Semi-Automatic Core Characterisation based on Geochemical Logging Data

Menno Bloemsma

TU Delft
Delft
University of
Technology

Challenge the future

Cover photo: Satellite image of the Betsiboka River estuary, Madagascar (NASA, 2002, ref. ISS005-E-9418). Brick-red lateritic soils, as a the result of tropical weathering, are responsible for the strong color of the river water.

Semi-Automatic Core Characterisation based on Geochemical Logging Data

Menno Bloemsma, 25 august 2010

Msc thesis Geomatics

Department of Geotechnology, section of Applied Geology
Faculty of Civil Engineering and Geotechnology
Delft University of Technology
The Netherlands

Graduation Committee:

Prof. Dr. ir. P.J.M van Oosterom	Chairman
Prof. dr. S.M. Luthi	Graduation Professor
Dr. G.J. Weltje	Supervisor
Dr. R.C. Lindenbergh	Co-reader
Rik Tjallingii	External Committee-member

Preface

The thesis project presented here has been carried out as part of the MSc Geomatics curriculum. The research was performed at the Faculty of Civil Engineering and Geo Sciences at the Delft University of Technology, section of Applied Geology. My involvement in geochemical logging started when I did some basic programming work for users of this geochemical data. The mathematical beauty of the methods in combination with the potential of this under-explored technique resulted that for my thesis project I wanted to do ’something with’ geochemical data. In retrospect I could not have chosen a more satisfying project.

However this would not be the case without all the support I received. I owe my deepest gratitude to my supervisor Gert Jan Weltje, who was always willing to help me out. Gert Jan’s door was literally always open for either a short question or a long discussion. I also appreciated his invitation to the WGSG congress in Calabria very much. The presentations and discussions were really inspiring, as was the social event in the vineyards. I would like to express my thanks to Roderik Lindenbergh for his useful and always positive feedback. Furthermore I would like to say that it is an honor for me to have Rik Tjallingii in my graduation committee. From TNO’s side I would like to show my gratitude to Frank van den Belt for his efforts to realize this project. Also the support of Greg van de Bilt and Mike Burns (Panterra BV) is fully appreciated. From the NIOZ (Dutch Institute for Sea Research) I would like to thank Rineke Gieles for running the XRF core scans. Also the support by Bert de Wijn of Wintershall Noordzee BV by making core E10-3 available for analysis has given us the opportunity to do a ’show-case’ with this beautiful piece of rock. Finally I would like to express my thanks to everyone I cannot mention individually, which includes my friends and colleague students for their moral support, the coffee breaks and the sporadic table soccer games.

The only thing I can add to this, is that I hope that all readers will enjoy this topic as much as I do.

Menno Bloemsma
Delft, August 2010

Abstract

Textural variations in sediment are considered to be one of the primary controls on the geochemical composition. This relation has been widely exploited for the climatological interpretation of unconsolidated Quartenary sediment cores. Despite the common objectives in respectively unconsolidated and consolidated sedimentological studies, for the latter the use of novel techniques to acquire geochemical data, such as non-destructive XRF on cores, has not been applied yet. Because there is an ever increasing need for detailed (textural) sedimentological data, this study investigated the possibilities of in-situ XRF for semi-automatic textural characterization.

An improvement to the current work flow of XRF core scanning data acquisition was suggested by a pre-calibration filtering step, in order to increase the signal to noise ratio. Apart from that, a sample selection routine was developed which, given the underlying calibration strategy, was expected to yield improved calibration results. Embedding the two algorithms into the calibration work flow proved that the sample selection algorithm as well as the pre-calibration filtering step result in improved calibration output. They are also found to be effective for real data; when applied to two datasets of unconsolidated cores, similar results were obtained.

The second part of this study included the design of a sedimentary basin infill model ”LINMIX”, based on the mixing of different sediment sources with a grain size dependent chemical composition. By doing so, the observed geochemical signal is decomposed into a portion that is the result of textural variations and a portion that reflects differences in provenance. When applying LINMIX to a record of Quaternary sediment (offshore Senegal) as a proof-of-concept, the model was able to reconstruct the geochemical record satisfyingly by linear mixing of 3 endmembers with 3 unique Grain Size Distributions (GSD’s) and 2 unique functions characterising the compositional change in the grain size spectrum (Transfer function or ’TF’). This result implies that the chemical variation induced by the relative mixing of two endmembers was only the result of grain size variation and not of source material variation. The two endmembers that were designated a common TF, have in a previous study been interpreted as both reflecting material from the same sediment source. This implies that the LINMIX model has potential for semi-automatic provenance interpretation of sediment cores. Additionally the endmember compositions correspond fairly well with the present-day composition of the Senegal river and African eolian dust.

Finally it was investigated how geochemical data can support semi-automatic grain size prediction of consolidated sediment, for which holds that there is no sediment source variation. This was done using a dataset of Carboniferous material, which is highly variable in terms of lithofacies. Around 11 meter of core has been geochemically logged with an XRF core scanner and successfully calibrated using 40 calibration samples (20 unique sample depths). The quality of the core scanning data was found to be high; the main rock-forming elements were calibrated with a signal to noise ratio larger than 4. Subsequently the textural information content of the geochemical data was investigated in two different setups. Initially the data was used as a quantitative tool to fill in the gaps between the grain size derived from plug data using Multi Variate Regression (MVR). Apart from that it was used as input for an unsupervised Bayesian classification scheme in terms of grain size classes. Whereas the former yielded a residual variance on the input data smaller than 1.5 ϕ-units in 66 % of the cases, the latter identified two core sections that, given RGB and geochemical data, should be classified differently. In both cases, an additional validation step should give more insight in the performance of the scanner as a grain size predictor. Recommendations w.r.t the scanner include (1) embedding the suggested algorithms in the data acquisition work flow and (2) installing a higher resolution camera to extract textural proxies from images.

List of symbols

$\mathcal{S}^{\mathcal{D}}$	Notation of the D-component simplex.
W	Geochemical record of M observations (rows) and D columns (components).
I	Measured intensities, M observations (rows) and D columns (components).
R^2	Parameter quantifying the relative signal to noise of calibrated geochemical record.
δN^2	Global intrinsic noise parameter, unique for a dataset and a set of elements.
V	Matrix with eigenvectors (columns).
$\mathbf{V}^*$	Reduced eigenspace of l eigenvectors.
l	Intrinsic dimensionallity of the intensity data.
λ^{cum}	Cumulative eigenvalues in ascending order.
$\lambda^{\%}$	Parameter denoting the percentage of variance in the data on a principal axis.
Ψ	Discrepancy parameter, median clr-distance of k observations.
$\mathbf{T}$	Transformation matrix of compositions to their barycentric equivalent.
$\mathbf{c}$	Membership coordinates in terms of the barycentric coordinate system.
$\mathbf{a}$	Linear coefficients of the compositional change with membership in terms of $\mathbf{c}$.
$\mathbf{A_1}$	Constraints matrix to prevent fraction of component to be larger than 1.
$\mathbf{A_0}$	Constraints matrix to prevent fraction of component to be smaller than 0.
$\mathbf{A_c}$	Constraints matrix including $\mathbf{A_0}$ and $\mathbf{A_1}$.
$\mathbf{q}$	$2N \times 1$ vector with N ones and N zeros.
$\mathbf{H}$	Matrix with the barycentric coordinates that correspond to the observations, on the diagonal.
$\check{\mathbf{c}}_i$	$M \times N$ matrix of membership observations in terms of their barycentric equivalent.
$\hat{\mathbf{b}}_\mathbf{p}$	Partial bulk chemistry obtained from geochemical record.
$\mathbf{C}$	$G \times D$ matrix with the discrete TF.
$\mathbf{f}$	Discrete grain size distribution.
$\check{\mathbf{b}}_\mathbf{p}$	Partial bulk chemistry obtained from GSD and chemistry of the size fractions.
$\mathbf{b}$	Bulk chemistry.
$\mathbf{M}$	Mass fraction matrix.
$\mathbf{F}$	Matrix including the GSD's of all endmembers.
$\check{\mathbf{b}}_\mathbf{g}$	Chemistry of a grain size class obtained from the analysis of the size fractions.
$\mathbf{d_0}$	Compositional linear intersect.
$\mathbf{d_1}$	Compositional linear slope.
$\bar{\mathbf{b}}_\mathbf{p}$	Weighted average of $\check{\mathbf{b}}_\mathbf{p}$ and $\hat{\mathbf{b}}_\mathbf{p}$.
$\mathbf{D}$	Matrix reflecting the TF as obtained using the compositional linear parameters.
φ_g	Components of matrix $\mathbf{D}$, reflecting a weighted grain size class composition g.
c_b	Acceptance criterion for the bulk chemistry.
c_g	Acceptance criterion for the Transfer Function.
r	Relaxation parameter of acceptance criteria and objective function $\mathcal{H}$.
G	Number of isolated grain size classes of which the geochemistry is determined.
N	Number of grain size endmembers.
D	Number of observed elements in the mixture.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Introduction

In Quaternary geological studies, geochemical data is widely applied for paleo-environmental analysis (Sinha et al., 2006). In these studies, the objective is to reconstruct the geological and climatological setting by analysis of the stored product; sediment. The relative concentration of Fe and K for instance has been applied as a measure for the humidity under which sediment has been formed (Mulitza et al., 2008). Similarly element ratios have been applied as a grain size proxy (Dinelli et al., 2007) and for the direct discrimination between facies (Svendsen et al., 2007). One of the primary catalysts for the wide-spread application of geochemical data is the development of relatively fast and inexpensive geochemical logging techniques.

Analogous to the studies on Quaternary sediment, the importance of geochemical data is also recognized in hydrocarbon exploration studies. In exploration projects different types of data are acquired that somehow reflect a proxy of the geochemistry. The Gamma Ray logging technique for instance, measures the amount of K, Th and U in the formation by sensing the natural radio-activity of the formation. The acquisition of Gamma Ray logs is applied to obtain textural properties of the subsurface. Additionally *density* measurements of rock can also be considered a geochemical proxy. Nevertheless the direct geochemical analysis of sedimentary rock is not widespread in the hydrocarbon industry.

For as far as the data acquisition is concerned, there are difference in techniques between the two fields of expertise. However it also arises that although different techniques are being used and the material being analyzed is different (unconsolidated vs. consolidated sediment), the two fields of expertise share common objectives (i.e. grain size prediction).

1.2 Problem Statement

Although geology includes many different fields of expertise which focus on different processes, different scales and consequently different material, the focus of this study is on *sedimentary* rock. Additionally we will only focus on sediment of non-organic origin, which is also referred to as *siliciclastic sediment*. Sediment is the product of rock being exposed to wind, water, temperature differences and vegetation. This process is referred to as *weathering* and is commonly subdivided into *chemical* and *physical* weathering.

The process of chemical weathering has an impact on the composition of the residue because it results in the removal of elements into solution. The breakdown of feldspar into illite and silica for instance:

$$\begin{aligned} 2\mathrm{KAlSi_3O_8} + 3\mathrm{H_2O} &\rightarrow \mathrm{Al_2(SiO_5)(OH_4)} + 4\mathrm{SiO_2} + 2\mathrm{K(OH)} \\ \text{feldspar} + \text{water} &\rightarrow \text{kaolinite} + \text{silica} + \text{solutes} \end{aligned}$$

results in the removal of potassium into solution. This means that the feldspar is turned into a new set of minerals which in turn reflect different mineralogical characteristics. Although the mineralogical change of the feldspar into kaolinite is trivial, it also results into the removal of potassium and silica into the solution, therefore leading to a *change* in the *chemical composition* of the solid residue.

On the other hand, this chemical reaction does not only reflect a change in bulk composition of the solid residue. Additionally the dominant grain size of the bulk will decrease when turning feldspar into clay minerals such as kaolinite. As a result the chemical and textural changes are correlated, which illustrates the potential of geochemical data as a textural proxy.

Physical weathering on the other hand includes volumetric changes of the rock by the introduction of material for instance. Commonly this involves water however also vegetation or salt may intrude the rock causing to fracture. Additionally volumetric changes of the rock itself, causing fracturing, is considered physical weathering (Allen, 1997).

Rates of both chemical and physical weathering are controlled by presence of water, biological activity and temperature. More general, it is determined by the tectonic setting. This implies that weathering rates depend on the environment in which the sediment is formed, transported and deposited.

Sediment characteristics reflect the environment under which it was formed.

An intuitive way of looking at this is by thinking of the earth's surface as being a *sedimentary routing system*, as described in the work of Allen (1997). A sedimentary routing system reflects the closed system within which the sediment is generated, transported and deposited. This concept is also referred to as the conceptual source-to-sink model (figure 1.1), that comprises different segments, each reflecting different conditions in terms of

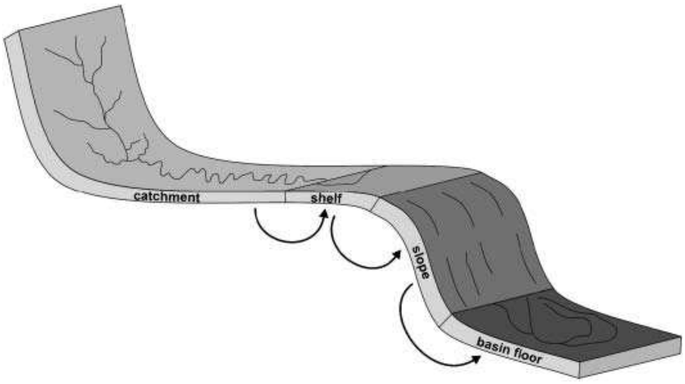

Figure 1.1: Schematic representation of the source-to-sink model of sediment generation and transport (modified from Somme et al. (2009)).

sediment generation and transport. Given that the rate of weathering is not (spatially and temporally) constant within this routing system, there is a gradient in terms of sediment generation. Additionally sediment transportation parameters (i.e. discharge) are not (spatially and temporally) constant, leading to a *fractionation* of sediment. Fractionation is the result of *selective transport*, which means that only size fractions of the weathering product are transported. For example, if the rainfall in a a sedimentary transport system is generally low, the expected sediment load of the river will be low and the sediment that will be transported under these energetically low conditions will primarily consist of fine material. On the other hand if discharge is high, more coarse material is transported to the basin. This means that:

> *Environmental conditions are reflected by both the grain size as well as the chemical composition of the sediment.*

Given the statement above, the chemical composition of a sedimentary basin contains valuable information with respect to the tectonic setting. However transport and sediment generation rates are not the only factors controlling the composition of the sediment. There is a whole set of processes involved that controls the generation of sediment which has been discussed in the work of Weltje and von Eynatten (2004). Figure 1.3 shows a graphical representation of their findings. As indicated in this figure, the source area or *parent lithology* also controls the composition of the sediment. A common parent material for sedimentary rock is igneous rock, which is formed by cooling and solidification of magma or lava. In figure 1.2, a general classification scheme of igneous rock is shown to illustrate the variable silica-content for the whole range of different igneous rock species. This means that if the parent material within one sedimentary routing system is not constant, the sedimentary record will show interfering signals of sediment originating from different parent material. The direct consequence is that one-to-one mapping of

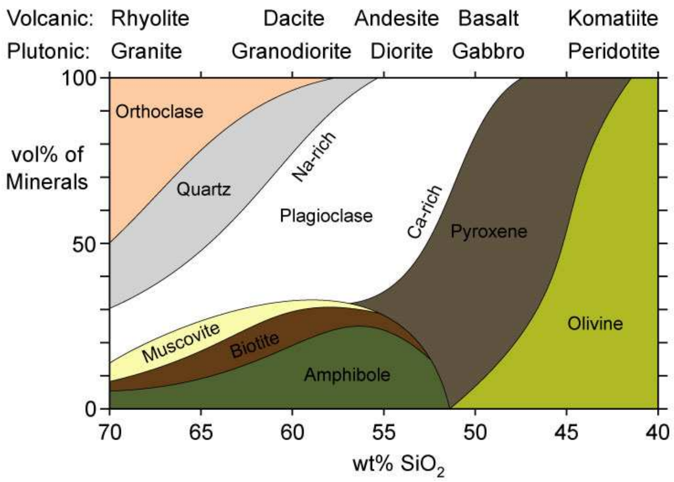

Figure 1.2: Volcanic/magmatic rock classification scheme.

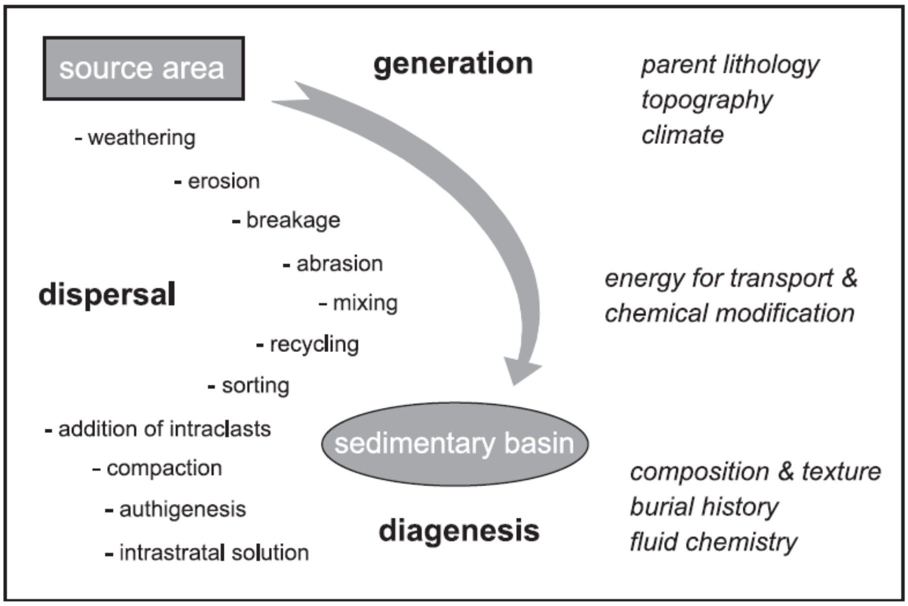

Figure 1.3: Schematic representation of the process of sediment generation, presented in the work from Weltje and von Eynatten (2004).

chemical composition to any of the components shown in figure 1.3 is only valid under the assumption that the effect of varying source rock composition is negligible. As a consequence:

> *The interpretation of a sedimentary record should incorporate a (spatial or temporal) variable source area.*

This final statement brings us to the objectives.

1.3 Objectives

The objectives of this study are threefold. From the statements given in the previous section we derive that although the geochemical signal is partially determined by the grain size, a direct conversion is more complicated. Therefore the general research question of this study is:

> **Can geochemical data be used to obtain textural properties of siliciclastic sediment?**

However first we will focus on the acquisition of geochemical data. For the acquisition of geochemical records a novel technique has proven itself in Quaternary geological studies as a fast and relatively inexpensive method to obtain high resolution geochemistry. However there is still room for improvement with respect to the data acquisition and processing. The first research question is therefore:

> **How can the current work flow of geochemical data acquisition be optimized to reduce costs and to increase accuracy and precision?**

The secondary objective in this study is to build a model that closely resembles the process of basin infill in terms of *both* geochemical and textural properties. The primary reason of building this model is that in order to derive textural properties from geochemical data, first a constructive *forward model* must be created. This forward model should incorporate the possibility of multiple sediment sources comprising different parent lithologies. Additionally it should be feasible in the sense that its *forward modeling* results are supported by real data. More formally:

> **Can we build a forward model of basin infill in a formal geochemical *and* textural context that successfully reconstructs real observations?**

Geochemistry-derived grain size prediction on consolidated sediment has some severe complications. On the other hand there is a strong commercial interest in this data. Although any direct grain size prediction is hampered by the factor of *parent lithology*, we want to find out how a direct prediction method using several data sources performs in the absence of parent lithology variation. More explicitly, this leads to the following research question:

> **Is it feasible to predict textural properties based on geochemistry, given multiple data sources and a known, straightforward provenance setting?**

We want to find an answer to this question in a quantitative sense. Recent developments have given the opportunity to acquire high resolution geochemical records in a non-destructive way. This method called X-Ray Fluorescence Spectroscopy (XRF) core scanning will be used throughout this thesis.

1.4 Methods

Most research projects start where previous work has stopped, which is also the case in this research project. Therefore the first step in finding answers to the research questions defined above is to investigate what is written about this topic in literature. This includes literature about the acquisition, processing and interpretation of (geo-)chemical data. After this literature study, we try to find answers to the research questions.

In order to answer the first question, we apply two different approaches. The objective includes the design and implementation of new steps in the current work flow in order to improve the quality of the geochemical record. We will initially test possible strategies and methods on synthetic data as a proof-of-concept. When a methodology is identified as a successful improvement of the work flow, it is tested on a real dataset. If the methodology is also found to be successful when applied to ’real’ data, it is adopted as a valuable processing step.

The second research question applies to the building of a model for basin infill. This model should have a physical foundation in the sense that no empirical methods are applied. Nevertheless the introduction of assumptions and simplification is inevitable, given the complex nature of the processes. Subsequently the model will be tested on a real dataset. For this experiment we apply a dataset of unconsolidated sediment that has been studied in great detail by geologists. Because of this, it is reasonable to say that the ’true’ history of the sediment is known. Furthermore an extensive dataset is available, comprising both textural and chemical data, which makes this core an ideal test case for the model. By applying this model to real data, we want to find out whether the model contributes to the geological interpretation. In other words, does the model yield results that are in the line with the geological interpretation. This would imply that it provides a semi-automatic interpretation.

The third and final research question stated that we want to investigate automatic grain size predictions and its quantitative performance. Previously we argued that direct estimation of grain size is not feasible because of differences in parent lithology. However we want to investigate the performance of a direct method given the absence of source rock variations. To investigate this, a core that originates from the Carboniferous is used for which this holds. The available data includes a comprehensive grain size and geochemical record. The geochemical record is obtained using XRF core scanning. For the grain size prediction, empirical and well-known classification methods will be used.

1.5 Thesis Structure

First a short introduction is presented about current methods to characterize sediment and sedimentary rock. Furthermore we will present a mathematical framework for the analysis of compositional data which, since geochemical data reflects composition as such, is used throughout the thesis. In chapter 3 some improvements are discussed to the current work flow of XRF core scanning data calibration (research question 1). In chapter 4 we will use XRF geochemical data in order to characterize sediment in terms of a basin infill model based on mixing (research question 2). Chapter 5 is about the possibilities of the XRF core scanning method for characterisation of consolidated cores. Therefore this chapter includes both the acquisition and processing of geochemical data as well as the application of this data for grain size prediction (research question 3). Finally chapter 6 will present the conclusions and recommendations.

Chapter 2

Sediment characterisation

In this chapter, we will discuss and illustrate the context of this study. First we will give a very general introduction on sediment generation and the current industry standard of subsurface characterisation. We will also sketch in very general terms why this research is relevant for hydrocarbon exploration and the methodology of geochemical logging of sedimentary cores.

2.1 The relevance of grain size information

A detailed understanding of sedimentary basins has economic and commercial importance. This because the majority of the water and hydrocarbon reservoirs is situated in ancient sedimentary basins, deep in the subsurface. The water or oil is situated in the *pore* structure of the sedimentary rock; a connected structure of open spaces between the sediment *grains*. Since the pore fluid (either water, oil, gas) is effectively trapped into this pore structure, the success of extracting the pore fluid from the sedimentary rock depends to a large extent on the properties of the sediment w.r.t. flow. We refer to these flow properties by the *permeability*; the ability of a material to transmit fluids.

The permeability is a quantitative measure for which holds that:

$$k \propto d^2 \tag{2.1}$$

given a material composed of perfectly spherical 'grains' (Graton and Fraser, 1935). Here the intrinsic permeability is denoted by k, where d denotes the mean pore diameter. It is not hard to imagine that the grain size and pore diameter are closely related. The proportionality constant in this equation is determined by other material properties such as path tortuosity (average path length of the flowing medium), porosity (fraction of air relative to the fraction of rock) and sorting (spread in grain diameters).

Although this relation holds for a situation of a packed bed (i.e. spherical objects

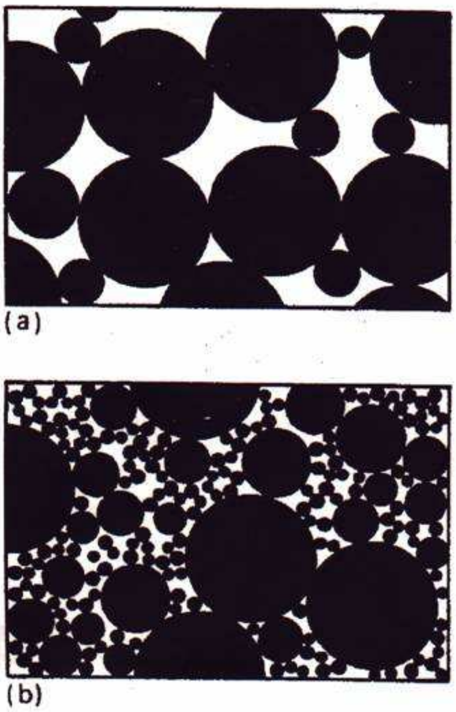

Figure 2.1: The sorting of the grains have a strong effect on the porosity. Figure (a) shows moderately sorted grains whereas figure (b) shows poorly sorted grains (modified from Brassington (1990).

and no cementation of/between these objects), it illustrates the importance of the grain diameter on the permeability. In a real reservoir though, these grains are not perfectly spherical and possibly cemented. Furthermore the pore structure is possibly cluttered with material leading to a severe decrease of permeability (Johnston and Beeson, 1945).

Another important formation parameter that is determined by the grain size is the *porosity*. Porosity is the fraction of open spaces in the bulk is referred to as the *porosity* and therefore ranges between 0 and 1, or alternatively 0 and 100 %. If the grains are poorly *sorted*, this implies a wide range of grain diameters in the sediment mixture. As a result, the small grains tend to fill up the pores, which leads to a strong decrease in the porosity and permeability (see figure 2.1).

A widely used unit system for grain size is the Krumbein ϕ-scale (Krumbein, 1941), which is a logarithmic scale of the grain diameter in the metric system. The grain size in ϕ-units is obtained as follows:

$$D = D_0 2^{-\phi} \tag{2.2}$$

where D is the grain diameter in mm and D_0 is a reference-diameter, which is usually 1 mm. Table 2.1 shows the ϕ-units with their equivalent grain diameters as well as their

ϕ	Particle Size	Name
> 8	$<4\mu$m	Clay
8 to 4	4-62 μm	Silt
4 to 3	62-125 μm	Very Fine Sand
3 to 2	125-250 μm	Fine Sand
2 to 1	0.25-0.5 mm	Medium Sand
1 to 0	0.5-1 mm	Coarse sand
0 to -1	1-2 mm	Very coarse sand
-1 to -6	2-64mm	Gravel

Table 2.1: Relation between ϕ-scale and grain diameters as well as the associated class-name

class-names.

2.2 Subsurface Characterisation

Reliable subsurface characterisation is important in the field of geotechnical project planning and hydrocarbon exploration (searching for hydrocarbons; oil and gas. The first step in characterising the subsurface consists usually of applying non-destructive geophysical methods. This includes electromagnetic techniques for the shallow subsurface and seismic techniques for the shallow and deep subsurface. However usually this data is complemented with data acquired in a well because of the limited horizontal and vertical resolution of seismic. Furthermore seismic processing partly relies on ground-truth data input. It also allows us to actually ’feel’ the rocks in the subsurface by analyzing the rock fragments present in the drilling mud.

In hydrocarbon exploration, it is common practice to perform wire line logging in order to characterize the drilled rock in terms of some physical parameters such as density, acoustic velocity, etc. Wire line logging is performed by lowering a sensor in a bore hole and record the data, referenced with the along-hole depth. There are many different types of detectors available that can be used on a wire line, but the ’standard’ logs include the density log, slowness log, resistivity log, and the gamma-ray (GR) log. The latter is primarily designed to characterize the formation in terms of texture.

A gamma-ray logging tool is a detector that records the natural radioactivity of the formation. In sedimentary formations this log normally reflects the shale (very fine-grained rock) content of the formation. This is because the radioactive (heavy) elements tend to concentrate in clays and shales (red. fine-grained sedimentary rock). Sandy formations on the other hand have a very low radioactivity, unless radioactive contaminants such as volcanic ash or granite wash are present, or when the formation water contains dissolved potassium salts (Serra, 1984). Gamma-rays are burst of high energy electromagnetic waves which are emitted spontaneously by some radioactive elements. Nearly all of the gamma radiation encountered in the earth is emitted by the radioactive potassium isotope of atomic weight 40 and the radioactive elements of the uranium and thorium series. Each of these elements emit gamma rays, the number and energies of which are distinctive of each element. Because in principle discrimination between K, Th and U is possible based on the energy spectrum of the gamma-rays, a Spectral Natural Gamma ray (SNG) tool was designed. This tool records the spectrum of the incoming gamma-ray after which the presence of K, U and Th is being determined.

Although wire line logging is still applied in most of the hydrocarbon and geotechnical projects, sometimes there is a need for a sample of the formation in its original form (in-situ). If the material of interest is still at the surface and unconsolidated, this can be obtained rather easily by using gravity coring. Gravity coring is the process of letting a tube penetrate the sediment by free fall from the ship to the sea bottom and retrieving it with the sediment still inside (see figure 2.3). For consolidated rock at depth the formation is being cored in a different way; the conventional drill bit is being replaced by a bit that extracts a piece of the formation in its original state (see figure 2.2). These formation samples are then subsequently transported ’up-hole’ for further analysis.

The major advantage of using cores instead of the log information is that the cores can be tested in a laboratory environment. They can be used to obtain reliable estimates of the porosity, permeability and mineralogical content of the formation. For consolidated cores, this data is obtained by analyzing ’core plugs’ which are in principal also cores; cylindrical samples taken from the cored rock. Core plugs are typically 1 to 1 1/2 inch in diameter and 1 inch to 2 inch long. They are ordinarily, when cut from a vertical well bore, cut perpendicular or parallel to the axis of the core, then called horizontal and vertical plugs, respectively.

Whereas the acquisition of this data is labor intensive and semi-automated, the opposite is true for the sedimentological characterisation. The sedimentological characterisation of a cored section is also referred to as the ’core description’ and consists of the analysis of the sedimentary rock w.r.t. several parameters. One of these parameters is the grain size and sorting (a measure of the standard deviation of the probability density function of the grains) which is being determined by visual inspection of the core surface. However also the presence of burrows and other sedimentological structures is determined in the same manner.

2.3 Compositional Data Analysis (CoDa)

The data obtained from either chemical analysis or XRF scanning methods yield element concentrations. These element concentration reflect the presence of certain elements in

Figure 2.2: Core drill bits (manufacturer Ceramco Pty Ltd) used to obtain cores of a formation.

Figure 2.3: Small gravity corer (manufacturer KC-Denmark) with in the middle a lead weight to increase the penetration performance.

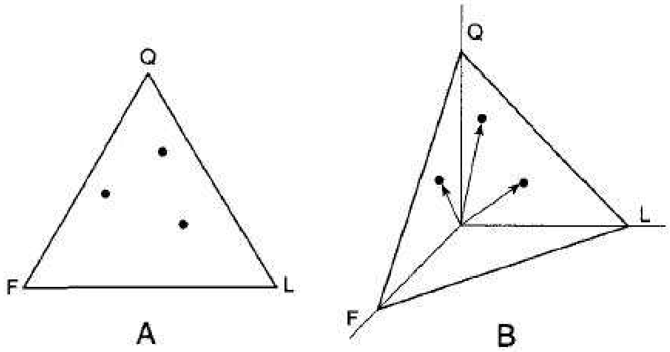

Figure 2.4: The position of compositional vectors on the plane originating from the unit-sum constraint. Modified from Weltje (1997).

the measured sample which represent the 'whole'. In chemical analysis these element proportions are given in parts per million (ppm), percentages or in $\frac{\mathrm{gr}}{\mathrm{kg}}$. The analysis of this so called compositional data is affected by its proportional nature such that the n-dimensional compositional space is essentially in $\mathbb{R}^{n-1}$. One way of illustrating this is by looking at a classical way of representing compositions made up of three *endmembers*. These endmembers represent the 'pure' components, either physical or not physical. With 'physical' endmembers we mean that the pure endmembers are also observed in the mixture; if we measure samples consisting of a mixture of Fe and K, there is no guarantee that we will find a sample that consists of 100% K or 100% Fe, respectively.

Since the relative proportions of all endmembers should sum up to either 1 (proportions), 100 (%) or 1.000.000 (ppm), all compositional vectors will be situated on a 2 dimensional plane restricted to the positive quadrant of a 3D Cartesian coordinate system (see figure 2.4). The position where these compositional vectors are on this plane reflect the relative proportions of the three endmembers. This is often called the simplex, of which a graphical representation, as shown in figure 2.4, is called a ternary diagram.

The compositional nature of the data also manifests itself in the dependency between the components; if the fraction of one of the components is perturbed, the fraction of the other components will by definition also be affected as a result of the unit-sum constraint.

To put these findings in more mathematical terms, the compositional space is being referred to as the simplex $\mathcal{S}^D$ which can be written as:

$$\mathcal{S}^D = \mathbf{x} = [x_1, x_2, ...x_D] \, | x_i > 0, i = 1, 2, ...D; \sum_{i=1}^{D} x_i = \kappa \tag{2.3}$$

where κ represents the sum-constraint; either unity, 100% or 1.000.000 for ppm units,

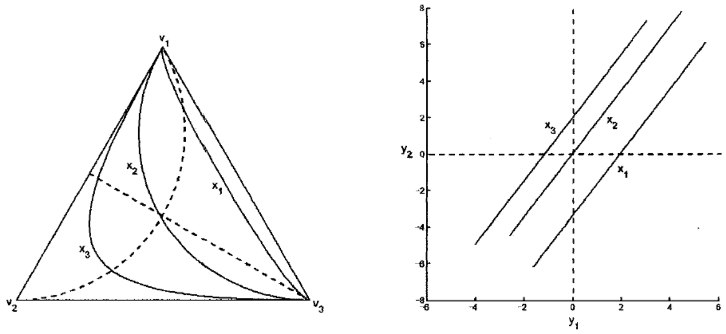

Figure 2.5: So called 'parallel' compositions (changing relative concentration of two endmembers) have a highly nonlinear appearance in compositional space (left). However this effect is eliminated in log ratio space (right).

and x_i represent the concentration of a certain element i expressed in one of the compositional units κ. As illustrated in figure 2.4, any vector $\mathbf{x}$ will lay on a simplex, whereas κ determines the relative position of this simplical plane w.r.t. the origin.

Whereas the choice of κ may seem an important convention, transforming the dataset from one compositional units system to another is a matter of *closing* the dataset to the chosen sum. This *closure* operation for all components $i = 1, 2, ...D$ follows from:

$$\mathcal{C}(\mathbf{z}) = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^{D} z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^{D} z_i}, ..., \frac{\kappa \cdot z_D}{\sum_{i=1}^{D} z_i} \right] \tag{2.4}$$

Due to the constrained compositional space and the fact that the individual compositions are intrinsically correlated, any direct analysis of compositional data is only possible when keeping track of these constraints. In order to overcome the compositional data constraints in the simplex, Aitchison (1986) built the log ratio framework which has been extended until today. It consists of a set of mathematical transformations and operations suitable for compositional quantities, which we refer to as the *CoDa* (Compositional Data Analysis) framework. The most intuitive compositional transformation is the additive log ratio, or 'alr'. In case of the alr-transformation, the data is being transformed from the simplex $\mathcal{S}^D$ into an unconstrained real space $\mathbb{R}^{D-1}$ as follows:

$$y_{i,D} = \ln\left(\frac{x_i}{x_D}\right), i = 1, 2, ...D - 1 \tag{2.5}$$

The main advantage of using the alr-transform is that, instead of the individual proportions, the transformed quantities are no longer hampered by their compositional nature and can be analyzed as independent variables. The convenience of the logratio-transformed quantities is illustrated in figure 2.5, where the right figures shows the same

3-component ’process’ as shown in the left figure, however now as a function of the 2 logratio transformed quantities y.

While the alr transformation is determined by the choice of the denominator and thereby not symmetric, the centered log ratio applies the geometric mean as denominator (with D the number of elements):

$$y_i = \log\left(\frac{x_i}{\text{geom}\,(x)}\right), i = 1, 2, \ldots. D - 1 \tag{2.6}$$

The transformation results for an compositional dataset comprising D elements and thereby spanning a constrained space in $\mathcal{S}^D$, into a dataset centered around zero and spanning an *unconstrained* $\mathbb{R}^D$ space. In contrast to the alr approach, the clr transformed data is again (as it was in compositional space) situated on a plane. However on this (hyper-)plane, the solution space stretches infinitely in all directions.

2.3.1 The compositional distance

Direct analysis of compositional data was considered difficult due to the simplicial constraints. However the introduced log ratio transformed dataset comprises an Euclidean vector space. Therefore we are able to apply all vector operations to the compositional dataset such as the norm, distance and inner product.

The ’compositional distance’ between two observations for instance can now be calculated in a straightforward manner without having to incorporate any of the simplicial constraints. In mathematical terms the compositional distance , or ’Aitchison distance’, between two observation is given by:

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{D} \left[\ln\frac{x_i}{\text{geom}(x)} - \ln\frac{y_i}{\text{geom}(x)}\right]^2} \tag{2.7}$$

which is simply the Euclidean distance or *norm* between the two logratio D-component data vectors x and y. The implications of logratio transformed data now comprising an Euclidean vector is that we may use multivariate techniques that do not rely on a full rank covariance matrix. Furthermore optimisation and adjustment techniques such as the least squares approach can now be applied by minimising the *compositional* distance.

2.3.2 The simplicial operators

Although the log ratio transformed data makes it possible to turn the compositional data into unconstrained data, we sometimes wish to ’stay in the simplex’. If we for instance wish to apply a perturbation in the *log ratio space*, however we wish to *apply* this shift to the data in its simplicial form, we need compositional equivalent operators to do this. Fortunately a framework exists to apply operations in log ratio space by making calculations with compositional data. It was found that a perturbation of two compositions for a vector space structure in $\mathcal{S}^D$ yield (Aitchison, 1986):

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1y_1, x_2y_2, ...x_Dy_D] \tag{2.8}$$

where $\mathcal{C}[..]$ represents the closure operation, as described in section. The simplicial equivalent of the power transformation by a scalar α of a composition in $\mathcal{S}^D$ was found to be:

$$\alpha \odot \mathbf{y} = \mathcal{C}[x_1^{\alpha}, x_2^{\alpha}, ...x_D^{\alpha}] \tag{2.9}$$

of which the purely compositional equivalent relation is as follows:

$$\mathbf{x} \odot \mathbf{y} = \mathcal{C}[x_1^{y}{}_1, x_2^{y}{}_2, ...x_D^{y}{}_D] \tag{2.10}$$

These relations between the operators in the simplicial vector space and the log ratio space facilitate to move freely between both space structures. If we want to use least squares adjustment for instance, we wish to minimize the distance between observed and predicted. We can apply curve fitting analogous to the 'classical' curve fitting by minimising the Aitchison distance for the following compositional-linear model:

$$\mathbf{y} = \mathbf{d_o} \oplus \mathbf{d_1} \odot \mathbf{x} \tag{2.11}$$

For the proof of equivalence between the Euclidean operators on logratio's and the simplicial operators to compositional quantities, I refer to the work of Aitchison (1986).

