	0.8751	0.88794	0.87524	0.87462	0.89199	0.88023	0.8742	0.91793	1	0.87553	0.88852	0.87919
FFV 118	0.87467	0.88796	0.8745	0.87469	0.88854	0.87414	0.87905	0.88961	0.87553	1	0.91807	0.88107
FFV 119	0.88681	0.92327	0.88589	0.88584	0.9252	0.88686	0.88843	0.92959	0.88852	0.91807	1	0.91782
FFV 120	0.87394	0.88649	0.87329	0.87324	0.88825	0.87483	0.87436	0.88986	0.87919	0.88107	0.91782	1

Table A 3.12: Correlation coefficient of VG 12, 0.2ms

	FFV 93	FFV 97	FFV 101	FFV 105	FFV 109	FFV 113	FFV 117	FFV 121
FFV 93	1	0.52546	0.52498	0.51649	0.51074	0.51044	0.51482	0.5093
FFV 97	0.52546	1	0.51421	0.52756	0.50888	0.51195	0.50845	0.51183
FFV 101	0.52498	0.51421	1	0.51312	0.52612	0.50896	0.51425	0.51101
FFV 105	0.51649	0.52756	0.51312	1	0.51102	0.52587	0.51278	0.51375
FFV 109	0.51074	0.50888	0.52612	0.51102	1	0.50843	0.52824	0.51312
FFV 113	0.51044	0.51195	0.50896	0.52587	0.50843	1	0.51369	0.52509
FFV 117	0.51482	0.50845	0.51425	0.51278	0.52824	0.51369	1	0.52724
FFV 121	0.5093	0.51183	0.51101	0.51375	0.51312	0.52509	0.52724	1

