Natural Resources Research, Vol. 19, No. 1, March 2010 (© 2010)
DOI: 10.1007/s11053-010-9110-4

A Comparison of Unstructured and Structured Principal Component Analyses and their Interpretation

Kristian Bjarnøe Brandsegg,[1,2,3] Erik Hammer,[1] and Richard Sinding-Larsen[1]

Received 17 September 2009; accepted 11 January 2010
Published online: 4 Feb. 2010

Multivariate analysis is employed to investigate the structure of variations within highly heterogeneous data. Traditionally, principal component analysis (PCA) is run by analyzing the entire wireline log and using PCA scores to characterize variability within and between lithologies. In this paper, we propose a technique using only specific subsets of all well records to quantify reservoir heterogeneity due to second order lithological variability. These subsets are chosen from uniform lithofacies parts of the wireline log in order to reduce the variability in the correlation matrix that otherwise would cause lithological changes. The purpose is to assess the efficiency of structured PCA in analyzing small-scale heterogeneity that is captured by wireline logs but often masked by traditional PCA approaches. This paper shows that a structured PCA procedure based upon special lithological units is superior to an unstructured PCA, when the focus is within lithology variations. This structured procedure is applied to data from the Heidrun field, offshore mid-Norway. The results demonstrate clear benefits from added insight into the variability of a complex fluviodeltaic heterolithic sequence that poses great challenges to hydrocarbon development.

KEY WORDS: Heterogeneous reservoir, fluviodeltaic deposits, well log interpretation.

INTRODUCTION

The quantification of heterogeneity in sandstone reservoirs is often challenging as the magnitude and type of heterogeneity is normally not known beforehand. Multivariate data analysis can be applied to study and visualize the data in a more comprehensive way and thereby ease the interpretation of heterogeneity. One common goal of multivariate data analysis is to reduce the dimensions of a specific dataset without losing information. Linear combinations of the original variables created through principal component analysis (PCA) define a smaller set of variables that extract successively the maximum of the remaining variability (Jolliffe, 2002). Another goal can be to seek the most representative multidimensional structure according to a given problem. This implies seeking an appropriate variance covariance matrix as input to the PCA. The general objectives can therefore be twofold; data reduction and interpretation (Davis, 2002). PCA has the potential to show relationships not previously suspected, and thereby uncover associations that are not readily seen.

Standard PCA, investigating the entire dataset to indicate gross-variability, has been applied to many disciplines. In geosciences, PCA has been widely used to evaluate geological processes using satellite images (c.f. Petrovic, Khan, and Chafetz, 2008) or outline different lithological types from wireline logs (c.f. Gupta and Johnson, 2001). Zhang and others (2007) show an example where PCA has successfully been applied to detect hydrocarbon bearing sands from satellite images.

[1] Department of Geology and Mineral Resources Engineering, Norwegian University of Science and Technology, N-7491, Trondheim, Norway.
[2] Exploro AS, Stiklestadveien 1, N-7041, Trondheim, Norway.
[3] To whom correspondence should be addressed; e-mail: kristbra@ntnu.no

1520-7439/10/0300-0045/0 © 2010 International Association for Mathematical Geology

In Zhang's study, the residual principal components (PCs) that are not disturbed by the non-hydrocarbon influence of the first PCs outlined different hydrocarbon bearing zones that could be a target for exploration.

Stanley and Sinclair (1988) introduced the term structured PCA for an analysis that only uses a specific subset of variables in a geochemical survey in order to better outline mineralized zones. Their results showed that the PCA using only the trace elements that were related to rock forming minerals outlined the major lithological units, whereas the mineralization was not delineated. Different modes in the frequency distribution of the polymodally distributed trace elements from the study area were identified and used to select the variables that in the structured approach pointed to the mineralized zones. This interpretation philosophy is in this paper extended to wireline log interpretation with some important modifications.

Normally in wireline log analysis the PCs are computed from the correlation matrix, where each input variable is equally weighted. The correlation matrix is chosen due to differences in scale of each wireline log variable (Doveton, 1994). Although PCA is a robust and powerful method for both visualizing and manipulating the multidimensional representation of wireline data, it cannot be used as a black box and should be carefully designed in order to obtain significant results. A major limitation of PCA is that the first few principal component axes extract max variability, but this does not guarantee the best subset of features (Nadler and Smith, 1993). This is due to the fact that PCA uncovers feature combinations that model the variance of a data set, but these may not be the same features that separate the different lithological changes. However, a structured PCA approach, analyzing a specific lithological unit or interval, includes only relevant variability and each of the PCs will therefore explain lithological effects that are not indicated when analyzing the entire data set which contain both relevant and not-relevant heterogeneity (Stanley and Sinclair, 1988).

Several methods have been proposed for the classification and grouping of lithologies (Davis, 2002). One method widely used is lithofacies analysis, introduced in 1980s using the name electrofacies, to characterize collective associations of wireline log responses that are linked to geological attributes (Serra and Abbott, 1982). It has been used in its standard form to characterize sequence stratigraphy (Eichenseer and Leduc, 1996), study heterolithic reservoirs (Gupta and Johnson, 2001), and for enhanced reservoir description (Pereira and others, 1990). Several studies of tidal and fluvial deposits have been analyzed in conjunction with PCA to enhance the understanding of heterolithic deposits separated into specific lithofacies (Avseth, Mukerji, and Mavko, 2005; Bourquin, Rigollet, and Bourges, 1998; Bridge and Tye, 2000; Hohn and others, 1997; Moline and Bahr, 1995; Singh, 2007). A common challenge in wireline log interpretation and petrophysics is to determine the relation and reliability of measurements of rock properties made at the borehole scale with the same property at the reservoir scale (Corbett, Jensen, and Sorbie, 1998). This challenge is particularly apparent in heterogeneous fluviodeltaic deposits (Martinius and others, 2005).

The present study uses PCA in a non-standard form in both unstructured and structured mode to characterize fluviodeltaic reservoir heterogeneity from wireline logs. The wireline logs represents separate measurements of different physical properties of the rock-fluid system and do not pose any simplex space constrains that in the case of compositional data violates the use of standard PCA. The objective is to describe and evaluate the benefits of using a modified structured PCA approach to show how petrophysical wireline log responses can be decomposed to reflect different orders of variability and how these can be differentially interpreted to provide additional insight into fluviodeltaic heterogeneity and its lithological complexity.

METHOD

Study Area, Wireline Data, and Software

The present study was carried out over a 300 m zone of the Upper Triassic to Lower Jurassic fluviodeltaic Åre Fm. (Dalland and others, 1988) from the Heidrun Field, offshore mid-Norway. A vertical water saturated well was selected where both core and petrophysical parameters have been thoroughly studied relative to five wireline logs (gamma ray, neutron porosity, bulk density, resistivity, and sonic logs). The computations have been performed within the R language, a free and open source software, which facilitates data manipulation, calculation, and graphical display (Dalgaard, 2008).

Univariate Analysis

The first step of an univariate analysis is to interpret the shape of the frequency distribution. The most prominent populations revealed by the five wireline logs can be identified by cumulative probability plots with the percentiles of the normal distribution as x-axis. Any polymodality in the five distributions might be caused by specific lithological processes. The identification of the number of modes or populations is determined by the inflection points in each of the probability plots (Stanley and Sinclair, 1988). The overlap between the neighboring populations is calculated separately for data with more than two populations and summed to portray the total magnitude of population overlap.

Unstructured and Structured PCA

Multivariate analysis of the five independent wireline log responses needs to be performed in order to supplement the univariate analysis. Eigenvectors representing orthogonal directions in space permit the viewing of data from a variety of perspectives. The aim of the modified structured PCA used in this study is not to reduce the dimensionality of the data, but to work on subsets of the wireline log and only include those samples in the correlation matrices that capture particular heterogeneity effects related to specific lithological units. PC loadings and scores are, according to this procedure, calculated from (1) a total unstructured analysis of all well records from all wireline log variables and (2) a structured subset of separate well records from specific lithological units. The analysis of the entire 300 m interval has been named TPC due to the use of the total number of records, whereas the structured approach is named according to the lithological units covered (sandstone, shale, coal, and cemented layers). A lithofacies classification is used to outline the geologic variation in rock types. The total wireline log interval was manually classified into four lithofacies based upon core analysis and wireline log responses according to the following rock types: sandstone (ss), shale (sh), coal (co), and cemented layers (cc). The choice of what samples to include in a subset was done on the basis of this facies interpretation in order to obtain apparently homogeneous lithological samples that could unmask the internal heterogeneity that otherwise is obstructed by intra-lithological variability. The loadings from the structured subsets are used to calculate PC scores that can be used to extrapolate the specific lithological signatures to the totality of well records. This calculation makes it possible to compare the log responses from the unstructured and the structured approach for the complete well.

Stability in the Eigenvectors

In order to ensure representativity of the computed eigenvectors for the sandstone (ss) subset, the following procedure was chosen: The subset was divided randomly into two groups and each group was analyzed. The loadings for the two subset groups were compared and expressed as a percentage difference between the initial subset and the two subset groups. The size of this percentage reflects the stability of the eigenvectors.

Visualization and Interpretation Methods

The populations defined by the PC scores, both resulting from unstructured and structured PCA, are evaluated by probability plots and histograms. Crossplots are used to visualize the relationship between components, both for the original data and for PC scores. Polymodal distributions are identified by selecting inflection points on each of the empirical cumulative frequency distributions, indicating a transition from one to the other population.

Standard PCA is often applied without interpreting the weighting (loadings) of each PC (Moline and Bahr, 1995). The interpretation of the loading values is however crucial as the loading signature represents a linear combination of variables that may or may not represent a process that make sense from a geological point of view. A comparison of the unstructured and structured loadings and their explained variability permits a detailed understanding of the relationship between geologic processes and the PC loadings and scores. Traditionally, loadings are only displayed in table form. In this study, two additional visualizations are carried out to enhance multidimensional similarity. The first uses visualization as star diagrams (Wegman, 1990) and the second is a glyph representation of Chernoff faces (Chernoff, 1973) mimicking human faces according to loading values.