Chapter 3

Improvements to the XRF core scanning work flow

In the previous chapter we illustrated the commercial and scientific importance of grain size. The first research question that we defined in chapter 1 was:

> **How can the current work flow of geochemical data acquisition be optimized to reduce costs and to increase accuracy and precision?**

However since the acquisition of geochemical records in this study is primarily done using non-destructive XRF, first a brief introduction about XRF is presented. Secondly the latest improvements in work flow are discussed, as published by Weltje and Tjallingii (2008). Based on their work, a number of improvements are suggested that may lead to the reduction of cost and increase in signal to noise ratio. Finally the discussion and conclusions are presented in the last section.

3.1 X-radiation Fluorescence Spectrometry

X-ray Fluorescence Spectrometry, or XRF is a widely used technique for the chemical analysis of specimens. The methodology is based on the phenomenon that when materials are exposed to short-wavelength (highly energetic) X-rays or gamma rays, ionisation of their component atoms may take place. Ionisation consists of the ejection of one or more electrons from the atom, and may take place if the atom is exposed to radiation with an energy greater than its ionisation potential.

X-rays and gamma rays can be energetic enough to expel tightly held electrons from the inner orbitals of the atom. The removal of an electron in this way renders the electronic structure of the atom unstable, and electrons in higher orbitals ’fall’ into the lower orbital to fill the hole left behind. In falling, energy is released in the form of a photon, the energy of which is equal to the energy difference of the two orbitals involved. Thus, the

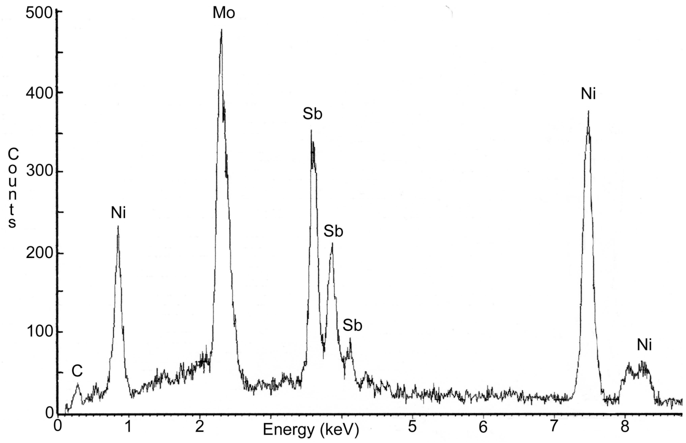

Figure 3.1: Energy spectrum with some distinct energy peaks with their associated elements (modified from Imrich (2001)).

material emits radiation, of which the energy spectrum provides a proxy of the presence of the different components (Jenkins, 1999). The term fluorescence in XRF, stands for the phenomenon in which the absorption of radiation of a specific energy results in the re-emission of radiation of a different energy (generally lower).

By removal of an inner electron by an energetic photon, an electron from an outer shell drops into its place. However there are a limited number of ways in which this can happen. The main transitions are given names: an L→K transition is traditionally called Kα, an M→ K transition is called Kβ. Each of these transitions yields a fluorescent photon with a characteristic energy equal to the difference in energy of the initial and final orbital. By sorting the energy channels and their corresponding counts, one obtains a spectrum of which an example is given in figure 3.1.

Typically these spectra are converted to 'element intensities' by associating these peaks to the elements using multivariate methods. One of the major problems of the interpretation of these spectra is the result of the frequency overlap between the characteristic lines of adjacent elements (Swerts and Van Espen, 1993). As a consequence, the method requires the use of sophisticated spectrum evaluation methods to obtain interference-free and background-corrected net peak areas. These spectrum evaluation methods mainly rely on the least squares principle; analytical functions describing the features in the spectrum or reference spectra of pure elements are used as a model and fitted the to observed spectrum.

A variety of quantisation procedures are in use, ranging from simple linear calibration curves, over empirical and semi-empirical influence coefficient to full fundamental parameter approaches (Dyck, van et al., 1986). An often used software package that supports these parametric curve fitting methodology is called 'Axil' (Analysis of X-ray

Figure 3.2: Avaatech XRF core scanner.

spectra by Iterative Least squares) which supports the use of the least squares framework for a variety of analytical functions. A windows-based version of Axil, i.e. WIN-Axil, is being used for the processing of the XRF data in this project.

3.2 XRF core scanning

Classical XRF spectrometry analysis is performed on pulverized and homogenized samples that are normally prepared as a flat disc, typically between 20 and 50 mm in diameter. However a *nondestructive* logging technique based on XRF methods has been developed for the analysis of split sediment cores (Jansen et al., 1998). This so called *XRF core scanner* (see figure 3.2) is able to acquire XRF measurements directly at the surface of the core. Typically reliable spectra are obtained relatively fast (1 measurement takes up to 30 seconds).

In contrast to ’classical’ XRF sample analysis, this core scanner is able to extract an almost continuous geochemical record; the down core resolution can go up to 1 millimeter. Finally, it provides data about the actual composition of the sediment at a very specific and confined area, which in contrast to tools such as natural-GR and color loggers that respectively average over a large area or produce merely a proxy of the actual composition superimposed with noise.

One of the drawbacks though of this in-situ methodology is the limited penetration of the X-radiation into the material. The incident X-ray beam interacts with small sample volumes, so the emitting characteristic X-rays contain information from a thin (microns) layer only (Jansen et al., 1998). The response depths of the elements in the sample material depend on the wavelength of the fluorescent radiation and the chemical composition of the matrix (Jenkins and De Vries, 1970). For the light elements Al and Si, this depth is about a few μm, for Ca several tenths of μm, and for Fe a few hundreds of μm.

With the core scanner one obtains energy spectra at all sampled locations with the counts (number of X-ray photons) and the frequency (energy) of these photons. However, as discussed in section 3.1, these spectra are not an easily interpretable quantity and therefore they undergo a processing step which yields an area quantity under a fitted curve. So in principle the dataset obtained after spectral analysis involves an area quantity which is assumed to correlate with the concentration of a specific element in the bulk. Stays the question though, how these area quantities can be turned into (relative) concentrations.

Although the area quantity has the appearance of a concentration, several aspects are not taken into account. The penetration depth for instance is not the same for all elements. Additionally the sensor does not have the same efficiency for all elements, leading to a bias. How the XRF core scanning data can be calibrated effectively will be discussed in the next section.

3.3 Calibration of the XRF core scanner

The traditional calibration approach in calibration of the XRF data is based on trying to estimate several coefficients that are related to measurement geometry, instrumental response, etc. In order to extract weight fractions from intensities with elements j and measured specimens i the following model is applied;

$$W_{ij} = K_i I_{ij} M_{ij} S_j \tag{3.1}$$

where W_{ij} represents the element concentration of element i in specimen j, K represents an element-specific device calibration coefficient, M_{ij} is the matrix effect that accounts for scatter, absorption and enhancement effects on I_{ij} caused by the presence of other elements in the specimen. Furthermore S_i represents the specimen effect which captures the measurement geometry and specimen homogeneity relative to the standard configuration. And finally I_{ij} is the intensity of element i in specimen j, obtained by post processing of the raw spectrum by background subtraction, sumpeak and escape-peak correction, deconvolution and peak integration (Weltje and Tjallingii, 2008).

However in the work of Weltje and Tjallingii (2008), a more robust calibration scheme was presented. This approach is based on the CoDa-framework and therefore on logratio's rather than the calibration on an element-by-element basis. Weltje and Tjallingii (2008) stated that "the fundamental problem in this approach is that S_i is very poorly constrained, owing to inhomogeneity of the specimens (e.g. variable water content and grain-size distribution, and the presence of burrows) and the irregular surface of a core". In some setups, spatial variations in thickness of an adhesive pore-water film, which forms directly below a protective foil covering the split core surface, also have to be considered a an error source (Tjallingii et al., 2007). As a result of these uncontrollable variations, S_i is unknown, and will cause biased compositional predictions W.

Weltje and Tjallingii (2008) found that the log ratio framework facilitates a more robust and unbiased calibration process. Given the components i and D in the specimen j, the following holds:

$$\frac{W_{ij}}{W_{Dj}} = \frac{K_i}{K_D}\frac{I_{ij}}{I_{jD}}\frac{M_{ij}}{M_{jD}}\frac{S_j}{S_j} \tag{3.2}$$

given that:

$$\frac{S_j}{S_j} = 1, \qquad \tilde{K}_{iD} = \frac{K_i}{K_D}, \qquad \tilde{M}_{iDj} = \frac{M_{ij}}{M_{Dj}} \tag{3.3}$$

this yields the following calibration equation:

$$\frac{W_{ij}}{W_{Dj}} = \tilde{K}_{iD}\left(\frac{I_{ij}}{I_{Dj}}\right)\tilde{M}_{ijD} \tag{3.4}$$

The relative detection efficiency $\tilde{K}$ is independent of j, and can be thought of as the average response of the measurement device during scanning of a core, i.e., a constant unique to elements i and D. The relative matrix effect $\tilde{M}$ varies with j, because it is a function of the chemical composition. Weltje and Tjallingii (2008) stated that if it is assumed that no prior information on the relative matrix effect is available, there is no choice but to derive a general expression for $\tilde{M}$ from the core-scanner measurements:

$$\tilde{M} = \left(\frac{I_{ij}}{I_{Dj}}\right)^{A_{Di}} \tag{3.5}$$

which, when plugged into equation 3.4 yields:

$$\left(\frac{W_{ij}}{W_{Dj}}\right) = \tilde{K}_{iD}\left(\frac{I_{ij}}{I_{Dj}}\right)^{A_{iD}+1} \tag{3.6}$$

They derived the more general expression by using the compositional operators and linear parameters α and β:

$$\begin{gathered} \alpha = A_{iD} + 1 \\ \\ \beta = \ln\left(\tilde{K}_{iD}\right) \end{gathered} \tag{3.7}$$

which yields the following Log-Ratio Calibration Equation (LRCE):

$$\ln\left(\frac{W_{ij}}{W_{Dj}}\right) = \alpha_{i,D}\ln\left(\frac{I_{ij}}{I_{Dj}}\right) + \beta_{i,D} \tag{3.8}$$

The practical application of this LRCE is illustrated in figure 3.3. Here can be observed that with the introduction of logratio's, the logratio intensities (x-axis) behave linear with respect to the logratio relative concentrations (y-axis). Whereas the depicted observations only represent the calibration samples, any new intensity observation can be turned into relative concentrations using the linear model (red dashed line). The resulting LRCE prediction as function of the composition of the calibration samples is shown in figure 3.3 from which can be derived that they line up perfectly along the line x=y. This implies that the compositional prediction based on XRF core logging is unbiased.

Summarizing, the application of the LRCE on XRF core scanning data of an unconsolidated core yielded good results. Not only did making use of element ratio's effectively cancel out all non-linearities induced by the compositional nature of the data (intrinsic correlation of component concentration) and the unknown measurement geometry, the LRCE is also expected to effectively reduce coherent noise caused by instrumentation. This expected noise reduction is caused by the fact that a certain part of noise and measurement error is uncorrelated with the transmitted wavelength. By being present in both the numerator and the denominator in the log ratio transformed data, this noise component is drastically reduced.

Furthermore the LRCE model is simple in terms of the number of coefficients that must be estimated from the data; only a intersect and a slope should be determined for every element pair *i*,*D* (see figure 3.3). This small number of coefficients facilitates redundancy w.r.t. the estimation of these coefficients. And although the calibration process is performed in the log ratio space, the results can be transformed back into relative concentrations using the alr inverse transformation (see figure 3.4).

By using the fact that the log ratio quantities span a real space, a robust statistical framework was defined by Weltje and Tjallingii (2008) using the distance between alr-transformed observed W and predicted composition $\hat{W}$ for the element-pair i and D. Given that n calibration samples are available, then this will yield n distances between observed and predicted for every element-pair of which the median value is being derived in the following manner:

$$s_{iD}^2 = \mathrm{med}_n \left[\ln\frac{\hat{W}_{in}}{\hat{W}_{Dn}} - \ln\frac{W_{in}}{W_{Dn}} \right]^2 \tag{3.9}$$

A parameter representing the signal-to-noise ratio could is defined by the total variance V of the signal of the element-pairs i and D as follows:

$$R_{iD}^2 = \frac{V_{iD} - s_{iD}^2}{V_{iD}} \tag{3.10}$$

We illustrate this signal to noise parameter with the following synthetic example comprising four observations and three components, calibrated by the LRCE;

$$W = \begin{matrix} 0.5412 & 0.1576 & 0.3012 \\ 0.3576 & 0.1915 & 0.4509 \\ 0.1871 & 0.4564 & 0.3565 \\ 0.3985 & 0.1152 & 0.4863 \end{matrix} \qquad \hat{W} = \begin{matrix} 0.5432 & 0.1191 & 0.3376 \\ 0.3146 & 0.1345 & 0.5510 \\ 0.1709 & 0.5371 & 0.2920 \\ 0.4217 & 0.1063 & 0.4720 \end{matrix} \tag{3.11}$$

the log ratio quantities (see equation 2.5) with the third column as denominator are obtained as follows;

$$\hat{W}^t = \begin{matrix} \ln\frac{0.5432}{0.3376} & \ln\frac{0.1191}{0.3376} \\ \ln\frac{0.3146}{0.5510} & \ln\frac{0.1345}{0.5510} \\ \ln\frac{0.1709}{0.2920} & \ln\frac{0.5622}{0.2920} \\ \ln\frac{0.4217}{0.4720} & \ln\frac{0.1063}{0.4720} \end{matrix} = \begin{matrix} 0.4756 & -1.0416 \\ -0.5604 & -1.4104 \\ -0.5360 & 0.6093 \\ -0.1128 & -1.4905 \end{matrix} \tag{3.12}$$

and:

$$W^t = \begin{matrix} \ln\frac{0.5412}{0.3012} & \ln\frac{0.1576}{0.3012} \\ \ln\frac{0.3576}{0.4509} & \ln\frac{0.1915}{0.4509} \\ \ln\frac{0.1871}{0.3565} & \ln\frac{0.4564}{0.3565} \\ \ln\frac{0.3985}{0.4863} & \ln\frac{0.1152}{0.4863} \end{matrix} = \begin{matrix} 0.5861 & -0.6478 \\ -0.2318 & -0.8562 \\ -0.6446 & 0.2477 \\ -0.1990 & -1.4398 \end{matrix} \tag{3.13}$$

The residual variance of the two alr transformed quantities is now determined by taking the median value of the squared difference between predicted and observed composition, which yields;

$$\begin{gathered} s^2_{1,3} = 0.0120 \qquad s^2_{2,3} = 0.1429 \\ V^2_{1,3} = 0.2352 \qquad V^2_{2,3} = 0.9631 \\ R^2_{1,3} = 0.949 \qquad R^2_{2,3} = 0.8516 \end{gathered} \tag{3.14}$$

This results should be interpreted as the first component having a significantly smaller residual variance as well as a significantly smaller total variance. Furthermore this first component reflects a much higher signal-to-noise parameter R^2.

For this synthetic example the number of calibration samples is small, making this parameter very sensitive to outliers. However when the number of samples is significantly larger, these statistical parameter can be considered a reliable estimate of the uncertainty of the data.

3.4 Improvements to the workflow

The work of Weltje and Tjallingii (2008) lead to the following three conclusions:

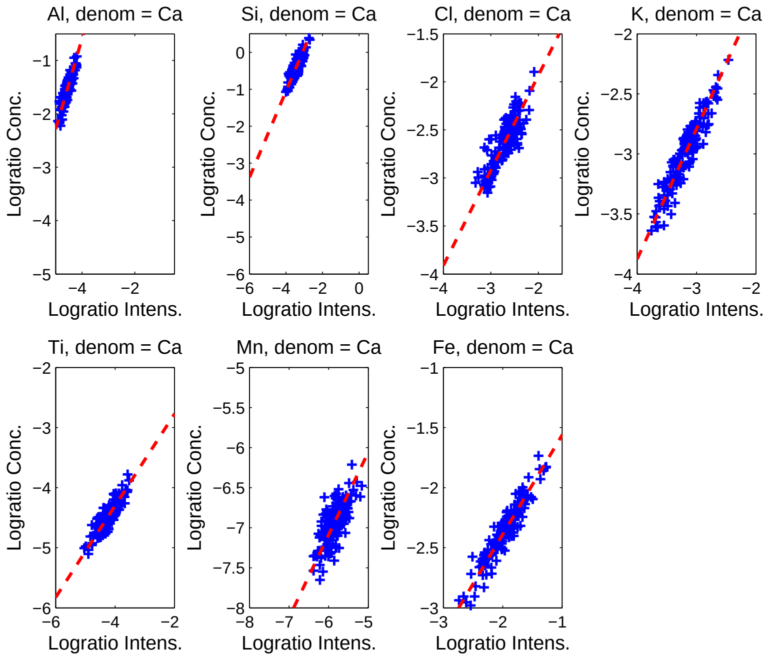

Figure 3.3: Linear fit between log ratio measured intensities and log ratio of the reference bulk chemistry of core 7920.

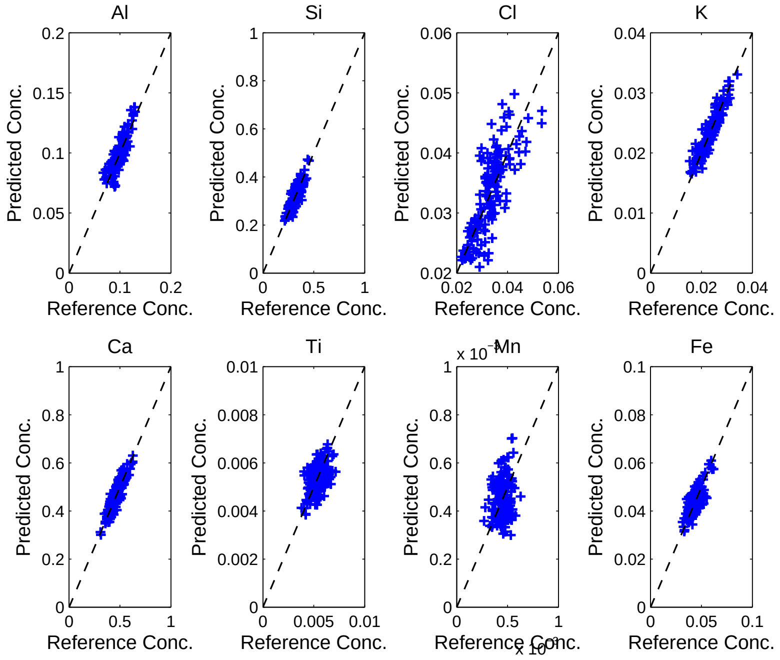

Figure 3.4: The reference and predicted bulk chemistry of core 7920.

1. Unbiased and accurate calibration results are possible using the LRCE.
2. A stabilisation of the prediction discrepancy is obtained using around 40 calibration samples.
3. By using the CoDa framework, 'classical' statistical techniques can provide insight in the quality of the process.

This study includes a direct follow-up on their recommendation that further research is needed to investigate the possibility to acquire a faster stabilisation of the prediction discrepancies. Therefore we have defined the following sub-questions:

> *Can clever use of multi variate methods result in an improved signal to noise ratio?*

and additionally a question concerning the sample selection:

> *Can a more robust sample selection strategy result in a faster stabilisation of the prediction discrepancies?*

In order to get an answer to the first research question, we will first discuss a method to determine the repeatability of the system.

3.5 Repeatability of the System

Because the acquisition of XRF core scanning data is relatively inexpensive compared to standard geochemical analysis, reliable estimates of the instrumental repeatability can be obtained with relatively low cost. It is also investigated how these reliable repeatability estimates can provide an improvement in the signal to noise ratio of the calibrated output.

The repeatability of a measurement system is a measure of the variation in measurements taken of the same sample and under the same conditions. For the XRF core scanning we can quantify this by keeping the scanning device at the same location while taking several measurements. The residual variance that we then observe after spectral analysis and quantification is considered as noise since it does not correlate with compositional variations in the specimen (or a location on the core). We can quantify this residual variance by taking the centered log ratio transform of a set of measurements, taken at the exactly same location on the core. Since the clr transformed data comprises an Euclidean space, we can simply use the Euclidean distance between these observations as an error estimate. Given that we have a set of m replicate measurements comprising D elements, we define the *mean* composition as follows:

$$\langle \mathrm{clr}(I_j) \rangle = \frac{1}{m} \sum_{i=1}^{m} \mathrm{clr}(I_j), \qquad j = 1, 2, \ldots D \tag{3.15}$$

For one set of m replicate measurements with D elements, the intrinsic noise of the system is then derived as:

$$\delta I^2 = \text{med}^m \left\{ \sum_{i=1}^{D} \left[\text{clr}(I_i) - \langle \text{clr}(I_i) \rangle\right]^2 \right\} \tag{3.16}$$

Which yields for k sets of replicates, k values of δI^2. A global noise parameter is then obtained by:

$$\delta N^2 = \text{med}_k \left\{ \delta I_k^2 \right\} \tag{3.17}$$

Even though this δN^2 reflects the intrinsic noise of the measurement system, the parameter will certainly be affected by the core characteristics as well; since δN^2 is obtained by evaluating the error of the residuals, it is also affected by the composition of the core. This because the presence of some elements can be quantified more precisely than others; if there is much overlap w.r.t. fluorescent energy of two elements X and Y, discriminating between the energy from component X and component Y is difficult. This means that in order to get a reliable estimate of δN^2, it is advisable to estimate δN^2 for every core individually.

One should keep in mind though that this noise parameter should not be confused with the calibration quality. The noise in the intensity data is a important factor in the sense that if the data is very noisy, the estimation of the linear coefficients will require more calibration samples. However the quality of the LRCE is determined by more than just the level of noise in the observed intensities, such as the reliability of the calibration sample chemistry. However there is also a sampling problem involved; the calibration samples do not reflect the same material that was measured by the scanner.

3.6 Pre-calibration filtering

Given the nature of the specimens that are being analyzed (sediment cores), we can expect a limited number of minerals in the sediment. This due to the simple reason that there are (1) only a limited number of minerals stable enough to withstand the process of weathering and transportation. On the other hand, XRF core scanners yield the composition in terms of a large number of elements; usually more than there are minerals in the specimen. This means that there is a high level of correlation between the different components. By definition all other signal is considered noise, which rises the challenge to filter the noise from the signal.

We suggest to apply a linear decomposition of the data based on the covariance structure of the variables. By applying this decomposition, a new set of variables is obtained, each reflecting coherent signal in the data. Additionally each variable in this set of variables reflects the variance in the original dataset in descending order. By subsequently throwing away the data which reflect only a small portion of the variance, the hypothesis is that the signal to noise ratio will be increased. If this data is subsequently applied in the LRCE in combination with calibration samples, the hypothesis is that improved results can be obtained.

This structure is implemented by using Singular Value Decomposition (SVD) on the *clr-transformed data*. SVD is a linear mapping of the data values onto a newly defined coordinate system represented by the eigenvectors. The SVD for a data matrix X with observations (rows) and components (columns) looks as follows:

$$X = USV^T \tag{3.18}$$

where U is an $m \times n$ matrix, S is $n \times n$ and V^T is also $n \times n$, given that n is the number of variables and m the number of observations. The columns of U are the coordinates of the data projected onto the newly defined basis. The rows of V^T contain the right singular values and form an orthonormal basis; the eigenvectors of the covariance matrix. The elements of S contain the singular values which are always larger than zero. The so called 'scores' of the data X are obtained by performing the following transformation:

$$Y^T = X^T V \tag{3.19}$$

An important implication of this linear transformation is that the data loss is minimal in a least square sense, given l the number of eigenvectors and $l \leq D$. The dimensionally reduced *scores* on this reduced space is obtained as follows:

$$Y_l^T = X_l^T V_l \tag{3.20}$$

where l reflects a subset of the eigenvectors, or the first l columns of V. For the singular values s_i, which are on the diagonal of matrix S, the following holds:

$$\lambda_i = s_i^2 \tag{3.21}$$

which is the variance of the initial data along the principal axis i.

This methodology is often being used to compress large multi-dimensional datasets for storage reasons or to evaluate the intrinsic dimensionality of the dataset. However in this case we suggest the dimension reduction as a filtering step of the core scanning data. If we would throw away some eigenvectors ($l \leq D$) comprising the l smallest eigenvalues of which we know that the dimensionality of the coherent signals in the data is smaller than the number of variables, this means that we throw away signal uncorrelated with S. Given that the variance of the signals S is sufficiently larger than the residuals of the measurement system, the principal direction that are omitted are more likely to be noise-dominated. In reality though, the problem is to determine this l and, more importantly this ’true’ dimensionality of the signal present in the data. This is where the δN^2 statistic, derived from the replicate analysis, comes into play.

A straightforward approach to determine l is to cutoff the smallest eigenvectors of which their combined sum is smaller than δN^2. To define this methodology in a more formal way, the D *ascending* eigenvalues are being transformed to a set of cumulative eigenvalues λ_i given by:

$$\lambda_i^{cum} = \sum_{j=1}^{i} \lambda_i, \qquad i = 1, 2, ...D \tag{3.22}$$

If we let $\mathbf{V}_i$ represent the i-th eigenvector in the set of eigenvectors $\mathbf{V}$, ordered on *ascending* eigenvalues, the set of principal axes $\mathbf{V}^*$ that can be considered to be above the noise level follows from:

$$\mathbf{V}_i \ni \mathbf{V}^* \quad | \quad \lambda_i^{cum} \leq \delta N^2, \qquad i = 1, 2, ...D \tag{3.23}$$

This means that the set $\mathbf{V} \ni \mathbf{V}^*$ will be omitted in the back-transformation to the original linear basis. The eigenvalues, cumulative eigenvalues and the component rejection is graphically depicted in figure 3.5 where the red line indicates the difference between data variance and noise level δN^2. Notice the strong correlation and negligible information loss while reducing the intrinsic dimensionality by 4.

The hypothesis is that this method of filtering out the PC’s with low variance will improve the signal to noise ratio and hence result in a better calibration model fit. However this is only the case for measurement noise *uncorrelated* with the composition; otherwise it will be part of the higher order PC’s. Furthermore one should keep in mind that the principal component with the smallest eigenvalue does not carry any information; variable D of the clr-transformed data will reflect the principal axis perpendicular to the hyperplane. As mentioned in section 2.3, D dimensional compositional quantities will, by definition, be situated on a $\mathbb{R}^D$ hyperplane, when clr-transformated.

Algorithm Implementation

The framework given above has been implemented in a MATLAB script. In pseudocode, the program performs the following steps where after it outputs the filtered data derived from l PCs. In order to facilitate the program to detect observation obtained at the same down core location, it is required to supply the down core depths together with the observations:

1. The data is clr-transformed,

2. Replicate samples are detected (using the depth attribute),

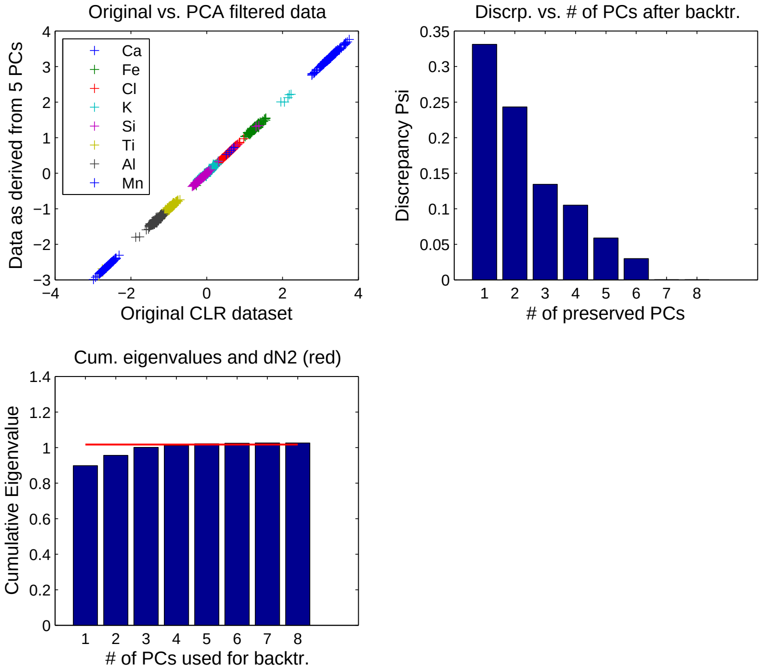

Figure 3.5: Plots showing the input scanning data of core 7920 versus the filtered data (upper left), the summed squared discrepancies as a function of the number of preserved PC's (upper right) and the cumulative eigenvalues λ^{cum} together with the cutoff value δN^2 depicted in red (lower left).

3. Calculate replicates variance parameter δI_k^2 at every replicate location k,
4. Calculate global noise parameter by δN^2,
5. The column-mean is subtracted from all observations,
6. SVD,
7. Calculate cumulative eigenvalues λ^{cum},
8. Evaluating equation 3.23 to identify the significant PC's,
9. Back-transform to original basis using the set of eigenvectors V^* and the corresponding scores and add the mean calculated in (5).

3.7 Sample Selection

Whereas the PCA-filtering step was mainly introduced to improve the final prediction, a good geochemical prediction is only obtained when there is an appropriate calibration sample set. Therefore we will further look into the current calibration methodology and how this can be improved.

In a previous study the prediction discrepancies for a dataset of an unconsolidated core were found to level out when using more than 40 calibration samples (Weltje and Tjallingii, 2008). However the hypothesis is that an improved result can be obtained with a more robust approach. This is expected to be primarily the case if (1) the dataset to which the model is fitted will be large (the probability of taking the 'good' points will decrease) and (2) when the variability of the element concentrations is large (if there is no variation at all, the choice of calibration sample is of no importance).

3.7.1 Sample selection; the concept

The basis for the approach that will be suggested here is that the best linear fit will be obtained when the points to which the linear curve for an element combination is fitted are (1) unclustered and (2) cover the entire spectrum of solutions. This originates from the fact that the effect of measurement error on the linear fit will be smallest when the points to which the fit is performed are far apart. The limiting cases are resp.:

1. If a line is fitted to two observations (subjected to noise $N(\mu,\sigma^2)$) that are infinitely far apart, the effect of the random noise will be infinitely small, and
2. If a line is fitted to two observation (subjected to noise $N(\mu,\sigma^2)$) infinitely close to each other, there will be an infinitely number of solutions for the linear model.

One way of selecting the locations to get the most effective fit of the linear model would be to select the minimum and maximum intensity of every observed component. However using this approach could result in the selection of high and low readings of a certain element that is dominated by noise (the element is almost undetectable). Constructing a linear fit by calibrating with the noise-extremes will yield unreliable LRCE model estimates! In addition, the selection of extremes for every element is also not efficient in case of highly correlated intensity measurements (for instance when the relative concentration of Ca directly correlates with the concentration of Zi). This problem of correlation can be solved by using PCA instead of the raw intensities.

3.7.2 Implementation

The suggested method is to assign a number of calibration samples to every PC proportional with the fraction of its eigenvalue λ_i w.r.t. the total variance of the data. In more mathematical terms with l the number of PC's in $\mathbf{V}^*$:

$$\lambda_j^{\%} = \frac{\lambda_j^{cum}}{\sum_{i=1}^{p} \lambda_i^{cum}}, \qquad j = 1, 2, ...l \tag{3.24}$$

Given this percentage of variance for each PC, the number of calibration samples are assigned proportional to this percentage. In order to get the best possible fit, the first two samples selected on all principal axes will reflect the two 'extremes'. However to prevent the selection of any negative or positive outliers, the samples are selected based on their distance to the 10th and 90th percentile. All other 'vacant' samples are distributed more or less uniform along the principal axis. In order to achieve this, the range of scores on a certain principal axis is being divided into p thresholds between the 20th and 80th percentile, where p reflects the number of samples to be selected. Within a certain threshold we wish to accomplish to select the sample that has the smallest sum of squared scores on all other than the current principal axis.

The identification of observations closest to the principal axis is being done in order to accomplish that sample-sets on a certain principal component will reflect only the 'process' along this principal axis without interference of the other 'processes'. However this sample selection strategy is merely based on some general idea rather than that we have determined explicitly whether this interference will harm the calibration process.

After this initiation step of distributing the samples over the PCs, the program will evaluate the spatial distribution down core of the selected samples. Because the routine tries to equally distribute the locations over the range of all principal components thereby neglecting the sample distance, sample locations might be too close each other to sample as individuals. Therefore a logical step in the process is to evaluate the down core locations and between-sample distance. In the implementation of the program, this minimum between-sample distance can be fed to the program.

The second extension of the program concerns the fact that the significance of improvement achieved using the sample selection routine is relatively small in the ’tail’ (so for a large number of calibration samples). From now on the number of samples for which the PCA-based locating of samples is considered effective S_{crit}, can be defined by the user. However if the user does not give an S_{crit}, it seems reasonable to put some more weight on the down core distribution rather than getting even more dense sample distributions on the principal axes. This down core uniform distribution is achieved by picking samples at positions right in between the already located samples using the PCA approach that have the largest between-sample distance.

A intuitive way of illustrating the sample selection is by looking at biplots. Biplot analysis is a commonly used way of evaluating and interpreting the results of a multi variate dataset. A biplot (figure 3.6) shows the normalized scores on respectively the first two or three principal axis, as obtained using PCA, which are indicates in figure 3.6 by red markers. The variable arrows on the other hand indicate the *loading* of a variable on the specific principal axis. Two parallel arrows indicate a high correlation between the two variables. A 90 degrees angle on the other hand means that the correlation between the two variables is absent whereas a 180 degrees angle indicates a negative correlation.

3.7.3 Geometrical context

An illustrative way of showing the concepts of the sample selection is by looking at a biplot. A biplot is a 2 or 3 dimensional plot showing the *loadings* of the original variables on the new principal axes. The axes of the variables are indicated by vectors which show the covariance structure of the data; a 90 degrees angle implies zero correlation whereas a small angle indicates a strong correlation. In a biplot, also the *scores* of the observation on the newly defined principal axes. The sample selection algorithm initially selects the extreme (10th and 90th percentile) scores along these principal axes, whereas the number of samples is proportional to the variance covered by the axis. An example of a biplot is shown in figure 3.6.

The biplot in figure 3.6 shows the results for core 7920 in which we observe roughly two highly correlated clusters in the clr transformed dataset; group 1 covering Ti and Fe, group 2 covering Si and K. What is also apparent in this figure is that the samples selected by the selection algorithm, indicated by green markers, include the most extreme scores in both the negative and the positive direction of the two principal axis. Note that this is only the case because the intrinsic dimensionality of the data was found to be two or larger. Otherwise the sample selection strategy would only be applied to the first principal

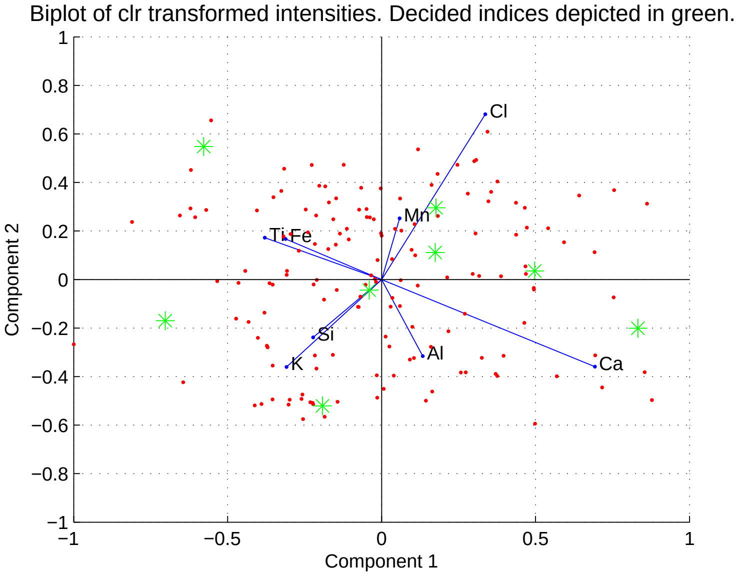

Figure 3.6: Biplot of the clr transformed data of Core 7920 with the selected calibration samples for $n = 8$*.*

axis, leading to a sample set mainly situated along the x-axis.

3.8 Module Performance

Given the two objectives discussed in the introduction, we suggest the following hypotheses:

1. A better element concentration prediction can be obtained when non-significant PCs are filtered out of the data before model fitting.

2. It is possible to achieve a better calibration result when the calibration points are carefully selected using the described method.

In order to quantify the performance of the resp. filtering and sample selection, a framework is being given here to quantify the performance. The calibration results are analyzed by calculating discrepancies for all elements D and all locations L that were not included in the calibration process. In order to be able to compare the spread in predictions with the spread in replicate intensity measurements, a comparison is performed in log ratio space in which euclidean distances are a valid distance. The distance between ’true’ and predicted concentration $\hat{W}$ can be used as a global discrepancy parameter for a specific

	A	B	C	D	E	F	G	H	I
Mineral I	0	0	0	1	0	2	0	0	1
Mineral II	0	2	0	1	0	1	0	0	1
Mineral III	1	3	1	0	1	0	2	1	0

Table 3.1: Minerals that were part of the synthetic rock, measured with the XRF core scanner.

set of D elements and k observations:

$$\Psi = \mathrm{med}_k \left\{ \sqrt{\sum_{i=1}^{D} \left[\mathrm{clr}\left(W_{ki}\right) - \mathrm{clr}\left(\hat{W}_{ki}\right)\right]^2} \right\} \tag{3.25}$$

which reflects a global measure for the prediction discrepancies; a high value of Ψ reflects a large distance in clr space and therefore a low-quality prediction. In order to get a global value for the random selection of calibration samples, the median value over all simulations yields a curve with the global Ψ as function of the number of calibration samples available. This curve was then compared to a Ψ curve obtained by selecting the calibration samples using the sample selection routine.

3.8.1 Algorithm performance for synthetic data

In order to assess whether the algorithms for filtering and sample selection are significantly better compared to the resp. unfiltered and random sample selection strategy, synthetic datasets were generated thereby allowing for applying the inverse as well as the forward model. This synthetic dataset is created using the fact that log ratio transformed data obey the rules of a ’normal’ dataset and can therefore be perturbed using an ’ordinary’ random number generator.

Creating synthetic dataset

The first step in order to obtain a synthetic dataset is to adopt a set of minerals that are going to be the components of mixture (see table 3.1). This table must be interpreted as rows representing mineral phases and the columns indicating the stoichiometry of the given phase in terms of the set of elements $\{A, B, C, ...\}$. After the random generation of mineral compositions, the corresponding element compositions are calculated which are subsequently transformed to mass fractions by using some arbitrary molar weights. In order to transform these compositions into XRF intensities, the alr-transform of the element *intensities* is determined using an arbitrary denominator element. To generate these intensities, we adopted the most straightforward model that has $\alpha = 1$ and $\beta = 0$

for all element-pairs. This yields the following model for the alr transformed quantities I^t and W^t where i denotes the numerator, D the common denominator and j the observation:

$$I^t_{iD,j} = \alpha_{iD} \cdot W^t_{iD,j} + \beta_{iD}, \quad i = 1,2,...D \tag{3.26}$$

subsequently we perform the reduced space transformation on the scores X derived from SVD of the clr-transformed intensities:

$$Y_l^T = X_l^T V_l \tag{3.27}$$

where l is smaller than the number of elements D in this synthetic mixture. Since we now know, by definition, that the data has an intrinsic dimensionality of l, we can now add noise by knowing that every PC that is above l is purely noise. Pseudo-observations subjected to measurement noise are generated as follows:

$$\tilde{X} = X + N(\mu, \sigma) \tag{3.28}$$

for a given μ and σ. This simply means that the concentrations are perturbed with noise with a given mean μ and standard deviation σ^2. For this synthetic experiment, we adopt $\mu = 0$ or more general, an unbiased prediction. Although this work flow mimics the forward model, in reality the noise component is not independent of the composition

Simulations

In order to analyze the algorithm performance, the sample selection routine was put to the test. In order to achieve this, datasets were simulated ($n = 500$) with a random mineral composition. Each dataset comprises 200 observations which were calibrated using an increasing number of pseudo-calibration samples. Additionally 60 randomly selected replicate samples were added to be used in the filtering step. Obviously the replicates within one replicate series reflect the same intensity reading, although they are *individually* and *independently* perturbed with noise. The calibration was performed both with the samples selected randomly as well as using the sample selection routine. The performance of the calibration process was also investigated for the case where the data was filtered using PCA-filtering respectively before and after it was feed to the sample selection routine.

