UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DATA-DRIVEN REAL-TIME GEOSTEERING USING SURFACE DRILLING DATA

A Thesis

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

Connor Holasek

Norman, Oklahoma

2019

DATA-DRIVEN REAL-TIME GEOSTEERING USING SURFACE DRILLING DATA

A THESIS APPROVED FOR THE
MEWBOURNE SCHOOL OF PETROLEUM AND GEOLOGICAL ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Deepak Devegowda, Chair

Dr. Chandra Rai

Dr. Ali Tinni

Acknowledgements

I would like to thank the members of my thesis committee, Dr. Deepak Devegowda, Dr. Chandra Rai, and Dr. Ali Tinni. I would like to give special thanks to Dr. Deepak Devegowda for his continued effort to help me further my education, and for his guidance and advice throughout my time in school.

I would also like to thank OU and the Mewbourne School of Petroleum and Geological Engineering (MPGE) for developing a proficient Petroleum Engineering program that harbors a highly academic environment to promote life-long learning.

I would like to acknowledge the research efforts from those involved with the OU-Marathon Oil research project and Marathon Oil for providing data used in this work. Efforts made towards rocktyping and geological modeling are utilized in this thesis, for which I am thankful for the work of other students involved with the project.

I would also like to acknowledge the efforts from Equinor to make the data set from the Volve Oilfield publicly available.

Lastly, I would like to thank my family and friends who supported and motivated me while pursuing my undergraduate and graduate degree.

Table of Contents

List of Figures

Abstract

In this thesis I present a method for estimating lithology or deriving formation properties from real-time surface drilling data. This information can then be used to enhance real-time geosteering capabilities. Current approaches for geosteering often rely on data from an MWD sub. Because of the position of the MWD sub relative to the bit, the MWD sub is relating information that is depth- and time-delayed relative to the bit. In this work, I use a data-driven approach that relies on the use of Hidden Markov Models (HMM) and Change Point Detection (CPD) algorithms to relate surface drilling signals to formation properties/lithology. My approach views the surface drilling data as a multichannel time-series signal which is an advantage over prior approaches that have attempted to derive variations in lithology from surface drilling data. I finally test my approach with two field datasets: the first is the Volve Oilfield in offshore Norway, and the second is the Meramec formation in the Oklahoma STACK. In both case studies, I demonstrate that the use of HMM and/or CPD can substantially enhance resolution of lithology changes from surface drilling data alone and can therefore be a promising approach for use in real-time geosteering decisions.

Chapter 1: Introduction

With automation and widespread sensor deployment, more data is being collected than ever before. The oil and gas industry is currently in a state of transition to incorporate workflows to harness the power of this data and to create actionable insights from the information content of the data. By capitalizing on recent developments in data mining and analytics, the industry continues to expand the power of data science to make meaningful impacts on day-to-day operations and improve economics. This chapter provides a brief overview of the applications of data-driven methods to the oil and gas industry.

1.1 Applications of Machine Learning in the Oil and Gas Industry

The SPE literature (www.onepetro.org) lists over 4000 papers on machine learning alone as of now and consequently, this review is not exhaustive. A few notable examples include Schuetter et al. (2015) who describes predictive modeling using regression to identify the impact of well construction, completions, and location on production performance for Wolfcamp wells. They also rank the input variables in terms of their impact and provide a list of variables that can be controlled by operators to improve well productivity. In another example, Bangert (2019) uses stochastic gradient-boosted decision trees to identify rod-pump dysfunction in real-time using dynamometer card which measures the load and position of the traveling valve in the subsurface assembly. He reports that dysfunction can be identified by abnormalities in the shape of the dynamometer card data when compared to that of a functional rod-pump. The machine learning described in the paper involves the identification of the statistical moments and geometric characteristics of the shape of the dynamometer card data and their relation to pump operating modes. Ounsakul et al. (2019) applies decision tree classification to select the optimal type of artificial lift, either rod-pump, gas

lift, electric submersible pump (ESP), or progressive cavity pump (PCP). In their study, the design considerations affecting artificial lift include well parameters such as depth, fluid properties (density and viscosity), surface facilities (power availability and pipeline proximity), and economic considerations. They conclude that an accurately trained model can provide suggestions for artificial lift to ensure long-term production enhancement. Machine learning has also been used to generate synthetic log data reducing the need for running additional costly logging suits. Tariq et al. (2016) have shown that Artificial Neural Network (ANN) can be used with triple combo log data to accurately predict compressional and shear wave velocity, and Yu et al. (2016) used triple combo, spectral gamma ray, compressional velocity, and photoelectric factor to estimate total organic carbon using Gaussian Process Regression. Both techniques proved to be more accurate than commonly used empirical correlations. In both examples, commonly available logs are used to generate more complex logs with high accuracy. Other applications include data-driven detection of factors governing well productivity by identifying sweet spots and hydraulic fracture locations (Gupta et al. 2019).

A related application of data analytics and machine learning to the work of this thesis is monitoring of artificial lift systems in real-time to detect and signal dysfunction to mitigate pump failure presented in Rensburg (2019). The repair or replacement of electrical submersible pump (ESP) assemblies can take a prolonged period of time and loss in production. ESPs are connected to monitoring systems that send data to a central control system. Traditionally, this data stream is monitored by trained personnel. Human intervention can be compromised by bias, subjectivity, and operator knowledge and is limited to monitoring only a few wells at a time and can present challenges for the analysis of large data streams (Rensburg 2019). This presents a great opportunity for predictive analytics to automate the ESP monitoring process. Rensburg (2019) describes the

idea of using deep learning to manage a large database of many ESPs deployed in the field for preventative surveillance to monitor pumps for abnormal behavior leading to failure. This type of monitoring system can be seen below in **Fig. 1**.

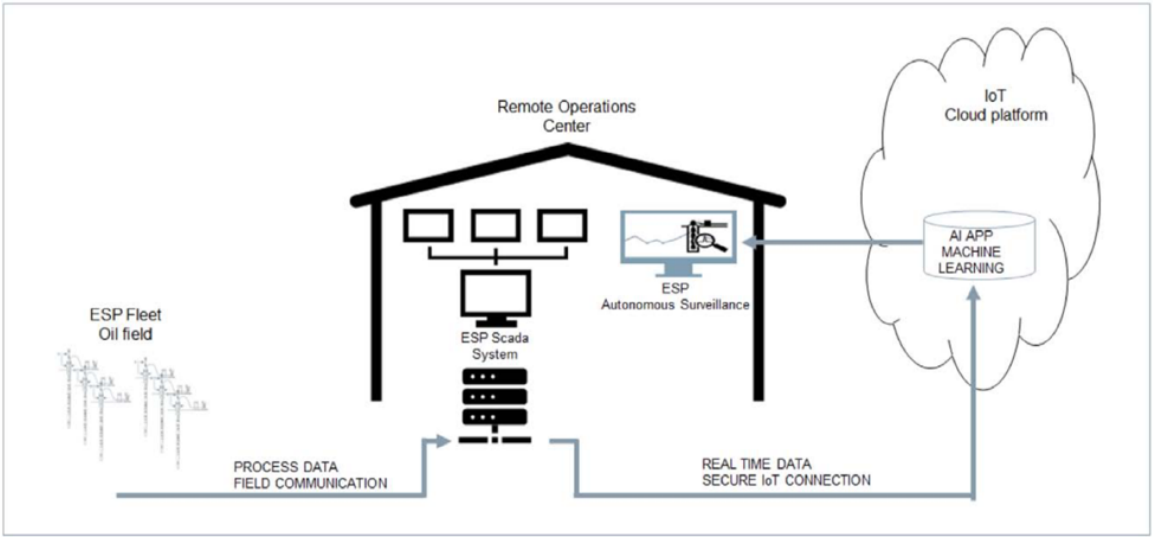

Fig. 1 - A model showing the ESP monitoring system utilizing real-time data collected from the field, and machine learning analytics to automate ESP surveillance. (Rensburg 2019)

In this model, data from the ESP fleet in the field is streamed to a remote operations center where the data is sent to a cloud platform to be analyzed by the deep learning algorithm that can detect ESP failure from anomalous behavior. If there is an impending failure, the trained model can send a signal to an operator to intervene and can therefore boost project efficiencies and reduce the overall need for well intervention.

1.2 Applications of Machine Learning Using Drilling Data

A considerable body of literature is also devoted to learning from drill bit-rock interactions. Most work focuses on training predictive models to relate drilling variables to formation properties to be utilized for future drilling projects. The work of Li (2019) describes machine learning

approaches for predicting the rate of penetration (ROP) to improve drilling efficiency. The work is motivated by the goal of reducing drilling costs by optimizing ROP and identifying causes for slowdowns. Drilling variables such as weight on bit (WOB), rotational speed (RPM), torque, hook load, flowrate, and standpipe pressure can help to describe bit performance and ROP. Many of these variables are used in analytical models such as the Bourgoyne model (Bourgoyne 1986) for prediction of ROP. These models are often limited in scope due to assumptions made within the model, as well as unknown downhole boundary conditions and can severely affect accurate ROP prediction. To overcome these limitations, Li et al. (2019) use an artificial neural network (ANN) to predict ROP using the variables listed previously. To predict ahead of the bit with ANN, a staggered layer approach of data input was proposed. To do this, the formation lithology is assumed to be the same for a short distance to predict ROP for a specific depth range ahead of the bit. The prediction interval ahead of the bit can be used as a tuning parameter to optimize the prediction quality. This data input scheme is shown in **Fig. 2**.

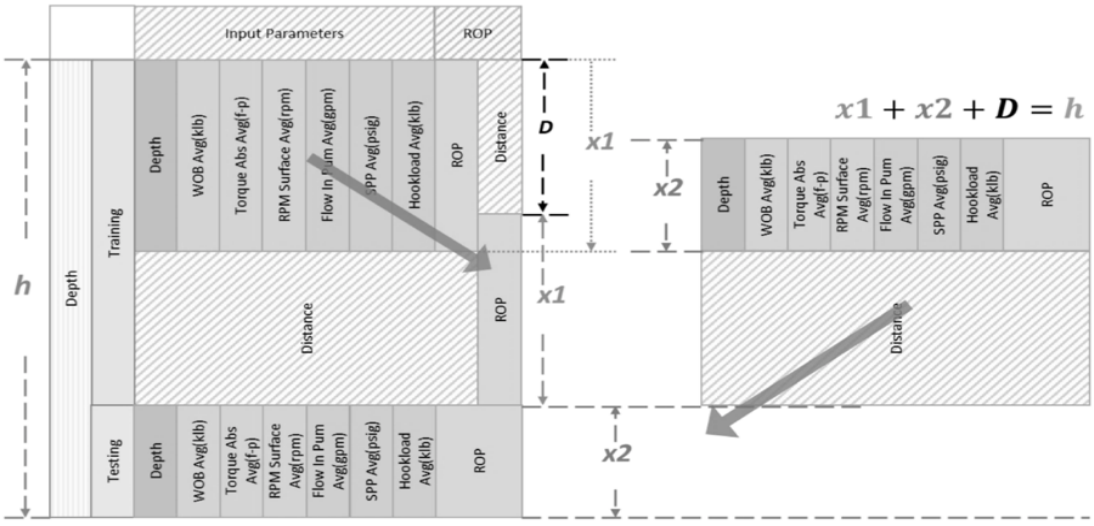

Fig. 2 - A schematic of the data entry procedure for looking ahead of the bit with ANN. In this diagram, the distance D is used as the prediction distance from which data from X1 and

X2 are used from the previous cycle of prediction. This distance D is a tuning parameter for the ANN to determine the greatest distance that the ANN can predict before predictions become unstable and inaccurate. (Li 2019)

Li (2019) applied three testing scenarios for prediction of ROP. One scenario uses downhole drilling data acquired from the measurement while drilling (MWD) alone; the other complements the MWD data with ROP data from intervals behind the bit, and the third approach complements the second approach with the addition of surface drilling data. Stability of the model is assessed based on the difference between predicted and actual ROP to determine a reliable distance for predictions ahead of the bit. The third scenario proved to be most accurate and was shown to be stable up to 40 ft. ahead of the bit. The model was able to achieve regression results of 0.91 but missed certain formation transitions seen in a comparison of predicted and actual ROP versus depth in **Fig. 3**. One of the drawbacks of their work is the assumption of a fixed lithology within the prediction interval that does account for the impact of the rocktype transitions on ROP.

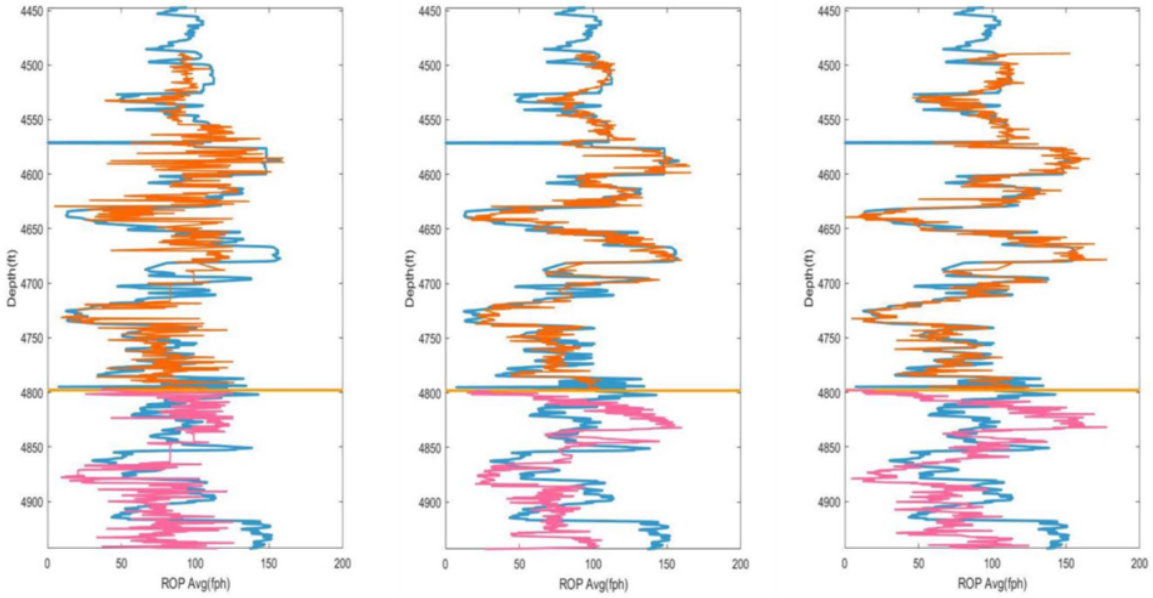

Fig. 3 - A figure showing the prediction of ROP as a function of depth, with the true ROP shown. The blue curve is the true ROP and the orange and purple are the ANN predicted

ROP's during the training and testing set respectively. Below the orange horizontal line is the transition from training to testing data. Towards the end of the log, the ROP has a sharp increase that is likely due to a bed boundary which is poorly identified by the ANN predicted ROP. (Li 2019)

In order to optimize ROP, bed boundaries need to be identified prior to drilling or in real-time while drilling. The different formations are often identified using seismic data, if available. Seismic velocities are used to build an earth model in the area for which the drilling trajectory will be planned (Mottahedeh 2008). This provides the geosteering team with information related to the expected formations and their corresponding depths. Due to the uncertainties and non-uniqueness of seismic earth models, identifying the correct depth for bed boundaries can be challenging (Osypov et al. 2011). Esmersoy et al. (2013) proposed a method of updating seismically derived earth models during drilling using checkshot measurements that periodically measure the velocity of the formation behind and ahead of the bit. After velocities are measured, the earth model can be updated to match the checkshot velocity data. By updating the earth model with more localized data from LWD checkshots, bed boundaries can more accurately be identified. Esmersoy et al. (2013) proposed using this method for fault location and pore pressure prediction ahead of the bit which can mitigate drilling risks and reduce the cost of drilling.

Lithology prediction from drilling data has also been a topic of interest in the industry. Identifying different lithologies during drilling operations can provide information to drillers about the well trajectory thereby enabling geosteering applications. Inoue et al. (2019) presents a method of identifying lithologies from MWD data using ANN. Lithology is determined after drilling using surface torque data to identify three lithologies: sand, silt-clay, and volcanic ash. Statistical moments (mean, standard deviation, skewness, kurtosis) and frequency content of the surface torque are used as features for model training. These features are correlated to lithology to help

the model learn to classify the various rocktypes. In their work, total classification accuracy achieved was 96% with a low number of false positives. However, when the rocktypes are more similar, such high prediction accuracies may not be possible. Inoue et al. (2019) also presents an application in another scenario with seven relatively similar lithologies where the classification accuracy is observed to be much lower, and the number of false positives is relatively higher.

Romanenkova et al. (2019) use a change point detection algorithm to identify sand and shale sequences, and the transitions from one rocktype to another. They use a gradient boosted decision tree as a classifier to identify rocktypes from well log data. A change point detection algorithm is applied to drilling data (ROP, WOB, torque) and the occurrence of change points is compared to transitions between the rocktypes seen on the well log. Thin layers within the rocktype sequence along the lateral are "dropped" or replaced with the previous majority rocktype with the assumption that the layers would be too thin for the change point detection algorithms to detect. The thickness of the thin beds is used as a tuning parameter for the change point detection algorithm and is selected as layers 1.5m or smaller. They use three change point detection algorithms: cumulative sums (Page 1954), Shiryaev-Roberts statistics (Roberts 1966), and posterior probabilities statistics (Smith 1975). The accuracy of the algorithm is measured by the number of predicted rocktype transitions divided by the number of true transitions. The rocktype transitions and the predictions from cumulative sums with and without the removal of thin layers for the prediction is shown in **Fig. 4**.

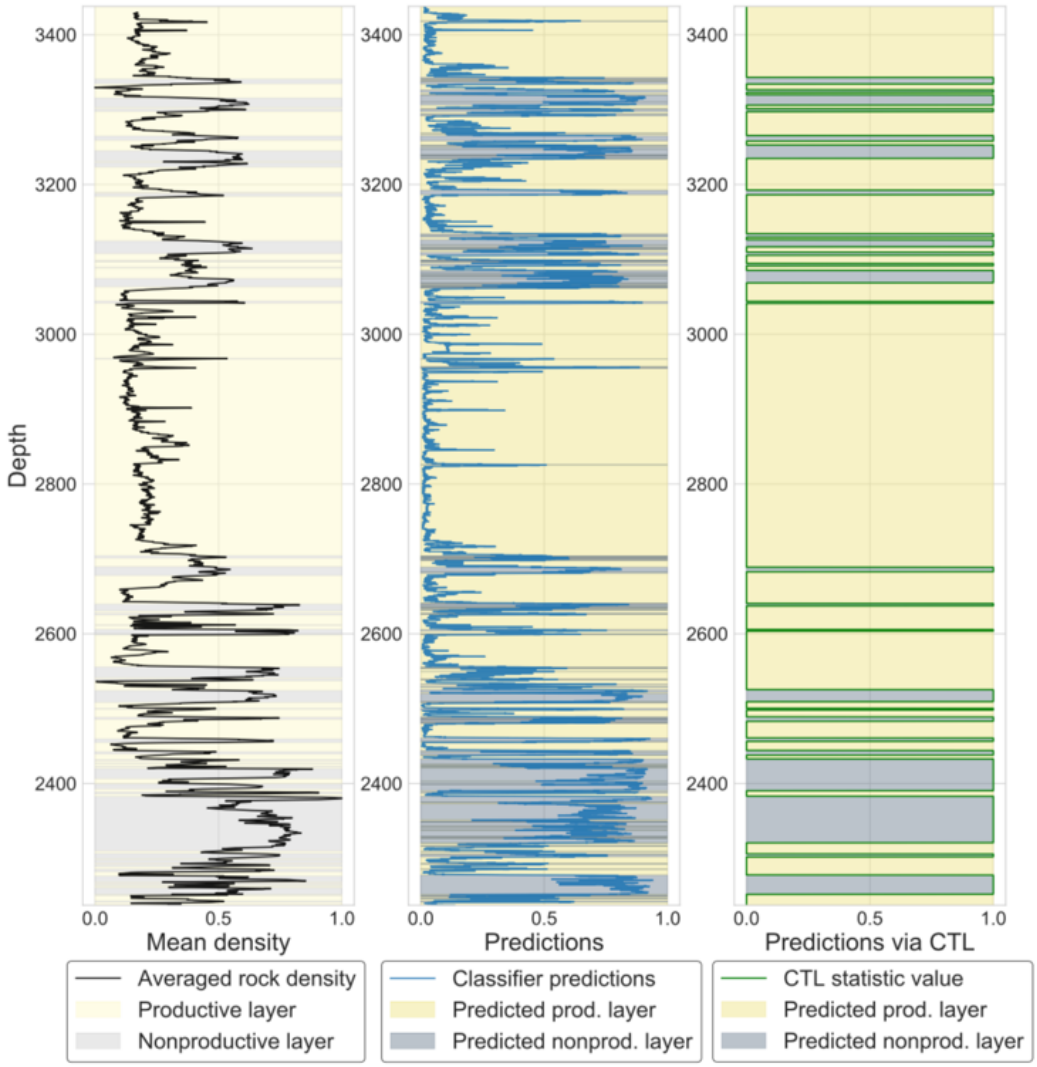

Fig. 4 - A log showing the results of the change point detection algorithm applied. The left most track shows a density log with the different rocktypes highlighted by the yellow and blue shading. The center track shows the change points detected on the data with the allowance of thin layers with the classifier probabilities. The right most track shows the results of the change point detection by cutting of thin layers (CTL). CTL reduces the rapid transitions between rocktypes and more closely matches the true classification. (Romanenkova et al. 2019)

The change point detection method for lithology identification is shown to be quite effective, especially with the removal of thin layers. In the far track log shown above, the predicted layers align well with the log data-derived layers shown in the far-left track.