RESULTS

Univariate Analysis

Initially, the statistical treatment of wireline data followed the procedure indicated by Stanley and Sinclair (1988) with the separation by univariate analysis of different populations of wireline log responses. Based upon these petrophysical responses, three major lithological units, sandstone (ss), shale (sh), and coal (co), in addition to cemented layers (cc), were manually identified on the basis of 15 cm well record intervals (Fig. 1 and Table 1). Probability plots of each wireline log were evaluated to identify the polymodality of the cumulative frequency distributions which in most cases reflects specific lithological populations. All five wireline logs exhibited polymodal distributions. The GR, RHOB, NPHI, and DT logs are visualized in Figure 2. The probability plot of the GR log indicates six populations, where the A population is interpreted to represent clean channel sands or cemented sandstone zones, B represents bayfill sands, F represents GR-rich spikes, and the C–E populations reflect coal and shale influenced sandstone intervals (Fig. 2a). Four populations are indicated on the RHOB log: A representing coal, B sand, C shale, and D cement (Fig. 2b). The largest

Table 1. Summary of the Lithofacies Description of the Studied Fluviodeltaic Well Interval

Lithofacies	Lithofacies ID	Description	Counts
Sand	1	Fine to medium grained sandstone	867
Shale	2	Shale and clay to very fine siltstone	830
Coal	3	Organic-rich coal to silt influenced coal	149
Cemented layer	4	Cemented layer	28

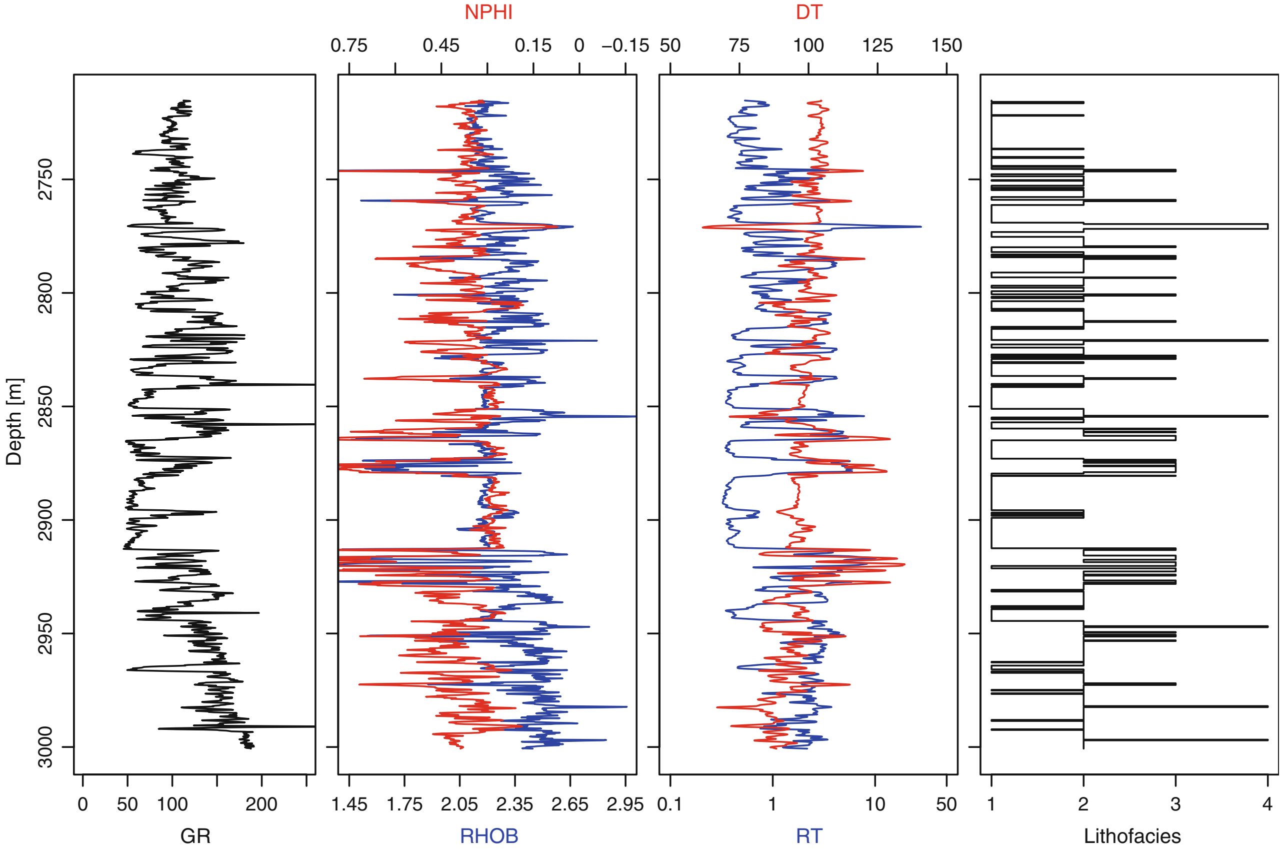

Figure 1. Five wireline logs (gamma ray, neutron porosity, density, resistivity, and sonic) that show a clear indication of vertical heterogeneity form the basis for the multivariate analysis. The lithofacies log derived from interpretation of cm scale core samples and the original wireline logs has the following notation: (1) sand, (2) shale, (3) coal, and (4) cemented layers.

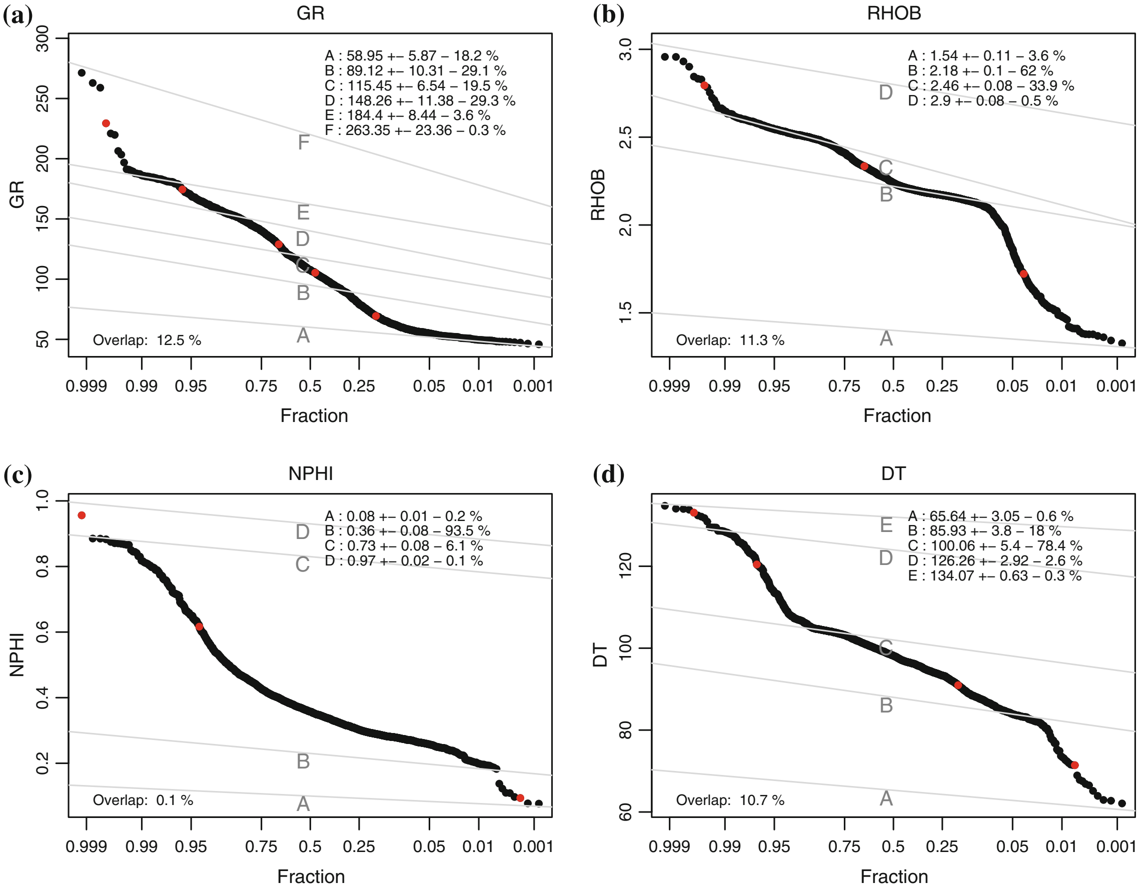

Figure 2. The probability plots, represented by four of the five wireline responses, indicate polymodal distributions. The mean and standard deviations for each distribution are specified including the percentile of the total records within each population. The red circles specify the inflection point between two populations and the lines indicate the average value for each population.

population of the NPHI log holds 93.5% of the total records (Fig. 2c) and reflects a range of sand/shale mixtures. The remaining three populations are interpreted to represent end-members with extreme low and high NPHI values related to cement and organic-rich coal, respectively. Interpretation of the DT log shows that the B–D populations reflect sand and shale intervals, whereas A indicates cement and E indicates organic rich coal (Fig. 2d). The two most exotic populations, organic-rich coal and cemented intervals, are clearly differentiated by these specific wireline responses. On the other hand, sand, shale, and impure coal intervals are found to be less distinguishable on the basis of only univariate analysis. An additional plot, describing the initial wireline log response distributions in relation to each lithological unit, indicates large variations of population overlap between the lithological units (Fig. 3).