In addition to the experimental setup described above, we also simulated data in the same manner, however now for a standard number of calibration samples while increasing the number of preserved principal axis. If the hypothesis is correct, there should be an optimal number of PC's (red. smallest median Ψ) which is smaller than raw data dimensionality. Additionally we expect this optimum to coincide with the findings of the filtering algorithm.

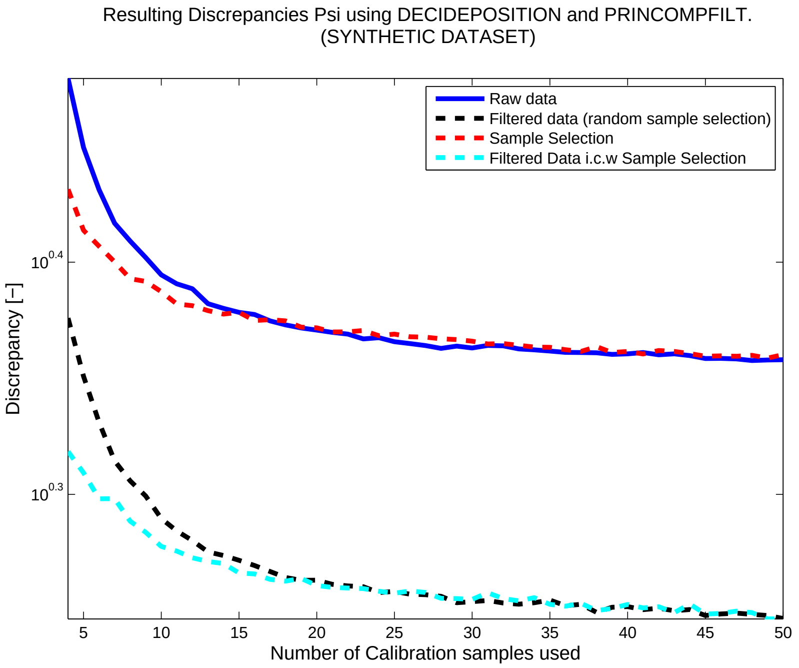

Figure 3.7: Median discrepancy values for the different number of calibration samples picked randomly. The plot shows both the sample selection and the results by using both sample selection and PCA-filtering methodology.

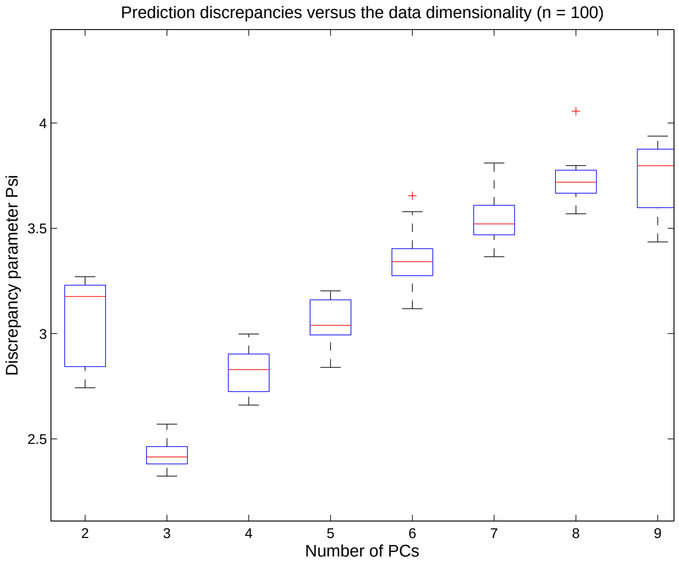

Figure 3.8: The model prediction discrepancies as determined from randomly generated synthetic data (n=100) with an intrinsic dimensionality of 3, versus the number of PC's preserved in the filtering step.

Results

In figure 3.7 the median Ψ-values for the number of calibration samples are shown. As can be seen in figure 3.7, the sample selection algorithm establishes a fast stabilisation of the discrepancies compared to the random approach. We can roughly say that there is an 20-30% faster decline of the prediction discrepancy Ψ.

The PCA-filtering is also a successful preprocessing step since the obtained discrepancies using the PCA filtering are much smaller than the discrepancies obtained with the unfiltered data. In general, 4 out of 9 principal components were preserved in the process of back transformation which results in the removal of effectively 4 principal components (one principal component only reflects the hyperplane the clr transform data is situated on). Although we generated the data with intrinsically 3 principal components, the filtering routine is at least conservative (it does not throw away too *many* PC's) and still gets rid of a significant number of principal components.

When we look at the prediction discrepancies compared to the number of preserved PC's (see figure 3.8), there can be seen that an optimal number of PC's is present. This optimum is located at the filtered data dimensionality of 3 which is equal to the intrinsic dimensionality of the dataset. The 'true' intrinsic dimensionality is known because it was used as value for l in the generation of the input data. In order to quantify whether the results obtained with the newly suggested methods are *significantly* better, a statistical tests will be applied. This will be discussed in the next section.

Statistical Significance

One of the major advantages of constructing a large number of synthetic datasets is that it allows us to asses the statistical significance of the results. Statistical tools yield a solid foundation to the overall idea that both modules (PCA-filtering and sample selection) give a better result. In order to apply a statistical test, the null hypotheses are defined as followed;

1. The sample selection procedure does not give lower values of Ψ compared to the random approach;
2. Using the PCA filtered data for the calibration does not yield lower values of Ψ compared to using the unfiltered data.

To test both null-hypothesis an Analysis Of Variance (ANOVA) was considered. Since I want to test equality between *two* groups, the ANOVA is equivalent to the t-test. This test works under the assumption of (Freedman et al., 2007);

- Independency of cases (outcome does not depend on the outcome of other realisations);
- Normality of the distribution of the residuals
- Equality of Variances within groups.

Unfortunately normally distributed noise on the clr transformed compositions will cause non-normal distribution of the prediction discrepancies in terms of Ψ.

In order to cope with this, a nonparametric test can be used. Such a test is based on testing the hypothesis based on ranks rather than the values itself. This results in assumption (2) being replaced by an equal distribution requirement. Given that we compare two equal quantities with each other (Ψ) which only differ w.r.t. the perturbation of noise, we can say that the equal distribution requirement is met.

A non-parametric test used for two groups is the Mann-Whitney (Hollander and Wolfe, 1973; Mann and Whitney, 1947) test which quantifies the probability that two groups are from the same population. Generally the null hypothesis is defined as the two populations being from the same population given a certain level of significance. The two-sample Mann-Whitney test can be used when the following requirements are met;

- Samples drawn randomly from the population,
- Within- and between-group independency,
- Ordinal or continuous measurement scale.

These requirements are all met, given the experimental setup. The setup was such that 500 random datasets were generated randomly and calibrated using a different number of calibration samples. Given this setup, the sample selection algorithm does not yield independent sets of samples for 1 dataset, which makes the results *for one dataset* not independent in the direction of the number of samples. However we wish to apply the test for *every number of calibration samples* and in this 'direction' the observations are independent since they are obtained from different randomly generated datasets. Given that the random number generator has a periodicity larger than 2^{10000}, it is reasonable to assume independency.

The continuous measurement scale requirement is also met; data is considered to be continuous if the values may take on any value within a finite or infinite interval, which is the case for Ψ. Furthermore you can count, order and measure Ψ, making this a continuous quantity.

Statistical Experiment

The Mann-Whitney test was applied by comparing the discrepancies derived from a random specimen selection with the PCA-based sample selection and the discrepancies obtained by applying the pre-calibration filtering step.

In figure 3.9 the discrepancies of both the random approach, the sample selection and the filtering i.c.w. the sample selection are shown for two cases: 5 calibration samples and 30 calibration samples. Here one can see that where the performance of the filtered results increases (Ψ decreases) with the number of calibration samples, the results obtained with the sample selection converges towards the random sample selection discrepancies with increasing number of samples. In the figure on the right, the distribution of discrepancies are shown where we observe that the sample selection results are very similar to the random sample selection in terms of Ψ.

Since the test is applied for every number of samples, it will result in two curves of the probability of equal mean as function of the number of calibration samples. In order to quantify the significance of the improvement by using PCA filtering and the sample selection routine, a confidence limit is adopted at 5% which reflects the probability of a type 1 error (probability of rejecting the null-hypothesis while it should be accepted).

Test results

The Mann-Whitney test of the Ψ values yields that the sample selection routine results in a significant improvement of the model fit up to 13 calibration specimens (when the number of calibration specimens is 14, the probability exceeds the level of significance 5%). The test results w.r.t. the combined use of sample selection and PCA-filtering show for all number of calibration specimens very small values of p ($p < 10^{-40}$). Based on

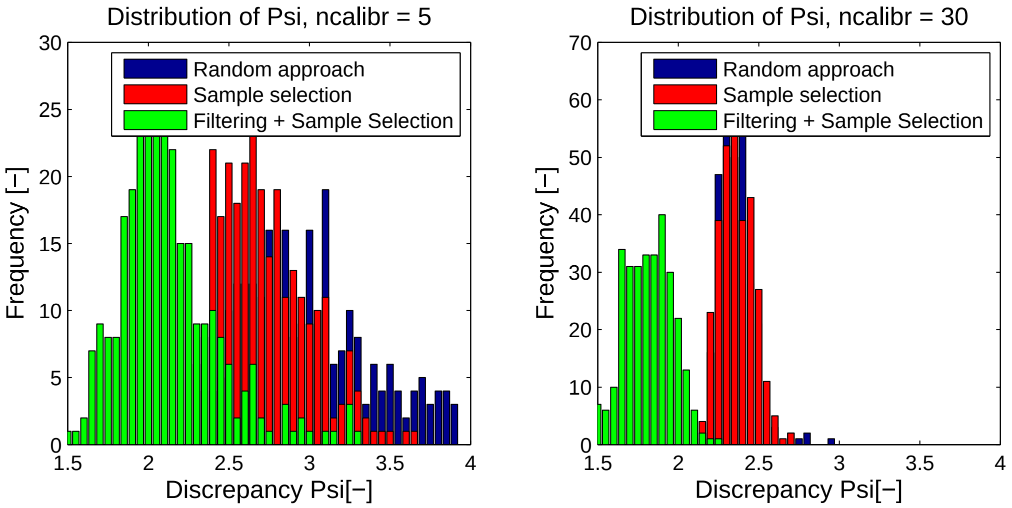

Figure 3.9: Distributions of the discrepancies Ψ *for the synthetic datasets for* $n = 5$ *and* $n = 30$.

these results we reject h_0 referring to the filtering step and h_0 of the calibration sample selection up to a maximum of 13 selected samples. Although for this (synthetic) dataset, the use of the two modules is beneficial for the prediction results for 13 selected samples, this value might well be affected by the nature of the dataset. In order to investigate the performance of the modules for the processing of ’real’ data, they were applied to data of unconsolidated sediment.

3.8.2 Application to core 7919 and 7920

The work flow is rather straightforward; first the routine of PCA-filtering was applied to define the number of significant PCs and to filter the data of core 7920 accordingly. After that the random sample selection strategy was tested using data of core 7919; in total 100 different permutations of the observation indices were generated. For every simulation, the number of calibration samples was increased, ranging from 3 to 50 samples, thereby creating $100 \times (50 - 3)$ different predictions. These results were compared to the results obtained by calibrating with samples selected by the selection routine, with the number of samples ranging again between 3 and 50.

In figure 3.10 the median discrepancies as a function of the number of calibration samples are shown for both the filtered and unfiltered data of soft-sediment core 7920. This filtering step was performed with the preservation of 5 dimensions of the 8 variables (elements). The test results w.r.t. the sample selection are shown in figure 3.11 which were determined using the data of core 7919.

The results in figure 3.10 indicate a significant improvement w.r.t. the prediction

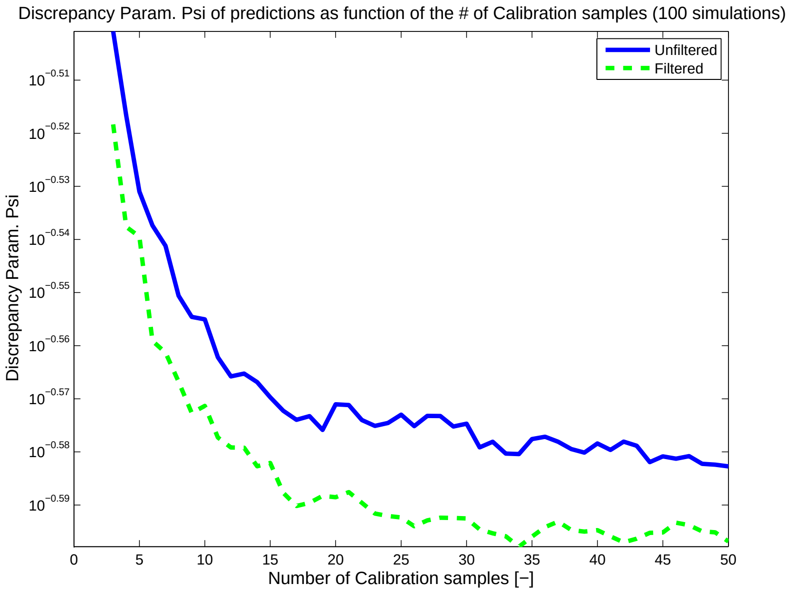

Figure 3.10: Discrepancy parameter as function of the number of calibration samples using the filtered and unfiltered data for the data of core 7920.

discrepancies whilst using the filtering step. Also the sample selection strategy shown in figure 3.11 yields good results; with the selection of 10 calibration samples, the expected Ψ is comparable to the Ψ obtained while calibrating with 40 randomly selected samples. From the plot with the discrepancy values as a function of the number of samples it may seem that the discrepancies become unstable for an increasing number of samples. However this is the result of the experimental setup; the prediction discrepancies are only evaluated for the observations which have *not* been used in the calibration process. This means that if the number of samples is increasing, the number of samples that are used to evaluate the performance decreases. Since the global discrepancy parameter is determined as the median of this set, Ψ becomes more unstable if the set of *calibration* samples increases.

3.9 Discussion and Conclusions

The research question that corresponds to the work that was presented in this chapter was:

> **How can the current work flow be optimized to reduce costs and to increase accuracy and or precision?**

which was subdivided into:

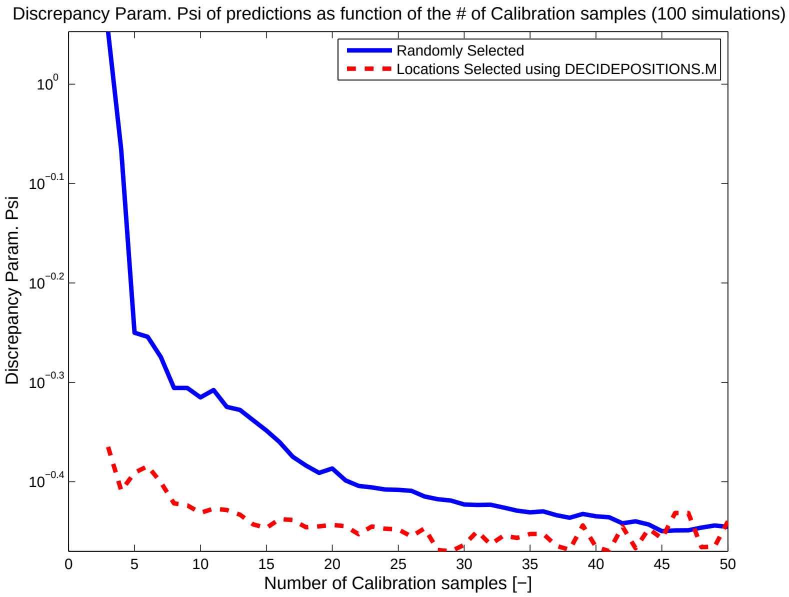

Figure 3.11: Discrepancy parameter as function of the number of calibration samples using the random approach (blue) and the sample selection algorithm. These results are derived from core 7919.

> *Can clever use of multi variate methods result in an improved signal to noise ratio?*
>
> *Can a more robust sample selection strategy result in a faster stabilisation of the prediction discrepancies?*

Based on the previous work of Weltje and Tjallingii (2008) and the CoDa framework, two modules were designed in order to achieve better calibration results. This included a module that could result in a significant signal to noise ratio of the intensity data. Furthermore a sample selection strategy was defined that should result in a faster stabilisation of prediction discrepancies.

The pre-calibration filtering step is found to provide improved results for both a real dataset as well as synthetic data. For the latter, the Mann-Whitney two-sampled test yielded significantly lower prediction discrepancies, given a 5% level of significance. The probability under the null-hypothesis though was much smaller than this 5% ($< 10^{-40}$). Additionaly a calibration sample selection strategy has been developed by using Principal Component Analysis. It was applied to the calibration of synthetic data which yielded that the algorithm is effective until 13 calibration samples. When the number of samples was larger, the sample selection was identified as being not significantly better, given an level of significance of 5 %. When applied to a real dataset of core 7919, calibration using the selected samples resulted in a very fast stabilisation of the prediction discrepancy, compared to the random sample selection. Therefore we conclude that the two modules are valuable and facilitate (1) calibrated output of higher quality and (2) the same prediction quality using a smaller sample set.

Chapter 4

Linear Mixing in a Sedimentary Basin

Whereas the first chapter discussed some new developments w.r.t. the processing of the geochemical data, we will now turn to the interpretation of geochemical data. In the section discussing the primary objective of this study we stated that the primary objective is to perform grain size predictions based on geochemistry. This implies that we somehow have to understand the relation between these two quantities. In the first chapter the we discussed the complexity of the process of basin infill and the factors controlling the composition of the sediment in the basin. We furthermore defined the research question:

> **Can we build a forward model of basin infill in a formal geochemical and textural context that successfully reconstructs real observations?**

The primary reason of building such a model is that the performance of the forward model provides insight in the *underlying physical* forward model configuration and parameters. If we have an answer to this question, inverse modeling can be applied in order to reconstruct the history of the basin. However because of the complexity of sedimentary basin infill, inverse modeling in terms of all these processes is by definition an under-determined problem that leaves us with many unknowns.

In this chapter we will investigate one of these components which is the process of mixing of chemically and texturally different sediment sources. A model is built for this process which is subsequently applied to a soft-sediment core in order to test the performance. The main reason of applying it to a soft-sediment core is that it is, in contrast to ’old’ sedimentary rock, not altered by diagenesis. Another important reason of applying this model to ’young’ sediment is that the history of this sediment is well-known and well-defined. Because of this, any model outcome can be validated with the rather accurate and well-defined ’reality’.

4.1 Core 9508; Offshore Africa, Senegal

Core 9508 was obtained from the continental slope off Senegal at about 15/29.90N/17/56.88W from 2384 m water depth using gravity coring (see figure 4.1). The continental slope off Northern Senegal is considered an ideal location to study the varying influx of eolian dust from the arid region in the North and fluvial sediment from the African continent, respectively. Historical records suggest that Sahel droughts result from changes in the large-scale distribution of sea surface temperature. The primary objective of the work of Mulitza et al. (2008) was to reconstruct the last 57.000 years in terms of climatology and to investigate whether the sediment in the core shows evidence for these large-scale variations in sea surface temperature distribution.

In the work of Mulitza et al. (2008), a large number of core samples, uniformly distributed along the core, were analyzed with respect to their geochemical composition. Furthermore the grain size distribution of these samples was determined. This was done because it turned out from previous previous work that whereas the fluvial portion typically has grain sizes below 10 μm (Gac and Kane, 1986), the eolian dust consists mostly (between 44% and 83%) of grains larger than 10 μm (Stuut et al., 2005). Based on these distributions the relative portion of fluvial sediment relative to the eolian sediment was derived based on the mass-percentage under 10 μm with respect to the bulk.

Additionally discrimination between the fluvial and eolian portion was done based on two element ratios (Fe/K and Al/Si resp.) of which the resulting Al/Si ratio is shown in figure 4.2. This figure also shows the mass percentage of the material smaller than 10 μm that is considered to be an indicator for the fluvial influx. From these results was derived that both the grain size related parameter, as well as the Al/Si ratio were found to coincide fairly well with so called Heinrich Stadials (H., 1988). These Heinrich Stadials coincide with the destruction of northern hemisphere ice shelves. The icebergs melting caused significant amounts of fresh water to be added to the North Atlantic causing a disturbance in the density-driven thermo-haline circulation of oceanic water. The result of these processes for the Sahel region is a series of periods with relatively low precipitation. As a result, the hypothesis was that sediment characteristics mark the occurrence of these Heinrich Stadials; influx from fluvial origin reflecting the *interstadials* where eolian sediment marking the Stadials. The approximate time intervals of the Heinrich Stadials as interpreted in the work of Mulitza et al. (2008), are indicated in grey and were determined from the age-model.

Based on these findings Mulitza et al. (2008) concluded that the relative influx of fluvial and eolian sediment has a climatological significance. Furthermore it was found that this signal is apparent in both the grain size data as well as the geochemical record.

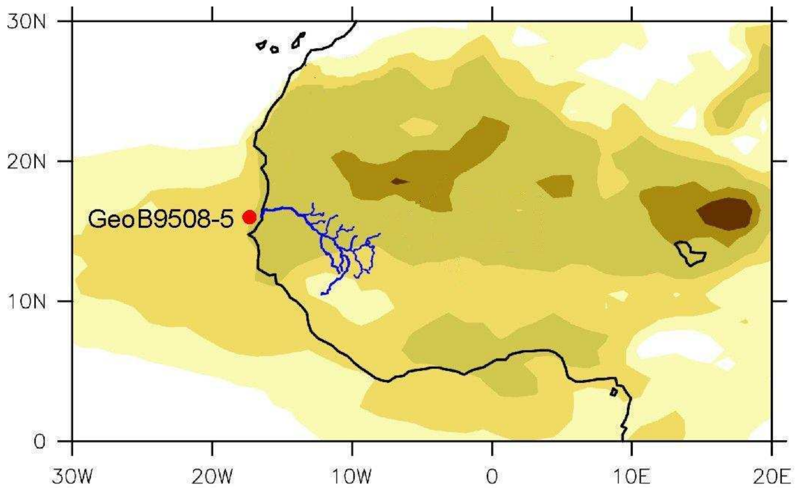

Figure 4.1: Location of core 9508 (modified from Mulitza et al. (2008)).

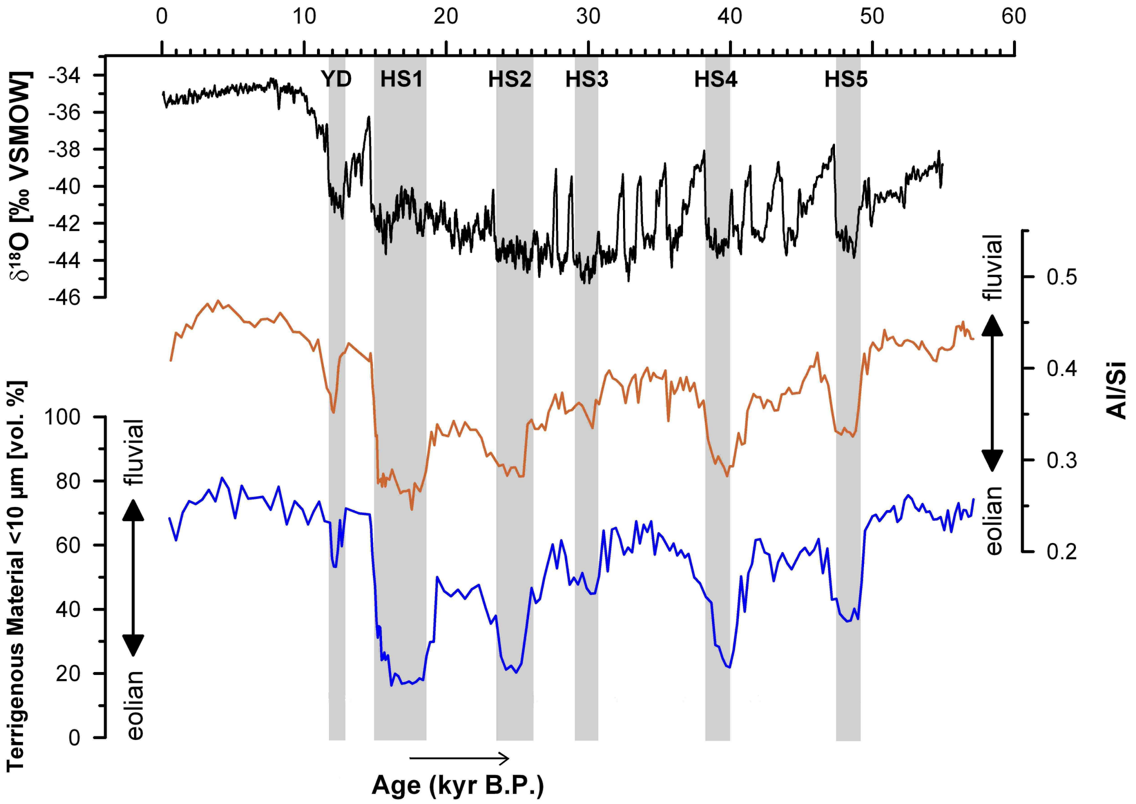

Figure 4.2: The Al/Si ratio (red), a record of the benthic σ^{18}O (black) and the volume percentage of all material smaller than μm. The Heinrich Stadials are indicates in grey (modified from Mulitza et al. (2008)).

4.2 LINMIX model

In chapter 1 we briefly discussed the framework of sedimentary basin infill and the primary concerns w.r.t. the direct mapping of the concentration of certain elements to a grain size signal. We stated that the one-to-one mapping of element concentrations or ratio's to grain size reflects a simplification of the 'true' history of sediment deposition. Also the application of grain size information in the way as was done in the work of Mulitza et al. (2008) (a simple 'hard' partitioning of the grain size spectrum), is considered to fail to provide insight into paleo-climate in many cases (Weltje and Prins, 2003).

The same can be said about provenance (i.e. parent lithology) analysis based on geochemical data; analogous to the grain size prediction, the direct application of geochemical data towards any sedimentological parameter is only justified when all other parameters have a negligible effect on the composition. More general, we can say that the observed chemical signal is the result of a convolution of a source signal and selective transport signal. Additionally, this convoluted signal can be perturbed after deposition by diagenetic processes.

In all other cases, we are dealing with an under-determined problem that cannot be readily solved for its components. However the fact that there are so many geochemical controls does not necessarily mean that the explanatory power of geochemistry is low; it only requires more data in order to *deconvolve* the signal. Still the introduction of a set of *assumptions* and *simplifications* of the reality are necessary to perform this deconvolution.

4.2.1 The concept

We will now introduce the concepts of the LINMIX model, which is based on a set of assumptions and simplifications. Even though the initial assumptions (i.e. that an element ratio is a direct indicator of a sediment property) are traded for a new set of assumptions, the new set reflects a more physical set, regarding the process of basin infill. The model is built on the following foundation:

1. The different sediment sources have a unique composition-versus-grain-size function.
2. The sedimentary record is the result of varying influx of different sediment sources, having their own unique grain size distributions.

which is described in great detail in the work of Weltje and Prins (2003), both the concepts as well as data that behaves accordingly.

Generally we can say that eolian sediment for instance, compared to fluvial sediment, is well sorted coarse material. Furthermore arid regions with primarily eolian sediment transport produce geochemically different sediment compared to the fluvial sediments, that are produced in a humid environment. Given that the two signals (the grain size and geochemical signal) are correlated, we want to find out whether there is an underlying relation that characterizes these two signals.

This relation will be applied in the model as Transfer Functions (TF's); functions describing the chemical composition as a function of the grain size. Although the processes underlying the transport and the associated compositional change of sediment are complex, it was found by Tolosana-Delgado and von Eynatten (2008) that some generalisations w.r.t. the (petrographic) composition is valid. They found that the petrographic composition could be approximated well by a linear model of the log ratio's for a wide range of grain sizes. The fact that a (log ratio) linear approximation yields acceptable results, makes it possible to estimate these linear parameters by using only a small number of observations (degrees of freedom = 2). As a result it is possible to estimate these TF's based on the chemical data of only 2 grain size fractions. Since the mutual relation between mineralogy is trivial, the following assumption is adopted:

> *The compositional change as a function of grain size can be sufficiently parameterised by a logratio linear model.*

Whereas isolating a grain size fraction is not possible for consolidated sediment, grain size fractions of unconsolidated sediment can be physically isolated by using dry sieving. This allows for chemical analysis of isolated size fractions and therefore the estimation of data points on the TF's.

The underlying TF model is important to characterize the compositional change of a certain endmember. However initially, the endmember composition is determined readily from the bulk chemistry using the inverse model of *emph simple linear mixing*.

4.2.2 Simple linear mixing

Since the suggested model is based on the concept of linear mixing, we will first discuss this concept in its most general form. Linear unmixing is based on the concept of expressing a composition **x** comprising D components, into N mixing coefficients m, and N 'pure' endmember compositions **y** (Weltje, 1997; Renner, 1991). In vector notation:

$$\mathbf{x} = \sum_{i=1}^{N} m_i \mathbf{y_i}, \qquad \mathbf{x} = \begin{bmatrix} x_{1,1} \\ x_{2,1} \\ \cdots \\ x_{D,1} \end{bmatrix} \tag{4.1}$$

If a set of observed compositions $\{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_k\}$ can be sufficiently expressed into linear mixing of two endmembers for instance, this means that the observed composition (of any entry of $\mathbf{x}_k$) is a linear function of $\{m_1, m_2\}$. Alternatively we are, given that we know the mixing proportions, able to extract the fraction of a component in the ’pure’ endmembers y_1 and y_2, by solving the linear system:

$$x = m_1 y_1 + m_2 y_2 \tag{4.2}$$

where for the endmember weights hold that:

$$m_1 + m_2 = 1 \tag{4.3}$$

It now appears that this system is easily extended towards a system of more endmembers. However by introducing more endmembers and additionally more elements D, problems may arise when applying this to real data. Given that the system is overdetermined (if $n \leq D$), it is unlikely that this system has a solution as a result of measurement error. However besides the fact there is most likely no solution, the solution is also not constrained to the nonnegative part of the solution space which may lead to a physically non-feasible solution.

One way of avoiding a violation of the unit-sum and non-negativity constraints would be to use log ratios. However whereas linear mixing problems behave accordingly in compositional mixing space, linear mixing is non-linear in log ratio space. In figure 4.3, the relations between the endmember weights and the observed compositions (in resp. weights, ratio’s and log ratio’s) are shown for a synthetic two endmember case. When the number of endmembers increase as well as the number of components, parameterising the behavior of these components as a function of the mixing proportions becomes even more difficult. Even though the mixing in log ratio/log ratio configuration seems to behave approximately linear, at the 100% membership locations it is definately not. Unfortunately this part of the mixing space is the most important part for the estimation of the endmember composition. Based on the results shown in figure 4.3 we conclude that the most suitable coordinate system in which the linear regression can be applied is the compositional space.

Whereas the linear mixing represented in figure 4.3 comprises a system with only two endmembers, this system is easily extended in a general form. The linear mixing for a system of three endmembers can be thought of as a ternary diagram framework (a representation of the $\mathcal{S}^3$ simplex in a constrained $\mathbb{R}^2$), above which the fraction of a certain component changes linearly in the z-direction by moving along the x,y-plane. Although we will sketch the problem comprising a three endmember system, this system can easily be extended to a multi-endmember system; see figure 4.3 for the 4-endmember case, where the color indicates the fraction of a component in the bulk.

One preliminary step before applying regression is to apply a coordinate transformation such that the membership is transformed into coordinates on the simplex $\mathcal{S}^n$. The reason for applying this transformation is that by doing so, the number of coefficients that must

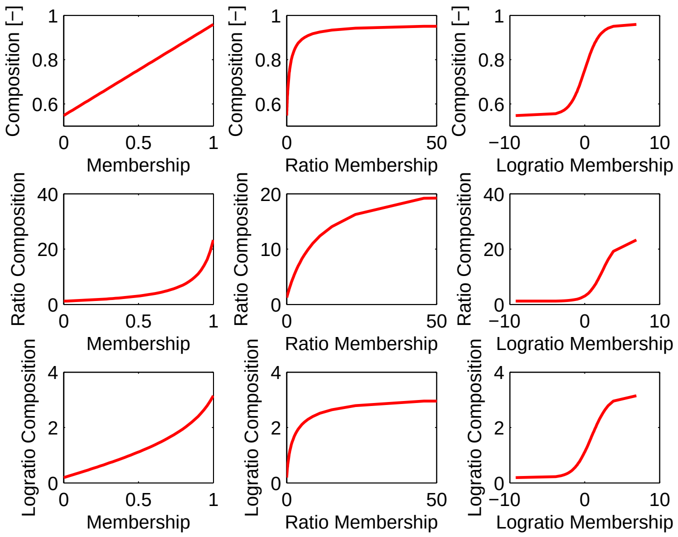

Figure 4.3: Linear mixing in compositional space, ratio space and log ratio space.

be estimated is reduced by 1 compared to the original mixing space. Or, more general, we transform the original mixing space, comprising N components into a Cartesian system in $\mathbb{R}^{N-1}$.

For a three endmember case the transformation matrix $\mathbf{T}$ is given by:

$$\mathbf{T_3} = \begin{bmatrix} -\cos\frac{\pi}{3} & \frac{1}{2} & \frac{1}{2} \\ -\sin\frac{\pi}{3} & -\frac{\cot\frac{\pi}{6}}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} \tag{4.4}$$

which is the transformation matrix for a ternary diagram representation. The derivation of this transformation matrix for any number of endmembers is described in many geometry textbooks such as the one from Schartz (1974), but can be summarized by subsequently transforming the data to barycentric coordinates and a rotation of the Cartesian coordinate system to the desired viewpoint.

If we let $\mathbf{m}$ reflect the relative fractions of the endmembers in the bulk in a $1 \times n$ vector, the following change of coordinates is applied:

$$\mathbf{c} = \mathbf{T} \cdot \mathbf{m} \tag{4.5}$$

As can be seen in figure 4.4, by staying in the simplex we are able to apply simple linear methods to estimate the compositional change with membership. The endmember compositions are determined by subsequently applying Least Squares Regression and evaluating the composition at the pure endmembers locations in the simplex s^n (the corner

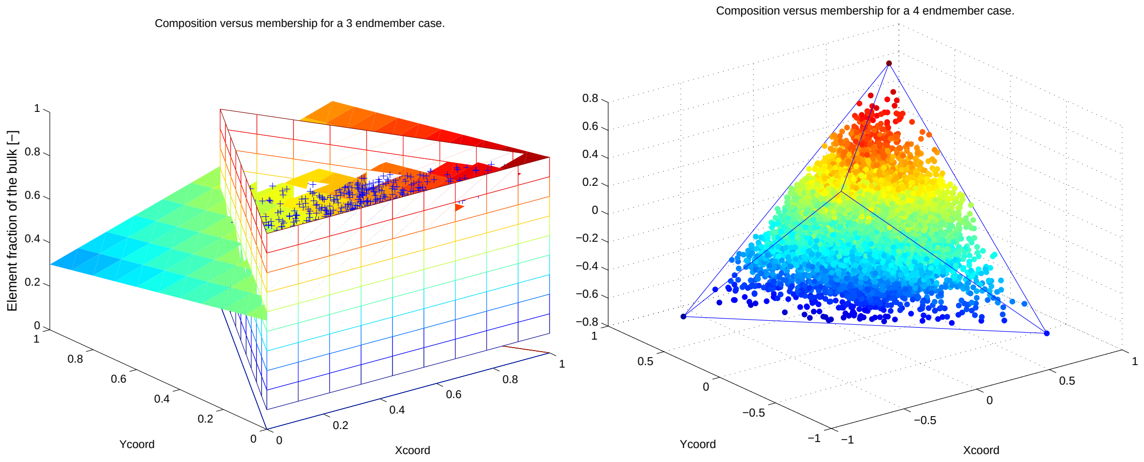

Figure 4.4: Linear mixing and the compositional response after coordinate transformation in resp. $\mathcal{S}^3$ and $\mathcal{S}^4$.

points of the ternary triangle).

There is however no free lunch; by staying in the simplex our linear regression model must be constrained such that the predicted composition will be physically meaningful (non-negative and not larger than 1, in case of fractions). In order to achieve this we will apply a Constrained Least Squares (CLS) methodology which results in the following model, given the number of endmembers N for a component j:

$$\begin{aligned} a_1 + a_2 c_{x,j} &= b_j & N &= 2 \\ a_1 + a_2 c_{y,j} + a_3 c_{x,j} &= b_j & N &= 3 \\ a_1 + a_2 c_{z,j} + a_3 c_{y,j} + a_4 c_{x,j} &= b_j & N &= 4 \end{aligned} \tag{4.6}$$

Given that there are M observations of this element j and N endmembers, we can rewrite this system into the following components:

$$\mathbf{c}_j = \begin{bmatrix} 1 & c_{1,1} & c_{1,2} & c_{1,3} & \cdots & c_{1,N-1} \\ 1 & c_{2,1} & c_{2,2} & c_{2,3} & \cdots & c_{2,N-1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots \\ 1 & c_{M,1} & c_{M,2} & c_{M,3} & \cdots & c_{M,N-1} \end{bmatrix} \quad \mathbf{a}_j = \begin{bmatrix} a_{1,j} \\ a_{2,j} \\ \vdots \\ a_{i,j} \end{bmatrix} \quad \mathbf{b}_j = \begin{bmatrix} b_{j,1} \\ b_{j,2} \\ \vdots \\ b_{j,M} \end{bmatrix} \tag{4.7}$$

with which equation 4.6 can be rewritten as:

$$\mathbf{c}_j \cdot \mathbf{a}_j = \mathbf{b}_j \tag{4.8}$$

This system can be solved in a least squares sense given that $M > N$. However in order to prevent the solution to become non-feasible (negative solution or fraction larger than unity) this system of equation must be solved using a set of constraints. In a case of 2 endmembers the constraints w.r.t. the solution space are:

$$\mathbf{A_1} \cdot \mathbf{a}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{A_0} \cdot \mathbf{a}_i = \begin{bmatrix} -1 & 0 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{4.9}$$

which can be written more general into the following form for a system of N endmembers and a given element j:

$$\mathbf{A_1} \cdot \mathbf{a}_j = \begin{bmatrix} \mathbf{l_1} \\ \vdots \\ \mathbf{l_N} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \leq \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{A_0} \cdot \mathbf{a}_j = \begin{bmatrix} -\mathbf{l_1} \\ \vdots \\ -\mathbf{l_N} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \leq \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.10}$$

where $\mathbf{l_N}$ is the location of endmember N in terms of the barycentric coordinates $\{c_1, c_2, ...c_{N-1}\}$. Where $\mathbf{A_1}$ and $\mathbf{A_0}$ are the compositional constraint matrices to estimate the fraction of one element, we must incorporate the compositional constraints such that the estimated composition *as a whole* is feasible. If we define the following building blocks:

$$\mathbf{A_c} = \begin{bmatrix} \mathbf{A_1} \\ \mathbf{A_0} \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} \mathbf{1^N} \\ \mathbf{0^N} \end{bmatrix} \tag{4.11}$$

we can define the full system of constraints for an N endmember system that is composed of D elements as follows:

$$\mathbf{A} \cdot \mathbf{a} = \begin{bmatrix} \mathbf{A_c} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A_c} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A_c} \\ \mathbf{A_c} & \mathbf{A_c} & \cdots & \mathbf{A_c} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_D \end{bmatrix} \leq \begin{bmatrix} \mathbf{q} \\ \mathbf{q} \\ \vdots \\ \mathbf{q} \\ \mathbf{q} \end{bmatrix} \tag{4.12}$$

In order to obtain the geochemistry of the endmembers we must solve the following linear system of equations using least squares:

$$\mathbf{H} \cdot \mathbf{a} = \begin{bmatrix} \check{\mathbf{c}}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \check{\mathbf{c}}_2 & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \check{\mathbf{c}}_D \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_D \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_D \end{bmatrix} \tag{4.13}$$

where $\check{\mathbf{c}}_j$ is $M \times N$ matrix with the observed membership transformed to the simplicial coordinate system, analogous to equation 4.7. Even though we have not isolated the endmembers in any way from the bulk we can now, by solving this constrained linear system, estimate the chemical composition $\hat{\mathbf{b}}_{\mathbf{p}}$ of every endmember i using the estimated linear coefficients $\hat{\mathbf{a}}$:

$$\hat{\mathbf{b}}_{\mathbf{p}} = \begin{bmatrix} \mathbf{l_i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{l_i} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{l_i} \end{bmatrix} \cdot \hat{\mathbf{a}} \tag{4.14}$$

For clarity, we shall refer to this *endmember* composition, as opposed to the *bulk* chemistry $\mathbf{b}$, by the *partial* bulk chemistry $\mathbf{b_p}$.