The geosteering team is required to make decisions to adjust the position of the bit to stay within the target formation. Such decisions involve steering the bit up or down within the formation or making no change and following the current bit trajectory. These decisions are made in real-time and can be prone to error. Poor bit trajectory monitoring can lead the bit out of the target zone leading to contact with unproductive sections of the reservoir. Kullawan et al. (2017) describes an approach using discretized stochastic dynamic programming (DSDP) that allow the driller to keep a greater percentage of the well within the target zone. For this approach, decision stages are predefined along the planned wellbore trajectory. These decisions stages are influenced using estimates of the bed boundary depths and locations of faults and the uncertainties associated with each of these quantities. When the drill bit reaches stage i, the estimates of bed boundary depths are updated. The stage-to-stage transition probabilities are modeled as a Markovian process with a constant correlation coefficient. Therefore, the estimation of bed boundary depth at stage $i+1$ is updated by determining the difference between the estimated and true depth of the bed boundary at stage i. With these updates, decisions can be made about steering the bit up or down to stay within zone. In the event of drilling out of zone, the team will need to decide whether the bit should be redirected into the reservoir by steering, or whether a sidetrack is a better option. These decisions are modeled by the DSDP to provide guidance towards the best-case scenario. To facilitate these decisions, the model is rewarded by keeping the wellbore within the reservoir and is penalized by drilling out of zone using a metric called the expected monetary value (EMV). The EMV of the project is determined by the length of the wellbore within the reservoir. Wellbores with more reservoir contact have a higher potential to access hydrocarbons, and therefore are assumed to be more valuable. In contrast, making the decision to sidetrack a well will incur additional costs not originally accounted for. However, in some situations, the decision to sidetrack

a well to reach the target zone instead of re-steering may be more beneficial because of the possibility of having an additional length of wellbore in the target zone. Rapid decision making in these scenarios is challenging and Kullawan et al. (2017) demonstrate how the DSDP algorithm optimizes decision making in real-time.

1.3 Motivation

Formations are targeted for rock properties that are favorable for production or stimulation. It is of vital importance that the geosteering team ensure the bit stays within the target formation. Target formations are often identified from seismic data or offset wells and mapped in 3D space (Solomon et al. 1994). However, the resolution and uncertainties associated with these data types makes geosteering a challenging task. Geosteering teams are generally equipped with MWD sensors that measure formation properties during the drilling process to help guide decisions about steering to stay within zone. Often, gamma ray measurements are utilized to identify the target formation, and deviations away from expected gamma ray trends can indicate drilling out of zone (Mottahedeh 2008). While these systems do provide vital information to the geosteering team, there are challenges associated with the use of MWD data. The MWD sub is generally placed farther back, sometimes as much as 30-90 feet, behind the bit (Moore et al. 1998). **Fig. 5** shows a normal setup of a MWD sub showing the distances between the bit and the MWD sub.

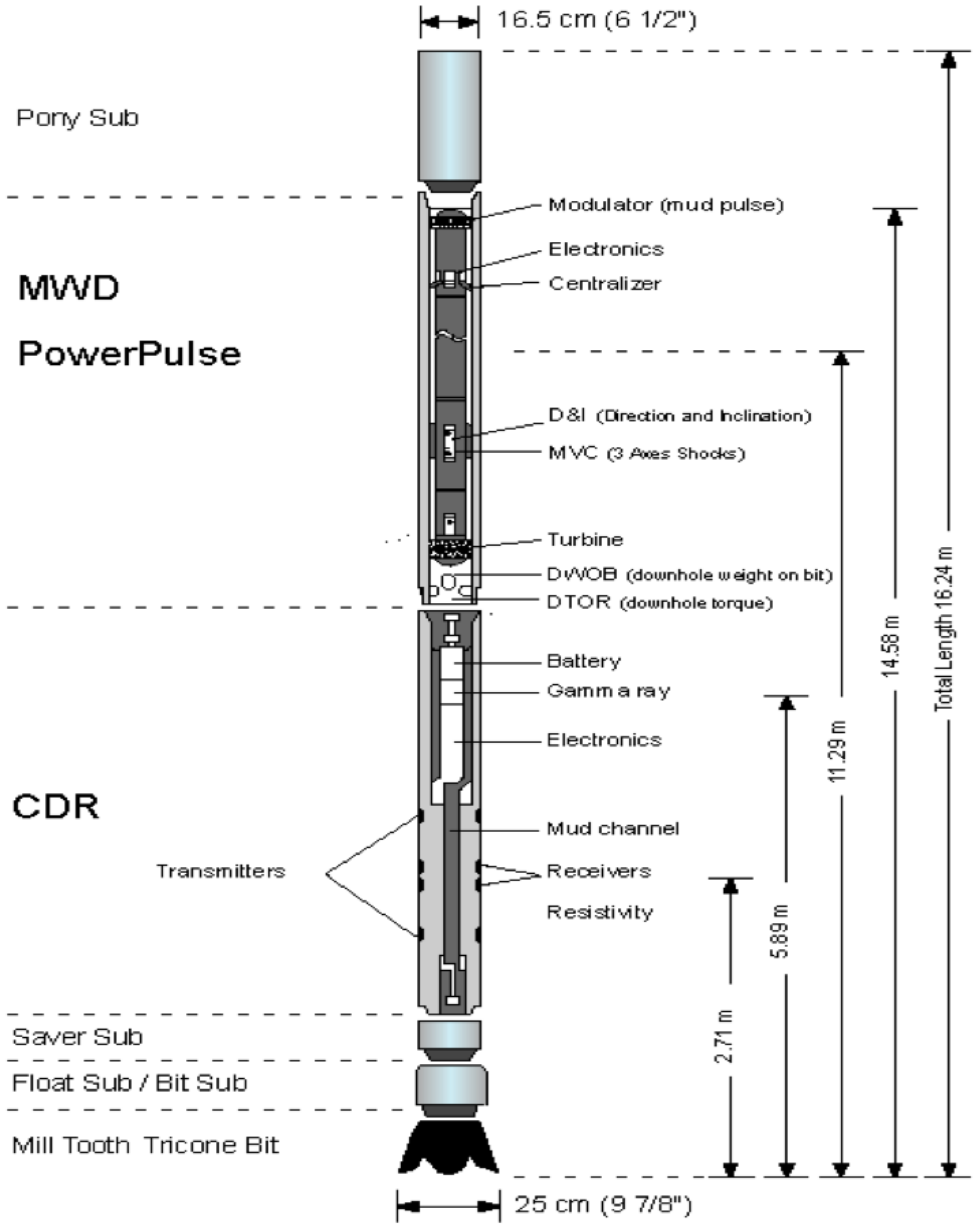

Fig. 5 - A figure of a typical MWD set up with various logging tools for measuring formation properties to inform drilling decisions. As can be seen from the figure, the sensors are often placed at different distances from the bit. As a result, there will be a lag from the data received from the logging tools, and the position of the bit. (Moore et al. 1998)

This distance implies that there can be a large time- and depth-lag between the MWD measurements and the location of the bit. Additionally, the mud pulse telemetry associated with MWD can take additional time for the information to reach the surface recorders. Consequently,

if the bit is out of the target zone, there may be a significant time delay before this is relayed to the surface. Information from a MWD sub can also be relayed to the surface by wired drill pipe that can transfer data quickly with high fidelity (Wilson 2013). Wired drill pipe provides a solution to the time-delay issue of MWD but can be expensive and is still depth-delayed with respect to bit performance.

Mud logging is another process used to identify formation properties during the drilling process. Initially, mud logging was a process of extracting gasses from the mud returning from downhole to identify zones with combustible hydrocarbon gasses (Whittaker 1987). Identifying hydrocarbon gasses within the return mud can signal productive formations with producible hydrocarbons. Rock cuttings from the bit are brought to surface in the return mud as well. These cuttings can be analyzed by a geologist to create a lithology log and can provide rapid analyses of reservoir properties such as porosity and permeability during drilling (Whittaker 1987). Information from mud logging can also be used to inform the geosteering team about the position of the bit within the various lithologic units, as well as identifying zones of various reservoir quality to improve landing zone determination. However, while mud logging can provide important information about formation properties, there is a lag associated with data collection because of the delay in cuttings retrieval. Additionally, in horizontal laterals, it is possible that the retrieval of cuttings may be compromised.

Due to the challenges with the time- and depth-lag and frequency of data recording with MWD and mud logging data, there is a real opportunity to explore the use of real-time surface drilling data to detect lithologies at the bit which is the focus of the work in this thesis. These methods have the potential to provide the geosteering team with higher frequency data in real-time to accelerate the decision making process.

1.4 Description of Surface Drilling Data

Surface drilling data is recorded in real-time and includes torque, hook load, RPM, mud flow rate and pressure, WOB, and ROP (Lesso et al. 2011) as a function of depth or time. Surface drilling data is recorded at high frequencies providing more utility than downhole drilling data whose frequency is limited by mud telemetry bandwidth (Lesso et al. 2011). Though surface drilling data benefits from higher frequency, the measurements are not directly measuring the interaction of the bit with the rock. However, surface drilling data can still be useful for analysis through the use of a metric called mechanical specific energy (MSE). MSE is a measurement that estimates the amount of energy required to remove a unit volume of rock (Teale 1965). MSE is calculated by **Eq. 1**:

$$MSE = \frac{SWOB}{A_{bit}} + \frac{120 \cdot \pi \cdot RPM \cdot STOR}{A_{bit} \cdot ROP} \cdot E_f \qquad \textbf{(Eq. 1)}$$

where SWOB is surface weight on bit, RPM is surface revolutions per minute, STOR is surface torque, ROP is rate of penetration, and A_{bit} is the cross sectional area of the drill bit. This quantity is multiplied by a drilling efficiency number E_f that is uniformly set to 0.35. This factor represents the inefficiencies of the drill bit (Pessier and Fear, 1992). Drilling operations typically focus on minimizing MSE and maximizing ROP simultaneously. However, while MSE is a useful metric, it is does not take into account other sources of energy loss required to describe true drilling efficiency. As described in Chen et al. (2019), there are other energy losses during drilling. Energy is lost in the drill string to elastic strain and kinetic energy, shock/vibration, and the interaction of the string with the rock surface along the walls of the wellbore. To optimize the drilling process,

these mechanisms of energy loss should be taken into account. **Fig. 6** shows a schematic of the energy input and losses during the drilling of a horizontal well.

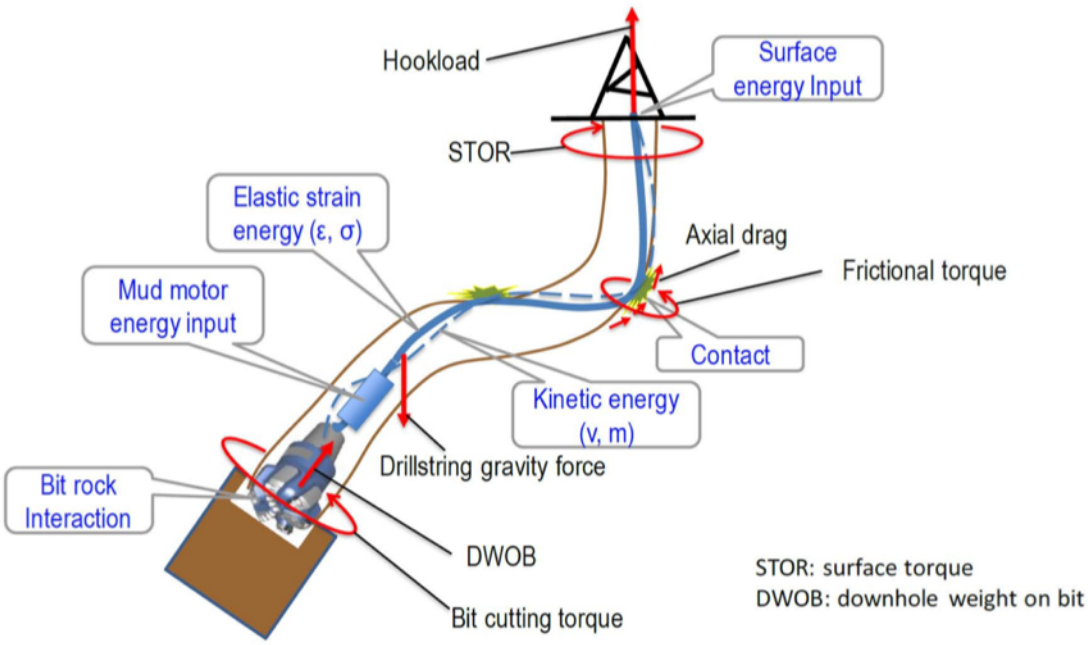

Fig. 6 - A schematic showing the energy input and losses during the drilling process. Energy is supplied from the rig at the surface and the mud motor downhole, and energy is lost from the interaction of the bit and the rock, and the strain, kinetic energy, and contacts of the drill string. (Chen et al. 2019)

Accurate identification of these other forms of energy loss is difficult and often ignored. Due to the difficulties in accounting for these complex forms of energy losses, MSE is often used to estimate efficiency.

1.5 Challenges and Pitfalls of Previous Attempts of Lithology Prediction with Drilling Data

As stated previously, there are inherent problems associated with data acquisition during the drilling process. MWD and mud logging data provide reliable estimates of formation properties but are associated with low data density, and a time delay from when the property is measured and recorded at the surface. In the literature review, I present published attempts to circumvent this problem. Nonetheless, these approaches are also limited either because of restrictive assumptions or the need for additional data. In Li (2019), the most accurate prediction was limited in range ahead of the bit and requires downhole MWD data to be accurate. This method also relies on the assumption that lithology does not change within the prediction window, which is very often invalid and severely impacts the prediction quality. In Esmersoy et al. (2013), check shot measurements are required to look ahead of the bit, in addition to a reasonably accurate geologic model, neither of which are always available. In Romanenkova et al. (2019), the accuracy of the model is limited by the thickness of the layers and will have limited utility in highly heterogeneous formations with thin layers.

In this work, I assume that bottom hole assembly stays constant throughout the entire well, that the influence of the surface drilling crew is minimal, and the vertical and horizontal sections are treated separately. Additionally, the effect of bit wear and transition to new bits is not considered. However, for operational purposes, a change of bit can easily be accounted for when analyzing real-time drilling data.

1.6 Outline

The format of this thesis is as follows: In this chapter, I presented a literature review of published attempts to identify lithology at the bit from drilling data. Chapter two provides a discussion of the algorithms used in this work. These are the time-series change point detection, and Hidden Markov Models. Chapter three describe the analysis and results of the methods applied to a public dataset for the Volve oilfield in the North Sea made available by Equinor (Equinor 2018). Chapter four will provide the analysis and results of the methods performed on a prominent unconventional reservoir in Oklahoma with data provided by Marathon Oil. Finally, Chapter five will provide a summary and conclusion of the research efforts I have conducted for this study.

Chapter 2: Mathematical Formulation

The previous chapter focused on the applications of machine learning and data analytics to gain insights from diverse data sources in the oilfield. This chapter focuses on the two different algorithms discussed in this thesis. These are Hidden Markov Models and time-series change point detection.

First, I present the change point detection algorithm from the Ruptures Python package (Truong et al. 2018) and describe the mathematical formulation. This package has multiple options for the configuration of change point detection algorithm which is described in detail. The second method described is the Hidden Markov Model (HMM). I implement HMMs with the R package depmixS4. Within this algorithm, the Forward-Backward algorithm, and the Baum-Welch algorithm are necessary for determination of the parameters for the HMM.

2.1 Change Point Detection

Change point detection is a technique that attempts to identify changes within a random non-stationary signal. A random signal is considered to be second-order stationary if the mean and variance are constant over time (Gagniuc 2017). A non-stationary signal, on the contrary, is characterized by a spatially or temporally varying mean and variance. Examples of non-stationary signals include porosity as a function of depth in a multilayered reservoir with different facies, or hourly wind speeds recorded in weather surveys. Other examples include time-series data containing non-periodic events. Non-stationary signals occur in many domains, such as bioinformatics, weather, and engineering (Adak 1993).

Real-time drilling data is representative of a non-stationary multichannel time-series signal that is influenced by operational decisions, depth, bottom-hole assembly architecture, changes in

lithologies, changes in well trajectory or even changes in the drilling crew. Some of the changes can contain information related to lithology changes, while others may be unrelated.

Change points occur when there is a shift in the statistical properties of a signal. To detect these changes, a select few statistical properties are monitored over time (Bai 1999). The assumption is that the signal is piece-wise stationary and that there are abrupt changes in properties at some unknown instances in time. In this thesis, I focus on the mathematical description of change point detection from the work of Truong et al. (2019).

Given a non-stationary random process $y = \{y_1, \dots, y_T\}$, for real values over T samples, changes in signal properties occur at instances $t_1, t_2, \dots, t_k$ where k is the number of change points. Depending on the application, the number of change points may or may not be known. The goal of change point detection is to accurately identify these changes over the given signal. To accomplish this, change point detection is formulated as a problem of identifying the optimal segmentation, τ, of the given signal. **Fig. 7** shows a time series signal segmented with different segments and change points in the optimal segmentation τ.

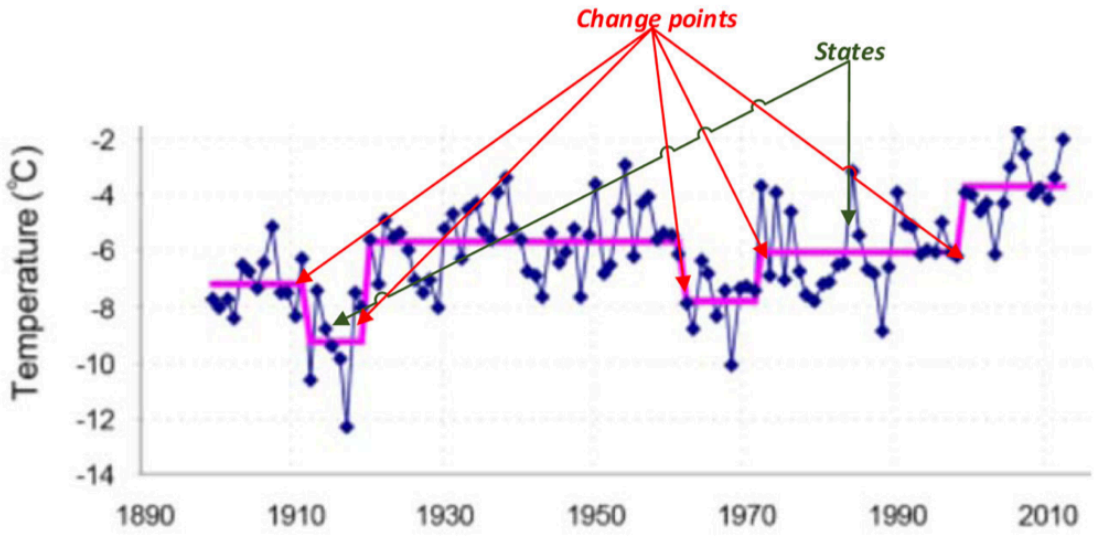

Fig. 7 - A plot showing a time series signal of yearly mean temperature of a Norwegian island. In this signal, the optimal segmentation τ is shown by different segments/states with the transitions between segments representing the change points. (Aminikhanghahi and Cook 2017)

To identify the optimal segmentation, a cost function is applied for each segment, and the sum of the cost function, $v(\tau, y)$ is minimized. This is represented generally by **Eq. 2**.

$$v(\tau, y) = \sum_{k=0}^{K} c(y_{t_k \ldots t_{k+1}}) \quad \textbf{(Eq. 2)}$$

In this formulation, $c(y_{t_k \ldots t_{k+1}})$ is a chosen cost function that is a measure the homogeneity of a segment from $\{t_k, t_{k+1}\}$, or the distance between two change points. Good segmentations have segments with homogenous data from the measured statistic in the cost function, and results in a low value of cost. Therefore, the best segmentation according to this formulation is the segmentation that minimizes the function v. For problems with an unknown number of change points, a penalty function is applied that represents a measure of complexity for each segment. This casts the problem into a discrete optimization problem represented by **Eq. 3**:

$$\min v(\tau) + pen(\tau) \quad \textbf{(Eq. 3)}$$

The most common penalty function, $pen(\tau)$, is the linear penalty. This penalty is described by:

$$pen(\tau) = \beta|\tau| \quad \textbf{(Eq. 4)}$$

where β is a smoothing parameter. This parameter controls the complexity of the model, where low values of β result in less smoothed signals and subsequently more change points, while high values of β result in smoother signals and discards more change points. The choice of β remains subjective and can be tuned to fit the specific application.

From the formulation of **Eq. 2**, the cost function needs to be selected for implementation of change point detection. There are several different possible choices of cost functions that can be selected based on the specific application. Cost functions are divided into two categories, parametric and non-parametric. Parametric cost functions operate on a finite-dimensional vector to derive segment costs. Non-parametric cost functions are applicable when the distribution family of the given signal is unknown or non-parametric.

A common parametric cost function category is the maximum likelihood estimation approach (Page 1955). For this approach, the signal is taken to be piece-wise stationary with segments of constant distribution. Statistical parameters, denoted by θ (such as the mean for example), change at point t_k which is the selected change point. There are many different cost functions designed to identify the changes in the parameter θ, and different statistical qualities used for determining change points. One such function is the quadratic error loss cost function

(Sen and Srivastava 1975), which is designed to detect a shift in mean of a piece-wise constant mean signal. This function is described by **Eq. 5**:

$$c(y_{a\ldots b}) = \sum_{t=a+1}^{b} \|y_t - \bar{y}_{a\ldots b}\|_2^2 \qquad \textbf{(Eq. 5)}$$

where $\bar{y}_{a\ldots b}$ is the mean of the sub-signal from a to b, and y_t is the value of the signal at time t. This function is used to detect change points that occur as a shift in mean from one sub-signal to another. This type of cost function has been applied in various settings including the analysis of well log data (Adams and McKay 2007) where NMR well log data from the Nile River Delta is used for detecting change points that correspond to transitions between geologic facies. The NMR response is assumed to be piece-wise constant mean that changes for each layer/lithology. Mean-shift change point detection is shown to accurately identify the different geologic layers from the log data (**Fig. 8**).