Unstructured and Structured PCA

An unstructured PCA based on the total number of well records was used to observe the major variability from all lithological units. The calculations are computed from standardized wireline log values so all variables have equal variability (Table 2). Separate analysis of the probability plot of the first two unstructured PCs (TPC1 and TPC2) indicates four populations each (Fig. 4). The TPC1 does not allow for a clear distinction between all the different lithological population types (Fig. 4a). This is especially evident for the sand-shale population overlap. The second PC, TPC2, identifies the major cemented layer with extreme low TPC2 scores, as well as the difference between sand and shale units (Fig. 4b). The TPC1-TPC2 crossplot (Fig. 4c), combining the principal two unstructured PCs, allows for only a rough discrimination of the principal lithological units. However, this unstructured PCA crossplot still permits a more precise separation of the lithological units that can be obtained from the crossplot of the RHOB and NPHI wireline logs (Fig. 4d).

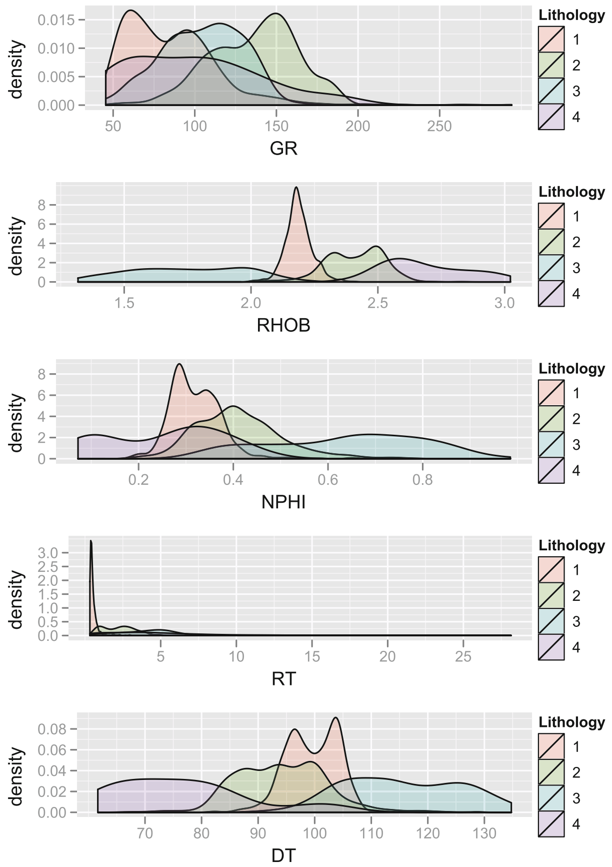

Figure 3. Separate plots displaying Kernel density plots of the initial wireline log responses within the records corresponding to four separate lithofacies units shown in Figure 2 with the following lithology notations: (1) sand, (2) shale, (3) coal, and (4) cemented layers.

The structured PCA that is based on the correlation matrix calculated from only a subset of well records highlights the internal variations within each of the interpreted lithological units, named PC_ss, PC_sh, PC_co, and PC_cc for sandstone, shale, coal, and cement, respectively (Table 2). All the probability plots for each separate lithological unit (Fig. 5), where the inter-lithological effect has been removed, still indicate polymodal distributions. However, the polymodality is caused by different intra-lithological populations characterized by the specificities of the structured loadings that now reflect the higher order variability once the inter-lithological variability has been removed. The principal sandstone lithological component, PC1_ss, can be separated into four subpopulations, where the A and B sub-populations comprise 3.0% of the total records. These populations are interpreted as GR-enriched sandstone and coal influenced sandstone, and the C and D sub-populations represent bay fill sand and channel fill deposited sandstone, respectively (Fig. 5a). Four sub-populations are also indicated by the PC2_ss (Fig. 5b): A represents a specific 4 m sandstone interval with low GR and NPHI values and higher RHOB values interpreted to be channel sands, B bay fill sand, C channel fill deposited sandstone, and D coal influenced sandstone. The principal shale component, PC1_sh (Fig. 5c), can be divided into two dominant subpopulations, assumed to be pure shale (B) and sand influenced shale (C). The dominant sub-population of PC2_sh (B), comprising 97% of the shale records, explains internal variations within the shale assumed to be related to porosity variations in contrast to the A and C sub-populations that respectively represent coal and cement influenced records (Fig. 5d). For the coal intervals, four sub-populations are indicated both for PC1_co and PC2_co. The A sub-population in PC1_co is pure coal and the remaining three populations are assumed to be related to the degree

Table 2. The First Three PC Loadings for the Total Unstructured PCA (TPC) and the Structured PCA of Each of the Lithofacies Groupings are Outlined

	TPC1	TPC2	TPC3	PC1_ss	PC2_ ss	PC3_ss	PC1_sh	PC2_sh	PC3_ sh	PC1_co	PC2_co	PC3_co	PC1_cc	PC2_cc	PC3_cc
GR	−0.169	−0.612	0.587	−0.741	0.171	0.645	−0.620	−0.196	0.734	0.043	−0.498	−0.164	0.106	0.190	0.713
RHOB	−0.592	−0.233	0.078	−0.221	−0.373	−0.076	−0.533	0.092	−0.245	0.573	−0.616	−0.079	−0.245	0.699	0.236
NPHI	0.468	−0.478	0.159	−0.461	0.226	−0.632	−0.112	−0.733	−0.214	−0.550	−0.501	−0.360	0.232	−0.039	0.518
RT	0.130	−0.578	−0.743	−0.435	−0.320	−0.394	−0.273	−0.379	−0.491	−0.361	−0.344	0.853	−0.844	−0.442	0.297
DT	0.621	0.091	0.270	0.011	0.823	−0.148	0.495	−0.522	0.338	−0.487	0.052	−0.332	0.402	−0.529	0.281
Variability	33.7%	29.7%	18.2%	47.5%	22.9%	17.4%	40.0%	26.5%	15.0%	61.1%	13.5%	11.5%	59.4%	21.4%	9.5%

Note: The percentage of the total variability accounted for by each PC is shown in %.

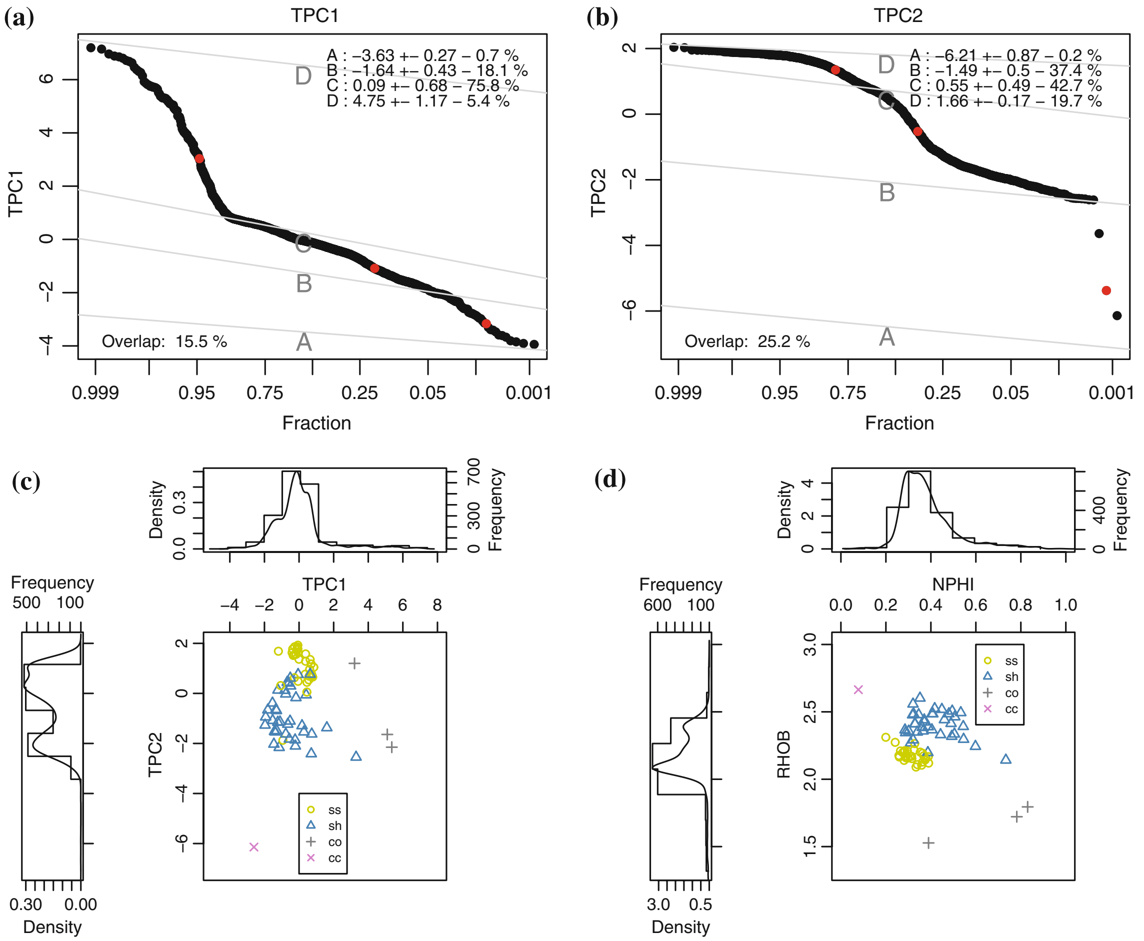

Figure 4. Analysis of the unstructured PCA. Polymodal distributions are indicated for the two first PCs. For visualization purposes only every fifth point is plotted in the crossplots. (a) TPC1 has four populations, where B and C populations contain 94% of the records, and (b) TPC2 also indicates four populations. (c) The TPC1-TPC2 crossplot illustrates that the TPC1 could not explain all major lithofacies variations itself as the sand–shale variations are discriminated by TPC2. (d) The crossplot of RHOB and NPHI wireline log variables has less discriminating power than the TPC1-TPC2 crossplot. In the crossplots, only every twentieth record is displayed.

of impurities (Fig. 5e). The PC2_co also contain four sub-populations, indicating coal–sand (B) and coal–shale (C) relations (Fig. 5f). The cemented interval outline four sub-populations of PC1_cc (Fig. 5g), where the A population represents records from the middle part of a 2 m cemented interval, the B population is related to the rim of this interval. The C–D populations are related to cement records influenced by nearby lithology types. For the PC2_cc (Fig. 5h), also interpreted to have four sub-populations, the A population is related to siderite cement, whereas the D population represents the middle part of the 2 m cemented interval. The B and C populations are assumed to be influenced by the nearby lithology types.