4.2.3 Generalized linear mixing model

Before moving on, let us revisit our objectives. In order to have the model be in line with the objectives, the model must be more complex then only being a matter of mixing endmembers with some arbitrary (best fitting) composition. In the objective was stated that the composition itself is again a function of grain size which therefore results in the partial bulk chemistry being determined by both a Transfer Function and a Grain Size Distribution (GSD). For the moment we assume that every endmember has its own unique TF and GSD.

A discrete version of the Transfer Function, denoted by $\mathbf{C}$, of a certain endmember i for D elements, discretized into G grain size classes can be written as:

$$\mathbf{C_i} = \begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,D} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ C_{G,1} & C_{G,2} & \cdots & C_{G,D} \end{bmatrix} \tag{4.15}$$

Because this Transfer Function reflects the composition in the grain size classes, the (horizontal) sum of all fractions should sum to one, representing the ’whole’:

$$\sum_{j=1}^{D} C_{k,j} = 1, \quad k = 1,2,...G \tag{4.16}$$

If we define the weight fractions $\mathbf{f}$ in all G grain size classes, which is in principle the GSD, as follows:

$$\mathbf{f} = \begin{bmatrix} f_{1,1} \\ f_{2,1} \\ \vdots \\ f_{G,1} \end{bmatrix}, \qquad \sum_{k=1}^{G} f_{k,1} = 1 \tag{4.17}$$

the bulk chemistry *of one endmember i* can now be written as:

$$\check{\mathbf{b}}_{\mathbf{p},i} = \mathbf{C}_{\mathbf{i}}^{\mathbf{T}} \cdot \mathbf{f}_{\mathbf{i}} \tag{4.18}$$

As mentioned before, this is only an approximation; in reality the TF's and the GSD's are both continuous functions. In that case the concentration of an element j in an endmember i is obtained by integration along the grain size spectrum denoted by x:

$$\check{b}_{p,ij} = \int_{-\infty}^{+\infty} c_{ij}(x) f_i(x) dx \tag{4.19}$$

where c_{ij} is the continuous function describing the fraction of element i in endmember j as function of the real variable x; grain size.

The partial bulk chemistry obtained here is equivalent to the partial bulk chemistry defined in equation 4.14. However it is now being expressed more explicitly as the result of the product between a compositional function and a function describing the relative importance (mass fraction). The ability to estimate this partial bulk chemistry by means of both the bulk chemistry and the more complex approach using the TF's and GSD implies that we have a way to validate the forward model with the inverse model. In other words, we are now able to unwrap the 1-dimensional geochemistry, into a grain size signal and a source component, determined by the TF.

4.2.4 LINMIX model

Suppose we would be able to measure the bulk chemistry **b** of a sample being a mixture of several endmembers, all having a unique Transfer Function **C**, the LINMIX model follows by plugging equation 4.18 into equation 4.1:

$$\begin{bmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{D,1} \end{bmatrix} = \mathbf{b} = \sum_{i=1}^{N} m_i \check{\mathbf{b}}_{\mathbf{p},i} = \sum_{i=1}^{N} \mathbf{C}_{\mathbf{i}}^{\mathbf{T}} \cdot [\mathbf{f}_{\mathbf{i}} \cdot m_i] \tag{4.20}$$

Theoretically the forward model (the right hand side) should correspond to the measured bulk (the left hand side) and, more importantly, this system of equation can be solved w.r.t. m_i, given that $G+N \leq D$. However it is unlikely that this system has a solution in case it is applied to real data. This because the quality of the measured chemical composition of the grain size fractions (i.e. the discrete TF denoted by $\mathbf{C_i}$) is low due to several reasons.

First of all we are not able to reconstruct the 'real' continuous GSD because we can only isolate a limited number of fractions. The system suggested in equation 4.20 also assumes a uniform distribution of grain sizes within an isolated class. This has turned out as being far from reality; there is between-class overlap and the within-class distribution is more or less bell-shaped. There are also limits w.r.t. the value of the compositional analysis of these size fractions because above a certain grain size, the grains become complex aggregates of individual minerals, making the data less reliable.

All these error sources contribute to the predicted bulk chemistry $\check{\mathbf{b}}$ and compositional matrix $\mathbf{C}$ from which the combined effect on the different components is denoted by ε_D. Therefore the system will most likely not have a solution and can only be approximated using regression techniques:

$$\begin{bmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{D,1} \end{bmatrix} = \check{\mathbf{b}} \approx \left[\sum_{i=1}^{N} \mathbf{C_i^T} \cdot [\mathbf{f_i} \cdot m_i] \right] + \varepsilon_\mathbf{D} \tag{4.21}$$

Similarly to the LINMIX model for the bulk chemistry, we can define an analogous model for the isolated grain size classes. If we have the grain size distributions of the endmembers and the bulk chemistry of an isolated grain size class, we can apply this linear system and thereby obtain a pseudo-observation of the TF. However in order to do so, the linear model must be slightly modified. This can be illustrated by looking more into the GSD vector $\mathbf{f}$. Let us define for an n endmember system the GSD *matrix* $\mathbf{F}$:

$$\mathbf{F} = \begin{bmatrix} \mathbf{f_1} & \mathbf{f_2} & \cdots & \mathbf{f_n} \end{bmatrix} \tag{4.22}$$

If we wish to apply the system suggested in equation 4.21 we have to rewrite the system such that the nett weights $\mathbf{f_i} \cdot m_i$ applied to the Transfer Functions $\mathbf{C_i}$ sum up to one; otherwise the mixing proportions of the different endmembers *within a grain size class* do not sum up to one. To achieve this unit-sum constraint, we introduce a Mass Fraction Matrix $\mathbf{M}$ which is uniquely defined for every observation and every grain size class:

$$\mathbf{M} = \mathcal{C}\left[(\mathbf{1}^G \cdot \mathbf{m}^T) \cdot \mathbf{F}\right] = \begin{bmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,N} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ M_{G,1} & M_{G,2} & \cdots & M_{G,N} \end{bmatrix} \tag{4.23}$$

where $\mathbf{m}$ is a vector of membership coefficients and $\mathbf{1^G}$ is a $G \times 1$ vector full of ones. If we now redefine our model for the case we have measured the bulk chemistry of an isolated grain size class g, denoted by $\check{\mathbf{b}}_{\mathbf{g}}$, we obtain the following results for an element j:

$$\begin{bmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{D,1} \end{bmatrix} = \check{\mathbf{b}}_{\mathbf{g}} \qquad \check{\mathbf{b}}_{\mathbf{g}(i)} = \left[\sum_{i=1}^{n} C_{i(g,j)} \cdot M_{(g,i)} \right] + \varepsilon_{\mathbf{D}} \tag{4.24}$$

where $\mathbf{M}_{(g,i)}$ and $\mathbf{c_i}_{(g,j)}$ represent resp. the mass contribution of endmember i in a given grain size class g and the fraction of element j in a grain size class g of 'pure' endmember i.

Now we have characterized the GSD's, membership and the composition of isolated grain size classes, we can solve for the discrete TF, denoted by $\mathbf{C_i}$, by applying the CLS regression model given in equation 4.6. However we prefer the TF in a functional form, rather than in the discrete matrix $\mathbf{C_i}$.

4.2.5 Fitting a transfer function

The tedious linear operations applied in the previous subsection yield compositions in different grain size classes for the different endmembers or endmember pairs. However since we would like to determine a continuous endmember composition, we have to transform these observations into a functional relation between composition and grain size. The next objective is therefore to obtain feasible Transfer Functions from this data. It was found by Tolosana-Delgado and von Eynatten (2008) that the compositional change w.r.t. the grain size can be reasonably well approximated by a linear function. Although his findings were based on a dataset of mineral composition rather than the geochemical composition, we assume the same behavior of the geochemistry based on the fact that they are very closely related (Posch and Kurz, 2007). The ordinary least squares adjustment may seem directly applicable in this situation, however the problems arise from the fact that it is a linear fit relating a compositional quantity as a function of a real variable (in this case the grain size). And since the compositional space is not an euclidean vector space, the minimisation of the distance requires a different framework.

As discussed also in section 2.3, the log ratio transformed data on the other hand will comprise an Euclidean vector space and can therefore be directly plugged into the ordinary least squares adjustment framework. Let us first define the log ratio transformed compositions $\check{x}_k$ as a function of a real variable k and the compositional slope and intersect

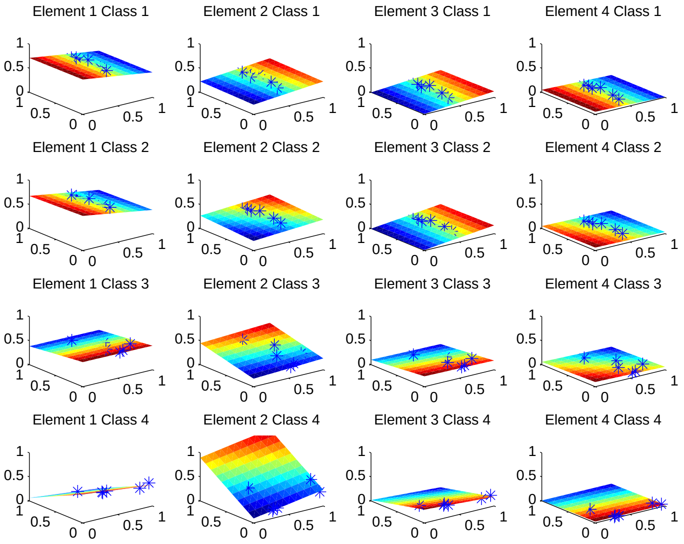

Figure 4.5: Graphic representation of applying linear regression to the composition vs membership in the simplex structure $\mathcal{S}^n$ ($D = 4$, $G = 4$, $N = 3$).

$\mathbf{d_1}$ and $\mathbf{d_0}$ for a combination of components i and j:

$$\begin{aligned}
\check{x}_{k,(ij)} &= \ln \frac{x_{k,i}}{x_{k,j}} \\
\check{y}_{k,(ij)} &= \ln \frac{y_{k,i}}{y_{n,j}} \\
\delta_{0,(ij)} &= \ln \frac{d_{0i}}{d_{0j}} \\
\delta_{1,(ij)} &= \ln \frac{d_{1i}}{d_{1j}}
\end{aligned} \tag{4.25}$$

from which follows that for any combination of i and j in $\{1, 2, ...D\}$, for which $i \neq j$ the compositional linear function is given by:

$$\check{y}_{k,(ij)} = \delta_{0,(ij)} + \delta_{1,(ij)} \cdot \check{x}_{k,(ij)} \tag{4.26}$$

If we want to perform this curve fitting 'in the simplex', we are forced to find a least squares solution to the following model (Daunis-i-Estadella et al., 2002):

$$\mathbf{y}_k = \mathbf{d}_0 \oplus x_k \odot \mathbf{d}_1 \oplus \varepsilon_k \tag{4.27}$$

with $\varepsilon_\mathbf{k} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_D]$ being the compositional discrepancies between model and observed composition. The compositional operators already indicate that a least squares solution is now obtained by minimising the Aitchison distance instead of the Euclidean distance. By doing so one yields a least-squares estimate of the compositional 'intersect'

$\mathbf{d}_0$ and 'slope' $\mathbf{d})1$.

The two curve fitting approaches are complementary in the sense that the two compositional fitting parameters relate to the log ratio parameters as followed:

$$\begin{array}{rcl} \delta_a & = & \mathrm{clr}(\mathbf{d}_a) \\ \mathbf{d}_a & = & \mathrm{clr}^{-1}(\delta_a) \end{array} \tag{4.28}$$

given $a = 0$ and $a = 1$. A synthetic example of this process is shown in figure 4.5 and 4.6. In figure 4.6 the relation between the additive log ratio composition (for all element combinations $i = 1,2,3$ and $j = 1,2,3$ for which holds that $i \neq j$) and some real variable is shown, which is in this case the grain size in ϕ-units. The plots in the middle column (element 2 is denominator element) indicate that the concentration of element 2 is very low for the high ϕ-values. However as the grain size increases, the concentration of element 2 increases. When putting the plots in figure 4.6 in the least squares framework, the fitted curves shown in the plots are obtained using ordinary least squares and characterized by δ_0 and δ_1.

The curve fitting results in the simplex are represented in figure 4.5; the uppermost plot shows the input data, which has a distinct compositional linear trend. As mentioned before, the observations in the rightmost corner (composition is dominated by component 2) represent the smallest ϕ-unit values. The plot in the middle shows the same results as obtained from the linear fit whereas the plot at the bottom shows the compositional slope and intersect. These last two compositions are the two compositions given by $\mathbf{d_0}$ and $\mathbf{d_1}$, given in equation 4.28.

4.2.6 The optimal transfer function

As stated in section 4.2.2, linear mixing problems behave perfectly linear when expressed in fractions and mixing proportions which can therefore be readily plugged into ordinary least squares regression. The same linear mixing returns a nonlinear compositional response when the composition is evaluated in the log ratio space. This nonlinearity is the most severe at high endmember membership, which is essentially the most important part in the suggested model because we obtain the endmember composition by *extrapolation*. On the other hand we try to stick to the model suggested by (Tolosana-Delgado and von Eynatten, 2008), in which is stated that these Transfer Functions can be approximated reasonably well by a linear model in log ratio space.

For the latter we have shown the framework to move freely between the simplex and the log ratio space by using the simplicial operators for perturbation and power transformation. By using this framework we can (1) apply linear curve-fitting w.r.t. the *log ratio* compositional response as a function of grain size and (2) apply linear curve-fitting w.r.t. the *simplicial* composition as a function of membership.

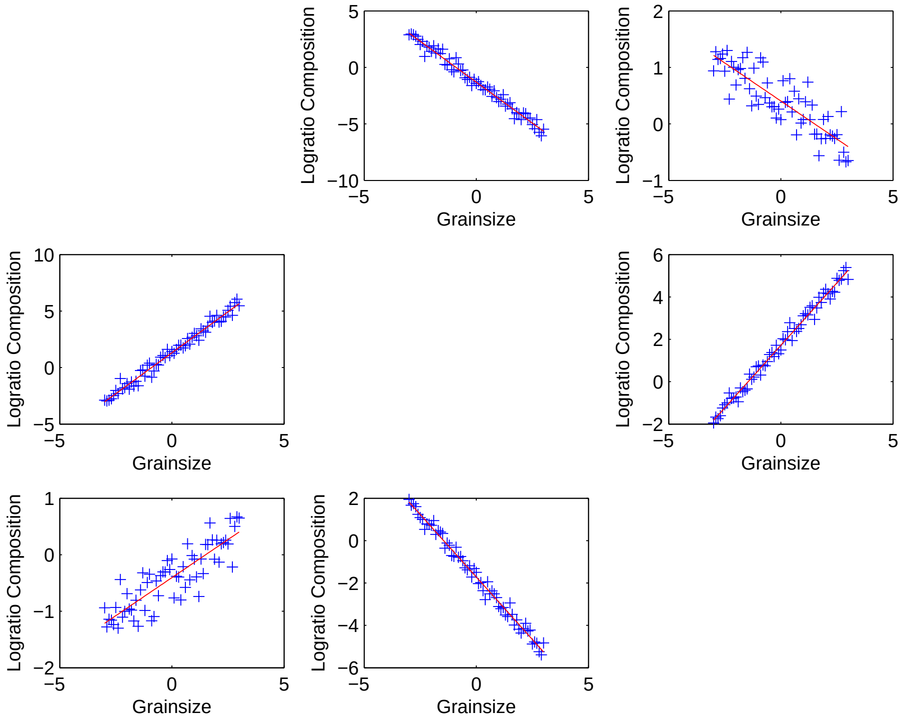

Figure 4.6: The results of applying a linear fit to all numerator/denominator pairs yields all the $[\delta_0, \delta_1]$ *pairs, given that* $i \neq j$ *(the diagonal).*

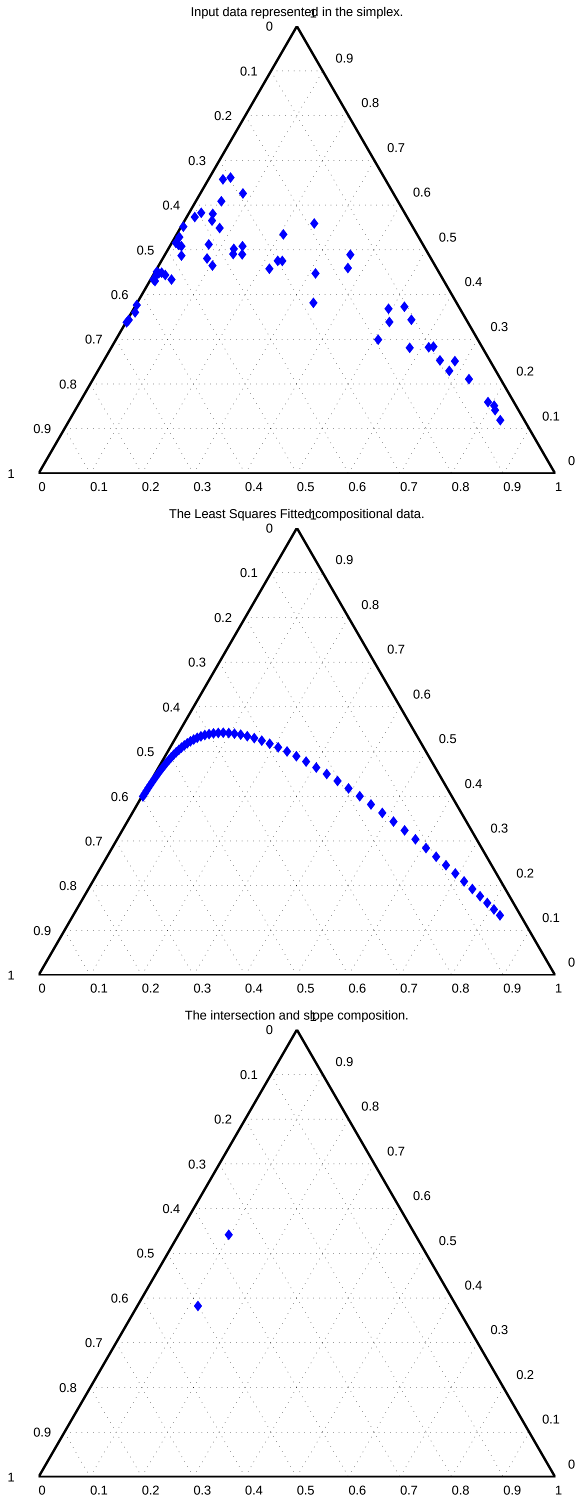

Figure 4.7: Input data (top), least squares linear regression results (middle) and the two compositional coefficients (bottom).

To revisit the primary objective; find a set of Transfer Functions with which the geochemical data can be reconstructed successfully and in the optimal sense. However the Transfer Functions affect both the geochemical data of the size fractions, the partial bulk and the bulk. Nevertheless we are able to estimate this Transfer Function from (1) the partial bulk chemistry we obtain from the fractionized data and (2) from the bulk chemistry i.c.w. the membership (see equation 4.20).

Because of this redundant information we adopt a weighted average between those two quantities as follows:

$$\bar{\mathbf{b}}_{\mathbf{p},i} = \text{mean}\{w\,\text{clr}[\hat{\mathbf{b}}_{\mathbf{p}_i}],\ \ (1-w)\,\text{clr}[\check{\mathbf{b}}_{\mathbf{p}_i}]\} \tag{4.29}$$

where w reflect the relative importance of the XRF-bulk-derived partial bulk w.r.t the partial bulk from the fractionized data. Now the 'optimal' endmember compositions are determined, we want to obtain an updated compositional slope and intersect $\mathbf{d_1}$ and $\mathbf{d_0}$ such that we reconstruct these 'optimal' endmember compositions as follows:

$$\bar{\mathbf{b}}_{\mathbf{p},i} = \mathbf{C_i}^T \cdot \mathbf{f_i} \tag{4.30}$$

where $\mathbf{C_i}$ is again a $D \times G$ matrix of the discretized TF of endmember i, discretized into the grain size classes with mean grain sizes $\{\phi_1, \phi_2, ...\phi_G\}$:

$$\mathbf{C_i} = [\mathbf{d}_0 \oplus \phi_1 \odot \mathbf{d}_1 \quad \mathbf{d}_0 \oplus \phi_2 \odot \mathbf{d}_1 \quad \cdots \quad \mathbf{d}_0 \oplus \phi_G \odot \mathbf{d}_1] \tag{4.31}$$

We can rewrite $\mathbf{C_i}$ using the perturbation and power transformation rules into the matrix $\mathbf{D}$ with column vectors φ_g ($g = 1, 2, ...G$; G the number of grain size classes) representing the composition of a infinitely small grain size class with mean ϕ_g and with f_g representing the mass-fraction in the specific class:

$$\mathbf{D} = \left[\begin{array}{cccc} \varphi_1 & \varphi_2 & \cdots & \varphi_G \end{array}\right], \qquad \varphi_g = f_g \cdot \mathcal{C}_v[\mathbf{d_0} \cdot \mathcal{C}[\mathbf{d_1}^{\phi(g)}]] \tag{4.32}$$

where $\mathcal{C}_v[..]$ denotes the *vertical* closure operation.

Even though this system can be solved or at least approximated (minimising the clr distance) using an optimisation toolkit, applying this system might become computationally expensive. Furthermore there are some limitations of solving this system for both the compositional slope and intersect $\mathbf{d_1}$ and $\mathbf{d_0}$. One of the limitations is that in order to have an overdetermined system in equation 4.32, the dataset should fulfill the requirement that $D \geq 2N$. Otherwise the optimisation can only be applied in a simplified form by solving either for the compositional intersect $\mathbf{d_0}$ or the compositional slope $\mathbf{d_1}$.

An advantage of applying a simplified approach by only estimating $\mathbf{d_0}$, is that the system not only becomes less computationally expensive, it can also be solved using standard linear linear methods. If we assume the compositional slope $\mathbf{d_1}$ (estimated from the XRF of grain size fractions) as being 'true', the term $\mathcal{C}[\mathbf{d_1}^{\phi(g)}]$ will be constant for a given grain size class, which we will now refer to as by the (closed) vector $\mathbf{h}$:

$$\varphi_g = f_g \cdot \mathcal{C}_v[\mathbf{d_0} \cdot \mathcal{C}[\mathbf{d_1}^{\phi(g)}]] = f_g \cdot \mathcal{C}_v[\mathbf{d_0} \cdot \mathbf{h}_{(g)}] = f_g \cdot \mathbf{h}_{(g)} \cdot \mathbf{d_0} \tag{4.33}$$

since matrix $\mathbf{D}$ does not have nonlinear terms anymore we can sum over all grain size classes $g = 1, 2, ...G$ of which we put the result on the diagonal of a matrix $\mathbf{E}$:

$$\mathbf{E} \cdot \bar{\mathbf{d}}_0 = \begin{bmatrix} \sum_{j=1}^{G} h_{j,1} \cdot f_j & 0 & 0 & 0 \\ 0 & \sum_{j=1}^{G} h_{j,2} \cdot f_j & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sum_{j=1}^{G} h_{j,D} \cdot f_j \end{bmatrix} \cdot \begin{bmatrix} d_{0,1} \\ d_{0,2} \\ \vdots \\ d_{0,D} \end{bmatrix} = \bar{\mathbf{b}}_{\mathbf{p}} \tag{4.34}$$

From this system the new perturbing composition $\bar{\mathbf{d}}_{\mathbf{0}}$ follows directly from:

$$\bar{\mathbf{d}}_{\mathbf{0}} = \mathcal{C}\left[\mathbf{E}^{-1} \cdot \bar{\mathbf{b}}_{\mathbf{p}}\right] \tag{4.35}$$

Surprisingly enough the nonlinear system of equation which could additionally only be solved under conditions w.r.t. the number of grain size classes and elements, can now be solved by simple element-by-element division.

Even though we have obtained a physically feasible linear mixing model structure, there are infinitely many model configuration possible. This because not only the EM-TF configuration can be altered; the model outcome is also determined by the given weight value in the weighted averaging step (see equation 4.29). How we determine this weight parameter will be discussed in the next section.

4.2.7 Model performance

Before moving on to the results of the linear mixing model, a framework must be defined that will determine the quality of the linear mixing (LINMIX) model. There are in principle three types of data sources that were used to obtain the model predictions; endmember data, chemical compositions of the grain size fractions and the XRF core scanning data. A satisfying model will yield discrepancies between model and data that are not significantly larger than the uncertainty of the input data itself.

An estimate of discrepancy between a model prediction and a data source of the compositional quantity W and comprising the same set of D elements is defined as follows:

$$F = \sqrt{\sum_{i=1}^{D} \left(\ln\left(\frac{\hat{W}_i}{\operatorname{geom}(\hat{W})}\right) - \ln\left(\frac{W_i}{\operatorname{geom}(W)}\right)\right)^2} \tag{4.36}$$

Estimates of the discrepancy parameter F can be determined for a whole series of different data sources and models. The same approach can be used to determine the quality of a data source by evaluating the dispersion of replicate samples analyzed and sampled in the same way and using the same sample material. In principle we can therefore distinguish between the F-parameters that relate to the quality of the input:

- Uncertainty of the XRF-bulk data;
- Uncertainty of the fractionized data;
- Uncertainty of the Endmember data,

whereas the discrepancy between data quality and the predictions by the LINMIX model:

- XRF-bulk down core composition vs. the LINMIX modeled down core composition,
- Input fractionized data vs. LINMIX modeled fractionized data,
- Input endmember data vs. LINMIX modeled endmembers.

will be denoted by $\hat{F}$. In principle the most optimal LINMIX model yields fractionized data, endmember data and XRF-bulk at the same accuracy of the input data itself. However this complete error modeling structure can only be adopted when there are reliable estimates of the uncertainties of all the input data sources.

However even if these uncertainty estimates are available, this error modeling structure becomes rather complex and computationally expensive when incorporating the accuracy of the GSD shape as well as the membership record. Also the error propagation with respect of the accuracy of the fractionized data and its effect of the final TF's is not easily solved analytically. This because the errors are determined by several aspects:

1. Quality of the endmember unmixing;
2. How well does the model behave according to 'reality' with respect to the compositional linear trend;
3. Quality of the geochemical data;
4. Within class distribution of mass;
5. Possible between-class overlap of mass.

For simplicity, we will adopt a simplified version of the error modeling by adopting the assumption that all error sources except (2) have a negligible effect on the final prediction error. The consequence of this assumptions is that the discrepancy F between pseudo-observations and (initial or optimized) compositional linear trend can be derived from the data and used to asses the model performance. Another assumption that is introduced is that the accurracy of the model input can be satisfyingly parameterised by their associated mean and standard deviation. Given that F_g denotes the integrity of the chemistry of the size classes, it is parameterised by $\{\mu_b, \sigma_b\}$. Obtaining a reliable estimate of the XRF-bulk integrity on the other hand is easily done by evaluating the residuals between the XRF calibration model (LRCE) and the calibration samples. Nevertheless this methodology is implicitly also based on the assumption that the log ratio linear calibration model is 'true' and the samples used to calibrate with are unbiased. Analogous to the chemistry of the size fractions, again the accuracy is modelled by two parameters; the mean and the standard deviation $\{\mu_x, \sigma_x\}$.

The second group of discrepancy values $\hat{F}$ that was introduced represented the performance of the LINMIX model; how far does the LINMIX model deviate from the input data. This subsequently determined for the bulk chemistry by evaluating the LINMIX model predictions and the predictions by XRF core scanning. Additionaly the discrepancies between the pseudo-observations, representing the theoretical compositional change with grain size, and the final logratio-linear TFs give an estimate of the goodness of fit between the data and the compositional linear functions. Now the framework is defined for both data and model prediction, model acceptance criteria are adopted. This includes one acceptance criterion that relates to the bulk chemistry, c_b, and one criterion that relates to the analysis of the size fractions c_g. If for a given model configuration and w the following holds, the model will be accepted:

$$\begin{aligned} c_g = 1 \quad &| \quad \operatorname{med}\{\hat{F}_g\} \leq \tfrac{r\sigma_g}{\sqrt{n_g}} + \mu_g \\ c_b = 1 \quad &| \quad \operatorname{med}\{\hat{F}_b\} \leq \tfrac{r\sigma_x}{\sqrt{n_b}} + \mu_x \end{aligned} \tag{4.37}$$

where $\{\mu_g, \sigma_g\}$ characterize the discrepancies between pseudo-observations and derived compositional linear TFs and $\{\mu_x, \sigma_x\}$ represent the mean and standard deviation of the XRF calibration residuals. The values of n_g and n_b represent the number of observations that were effectively used to calibrate the linear mixing model and r is an integer relaxation-parameter quantifying the strictness.

The 'optimal' solution for the TF-EM configuration and w that is accepted according to the criteria given in equation 4.37 follows by minimising the objective function $\mathcal{H}_{(w,r)}$:

$$\min\{\mathcal{H}\} = \min\left[\frac{\sqrt{n_g}}{r\sigma_g}\left(\operatorname{med}\{\hat{F}_g\} - \mu_g\right)^2 - \frac{\sqrt{n_b}}{r\sigma_b}\left(\operatorname{med}\{\hat{F}_b\} - \mu_x\right)^2\right] \tag{4.38}$$

This relations means in principle that we want to keep the distance between LINMIX-bulk and XRF-bulk data equally large as the distance between LINMIX-fractionized and XRF-fractionized geochemistry. Analogous to the acceptance criteria the number of observations is also incorporated; if a large number of observations is available we accept a smaller distance between model prediction and data. The model that is (1) accepted according to the acceptance criteria for any w and that (2) shows the minimal value for $\mathcal{H}$ for a specific (range of) w is now identified as the 'optimal' model.

4.2.8 Summary of the work flow

To summarize the whole work flow, a flowchart of the unmixing of the geochemical data is shown in figure 4.8. Here one can see that the geochemical data determined by using calibrated XRF core scanning is the guide for the optimisation process. This data will represent the bulk chemistry of the mixture from which the partial bulk chemistry $\hat{\mathbf{b}}_p$ can be obtained using the unmixed GSD's and the inverse model (see equation 4.14). In order to do this, the membership is plugged into the $\mathcal{S}^n$ simplicial framework, from which the partial bulk is determined using constrained least squares methods. The calculation of the first guess Transfer Function matrix $\mathbf{C}$ is also done by applying the inverse linear mixing framework, however now by applying this framework to every grain size class *individually*, which yields $\mathbf{b_g}$ (see equation 4.24).
After this initialisation step the forward model is applied such that the partial bulk chemistry is obtained by applying linear mixing using the GSD's,TF's and Memberships. This results in an estimate of the partial bulk given by $\hat{\mathbf{b}}_\mathbf{p}$ (see equation 4.21). A weighted average between the partial bulk from XRF $\hat{\mathbf{b}}_{\mathbf{p},i}$ and the derived partial bulk chemistry from the fractionized data $\check{\mathbf{b}}_{\mathbf{p},i}$ yields an updated version of $\overline{\mathbf{b}}_{\mathbf{p},i}$ which will be the input of the compositional least square fitting, given in equation 4.35. By calculating the necessary *shift* (the compositional slope is not altered) in log ratio space of the Transfer Functions in order to reproduce $\overline{\mathbf{b}}_{\mathbf{p},i}$, an updated Transfer Function is obtained which yields the geochemical prediction in the optimal sense, as shown in equation 4.21.

One should keep in mind though that the LINMIX model, as it is defined here, is based on two assumptions:

> *The compositional change as a function of grain size can be sufficiently parameterised by a logratio linear model.*

> *The compositional slope, as determined from the fractionized data is 'true'*

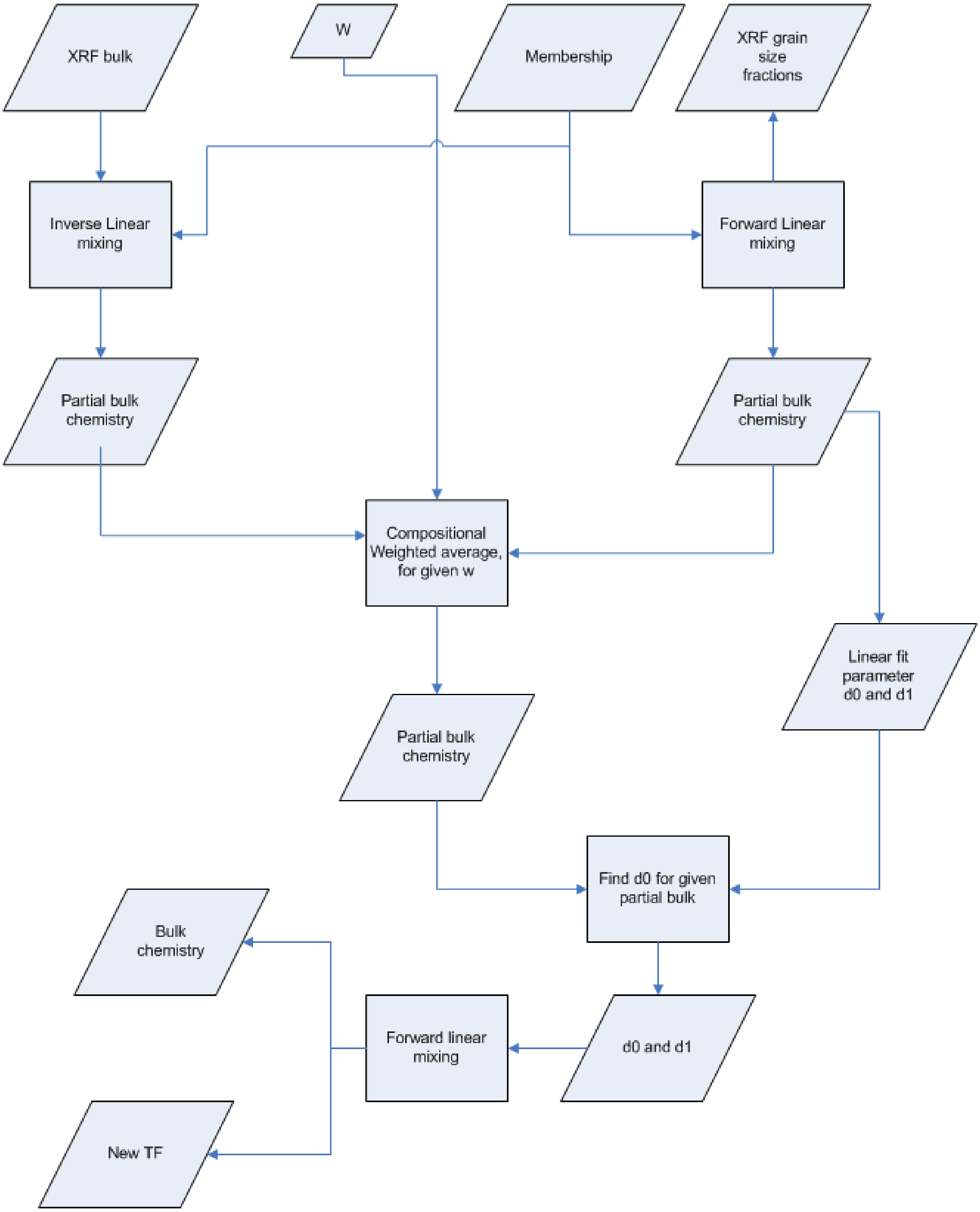

Figure 4.8: Flowchart of LINMIX model.

Fraction	grain size $[\mu m]$	phi-units
1	< 2	10.6267
2	$2 - 10$	7.8048
3	$10 - 40$	5.6438
4	$40 - 100$	3.9829
5	> 100	2.8219

Table 4.1: Isolated grain size fractions with the grain diameter range and average grain size in phi-units.

	n_b	n_g	n_m
n_b	444	6	92
n_g	-	6	6
n_m	-	-	191

Table 4.2: Number of observations of each dataset (b = XRF-bulk, g = analysis of grain size fractions and m = GSD analysis) and their down core intersections.

4.3 Application to Core 9508; Offshore Senegal

In this section the LINMIX model was applied to a real dataset; the data of core 9508 that was described in more detail in section (tag). The core was obtained offshore Africa and was studied in great detail w.r.t. its paleo-climatological information content. Additionally grain size data, XRF core scanning data as well as XRF of isolated grain size fractions has been acquired; *fractionized data*. A summary of the number of observations in each dataset as well as their down core intersections is shown in table 4.2.

4.3.1 Unmixing grain size distributions

Representative samples were taken from the core which were analyzed w.r.t. the grain size distribution. Grain size distributions were numerically unmixed by applying the endmember unmixing algorithm EMMA (Weltje, 1997). This algorithm searches for the optimal set of grain size distributions that sufficiently reconstructs all the observed GSD's. Additionally the number of endmember-GSD's used to reconstruct the observations is minimized. Analogous to PCA methodologies, the discrepancy between the mixed GSD's and the observed GSD's descreasing while increasing the number of endmember-GSD's. Given that a goodness-of-fit is adopted, a cutoff can be applied such that the number of endmember-GSD's is manageable. For the data of core 9508 it was found that a three endmember system (see figure 4.10) yields an acceptable (mean coefficient of determination of $R^2 = 0.84$) approximation of the observed grain size signal (see figure 4.9).