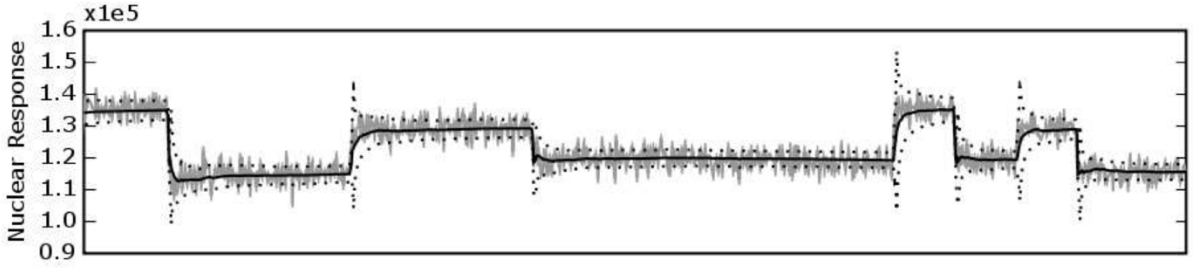

Fig. 8 - NMR response from the subsurface formations of the Nile River Delta. The various facies have varying NMR response and have constant mean within each formation. A mean shift change point detection algorithm is used to identify change points. The solid black line shows the mean, and the dashed line is a one standard deviation error. The change points are detected when the mean has a sudden change. (Adams and McKay 2007)

Another type of parametric cost function is the piecewise linear regression cost function (Kim and Siegmund 1989). In these types of cost functions, there is a response signal dependent on covariates, and at some instance t_k the relationship between covariates and response abruptly changes. In this type of cost, a linear regression (or multiple regression) is fit to each segment of the signal and the change points occur where the regression coefficients abruptly change.

Finally, the last element of change point detection is the search method for identifying the changes within the signal. There are two different categories of search methods: exact and approximate methods (Truong et al. 2019). Exact methods determine the exact solution to the discrete optimization problem, while approximate methods provide a segmentation that approximates the optimal solution. A naïve approach to optimal search would be complete enumeration of the signal into all possible segmentations to exhaustively search for the optimal segmentation. Unfortunately, this is often far too computationally exhaustive, especially for problems with unknown amount of change points and multichannel signals. As a result, approximate methods are often employed to reach reasonable solutions in an acceptable amount of time. There is an optimal search method that does not require the number of change points to be known and has low computational complexity on the order of $\mathcal{O}(T)$, known as Pruned Exact Linear Time (PELT) (Killick et al. 2012), where *T* is the number of data points in the signal. This algorithm works when the given penalty function is linear. The purpose of this algorithm is to disregard points that do not reduce the value of the objective function as possible change points. A pruning rule is used to make this decision and the rule is defined as follows: (Killick et al. 2012)

$$if\ [\min \nu(\tau, y_{0\ldots t}) + \beta|\tau|] + c(y_{t\ldots s}) \geq [\min \nu(\tau, y_{0\ldots s}) + \beta|\tau|]; \qquad \textbf{(Eq. 6)}$$

$$then\ t\ cannot\ be\ the\ last\ change\ point\ prior\ to\ \tau$$

The indexes *t* and s are given as (0<t<*s*<*T*). This implies that if the segmentation from $y_0 \ldots y_t$ plus the cost of the segment from $y_t \ldots y_s$ is greater than the minimization of the cost of the segmentation from $y_0 \ldots y_s$, there have to be additional change points between *t* and *T*. Essentially, if the segmentation from $y_0 \ldots y_t$ plus the cost of an additional segment to y_s is worse than the optimal segmentation from $y_0 \ldots y_s$, there must exist additional change points between $y_t \ldots y_T$ that further reduce the overall cost. This is then followed by pruning that removes points which cannot possibly be change points, thereby significantly reducing the computational cost compared to other exact methods. This type of segmentation can be seen in **Fig. 9**.

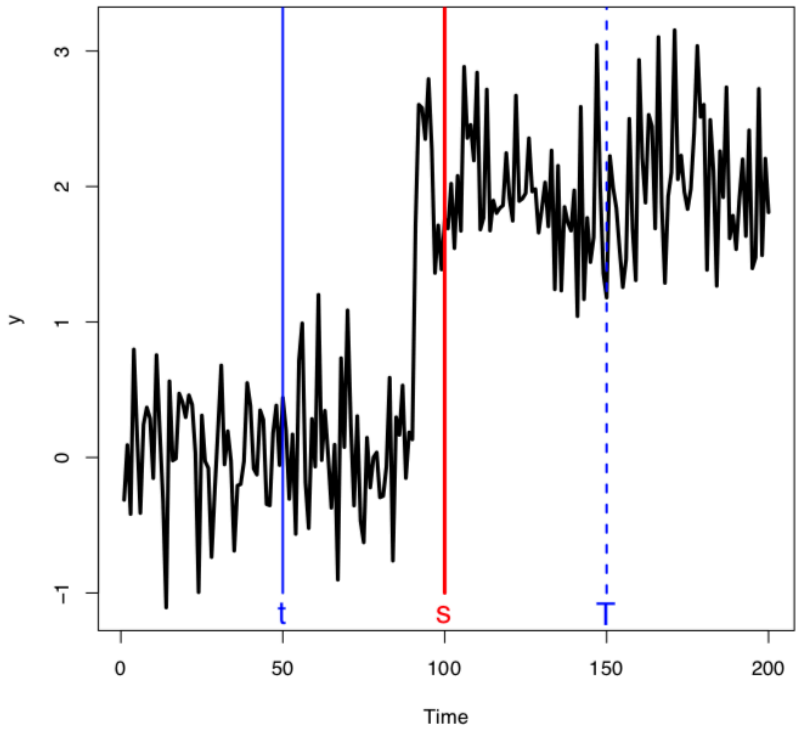

Fig. 9 - An example with simulated data showing the mechanics of the PELT algorithm. The segmentation from *0* to *t* plus the cost of the segmentation from *t* to *s*, is greater than optimal segmentation from *0* to *s*. This means that there must be another change point between *t* and *s* (just before *s*). (Killick 2017)

This algorithm selects a point s on a segment from $y_t \dots y_T$ that eliminates the data preceding s thereby significantly reducing the amount of data needed to search for change points. Once a change point is selected at s, the algorithm performs another iteration by shifting t to the location of s. The algorithm then either finds another change point between the new t and T, or does not find a change point that reduces the overall cost and ends with the optimal segmentation. This improves the computational efficiency of the algorithm significantly by eliminating a portion of the signal before searching for change points. The algorithm is able to determine the global minima for the associated cost and penalty function being applied to the signal (Killick et al. 2012).

There are multiple different approximate search methods that can be used to identify a known or unknown number of change points (Truong et al. 2019). While these methods compute an approximate solution to the optimization problem, many are able to achieve acceptable segmentations. One such method is the window-based search algorithm (Basseville and Nikiforov 1993). This algorithm uses a sliding window that moves across the signal and computes the statistical property of interest to identify change points. The window is divided into half, with the first window sliding along the newest points within the signal, and the second window following just behind the first. For each window, a chosen statistical property is calculated in both half-windows and compared. When there is a large discrepancy between the properties of the two half-windows, I detect a change point. This involves computing a discrepancy measure which is defined as:

$$d(y_{u\dots v}, y_{v\dots w}) = c(y_{u\dots w}) - c(y_{u\dots v}) - c(y_{v\dots w}) \qquad \textbf{(Eq. 7)}$$

with the indexes $u < v < w$. This measure provides a quantitative comparison of the two windows by splitting the signal at v. The pre-defined statistical property is compared for each segment, u to v and v to w. For example, in **Fig. 10** the sliding window is split at v where the two windows are the dark blue (u to v) and light blue rectangles (v to w). In the left-most location of the windows, the split at v does not create two segments that are different, and the resulting discrepancy curve is low. However, in the right-most location, the split at v does result in two segments with different mean and the discrepancy curve is high. When the discrepancy curve exceeds a threshold and reaches a peak, the point is selected as a change point.

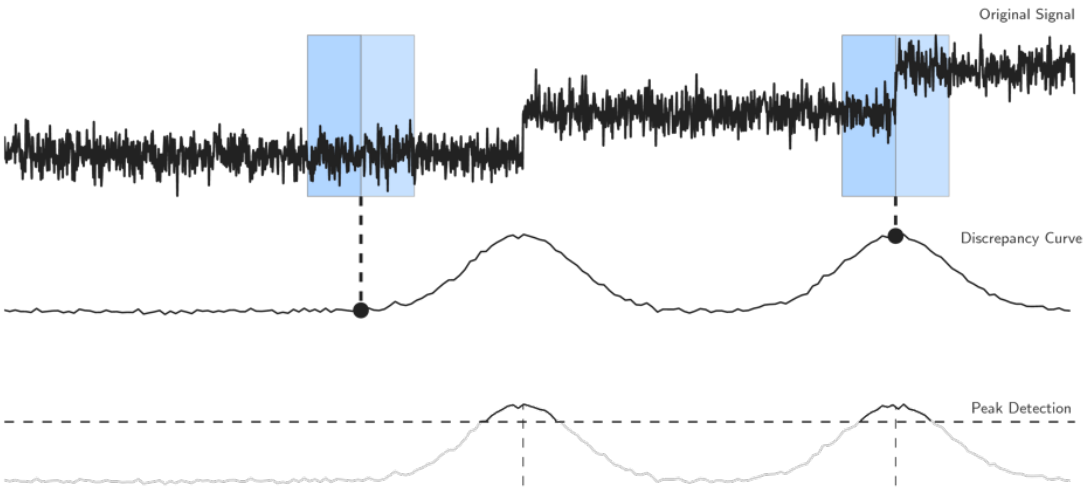

Fig. 10 - A schematic of the sliding window search method for change point detection. In the top of the figure, the signal and sliding windows are shown. At each point the statistical properties between the light and dark blue windows are computed and compared by the discrepancy measure, which is plotted below. For the two obvious change points in the signal, the discrepancy measure reaches a peak and is detected by the peak detection algorithm, shown at the bottom. (Truong et al. 2019)

This method requires the input of the width of the window size w, which is subjective but can be tuned to fit the type of signal being processed. This algorithm is convenient to use because

of the low computational complexity on the order of $\mathcal{O}(nw)$, where n is the number of points in the signal, and w is the width of each window, and because it can be used on signals with an unknown number of change points (Basseville and Nikiforov 1993).

Another approximate search method is the binary search algorithm which is a greedy segmentation of the signal into different sub-signals. For this approach, the signal is examined in its entirety before being segmented into two segments where the statistical properties of the segments differ. Following the first segmentation, additional splits are made on the two new segments if any other change points are detected. This process is repeated sequentially for each new segment as shown in **Fig. 11** which shows binary segmentation of a signal into three different segments.

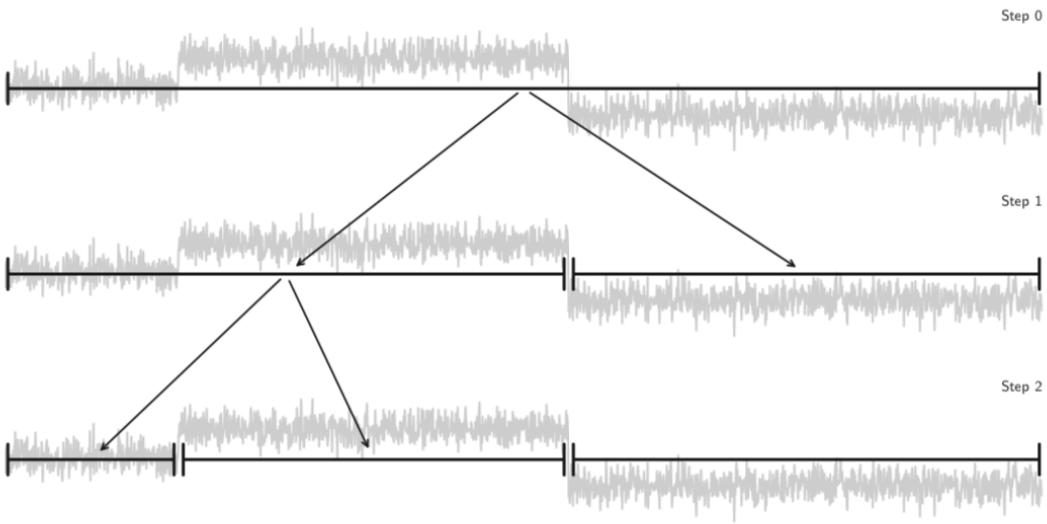

Fig. 11 - An example of binary segmentation of a signal with two change points, and three segments. In step 0, the signal is processed in its entirety, and a change point that reduces the overall cost of segmentation is chosen. The first change point corresponds to the largest change within the signal. The next step performs the same process for both new segments and selects a second change point (Truong et al 2019).

Because the binary search method is approximate, the resulting segmentation may be sub-optimal, and has been shown to create poor segmentations for signals with two or more change points (Hawkins 2001). Additionally, Killick et al. (2012) explains that when change points are close to one another, the search method often fails to correctly identify the change points. The sequential search for the identification of additional change points depends on the location of the previously identified change points and therefore, when change points are close together, the error associated with approximating change point location can become more pronounced. For geoscience applications, there will likely be several change points, possibly occurring in close proximity to one another and therefore this search method is unlikely to be promising in practice.

2.2 Hidden Markov Model

Hidden Markov Models (HMM) are used to detect a series of states along a time-series signal and do not directly detect change points. These states can be viewed as sub-signals that correspond to responses from some model or process. When there is a change of state, this can be interpreted as a change in the underlying process.

The fundamental concept underlying Hidden Markov Models is the Markov property that predicates that the current state of the system solely relies on the immediate past state and not the earlier history (Gagniuc 2017). This property is written as:

$$P(C_{t+1}|C_t, \dots, C_1) = P(C_{t+1}|C_t) \qquad \textbf{(Eq. 8)}$$

Where $\{C_t : t \in N\}$ represents the sequence of random variables for $t = 0$ to N. In other words, the next state C_{t+1}, is not dependent on the entire sequence history but only on the present C_t. This

can be easily displayed by a directed graph as seen in **Fig. 12**, showing the dependencies of each variable in the sequence only on the immediate past.

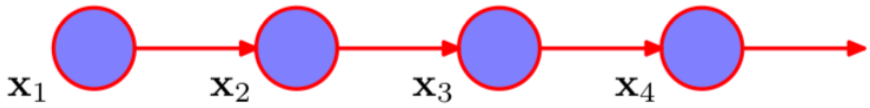

Fig. 12. A directed graph showing the first order Markov property. In this graph, the sequence moves to the right with random variables X occurring at each time step. For each time, the variable is dependent only on the immediate past. (Bishop 2006)

These random variables are a discrete sequence of states that represent some discernable properties. A simple example of such states for a time-series of weather patterns is a sequence of sunny and rainy days. Lithology variations can also be represented as a Markovian process because depositional processes typically tend to follow ordered patterns such as transgressive to progressive sequences.

For a transition from one state to another, I define a probability called a transition probability, which explains the dependency of the future on the present. The transition probability is written as:

$$a_{ij} = P(C_{t+1} = j | C_t = i) \qquad \textbf{(Eq. 9)}$$

where i and j refer to the current state and future state respectively and represents the probability of transitioning to state j, given that you are currently in state i. Note that a transition to a different state is not necessarily required, and that i can be equal to j. Therefore, for a Markov chain with

m states, there is an $m \times m$ matrix of transition probabilities to describe the possible transitions. This matrix can be written as:

$$\Gamma = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix}$$

where the sum of each row is equal to 1 (sum of transition probabilities from one particular state to all possibilities). The diagonal of the matrix represents the probability of staying in the same state for the future time step. For example, in a scenario with three states given as $j_1 - j_3$, the transition probabilities, a_{ij}, can be visualized by **Fig. 13**.

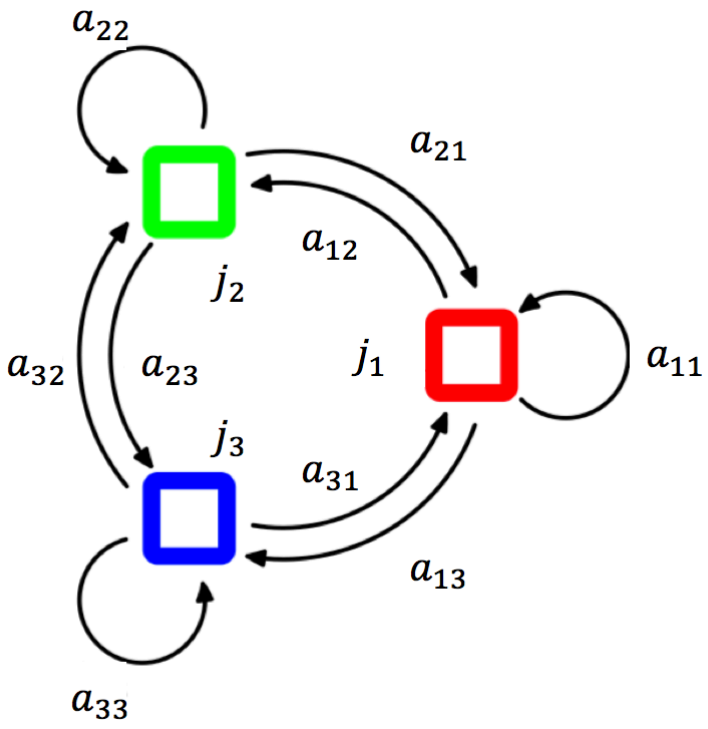

Fig. 13 - A visualization of the transition probabilities a_{ij} for a three state Markov chain. The probability of transitioning from one state j_i to another is shown by the directed arrows.

For example, a_{12} represents the probability of transitioning from state 1 to 2 shown as an arrow leading from the red to green state. (Bishop 2006)

These transition probabilities are a key element to the formulation of the HMM. Markov chains can be useful to derive the probability of being in a certain state at a time t in the future. This is referred to as the unconditional probability and can be written as:

$$u(t) = \big(P(C_t = 1), \dots, P(C_t = m)\big), \quad t \in N \qquad \textbf{(Eq. 10)}$$

This a vector of probabilities with $m \times N$ entries to represent the probability distribution for being in each of the m states at time t. With this, the probability distribution for $t + 1$ can be deduced from the distribution at t and Γ. To obtain $u(t + 1)$ the calculation is as follows:

$$u(t + 1) = u(t)\Gamma \qquad \textbf{(Eq. 11)}$$

More generally, given the initial probability distribution π, that describes the probability of initially being in each of the m states, you can compute the probability distribution of u at any time t by:

$$u(t) = \pi^T \Gamma^t \qquad \textbf{(Eq. 12)}$$

This is the initial probability distribution transposed into a row vector, multiplied by the transition probability matrix raised to the *t* power.

I now discuss the concept of a 'hidden' Markov chain. In an HMM, the process responsible for the states is hidden and not observed directly. I can however detect some output or observations generated by the underlying hidden process. These observations (dependent on the states) are used to identify a succession of states as shown in a directed graph of a first order Markov chain (**Fig. 14**).

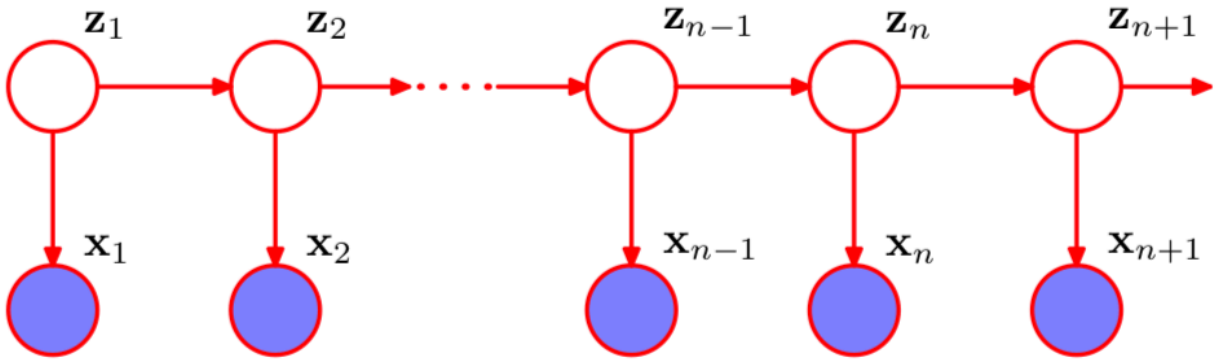

Fig. 14 - A sequential directed graph showing the basis of the HMM. Each state in the sequence represented by Z is connected to the previous state, and an observation, X, occurring at the same time. This is an extension of the first order Markov chain presented above that includes observations conditioned on the state. (Bishop 2006)

In the graph shown above, the states represented by Z, exhibit the first order Markov property of being connected to only the immediately preceding state. However, these states are not observed, rather there are observations, given by X, conditioned to the states at each time step. With these observations X, the unobserved states Z can be predicted in the immediately following time step.

There are three elements of the HMM that I now describe in detail. The first element is a finite number of states, N, that have some distinct properties. The second element are the transition probabilities that represents the probability of transitioning from one state to the next from time t

to $t+1$. It Is important to note that there is no requirement to change steps at each time-step; rather the model can remain within the same state over successive time-steps. For the first time step in the sequence, I assume some initial state probabilities that are later refined by the HMM algorithm.