As the different PC within a specific PCA are independent of each other, crossplots are introduced to show how the sub-populations of the PC scores are interacting. The crossplot of the structured PCA, PC1_ss, and PC2_ss shows the internal variations of the 867 records representing the sandstone lithological unit, where clean sand is plotted in the right part of the diagram, while GR-rich sand, silt and coal influenced records are plotted in the left, lower left, and upper parts, respectively (Fig. 6a). The trend lines illustrate the intra-lithology variations. The populations of the PC1_ss and the PC2_ss generate a more precise description of the within sandstone lithological variations than the unstructured PCA can provide. The crossplot of the shale

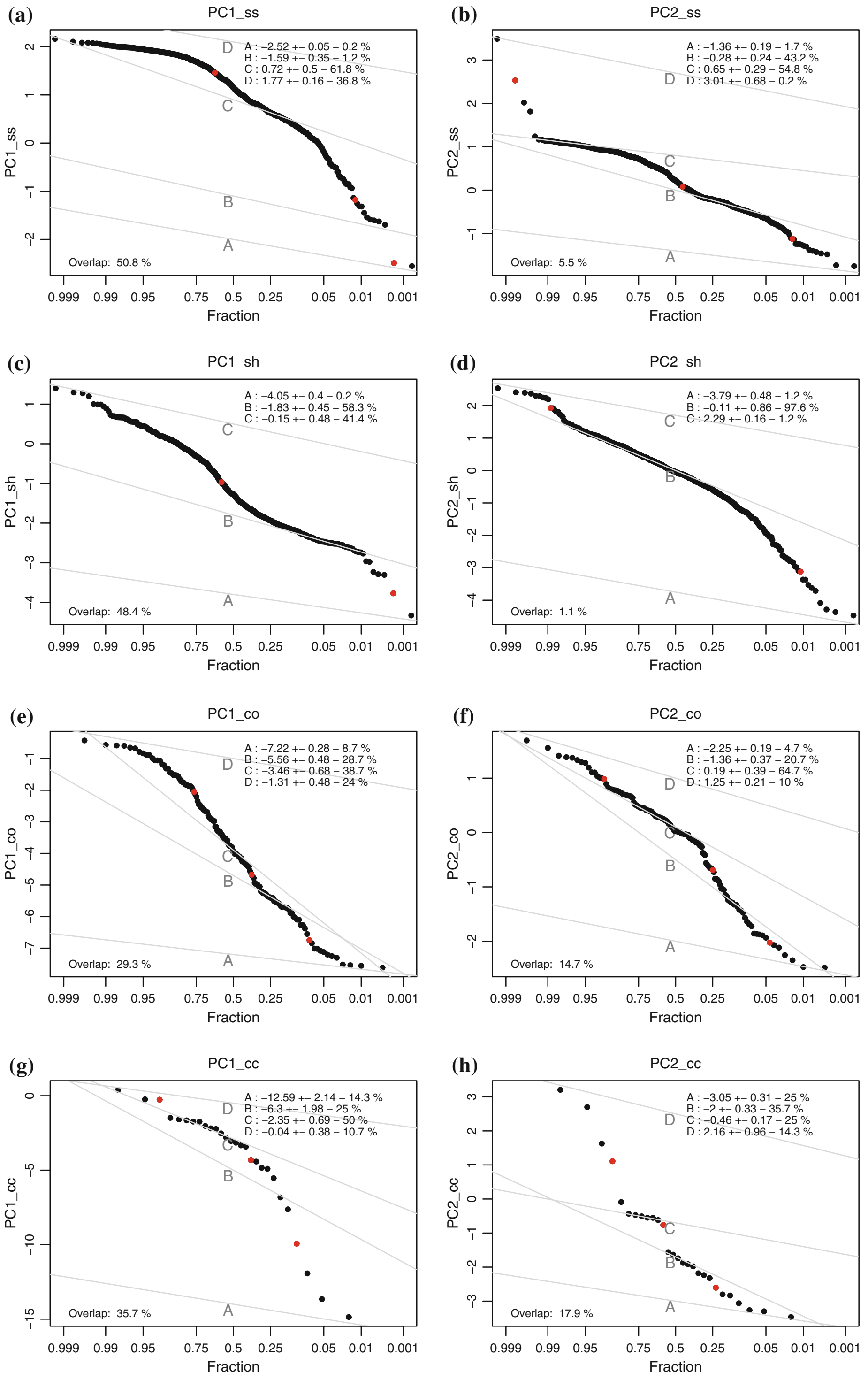

Figure 5. Separate plots displaying probability plots with interpreted populations of the first two PCs for each of the four separate lithofacies analysis performed in this structured PCA; sandstone (ss), shale (sh), coal (co), and cement (cc), respectively.

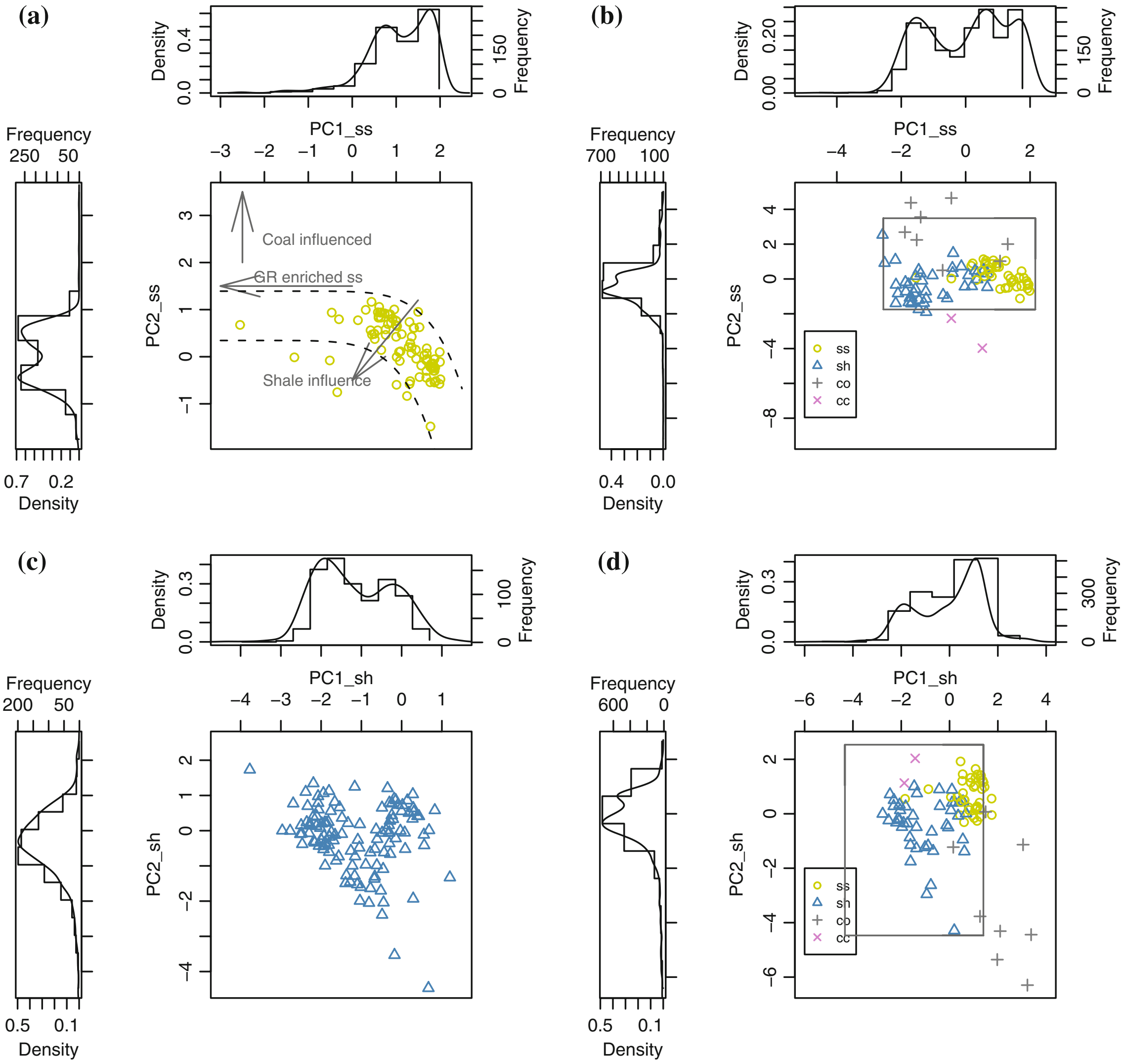

Figure 6. Crossplots of the first two PCs of the (a) sandstone and (b) shale lithological units, including crossplots where the loadings of each specific lithological unit are applied to the entire study interval (c and d) to illustrate the difference between the structured PCAs. In (a) and (b) only every tenth record is displayed, whereas (c) and (d) display every twentieth record.

lithological units indicate that PC1_sh separate sand–shale variations and the low PC2_sh scores outline coal influenced shale (Fig. 6b). The interpretation of the crossplot of the coal records show that low PC1_co scores represent organic-rich coal, whereas the PC2_co discriminates between sandstone without impurities and shale influenced sandstone (Fig. 7a). The crossplot of the cemented interval separates both cement types and the thickness of the cement interval that is not discovered by the unstructured PCA (Fig. 7b). The internal variations within the specific lithological units give a more precise picture of the intra-lithological variability than the unstructured PCA.