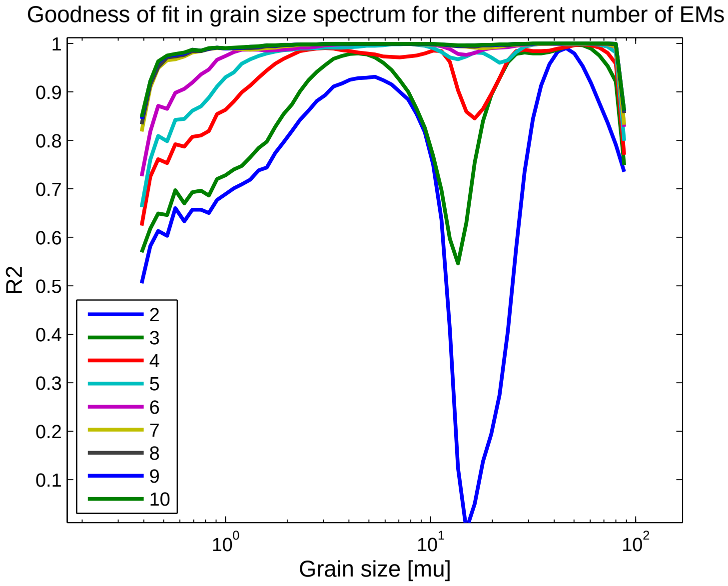

Figure 4.9: Goodness-of-fit between grain size observations and the EMMA unmixing results for different number of endmembers. Adopting three endmembers results in a mean R^2 value of 0.84.

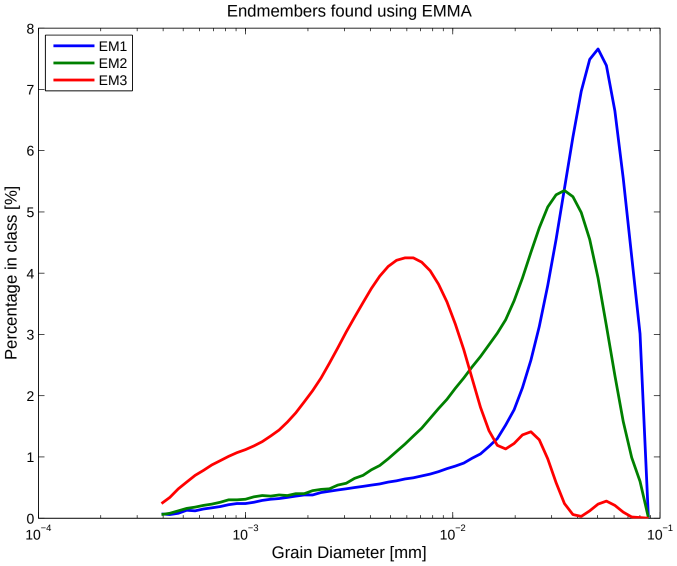

Figure 4.10: Endmembers found for Core 9508 using the EMMA algorithm.

	Al	Fe	K	Si	Ti
Al	-	0.8587	0.7686	0.9389	0.5151
Fe	0.8587	-	0.9518	0.9440	0.9200
K	0.7686	0.9518	-	0.7415	0.7446
Si	0.9389	0.9440	0.7415	-	0.7450
Ti	0.5151	0.9200	0.7446	0.7450	-

Table 4.3: R^2 values for all log ratio pairs for the 9508-dataset.

The result of this unmixing step is a membership-record comprising three compositional variables.

4.3.2 Calibration of the core Scanning data

The XRF core scanning data was calibrated and thereby turned into a semi-quantitative dataset by using the LRCE discussed in section 3.1. A useful statistic to quantify the performance of the log ratio calibration models is the R^2 statistic. The R^2 statistic, which ranges from zero to unity, may be interpreted as the magnitude of down-core variability relative to the width of the error bar. The R^2 statistic for all log-ratio pairs is shown in table 4.3 where the columns represent the denominator.

According to the R^2 statistics in core 9508 the best calibration results are obtained by using a log ratio model with Fe as common denominator which yielded $R^2 \geq 0.85$ for all log ratio pairs. The down core compositional record, obtained by applying this calibration model are shown in figure 4.11. In the data of core 9508 we observe, analogous to the work of Mulitza et al. (2008), several distinct and sudden decreases in the Fe-concentration, leading to an increase of the logratio's, which were ascribed to be the result of an arid climate in the Sahel. The predicted element concentrations obtained using the calibration step are now ready to be used in the linear mixing model. The major advantage of using this calibrated XRF core scanning data over the calibration samples alone is that we have obtained a relatively high resolution geochemical record with many overlapping data points with the grain size record ($n_{m,b} = 92$).

4.3.3 Model acceptance

We would like to investigate whether the acceptance criteria as well as the values of the objective function have potential to provide an semi-automatic geological interpretation. In other words, will the model identify endmembers 1 and 2 as being material of the same source (eolian material). In order to test this we setup the following objective:

Find the model that is accepted according to the acceptance criteria and that

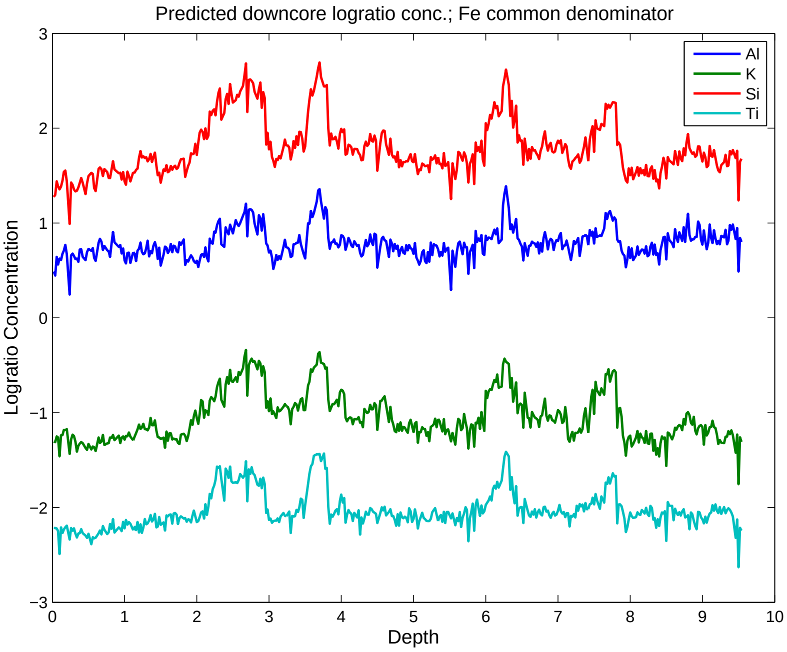

Figure 4.11: Down core logratio (alr) composition of core 9508 with Fe as common denominator.

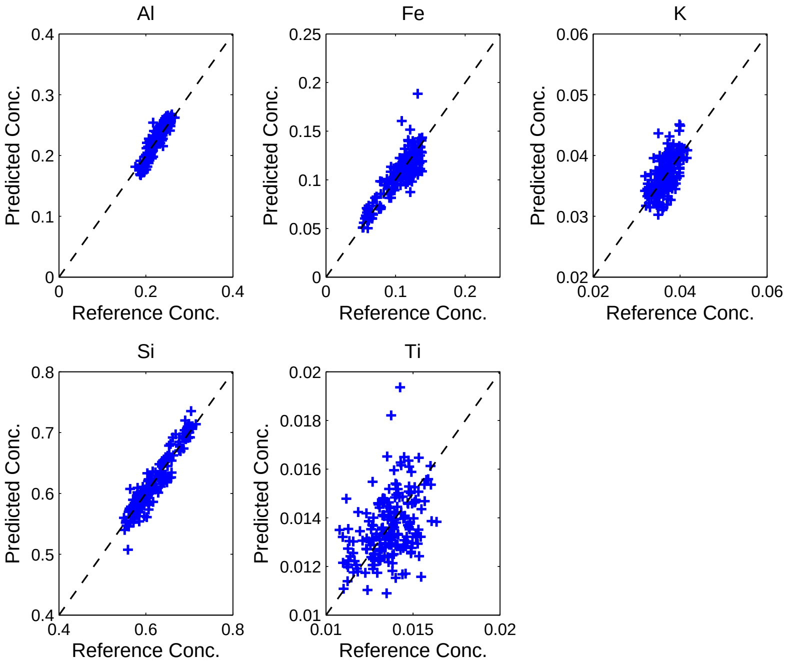

Figure 4.12: Results of the calibration of the XRF data with the calibration sample composition on the x-axis and predicted composition on the y-axis.

TF1	TF2	TF3	$\text{med}\{F_g\}$ $w = 0$	$\text{med}\{F_g\}$ $w = 1$	$\text{med}\{F_b\}$ $w = 0$	$\text{med}\{F_b\}$ $w = 1$
{1,2,3}	-	-	0.2869	0.5311	0.4382	0.1212
{1,2}	3	-	0.4903	0.6532	0.3595	0.1169
{1,3}	2	-	0.3745	0.5148	0.2883	0.1511
{2,3}	1	-	0.5658	0.9546	0.3126	0.1506
1	2	3	0.6118	0.8248	0.4120	0.1131
						($\mu_x = 0.0960$)

Table 4.4: Different configurations of endmember-TF pairs with the discrepancies.

Model #	Complexity	TF1	TF2	TF3	$c_b = 1 \vert c_g = 1$ $r = 1$	$c_b = 1 \vert c_g = 1$ $r = 2$
1	1	{1,2,3}	-	-	empty	empty
2	2	{1,2}	3	-	empty	$w = 0.75$
3	2	{1,3}	2	-	empty	empty
4	2	{2,3}	1	-	empty	empty
5	3	1	2	3	empty	empty

Table 4.5: Configurations and the results w.r.t. the acceptance criteria as well as w for the optimal solution in terms of $\mathcal{H}$.

comprises a minimal complexity.

where the complexity is the number of TF's that are used to reconstruct the bulk chemistry and fractionized data. By doing so, we want to find out whether a prediction quality parameter could assist in the geological interpretation. The methodology we apply to investigate this is to run the model using every possible TF-endmember set. For every possible model configuration the F-distance parameter was determined for both $w = 0$ and $w = 1$ which, in words, simply means that we asses how 'closely' the model can resemble the fractionized data and bulk chemistry, respectively.

4.3.4 Acceptance results

Table 4.4 shows the results of the discrepancy parameter for the different LINMIX model configurations. There are a couple of things we observe in these results. First of all we observe that the least complex model $\{EM1, EM2, EM3\}$, has the best fit w.r.t. the compositional linear trend. This follows from the discrepancy value of 0.2869 for $w = 0$. Additionally we observe that models 3 and 4 are, for as far as the bulk chemistry is concerned, the least capable of reconstructing the bulk chemistry ($\text{med}\{F_b\} = 0.1511$ where the XRF data has $\mu_x = 0.0960$). Although model 1 is the most respectful for the compositional linear trend, it is not capable of reconstructing the bulk-chemistry satisfyingly which results in the rejection of this model for all w.

However, as indicated in table 4.5, model 2 is accepted which has a minimum $\mathcal{H}$-value for $w = 0.75$. Given the interpretation of the different endmembers, this is in line with the findings of Mulitza et al. (2008). They state that endmembers 1 and 2 reflect both eolian sediment whereas endmember 3 is the fluvial portion. The model outcome is that the compositional change that is observed with a *relative* change of endmembers 1 and 2 is only the result of grain size variations. In other words, that endmembers 1 and 2 have in principle the same source material and were transported in the same way. However to reconstruct the geochemistry and fractionized data, endmember 3 must be modelled using an additional TF indicating different a different source.

4.3.5 Unmixing results

Shown in figure 4.13 are the partial bulk chemistry from the XRF-bulk as well as the bulk chemistry as determined using model 1 and $w = 1$. At the corner points of the simplex in the xy-plane plane, the partial bulk chemistries are shown from respectively the XRF-bulk (green) and the forward linear mixing (black). On the other hand the blue and red markers indicate the linear mixing model prediction for all membership observations by mixing of the two different sets of partial bulk chemistries (XRF and fractionized data, respectively).

The *initial* predictions by the two different sets of partial bulk chemistries are not in correspondence; for K for instance, the initial guess of the two bulk chemistries are not overlapping w.r.t. the fraction of bulk (red. the z-direction of the figures). Also the ’red’ prediction of Al is significantly lower compared to the ’blue’ prediction. However when looking at the optimized partial bulk in figure 4.14 which was obtained by applying $w = 1$, one can immediately see the strong improvement between predicted bulk and observed XRF-bulk. This is the most apparent for the Al, Ti and K concentration. There seems to be no significant change in the Fe-concentration which is, given the fairly good initial guess, as expected.

The down core bulk chemistry as obtained using the same model and model parameters is shown in figure 4.15, where one can see that the LINMIX estimate of the bulk chemistry has a lower variance compared to the XRF-bulk. However when looking at the initial guessed bulk chemistry, the optimisation of the TFs has definately resulted in an improved prediction.

In order to further investigate the improvement, it is also worthwhile to look at the predicted bulk chemistry down core in log ratio space (figure 4.18). What becomes clear from this figure is that whereas Al/K and Fe/K are estimated rather accurately, the Ti/K ratio is poorly reconstructed. There even seems to be negative correlation between the

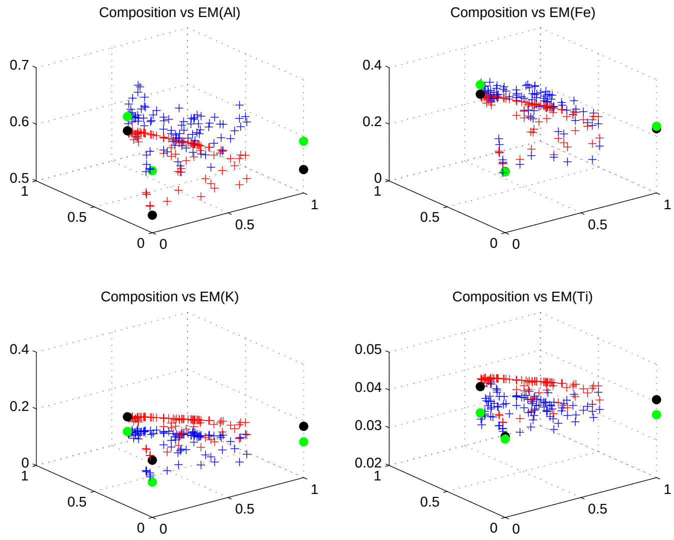

Figure 4.13: Initial guess of the partial bulk chemistry from XRF-bulk (blue) and the results from the forward linear mixing model (red) using 1 TF. The composition of the pure endmembers are shown at the corner points as derived from the XRF-bulk $\hat{\mathbf{b}}_{\mathbf{p}_i}$ *(green) and the forward linear mixing* $\check{\mathbf{b}}_{\mathbf{p}_i}$ *(black).*

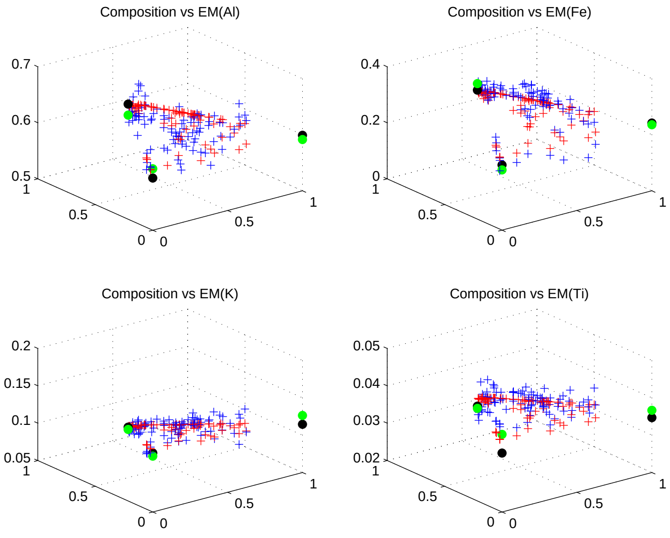

Figure 4.14: Optimized results of the partial bulk chemistry from XRF-bulk (blue) and the results from the forward linear mixing model (red) using 1 TF. The composition of the pure endmembers are shown at the corner points as derived from the XRF-bulk $\hat{\mathbf{b}}_{\mathbf{p}_i}$ *(green) and the forward linear mixing* $\check{\mathbf{b}}_{\mathbf{p}_i}$ *(black).*

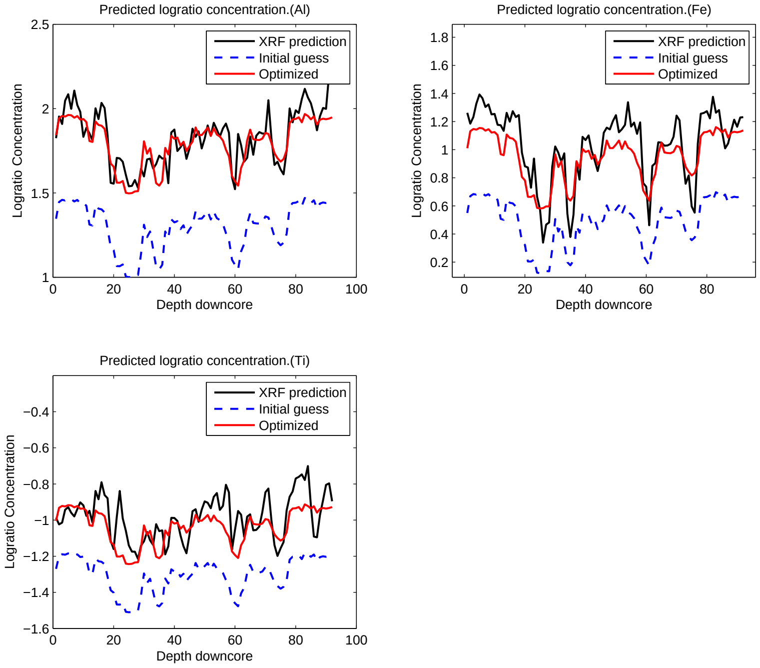

Figure 4.15: Reference, initial guess and optimized bulk chemistry obtained using a 1-TF forward linear mixing. Shown here are only the locations where both GSD and XRF-bulk was observed.

observed and predicted Ti/K concentration for some core sections. Since in model 1 only one TF was adopted, the level of variance that can be achieved in the predicted bulk chemistry is limited; we are only altering the *intersect* and not the *slope* of the TF's which implies that with the fixed GSD's, the variance of the bulk that can be achieved is restricted.

When looking at the results of model 2 in figure 4.16, it appears that the down core prediction is in correspondence with the observed bulk chemistry. The peaks and troughs are reconstructed effectively than for the 1-TF model. Furthermore in figure 4.20, the improvement of the optimisation step is very clear in the discrepancies between LINMIX and XRF-bulk. Also the mean of the intrinsic data uncertainty and the discrepancy between the XRF-bulk and LINMIX are almost the same, which lead to the acceptance of this model. In figure 4.17, the 2 determined TF's of model 2 are shown which indicate that the two TF's are very similar, except for the sign difference in slope of the Fe curve. Whereas the TF that describes the two coarsest endmembers has an increasing Fe concentration with fining sediment, the other TF shows a decrease in Fe content with fining sediment. Given that these TF's reflect the history and mode of transport of the sediment,

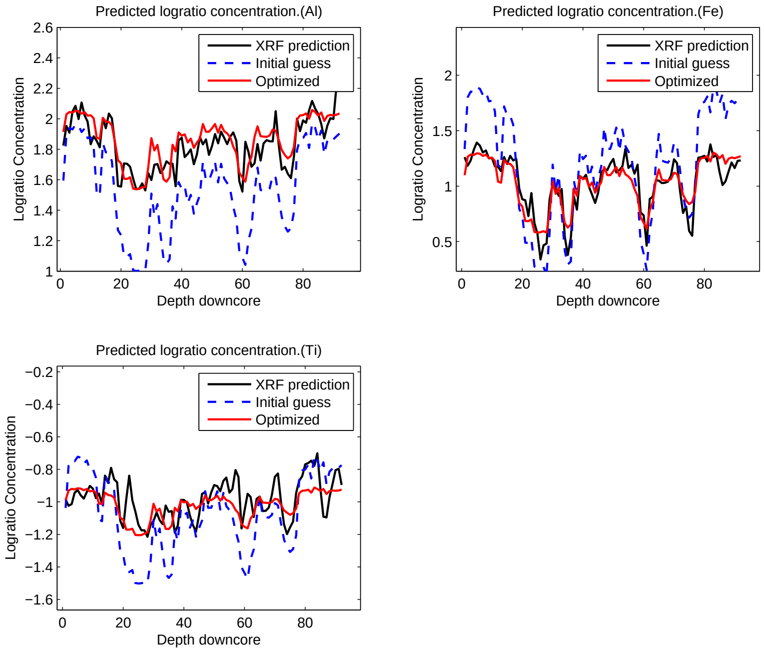

Figure 4.16: Reference, initial guess and optimized bulk chemistry obtained using a 2-TF forward linear mixing. Shown here are only the locations where both GSD and XRF-bulk was observed.

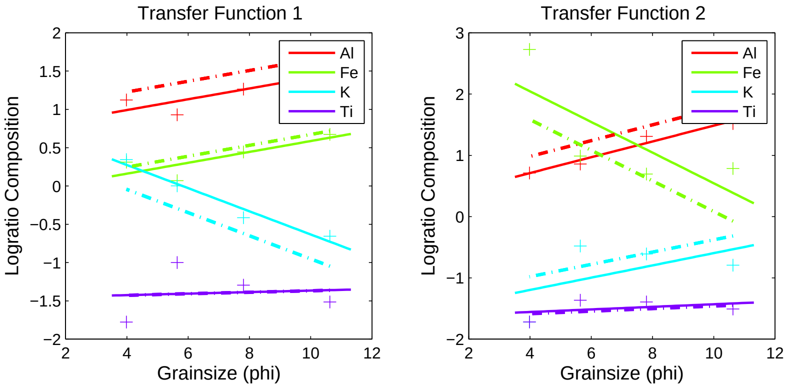

Figure 4.17: Initial (solid line) and optimized (dashed line) Transfer Functions of model 2 with $\{EM1, EM2\}$ (left) and EM3 (right) and $w = 0.75$.

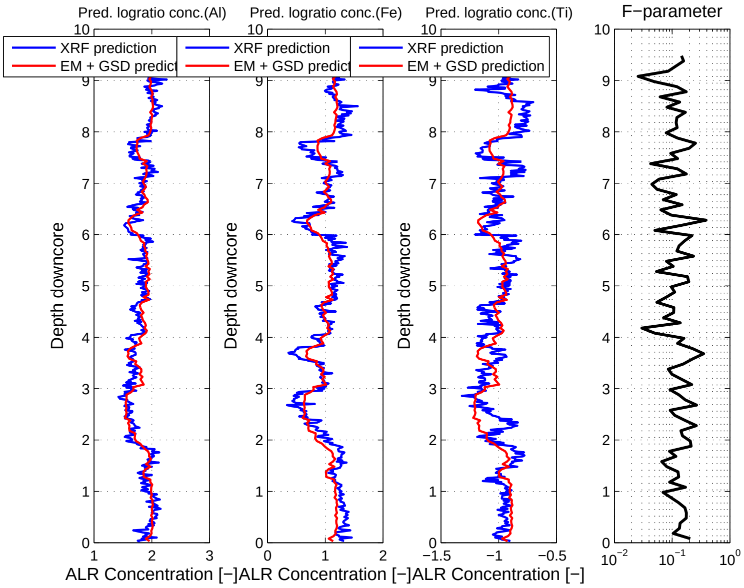

Figure 4.18: XRF-bulk and 1-TF linear mixing results (expressed in log ratio composition; K as common denominator), obtained using $w = 1$ *and model 1.*

the next section discusses the geological context of the derived TF's.

4.3.6 Interpretation

Given that the interpretation by Mulitza et al. (2008) is true, we would expect this to come forward in the chemical fingerprint. If we now turn to the endmember compositions (figure 4.21) found for model 2, it turns out that for the eolian endmembers the Fe/K proportion is around 2. For the fluvial portion on the other hand, the Fe/K ratio is around 4. If we compare these values with the present-day sediment compositions for the fluvial (Gac and Kane, 1986) and eolian (Orange et al., 1993) material of resp. 4.83 and 2.35 we can conclude that the values derived from the LINMIX model are realistic in that sense. Furthermore it turns out that the choice of either model 1 (figure 4.21, lower plots) or model 2 (figure 4.21, upper plots) has a negligible effect on the final endmember composition. In order to assess the overall behavior of the TF, we convert the rather difficult to interpret TF to the composition of two infinitely small grain size classes (see figure 4.22). The selected grain size classes ($\phi = 3$ and $\phi = 10$) correspond respectively to the coarsest and finest sediment in the mixture. From these 2-TF results can be derived that, in contrast to the fluvial TF, the eolian TF (TF 1; the uppermost pie-plots) is rather constant along the grain size spectrum. This observations is in line with what we expect; in an arid environment, chemical weathering is almost absent leaving the compositional change with grain size primarily being determined by the physical breakdown of grains.

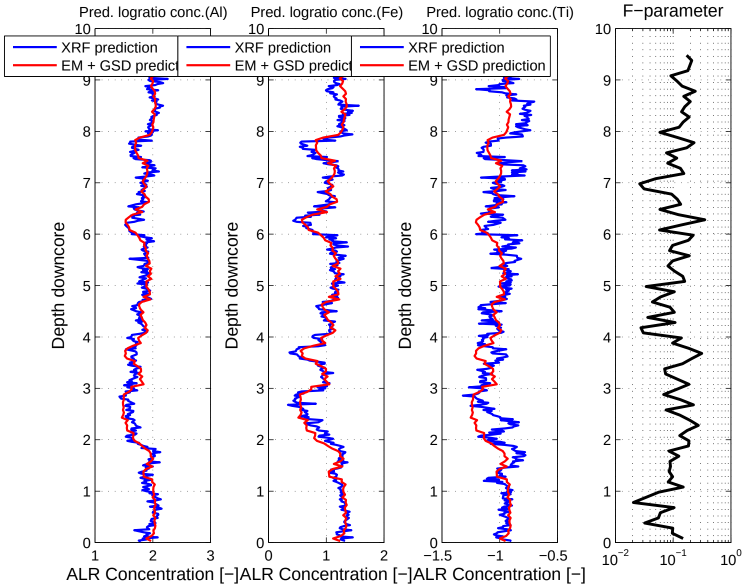

Figure 4.19: XRF-bulk and 2-TF linear mixing results (expressed in log ratio composition; K as common denominator), obtained using $w = 0.75$ *and model 2.*

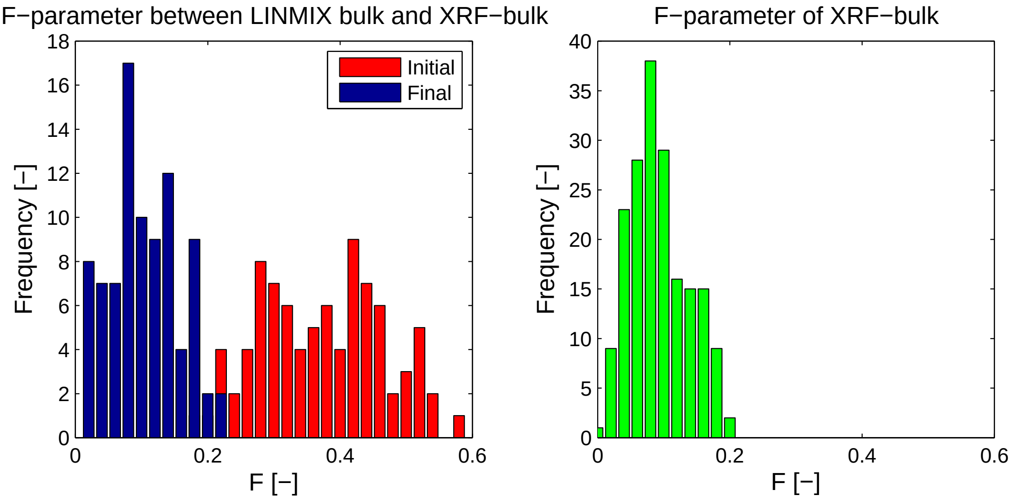

Figure 4.20: Histogram of the initial and final bulk discrepancies(core 9508) and the discrepancies as obtained in the calibration process for model 2 and $w = 0.75$*.*

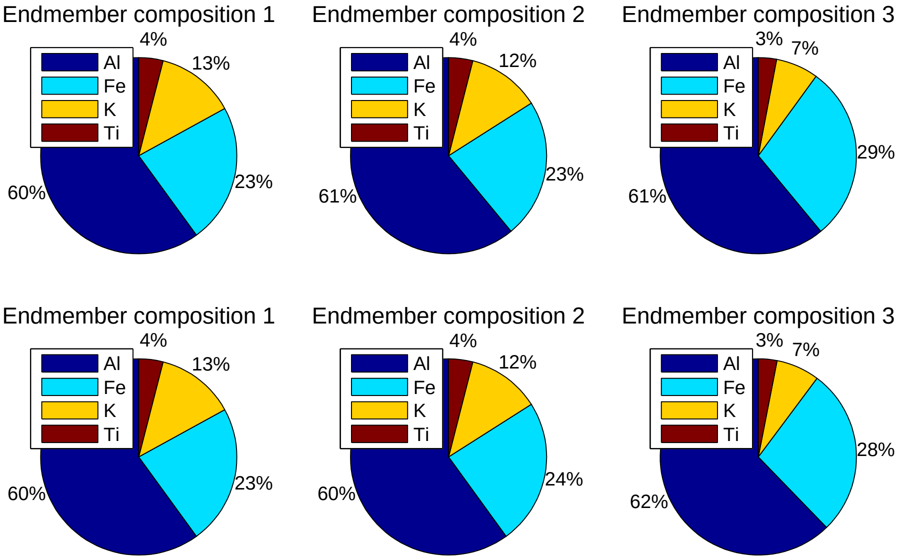

Figure 4.21: Compositions of the pure endmembers for the configuration with TF's for resp. $\{EM1,EM2\}$ *and* $\{EM3\}$ *(model 2) and* $w = 0.75$ *(upper) and* $\{EM1,EM2,EM3\}$ *(model 1) and* $w = 1$ *(lower).*

The opposite is true for very humid environments of which the fluvial portion of the sediment is considered the product. Observations show that sediment exposed to a very humid environment is often exposed to severe chemical weathering. Typical for intensive chemical weathering is the residual accumulation of (hydr)oxides of Fe, (Mn), and Al (see cover photo). When looking at the Al concentration of the fluvial TF, relative enrichment in Al and K is observed in the fine portion. Fe on the other hand is decreasing with grain size. With respect to the Ti concentration can be stated that the compositional change with grain size is low, as is the overall fraction of Ti in the bulk. Given that the LINMIX-model yields a composition with grainsize in a functional form, this function can be extrapolated to a considerable grain size which leads to the composition of the parent rock. Although the extrapolation towards a source rock composition may work in theory, the composition is expressed in only a small number of elements (red. rock-forming elements) which does not provide enough information to classify it accordingly. Furthermore, this logratio-linear model is adopted merely for mathematical convenience and does not necessarily reflect reality. Especially near the edges of the sedimentological grain size spectrum, the validity of the logratio-linear model is not ensured, as found by Tolosana-Delgado and von Eynatten (2010).

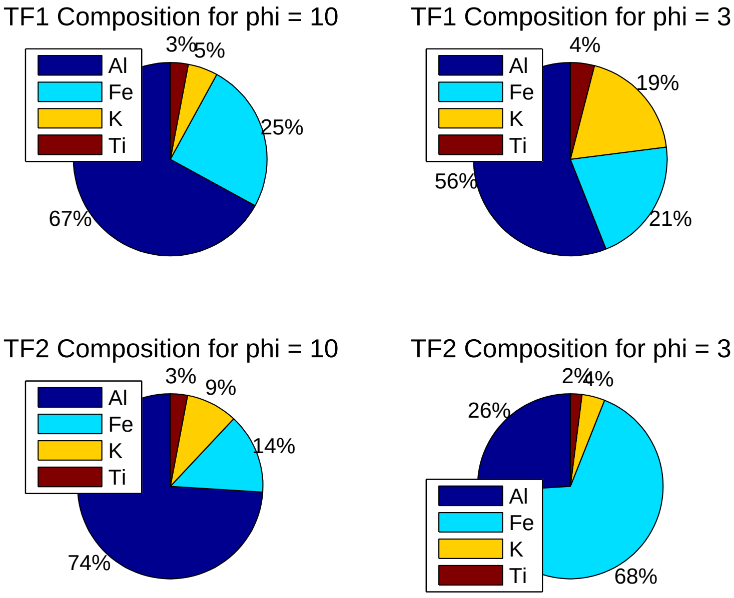

Figure 4.22: Model 2 TF compositions for $\phi = 10$ *and* $\phi = 3$.

4.4 Discussion and Conclusions

This chapter discussed the following research question:

> **Can we build a forward model of basin infill in a formal geochemical and textural context?**

The answer to this question is that we have successfully parameterised a forward model for the process of sediment mixing. Additionally it was applied to data of a Quaternary core of unconsolidated sediment of which was known that it reflects two different sediment sources. According to the defined acceptance criteria and relaxation parameter r, it is the only model that was accepted. Given the results of the acceptance framework, we conclude that model 2 is the most likely model that underlies the sedimentation. This conclusion is based on the fact that it provides a satisfying reconstruction of the input data. This model outcome is in line with the geological interpretation since the only accepted and optimal model configuration corresponds with the geological interpretation. Furthermore the composition of the endmembers was found to corresponds reasonably well with present-day values. This includes the composition of eolian dust from the Sahel mountain (model; Fe/K=4, observations; Fe/K=4.8) and the composition of fluvial sediment from the Senegal river (model; Fe/K=2, observations; Fe/K=2.3).

Although this is what follows from the data, one must keep in mind that the model is based on several assumptions. One of these assumptions is that the underlying TF can be sufficiently described with a compositional linear function. Additionally the compositional slope was assumed to be correctly derived from the fractionized data. This implies that if other or more data is available, the model acceptance could lead to different results. Nevertheless the adoption of the most complex model is not *necessary* to reconstruct the available data.

Chapter 5

In-situ Grain size Prediction of Core E10-3

5.1 Introduction

This chapter will discuss the third and last research question that we stated in this study:

> **Is it feasible to predict textural properties of consolidated sediment satisfyingly, given multiple data sources and a known, straightforward provenance setting?**

In the previous section we discussed the forward model that incorporates linear mixing. This process of linear mixing is very common in sedimentary basins. Furthermore, as illustrated in the previous section, the inverse modeling of a mixed quantity is mathematically straightforward, although the introduction of the TF results in a small complication.

For the unconsolidated sediment it was found that the sediment of the Quaternary core, that was known to reflect a mixture of 2 different sediment sources, could be satisfyingly decomposed into three grain size endmembers. Apart from that, it was found that an acceptable geochemical reconstruction was obtained when 2 Transfer Functions were adopted.

However in order to apply this framework to a core of consolidated sediment, there are some problems that arise. First of all, GSD's cannot be retrieved and unmixed due to the simple fact that the grains in this core are physically attached to each other. In this core not the grains itself are the building blocks of the material, but the grains are all connected and turned into one structure; the *matrix*. Furthermore diagenetic processes might have altered the chemical composition, leading to a composition that does not 'obey' the rules of linear mixing. Therefore the grain size prediction for this core requires a different approach.

5.1.1 Research Question

There are effectively three research questions w.r.t the consolidated core. Whereas the in-situ geochemical logging technique is widely exploited in Quaternary geological studies, its performance with respect to the analysis of consolidated, ’hard rock’, cores is unknown. Given some differences w.r.t. the data acquisition, the first question that must be answered is:

> *Can XRF core scanning data of consolidated sediment be successfully calibrated?*

If the calibration of this data yields satisfying results, we get to the following questions. First we want to investigate how the XRF methodology, in combination with the accurate grain size data derived from the plugs, performs as an almost continuous grain size predictor. In other words:

> *Is it possible to fill-in the gaps in between the sparse plug observations in terms of mean grain size?*

Finally we want to investigate how the methodology can facilitate the semi-automatic classification of the core. Since this classification is labor intensive and is limited w.r.t. resolution, a way of combining the sources of data could work as a quality control and additionally result in improved classification results:

> *What is the added value of XRF core scanning data in the process of down core lithofacies classification?*

5.1.2 Methodology

In order to get an answer to the first question, a complete LRCE work flow is followed for the data of an consolidated core. If the data is calibrated successfully, we will investigate the other two questions. To find an answer to these two questions, the geochemical data must somehow be embedded in a model, either empirical or with a physical foundation. Any model which is applied may rely on input data as it is being acquired in a full core analysis work flow, which includes:

1. Spectral Gamma Ray

2. Core Description

3. Plug Analysis (grain size, among others)

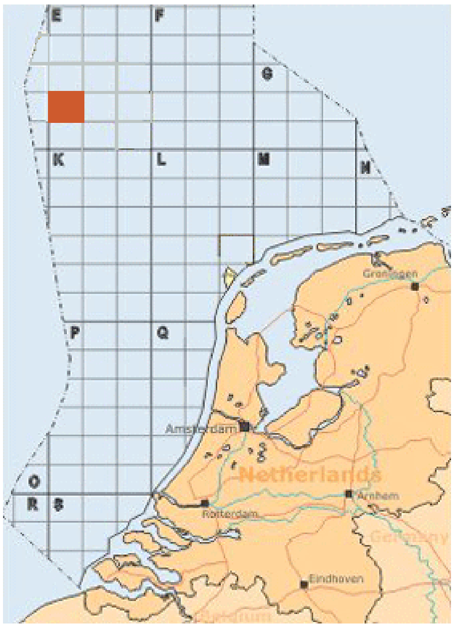

Figure 5.1: Block E-10 depicted in the upper left in red. Source: www.tullowoil.com.

As mentioned already in the introduction, the application of a physically feasible model such as the linear mixing model described in section 4 is complicated by the fact that we do not have fractionized geochemistry and Grain size Distribution to unmix accordingly.

However for this study, a dataset of a consolidated core is made available that consists of sediment that originates from one common source rock (pers. com. F. van den Belt, TNO). Because of this, the LINMIX model is drastically simplified and a direct approach to model grain size from geochemistry is justified.

5.2 Carboniferous core E10-3

In order to asses the methodologies as discussed in the previous sections, a consolidated core was used as a test case. This core is acquired in an offshore well that was drilled for hydrocarbon exploration purposes by Wintershall Noordzee BV. The depth of the cored section is around 4000 m below sea-level and originates from the Carboniferous. The well is situated on the North Sea (Dutch sector) in block E10-3, which is indicated in figure 5.1. A complete core analysis was performed by Panterra BV of which the findings were presented in the core analysis report by Boels (2003). However before going further into the core data and its purpose for geochemical and textural analysis, we will present a brief introduction about the core.

5.2.1 Geological Setting

In general, the Early Carboniferous (360-325 Ma) is characterized by the deposition of carbonates on shelves and isolated platforms while in intervening basins calciturbidites and mudstones were deposited (Freedman et al., 1990). The Namurian (326-313 Ma) is characterized by increasing siliciclastic influx from the rising Variscan Mountains in the south and the Fennoscandian Shield in the north (Ricken et al., 2000). In this way, Namurian sediments progressively in-filled and blanketed the submarine relief that had existed during the Early Carboniferous. Due to an ongoing regressive trend, Namurian basinal shales and turbidites were gradually replaced by delta-plain conditions at the end of the Namurian which marks the onset of the Westphalian (318-303 Ma). Most of the economically interesting coal seams formed during the Westphalian A and B in a lower delta plain environment (Drozdzewski, 2005). An increase in sand content and a shift towards dryer conditions, leading to a decrease in the number of coal seams, took place during the late Westphalian and Stephanian (Besly et al., 1993).