The third element is the emission probability which represents the probability that a certain observation is a result of a particular state. Therefore, there are N such emission probabilities, one for each state. The following will give the notation for the HMM: (Rabiner and Juang 1986):

$$T = Length\ of\ observation\ sequence$$

$$N = Number\ of\ states$$

$$M = Number\ of\ observations$$

$$Q = \{q_1, q_2, \dots, q_N\}, sequence\ of\ states$$

$$O = \{o_1, o_2, \dots, o_N\}, sequence\ of\ observations$$

$$A = \{a_{ij}\}, a_{ij} = P(q_j\ at\ t+1 | q_i\ at\ t), state\ transition\ probability$$

$$B = \{b_j(k)\} = P(o_k\ at\ t | q_j\ at\ t), emission\ probability\ distribution$$

$$\pi = \{\pi_i\} = P(q_i\ at\ t = 1), initial\ state\ probability$$

With this notation, a HMM can be described as $\lambda =< \pi, A, B >$, where λ represents the HMM. The mathematical description of HMM follows the work of Petrushin (2000). The goal of using HMM is to determine the likelihood that the sequence of observations O is produced by the HMM λ. This is referred to as the likelihood calculation given as $P(O|\lambda)$. To compute this, first assume that the sequence of states Q that produces the observation sequence O is known. Then the joint probability of O and Q can be calculated by:

$$P(O, Q|\lambda) = \pi_{q1} B_{q1}(O_{q1}) \prod_{i=2}^{T} A_{qi-1} B_{qi}(O_i) \qquad \textbf{(Eq. 13)}$$

This is the formula for just one state and for the full likelihood, this must be repeated and summed over all states given by:

$$P(O|\lambda) = \sum_{i=1}^{Q} P(O, Q|\lambda) \quad \textbf{(Eq. 14)}$$

However, this is only viable for short observation sequences as the number of computations required is given by N^T (i.e. for 3 states with 1000 observations, 3^{1000} calculations required). This can be avoided, however, with the use of dynamic programming. Dynamic programing is a technique that stores the results of computations to be used later when needed (Petrushin 2000). Dynamic programming is utilized with the forward-backward algorithm which uses recursive calculations to be able to solve the calculations for the next instance using the current one (Baum 1972). The forward-backward algorithm is used to compute the probability distribution for each state given a sequence of observations, $P(Q_t|O_{1:T})$. For the forward calculation, let $\alpha_k(i)$ be the probability of the observation sequence O, for t increasing in length from 1 to k resulting in state i. First, the observation probability for the first observation is calculated for each state $j = 1\ to\ N$ by:

$$\alpha_1(j) = \pi_j B_j(o_1) \quad for\ j = 1\ to\ N \quad \textbf{(Eq. 15)}$$

This step is the initialization step. This is followed by the recursion step which calculates α_k for the remaining observations in the sequence. This involves calculations for each state, and for each possible transition between states. As mentioned previously, the recursion step uses the

calculation from the previous observation to compute the present observation. Therefore, α_{k+1} can be calculated by:

$$\alpha_{k+1}(j) = \left[\sum_{i=1}^{N} \alpha_k(i) a_{ij}\right] \cdot B_j(o_{k+1}) \quad For\ j = 1\ to\ N, k = 1\ to\ T-1 \qquad \textbf{(Eq. 16)}$$

This calculation can be more easily understood visually with a directed diagram showing the possible transitions to one state j. Given a scenario identical to that of **Fig. 13** the calculation of $\alpha_{k+1}(j)$ is shown in **Fig. 15** below.

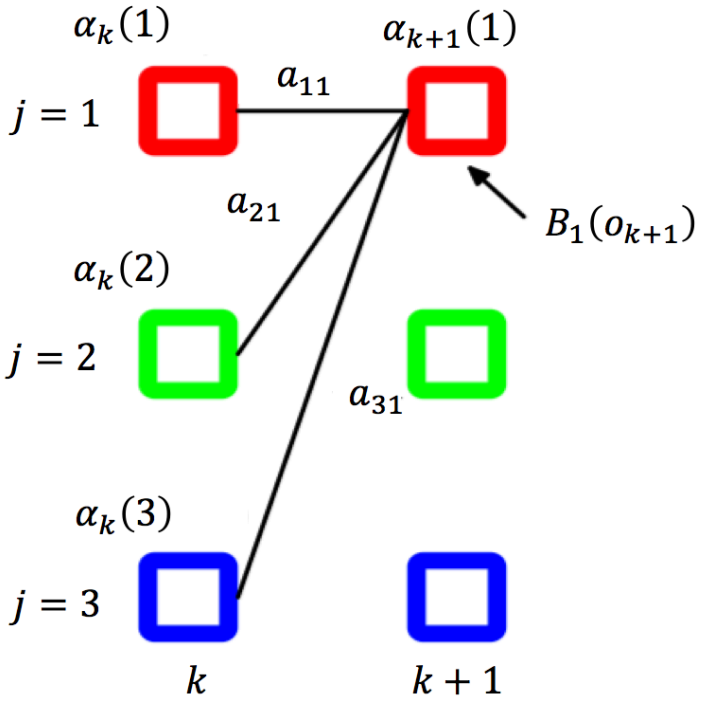

Fig. 15 - The forward recursion step of the forward backward algorithm for calculation of $\alpha_{k+1}(1)$. To calculate $\alpha_{k+1}(1)$, the values of $\alpha_k(j)$ are required which are calculated recursively starting with the initial probabilities π_j from the first time step forward (Eq. 16). Summing over the $\alpha_k(j)$ multiplied by their weights a_{j1} (transition probability from state *j* to 1), and multiplied by the emission probability B_1 for the observation occurring at time *k+1* given state 1 allows for calculation of $\alpha_{k+1}(1)$.

Finally, the required probability, $P(O|\lambda)$, is calculated as the sum of the $\alpha_T(i)$ for each state in the HMM. This is the procedure for determining the probability that an observation sequence is produced by the HMM λ. This probability can be similarly calculated using the backward calculation from the forward-backward algorithm. The backward algorithm calculates a variable symmetric to α_k denoted as β_k given by $P(o_{k+1}, o_{k+2} \ldots o_T|q_k = i, \lambda)$. For the calculation of this variable, the calculations are done recursively backwards along the observation sequence. The backward calculation begins with an initialization step by setting $\beta_T(i) = 1$ for all states $i = 1\ to\ N$. Next, the recursion step is performed, which for the backward algorithm implies using β_{k+1} to calculate β_k. This calculation is given by:

$$\beta_k(i) = \sum_{j=1}^{N} a_{ij} B_j(o_{k+1}) \beta_{k+1}(j) \quad For\ i = 1\ to\ N, k = T - 1\ to\ 1 \qquad \textbf{(Eq. 17)}$$

This calculation can be visualized similarly to the forward algorithm with a directed graph of states. **Fig. 16** shows a graphical representation of the backwards algorithm to recursively calculate $\beta_k(i)$.

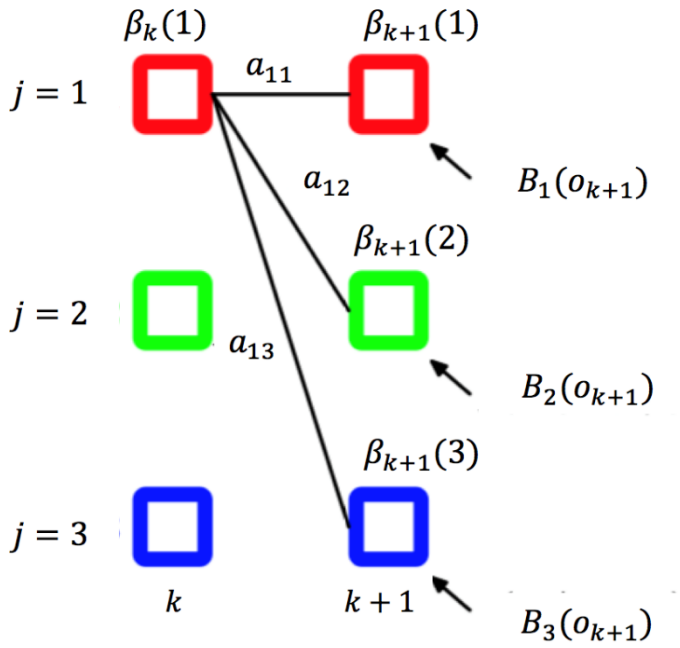

Fig. 16 - The backward recursion step of the forward backward algorithm for calculating $\boldsymbol{\beta_k(1)}$. To calculate $\boldsymbol{\beta_k(1)}$, the $\boldsymbol{\beta_{k+1}(i)}$ values are required which are initialized as 1 for the final time step, and the subsequent values are recursively calculated backwards through the time steps. The emission probabilities $\boldsymbol{B_{k+1}(j)}$ are required for each state, which are multiplied by their transition probabilities $\boldsymbol{a_{1i}}$ and summed then multiplied by $\boldsymbol{\beta_{k+1}(j)}$.

After computing through the observation sequence with estimates of β_1, the required probability can be computed for the HMM. The probability of the observation sequence by the HMM is given by:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \qquad \textbf{(Eq. 18)}$$

With these calculations, the parameters α_k and β_k can be used to help identify the most probable sequence of states given the observation sequence. This is the overall goal of the HMM

process: to accurately identify the sequence of states from only the sequence of observations. This can be done with posterior decoding (Petrushin 2000). That is, for each observation, given the transition and emission probabilities, and the state at the current observation, you can determine which state the system is most likely to be in at the next observation. To describe this, let $\lambda_k(i)$ represent the probability that the HMM emitting an observation at $t = k$ is in state i. Calculation of the parameter $\lambda_k(i)$ is given by:

$$\lambda_k(i) = \frac{\alpha_k(i)\beta_k(i)}{P(O|\lambda)} \quad For\ k = 1\ to\ T, i = 1\ to\ N \qquad \textbf{(Eq. 19)}$$

Using this calculation, I can now determine which state is the most probable for time k. This is given by:

$$q_k = \arg\max_{1 \le i \le N} \{\lambda_k(i)\} \quad For\ k = 1\ to\ N \qquad \textbf{(Eq. 20)}$$

To determine the most probable states for each time in the observation sequence based on the observations and the HMM parameters, I now need to have reliable estimates for the transition and emission probabilities for the calculations of the sequence of hidden states as shown above.

To estimate the transition and emission probabilities, if you have an observation sequence O of length T, and you also know the sequence of states Q, you could easily estimate the probabilities by counting the transitions from state i to j, and the number of times an observation k is made while in state i, compared to the total number of each. This would give you a fractional amount for each state transition, and each signal emission that would be used as the probabilities. This process is similar to the parameter definitions for Markov chains where the state sequence is

known. However, this is not possible as the sequence of states Q is not known (hidden). Therefore, these parameters need to be estimated without knowing the state sequence, and only from the observation sequence. This is made possible by the Baum-Welch algorithm (Baum 1972). The goal is to estimate the number of times that there are transitions from state i to state j, and the number of times that there are transitions away from state i for observation o. To accomplish this, the probabilities of being in the i^{th} state at time k, and transitioning to the j^{th} state at time $k+1$ are required. The probability of being in state i at time k is defined above as λ_k. The probability of transitioning from state i to j at time $k+1$ is given by $\xi_k(i,j)$. This probability can be written as:

$$\xi_k(i,j) = P(q_k = i, q_{k+1} = j|O,\lambda) = \frac{\alpha_k(i)a_{ij}B_j(o_{k+1})\beta_{k+1}(j)}{P(O|\lambda)} \qquad \textbf{(Eq. 21)}$$

This uses the forward and backwards probabilities as shown above, and the transition and emission probabilities that have not been calculated. The Baum-Welch algorithm computes approximations of these probabilities in an iterative fashion and compares the approximations progressively. First, random values of the transition and emission probabilities, a_{ij} and b_j, are chosen. With these, the parameters α and β can be calculated with the Forward-Backward algorithm. Given these parameters, the number of transitions from state i to j, and the number of transitions away from state i associated with observation o can be calculated. Let A_{ij} represent the number of transitions from state i to j, and $B_i(k)$ be the transitions from state i at time k associated with observation o. These are calculated by:

$$A_{ij} = \frac{1}{P(O|\lambda)} \sum_{l=1}^{1-T} \alpha_i(l) a_{ij} b_j(o_{l+1}) \beta_j(l+1) \quad For\ i = 1\ to\ N, l = 1\ to\ T-1 \quad \textbf{(Eq. 22)}$$

$$B_i(k) = \frac{1}{P(O|\lambda)} \sum_{\{l|o_l=k\}} \alpha_i \beta_i(l) \quad For\ i = 1\ to\ N, l = 1\ to\ T \quad \textbf{(Eq. 23)}$$

These calculations are referred to as pseudo-counts, which are the number of transitions from i to j, and the transitions from i associated with observation o. This generates a pseudo-state sequence that can be used to calculate the transition and emission probabilities. These probabilities are then calculated as:

$$a_{ij} = \frac{A_{ij}}{\sum_{i=1}^{N} A_{ij}} \quad For\ i, j = 1\ to\ N \quad \textbf{(Eq. 24)}$$

$$b_i(k) = \frac{B_i(k)}{\sum_{l=1}^{M} B_i(l)} \quad For\ i = 1\ to\ N, k = 1\ to\ M \quad \textbf{(Eq. 25)}$$

For the emission probabilities, M is the number of discrete observations in the observation sequence. This calculation process is repeated with the new estimations of a_{ij} and b_{ij}, and compared with the previous estimations of a_{ij} and b_j (initially randomly chosen). With each new estimation, the log likelihood is calculated for the consecutive estimates. This process is repeated until the difference between two values of likelihood are within a certain threshold, or the maximum number of iterations has occurred. This ensures that the estimation of the HMM parameters are near optimal because the exponential family of probability functions are logarithmically concave, and small differences in likelihood show that the estimations are near the global minima (Kass and Vos, 1997).

With the discussion above, I can now calculate all components of the HMM model to describe a scenario where the observation sequence is a set of discrete values, such as “hot” or

"cold" for example. HMMs can, however, be used for continuous observation sequences with the use of Gaussian mixture models to describe the observation sequences. Gaussian mixture models are a way of representing continuous data with a number of Gaussian distributions (Bishop 2006) that are combined linearly to describe more complex densities with something called mixing coefficients. These mixing coefficients are, in essence, probabilities that a data point comes from a specific Gaussian distribution. For HMM with continuous observation densities, the emission probabilities are replaced by the Gaussian mixture models and become a function of the mean and covariance of the mixtures.

I will present a simplistic example to show how HMM works in an easy to understand scenario. This example follows the work presented by Oehm (2018) where there is a game with two players, Bob and Alice, who roll two dice and win jellybeans. The game is set so that if Bob rolls the dice, and the resulting sum is greater than four, Bob takes a handful of jellybeans and gets to roll again. If Bob rolls less than four, he passes the dice to Alice. However, if the roll is a two, he takes a handful of jellybeans before passing the dice to Alice. When Alice has the dice and rolls greater than four, she takes a handful of jellybeans. However, Alice only likes black jellybeans, so when she takes a handful, she eats only the black beans and puts the other colors back into the jar. Therefore, it would be expected that the number of jellybeans removed from the jar will be less when Alice takes a handful versus Bob. Now the goal for the HMM is to identify which player took jellybeans from the jar by observing only the number of jellybeans removed. In this case, the players are the hidden states (Bob or Alice), and the number of beans removed are the observations. **Fig. 17** shows simulated data for the number of jellybeans taken per roll for 100 dice rolls.

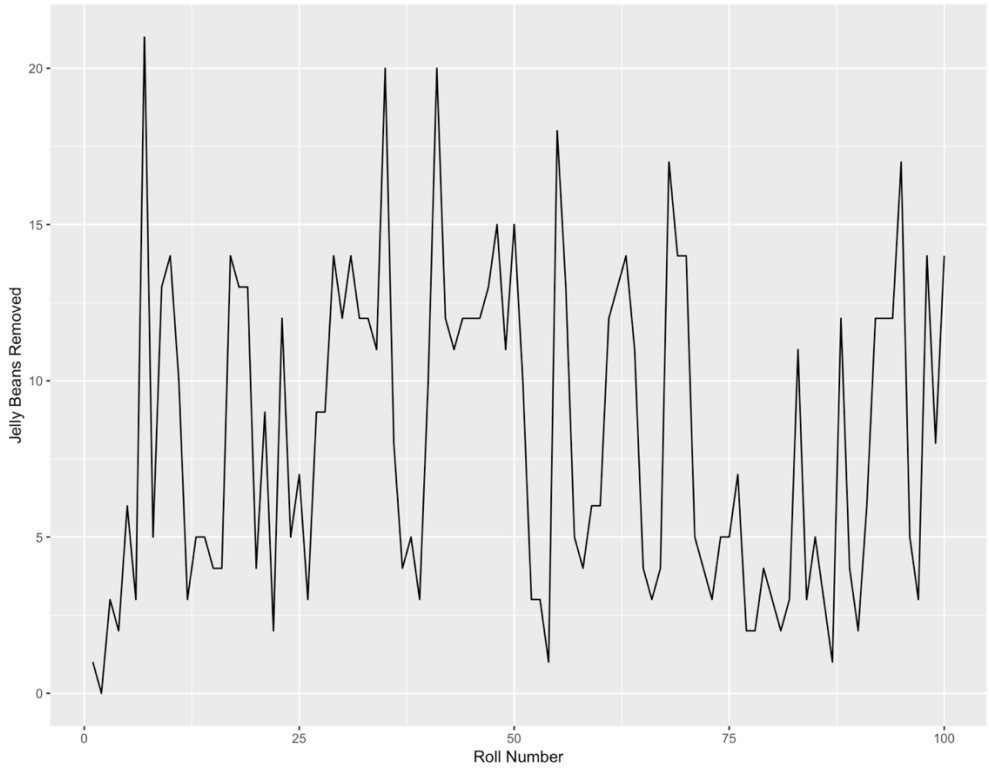

Fig. 17 - Simulation of the number of jelly beans taken for each dice roll for 100 rolls. By this data alone, it would be difficult to say which player was taking jelly beans from the jar. This is the series of observations to uncover the sequence of hidden states (Bob or Alice). (Oehm 2018)

Looking at the simulated data to determine who rolled the dice is rather challenging. A first approach is to assume that when a low number of beans are taken, the roller is Alice. Because of the sequential nature of the game (rules for giving dice to the other player), HMM can be used to estimate the probability of the player rolling the dice at each roll. In **Fig. 18** shown below, the observations, true sequence of states, HMM predicted states, and the HMM state probabilities are shown for each roll number in the game.

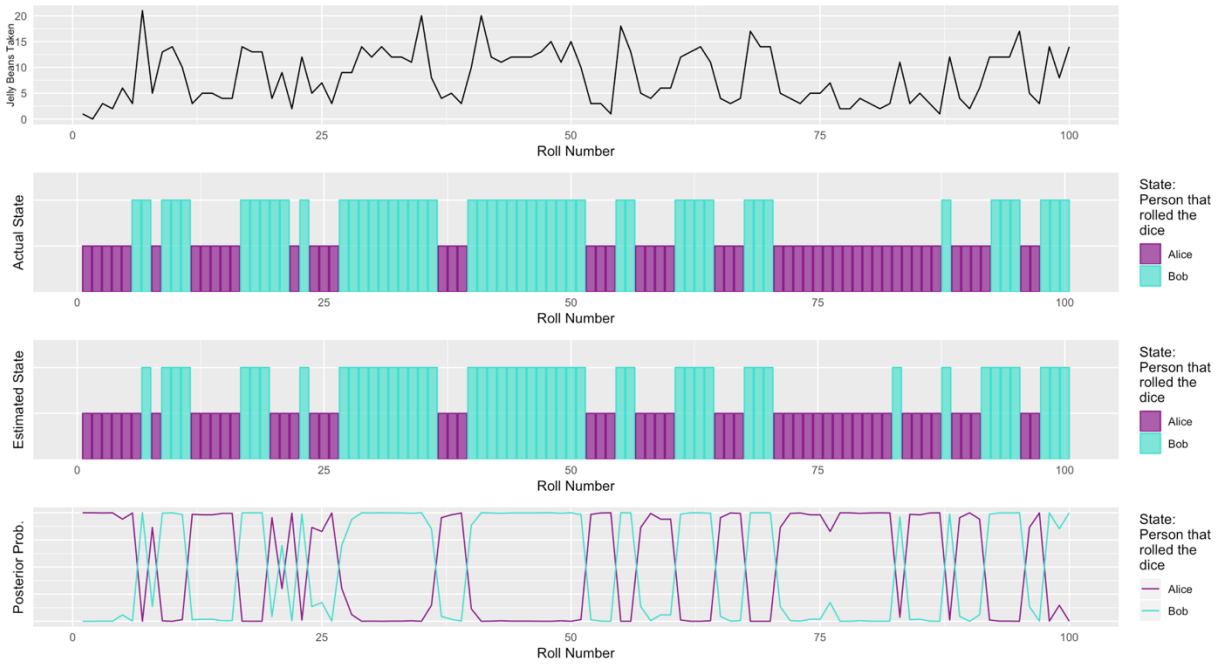

Fig. 18 - A plot showing the number of beans taken, the actual roller, the HMM predicted roller, and the probability of each player rolling for each roll number in the game. The HMM is able to identify the actual roller with near perfect accuracy. (Oehm 2018)

The HMM does a near perfect job identifying the actual roller from just the number of beans taken from the jar. In the setup of the simulation, Alice takes on average 4 jellybeans while Bob takes 12 on average. While this is a relatively large difference that makes classification of states easier, this information is not known by the HMM.