Table A 3.13: Correlation coefficient of VG 13, 0.2ms

	FFV 122	FFV 123
FFV 122	1	0.74743448
FFV 123	0.74743448	1

Table A 3.14: Correlation coefficient of VG 14, 0.2ms

	FFV 3	FFV 4
FFV 3	1	0.9089
FFV 4	0.9089	1

Table A 3.15: Correlation coefficient of VG 15, 0.2ms

	FFV 98	FFV 99
FFV 98	1	0.6011
FFV 99	0.6011	1

Table A 3.16: Correlation coefficient of VG 15 (Singletons), 0.2ms

	FFV 1	FFV 2	FFV 29	FFV 92	FFV 100	FFV 106	FFV 107	FFV 108	FFV 114	FFV 115	FFV 116	FFV 124	FFV 125	FFV 126	FFV 127	FFV 128	FFV 129	FFV 130	FFV 131	FFV 132
FFV 1	1	0.955	-0.103	0.033	-0.002	0.082	-0.033	0.079	0.101	0.025	0.099	-0.158	-0.347	0.063	0.112	0.392	0.386	0.366	0.477	0.284
FFV 2	0.955	1	0.072	0.057	0.037	-0.040	-0.047	-0.042	-0.039	-0.057	-0.042	0.061	-0.293	-0.118	-0.067	0.237	0.261	0.276	0.420	0.169
FFV 29	-0.103	0.072	1	0.073	0.154	-0.282	0.110	-0.283	-0.348	-0.085	-0.349	0.567	0.173	-0.506	-0.500	-0.385	-0.303	-0.187	-0.056	-0.383
FFV 92	0.033	0.057	0.073	1	-0.363	-0.173	-0.291	-0.172	-0.132	-0.225	-0.131	0.101	0.005	-0.091	-0.079	-0.053	-0.036	-0.027	-0.001	0.076
FFV 100	-0.002	0.037	0.154	-0.363	1	-0.200	-0.182	-0.180	-0.196	-0.221	-0.191	0.092	-0.004	-0.087	-0.067	-0.040	-0.029	-0.013	-0.006	-0.203
FFV 106	0.082	-0.040	-0.282	-0.173	-0.200	1	0.507	0.346	0.269	0.146	0.251	-0.432	-0.118	0.381	0.363	0.301	0.219	0.151	0.076	0.241
FFV 107	-0.033	-0.047	0.110	-0.291	-0.182	0.507	1	0.506	-0.010	0.058	-0.008	-0.059	0.028	0.059	0.028	0.014	0.003	0.005	0.006	0.018
FFV 108	0.079	-0.042	-0.283	-0.172	-0.180	0.346	0.506	1	0.252	0.149	0.274	-0.432	-0.117	0.383	0.363	0.299	0.219	0.149	0.072	0.247
FFV 114	0.101	-0.039	-0.348	-0.132	-0.196	0.269	-0.010	0.252	1	0.606	0.402	-0.500	-0.144	0.443	0.425	0.350	0.257	0.175	0.087	0.301
FFV 115	0.025	-0.057	-0.085	-0.225	-0.221	0.146	0.058	0.149	0.606	1	0.607	-0.309	-0.055	0.281	0.244	0.194	0.135	0.097	0.057	0.170
FFV 116	0.099	-0.042	-0.349	-0.131	-0.191	0.251	-0.008	0.274	0.402	0.607	1	-0.501	-0.142	0.444	0.425	0.349	0.255	0.175	0.087	0.301
FFV 124	-0.158	0.061	0.567	0.101	0.092	-0.432	-0.059	-0.432	-0.500	-0.309	-0.501	1	0.263	-0.829	-0.827	-0.709	-0.570	-0.440	-0.248	-0.446
FFV 125	-0.347	-0.293	0.173	0.005	-0.004	-0.118	0.028	-0.117	-0.144	-0.055	-0.142	0.263	1	-0.160	-0.560	-0.730	-0.711	-0.624	-0.462	-0.143
FFV 126	0.063	-0.118	-0.506	-0.091	-0.087	0.381	0.059	0.383	0.443	0.281	0.444	-0.829	-0.160	1	0.552	0.458	0.267	0.064	-0.029	0.397
FFV 127	0.112	-0.067	-0.500	-0.079	-0.067	0.363	0.028	0.363	0.425	0.244	0.425	-0.827	-0.560	0.552	1	0.729	0.623	0.551	0.258	0.357
FFV 128	0.392	0.237	-0.385	-0.053	-0.040	0.301	0.014	0.299	0.350	0.194	0.349	-0.709	-0.730	0.458	0.729	1	0.772	0.660	0.484	0.335
FFV 129	0.386	0.261	-0.303	-0.036	-0.029	0.219	0.003	0.219	0.257	0.135	0.255	-0.570	-0.711	0.267	0.623	0.772	1	0.791	0.575	0.256
FFV 130	0.366	0.276	-0.187	-0.027	-0.013	0.151	0.005	0.149	0.175	0.097	0.175	-0.440	-0.624	0.064	0.551	0.660	0.791	1	0.684	0.182
FFV 131	0.477	0.420	-0.056	-0.001	-0.006	0.076	0.006	0.072	0.087	0.057	0.087	-0.248	-0.462	-0.029	0.258	0.484	0.575	0.684	1	0.127
FFV 132	0.284	0.169	-0.383	0.076	-0.203	0.241	0.018	0.247	0.301	0.170	0.301	-0.446	-0.143	0.397	0.357	0.335	0.256	0.182	0.127	1

Table A 3.17: Correlation coefficient of VG 1, 0.7ms

	FFV 5	FFV 6	FFV 7	FFV 8	FFV 9	FFV 10	FFV 11	FFV 12	FFV 13
FFV 5	1	0.9958	0.9877	0.9762	0.9597	0.9373	0.9065	0.864	0.7823
FFV 6	0.9958	1	0.9964	0.9893	0.9775	0.9597	0.9327	0.8934	0.8161
FFV 7	0.9877	0.9964	1	0.9966	0.9885	0.975	0.9533	0.9184	0.8443
FFV 8	0.9762	0.9893	0.9966	1	0.996	0.9862	0.9684	0.9389	0.8712
FFV 9	0.9597	0.9775	0.9885	0.996	1	0.995	0.9816	0.9567	0.8971
FFV 10	0.9373	0.9597	0.975	0.9862	0.995	1	0.9931	0.9737	0.9215
FFV 11	0.9065	0.9327	0.9533	0.9684	0.9816	0.9931	1	0.9897	0.9466
FFV 12	0.864	0.8934	0.9184	0.9389	0.9567	0.9737	0.9897	1	0.9746
FFV 13	0.7823	0.8161	0.8443	0.8712	0.8971	0.9215	0.9466	0.9746	1