5.2.2 Core description

Panterra BV Netherlands has performed a study of core E10-3 w.r.t. the sedimentology, petrography and reservoir quality which is referred to as the *core description*. The results of this core description were reported in (Boels, 2003). Core E10-3 was characterized as reflecting a deltaic system of braided rivers with varying marine influence. In this core Panterra recognized a couple of different lithofacies associations: poorly drained floodplain, floodplain, swamp, crevasse splays, interdistributary bay and braided channel deposits. A detailed description of each lithofacies association is given in appendix (tag). Additionally they recognized a number of grain size classes, ranging from mudstone to very coarse sandstones, found in the braided river beds. An overview of the observed grain size classes in given in table 5.1.

5.2.3 Diagenesis

Panterra stated that the diagenetic elements that are encountered in the core can be subdivided into three groups;

1. Early Diagenesis;

2. Intermediate Diagenesis;

3. Late Diagenesis.

Class	Textural Classification	Mean Grain Size [μm] (ϕ-units)	Sorting
M	Mudstone	6 (7.4)	(1.6)
J	Siltstone	31 (5)	1.1
S1l	Very fine lower sandstone	74 (3.7)	1.5
S1u	Very fine upper sandstone	105 (3.2)	1.7
S2l	Fine lower sandstone	149 (2.75)	1.7
S2u	Fine upper sandstone	210 (2.25)	1.7
S3l	Medium lower sandstone	297 (1.75)	1.9
S3u	Medium upper sandstone	420 (1.25)	1.7
S4l	Coarse lower sandstone	595 (0.75)	2.0
S4u	Coarse upper sandstone	841 (0.25)	2.2
S5l	Very Coarse lower sandstone	1189 (0)	2.4

Table 5.1: Grain size classes and the textural properties as encountered in core E10-3.

The early phase of diagenesis was marked by the replacement of weak, unstable minerals by clays (kaolinite and illite). This replacement would have stimulated compaction in the early burial phase.

The following phase consists of the formation of hematite. In a study of the Upper Carboniferous of the Southern North Sea (Belt, van den, 1999) it was suggested that hematite formed during 2 phases: primarily, directly after deposition, and secondarily, during the penetration of oxidising meteoric waters as a result of Permian uplift. In core E10-3, the reddened braided channel sandstones are embedded in grey, poorly drained floodplain deposits. This suggests initial environmental conditions were reducing and therefore late hematite formation is possibly related to Permian uplift. Late diagenesis comprises the formation of ferroan dolomite, anhydrite, siderite, and barite.

5.2.4 Core selection

Of the complete core E10-3, a subset of core sections has been used in this study. The selection of these sections was based on the inclusion of most lithofacies and grain size classes identified in the core. Furthermore there are two sets, two meter in length, that are adjacent w.r.t. depth (core 1 box 13 and 14 and core 2 box 13 and 14 resp.).

The down core lithofacies classification (with the depth increasing towards the right) of the selected core is shown in figure 5.2. As depicted there, the core selection can roughly be subdivided into three parts. The first part consists primarily of the Crevasse Splay and Interdistributary Bay lithofacies. This simply means that this first regions reflects resp. proximal and distal deposits of this deltaic river system. The second section on the other hand reflects the coarse grained deposits of a braided river system. The third and last section reflects predominantly Crevasse Splay lithofacies as well as Swamps which partly consists of coal seems. Therefore we can state this third sections mainly reflect the

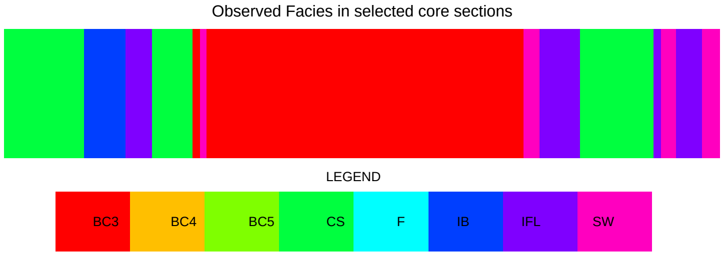

Figure 5.2: Lithofacies in the core sections applied in this study (BC = braided channel, CS = crevasse splay, F = Well-drained Floodplain, IB = inter-distributary bay, IFL = Poorly-drained Floodplain and SW = swamp.*

distal deposits of the deltaic system.

With respect to the dominant grain size in these three sections we would expect the finest sediment in the distal deposits found in the Poorly Drained Floodplain whereas the coarse to very coarse material will be found in the braided river sediment in the middle part. As far as the Swamp is concerned, since it is not dominated by clastic sediment, a classification in terms of grain size is difficult since the organic matter is not composed of grain (mineral assemblages) as such.

5.2.5 Data Acquisition

Core E10-3 has been analyzed using an XRF core scanner at the Dutch National Institute for Marine Research (NIOZ) in November 2009. The core has been scanned at two different voltages (10kV and 30kV) which yielded an extensive compositional record of the core (see table 5.2 for all elements). Whereas the Quaternary cores described in the previous part primarily consisted of fine-grained sediment mixed with water, these cores are completely dry. In addition to this, the core surface has been polished. Therefore the measurement could be obtained directly at the core instead of taking measurements with a thin plastic seal in between in order to protect the sensor.

In addition to the XRF data (intensities), the core scanner has the ability to obtain high resolution RGB-images of the cores. This ability was exploited which yielded RGB images (8 bits per channel) with a resolution of 10000 × 2048 pixels, which corresponds in the along core direction to around 100 pixels per centimeter core.

The data that is obtained with the core scanner reflects a proxy of the relative concentration of each element in the measured specimen (in this case a core section). This intensity is semi-quantitative in the sense that a high intensity reading of Fe as compared to a lower reading reflects a higher Fe content in the specimen. However the semi-quantitative performance of the core scanning data is only exploited after a series of processing steps is performed.

10 kV	{Al,Si,P,S,Cl,K,Ca,Ti,Cr,Mn,Fe,Co,Rh}
30 kV	{Zn,Ga,Br,Rb,Sr,Y,Zr,Au,Pb,Bi}

Table 5.2: Elements that have been quantified in Core E10-3 by XRF core scanning with two different voltages.

5.3 Data Processing

The work flow of the data processing consists of four parts:

1. Pre-calibration filtering of the intensity data,
2. Calibration sample selection and acquisition,
3. First-pass calibration,
4. Second calibration step without elements that have a mean concentration < 1 % or $R^2 < 0.5$,
5. Post-calibration evaluation of the calibrated data with the PCA filtering framework.

The first step is there to reduce the noise in the data induced by the measurement system. Subsequently the filtered data is being used to obtain a set of calibration sample locations that yield the most effective log ratio calibration process. When the data of these calibration samples is available, a first pass calibration is being performed to identify which components are effectively calibrated using the log ratio linear model. Furthermore this step is there to identify the mean (log ratio) concentration of each element.

Then a second calibration step is performed, however now by only applying it to a subset of the elements. After this final calibration run, the calibrated data is being evaluated, analogous to the initial filtering step however now based on a global noise parameter δN^2 derived from replicate analysis *of the calibration samples*.

5.3.1 Preprocessing

One of the pre-processing steps that is performed before calibrating the XRF data is the filtering using the PCA approach (see section (tag)). This is done by applying a Singular Value Decomposition (SVD) and back transform to the original basis using a subset of all the eigenvectors. This number of eigenvectors is determined by using the global noise parameter δN^2; the eigenvectors with the smallest eigenvalues which have a combined variance smaller than or equal to δN^2 are omitted in the back-transformation. The number of eigenvectors that is being preserved is being referred to as the ’intrinsic dimensionality’ of the data.

In section 4 , we described the results for the unconsolidated sediment core. It was found that the intrinsic dimensionality of the 7920 dataset was only 5. For this consolidated core, we expect a comparable or higher relative information content because of advantageous measurement conditions.

The filtering was performed by making use of a set of replicate measurements along core on both 10 kV and 30 kV. From these replicate samples the global noise parameter δN^2 was derived that determined the cumulative eigenvalue cutoff. This cutoff was applied to the clr transformed data decomposed into Principal Components. The filtered dataset in terms of compositions was then generated by back-transformation using a subset of the initial eigenvectors.

We found that a total number of 15 principal components is effectively above the global noise level δN^2. This result confirms the hypothesis stated in (Weltje and Tjallingii, 2008) that improved calibration results can be achieved with consolidated cores due to the simple fact that there is (1) less attenuation and scatter of the X-ray and (2) higher repeatability due to the lack of pore water.

5.3.2 Calibration sample selection and acquisition

As mentioned before, the calibration samples were selected based on the raw XRF core scanning data using the sample selection algorithm. This selection algorithm is based on the simple geometric principle that a least squares linear fit is more reliable if the observations are distributed along the whole range of observed values. Therefore the algorithm selected observations on the principal axes that are uniformly distributed along the range of observed scores. The number of samples per principal axis that are being selected is proportional to the variance of this component.

A summary of the results of the selected sample can be seen in figure 5.4. From these results we conclude that that the program has successfully selected locations that cover the entire range of values for almost all principal components. For some components this is not the case, which is induced by the fact that the algorithm does not select the minimum score and maximum score, but the 5% and 95% percentile, thereby reducing the chance of selecting outliers.

In figure 5.4 the observations as determined by the subroutine that selects locations using PCA are depicted as well as the locations that were selected to achieve a good down core distribution of samples. From this figure we derive that the spatial spreading routine nicely fills up the underexplored parts, where the red markers cover the entire span of the principal component scores. Furthermore there are no locations selected that are too close to each other which is achieved by selecting a minimum between-sample distance of 8 cm.

It was suggested by Weltje and Tjallingii (2008) that selecting a set of calibration

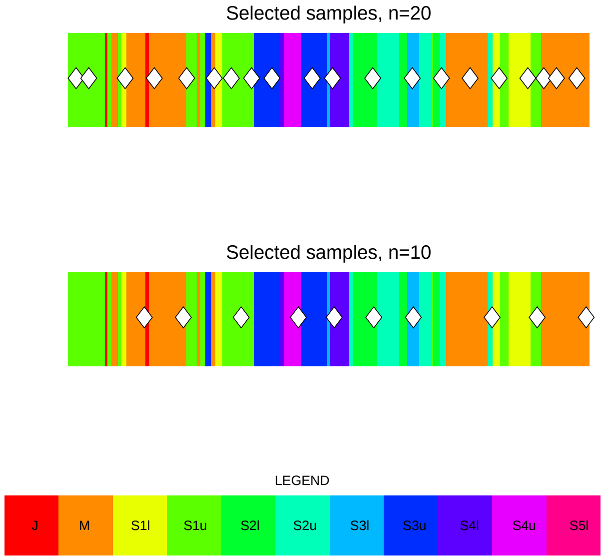

Figure 5.3: Selection of core sections and the recognized lithofacies.

samples that include all lithological units would give the best calibration result. In order to asses whether this is the case, we applied the set of grain size classes (see table 5.1), as identified by Panterra. When looking at table 5.3 it becomes clear this selection of lithological extremes is partly achieved. For 20 calibration samples 8 out of 10 grain size classes are explored. However since the number of selected samples is larger than the number of lithofacies, the sample selection was also performed for only a small set of samples (number of samples is equal to number of classes). For this selection 7 our of 10 grain size classes were explored. Based on these results we may conclude that the sample selection establishes the inclusion of lithological extremes which, as suggested in the work of Weltje and Tjallingii (2008), yields the best calibration results.

The calibration samples were acquired by simply using a hammer. In order to get an idea of the variability of replicate calibration samples, all 20 samples were physically divided into two parts and analyzed separately. The chips and/or fragments were measured w.r.t. their chemical composition using ICP which, in contrast to the XRF core scanner, yields semi-quantitative geochemical estimates (i.e. ppm or gr/kg).

5.3.3 Calibration of the XRF data

Since we now the set of calibration specimens analyzed w.r.t. their composition, we are able to perform a first-pass calibration step. The first pass calibration yields the quality as well as mean down core relative concentrations. This quality is quantified by the signal to noise parameter R^2.

Grainsize class	J	M	S1l	S1u	S2l	S2u	S3l	S3u	S4l	S4u	
$n = 10$	0	3	0	2	1	1	1	0	1	1	7/10
$n = 20$	0	6	3	5	1	1	1	2	1	0	8/10

Table 5.3: The observed lithofacies at the positions of the selected indices for resp. $n = 10$ *(n equals the number of unique classes) and* $n = 20$*, which is the amount of samples that was acquired from core E10-3.*

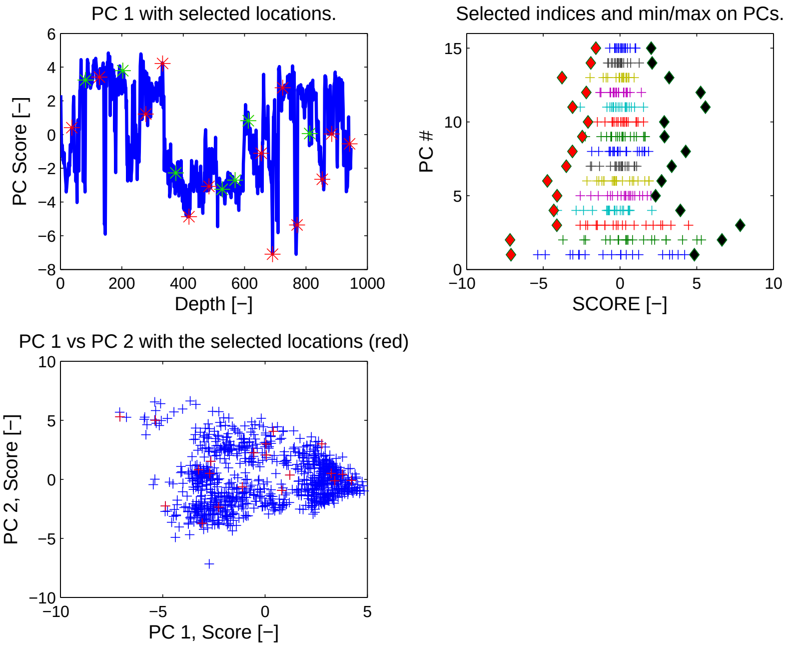

Figure 5.4: Summary of the decideposition algorithm with the 20 selected locations down core (upper left), in PCA1-PCA2 (lower left) and indicated on the score range of all PC's (upper right, with minimum observed score in red and maximum observed score in black).

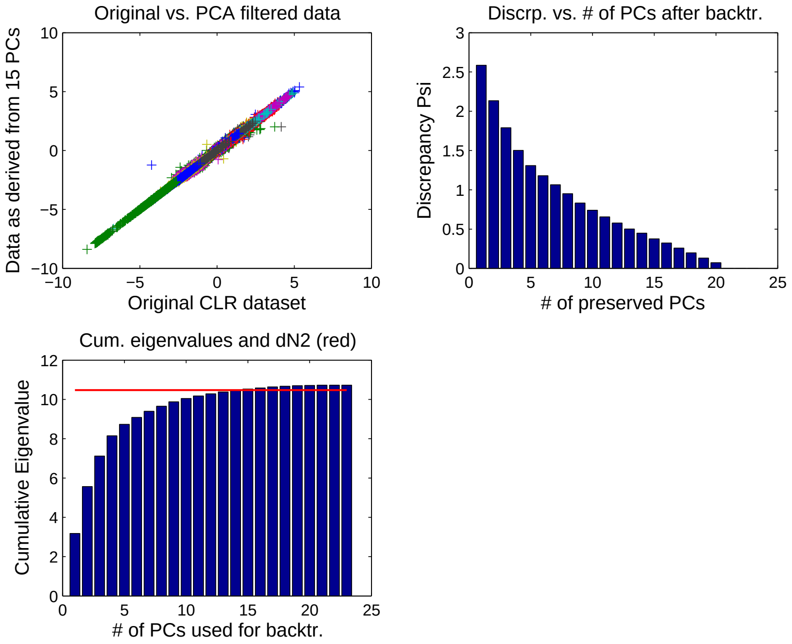

Figure 5.5: Summary of the princompfilt algorithm which yielded that the data could be satisfyingly reconstructed using a subset of 15 PC's.

If the R^2 value for a set of elements (numerator j and denominator element D) is found to be smaller than 0.5 this component is omitted in the second run of core scanning calibration. If $R^2_{jD} < 0.5$ this simply means that the magnitude of the residual variance is smaller than half the magnitude of the variance of the log ratio signal, which we found a reasonable cutoff. In addition to this criterion, the elements of which the maximum down core fraction is smaller than 1 % were also omitted in the secondary calibration process. We applied this cutoff because we are primarily interested in the major rock-forming elements and not in trace-elements. The final set of elements that were 'preserved' included Al, Fe, K, Si, Ti which represent an average fraction of the bulk of 98 % (given that the results from the initial calibration step is 'true').

In table 5.4 the R^2 values are shown for all log ratio combinations; rows represent the numerator and columns represent the denominator. It was found that Si as common denominator yielded the best results by means of the resulting R^2. The signal to noise values for the model with Si as common denominator yields for Al 0.96 and for K 0.92. Given that these three components already make up on average 94 % of the core, we selected this as the 'best' model.

The logratio-linear model shown in figure 5.7 shows that the concentration of the different components is always increasing w.r.t. the measured log ratio intensities. This makes the obtained results plausible in the sense that increasing log ratio intensities yield

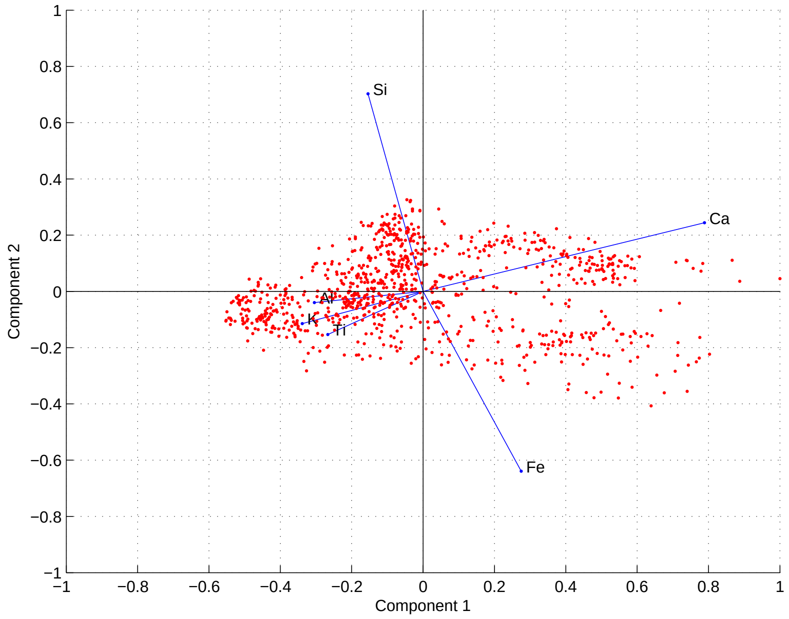

Figure 5.6: Biplot of the calibrated geochemical data of Core E10-3.

increasing log ratio concentrations. Furthermore the slope that was found is significantly large which means that the dynamic range of the log ratio intensities is of comparable magnitude compared to the dynamic range of the concentrations.

In figure 5.6, a biplot is shown of the clr-transformed elements that were selected according to the quality criteria. This biplot illustrates the covariance structure such that parallel variable 'arrows' indicate a strong correlation between the two variables. This means that there is one apparent signal in the dataset that comprises the clr-transformed concentration of Al, Ti and K. Calcium on the other hand is negatively correlated to the signal of this cluster. The other coherent signal represent the Si concentration which is negatively correlated with Fe. Furthermore we can conclude that the cluster comprising Fe and Si is not correlated to the Al/Ti/K cluster because of their 90 degrees angle between the variable indicators.

5.3.4 Post-calibration filtering

The third step of the data processing is the evaluation of the the compositional predictions by using the calibration sample quality. This quality was derived from the replicate sets of calibration samples. These replicate calibration samples provide a rough estimate of the residual variance induced by sampling and the chemical analysis techniques that are applied. Whereas the 'initial' signal to noise parameter R^2 was adopted to provide a signal to noise parameter of the XRF data itself, we can in principle derive another signal to

	Al	Ca	Fe	K	Si	Ti
Al	0	0.7346	0.8645	0.3372	0.9610	0.6336
Ca	0.7346	0	0.3811	0.7897	0.7212	0.6745
Fe	0.8645	0.3811	0	0.8855	0.8735	0.9311
K	0.3372	0.7897	0.8855	0	0.9752	0.8519
Si	0.9610	0.7212	0.8735	0.9752	0	0.9739
Ti	0.6336	0.6745	0.9311	0.8519	0.9739	0

Table 5.4: Values of the signal to noise parameter R^2 *found in the second-pass calibration step of the E10-3 data.*

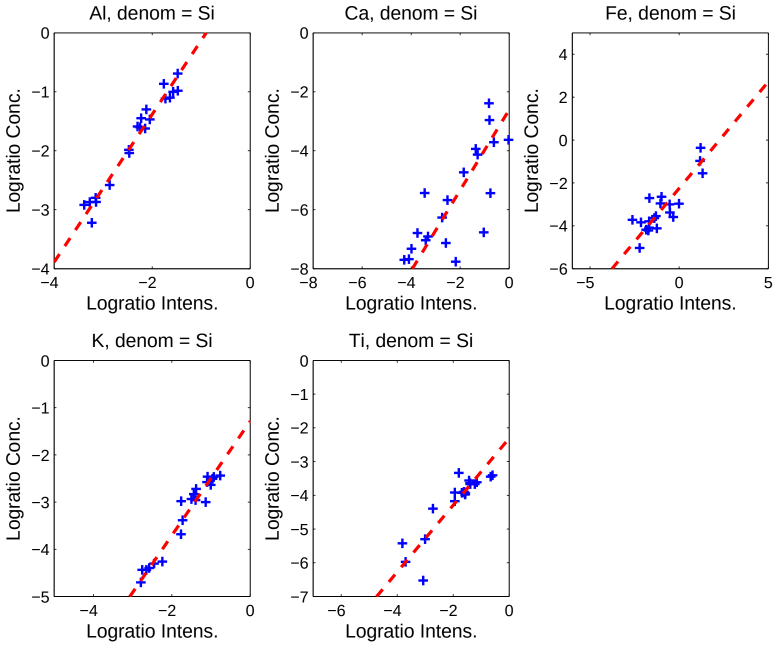

Figure 5.7: The log ratio linear models as obtained in the second-pass calibration step, using the model with Si as common denominator.

noise parameter that is related to the calibration process. Even though the intrinsic dimensionality of the XRF data itself is significantly higher for this core due to the favorable measurement conditions, the sampling conditions are definitely less favorable for these 'hard rock' cores. There are two main problems that arise when acquiring calibration samples for consolidated sediment.

The first problem mainly has to do with down core positioning. The acquisition of a sample on a soft sediment core can be performed rather accurate with reference to the down core position; a small (diameter) sampling cylinder is pressed into the soil and retracted from the sediment with the sediment content inside. Replicate samples are obtained by performing the same procedure on a location at the same stratigraphic depth such that is reflects more or less the same material. However sample acquisition as it is done in this study has a much lower down core positioning accuracy.

Another difficulty with this hard rock arises from the nature of the rock. The samples obtained from this consolidated are, in contrast to the soft-sediment cores, a collection of rock fragments with a rather variable size and shape. Due to this variable size and shape, replicate samples reflect different intra-sample variability. In order to implement the variability of the replicate samples into the calibration process, the methodology of the pre-processing PCA-filtering step is applied. This means that we derived an estimate of the intrinsic variability of the calibration samples by deriving a median clr-distance between replicate sets. However whereas the prior framework was related to the replicate variance of the intensities I in the form of δI^2, we now replace the I for an W that reflects the compositional distance within a replicate set of m calibration specimen replicates:

$$\delta W^2 = \mathrm{med}^m \left\{ \sum_{i=1}^{D} [\mathrm{clr}(W_i) - \langle \mathrm{clr}(W_i) \rangle]^2 \right\} \tag{5.1}$$

a global intrinsic noise parameter is obtained by taking the median value over all replicate sets k:

$$\delta N_w^2 = \mathrm{med}_k \left\{ \delta W_k^2 \right\} \tag{5.2}$$

This framework yielded that in principle only 2 principal components can be considered to be above the global noise level, given the variable subset of Al, Ca, Fe, K, Si, Ti. In other words; two principal components can effectively describe the information content of the data. This means that the 2-dimensional biplot shown in figure 5.6 in principle shows the total information content of the data that can be considered to be above the noise level.

5.4 Grain Size Modeling

Core sections are usually analyzed using a large set of measurement methods. In this study we will focus on the XRF-bulk and the images which will be the primary grain size *proxies*. The grain size data itself consists of (1) a down core classification in terms of grain size classes and (2) the more detailed analysis of core plugs.

Core Description

The core description represents the data that was obtained by visual inspection of the core. The grain size component of the core description includes a continuous classification record of the core in terms of grain size classes. Although on appearance the resolution of the classification reported in the core analysis report has a high resolution (red. it is a continuous record), the intrinsic resolution is low. This is the result of the fact that within a core section that is assigned a certain class, the 'true' classification will most probably yield a mixture of different classes. This implies that the core description has, although not quantified as such, a limited resolution which can be translated into a *minimum down core class-size*. In other words; if there is a small contrasting bed of 1 cm thickness, this will not be classified as such because of its small 'size'. Nevertheless in some cases core sections are classified as mixtures, however any information regarding the mixing coefficients is lacking. The unquantifiable resolution and the lack of mixing coefficients in case of a mixed class, implies that there is a level of subjectivity involved in this data.

Images

Besides the data that has been acquired as part of the core analysis work flow, images have been obtained as part of the XRF core scanning. These images hold valuable information in terms of color and texture; it is not hard to imagine that the color and texture of the rock holds information concerning the lithotype as well as grain size.

The resolution of the images is high compared to consumer camera's (i.e. 100 pixels per centimeter in the down core direction). However the images are not suitable for the detection of individual grains. Given the fact that 100 μm corresponds to 3 to 2 phi-units, which is classified in the Wentworth classification as a fine sand, the possibilities to identify individual grains are limited to the coarse part of the grain size spectrum.

Although the application of the textural properties is limited due to the resolution, the colors on the other hand contain information w.r.t. the lithologies. Given the nature of RGB data, we will apply some adjustments to the RGB data for processing convenience. One of these adjustments is the fact that the color, in terms of 8 bit RGB values, is only partially a compositional quantity. The observation in terms of each of these channels range (for 8 bit values) between 0 and 255 where 255 means that a pixel is fully saturated

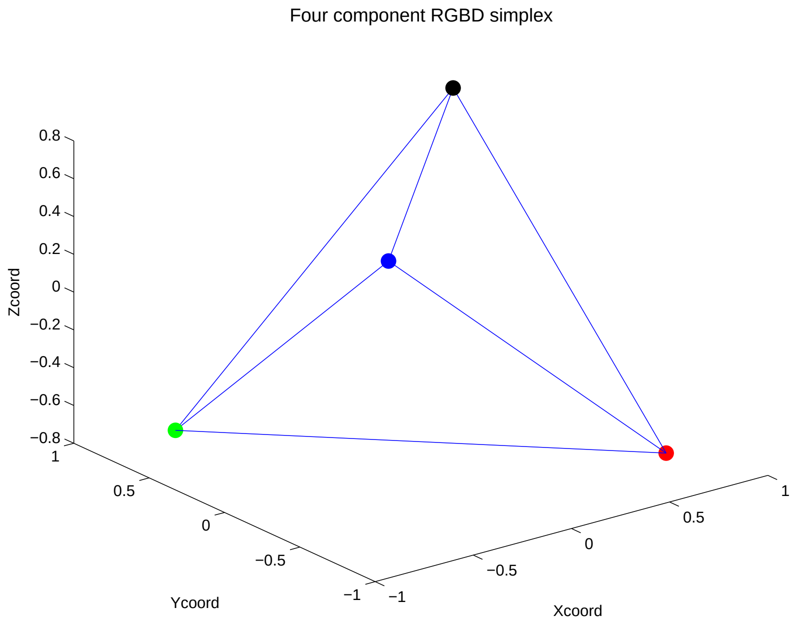

Figure 5.8: The RGBD solution space spanning a 4-component simplex.

with either red, green or blue (red. a white pixel). Although the fraction of each channel (RGB) represents a compositional quantity, the sum of the value of each channel does not necessarily have to sum up to an arbitrary but fixed value. In other words, there is an additional component that determines the ’darkness’ which follows directly from the sum of the RGB-values. This darkness quantity, denoted by X_4 follows from the RGB-values:

$$X_4 = 768 - \sum_{i=1}^{3} X_i \tag{5.3}$$

for an 8 bit system. The value of 768 is the maximum possible value of the summed RGB-values; $3 \times 2^8 = 768$. However the contrast in X_4 is optimal if the RGB-data have been stretched such that they range between 0-255. Otherwise the darkness-parameter X_4 will dominate the RGBD composition.

By performing this transformation the pixel values $\{X_1, X_2, X_3, X_4\}$ become a compositional quantity in the sense that they sum up to 768 for the 8 bit case. Therefore the solution space now spans a 4-component simplex (see figure 5.8 and 5.9), or $\mathcal{S}^4$. This also means that by performing a logratio-transformation, the RGB data is (1) in an unconstrained space and (2) in the same units as the geochemical data (clr-transformed composition).

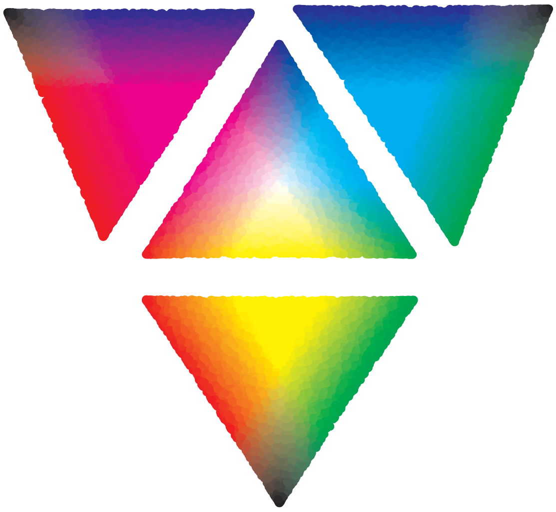

Figure 5.9: RGBD color scheme.

RGBD of Core E10-3

After the primary processing step of transforming the RGB to RGBD, the data is being turned into a dataset equally sampled as the XRF core scanning data. Since the logratio transformed data now comprises an unconstrained space, we can simply take the empirical mean of the clr-transformed RGBD-data within the measured down core locations. These measured locations are modeled as 1×1 cm building blocks which is reasonable to assume, given the sensor's footprint. Based on the instrumental setup, it is reasonable to assume a uniform effectiveness of the X-ray source and receiver on the 1×1 cm blocks. The down core strips of 1 cm in width are shown in figure 5.10. The down core RGBD data on the other hand is shown in figure 5.8. In this figure there are a couple of different processes we observe in the RGBD-composition.

First of all there is a strong increase in red for the first 100 observations, which corresponds to the first core, core 1 Box 13. This is not surprising since the corresponding core section (core 1; box 13) contains oxidized sands/silts which yield a relative increase in redness. The other clear events occur between 700-800 and 850-950 which correspond to core 2 box 18 and core 2 box 33, respectively. Also for these cores it is not surprising that there is a tendency of increasing 'darkness'; the coal seems in these cores have a very low reflectance w.r.t. all colors.

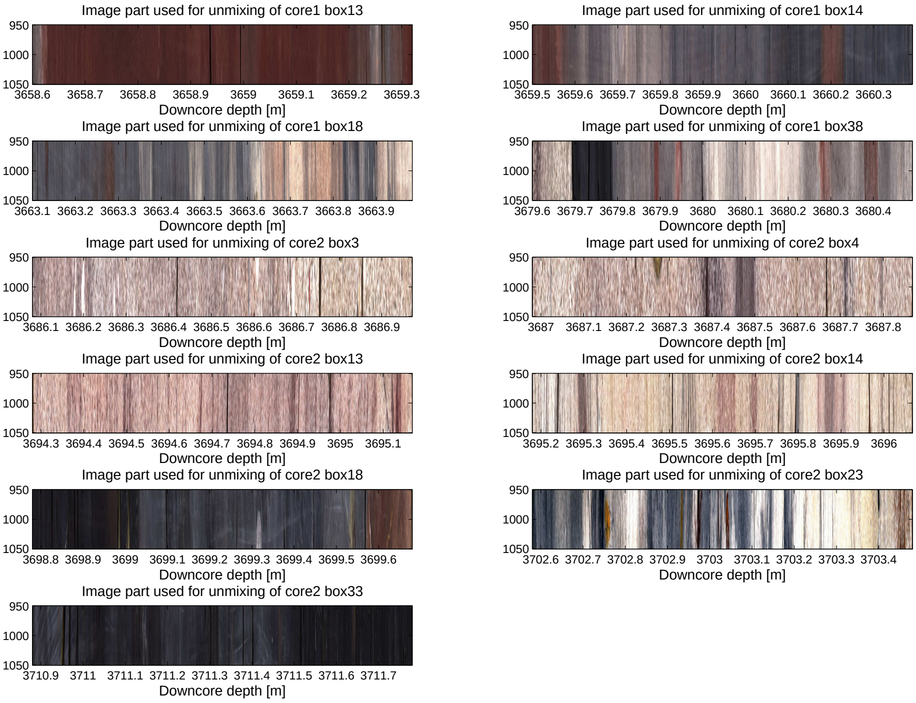

Figure 5.10: Cropped core sections, 1 cm in the cross-core direction.

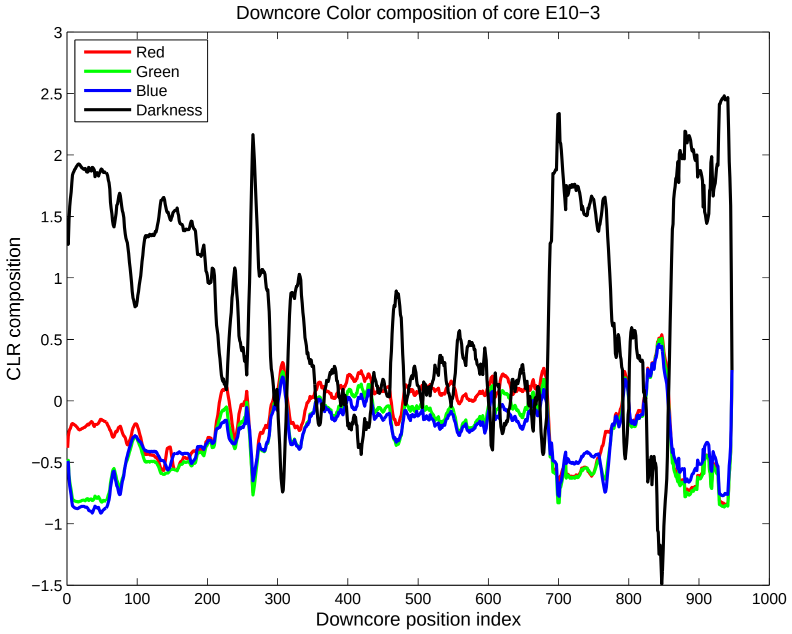

Figure 5.11: Down core record of the logratio RGBD-data.

RGBD Information Content

Given that we mitigated the semi-compositional nature of the RGB-data, the RGBD-values can now be assessed w.r.t. their covariance structure. In figure 5.11 we observed some level of correlation between green and blue for instance. A proper way of evaluating the information content of a set of correlated variables is to perform PCA, thereby reducing the dimensionality of the data. The loadings of the variables on the first two eigenvectors obtained with PCA are shown in figure 5.12. Given that the first two principal components represent 99% of the variance, it is reasonable to say that this 2-D plot summarizes the information content of the RGBD data.

In the biplot can be derived that the first principal component is dominated by the darkness variable. The second component on the other hand is a measure of the amount of blue, relative to the amount of red. Given the nature of the material, we conclude that this is in line with the expectations. The dominant signal (red. the first principal component) is merely a grain size signal; highly reflective sand or low reflective shale. The second principal component on the other hand reflects the red shales vs the blue shales.

Given this RGBD record, a straightforward classification based on the RGBD data is able to discriminate reasonably well between red shale, 'blue' shale and coal layers. Additionally the 'darkness' value is expected to correlate fairly well with the grain size.

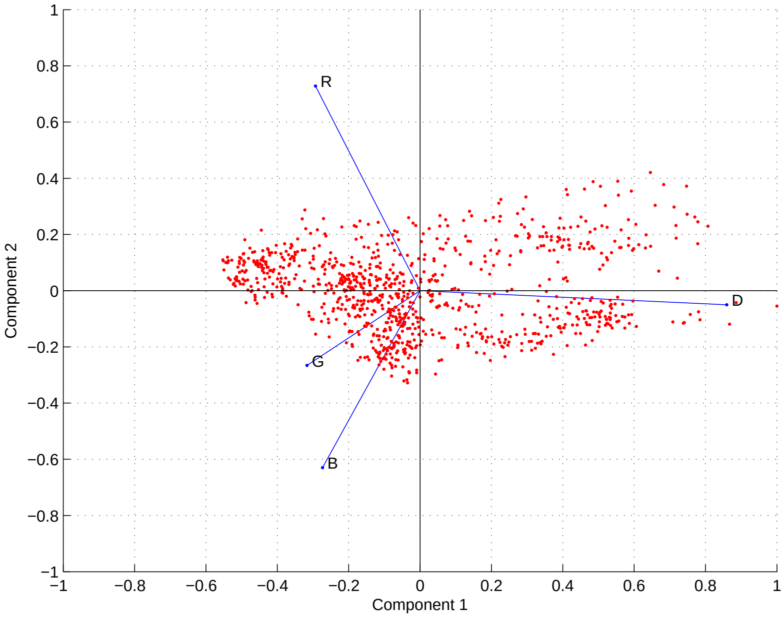

Figure 5.12: Biplot of the RGBD data of core E10-3.

XRF core scanning data

The XRF core scanning data has been calibrated using the LRCE. Analogous as was found for the measured intensities, the calibrated data also represent a limited dimensionality. The resulting concentrations for the subset of elements is shown in figure 5.13.

However in order to evaluate the performance of the geochemistry as a grain size predictor, we should determine the correlation coefficients *between* geochemistry and grain size. A straightforward way to asses this is to obtain the correlation between grain size and geochemistry in terms of additive logratio pairs with elements i and j where $i = 1,2,..D$, $j = 1,2,...D$ and $i \neq j$. The results of this are shown in figure 5.5.

Here we observe three element pairs showing a strong intrinsic grain size signal which includes Al/Si, Ti/Si and K/Si. The magnitude of these logratio's are negatively correlated with the grain size which is in line with what we would expect; an increase in ϕ-units is analogue to a decrease in grain diameter. In general we expect an enrichment of Al and K in the fine sediment whereas Si and Ti are concentrated in the coarse fraction. This is the result of the relatively high resistance of the minerals in which Si and Ti primarily occur. On the other hand the fine fraction is expected to reflect high Al/Si ratios as a result of the relative low resistance against chemical weathering of mineral phases containing Al (i.e. feldspar).