In an analogous manner, an HMM can be used to detect states (representing lithologies) from surface drilling data. If the surface drilling data is related to rock properties (beans removed from jar), and rocktypes have varying rock properties (number of beans removed on average for each player), HMM can be used to identify states within the drilling data that correspond to distinct rock properties. Choosing the appropriate number of states is important and a trial-and-error approach should be used to determine the number of states that maximizes variance between states. In this work, three states were determined to be optimal.

Chapter 3: Application to the Volve Oilfield

This chapter provides a demonstration of the methods described in Chapter 2 to the Volve oilfield in the North Sea. The Volve field is an offshore decommissioned oilfield off the coast of Norway developed by Statoil and data related to this field is now publicly available (Equinor 2018). The field was discovered in 1993 by Statoil and commercially developed in 2008 and produced until 2016 (Equinor 2018). The field is estimated to have recovered 54% of the original oil in place at the point of abandonment, which was 3 to 5 years longer than the expected life of the field (Equinor 2018). The major formation in the field is the Hugin formation which is a sandstone formation of Jurassic age. A map of the Hugin field with the Statoil wells producing from the formation is shown in **Fig. 19**.

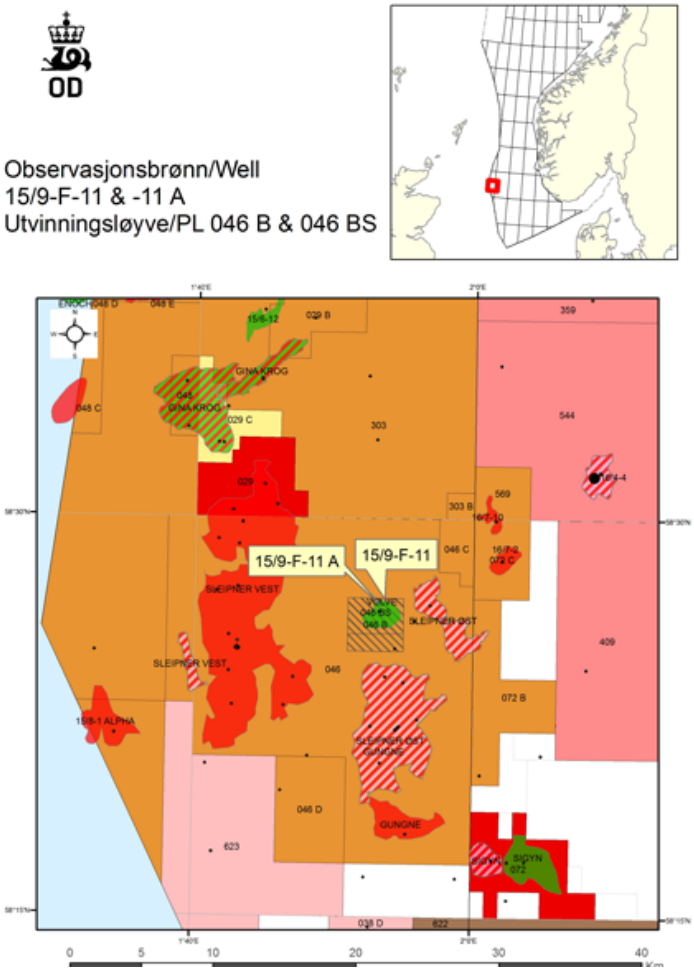

Fig. 19 - A map showing the Volve field off the coast of Norway. The location of the Statoil discovery well, the 15/9-F-11, is shown. (Melburg 2013)

The rock is made of fine to coarse grained sand with varying amounts of clay. There are variations in reservoir quality controlled primarily by grain size and clay content that affects the porosity and permeability (Equinor 2018). The publicly available dataset for the Volve field includes real-time surface drilling data, as well as open-hole logging data for 15 wells. Additional data available includes geological description and core data supporting the description of the formation.

In the workflow applied here, I first perform rocktyping from log data. I then compare the drilling-derived states to these rocktypes and also verify the correspondence between the change

points and transitions between rocktypes. K-means clustering is often used to derive electrofacies/rocktypes from well logs (Gupta et al. 2017) and defines groups/clusters in multivariate data sets that have similar properties (MacQueen 1967). Optimal groupings result in low intra-cluster variance and high variance between clusters (MacQueen 1967). K-Means clustering is displayed in an example with the Iris data set ((Fisher 1936) shown below in **Fig. 20**.

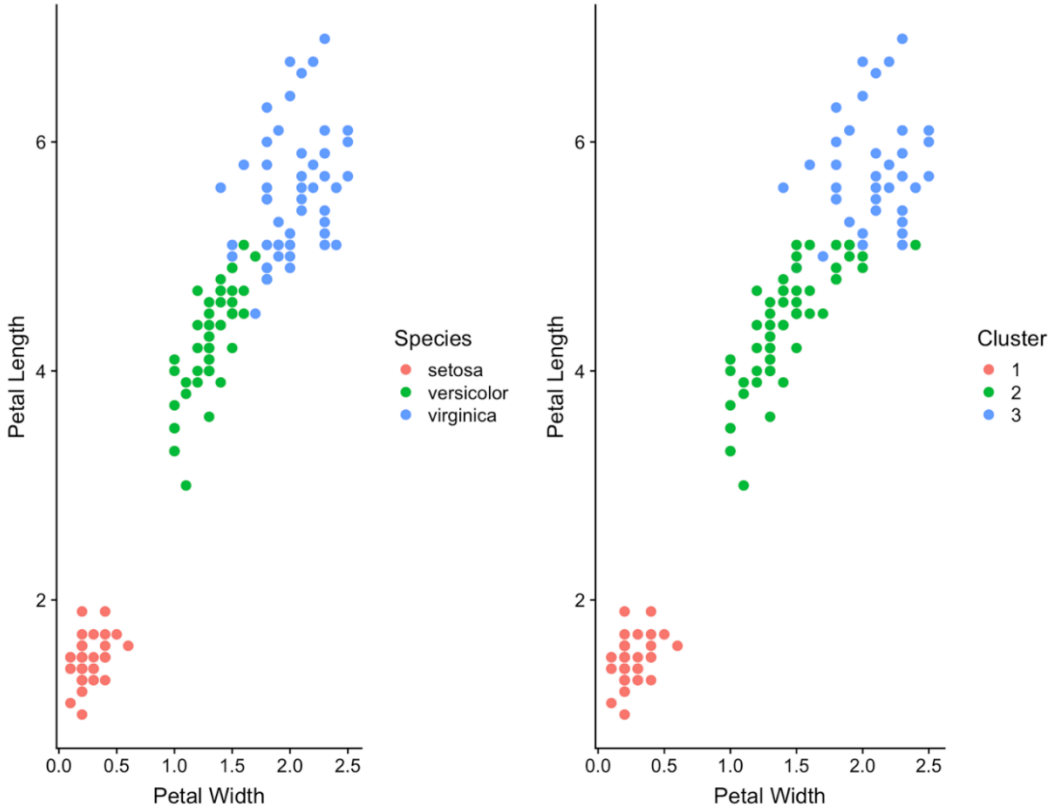

Fig. 20 - An example of K-Means clustering for the Iris data set showing the petal width and length for three species of Iris flower. The K-Means clustering algorithm creates groups with low variance within clusters, and high variance between clusters. (Fisher 1936)

Following rocktyping using K-means, I define drilling states from the Hidden Markov Model (HMM) and identify change points with the change-point detection algorithm (CPD). I then check the correspondence between the HMM states and log-derived rocktypes and the occurrence

of change points with transitions from one rocktype to another. I present the results and analysis of the described workflow below for Well 15-9 from the dataset. Results for additional wells are displayed in Appendix A.

Fig. 21 shows the box plots corresponding to rocktypes derived from K-means clustering of the bulk density (RHOB), gamma ray (GR), and compressional sonic travel time (DT) logs.

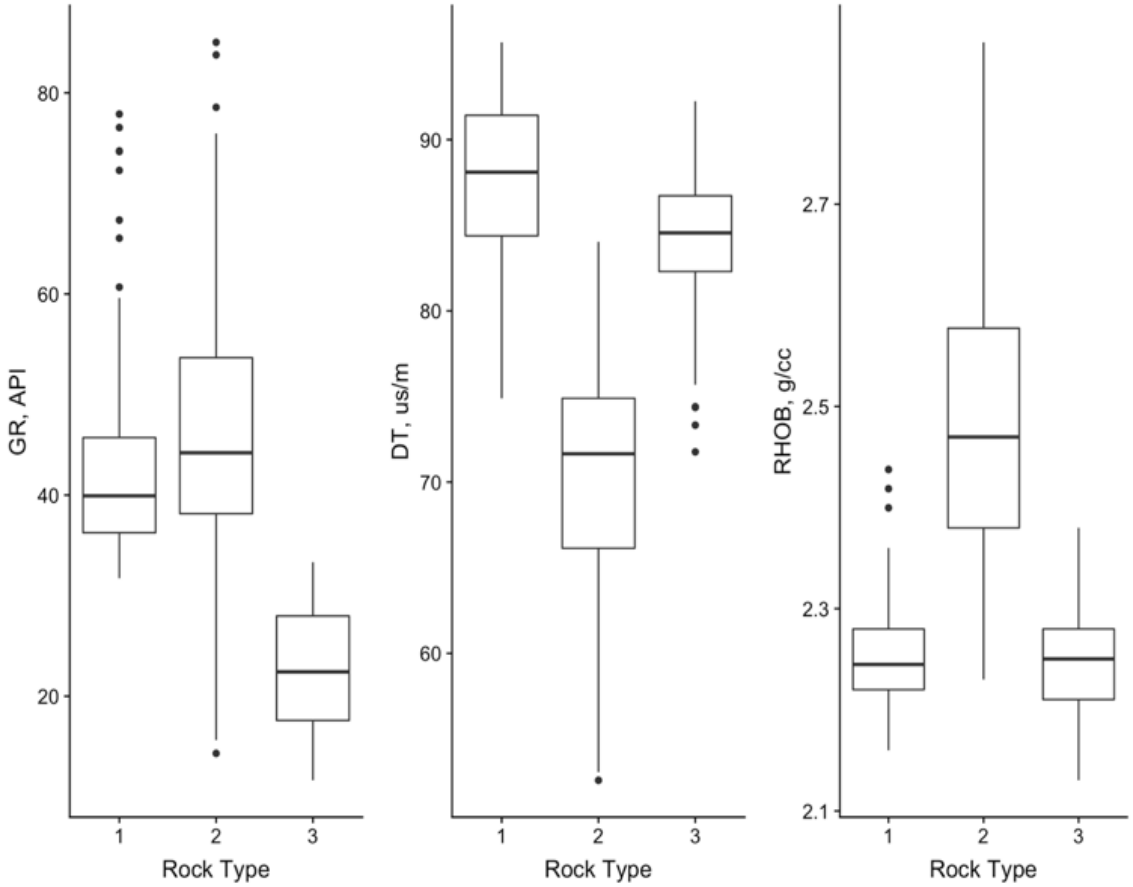

Fig. 21 - Boxplots showing the distribution of log data for each of the rocktypes derived from K-Means clustering. Each rocktype has distinctly different signatures.

3.1 Application of HMM to Surface Drilling Data

Fig. 22 shows boxplots of the drilling data corresponding to each of the states derived from an HMM model while **Fig. 23** shows a scatterplot matrix of the drilling data colored by the HMM states. The HMM-derived states show distinctly different signatures in terms of the surface drilling data.

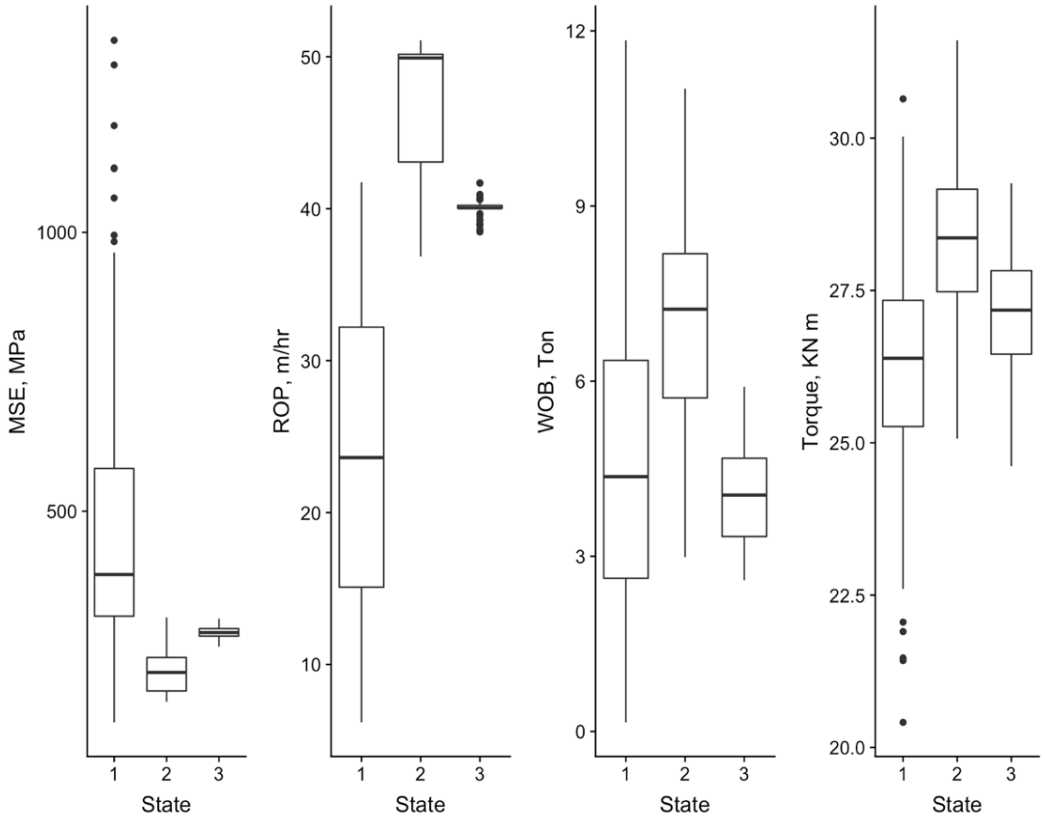

Fig. 22 - Boxplots showing the surface drilling data for each of the HMM derived states. The HMM states correspond to different drilling data signatures.

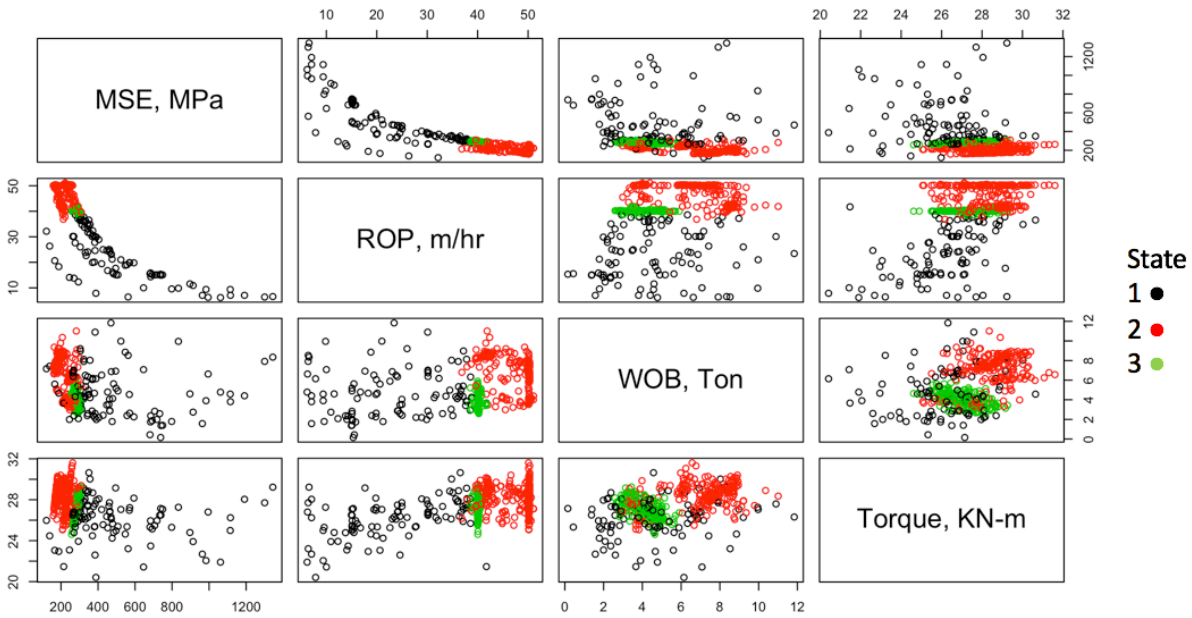

Fig. 23 - A scatterplot matrix of the surface drilling variables colored by HMM states.

Transitions in the drilling-derived states potentially indicate that some variable or combination of variables from the drilling data are responding to some change in formation properties. To investigate this, I plot the distribution of log data corresponding to each drilling-derived state in **Fig. 24**. The differences between the HMM-derived states seen in **Fig. 24** indicate that surface drilling data can distinguish between zones of different properties and can therefore provide real-time lithology-related information while drilling.

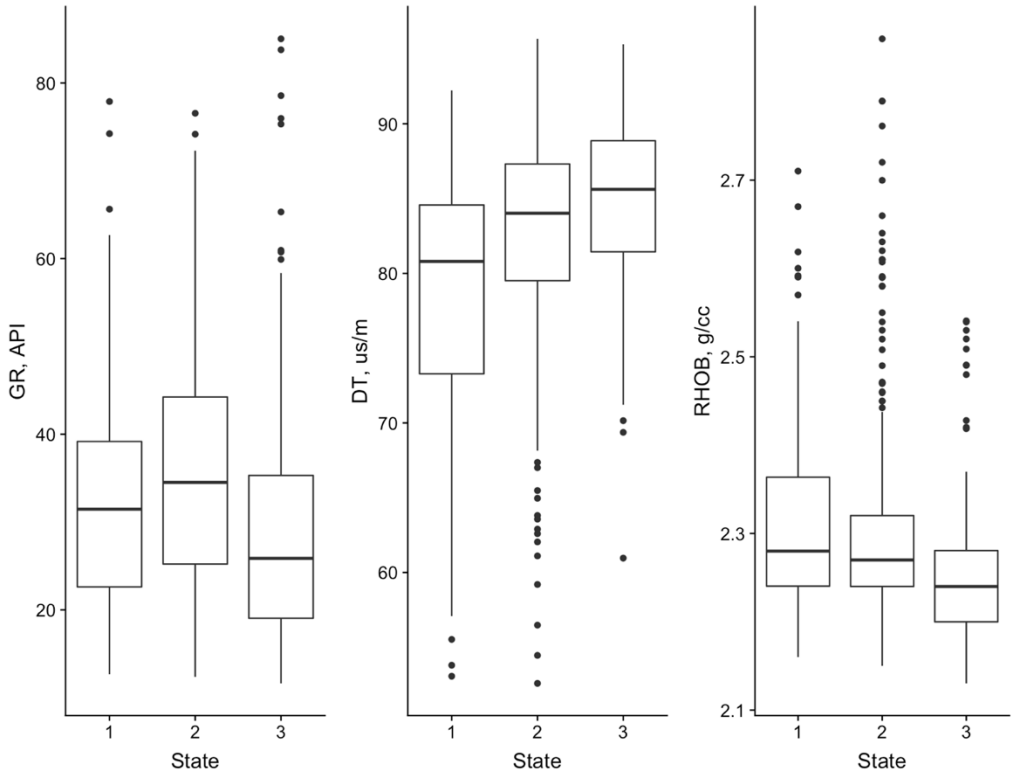

Fig. 24 - Boxplots showing the distribution of log-derived properties drilling-derived HMM states. A good separation is observed on the DT and RHOB logs.

The HMM state transitions represent a change in drilling operations or formation properties that can potentially be utilized to improve drilling efficiency. **Fig. 25** shows a plot of the ROP and MSE corresponding to the different drilling states inferred by HMM showing that a knowledge of the drilling state a priori can help optimize ROP and/or minimize damage and wear to the bit. It is however important to note that the drilling states do not show a one-to-one correspondence with the log-derived rocktypes which will be discussed in greater detail later.