Table A 3.18: Correlation coefficient of VG 2, 0.7ms

	FFV 14	FFV 15	FFV 16	FFV 17	FFV 18	FFV 19	FFV 20	FFV 21	FFV 22	FFV 23	FFV 24	FFV 25	FFV 26	FFV 27	FFV 28
FFV 14	1	0.9774	0.9494	0.9214	0.8923	0.9932	0.9856	0.9667	0.9431	0.9177	0.9246	0.92	0.918	0.914	0.9119
FFV 15	0.9774	1	0.9883	0.9714	0.9504	0.9587	0.9943	0.9921	0.9815	0.9651	0.9658	0.9643	0.9635	0.9617	0.9608
FFV 16	0.9494	0.9883	1	0.9931	0.9802	0.9233	0.9748	0.9944	0.9948	0.9869	0.9835	0.9842	0.9843	0.9841	0.9839
FFV 17	0.9214	0.9714	0.9931	1	0.9939	0.8908	0.9528	0.9818	0.9956	0.9947	0.9886	0.9908	0.9915	0.9925	0.9928
FFV 18	0.8923	0.9504	0.9802	0.9939	1	0.859	0.9282	0.9646	0.9853	0.996	0.9882	0.9915	0.9927	0.9945	0.9952
FFV 19	0.9932	0.9587	0.9233	0.8908	0.859	1	0.9783	0.951	0.9209	0.8918	0.9057	0.8979	0.8946	0.8884	0.8855
FFV 20	0.9856	0.9943	0.9748	0.9528	0.9282	0.9783	1	0.9894	0.9719	0.9511	0.9578	0.954	0.952	0.948	0.9461
FFV 21	0.9667	0.9921	0.9944	0.9818	0.9646	0.951	0.9894	1	0.9929	0.9801	0.9823	0.9807	0.9799	0.9775	0.9763
FFV 22	0.9431	0.9815	0.9948	0.9956	0.9853	0.9209	0.9719	0.9929	1	0.9943	0.9928	0.993	0.9928	0.9923	0.9917
FFV 23	0.9177	0.9651	0.9869	0.9947	0.996	0.8918	0.9511	0.9801	0.9943	1	0.9963	0.9976	0.998	0.9984	0.9984
FFV 24	0.9246	0.9658	0.9835	0.9886	0.9882	0.9057	0.9578	0.9823	0.9928	0.9963	1	0.9992	0.9988	0.9978	0.9972
FFV 25	0.92	0.9643	0.9842	0.9908	0.9915	0.8979	0.954	0.9807	0.993	0.9976	0.9992	1	0.9998	0.9992	0.9988
FFV 26	0.918	0.9635	0.9843	0.9915	0.9927	0.8946	0.952	0.9799	0.9928	0.998	0.9988	0.9998	1	0.9996	0.9993
FFV 27	0.914	0.9617	0.9841	0.9925	0.9945	0.8884	0.948	0.9775	0.9923	0.9984	0.9978	0.9992	0.9996	1	0.9999
FFV 28	0.9119	0.9608	0.9839	0.9928	0.9952	0.8855	0.9461	0.9763	0.9917	0.9984	0.9972	0.9988	0.9993	0.9999	1