Although the Al/Si ratio is found to partly reflect a grain size signal, other logratio pairs also show considerable correlation with the grain size. Additionally the RGBD-data

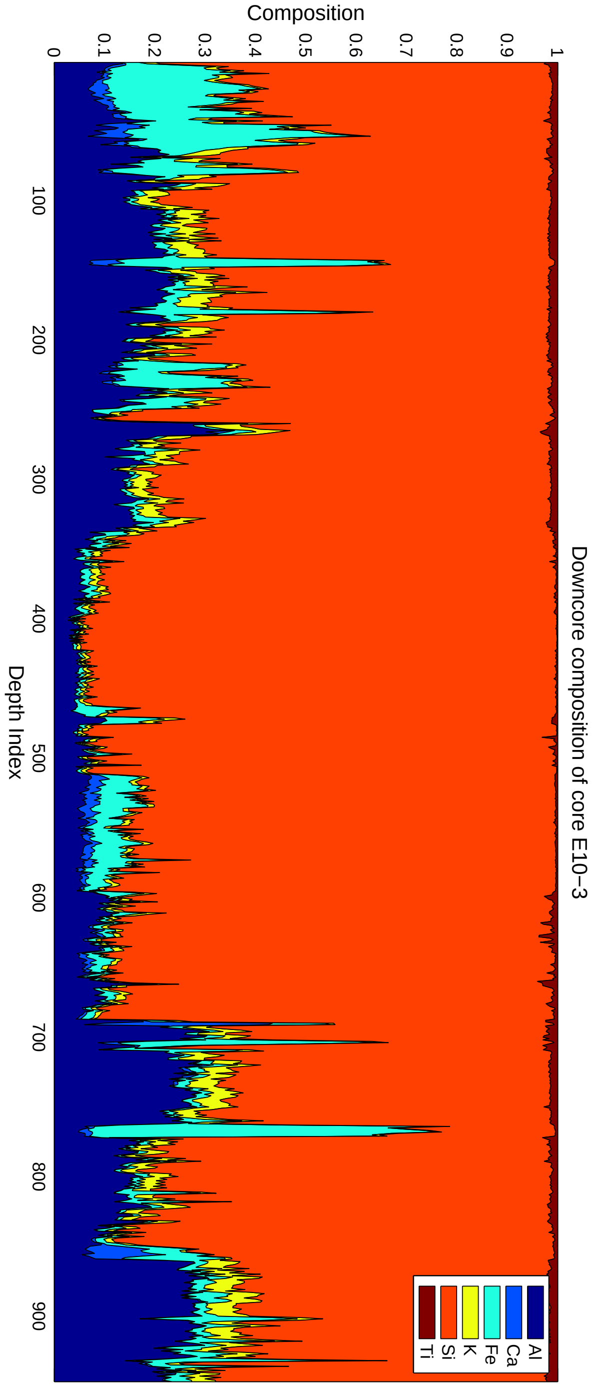

Figure 5.13: Down core geochemical record of the scanned core E10-3 sections.

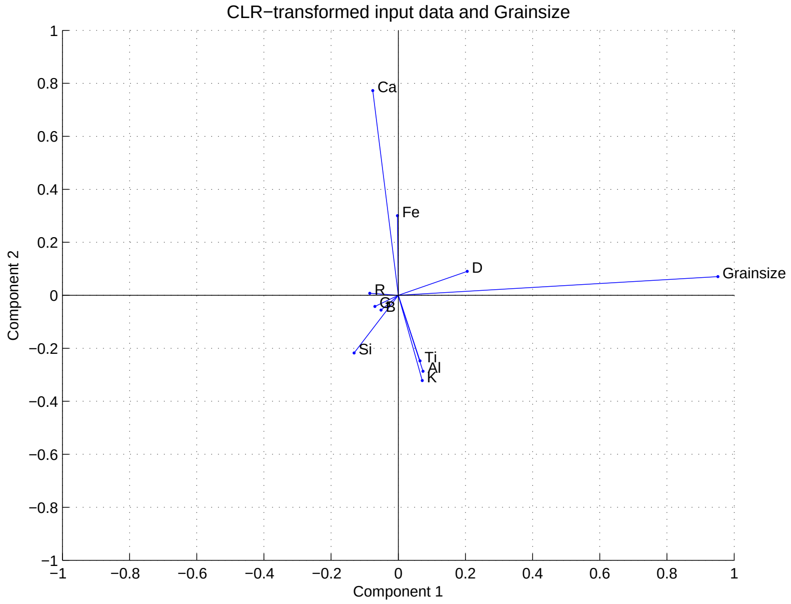

Figure 5.14: Biplot of the logratio calibrated XRF, RGBD and grain size data.

was found to contain lithofacies information by (1) the ability to discriminate between redox-environments and (2) the ability to discriminate between sand and shale. The co-variance structure of all input quantities (RGBD and XRF-bulk) is presented in figure 5.14 in a biplot. From this figure can be derived that the only parameter correlating positively with the grain size is the darkness. As far as D is concerned, this is in line with what we expect, given the relative 'darkness' of the shales. Also the negative correlation of Si with the grain diameter is analogous to what we would expect. These results imply that the principal components can be given a geological/sedimentological context. Whereas the first principal components is a grain size signal, the second principal components reflects the *redox-environment*; high score is an oxidising environment whereas a low score indicates a reducing environment. This interpretation is based on the loading of red on the second principal components and its negative correlation with blue.

Summarising we conclude that we have many indicators which are valuable as a grain size proxy. Although linear regression with Al/Si for instance will yield proper grain size predictions, we expect to get the best results if the grain size is predicted while using all variables, geochemistry and RGBD, simultaneously. Given the fact that additional data sources (red. grain size proxies) can be applied in order to support the XRF-based prediction, there are endless possibilities w.r.t the model architecture and complexity. A widely used technique to apply when multiple predictors are available for the estimation of a single variable is Multi Variate Regression, or MVR.

	Al	Ca	Fe	K	Si	Ti
Al	-	0.2949	0.2896	0.0590	0.8638	0.0554
Ca	-02949	-	-0.1698	-0.2781	0.0682	-0.2891
Fe	-0.2896	0.1698	-	-0.2793	0.2633	-0.2753
K	-0.0590	0.2781	0.2793	-	0.8221	0.0208
Si	-0.8638	-0.0682	-0.2633	-0.8221	-	-0.7252
Ti	-0.0554	0.2891	0.2753	-0.0208	0.7252	-

Table 5.5: Correlation coefficient between the different alr-transformed element pairs and the grain size record. Columns represent the denominator, rows the numerator.

5.4.1 Multi Variate Regression (MVR)

Multi-Variate Regression is probably the most straightforward model to adopt when there are multiple predictors for one variable. In this case the predictors may include any of the available data; XRF, RGBD, spectral gamma ray, etc. The multi variate regression model then follows for n predictors x by finding a least squares solution w.r.t. y as follows:

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... a_n x_n + \varepsilon \tag{5.4}$$

However whereas the model shown above is a first order polynomial, also a model with the variables in an higher order form x^n comprise a linear model. However in this case we will use the first-order multi-variate model.

The reason of choosing this first-order model is that, given the fact that the compositional change with grain size (i.e. the TF) was successfully modeled as a compositional linear function, a first-order regression approach to predict grain size from composition makes sense. Nevertheless this methodology relies on an intrinsic assumption about the grain size distribution. Since it maps the composition directly on the grain size, there is an intrinsic assumption about the grain size distribution; the grain size distribution is such that the integral of the product between GSD and TF equals the composition at the mean grain size μ.

$$\begin{bmatrix} c_{1,\mu} \\ c_{2,\mu} \\ \vdots \\ c_{D,\mu} \end{bmatrix} = \int_{-\infty}^{+\infty} c_{ij}(x) f_i(x) dx \tag{5.5}$$

Or, more general, the grain size distribution is infinitely small in the grain size spectrum, resulting in the composition as a function of grain size following the intrinsic Transfer Function. However given that we are not able to estimate the GSD's from the data, we think this assumption is reasonable. The multi-variable linear system given above can be solved in a straightforward way using least squares methodologies or any other optimisation strategies. Here the Expectation Maximisation (EM) algorithm has been used, however the principles of this optimisation strategy fall outside the scope of this study and are described in the work of Meng (1993).

The input data of the MVR includes both the calibrated XRF as well as the RGBD. However both quantities are alr-transformed, in order to utilize them in their unconstrained form. For the XRF-data we apply Si as common denominator which was found to yield logratio variables with the most pronounced correlation with grain size, given the high level of correlation of the Si/Al ratio. Apart from that Si has an overall high concentration of and it showed a high performance in the XRF calibration step. Furthermore we apply the alr-transformation to the RGBD-data by using D as common denominator. This decision is based on the overall high level of correlation between lithology and darkness. Furthermore the fact that the darkness is the dominant signal in the RGBD data made us decide to use it as such.

5.4.2 Classification

Besides the quantitative approach, we also applied a categorical prediction. The categories or classes with which we wish to characterize the core are the grain size classes that were identified in the core report (Boels, 2003). Although any other categorical dataset can be applied, the success of the classification is obviously determined by the geochemical variation between the different classes. The primary reasons for performing the classification is to get a quantitative classification for core sections that comprise a mixture of different classes. However also it can also provide a way of quality control on the final classification results. A well-established technique to cope with classification as well as *mixtures* of classes is based on the Bayesian theorem.

Classification

Given that we have a partitioning of the core in terms of classes, a classification strategy is defined with the (clr-transformed) XRF and RGBD data as input. This classification scheme is based on the Bayesian theorem:

$$P(C_i|X) = P(X|C_i) \cdot \frac{P(C_i)}{P(X)} \tag{5.6}$$

This theorem implies, in words, that the probability of the observation X to be of class i is determined by the likelihood of X (red. the marginal distribution of X given a class C_i) and the probability of encountering class i at all (the *prior* probability) divided by the marginal distribution of X. The practical application of this Bayesian framework to classification is performed by fitting Probability Density Functions (PDF's) to the data X in the different classes, thereby obtaining a functional form of the likelihood.

The input is obtained by supplying the observations with the designated class, which represent the 'training data'. From this training data the marginal distributions of the variables in X, given they are member of class i, are derived. Subsequently a global class-probability, or *prior* probability $P(C_i)$, is empirically derived. Together with the marginal distributions of X within the solution space, that also follow from this training data set, this can be readily plugged into the Bayesian framework.

After this training step, an observation X that is situated somewhere in the solution space can now be assigned a *posterior* probability $P(C_i|X)$ for a class i. However, more general, this probability can be interpreted as *mixing proportions* of the different classes. Using this mixing proportions, the forward linear mixing model for a given quantity y for observation k, with the values of this quantity for the classes $i = 1,2,...C$ denoted by y_1,y_2,$...y_C$:

$$y_k = \sum_{i=1}^{C} y_i \cdot P_k(C_i|X) \tag{5.7}$$

An illustration of this classification technique is shown in figure 5.15. The two figures show the observations within X that are known to reflect class 1 and 2 , respectively. A PDF is fitted to this subset of the data which yields a functional form of the probabilities within the solution space of X. By plugging the result of this function for an arbitrary outcome of X, together with the prior probabilities of C_i and X into the Bayesian framework, the posterior probabilities for all classes are obtained.

The improvement of using this approach versus the classification as determined in the core description is that now there are mixing proportions of the different grain size classes for every observation. Therefore this approach facilitates a quantitative approach to a categorical data type which, in contrast to the currently used approach, also yields relative mixing proportions. If necessary, it can still be transformed into a 'hard' classifier by selecting the class for an observation k for which the posterior probability is maximal.

5.4.3 Results

The two methods (MVR and classification) have been applied to the data of core E10-3. Both methods yield a down core grain size record that complements the currently available grain size record, as obtained from the classification. However whereas the MVR is 'trained' with the plug analysis, the classification is based on the continuous classification record.

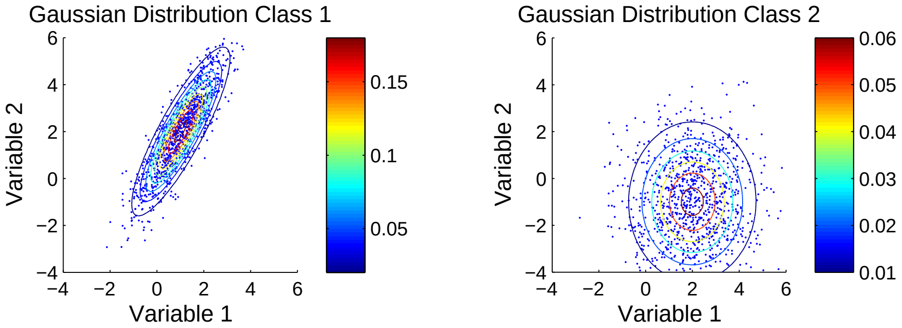

Figure 5.15: Graphical representation of the methodology of fitting a multivariate Gaussian distribution to the data X for two classes.

MVR results

The results of the MVR for core 1 and core 2 are shown in figure 5.16 and 5.17, respectively. In the figures are the core images, grain size derived from the core description (continuous green curve) and the plug analysis depicted (green markers).

What can be observed in the MVR predictions is that the *overall* prediction is unbiased within acceptable limits (see figure 5.18). Nevertheless there are some core sections that show a biased prediction w.r.t. the core description-derived grain size. However this bias is the direct consequence of the discrepancies between the two data sources as a result of ambiguous classification; observations that are classified as a mixture however without the mixing coefficients. Overall though it is reasonable to assume that the plug analysis results are closest to the 'true' grain size.

Furthermore the MVR results show a significantly higher resolution; it detects small scale variability w.r.t the grain size such as 3702.7 m. Additionally the sudden change in lithotype at depth 3699.6 is well reconstructed, despite the absence of any magnitude control on the grain size by neighboring plug observations.

With respect to the prediction discrepancies a standard deviation of the residuals is found of 1.0 ϕ-unit (see figure 5.18). The consequence of this value is that, given a Gaussian distribution of the residuals, is that 66 % of the observations will be not more than 1 ϕ-unit discrepant with the 'real' grain size. However this is only true under the assumption that the grain size derived from the plugs is 'true'.

Classification Results

The results of the classification are shown in figures 5.20 5.21 and 5.22 which shows the original classification as well as the grain size prediction, based on the posterior probabilities and the grain size of the corresponding class.

The predicted grain size record in figures 5.20,5.21 and 5.22, shows an overall good

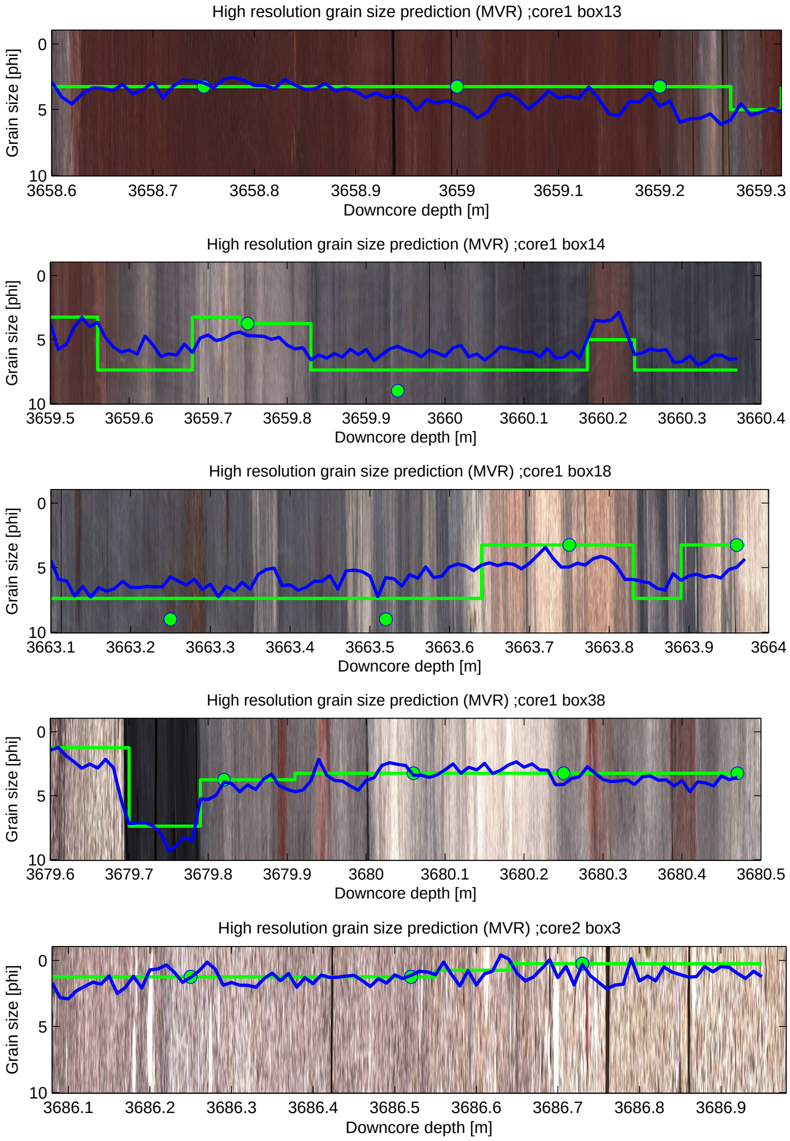

Figure 5.16: Core sections $\{(1,13),(1,14),(1,18),(1,38),(2,3)\}$ *(core, box respectively) of E10-3 together with the grain size, as obtained from plug analysis (green markers) and the MVR predictions (blue).*

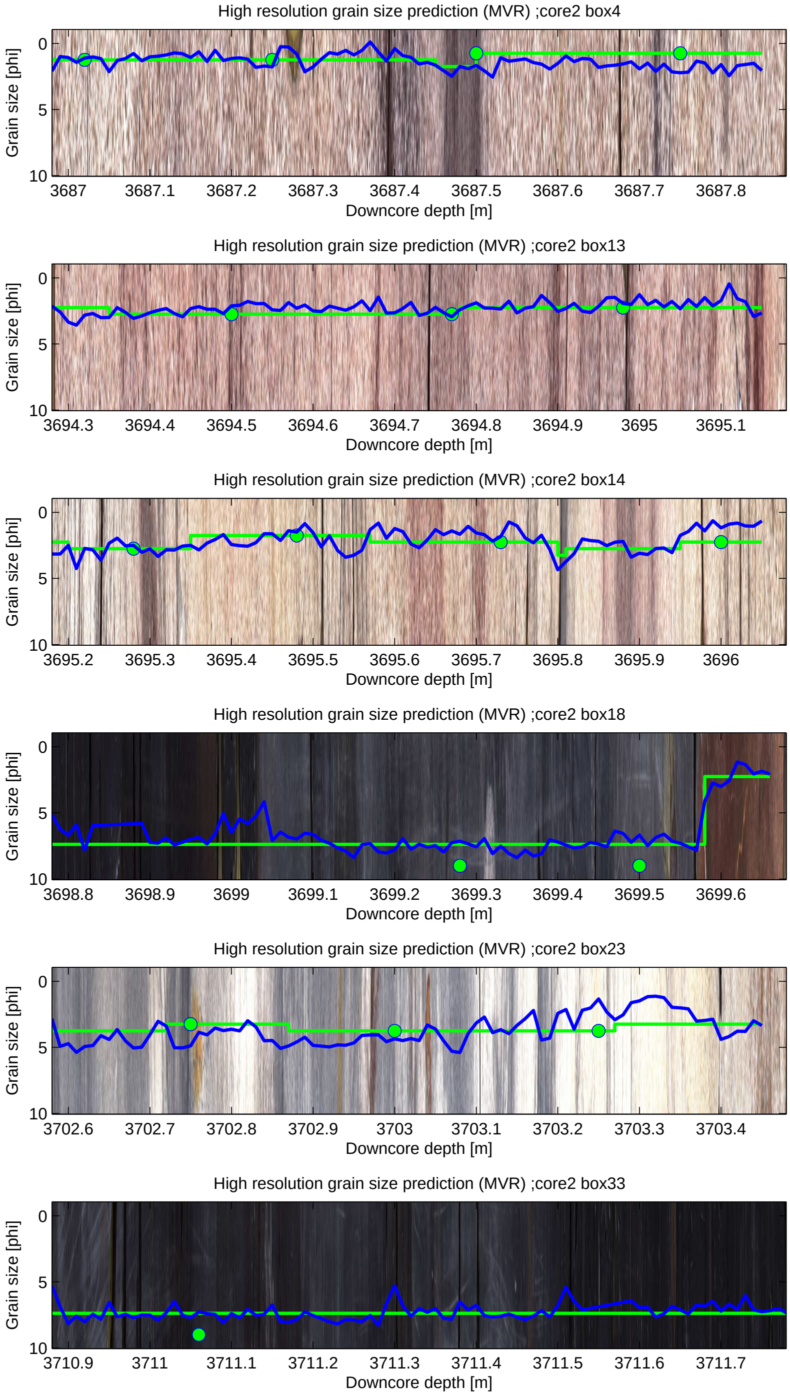

Figure 5.17: Core sections $\{(2,4),(2,13),(2,14),(2,18),(2,33)\}$ *(core, box respectively) of E10-3 together with the grain size, as obtained from plug analysis (green markers) and the MVR predictions (blue).*

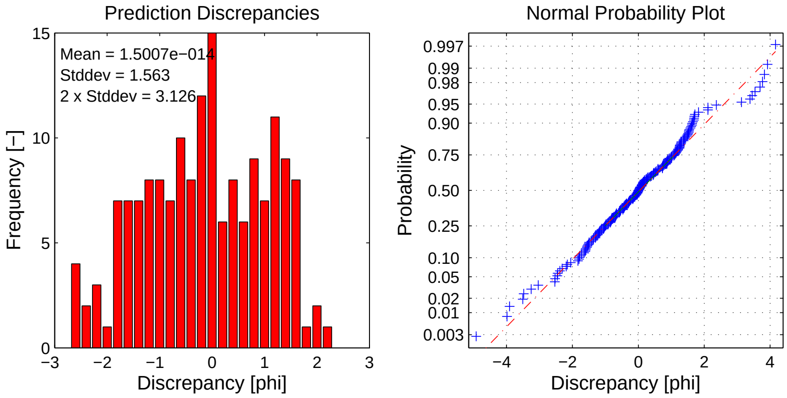

Figure 5.18: Histogram of the prediction discrepancies of the MVR and a normal probability plot.

match with both the plug analysis grain size and the core description results. Furthermore the overall grain size prediction obtained with the Bayesian classification is unbiased; figure 5.19 displays the residuals between plug data and predicted grain size. However the predicted grain size record in core section 18 (core 1) seems to be deviating from the core description results. There are some bands in the core which are classified by the Bayesian scheme as small bands of coarse material in a matrix of fine-grained shale. Also the transitions at 3663.6 and 3663.9 m downcore depth are (1) classified as smooth transitions and (2) sometimes repositioned in the down core direction with respect to the core description.

Another core section for which the classification result is significantly different from the core description is core section 14 (core 1). Here a small section of 10 cm is observed for which the predicted grain size is relatively coarse compared to the core description. The same result, but less pronounced, was obtained with the MVR approach; the small band with light-colored material is classified as a relative coarse-grained rock.

The predicted grain size in the other core sections is in general not deviating much from the initial grain size. However this finding is rather trivial since the unsupervised classification has been *trained* with the input data. Therefore one should keep in mind that the XRF-guided classification is in the absolute sense not closer to the reality than the input data, however it indicates regions where the prediction based on XRF and RGBD deviates from the initial grain size record.

5.5 Discussion and Conclusions

Question 1:

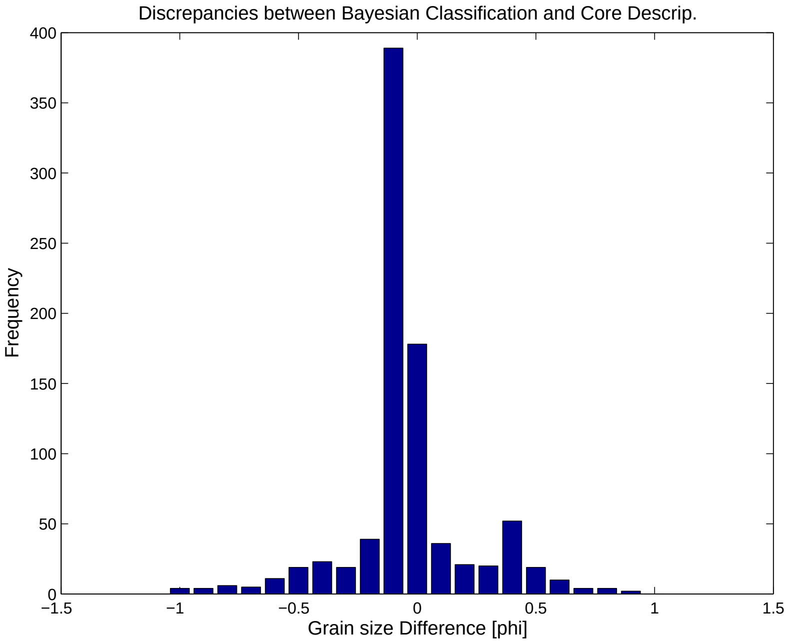

Figure 5.19: Histogram of the prediction discrepancies of the Bayesian Classification.

> *Can XRF core scanning data of consolidated sediment be successfully calibrated?*

XRF core scanning has shown to be an effective geochemical logging device for consolidated cores. When applying the PCA-filtering to data of core E10-3, 15 principal components (of the 23 elements measured) were found to be above the noise level. From the calibration of the core itself we conclude that the XRF core scanning data of core E10-3 has been successfully calibrated. A second-pass calibration step whereby only preserving the elements for which the signal to noise ratio is higher than 4 and the down core concentration is higher than 1%, returned the elements Al, Ca, Fe, K, Si, Ti. Nevertheless the calibration samples in this study have a significantly lower quality than the intensity data, leading to an ultimate calibrated data dimensionality above the noise level of 2. This is the result of the rather inaccurate sample acquisition.

Question 2:

> *Is it possible to fill-in the gaps in between the sparse plug observations in terms of mean grain size?*

The XRF core scanning data and the RGB data have been embedded in the grain size prediction work flow. The quantitative and direct approach of the grain size prediction by using the plug data yields a high resolution grain size record with a minimal bias.

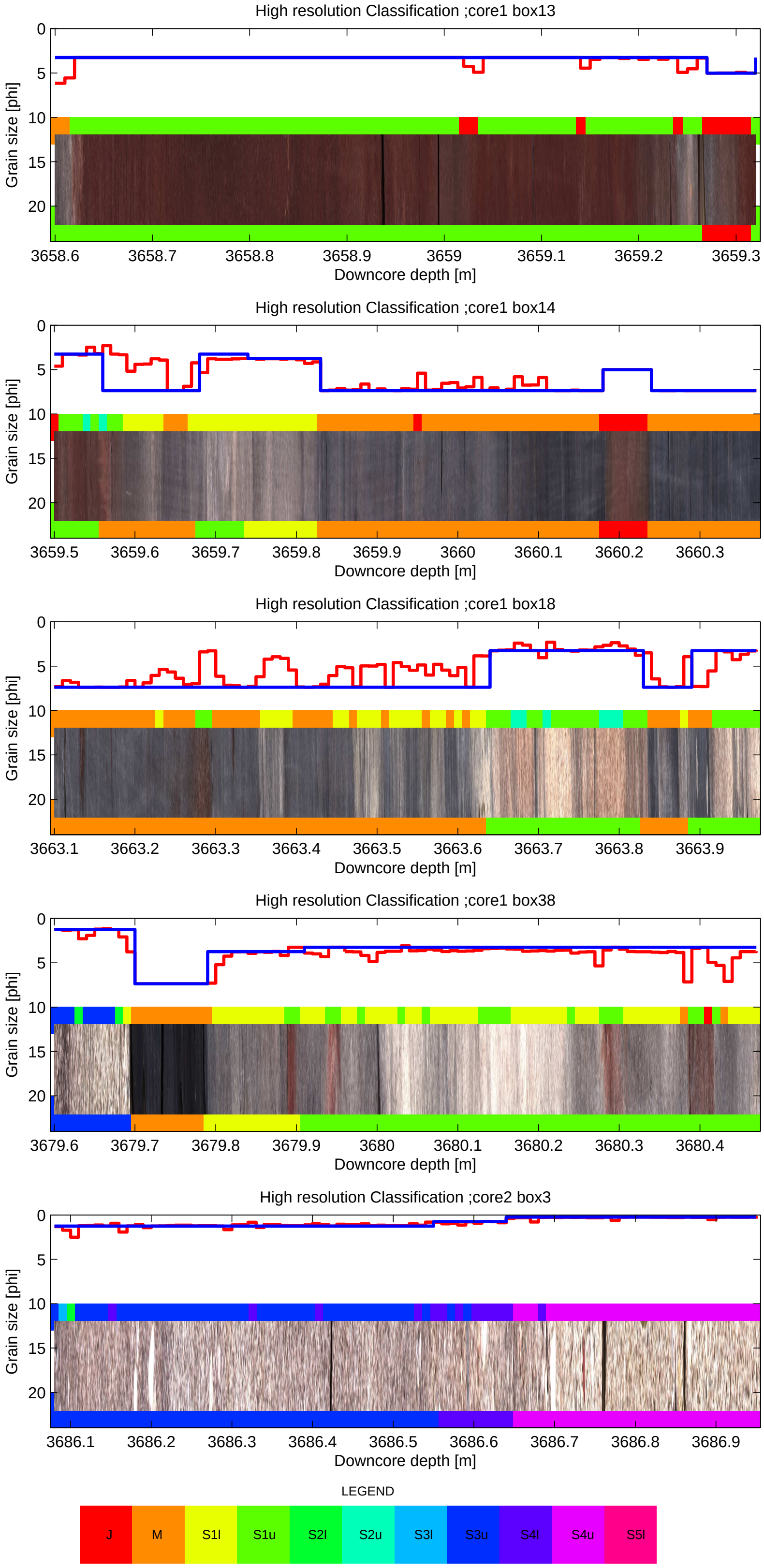

Figure 5.20: Core sections $\{(1,13),(1,14),(1,18),(1,38),(2,3)\}$ (core, box respectively) of E10-3 with initial classification (above core image), the 'hard' Bayesian classification results (above core image), the input grain size record (blue curve) and the predicted grain size record (red curve).

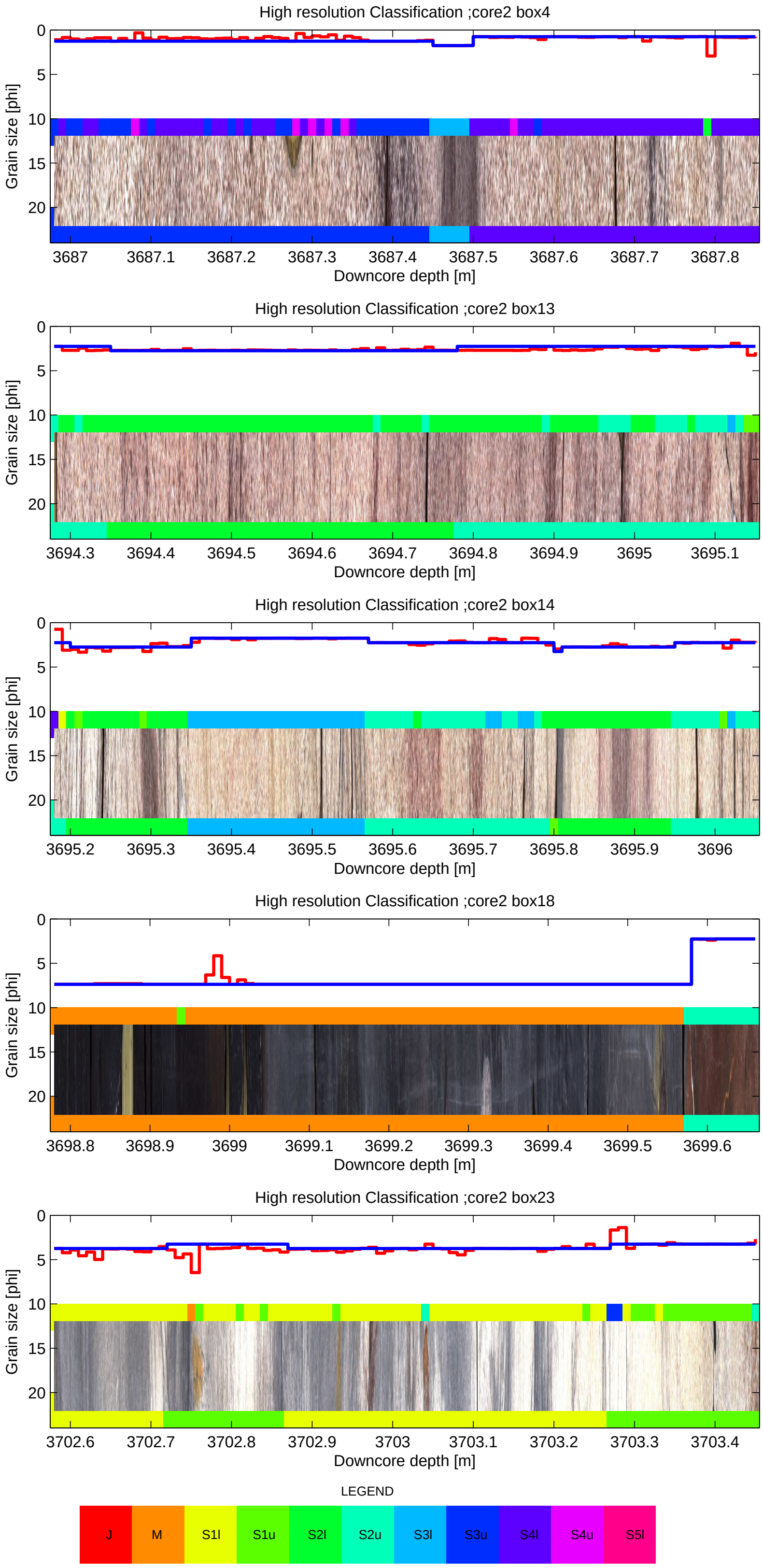

Figure 5.21: Core sections $\{(2,4),(2,13),(2,14),(2,18),(2,23)\}$ *(core, box respectively)of E10-3 with initial classification (under core image), the 'hard' Bayesian classification results (above core image), the input grain size record (blue curve) and the predicted grain size record (red curve).*

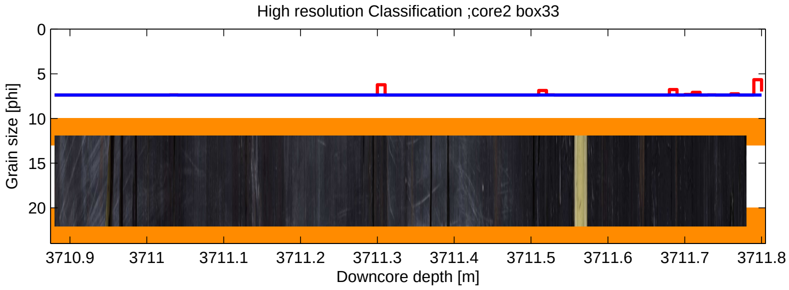

Figure 5.22: Core sections $\{(2,33)\}$ *of E10-3 with initial classification (under core image), the 'hard' Bayesian classification results (above core image), the input grain size record (blue curve) and the predicted grain size record (red curve).*

Although it followed from the data that some skewness is present, the discrepancies between the plug data and the predictions stays roughly within 1.5 phi-units in 66 % of the cases.

Question 3:

> *What is the added value of XRF core scanning data in the process of down core lithofacies classification?*

A down core classification based on geochemistry and RGBD-data have been applied in order to get a continuous record in terms of grain size classes. From the correlation coefficients of the different predictors w.r.t. the grain size signal we conclude that the elements Al, Si, K and Ti are the most valuable contributors to the grain size prediction. The tendency of relative enrichment w.r.t. Si with an increasing grain size and depletion of Al is widely accepted.

The classification scheme identified 2 core sections as being mis-classified. This includes core 1 box 14 and core 1 box 18 where the XRF-guided classification has yielded a relatively coarse-grained classification w.r.t. the input data. Additionally in these core the XRF-guided classification has identified some small down core shifts of the class-boundaries. Since the classification scheme has a probabilistic foundation, proxies for the *class mixing proportions* were obtained rather than the 'hard' classification in the core description. This resulted in smooth down core transitions in terms of grain size. These observations illustrate how the XRF core scanning can be useful as a quality control. Furthermore the ability to increase the resolution of the classification with an order of magnitude can be considered an improvement. Nevertheless a validation of the core predictions in terms of grain size (classes) is necessary to estimate the performance in a quantitative sense.

Chapter 6

Conclusions and Recommendations

6.1 Conclusions

In this study we tried to find answers to three main questions. We will conclude on these research questions one by one.
Question 1:

> **How can the current work flow be optimized to reduce costs and to increase accuracy and or precision?**

Based on the log ratio framework suggested by Aitchison, a filtering step was suggested that applies the uncertainty estimates of the XRF by replicate analysis. Additionally a sample selection criterion was suggested in order to achieve an effective XRF calibration. By applying these two algorithms to core 7919 and 7920, it was found that both algorithms are effective. The sample selection criterion yielded a reduction in the amount of samples needed to obtain the same level of prediction quality that is obtained when the samples are selected randomly. Additionally the pre-calibration filtering step was found to result in lower discrepancies between predicted and 'true' geochemistry.

When applying the sample selection algorithm to synthetic data, it was found that the same calibration quality was obtained with around 20-30% less samples compared to the random approach. Furthermore the filtering step resulted in a significant increase in the signal to noise ratio. The performance of both routines was given a statistical foundation; the sample selection criterium is effective for all number of calibration samples and the filtering step was found to be effective up to a maximum of 13 samples *for this dataset* (level of significance 5 %).

When applying the sample selection algorithm to the data of core E10-3, it designated 20 sample locations on which covered 8/10 of the identified grain size classes. When the number of requested samples was the same as the number of lithotypes, 7/10 grain size classes were explored. These results illustrate the ability of the algorithm to identify lithological extremes, although it has no lithological data input.

Question 2:

> **Can we build a forward model of basin infill in a formal geochemical and textural context?**

Based on the concepts of linear mixing of endmembers with a grain size dependent composition, the LINMIX model was built. The forward LINMIX modeling for core 9508 has been successful in the sense that it yields compositional predictions of comparable quality of the XRF-core scanning data. These predictions were found to be obtained for 1 specific EM-TF configuration. This model was found to have a geological significance because the two endmembers that were assigned 1 TF, have been identified in previous work as reflecting sediment of a different sediment source (eolian vs fluvial). Chemically the obtained TF's correspond to the model findings; fluvial sediment has in general, in contrast to eolian sediment, a tendency to be more rich in Fe. The concentration of K on the other hand is generally lower in fluvial sediment. Apart from the TF's itself, the corresponding endmember composition are within acceptable limits of present-day observations of fluvial and eolian sediment collected in Central Africa.

Question 3:

> **Is it feasible to predict textural properties based on geochemistry, given multiple data sources and a known, straightforward provenance setting?**

Since the prediction of textural properties was based on XRF core scanning, first the performance of the scanner was investigated when applied to consolidated sediment, the result of which are encouraging. The PCA-filtering step preserved 15 principal components of the 23 elements measured.