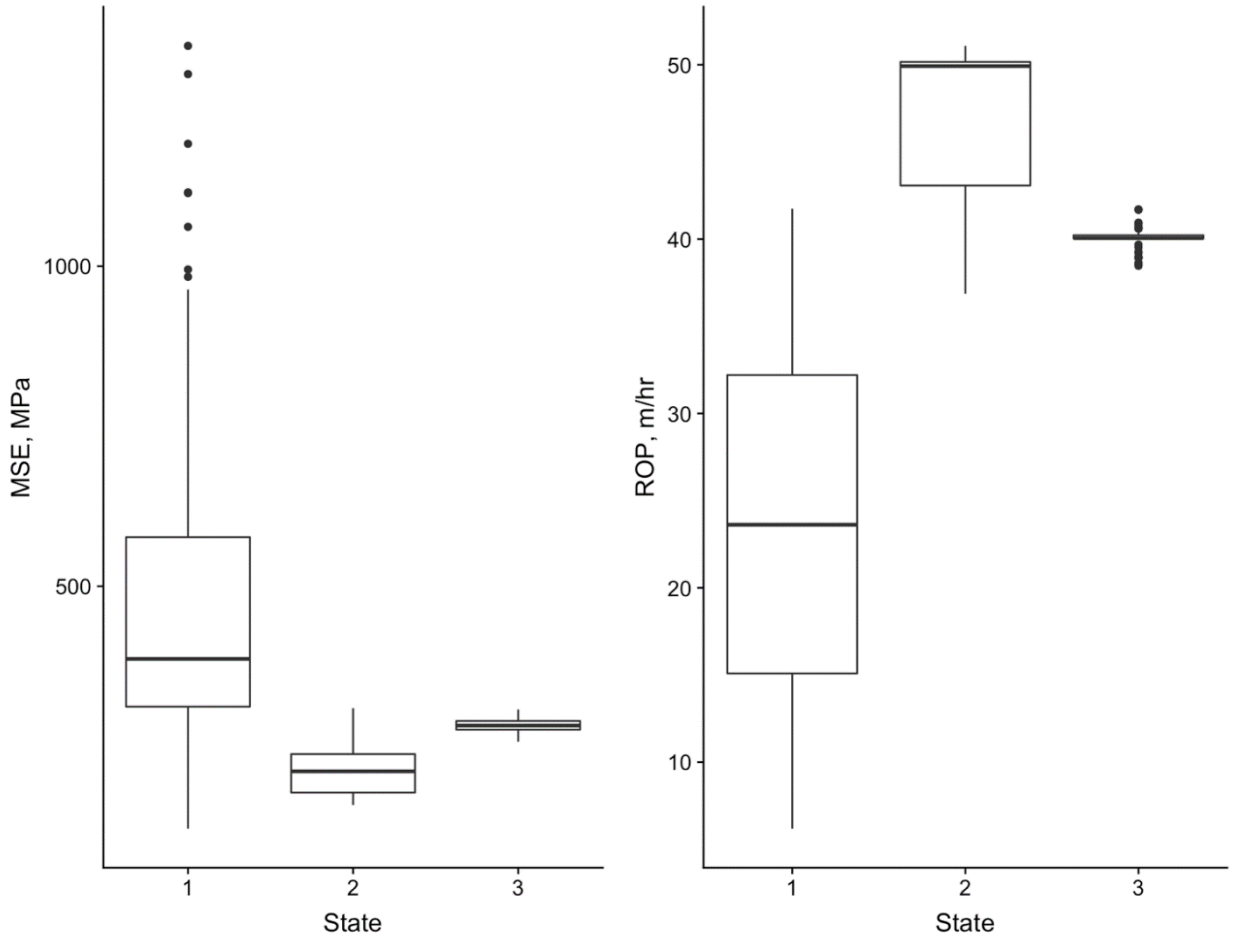

Fig. 25 - Boxplots showing the MSE and ROP response for the surface drilling-derived HMM states. These variables are used to monitor drilling efficiency and the HMM states show good separation for both variables. Monitoring HMM states can be used to optimize drilling efficiency.

3.2 Application of CPD to Surface Drilling Data

I then apply a change point detection algorithm to the surface drilling data to detect changes in drilling modes. I use a moving window approach with linear penalty and mean-shift detection (Basseville and Nikiforov 1993, and Sen and Srivastava 1974) on the multivariate surface drilling time-series signal. The linear penalty value for the change point detection algorithm (as described in Chapter 2) used for this well is 0.005, and the size of the sliding window is 25 data points (0.5

ft. each) for each of the two windows (12.5 ft. window). **Fig. 26** shows a visualization of the sliding window overlaid on a plot of MSE vs. Depth.

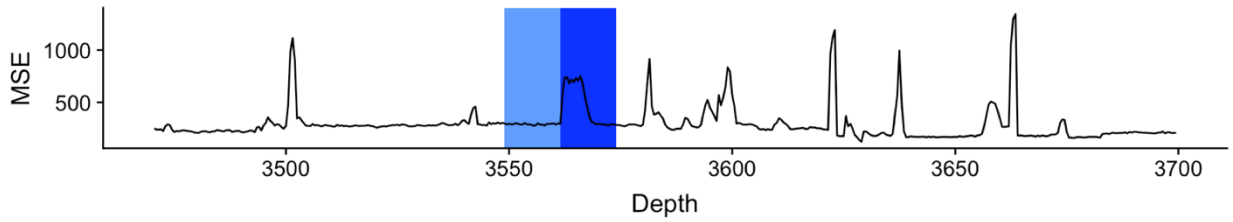

Fig. 26 - A plot of MSE vs. Depth for well 9. The blue shaded boxes represent the sliding window used for the search algorithm for CPD. The two windows are 12.5 ft. wide. These windows slide along the signal and the mean of the signal is calculated in both windows and compared to identify change points. When the difference of mean of the two windows is large, the discrepancy measure is high and a change points is detected.

Change points are derived from the multivariate signal of surface drilling data using the approach described in Chapter 2. I then compare the occurrence of change points to the state-transitions derived from the HMM model as well as the log-derived rocktypes.

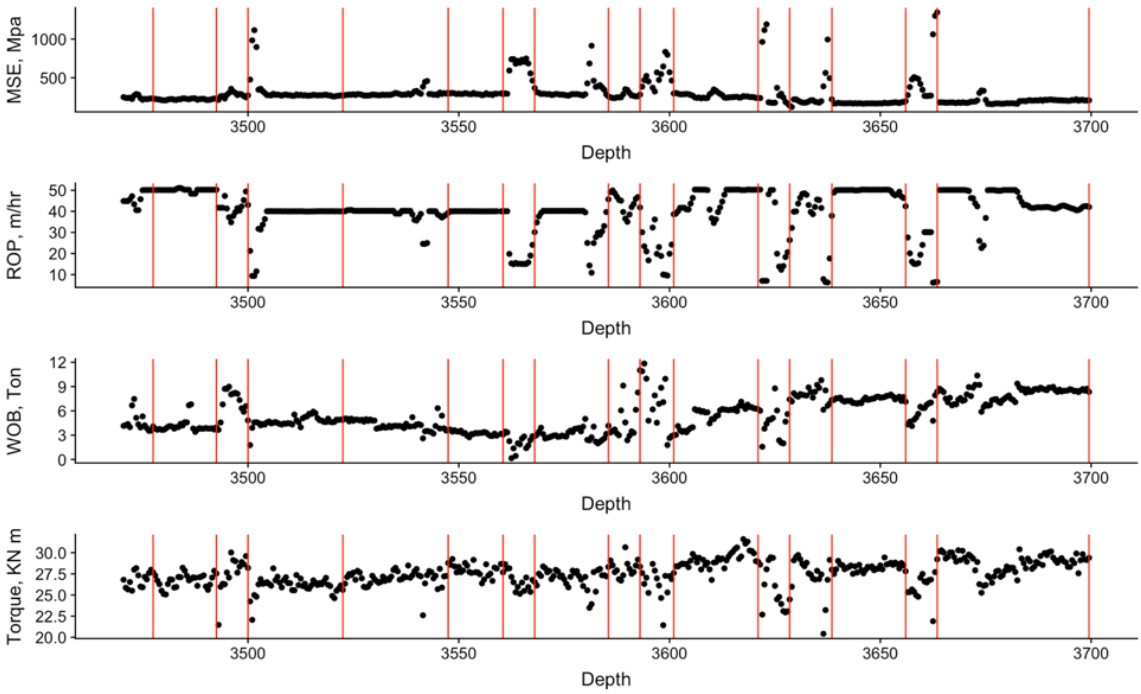

Fig. 27 – Change points in red vertical lines overlain on the surface drilling variables. Each change point can be visually correlated with an abrupt change within one or more of the input signals.

For Well 15-9, **Fig. 27** shows the change points detected in the drilling signal as vertical red lines. The change points are detecting a shift in the mean of the multivariate signal and, in this specific example, can also be picked up visually.

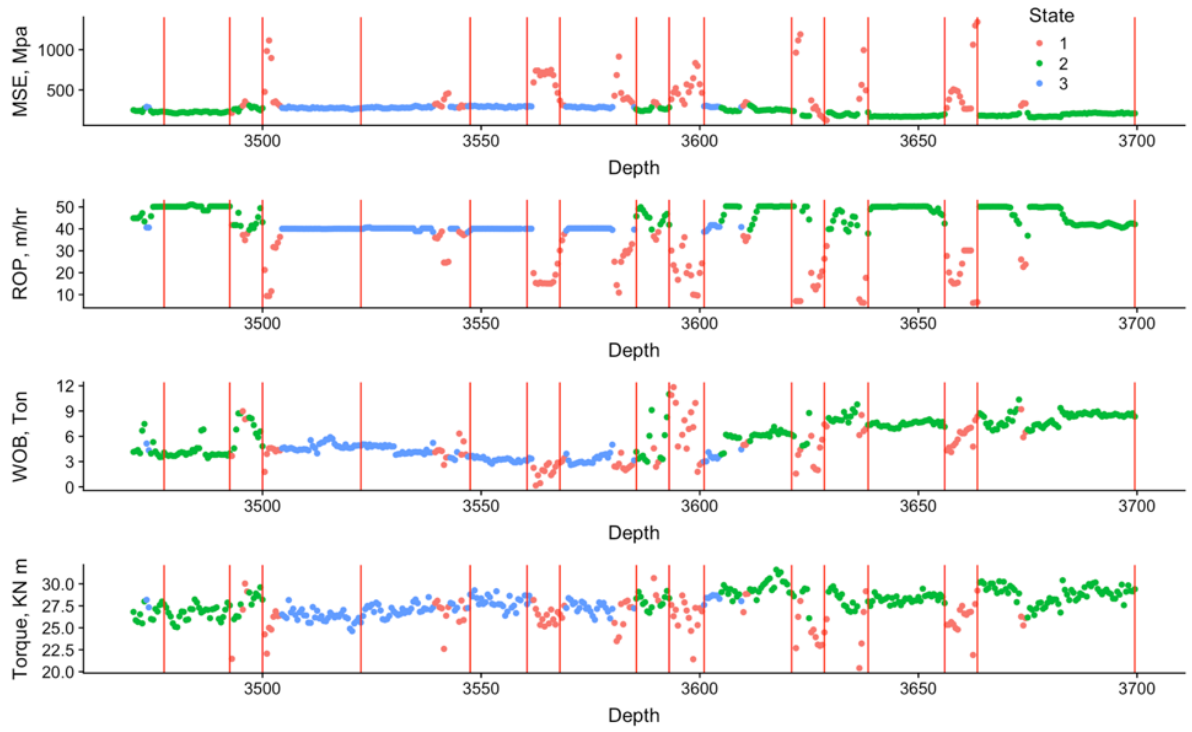

Fig. 28 - Plots of the real-time drilling signal colored by the HMM states. In many occurrences, the change points detect the shift from one state to another, for example the change points picked at 3655 ft. and 3660 ft. identifying the transition from state 2 to 1 then back to 2.

Fig. 28 shows the drilling data colored by the HMM states with the change points shown in vertical red lines to assess whether change points correspond to changes in the drilling states. The change points align very well with the locations where there are transitions in drilling states providing a measure of confidence that both approaches are responding to similar changes in drilling modes and/or formation properties. As described in Chapter 2, the HMM calculates a probability for each state at each prediction location. The state with the greatest predicted probability is taken to be the predicted state. For Well 15-9, the state probabilities are shown in **Fig. 29**. The states are estimated from the predicted probabilities shown below. In this example, the HMM is able to predict states with high confidence with the predicted state probabilities at most locations near 1.0.

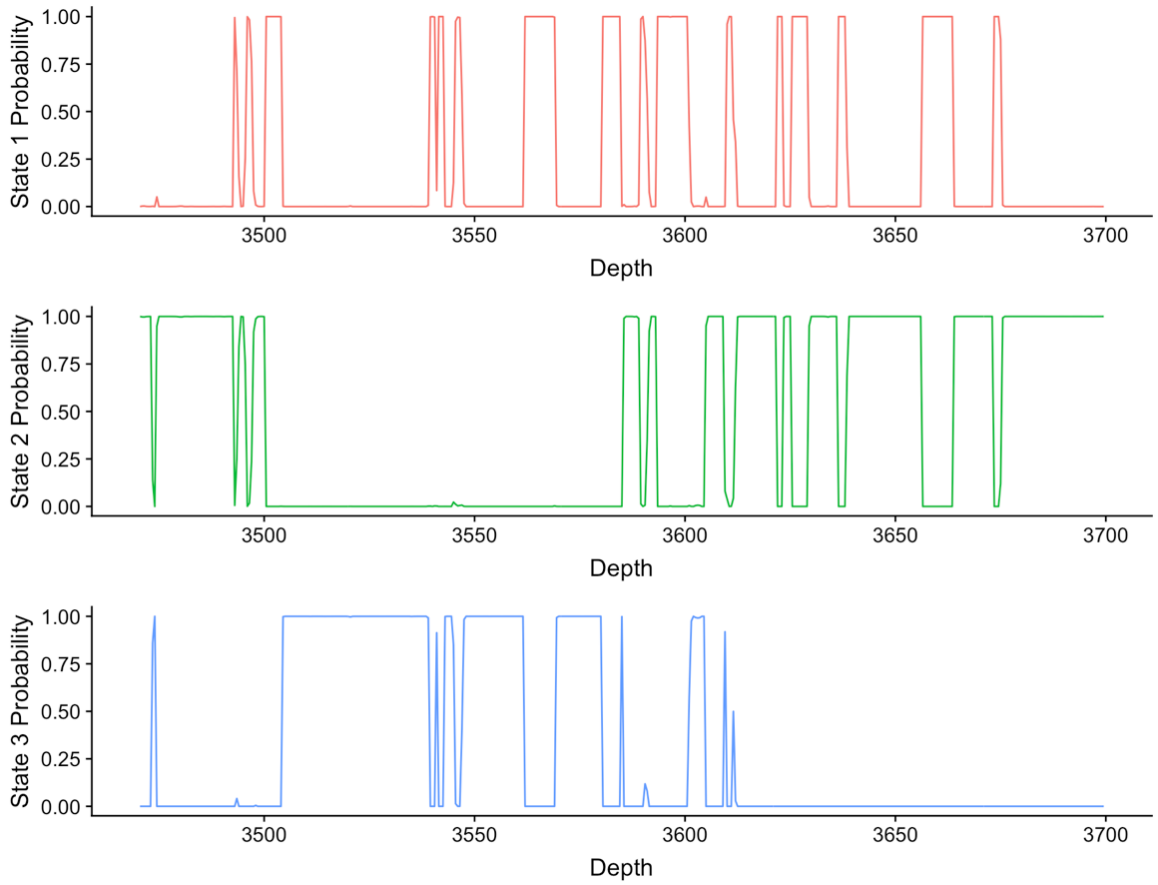

Fig. 29 - Plots showing the predicted state probabilities for each prediction depth. At each location, the state with the highest predicted probability is taken as the predicted state. The HMM is able to predict states with high confidence.

Change points detected from surface drilling data in Well 15-9 are also compared with the K-means predicted rocktypes. **Fig. 30** shows the log data colored by K-Means derived rocktypes to compare with the surface drilling-derived change points.

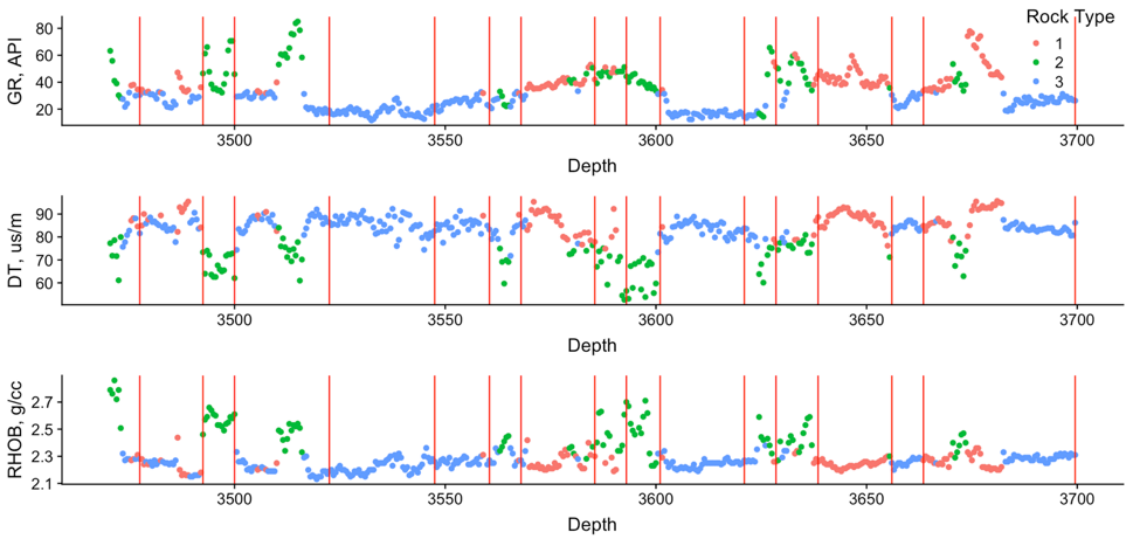

Fig. 30 - Plots showing the logging data from Well 15-9 used to derive the K-Means rocktypes with the change points shown in red vertical lines. In many occurrences the change points are able to accurately identify the changes in rocktypes such as from 3630 ft. to 3660 ft. and 3670 ft. where the rocktype transitions from 2 to 1 to 3.

The change points show good correspondence to transitions from one rocktype to another. The correspondence is not exact because as mentioned earlier, surface drilling data is often not responsive to formation properties alone but also operational considerations such as mud weight, driller experience, depth of the well and other factors. For instance, between 3650 ft. and 3700 ft., no change point is detected although there are apparent transitions in rocktype (**Fig. 31**).

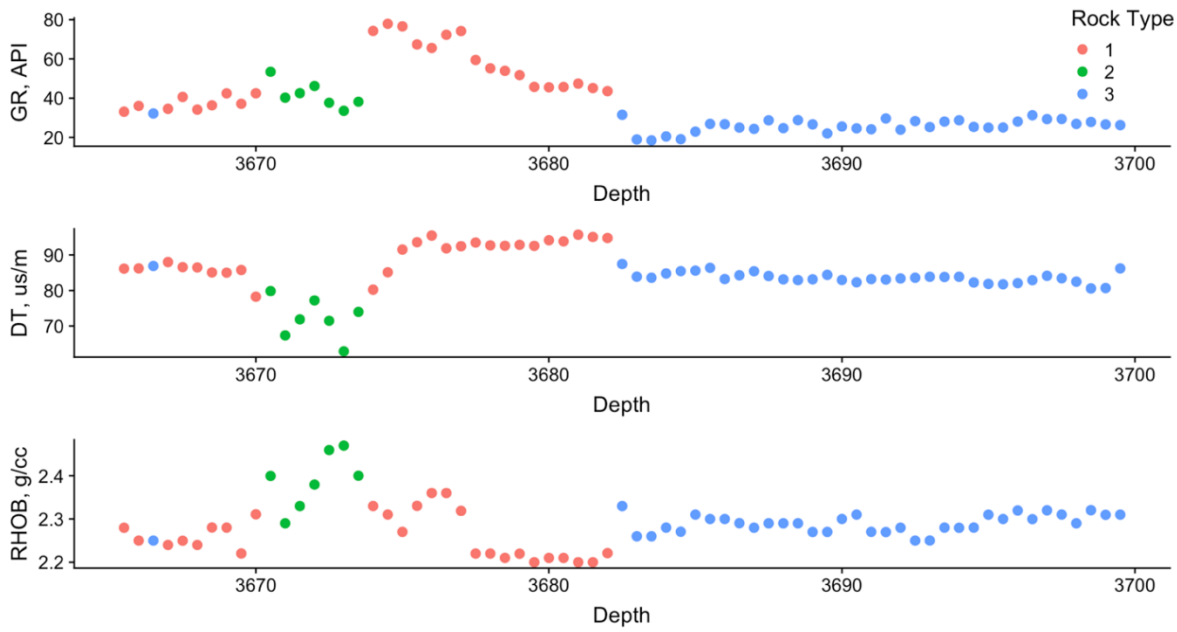

Fig. 31 - The log data for well 15-9 used for rocktyping, shown over the interval 3665 ft. to 3700 ft. where there are transitions in rocktype that are not detected by the CPD algorithm.

Fig. 32 shown below is a comparison of the HMM derived states, K-Means derived rocktypes, and change points for Well 15-9 on a plot of GR versus depth. For both HMM states and rocktypes, the change points do a good job of identifying transitions from one state/rocktype to another. HMM state transitions and change points do not always correspond with rocktype transitions because surface drilling data can vary for operational reasons also. Nonetheless HMM states and change points do correspond with many of the rocktype transitions.

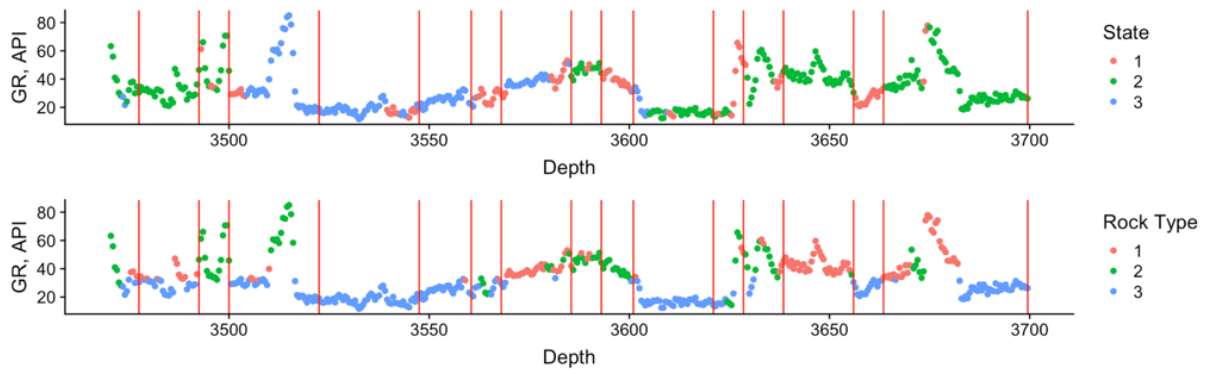

Fig. 32 - Plots showing the GR response colored by the HMM states and rocktypes for comparison of states to rocktypes. While the state transitions often correspond with the changes in rocktype, the particular state does not always align with only one rocktype.