Table A 3.19: Correlation coefficient of VG 3, 0.7ms

	FFV 30	FFV 31
FFV 30	1	0.9895
FFV 31	0.9895	1

Table A 3.20: Correlation coefficient of VG 4, 0.7ms

	FFV 33	FFV 34
FFV 33	1	0.9460
FFV 34	0.9460	1

Table A 3.21: Correlation coefficient of VG 5, 0.7ms

	FFV 32	FFV 35
FFV 32	1	0.9717
FFV 35	0.9717	1

Table A 3.22: Correlation coefficient of VG 6, 0.7ms

	FFV 36	FFV 37
FFV 36	1	0.8740
FFV 37	0.8740	1

Table A 3.23: Correlation coefficient of VG 7, 0.7ms

	FFV 38	FFV 39	FFV 40	FFV 41	FFV 42	FFV 43	FFV 44	FFV 45	FFV 46	FFV 47	FFV 48	FFV 49	FFV 50	FFV 51	FFV 52	FFV 53	FFV 54	FFV 55
FFV 38	1	0.9783	0.9786	0.9998	0.9733	0.9739	0.9993	0.9679	0.9687	0.9934	0.9861	0.9862	0.9934	0.985	0.9851	0.9932	0.9835	0.9836
FFV 39	0.9783	1	0.9987	0.9779	0.9566	0.9572	0.9772	0.9484	0.9492	0.9685	0.9717	0.9714	0.9685	0.961	0.9611	0.9682	0.9584	0.9585
FFV 40	0.9786	0.9987	1	0.9781	0.9569	0.9575	0.9774	0.9486	0.9496	0.9689	0.9718	0.9721	0.9688	0.9614	0.9615	0.9686	0.9588	0.9589
FFV 41	0.9998	0.9779	0.9781	1	0.9728	0.9734	0.9998	0.9674	0.9681	0.9938	0.9867	0.9868	0.9939	0.9855	0.9856	0.9939	0.9841	0.9842
FFV 42	0.9733	0.9566	0.9569	0.9728	1	0.9941	0.9719	0.9475	0.9482	0.9642	0.958	0.958	0.9642	0.9687	0.9676	0.9639	0.956	0.956
FFV 43	0.9739	0.9572	0.9575	0.9734	0.9941	1	0.9725	0.9478	0.9487	0.9648	0.9587	0.9587	0.9648	0.9682	0.9693	0.9645	0.9566	0.9566
FFV 44	0.9993	0.9772	0.9774	0.9998	0.9719	0.9725	1	0.9664	0.9671	0.9934	0.9863	0.9863	0.9936	0.985	0.985	0.9938	0.9836	0.9836
FFV 45	0.9679	0.9484	0.9486	0.9674	0.9475	0.9478	0.9664	1	0.9852	0.9601	0.9532	0.9532	0.9601	0.9537	0.9538	0.9598	0.9652	0.9625
FFV 46	0.9687	0.9492	0.9496	0.9681	0.9482	0.9487	0.9671	0.9852	1	0.9609	0.9541	0.9541	0.9609	0.9546	0.9546	0.9606	0.9633	0.9659
FFV 47	0.9934	0.9685	0.9689	0.9938	0.9642	0.9648	0.9934	0.9601	0.9609	1	0.9958	0.9958	1	0.9949	0.9949	0.9998	0.994	0.9941
FFV 48	0.9861	0.9717	0.9718	0.9867	0.958	0.9587	0.9863	0.9532	0.9541	0.9958	1	0.9999	0.9957	0.9948	0.9949	0.9955	0.9936	0.9937
FFV 49	0.9862	0.9714	0.9721	0.9868	0.958	0.9587	0.9863	0.9532	0.9541	0.9958	0.9999	1	0.9957	0.9948	0.9949	0.9955	0.9936	0.9937
FFV 50	0.9934	0.9685	0.9688	0.9939	0.9642	0.9648	0.9936	0.9601	0.9609	1	0.9957	0.9957	1	0.9948	0.9948	0.9999	0.9939	0.994
FFV 51	0.985	0.961	0.9614	0.9855	0.9687	0.9682	0.985	0.9537	0.9546	0.9949	0.9948	0.9948	0.9948	1	0.9994	0.9945	0.9943	0.9943
FFV 52	0.9851	0.9611	0.9615	0.9856	0.9676	0.9693	0.985	0.9538	0.9546	0.9949	0.9949	0.9949	0.9948	0.9994	1	0.9945	0.9943	0.9943
FFV 53	0.9932	0.9682	0.9686	0.9939	0.9639	0.9645	0.9938	0.9598	0.9606	0.9998	0.9955	0.9955	0.9999	0.9945	0.9945	1	0.9937	0.9937
FFV 54	0.9835	0.9584	0.9588	0.9841	0.956	0.9566	0.9836	0.9652	0.9633	0.994	0.9936	0.9936	0.9939	0.9943	0.9943	0.9937	1	0.9987
FFV 55	0.9836	0.9585	0.9589	0.9842	0.956	0.9566	0.9836	0.9625	0.9659	0.9941	0.9937	0.9937	0.994	0.9943	0.9943	0.9937	0.9987	1