The loadings of the two first PCs of each separate lithological unit are separately applied to calculate new wireline records to visualize how the specific score values used to explain within lithological variations will perform when applied to the entire study interval. These new PC scores covering all records of the study interval help to determine the variations. The crossplot of PC1_ss and PC2_ss scores include all lithologic units using the sandstone lithological loadings. This sandstone view allows us

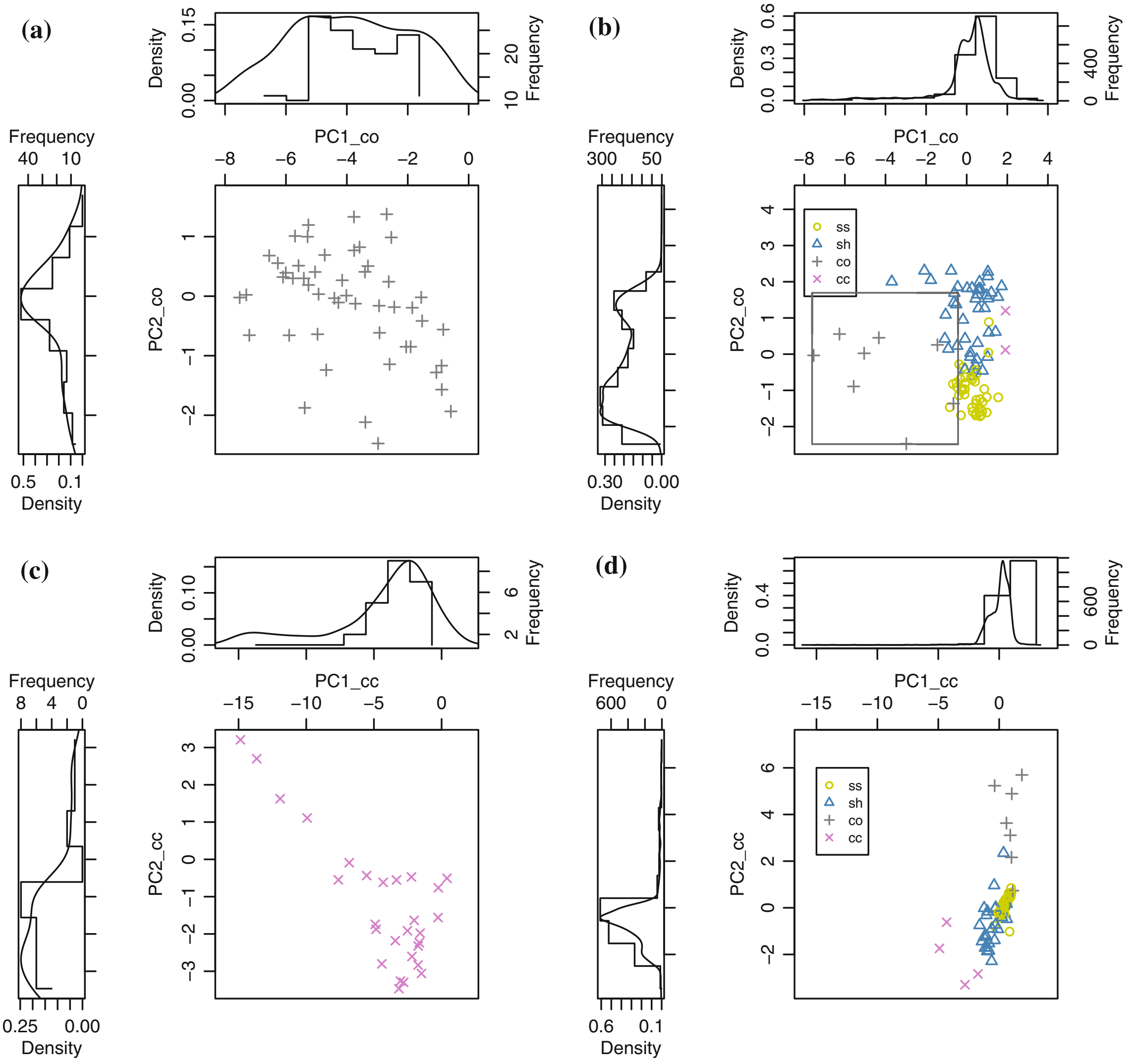

Figure 7. Crossplots of the first two PCs of the (a) coal and (b) cement lithological units, including crossplots where the loadings of each specific lithological unit are applied to the entire study interval (c and d) to illustrate the difference between the structured PCAs. In (a) and (b) only every tenth record is displayed, whereas (c) and (d) display every twentieth record.

to differentiate between cemented intervals, variations in coal and shale records in addition to the shale and within sandstone variations (Fig. 6c). Similar crossplot using shale loadings (Fig. 6d) illustrates how sandstone can be discriminated from shale as well as displaying gradations of shale variation including cemented and coal intervals. The crossplot of PC1_co and PC2_co differentiates between coal and other lithologies, including coal quality along the x-axis and sand influence along the y-axis (Fig. 7c). Similar crossplot of cemented loadings (Fig. 7d) indicates that the 2 m thick cemented interval has its own signature compared to the other cemented records plotted along the sand–shale–coal line. This result shows that applying structured PCA and later using these specific PC loadings to include all study interval records can go beyond the interpretation of both univariate and unstructured PCA when the separation of petrophysical variations are in focus. In order to ensure the representative of the computed eigenvectors, the sandstone (ss) subset was divided randomly into two groups. The loadings for these two subset groups of eigenvectors were compared with the initial sandstone subset and the results show that only loadings between −0.1 and 0.1 give percentile variation above 10% (Table 3). This test shows that there is stability in the eigenvectors.

Table 3. The Stability in Eigenvectors was Tested by Selecting at Random Half of the Samples within the Sandstone Lithofacies

	PC1_ss1	PC2_ss1	PC3_ss1	PC1_ss2	PC2_ss2	PC3_ss2	PC1_ss01	PC2_ss01	PC3_ss01	PC1_ss02	PC2_ss02	PC3_ss02
GR	−0.744	0.079	0.661	−0.739	0.269	0.612	−0.10%	18.40%	−0.61%	0.07%	−11.14%	1.31%
RHOB	−0.219	−0.365	−0.157	−0.224	−0.368	0.006	0.23%	0.54%	−17.38%	−0.34%	0.34%	58.57%
NPHI	−0.457	0.312	−0.601	−0.465	0.128	−0.653	0.22%	−7.99%	1.26%	−0.22%	13.84%	−0.82%
RT	−0.437	−0.261	−0.419	−0.434	−0.381	−0.355	−0.11%	5.08%	−1.54%	0.06%	−4.35%	2.60%
DT	0.009	0.834	−0.038	0.014	0.794	−0.270	5.00%	−0.33%	29.57%	−6.00%	0.90%	−14.59%
Variability	46.9%	23.2%	17.3%	48.1%	22.8%	17.3%						

Note: Columns 2–7 outline loadings for the two specific halves and show only marginal difference. Columns 8–13, expressing the percentile loading variation between all sandstone samples (PC_ss in Table 2), and the two specific halves, respectively, show that only loadings between −0.1 and 0.1 give percentile variation above 10%.

Comparison of Unstructured and Structured PCA Loadings

The difference in loading values, including their ability to explain the total data variability, is distinct when comparing unstructured PCA and the four separate structured PCAs (Table 2). The bar plot (Fig. 8a) shows that the first two PCs of the unstructured PCA explain less of the total variability than the structural PCAs. This implies that unstructured PCA uses a correlation matrix that has less strong correlations due to a larger part of heterogeneity from inter-lithological variations. The separate structured PCA analyzes specific lithological units avoids interactions from intra-lithological variations.

The star diagrams (Fig. 8b) visualize the relation between the unstructured and structured PC loadings expressed in Table 2. The PC1_ss has about identical loadings as TPC2, indicating that TPC2 represents the residual sandstone variability due to internal sandstone variations once the major lithofacies variability has been removed by TPC1. The similarity between PC2_ss and TPC1 shows that the residual variability once the intra-lithological sand variability is removed contains much of the same heterogeneity as shown in the totality of the well records. This indicates a sort of fractal behavior of the lithological mix at the Åre Fm. scale (300 m) and the scale of the combined sandstone layers (130 m). The TPC3 signature is related to the PC1_cc, indicating variation due to the cemented records. A second graphical visualization of the PC loadings in Table 2 is represented by Chernoff faces (Fig. 8c); these faces that mimic human faces are drawn based upon the loading values of the five wireline variables and can discriminate similar PC loading patterns and correspond to the results of the star diagrams.