Section 3.3 Summary

In this chapter, I demonstrated the use of K-Means clustering to define rocktypes from open hole logging data, HMM to derive drilling states from surface drilling data, and CPD to identify changes points in the surface drilling data signals. The following conclusions can be drawn from this application.

1. HMM is able to define groups with distinct formation properties using surface drilling data.
2. CPD can identify change points within the surface drilling signal that correspond with both the HMM states and the rocktypes derived from logging data.
3. Though HMM state transitions and change points have good correspondence, the correspondence with rocktypes is weaker likely because of operational changes as opposed to changes in formation properties.

The practical significance of this work is that it provides a reliable, real-time approach to detecting changes in formation properties using surface drilling data. While the accuracy is only around 75% in identifying rocktypes or detecting rocktype transitions, this workflow provides

instantaneous feedback to the operator about formation properties and potentially can be used to steer the bit to stay in zone in horizontal wells. The next chapter discusses the application of these methods to horizontal wells in an unconventional reservoir in Oklahoma.

Chapter 4: Oklahoma Unconventional Reservoir

The area of interest in this study consists of the Mid-Continent STACK in the Anadarko Basin in Oklahoma. The formation of interest is the Meramec formation that is characterized by varying lithology and a high heterogeneity.

The Meramec series in the Sooner Trend of the Anadarko Basin of Canadian and Kingfisher counties (STACK) are primarily siliclastic systems of quartz and carbonate deposited as fine grained sandstones to mudstones. These rocks were deposited in varying conditions throughout the Mississippian that produced a range of lithofacies. Along the shallow ramp of the Anadarko basin, deposits slowly transition from shallow-water deposits to deep-water, increasing the overall proportion of shale volume (Dutton 1985). Gradual dipping basinward follows a trend of increasing porosity as clay content increases (Miller 2018). In addition to depositional variation within the basin, sequences of sea level rise and fall introduce depositional cycles of coarsening or fining upward strata. The strata deposited in the lower Mississippian show characteristics of greenhouse conditions with lesser sea-level fluctuation, while the upper Mississippian is more representative of an icehouse environment with large sea-level variation (Haq and Schutter 2008). These sea-level variations create fining upwards sequences when deposited during sea-level rise, and coarsening upward when sea-level is falling that produce varying lithofacies with more calcareous facies distally and chert rich facies proximally (Drummond 2018). For the more distal facies, higher porosity is found in rocks with greater clay content that reduces calcite cement, and proximally diagenetically altered chert represent the higher porosity rocks (Drummond 2018). Therefore, detecting presence of clay in the distal deposited facies from logs or drilling data can improve landing zone identification by targeting higher porosity zones.

Data for the Meramec formation is provided by Marathon Oil including MWD data, real-time surface drilling data, and 3D geological modeling data of the study area. Additionally, one set of triple combo logs are available in the vertical section of one well. The geomodel is populated with rocktypes using a typical rocktyping workflow based around defining rocktypes on porosity and permeability, as well as mineralogy from core data, then up-scaled to well log data and populated throughout a geomodel. The wells in this study are horizontal wells, and the workflow will be applied to surface drilling data recorded during the drilling of the horizontal section of the wells. Drilling derived states, and change points, will be assessed with MWD-recorded gamma ray, as well as rocktypes obtained from the 3D geomodel and well deviation surveys. The population of the geomodel with rocktypes is done with grid cells of 110 ft. in length and width, and 2 ft. thickness. Without logging data available through the lateral, I use rocktype logs generated from the geomodel to compare to the drilling states for state to rocktype comparison. An important point to note is that the rocktypes from the geomodel are likely associated with huge uncertainties because of the lack of nearby well control. I also use MWD-recorded gamma ray along the lateral to compare HMM states to formation properties. The drilling data I used for this analysis from the surface drilling data includes ROP, WOB, MSE, torque, and differential pressure (DP).

4.1 Vertical Section with Triple Combo Log

One of the wells in the study area has a triple combo log available through the vertical section. I first use a simple K-Means approach on the surface drilling data to derive 3 drilling clusters. **Fig. 33** shows a scatterplot matrix of the drilling variables colored by the different K-Means clusters.

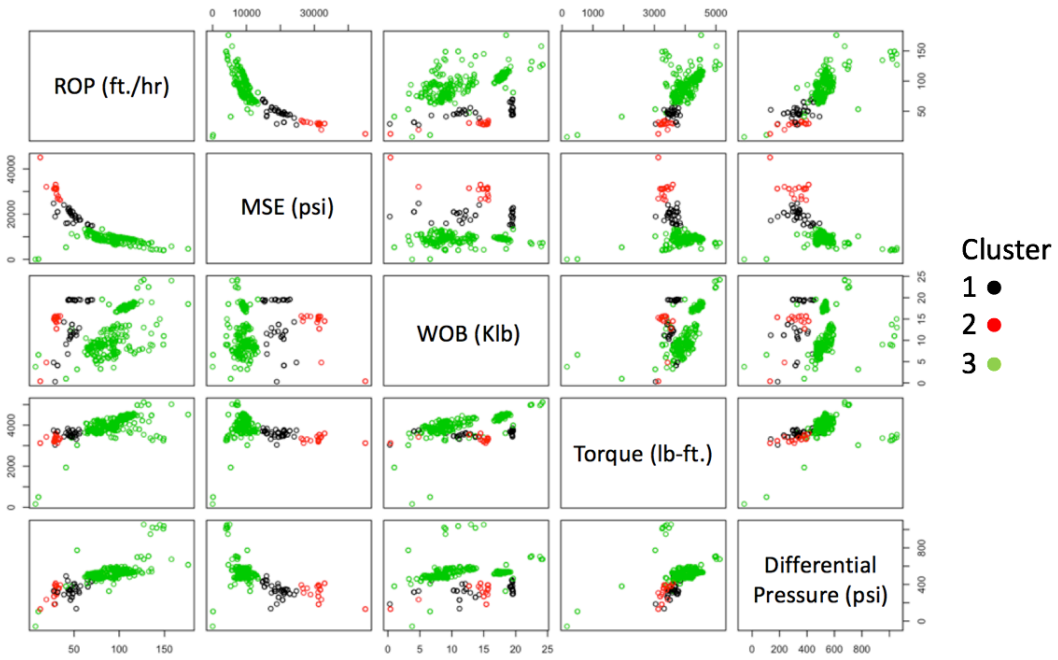

Fig. 33 - A scatter plot matrix showing the drilling variables and the K-Means derived clusters.

The clusters are able to achieve distinct groups in the drilling data that typically range from high to low values for each variable, like MSE for example. The formation properties corresponding to the clusters are shown in **Fig. 34**.

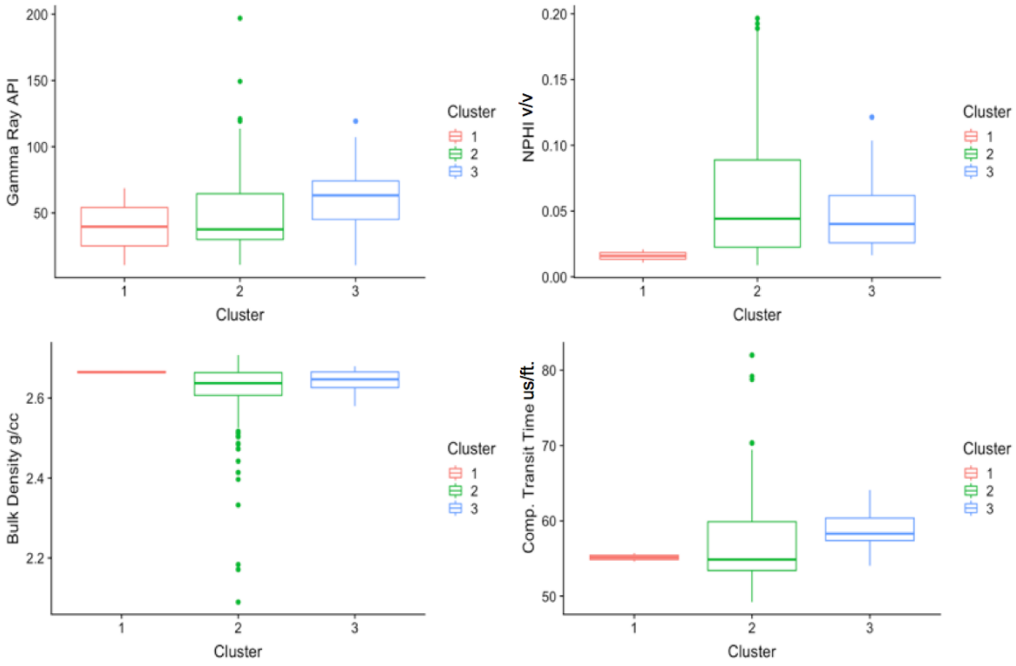

Fig. 34 - Box plots showing the distribution of open hole logging data for each of the K-Means derived drilling clusters. The drilling clusters do a poor job of differentiating formation properties measured by the four logs, and show that more advanced methods will need to be tested.

While the unsupervised K-means algorithm does satisfactorily in identifying zones with different log signatures, I will demonstrate at a later stage in this chapter, that the HMM-based approach shows a better performance. Next, I use the drilling variables to train an HMM model and derive three drilling states. These drilling variables can be seen in the scatterplot matrix in **Fig. 35** colored by HMM state.

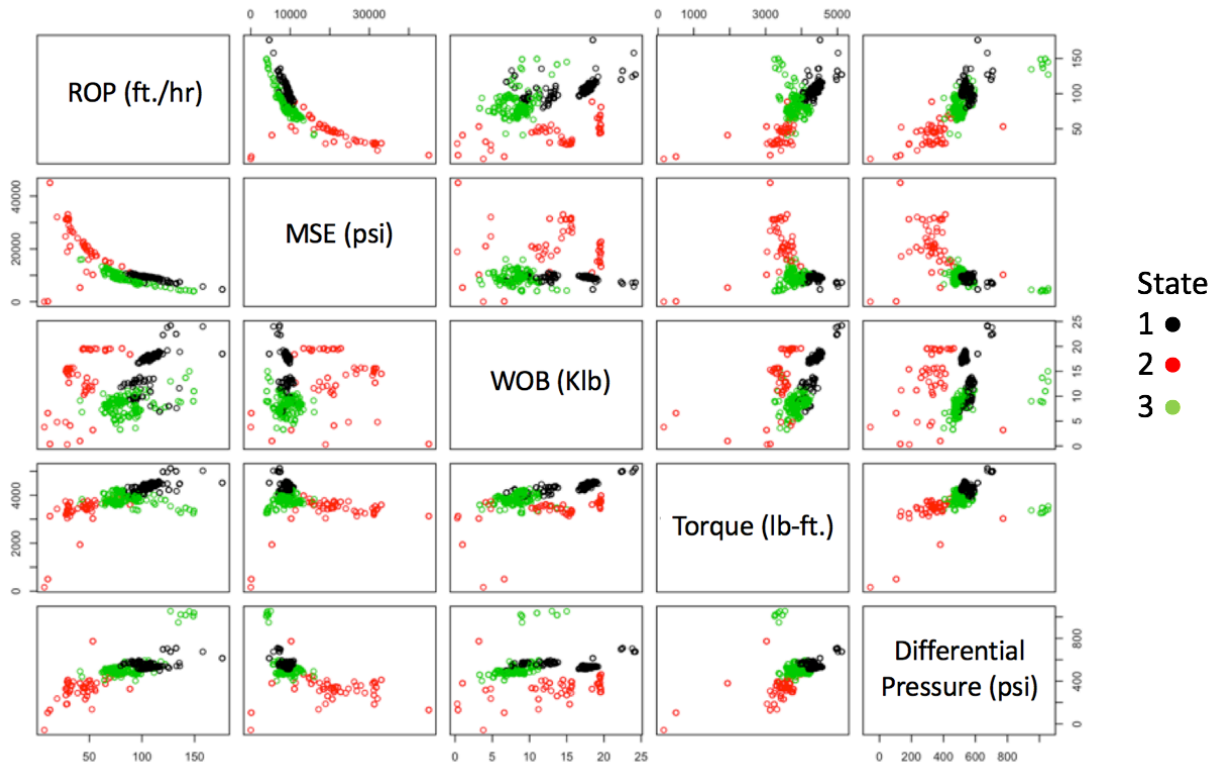

Fig. 35 - A scatterplot matrix showing the drilling variables colored by HMM state for the vertical section of one well.

The scatterplot of drilling variables derived from the HMM appears to capture the variability in the drilling behavior better than the K-Means approach. For instance, in **Fig. 36** below, in the crossplot of MSE and ROP, the K-Means clusters simply correspond to high to low values along the curve, while the HMM is able to pick the distinct groups belonging to different (nonlinear) trends. This will become more apparent when dealing with the lateral section of the well and will be discussed in greater detail in a later section.

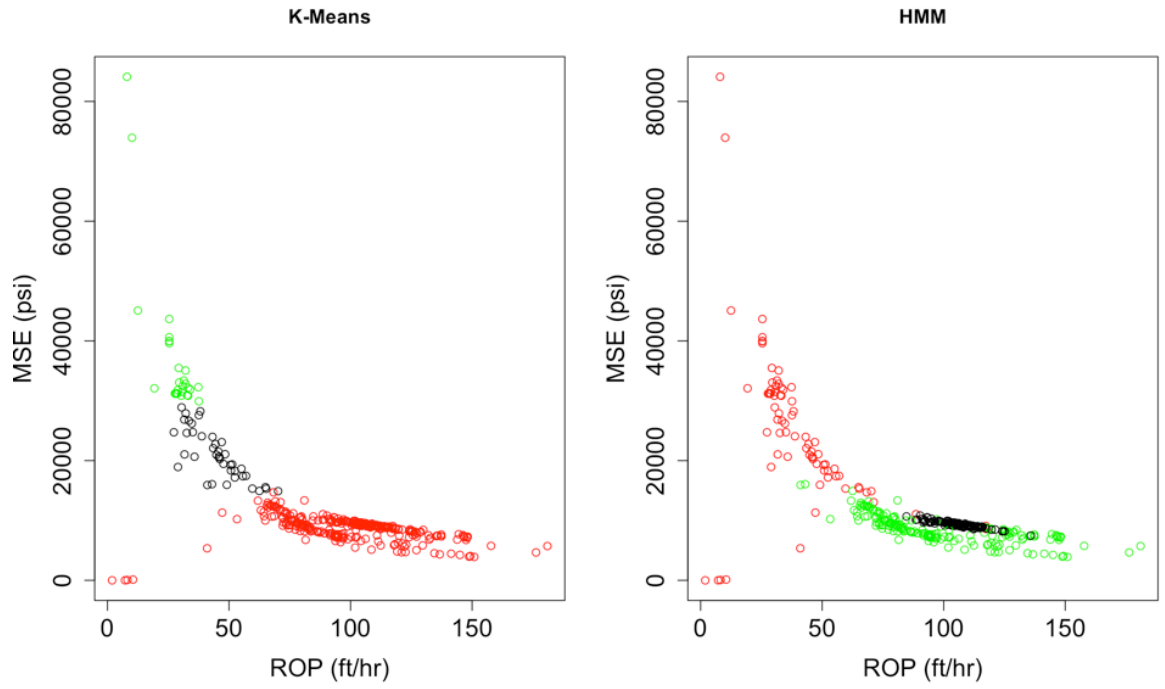

Fig. 36 - Crossplots of ROP vs. MSE for HMM and K-Means. For K-Means the clusters are picked by magnitude alone, and the HMM states correspond to different curves within the data.

The formation properties associated with the HMM derived states are shown below in **Fig. 37**. The HMM-derived states are able to achieve better distinction than K-Means in the log measured formation properties, such as gamma ray and compressional transit time. In this specific example, HMM-derived states show more distinctions than the K-Means approach and provides confidence in the use of HMM to derive drilling states corresponding to different formation properties,

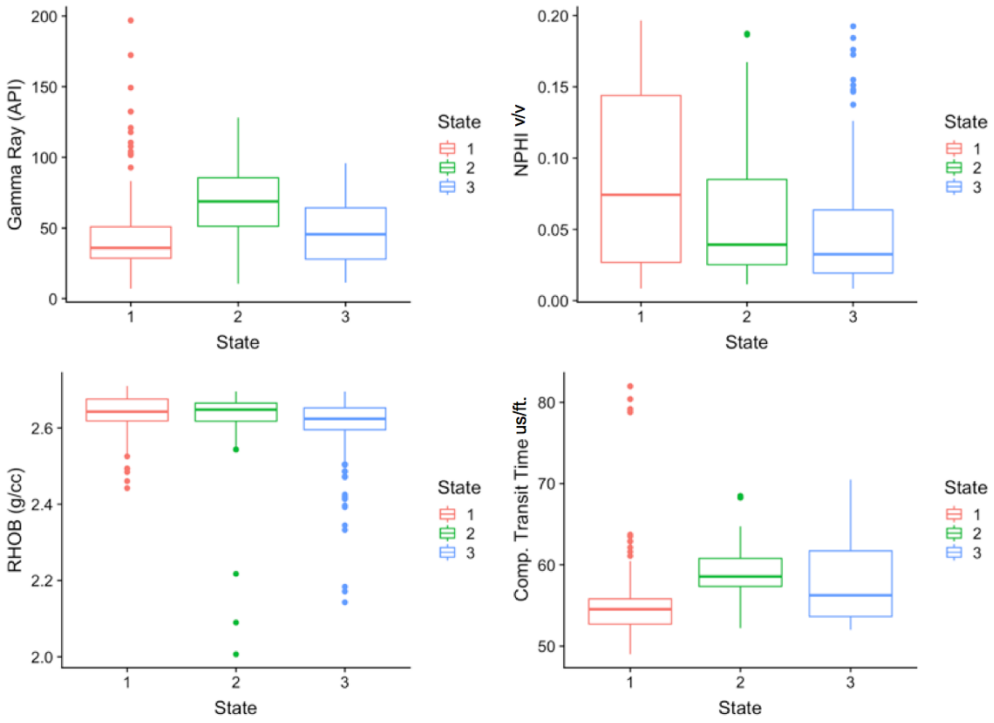

Fig. 37 - Boxplots showing the distribution of the open hole logging data for each of the HMM derived drilling states. Compared to the K-Means derived drilling clusters, the HMM states are able to provide a more satisfactory distinction.

4.2 Application to Horizontal Wells

The dataset comprises of surface drilling data corresponding to the lateral section of five wells and MWD-measured gamma ray values. Prior to analysis, I scale the MWD-measured gamma ray, and surface drilling variables using min-max scaling so the gamma ray values for each well are on the same scale to allow for direct well-to-well comparison. I first use a K-Means clustering approach of the surface drilling variables. **Fig. 38** shows the results from five wells on a scatterplot matrix of the drilling variables.

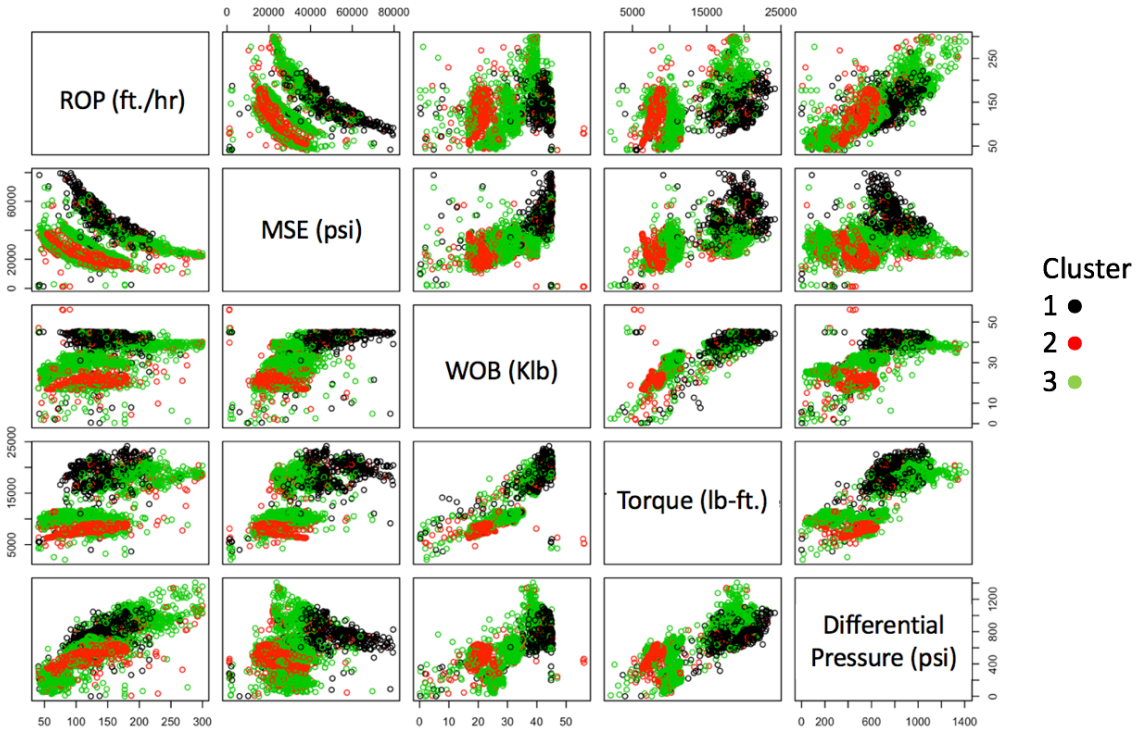

Fig. 38 - A scatterplot matrix for the drilling variables for the horizontal section of five wells colored by cluster. I observe several overlapping clusters.