Table A 3.24: Correlation coefficient of VG 8, 0.7ms

	FFV 56	FFV 57	FFV 58	FFV 59	FFV 60	FFV 61	FFV 62	FFV 63	FFV 64
FFV 56	1	0.8615	0.8615	0.9382	0.8583	0.8584	0.917	0.8544	0.8542
FFV 57	0.8615	1	0.9998	0.9568	0.9847	0.9848	0.9705	0.9732	0.9735
FFV 58	0.8615	0.9998	1	0.9569	0.9847	0.9849	0.9706	0.9732	0.9735
FFV 59	0.9382	0.9568	0.9569	1	0.9496	0.9497	0.9967	0.9412	0.9413
FFV 60	0.8583	0.9847	0.9847	0.9496	1	0.9991	0.9623	0.9836	0.9838
FFV 61	0.8584	0.9848	0.9849	0.9497	0.9991	1	0.9625	0.9836	0.9839
FFV 62	0.917	0.9705	0.9706	0.9967	0.9623	0.9625	1	0.9528	0.953
FFV 63	0.8544	0.9732	0.9732	0.9412	0.9836	0.9836	0.9528	1	0.998
FFV 64	0.8542	0.9735	0.9735	0.9413	0.9838	0.9839	0.953	0.998	1

Table A 3.25: Correlation coefficient of VG 9, 0.7ms

	FFV 65	FFV 66	FFV 67	FFV 68	FFV 69	FFV 70	FFV 71	FFV 72	FFV 73
FFV 65	1	0.9532	0.9532	0.9976	0.9434	0.9438	0.9932	0.937	0.9377
FFV 66	0.9532	1	0.9998	0.9656	0.995	0.9951	0.977	0.9926	0.9925
FFV 67	0.9532	0.9998	1	0.9657	0.9951	0.9951	0.9771	0.9926	0.9925
FFV 68	0.9976	0.9656	0.9657	1	0.9565	0.9569	0.9984	0.9502	0.9511
FFV 69	0.9434	0.995	0.9951	0.9565	1	0.9995	0.9692	0.9946	0.9948
FFV 70	0.9438	0.9951	0.9951	0.9569	0.9995	1	0.9695	0.9943	0.9948
FFV 71	0.9932	0.977	0.9771	0.9984	0.9692	0.9695	1	0.9634	0.9643
FFV 72	0.937	0.9926	0.9926	0.9502	0.9946	0.9943	0.9634	1	0.9988
FFV 73	0.9377	0.9925	0.9925	0.9511	0.9948	0.9948	0.9643	0.9988	1