Subsequently the XRF core scanning data and the RGBD data has been embedded in a grain size prediction work flow using Multi Variate Regression (MVR). The discrepancies between the plug data and the predictions stays within 1.5 ϕ-units in 66 % of the cases. Additionally a classification scheme based on Bayesian inference identified two distinct core sections that should be classified differently, given the XRF and RGBD data. By applying this Bayesian classification *mixing* proportions of the different classes are obtained; something which the core description is lacking. The core sections of which the Bayesian classification deviates from the core description includes core 1 box 14 and core 1 box 18 where the XRF-guided classification yields a relatively coarse-grained classification w.r.t. the input data. Additionally the XRF-guided classification resulted in small down core shifts and smoothening of the class-boundaries. These results illustrate the value of XRF core scanning as a quality control. Furthermore the ability to increase the resolution of the classification with an order of magnitude can be considered an improvement. Nevertheless a validation of the core predictions in terms of grain size (classes) is necessary to validate the predictions quantitatively.

General research question:

> **Can geochemical data be used to obtain textural properties of siliciclastic sediment?**

The answer to this question is not so straightforward and depends on the available geological information. For instance in case there is evidence that the sediment originates from the same source, a fairly good grain size prediction is possible. Carboniferous core E10-3 was texturally characterised rather accurate.

However in the absence of source-information, the answer to this question is less straightforward. Although the LINMIX model is capable of incorporating source information, factors like the validity of the compositional trend and the *belief* in the correspondence between EMMA-derived grain size endmembers and the ’true’ endmembers affect the answer to this question. In the test-case that was discussed in this thesis, the LINMIX model was able to reconstruct the input data. However in order to apply this model, additional information is necessary to predict grain size. The ’direct’ grain size prediction based on geochemistry, which is essentially the 1-TF model, proved to be insufficient.

6.2 Recommendations

Based on the conclusions and the findings presented in this thesis, we suggest a couple of recommendations. These recommendations are subdivided into different topics knowing; sampling-related, model-related, scanner-related and software-related suggestions.

Sampling

XRF core scanning data relies on two aspect; intensity data as obtained with the scanner and geochemical data of the calibration samples. Whereas the primary source of error for the unconsolidated cores is on the intensity data side, for the consolidated core the final geochemical prediction is primarily affected by the calibration sample quality. This ’quality’ is not only determined by the quality of the geochemical analysis techniques, as well to the sampling itself; variability of sample size and accuracy of its corresponding down core position.

The used sampling methodology (using a slash-hammer) in this study is definitely not the most accurate one, because chips will come off at places at low cohesions which in itself is not independent of the composition. Future sampling should be performed similar to the acquisition of plugs. Using a small drill or some other sampling device will most probably give a direct reduction of the residual variance.

Another important question that should be answered for future use of in-situ XRF on consolidated cores is how many calibration samples are needed in order to perform proper grain size prediction/classification. Given the results we obtained w.r.t. the number of explored lithotypes, we suggest to use at least as many calibration samples as there are lithotypes being identified. This is based on the findings that for n = 20 and 12 lithotypes, almost all lithotypes were explored (8/10). Furthermore the prediction discrepancies found for the synthetic data stabilized after around 15 samples.

Nevertheless this sampling strategy is merely a rule-of-thumb and has no strong physical foundation. Therefore we recommend further research on sampling strategy and additionally on the general applicability of calibration samples for this purpose. For instance we can imagine that a clean sandstone without any ’contaminants’ has a pretty much standard composition, at least within the same sedimentary basin and stratigraphic level.

Linear Unmixing

A strong recommendation about the LINMIX model is the fact that the reliability of the endmember compositions can be drastically increased. By sampling the *pure* endmembers, the estimation of the TF’s is simplified, making the model output more reliable. Another topic of discussion is the information content of the TFs. The extrapolation of the geochemistry towards the composition of the source material as applied in this study is disputable due to several reasons. First of all the number of elements that have been analyzed is very small leading to a large blind spot. Especially the fact that the geochemistry of the size fractions was not expressed in terms of Si concentrations can be considered a hiatus. This because Silica is one of the major rock-forming elements. Secondly it is the general idea that the chemical alteration of the material has a large impact on the composition of the weathering products, making the reconstruction of the source rock composition from sediment difficult. Especially elements that are readily soluble will not record their crustal abundances in any simple fashion (Taylor and McLennan, 1981). Finally the extrapolation is only valid under the assumption that the compositional linear model of the TF is valid along the entire grain size spectrum. Recent work on field data has shown that the compositional linear trend shows some nonlinearities which were found to occur at specific grain size values (Tolosana-Delgado and von Eynatten, 2010).

This implies that if the LINMIX approach is applied on another core, the set of elements in which the composition is being expressed should comprise a large set of elements that includes relatively insoluble elements. Secondly more data should give more insight into the compositional behavior along the grain size spectrum. This could lead to a more in-depth understanding. Furthermore it should be investigated how applicable and valid

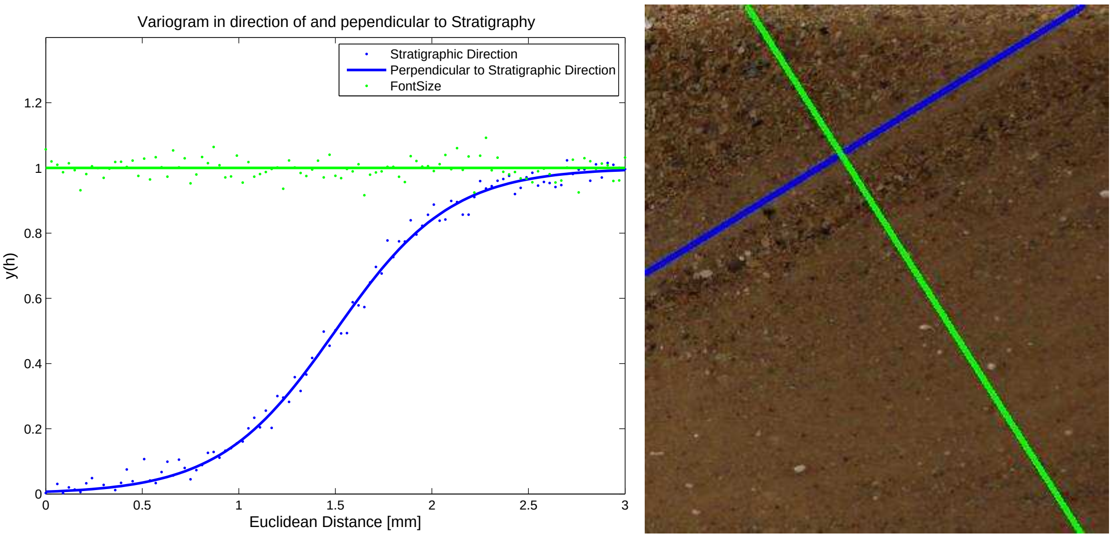

Figure 6.1: The calculation of a variogram in the stratigraphic directions yields grain size information while perpendicular to this direction, no spatial correlation will be found (synthetic example!).

these TF's are in the context of their spatial and temporal interval.

The scanner

The quantitative grain size prediction by XRF core scanning has shown to provide information that cannot be estimated in an acceptable time span using the current methodologies. Especially because the scanner can even go beyond the here applied XRF-sensor resolution of 1 cm. What need to be investigated though is whether there is a need for higher resolution geochemical record. If so, we suggest an improvement of the scanner in its current setup w.r.t. its camera. One of the limitations of the scanner's camera is the resolution. This because it does not facilitate grain size prediction supported by identification and/or counting of individual grains. With the current resolution this methodology is restricted to the coarse end of the grain size spectrum; sand-sized and coarser.

If the resolution of the camera would be higher, there are well-established multivariate techniques for the derivation of grain size. A possible approach in this would be to first determine the direction of the stratigraphic level (see figure 6.1). Secondly the calculation of a directional variogram should be performed to map the spatial correlation along the core. Subsequently the range of the variogram could then lead to a proxy of the grain size. However as mentioned, it is important that the stratigraphic direction is taken into account. This is especially important in presence of sedimentological structures such as cross-bedding because if one would determine a variogram under a direction other than the stratigraphic level, one will end up with an unexpectedly low spatial correlation. This because in a cross-bedding structure there is bedding of alternating fine and coarse sediment and if the variogram is not determined parallel to this bedding, the spatial variation is absent leading to a mis-classification.

Another way of extracting the grain size distribution would be to apply an unmixing strategy on the RGBD data for a given XRF observation. Given that RGBD data comprises a compositional quantity, it can be unmixed using the EMMA algorithms into the number and fingerprint of these color-endmembers. We imagine that if there is a homogeneous mud at the specific section of the image, the unmixing will not be very effective. On the other hand if the material has a strong texture, it leads to the identification of specific endmembers; either grains, pores or possible even cement. If subsequently the unmixing strategy is performed, analogous to the unmixing in terms of grain size endmembers, the chemical fingerprints of these endmembers can be derived. This means that if the XRF scanner has measured a transition from pure mud to sand, the transitional zone of the grain size record can be filled in by support of the color data.

The Software

A recommendation w.r.t. the data acquisition follows directly from the suggested filtering and sample selection approach. One of the problems with all the suggested methods and techniques is that it complicates the whole work flow. Users must take time to filter their data, subsequently do the sample selection. Furthermore they have to decide about a strategy of replicate sample acquisition; where should they be acquired and how many replicate analyses are necessary for a reliable repeatability estimate.However the algorithms suggested in this thesis can be implemented on the acquisition side, rather than the user side. If the scanner has scanned a core section, the covariance structure of the data can already be estimated. Given that the covariance structure of the intensities is known, although not calibrated as such, the selection of ’extremes’, analogous to the calibration sample selection, can be performed. The replicate analysis of these ’extremes’ seems a reasonable approach because the repeatability of the scanner is partly determined by the composition of the bulk. As such, this approach will provide a reliable estimate of the repeatability. Nevertheless this approach asks for a direct post-scan spectrum quantification; the calculation of element intensities from the observed spectra. Given that this approach is more user-friendly since it is full-automatic, we recommend a direct and embedded spectrum analysis in order to facilitate this full-automatic work flow.

Appendix A

Data Cores 7920 and 7919

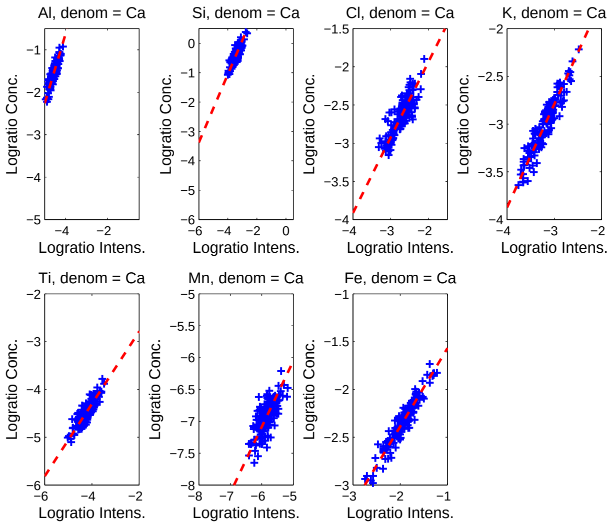

Figure A.1: Composition of calibration sample vs. predicted composition of core 7920.

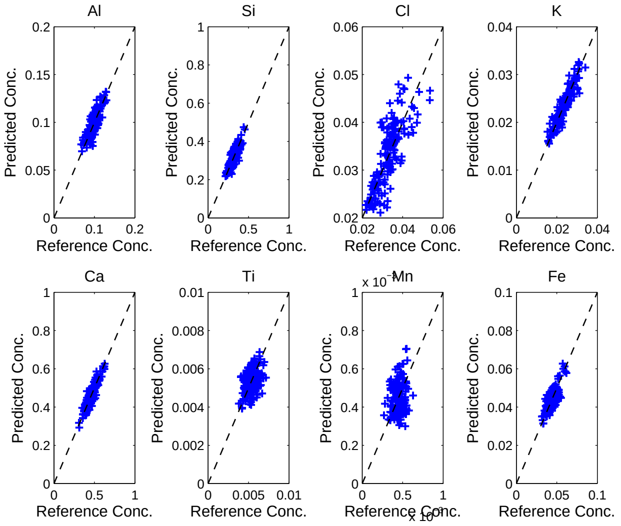

Figure A.2: LRCE model results of 7920.

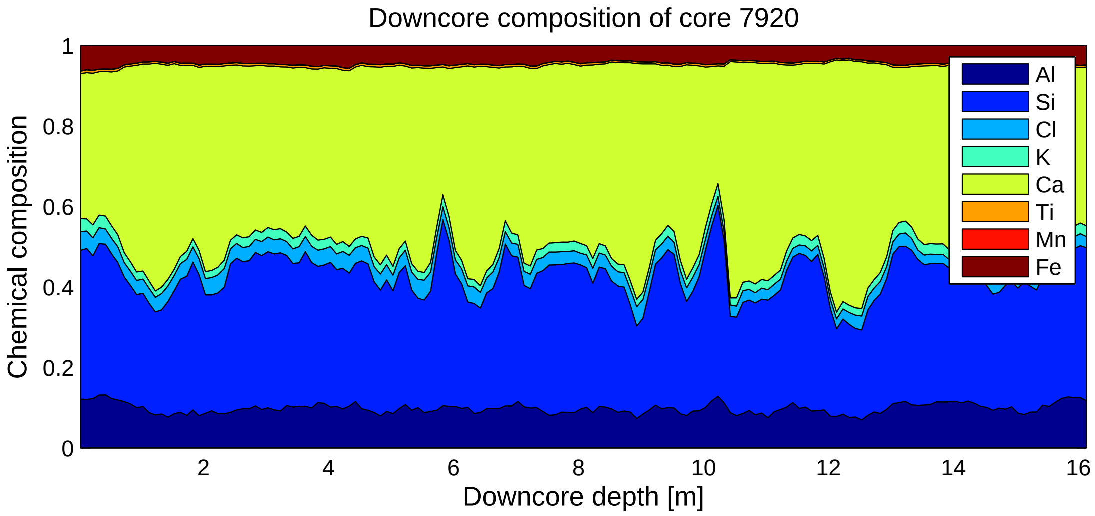

Figure A.3: Down core compositional record of 7920.

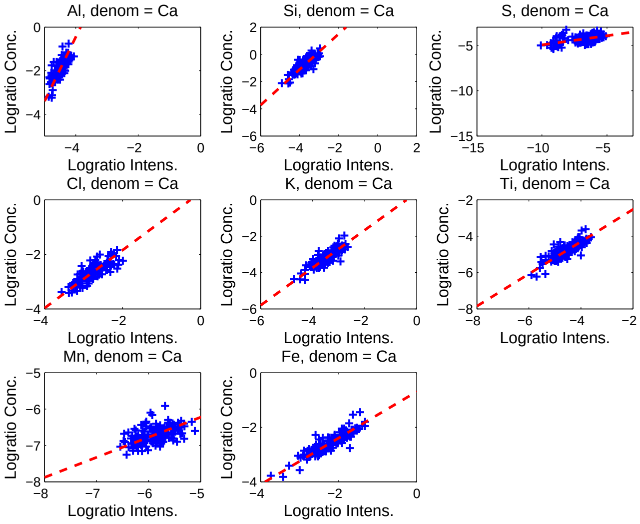

Figure A.4: Composition of calibration sample vs. predicted composition of core 7919.

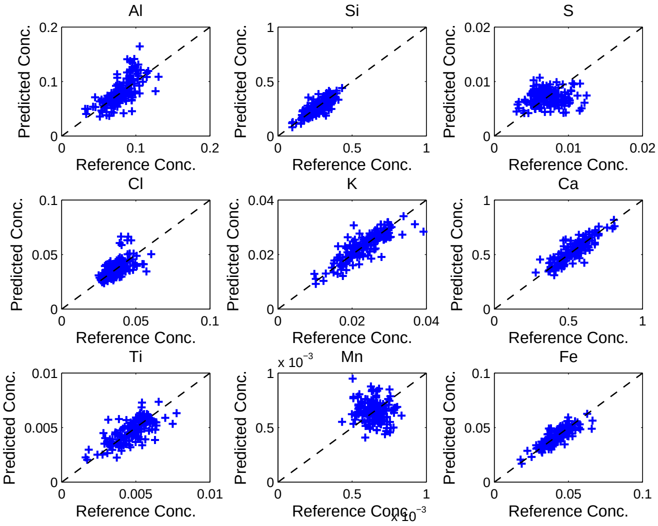

Figure A.5: LRCE model results of 7919.

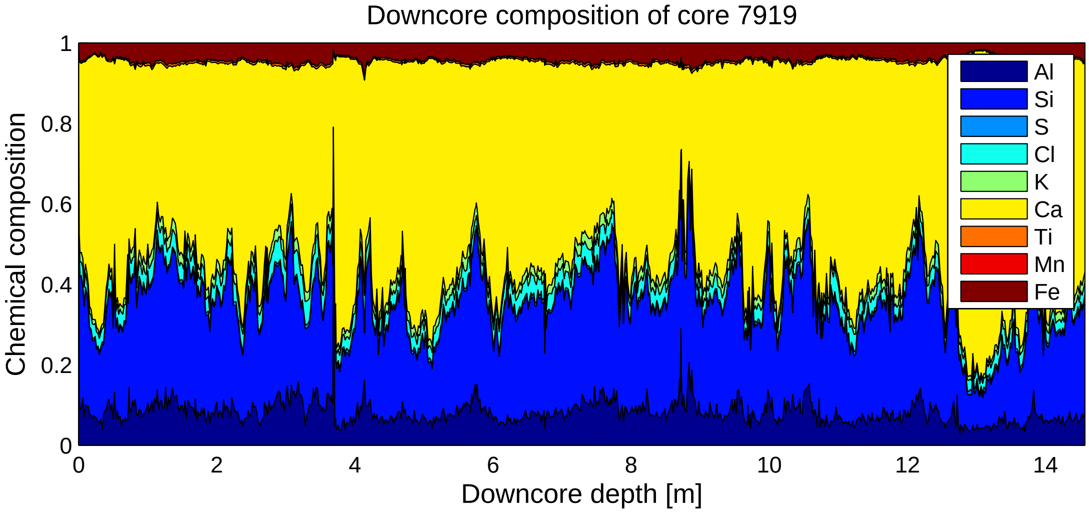

Figure A.6: Down core compositional record of 7919.

Appendix B

Lithofacies in core E10-3

Described below are the different lithofacies, as identified in the core report of Core E10-3 (Boels, 2003).

Braided channel complex (BC3)

Roughly 55% of the cores from Well E10-3 consists of the braided channel lithofacies association, making it the dominant lithofacies association. The grain size of the sandstones is mainly upper fine to medium sand but varies from pebbles- to silt-sized sand. The sedimentary structures observed within the beds comprise alternations of massive, high-angle cross-bedding, low-angle cross-bedding, tangential cross-bedding, horizontal lamination, and ripple lamination.

The sandstones of the braided channel lithofacies association are interpreted as the bed load deposits of a high-energy braided channel complex system. The large continuous sandbody thickness, the general absence of a clear grain size trend, the abundance of tangential and low-angle cross lamination, and the coarse-grained internal erosive reactivation surfaces all suggest that these units make up deposits from high-energy braided streams.

Crevasse splay (CS)

About 11% of the cored interval was assigned to the Crevasse Splay deposits. The lithofacies association consists of parallel, low-angle, and ripple laminated very fine-grained sandstone beds with moderate amounts of carbonaceous matter and clay (5-35%) occurring as intraclasts and laminations. Also, these deposits are moderately often rootletted at the top. The average bed thickness is 0.89 m. This association commonly marks the transition of IB (interdistributary bay) to IFL (poorly drained floodplain) facies associations.

These deposits were interpreted as crevasse splay deposits on basis of the fine-grained nature of the sandstone beds, the relatively thin bed thickness, the clay occurring as laminae and the occurrence within floodplain sediments (IFL and IB, see below). These deposits formed on the floodplain as the result of a breaching of channel margins during high water discharge. The rootletting in the upper part of the deposits would imply an abrupt cessation of sediment deposition allowing development of vegetation on the sediment surface.

Poorly drained floodplain (IFL)

Poorly drained Floodplain deposits (IFL) comprise nearly 18% of the cored interval. The association consists of dark grey to black, abundantly rootletted claystones with regular coal laminae, frequent siderite nodules, and comon preserved organic matter. The poorly drained floodplain contains abundant intercalations of swamp (SW) and crevasse splay (CS) deposits. The poorly drained floodplain is recognized as such by the dark grey colour, the abundant rootletting and good preservation of organic matter. This all indicates a waterlogged or slightly submerged, reducing environment with intense plant growth. Being a lateral equivalent of the braided channel sandstones, the dominance of clay-sized deposits implies a setting distal to the active distributary.

Interdistributary bay deposits (IB)

Nearly 12% of the cored interval consists of Interdistributary Bay deposits. The association consists of dark grey, relatively undisturbed laminated claystones with a few silt or sand laminae. Some beds have a yellowish coloration due to siderite cementation. Distortion through soft sediment deformation occurs sporadically. The transition from IB (Interdistributary bay) to IFL (poorly drained floodplain) lithofacies association in the cored interval systematically occurs through CS lithofacies association (crevasse splay).

The well-preserved parallel lamination of the claystones, and the absence of rootletting imply deposition within a permanently standing body of water, hence the deposits classify as Interdistributary bay deposits. The upward transition from Interdistributary Bay to Poorly drained Floodplain deposits through Crevasse Splays is considered to reflect the progradation of fluvial facies.

Swamp (SW)

Swamps comprise a small percentage of the cored interval (ca. 3parallel laminated organic rich claystones with cm-thick coal laminae. The coal beds, and the abundance of organic material suggest a waterlogged, anoxic environment with abundant floral inhabitance and no clastic input. Hence, these deposits are classified as Swamp (SW).

Well drained Floodplain (F)

3% of the cored interval consists Well-drained Floodplain, which only occurs in the lowermost section of core 2. It consists of thoroughly rootletted sandy claystones with a mottled appearance and an overall reddish coloration. Dark fragments occur which may represent the former host sediments give the rock its brecciated appearance.

These sediments are termed Well-drained Floodplain on basis of the following arguments. The strong disturbance of the sediment suggests slow sedimentation rates and pedogenesis. The primary red coloration indicated that at time of deposition the ground water table was periodically below the sediment surface at the time of deposition. For example, the environment could have been subject to seasonal flooding.

Appendix C

E10-3 Compositional record

Core	Begin [km]	End [km]	Index range
Core 1 Box 13	3.6586	3.6593	1-73
Core 1 Box 14	3.6595	3.6604	74-161
Core 1 Box 18	3.6631	3.6640	162-249
Core 1 Box 38	3.6796	3.6805	250-337
Core 2 Box 3	3.6861	3.6870	338-423
Core 2 Box 4	3.6870	3.6879	424-511
Core 2 Box 13	3.6943	3.6952	512-599
Core 2 Box 14	3.6952	3.6961	600-687
Core 2 Box 18	3.6988	3.6997	688-771
Core 2 Box 23	3.7026	3.7035	772-859
Core 2 Box 33	3.7109	3.7118	860-947

Table C.1: Down core ranges in terms of meters and indices applied in this study.

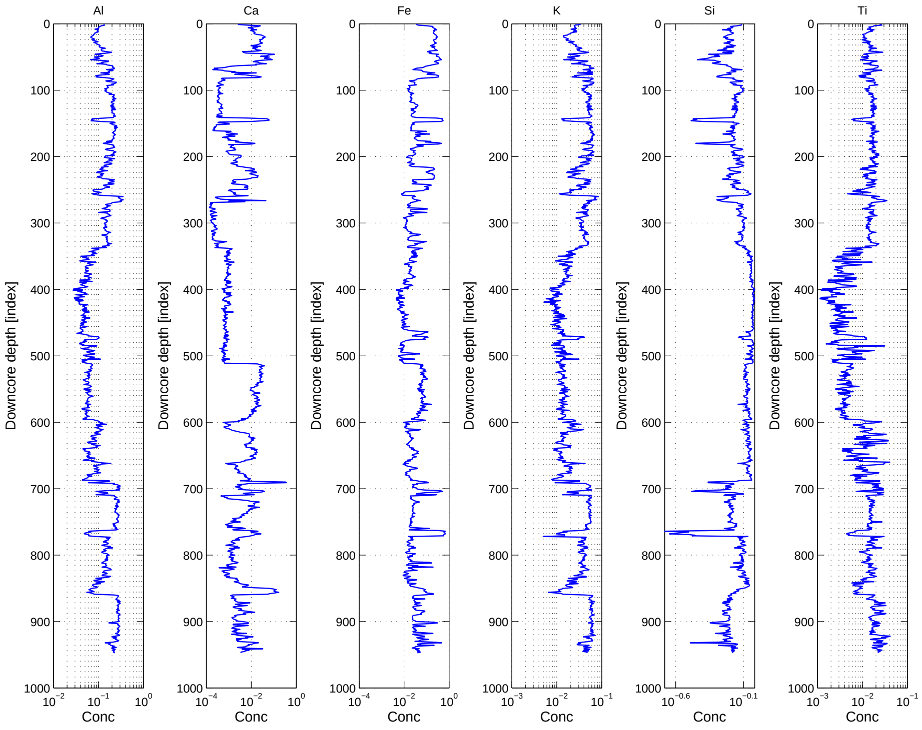

Figure C.1: Down core compositional record of Core E10-3.

Appendix D

Data Acquisition of core 9508

This section describes the data acquisition steps that were performed on core 9508, as performed in the work of Mulitza et al. (2008). Before the geochemical analysis of the sediment was performed, the terrigenous fraction was isolated; the primary source of calcium is of marine origin which is in this study of minor interest. The necessary pretreatment steps to isolate the terrigenous fraction are:

1. Organic Carbon was removed by adding 10 ml H2O2 (35%) to approximately 750 mg of bulk sediment.
2. Calcium Carbonate was removed by adding 10 ml HCL (100%) to the Carbon-free sediment in 100 ml demineralised water and boiled for 1 minute to speed up the reaction.
3. In order to remove the biogenic silica, NaOH pellets are added.

Grain Size Distributions of Core 9508 were obtained using a Coulter Laser Particle Sizer LS200, resulting in 59 grain size classes ranging from 0.39 to 88 μm.

A continuous record of the geochemistry was obtained by running an XRF core scan at 10 kV with measurement time of 30 seconds and an amperage of 350 μA. The XRF data was calibrated by using geochemical data derived from sediment samples. These samples were taken uniformly distributed along-core at 4 cm intervals. The element concentrations in these samples were determined on 4 g of dry subsamples by Energy Dispersive Polarisation X-ray fluorescence (EDP-XRF).

Grain size fractions have been physically isolated and analyzed w.r.t. the chemical composition, again by using EPD-XRF. This yielded the composition in terms of 6 elements which together represented on average 40 % (40 gr/kg sample) of the total mass of the samples. The grain size ranges of the isolated fractions are shown in table 4.1.

In order to obtain a down core age model, radiometric dating was performed on foraminifera (small organisms) picked from the coarsest grain size fraction.

Appendix E

Running Example CLS

To illustrate the Constrained Least Squares procedure, we will now present a small example of a 2-endmember system and 3 elements. We have the following observations (rows):

$$\mathbf{b} = \begin{bmatrix} 0.3 & 0.4 & 0.3 \\ 0.1 & 0.32 & 0.58 \\ 0.35 & 0.5 & 0.15 \end{bmatrix}, \qquad \mathbf{m} = \begin{bmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$$

The objective is now to estimate the coefficients a that determine the compositional change with membership. The number of coefficients is equal to $D \times N$ which sums up to 8 coefficients. Let us first present the system of constraints, denoted by $\mathbf{A}$:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Analogous the matrix with which we solve for the coefficients **H** is now given by:

$$\begin{bmatrix} 0.3 & 0.7 & 0 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 0.6 & 0.4 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.3 & 0.7 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0.4 & 0.8 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.1 \\ 0.35 \\ 0.4 \\ 0.32 \\ 0.5 \\ 0.3 \\ 0.58 \\ 0.15 \end{bmatrix}$$

By solving this CLS system (using the built-in MATLAB function *lsqlin* for instance), the coefficients for a are obtained. The partial bulk chemistry now follows from:

$$\hat{\mathbf{b}}_{\mathbf{p}_1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{a}_6 \end{bmatrix}, \quad \hat{\mathbf{b}}_{\mathbf{p}_2} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{a}_6 \end{bmatrix}$$

By applying the same methodology, also the composition of a grain size class can be determined. Let us define the following discrete grain size distributions, shown in figure E.1. Although in reality a grain size distribution is discretized into much more classes, we assume that the synthetic GSD can be satisfyingly reproduced using only 2 classes. If we would mix these two GSD's into the mixing proportions given below, we will end up with the relative proportion of each endmember. However these proportions do not necessarily sum up to one. Therefore the proportions are *closed* which leads, for the 3 observations, to the closed proportions shown in figure E.2.

Since the proportions sum up to one, the CLS system can be readily used which yields the endmember compositions *within* a grain size class.

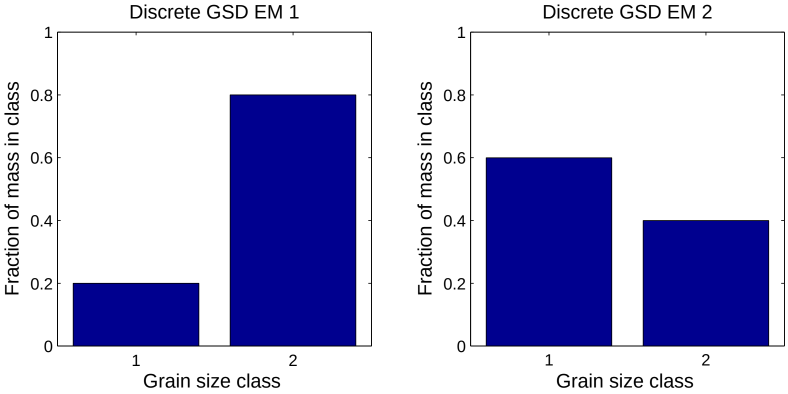

Figure E.1: Discrete GSD's of the synthetic 2-endmember system.

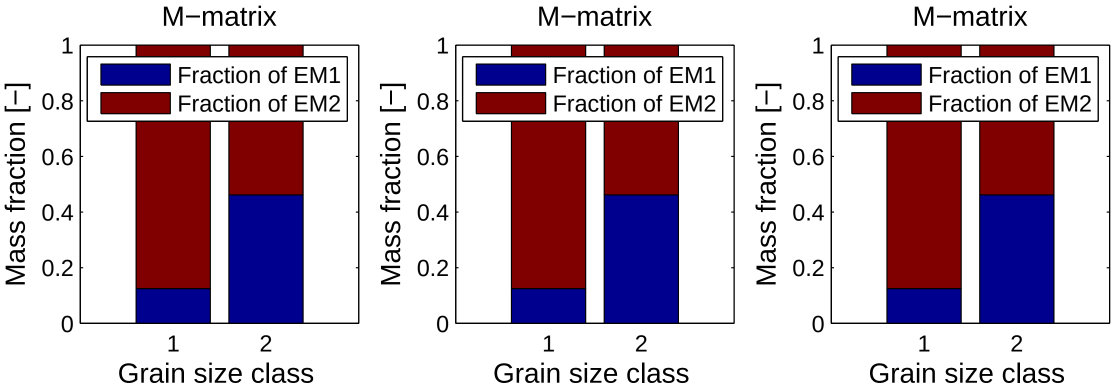

Figure E.2: The closed proportions **M** *for the three observations, respectively.*

Bibliography

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall.

Allen, P. (1997). *Earth surface processes*. Wiley-Blackwell.

Belt, van den, F. (1999). Sedimentology, petrography, and reservoir properties of well d12-5x, warmond. Technical report, Panterra Nederland BV.

Besly, B., Burley, S., and Turner, P. (1993). The late carboniferous 'barren red bed' play of the silver pit area, southern north sea. In Parker, J., editor, *Petroleum Geology of Northwest Europe: Proceedings of the 4th conference*, volume 4, pages 727–740. The Geological Society (London).

Boels, J. (2003). Sedimentology, petrography and reservoir quality of the upper carboniferous in well E10-3. Technical report, Panterra Nederland BV.

Brassington, F., editor (1990). *Rising groundwater levels in the United Kingdom*. Proceedings of the Institution of Civil Engineers.

Daunis-i-Estadella, J., Egozcue, J., and Pawlowsky-Glahn, V. (2002). Least squares regression in the simplex. In Bayer, U., Burger, H., and Skala, W., editors, *Proceedings of IAMG02 The eighth annual conference of the International Association for Mathematical Geology*, pages 411–416. Berlin.

Dinelli, E., Tateo, F., and Summa, V. (2007). Geochemical and mineralogical proxies for grain size in mudstones and siltstones from the pleistocene and holocene of the po river alluvial plain, italy. In Arribas, J., Critelli, S., and Johnsson, M. J., editors, *Sedimentary Provenance and Petrogenesis: Perspectives from Petrography and Geochemistry*, pages 25–36. Geological Society of America Special Papers.

Drozdzewski, G. (2005). Zur sedimentaren entwicklung des subvariscikums im namuriumund westfalium nordwestdeutschlands. In Wrede, V., editor, *Stratigraphie von Deutschland V, Das Oberkarbon (Pennsylvanium) in Deutschland*, volume 254, pages 151–203. Forsch.-Inst. Senckenberg (Frankfurt a. M.).

Dyck, van, P. M., Torok, S. B., and Grieken, R. E. V. (1986). Enhancement effect in X-ray Fluorescence analysis of environmental samples of medium thickness. *Analytical Chemistry*, 58:1761–1766.

Freedman, D., Pisani, R., and Purves, R. (1990). *Geological Atlas of Western and Central Europe*. Geological Society Publishing House, third edition.

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics*. W.W. Norton & Company, third edition.

Gac, J. and Kane, A. (1986). Le fleuve senegal: I. bilan hydrologique et flux continentaux de matieres particulaire a l'embouchure. *Sci. Geol. Bull.*, 39:99–130.

Graton, L. and Fraser, H. (1935). Systematic packing of spheres with particular relation to porosity and permeability. *Journal of Geology*, 43:795–909.

H., H. (1988). Origin and consequences of cyclic ice rafting in the northeast atlantic ocean during the past 130,000 years. *Quaternary Research*, 29:142–152.

Hollander, M. and Wolfe, D. (1973). *Nonparametric Statistical Methods*. Wiley.

Imrich, K. (2001). Failure analysis of a pilot scale metler. Technical report, Westinghouse Savannah River Company.

Jansen, J. H. F., der Gaast, S. J. V., Koster, B., and Vaars, A. J. (1998). Cortex, a shipboard XRF-scanner for element analyses in split sediment cores. *Marine Geology*, 151:143–153.

Jenkins, R. (1999). *X-Ray Fluorescence Spectroscopy*. Wiley and Sons, New York, second edition.

Jenkins, R. and De Vries, J. (1970). *Practical X-ray Spectrometry*. MacMillan.

Johnston, N. and Beeson, C. M. (1945). Water permeability of reservoir sands. *Trans. A.I.M.E.*, 160:43–55.

Krumbein, W. (1941). Measurement and geological significance of shape and roundness of sedimentary particles. *Journal of Sedimentary Research*, 22:64–72.

Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

Meng, X., D. B. R. (1993). Maximum likelihood estimation via the ECM algorithm. *Biometrika*, 80:267–278.

Mulitza, S., Prange, M., Stuut, J.-B., Zabel, M., von Dobeneck, T., Itambi, A. C., Nizou, J., Schulz, M., and Wefer, G. (2008). Sahel megadroughts triggered by glacial slowdowns of atlantic meridional overturning. *Paleoceanography*, 23,doi:10.1029/2008PA001637.

Orange, D., Gac, J., and Diallo, M. (1993). Geochemical assessment of atmospheric deposition including harmattan dust in continental West-Africa. *Tracers in Hydrology*, 215:303–312.

Posch, M. and Kurz, D. (2007). A2M - a program to compute all possible mineral modes from geochemical analysis. *Computers and Geosciences*, 33:563–572.

Renner, R. (1991). An examination of the use of the logratio transformation for the testing of endmember hypothesis. *Mathematical Geology*, 23(4):549–563.

Ricken, W., Schrader, S., Oncken, O., and Plesch, A. (2000). Turbidite basin and mass dynamics related to orogenic wedge growth; the rheno-hercynian case. In Franke, W., Haak, V., Oncken, O., and Tanner, D., editors, *Orogenic Processes: Quantification and Modelling in the Variscan Belt*, volume 179, pages 257–280. Geological Society Special Publication (London).

Schartz, A. (1974). *Calculus and analytic geometry*. Holt, Rinehart and Winston.

Serra, O. (1984). *Fundamentals of Well-Log Interpretation. I: The Acquisition of Logging Data.*. Elsevier Science Publishing Co.

Sinha, R., Smykatz-Kloss, W., Stuben, D., Harrison, S., Berner, Z., and Kramar, U. (2006). Late quaternary palaeoclimatic reconstruction from the lacustrine sediments of the sambhar playa core, thar desert margin, india. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 233:252–270.

Somme, T., Helland-Hansen, W., Martinsen, O., and Thurmond, J. (2009). Relationships between morphological and sedimentological parameters in source-to-sink systems a basis for predicting semi-quantitative characteristics in subsurface systems. *Basin Res.*, 21:361–387.

Stuut, J., Zabel, V., Ratmeyer, P., E., H., Schefuss, G., L., and R., S. (2005). Provenance of present-day eolian dust collected off north-africa. *J. Geophys. Res.*, 110. Doi:10.1029/2004JD005161.

Svendsen, J., Harald, H. F., Stollhofen, and Hartley, N. (2007). Facies discrimination in a mixed fluvio-eolian setting using elemental whole-rock geochemistryapplications for reservoir characterization. *Journal of Sedimentary Research*, 77:23–33.

Swerts, J. and Van Espen, P. (1993). Partial least squares techniques in the energy-dispersive x-ray fluorescence determination of sulfur-graphite mixtures. *Analytical Chemistry*, 65:1181–1185.

Taylor, S. and McLennan, S. (1981). *The composition and evolution of the continental crust, Rare Earth Element evidence from Sedimentary Rock*, volume 301. Blackwell Scientific Publications.

Tjallingii, R., Rohl, U., Kolling, M., and Bickert, T. (2007). Influence of the water content on XRF core-scanning measurements in soft marine sediments. *Geochem., Geophys.*, 8. Doi:10.1029/2006GC001393.

Tolosana-Delgado, R. and von Eynatten, H. (2008). Grain-size control on petrographic composition of sediments: Compositional regression and rounded zeros. *Mathematical Geosciences*, 41:869–886.

Tolosana-Delgado, R. and von Eynatten, H. (2010). Simplifying compositional multiple regression: Application to grain size controls on sediment geochemistry. *Computers and Geosciences*, 36:577–589.

Weltje, G. (1997). End-member modeling of compositional data: Numerical-statistical algorithms for solving the explicit mixing problem. *Mathematical Geology*, 29:503–549.

Weltje, G. and Prins, M. (2003). Muddled or mixed? inferring palaeoclimate from size distributions of deep-sea clastics. *Sedimentary Geology*, 162:39–62.

Weltje, G. and Tjallingii, R. (2008). Calibration of XRF core scanners for quantitative geochemical logging of sediment cores: Theory and application. *Earth and Planetary science letters*, 274:423–438.

Weltje, G. and von Eynatten, H. (2004). Quantitative provenance analysis of sediments review and outlook. *Sedimentary Geology*, 171:1–11.