Comparison of Unstructured and Structured PCA Scores

PCA can be regarded as a data-driven method because of the dependency between the position of the eigenvector and the gravity field of the samples. PCA can therefore give different results according to a specific selection of input variables and/or samples. It is therefore important to ensure that as much of the unwanted heterogeneity is removed by including a proper choice of samples representative for each lithological subset. Similarities in PCA loadings of unstructured and structured PCA can either be related to pure luck or, if properly designed, driven by specific geologic phenomena. In the following, the relationship between unstructured and structured PCA is illustrated by plotting the associations in crossplots. In this paper, only differences in eigenvector loadings between the structured and unstructured approach for each subset are considered. However, Figure 9 portrays how the individual score values of the unstructured PCA in the sandstone subset match the scores calculated with a structured correlation matrix based only on the subset samples. The similarity of the sandstone records of TPC1 and PC2_ss (Fig. 9a) could give the impression that the total unstructured analysis is as good as obtained with structured loadings, but this is only a consequence of the difference in the loadings for GR and RT being cancelled out because of close to zero standardized values in the sandstone for the wireline logs and similar loadings for RHOB, NPHI, and DT resulting in an alignment along the pure sand–shale trend line. The TPC2-PC1_ss plot (Fig. 9b) shows an alignment along a coal–sand–shale–cement trend along the structured PC1_ss vector. The TPC2-PC1_ss plot shows the close correlation between the variables with a marginal

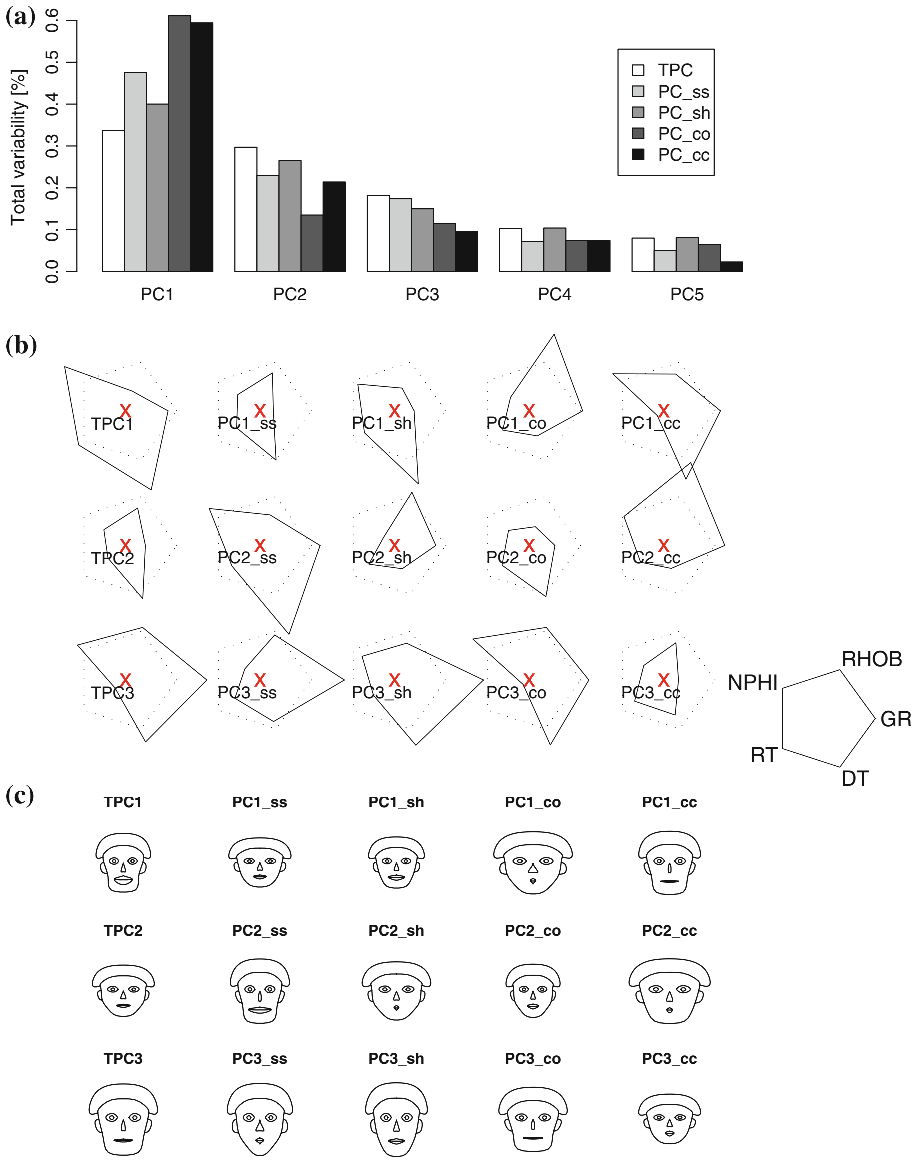

Figure 8. Separate plots displaying and comparing unstructured and structured PCA loadings and their magnitude of variability visualized from the data in Table 2. (a) The bar plot explains the total variability of each PC both for unstructured and structured PCA. (b) The star diagrams show that each PC loading has its own specific signature that can be compared to other loadings. The most prominent loadings are easily identified by their high negative or positive values confirming that similar loadings have different PC rank. Zero loading is plotted at the dotted line and high negative loading is at the center point. The dotted line of the diagram is where PC loadings are zero. (c) Another visualization of the PC loadings is Chernoff faces, which use faces to display five variables in one plot; GR, height of face; RHOB, width of face; NPHI, shape of face; RT, height of mouth; DT, width of mouth.

difference in the lower values interpreted to be related to GR-rich sandstone records. The sandstone records deviating perpendicular to the trendline is interpreted to be related to larger standardized values of RT. In the two crossplots, the two modes of each of the PCs, illustrated by the gray lines, express the similarity between the populations and show that the structured PCA modes have a wider separation, even if the gross lithology relation is similar.

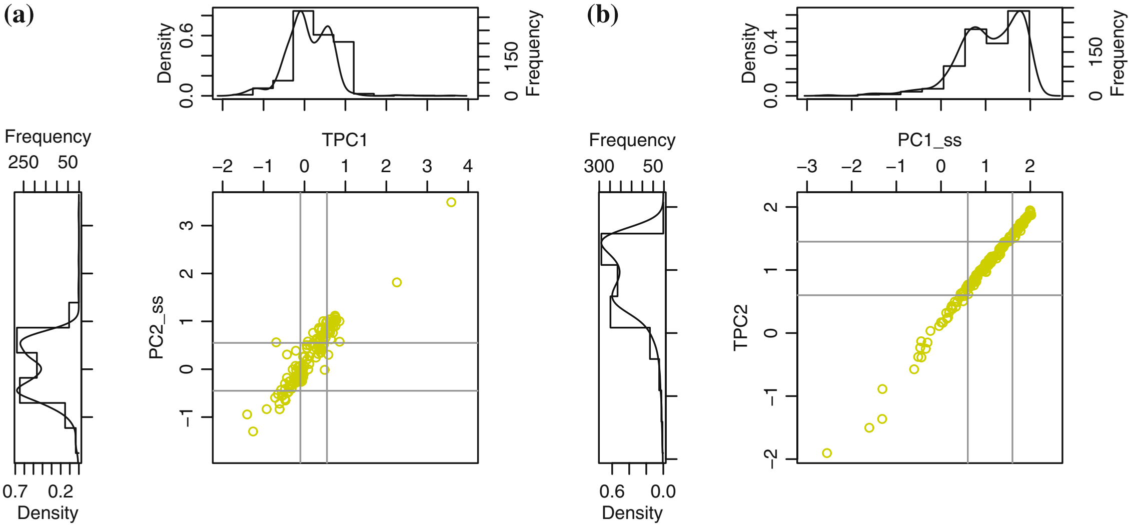

Figure 9. The crossplots of only sandstone lithology units of TPC and PC_ss illustrate the difference between unstructured and structured PCA with close to similar loading signatures. (a) The GR-rich sandstone records are classified as shale influenced sand by TPC1, whereas it is within the channel sand population of PC2_ss. (b) The crossplot of TPC2 and PC1_ss indicates similar scores for values over zero interpreted to be channel and bayfill sandstone, whereas there is a slight difference in the negative scores. The difference in the negative scores is related to GR-enriched sandstone and shale influenced sandstone. Only every tenth record is displayed.

The near perfect correlation between the TPC2 and PC1_ss scores indicates that the largest contribution to the total variability captured by TPC1 comes from overall shale, coal, cement vs. sandstone contrasts and that the residual variability captured by TPC2 reflects the dominant intra sandstone variability portrayed by PC1_ss. The scatter from the less perfect correlation between TPC1 and PC2_ss is an indicator of the fractal nature of the variability where the dominant shale, coal, cement vs. sandstone contrasts in a fractal way is representative both by the gross 300 m Åre Fm. interval as well as residual sandstone variability, PC2_ss, once the dominant intra-sandstone variability is removed.

In the TPC1-PC2_ss crossplot, the deviating samples perpendicular to the general trendline are related to minor lithology variations interpreted to be caused by extreme GR-enriched sandstone (>200 API units). Even if these points are related to the GR-enriched population of TPC1, these points fall within the two modes of the structured PCA, PC2_ss. This indicates that the GR-enriched sandstone variations are entirely captured by the principal structured PC, PC1_ss, whereas for the unstructured PCA both the two first PCs are needed to express this phenomenon.

A more in-depth analysis of the structured approach where the original log responses are recalculated based upon differences in heterogeneity will be published in a separate paper that is based upon the preliminary results presented in Brandsegg, Hammer, and Sinding-Larsen (2008).

Comparison Between Univariate and Multivariate Overlap

The distance between the mean value of each sub-population can indicate their separation. An overlap criterion is introduced to evaluate the difference between univariate, unstructured, and structured PCA. The population overlap between each component is determined by the percentage of data which falls between the mean plus two standard deviations of the lower population and the mean minus two standard deviations of the upper population (Stanley and Sinclair, 1988). The information conveyed in Table 4 shows that there is a marked difference in the amount of overlap between the primary and secondary PCs for the unstructured and structured PCA. The unstructured PCA has little overlap between populations because the variability is spanning the full variability space and thereby

Table 4. Comparison of Component Population Overlaps Including the Number of Populations of All Initial Wireline Log Variables and the Most Significant PCs of Both Unstructured and Structured PCA

Variable	Populations	Percent Overlap (%)
GR	6	12.5
RHOB	4	11.3
NPHI	4	0.1
RT	3	0.9
DT	5	10.7
TPC1	4	15.5
TPC2	4	25.2
TPC3	4	67.8
TPC1_ss	4	7.6
TPC2_ss	3	56.2
PC1_ss	4	50.8
PC2_ss	4	5.5
PC3_ss	3	0.7
PC1_sh	3	48.4
PC2_sh	3	1.1
PC1_co	4	29.3
PC2_co	4	14.7
PC1_cc	4	35.7
PC2_cc	4	17.9

identifies end member populations focusing on inter-population variability rather than intra-population variability. The increase in the degree of overlap with higher order PCs reflects the increasing compactness of the variability space and hence the increasing overlap of the lithological populations. The structured PCs show the opposite trend whereby the first PC displays a large overlap between what is now differences within the lithological population due to an expansion of internal heterogeneity in the respective lithological unit. These observations permit us to break the apparent uniform lithological population defined from the unstructured into sub-populations reflecting the local petrophysical contrast within the lithological unit.