Table A 3.26: Correlation coefficient of VG 10, 0.7ms

	FFV 74	FFV 75	FFV 76	FFV 77	FFV 78	FFV 79	FFV 80	FFV 81	FFV 82	FFV 83	FFV 84	FFV 85	FFV 86	FFV 87	FFV 88	FFV 89	FFV 90	FFV 91
FFV 74	1	0.9789	0.9791	0.9901	0.9753	0.9757	0.9851	0.9712	0.9719	0.9694	0.9509	0.9507	0.9711	0.9492	0.9494	0.9606	0.9452	0.9457
FFV 75	0.9789	1	0.9983	0.993	0.9961	0.9962	0.9944	0.9937	0.994	0.9228	0.9704	0.9701	0.962	0.9665	0.9667	0.9615	0.9597	0.9602
FFV 76	0.9791	0.9983	1	0.9931	0.996	0.9962	0.9945	0.9936	0.994	0.9231	0.9704	0.9704	0.9623	0.9665	0.9669	0.9616	0.9599	0.9603
FFV 77	0.9901	0.993	0.9931	1	0.9903	0.9906	0.9983	0.9877	0.9882	0.9424	0.9642	0.964	0.9755	0.9596	0.9598	0.9725	0.9532	0.9537
FFV 78	0.9753	0.9961	0.996	0.9903	1	0.9975	0.9924	0.9949	0.9951	0.9182	0.9672	0.9671	0.9589	0.9669	0.967	0.9592	0.9591	0.9596
FFV 79	0.9757	0.9962	0.9962	0.9906	0.9975	1	0.9925	0.9948	0.9951	0.9189	0.9675	0.9673	0.9593	0.9669	0.9674	0.9594	0.9594	0.9599
FFV 80	0.9851	0.9944	0.9945	0.9983	0.9924	0.9925	1	0.9904	0.9908	0.9344	0.9656	0.9654	0.9736	0.9599	0.9601	0.9732	0.9528	0.9533
FFV 81	0.9712	0.9937	0.9936	0.9877	0.9949	0.9948	0.9904	1	0.9963	0.9136	0.9652	0.9649	0.9566	0.9623	0.9625	0.9584	0.9576	0.9576
FFV 82	0.9719	0.994	0.994	0.9882	0.9951	0.9951	0.9908	0.9963	1	0.9147	0.9655	0.9654	0.9573	0.9628	0.963	0.9589	0.9576	0.9586
FFV 83	0.9694	0.9228	0.9231	0.9424	0.9182	0.9189	0.9344	0.9136	0.9147	1	0.9084	0.9082	0.9659	0.901	0.9012	0.9478	0.8955	0.896
FFV 84	0.9509	0.9704	0.9704	0.9642	0.9672	0.9675	0.9656	0.9652	0.9655	0.9084	1	0.9865	0.9478	0.958	0.9582	0.9482	0.9472	0.9475
FFV 85	0.9507	0.9701	0.9704	0.964	0.9671	0.9673	0.9654	0.9649	0.9654	0.9082	0.9865	1	0.9476	0.9578	0.9581	0.948	0.947	0.9475
FFV 86	0.9711	0.962	0.9623	0.9755	0.9589	0.9593	0.9736	0.9566	0.9573	0.9659	0.9478	0.9476	1	0.9346	0.9349	0.9897	0.9246	0.925
FFV 87	0.9492	0.9665	0.9665	0.9596	0.9669	0.9669	0.9599	0.9623	0.9628	0.901	0.958	0.9578	0.9346	1	0.9796	0.9323	0.9561	0.9564
FFV 88	0.9494	0.9667	0.9669	0.9598	0.967	0.9674	0.9601	0.9625	0.963	0.9012	0.9582	0.9581	0.9349	0.9796	1	0.9326	0.9563	0.9566
FFV 89	0.9606	0.9615	0.9616	0.9725	0.9592	0.9594	0.9732	0.9584	0.9589	0.9478	0.9482	0.948	0.9897	0.9323	0.9326	1	0.9204	0.9208
FFV 90	0.9452	0.9597	0.9599	0.9532	0.9591	0.9594	0.9528	0.9576	0.9576	0.8955	0.9472	0.947	0.9246	0.9561	0.9563	0.9204	1	0.9714
FFV 91	0.9457	0.9602	0.9603	0.9537	0.9596	0.9599	0.9533	0.9576	0.9586	0.896	0.9475	0.9475	0.925	0.9564	0.9566	0.9208	0.9714	1

Table A 3.27: Correlation coefficient of VG 11, 0.7ms

	FFV 94	FFV 95	FFV 96	FFV 102	FFV 103	FFV 104	FFV 110	FFV 111	FFV 112	FFV 118	FFV 119	FFV 120
FFV 94	1	0.9653	0.9367	0.9367	0.9401	0.9313	0.9309	0.9381	0.9317	0.9303	0.9359	0.9318
FFV 95	0.9653	1	0.9648	0.9394	0.953	0.9376	0.936	0.95	0.9366	0.936	0.9483	0.9376
FFV 96	0.9367	0.9648	1	0.9318	0.9399	0.9358	0.9305	0.9375	0.9315	0.9312	0.9365	0.9315
FFV 102	0.9367	0.9394	0.9318	1	0.965	0.9318	0.9353	0.9404	0.9318	0.9322	0.9364	0.9318
FFV 103	0.9401	0.953	0.9399	0.965	1	0.9645	0.9399	0.9536	0.9397	0.9379	0.9494	0.9382
FFV 104	0.9313	0.9376	0.9358	0.9318	0.9645	1	0.9311	0.939	0.9354	0.9298	0.9348	0.9301
FFV 110	0.9309	0.936	0.9305	0.9353	0.9399	0.9311	1	0.9646	0.9325	0.9356	0.9382	0.9309
FFV 111	0.9381	0.95	0.9375	0.9404	0.9536	0.939	0.9646	1	0.9649	0.9401	0.9532	0.9395
FFV 112	0.9317	0.9366	0.9315	0.9318	0.9397	0.9354	0.9325	0.9649	1	0.9311	0.9375	0.9349
FFV 118	0.9303	0.936	0.9312	0.9322	0.9379	0.9298	0.9356	0.9401	0.9311	1	0.9648	0.9358
FFV 119	0.9359	0.9483	0.9365	0.9364	0.9494	0.9348	0.9382	0.9532	0.9375	0.9648	1	0.9653
FFV 120	0.9318	0.9376	0.9315	0.9318	0.9382	0.9301	0.9309	0.9395	0.9349	0.9358	0.9653	1

Table A 3.28: Correlation coefficient of VG 12, 0.7ms

	FFV 93	FFV 97	FFV 101	FFV 105	FFV 109	FFV 113	FFV 117	FFV 121
FFV 93	1	0.6221	0.6184	0.6108	0.6166	0.6129	0.6131	0.6105
FFV 97	0.6221	1	0.6167	0.6207	0.6164	0.6174	0.6125	0.6132
FFV 101	0.6184	0.6167	1	0.6146	0.6207	0.6123	0.6138	0.6123
FFV 105	0.6108	0.6207	0.6146	1	0.6138	0.619	0.6108	0.614
FFV 109	0.6166	0.6164	0.6207	0.6138	1	0.6154	0.6233	0.6164
FFV 113	0.6129	0.6174	0.6123	0.619	0.6154	1	0.6166	0.6199
FFV 117	0.6131	0.6125	0.6138	0.6108	0.6233	0.6166	1	0.6206
FFV 121	0.6105	0.6132	0.6123	0.614	0.6164	0.6199	0.6206	1

Table A 3.29: Correlation coefficient of VG 13, 0.7ms

	FFV 122	FFV 123
FFV 122	1	0.9077
FFV123	0.9077	1

Table A 3.30: Correlation coefficient of VG 14, 0.7ms

	FFV 3	FFV 4
FFV 3	1	0.8491
FFV 4	0.8491	1

Table A 3.31: Correlation coefficient of VG 15, 0.2ms

	FFV 98	FFV 99
FFV 98	1	0.7432
FFV 99	0.7432	1

Table A 3.32: Correlation coefficient of VG 15 (Singletons), 0.2ms

	FFV 1	FFV 2	FFV 29	FFV 92	FFV 100	FFV 106	FFV 107	FFV 108	FFV 114	FFV 115	FFV 116	FFV 124	FFV 125	FFV 126	FFV 127	FFV 128	FFV 129	FFV 130	FFV 131	FFV 132
FFV 1	1	0.8648	-0.0046	0.0205	0.0392	-0.006	-0.0265	-0.0074	-0.016	-0.0476	-0.0193	-0.0435	-0.269	-0.2487	-0.1128	0.1577	0.2668	0.4016	0.5738	0.1224
FFV 2	0.8648	1	0.0934	-0.0151	0.0499	-0.0248	-0.0088	-0.0263	-0.0401	-0.0384	-0.0419	0.0574	-0.1455	-0.2003	-0.1437	0.0791	0.1494	0.2576	0.3384	0.1494
FFV 29	-0.0046	0.0934	1	-0.0422	0.0749	-0.0206	0.1016	-0.0216	-0.0535	0.0528	-0.0538	0.1946	0.1265	-0.085	-0.1398	-0.1447	-0.0458	-0.0239	-0.1231	-0.1055
FFV 92	0.0205	-0.0151	-0.0422	1	-0.2249	-0.0918	-0.1904	-0.0924	-0.0562	-0.1376	-0.0618	-0.1159	-0.0478	0.0659	0.0786	0.0612	0.0429	0.024	0.0185	0.0194
FFV 100	0.0392	0.0499	0.0749	-0.2249	1	-0.2818	-0.3155	-0.2645	-0.2794	-0.3382	-0.2774	0.1611	0.0146	-0.1414	-0.1056	-0.0483	-0.0086	0.018	0.0373	-0.2624
FFV 106	-0.006	-0.0248	-0.0206	-0.0918	-0.2818	1	0.7034	0.3028	-0.0402	-0.1175	-0.0486	-0.1715	-0.0888	0.1053	0.1372	0.1031	0.0681	0.0306	0.0052	0.1333