Comparison Between Two Lithofacies Classifications

In the previous analysis, with four lithofacies classifications related to rock types that have been separately calculated, an increased separation between each of the lithofacies was achieved. In order to portray the effect of sedimentary features related to depositional environment, a new lithofacies classification was introduced, following the work of Kjærefjord (1999) and Hammer, Mørk, and Næss (2009). The new lithofacies types, predominantly based upon core analysis, were segmented into four sedimentary features related to deposition environment: fluvial channel (FCH), floodplain fines (FF), sandy bay-fill (SBF), and muddy bay-fill (MBF). The RHOB/NPHI crossplot outlines the high and low RHOB values of cemented and coal influenced intervals, in addition to portraying an overlap of the sandy and muddy bay-fill deposition feature populations (Fig. 10a). The overlapping of sandy and muddy bay-fill is related to the highly heterolithic deposition of a bay-fill environment which is difficult to differentiate (Svela, 2001; Hammer, Mørk, and Næss, 2009). The structured sandstone PCA loadings, indexed by the four sedimentary features, allow considerable additional differentiation to be mapped out, which otherwise would have been missed (Fig. 10b). When applied to all interval records studied, the first PC, PC1_ss, separates two populations of sand and one population of shale, whereas the PC2_ss separates two sandstone populations and the end-members of coal and cemented records. The crossplot of these two PCs points to two sandstone populations related to FCH and SBF, with a more pronounced separation than the initial NPHI/RHOB crossplot. The PC2_ss separates FCH and SBF, whereas PC1_ss explains the internal variations within these records. An increased separation between SBF and MBF is generated when applying the two PCs of the structured shale PCA (Fig. 11), as the SBF records are clustered, surrounded by the MBF records.

The structured PCA crossplot is superior to the initial wireline log responses when focusing on specific variations within a specific depositional setting. The calculations and graphical visualization of data using loadings expressing variations in the sandstone population has enhanced the differentiation between the different depositional environments without interfering with the other specific lithologies, such as coal and cement influenced intervals.

DISCUSSION

A basic requirement for using multivariate analysis on geologic data is to reflect on the quantification procedure measuring the geologic processes that constitute the input data for your analysis (Davis, 2002). By the use of PCA, analyzing patterns within the data aim to translate geologic objects that are described by a set of indirect information (e.g. wireline logs) into categorical information, which refers to a given geologic property (e.g. lithology

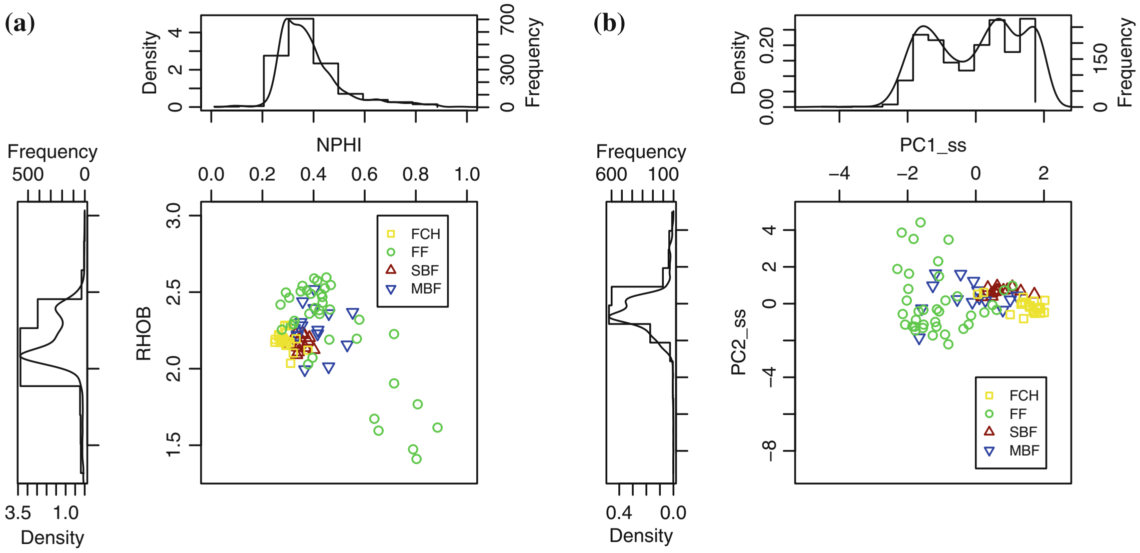

Figure 10. Comparison of initial wireline logs, NPHI and RHOB, and the two first PC scores, PC1_ss and PC2_ss, calculated by the structured sandstone intervals with lithofacies types classified according to Hammer, Mørk, and Næss (2009). There is more distinct separation between the fluvial channel (FCH) and sandy bay-fill (SBF) when structured PCA is applied. Only every tenth record is displayed.

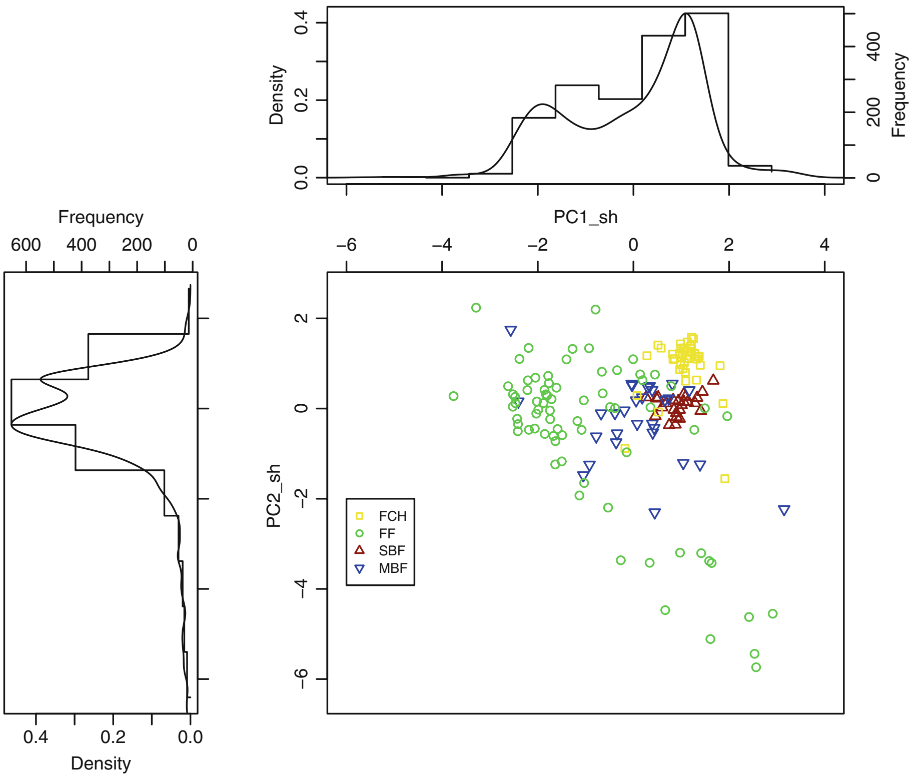

Figure 11. Crossplot of the two first PC scores, PC1_sh and PC2_sh, calculated by the structured PCA of the shale intervals with a lithofacies types classified according to Hammer, Mørk, and Næss (2009). There is more distinct separation between the sandy bay-fill (SBF) and muddy bay-fill when structured PCA using only shale records are applied. Only every tenth record is displayed.

type and porosity). In this study, petrophysical wireline responses in combination with core analysis have outlined four lithofacies types to be separately evaluated to determine their independent signatures that can express geologic processes that operate within lithofacies scale. It should therefore be noted that the lithofacies types used here are based upon manual lithofacies classifications interpreted from cores, supplemented by wireline log analysis in noncored intervals, and not an automatic pattern recognition identification of lithofacies types. However, our aim has been to identify the merit of using specific lithofacies weights to explain variations within specific lithofacies and to show how these weights can be used to enhance lithofacies interpretation. The populations identified from the initial wireline logs do not independently outline specific rock types and/or lithological processes in contrast to the unstructured/structured PCA procedure. The different populations in the probability plot analysis have the potential to have large overlaps if the PCs are polymodally distributed with component population means of roughly equal magnitude, but with large standard deviations. Ignoring discernable univariate patterns in the design of subsequent multivariate analysis may lead to unnecessary ambiguities and/or complexities in the multivariate results and the subsequent interpretation (Stanley and Sinclair, 1988). New variables calculated from structured PCA can be powerful discriminators for visualizing within lithofacies signatures that are not achievable with a standard PCA approach. Griffiths (1988) stated that a question asked may be unanswerable within the system in which it is formulated and to solve such a problem it is necessary to enlarge the system, creating a meta-language, and find the solution, if any, within this enlarged system. In this study, the reasoning of Griffiths is applied by using separate loadings from lithological units as a form of meta-language to explain within lithological contrasts. This is further exemplified by the principal structured PCA sandstone loading that had close to identical loading of the second unstructured PC: Despite the fact that the unstructured PCA included other lithofacies wireline responses, its similarity to sandstone processes could not be identified prior to the structured PCA.