FFV 107	-0.0265	-0.0088	0.1016	-0.1904	-0.3155	0.7034	1	0.7053	-0.1513	-0.1811	-0.1453	-0.0321	0.0094	0.0431	0.0214	0.0011	0.0022	-0.0109	-0.04	0.1266
FFV 108	-0.0074	-0.0263	-0.0216	-0.0924	-0.2645	0.3028	0.7053	1	-0.0538	-0.1192	-0.0374	-0.175	-0.0917	0.1057	0.1421	0.1046	0.0721	0.0341	0.0041	0.1389
FFV 114	-0.016	-0.0401	-0.0535	-0.0562	-0.2794	-0.0402	-0.1513	-0.0538	1	0.7168	0.3123	-0.2355	-0.1096	0.1505	0.1877	0.1328	0.0773	0.0355	0.0007	0.1535
FFV 115	-0.0476	-0.0384	0.0528	-0.1376	-0.3382	-0.1175	-0.1811	-0.1192	0.7168	1	0.7171	-0.1458	-0.0237	0.1268	0.1081	0.052	0.0173	-0.0077	-0.0501	0.1585
FFV 116	-0.0193	-0.0419	-0.0538	-0.0618	-0.2774	-0.0486	-0.1453	-0.0374	0.3123	0.7171	1	-0.2331	-0.1045	0.15	0.1863	0.1294	0.0711	0.0297	-0.0007	0.1622
FFV 124	-0.0435	0.0574	0.1946	-0.1159	0.1611	-0.1715	-0.0321	-0.175	-0.2355	-0.1458	-0.2331	1	0.5151	-0.4847	-0.7139	-0.6199	-0.4994	-0.384	-0.1459	-0.2397
FFV 125	-0.269	-0.1455	0.1265	-0.0478	0.0146	-0.0888	0.0094	-0.0917	-0.1096	-0.0237	-0.1045	0.5151	1	-0.0054	-0.6222	-0.8051	-0.7609	-0.6569	-0.3422	-0.0839
FFV 126	-0.2487	-0.2003	-0.085	0.0659	-0.1414	0.1053	0.0431	0.1057	0.1505	0.1268	0.15	-0.4847	-0.0054	1	0.1495	0.0278	-0.1659	-0.3499	-0.3349	0.1757
FFV 127	-0.1128	-0.1437	-0.1398	0.0786	-0.1056	0.1372	0.0214	0.1421	0.1877	0.1081	0.1863	-0.7139	-0.6222	0.1495	1	0.5693	0.4271	0.3378	-0.0747	0.1544
FFV 128	0.1577	0.0791	-0.1447	0.0612	-0.0483	0.1031	0.0011	0.1046	0.1328	0.052	0.1294	-0.6199	-0.8051	0.0278	0.5693	1	0.7033	0.5315	0.1118	0.123
FFV 129	0.2668	0.1494	-0.0458	0.0429	-0.0086	0.0681	0.0022	0.0721	0.0773	0.0173	0.0711	-0.4994	-0.7609	-0.1659	0.4271	0.7033	1	0.7392	0.2763	0.061
FFV 130	0.4016	0.2576	-0.0239	0.024	0.018	0.0306	-0.0109	0.0341	0.0355	-0.0077	0.0297	-0.384	-0.6569	-0.3499	0.3378	0.5315	0.7392	1	0.4903	0.0281
FFV 131	0.5738	0.3384	-0.1231	0.0185	0.0373	0.0052	-0.04	0.0041	0.0007	-0.0501	-0.0007	-0.1459	-0.3422	-0.3349	-0.0747	0.1118	0.2763	0.4903	1	0.0003
FFV 132	0.1224	0.1494	-0.1055	0.0194	-0.2624	0.1333	0.1266	0.1389	0.1535	0.1585	0.1622	-0.2397	-0.0839	0.1757	0.1544	0.123	0.061	0.0281	0.0003	1