The limitation scale for this study is related to the sampling interval of the wireline logs. Even though the sample interval is 15 cm, some of the different wireline log measuring tools can have larger distance between transponder and receiver resulting that not small-scale heterogeneities are captured. Nordahl and Ringrose (2008) concluded that, by using the representative elementary volume (REV) concept as a basis, it is important to incorporate lamina scale (mm) and lithofacies scale (dm-m) heterogeneities into full field reservoir scale heterogeneity to reduce uncertainty in reservoir modeling. This structural PCA approach has been applied on wireline logs to enhance the separation of petrophysical contrasts in fluviodeltaic deposits and support the estimated lithofacies REV around 20 cm lengths stated by Nordahl and Ringrose (2008).

CONCLUSIONS

We have evaluated separate PCAs derived from different lithological subsets of the well records to detect and interpret for higher order heterogeneity within the different lithologies. This procedure has allowed us to gain a clearer and more comprehensive interpretation of the data than by use of traditional PCA procedures. A case study analyzing higher order lithological effects from the fluviodeltaic environment of the Heidrun Field, offshore mid-Norway, has indicated that our ability to map and interpret higher order variability will improve the fluviodeltaic reservoir heterogeneity description that is important for production scheduling. The structured/unstructured PCA method adds to the standard interpretation of the wireline log data by identifying specific intra-lithological processes that are not outlined by traditional approaches. This workflow can easily be applied to isolate other depositional environments and is assumed to be particularly valuable in other studies involving heterolithic deposits. The structured/unstructured PCA method permits the effective removal of variability due to gross lithological effects and allows for differential interpretation of heterogeneity. This procedure can further be applied into lithofacies classification routines for incorporating small-scale heterogeneities that potentially can be used to decrease the misclassification records. The use of separate PCAs through the examples given has been effective in portraying petrophysical variability of reservoir properties within different lithological units. We suggest that this procedure should be used to pre-process effective reservoir properties in order to enhance the choice of reservoir drainage strategies.

ACKNOWLEDGMENTS

Heidrun Unit (Statoil Petroleum AS (operator), Petoro AS, ConocoPhillips Skandinavia AS, Eni Norge AS) is acknowledged for providing well data and permission to present case examples from the Heidrun Field. This paper represents a contribution to the GeoEnhance consortium on reservoir characterization at the Norwegian University of Science and Technology (NTNU). The research stems from the post-graduate work of K. B. Brandsegg and E. Hammer at the Faculty of Engineering Science and Technology at NTNU. Special thanks to Arve Næss and Mali Brekken at Statoil and Steinar Ellefmo for helpful comments. Two anonymous reviewers gave valuable comments which greatly improved this paper.

REFERENCES

Avseth, P., Mukerji, T., and Mavko, G., 2005, Quantitative seismic interpretation: applying rock physics tools to reduce interpretation risk: Cambridge University Press, Cambridge.

Bourquin, S., Rigollet, C., and Bourges, P., 1998, High-resolution sequence stratigraphy of an alluvial fan-fan delta environment: stratigraphic and geodynamic implications—an example from the Keuper Chaunoy sandstones, Paris basin: Sed. Geol., v. 121, no. 3–4, p. 207–237.

Brandsegg, K. B., Hammer, E., and Sinding-Larsen, R., 2008, Quantifying fluvial sandstone heterogeneity by using multivariate analysis, *in* Sirum, H. J. H., and Haukdal, G. K., eds., NGF Abstracts and Proceedings of the Geological Society of Norway: Stavanger, Norway, p. 7–9.

Bridge, J. S., and Tye, R. S., 2000, Interpreting the dimensions of ancient fluvial channel bars, channels, and channel belts from wireline-logs and cores: AAPG Bull., v. 84, no. 8, p. 1205–1228.

Chernoff, H., 1973, The use of faces to represent statistical association: J. Am. Stat. Assoc., v. 68, p. 361–368.

Corbett, P., Jensen, J., and Sorbie, K., 1998, A review of up-scaling and cross-scaling issues in core and log data interpretation and prediction, *in* Harvey, P., and Lovell, M., eds., Core-Log Integration. Vol. 136 of Geological Society Special Publication, 136: Springer, London, p. 9–16.

Dalgaard, P., 2008, Introductory Statistics with R (2nd edn.): Springer, London.

Dalland, A., Augedahl, H., Bomstad, K., and Ofstad, K., 1988, The post-Triassic succession of the Mid-Norwegian Shelf, *in* Dalland, A., Worsley, D., and Ofstad, K., eds., A Lithostratigraphic Scheme for the Mesozoic and Cenozoic Succession Offshore Mid- and Northern Norway. Vol. 4, Norwegian Petroleum Directorate Bulletin: Springer, Stavanger, p. 5–42.

Davis, J. C., 2002, Statistics and data analysis in geology: Wiley, NewYork.

Doveton, J. H., 1994, Geological log analysis using computer methods. Vol. 2, AAPG Computer Applications in Geology.

Eichenseer, H. T., and Leduc, J. P., 1996, Automated genetic sequence stratigraphy applied to wireline logs: Bulletin Des Centres De Recherches Exploration-Production Elf Aquitaine., v. 20, no. 2, p. 277–307.

Griffiths, J., 1988, Measurement, sampling and interpretation, *in* Chung, C. F., Fabbi, A. G., and Sinding-Larsen, R., eds., Quantitative Analysis of Mineral and Energy Resources. NATO ASI Series, 82: D. Reidel Publishing Company, Boston, p. 37–56.

Gupta, R., and Johnson, H. D., 2001, Characterization of heterolithic deposits using electrofacies analysis in the tide-dominated Lower Jurassic Cook Formation (Gullfaks Field, offshore Norway): Petrol. Geosci., v. 7, no. 3, p. 321–330.

Hammer, E., Mørk, M. B. E., and Næss, A., 2009, Facies controls on the distribution of diagenesis and compaction in fluvial-deltaic deposits. Marine Petrol. Geol. Corrected proof (in press). doi:10.1016/j.marpetgeo.2009.11.002.

Hohn, M. E., McDowell, R. R., Matchen, D. L., and Vargo, A. G., 1997, Heterogeneity of fluvial-deltaic reservoirs in the Appalachian basin: a case study from a Lower Mississippian oil field in central West Virginia: AAPG Bull., v. 81, no. 6, p. 918–936.

Jolliffe, I., 2002, Principal component analysis (2nd edn.): Springer, New York.

Kjærefjord, J., 1999. Bayfill successions in the lower Jurassic Åre formation, Offshore Norway: sedimentology and heterogeneity based on subsurface data from the Heidrun Field and analog data from the Upper Cretaceous Neslen Formation, eastern Book Cliffs, Utha, *in* Hentz, T., ed., 19th Annual Research Conference. Advanced Reservoir Characterization for the Twenty-First Century. Gulf Coast Section and Society Economic Paleontologists and Mineralogists Foundation, Special Publication, p. 149–157.

Martinius, A. W., Ringrose, P. S., Brostrøm, C., Elfenbein, C., Næss, A., and Ringås, J. E., 2005, Reservoir challenges of heterolithic tidal sandstone reservoirs in the Halten Terrace, mid-Norway: Petrol. Geosci., v. 11, no. 1, p. 3–16.

Moline, G. R., and Bahr, J. M., 1995, Estimating spatial distributions of heterogeneous subsurface characteristics by regionalized classification of electrofacies: Math. Geol., v. 27, no. 1, p. 3–22.

Nadler, M., and Smith, E. P., 1993, Pattern recognition engineering: Wiley, New York.

Nordahl, K., and Ringrose, P. S., 2008, Identifying the Representative Elementary Volume for permeability in heterolithic deposits using numerical rock models: Math. Geosci., v. 40, no. 7, p. 753–771.

Pereira, H. G., Silva, A. C. E., Soares, A., Ribeiro, L., and Decarvalho, J., 1990, Improving reservoir description by using geostatistical and multivariate data-analysis techniques: Math. Geol., v. 22, no. 8, p. 879–913.

Petrovic, A., Khan, S. D., and Chafetz, H. S., 2008, Remote detection and geochemical studies for finding hydrocarbon-induced alterations in Lisbon Valley, Utah: Marine Petrol. Geol., v. 25, no. 8, p. 696–705.

Serra, O., and Abbott, H. T., 1982, The contribution of logging data to sedimentology and stratigraphy: Soc. Petrol. Eng. J., v. 22, no. 1, p. 117–131.

Singh, Y., 2007, Lithofacies detection through simultaneous inversion and principal component attributes: The Leading Edge, v. 26, no. 12, p. 1568–1575.

Stanley, C. R., and Sinclair, A. J., 1988, Univariate patterns in the design of multivariate analysis techniques for geochemical data evaluation, *in* Chung, C. F., Fabbi, A. G., and Sinding-Larsen, R., eds., Quantitative Analysis of Mineral and Energy Resources. NATO ASI series, 82: D. Reidel Publishing Company, Boston, p. 113–130.

Svela, K., 2001, Sedimentary facies in the fluvial-dominated Åre formation as seen in the Åre 1 member in the Heidrun Field, *in* Martinsen, O., and Dreyer, T., eds., Sedimentary

Environments Offshore Norway—Paleozoic to Recent. Vol. 10. Norwegian Petroleum Society Special Publication: Elsevier Science B.V., Amsterdam, p. 87–102.

Wegman, E. J., 1990, Hyperdimensional data-analysis using parallel coordinates: J. Am. Stat. Assoc., v. 85, no. 411, p. 664–675.

Zhang, G. F., Shen, X. H., Zou, L. J., Li, C. J., Wang, Y. L., and Lu, S. L., 2007, Detection of hydrocarbon bearing sand through remote sensing techniques in the western slope zone of Songliao basin, China: Int. J. Remote Sens., v. 28, no. 7–8, p. 1819–1833.