In the plot of ROP vs. MSE, for example, the K-Means derived clusters have a significant amount of overlap resulting in very little distinction. To observe variation of formation properties for the various K-Means clusters, the MWD-measured gamma ray is shown on a boxplot for each of the three clusters in **Fig. 39**.

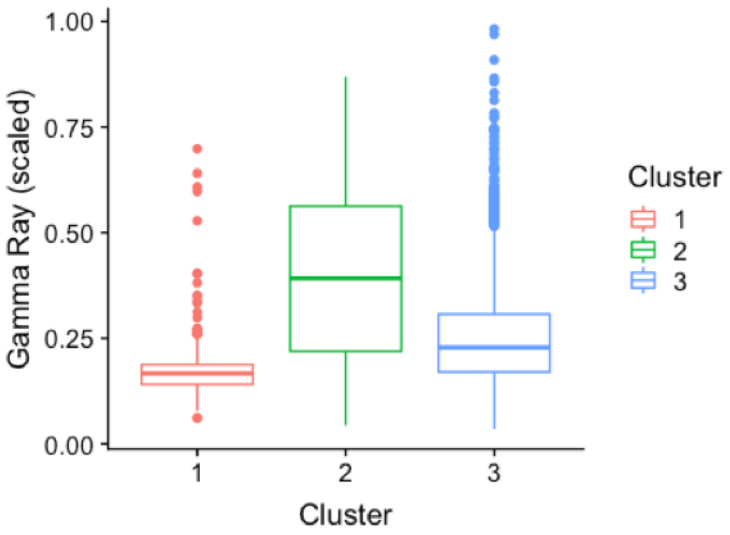

Fig. 39 - Boxplots showing the scaled MWD measured gamma ray from all five wells for each K-Means derived cluster. The clustering provides little distinction for gamma ray and thus unsatisfactory results.

Although there is some distinction in the measured gamma ray which underscores the viability of K-Means to identify lithological changes from surface drilling data, the overlapping ranges in the drilling variables seen in **Fig.** 38 will render the grouping susceptible to noise in the drilling data. I therefore follow this work with training a HMM model to detect drilling states and a corresponding matrix scatterplot is shown in **Fig. 40** colored by the different drilling states.

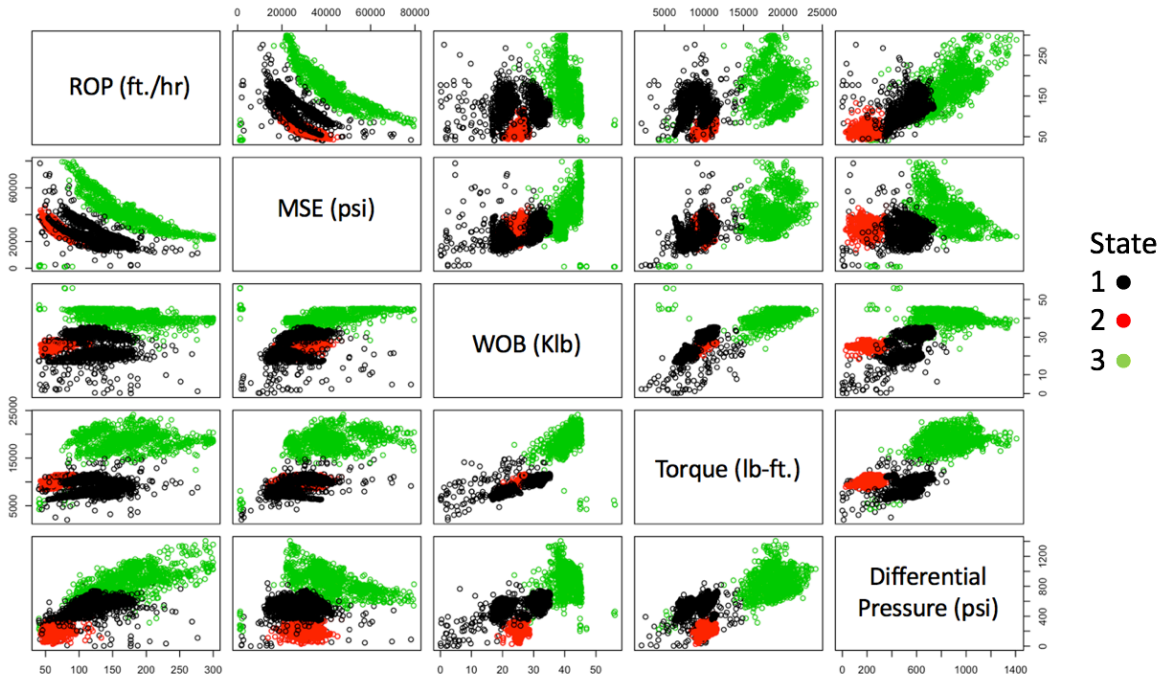

Fig. 40 - A scatterplot matrix for the five wells colored by the HMM states. The HMM derived drilling states are able to capture the different behavior of the drilling data much better than K-Means. In the plot of ROP vs. MSE, the HMM states are able to identify the different curves with significantly less overlap.

The figure above shows that the HMM capture the different behavior of the various drilling modes with better fidelity than K-Means. In **Fig. 41** I show the crossplot of ROP and MSE comparing K-Means and HMM results indicating the superior performance of the HMM to capture variations in drilling variables.

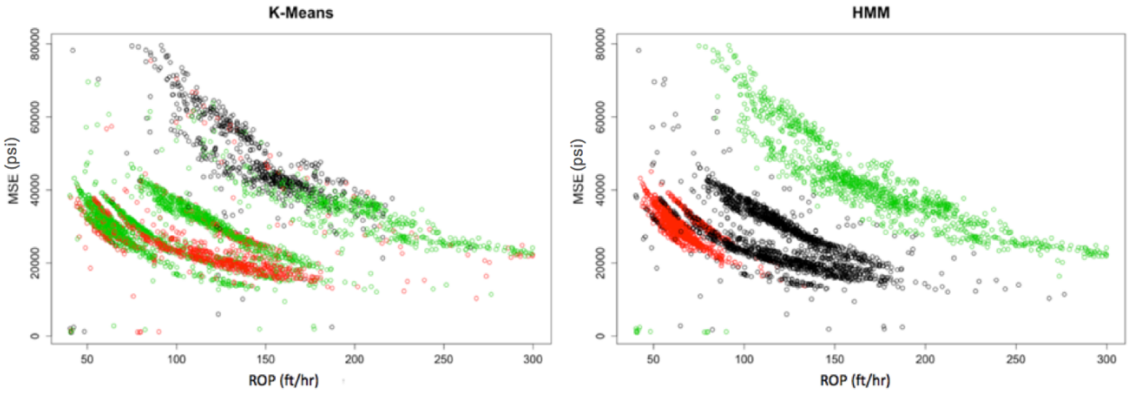

Fig. 41 - Crossplots of ROP vs. MSE for the five horizontal wells colored by K-Means clusters or HMM states. This shows how the HMM states are able to describe the behavior of the drilling data much better than K-Means, with significantly less overlap between clusters.

Additionally, the gamma ray values corresponding to the HMM states show much better separation as seen in **Fig. 42** when compared to the K-Means clusters. This offers a promising approach to geosteering based on drill string/bit performance as defined by HMM states. In this specific play, higher gamma ray values are associated with higher clay content and porosity and are considered to be the more productive intervals. Consequently, by using HMM, I am able to detect distinct rocktypes corresponding to different gamma ray values from surface drilling data alone.

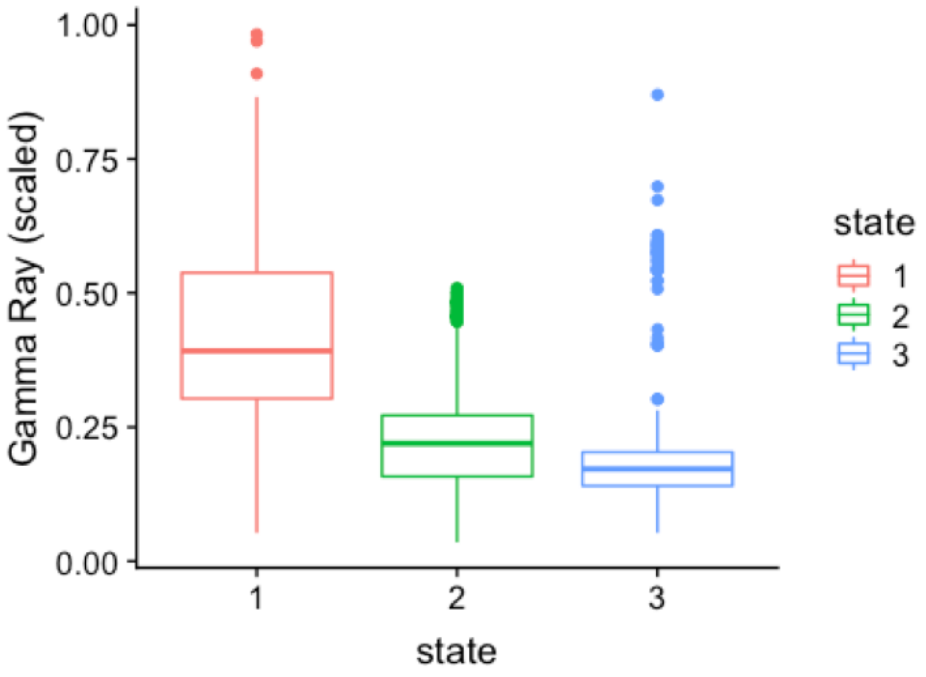

Fig. 42 - Box plots showing the MWD-measured scaled gamma ray from each of the wells for each HMM state. HMM is able to achieve much better distinction of gamma ray than K-Means.

I also apply the CPD algorithm to the surface drilling data to assess the correspondence of change points with transitions between the HMM-derived states and geomodel-derived rocktypes. Unfortunately, the surface drilling data is associated with considerable noise rendering the CPD algorithm overly sensitive to random fluctuations within the data. To mitigate this problem, the drilling signals are smoothed with a 110 ft. moving average which is the same width as the grid cells in the 3D geomodel. Doing this removes much of the random noise, but still preserves the trends in the data to prevent obscuring the analysis. The window size of the sliding window for the CPD algorithm is also taken to be 110 ft., and the linear penalty value for the algorithm is set at 0.05. **Fig. 43** shows the results of CPD shown on the raw non-smoothed input signal for one well for the horizontal section.

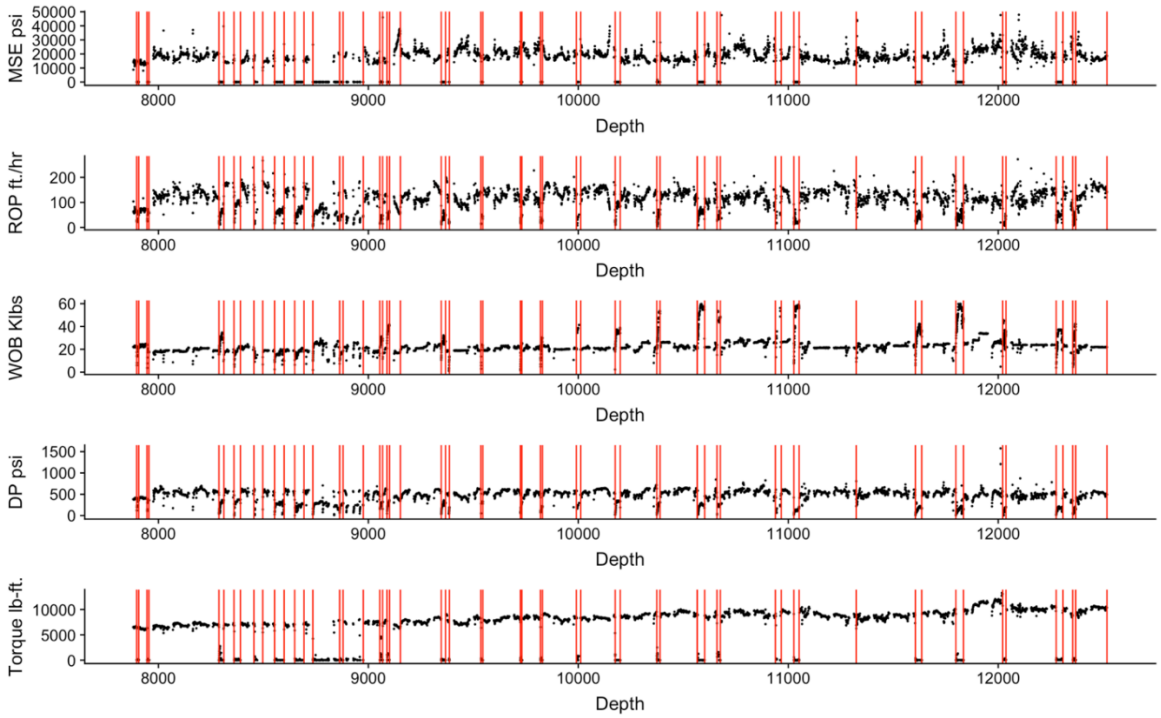

Fig. 43 - Plots showing the raw surface drilling data used for CPD with the change points shown as red vertical lines. The change points detected align with many changes in the multivariate signal based on a change in mean within the signal.

The CPD algorithm is able to detect changes within the multichannel drilling signal. While visual inspection of the data confirms these change points, it would be challenging to accomplish this manually in real-time or near real-time. I then compare the occurrence of change points with the HMM-derived states to provide insight into the relative merits of each method. The results are shown in **Fig. 44** with the drilling signal colored by the HMM-derived states.

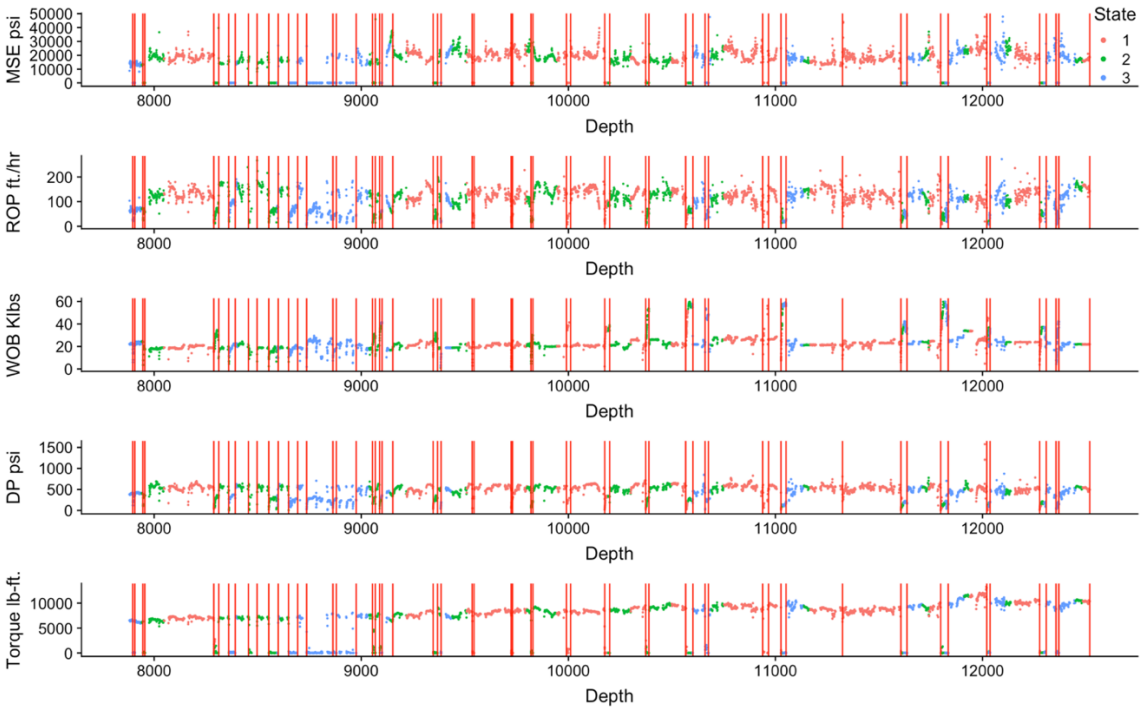

Fig. 44 - A plot showing the raw drilling data colored by HMM state, with the change points shown in red vertical lines. The change points are able to pick the transitions between drilling states very well, such as in the rapid fluctuations at 8500 ft. and the transition from state 2 to 3 at 11800 ft. This provides some assurance that the HMM states and change points are detecting the same type of shift within drilling modes or formation properties.

Overall, there appears to be excellent agreement between the location of the change points and transitions from one HMM-derived state to another. There are, however, a few locations where no change points are detected, and this is a consequence of the smoothing described earlier. My work in this section relies on smoothing over a 110 ft. window to ensure that the drilling states/change points are being applied at the same resolution as the regional geomodel. Nevertheless, the occurrence of change points and HMM state transitions show good correspondence and are responding to similar shifts in the drilling operations or formation properties. The HMM state probabilities for this well are shown below in **Fig. 45**.

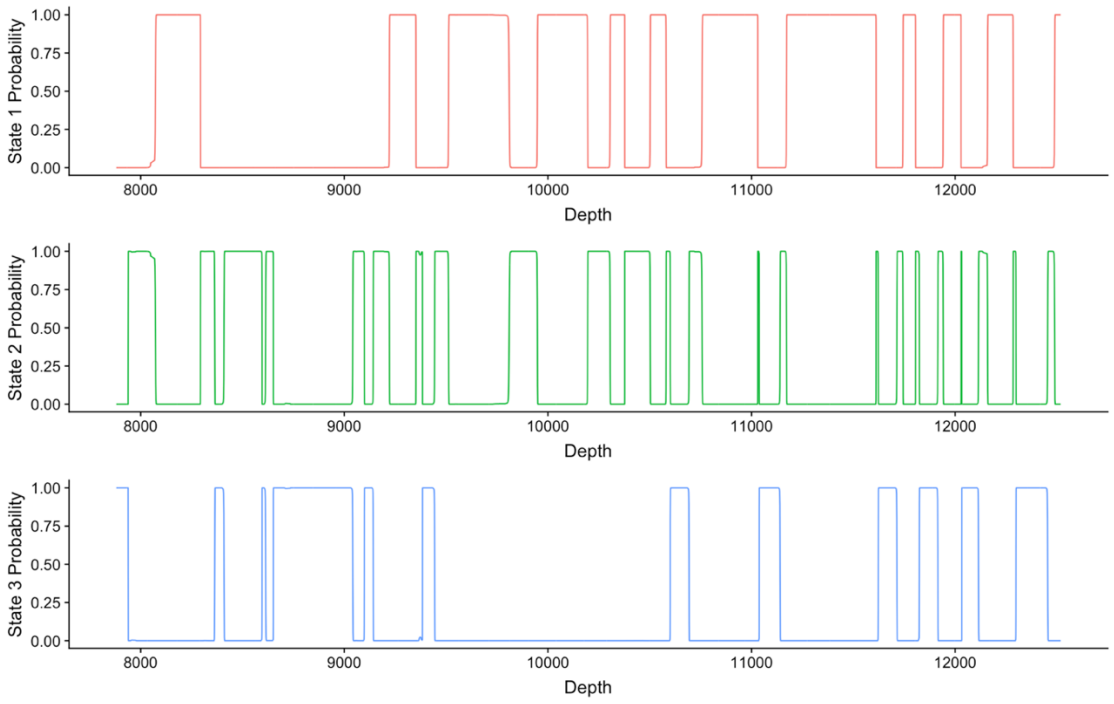

Fig. 45 - Plots showing the HMM state probabilities for each of the three states versus depth for well 1. The state with the highest probability is taken as the state at that prediction interval.

As shown previously in **Fig. 42,** the HMM-derived drilling states correspond to different gamma ray values which implies that the change points correspond to changes in formation properties expressed by gamma ray values.

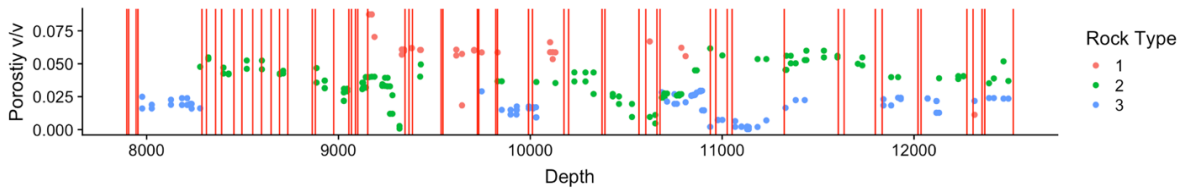

Fig. 46 - A plot of the porosity obtained from the 3D geomodel colored by the geostatistical rocktypes, with the change points from drilling data shown as red vertical lines. There is little to no correspondence of the change points with the rock types in this specific case.

In the next step, I compare the location of the change points to the transitions between rocktypes derived from the geomodel. A key point to note is that the geomodel is one select realization from a suite of many equiprobable models. Also, the rocktypes derived from the geomodel are associated with high uncertainty because of the lack of nearby well control. A high degree of correspondence between the methods shown in this work and the geomodel is unlikely. **Fig. 46** shows the porosity values obtained from the geomodel colored by rocktype with the change points shown.

As discussed earlier, the change points do not correspond to the transitions between the geomodel-derived rocktypes. Additionally, the rocktypes derived for this geomodel are obtained from an analysis of porosity-permeability core data. Because the surface expression of bit performance is likely related to the mechanical properties of the rock, the surface drilling data may not correspond to the petrophysical rocktypes. Ongoing work to re-define the rocktypes based on mineralogy and shear and compressional velocities will likely ensure better agreement. However, because the change point algorithm and the HMM model agree satisfactorily, the basis of correspondence should be the gamma ray response along the lateral as shown in **Fig. 47**.

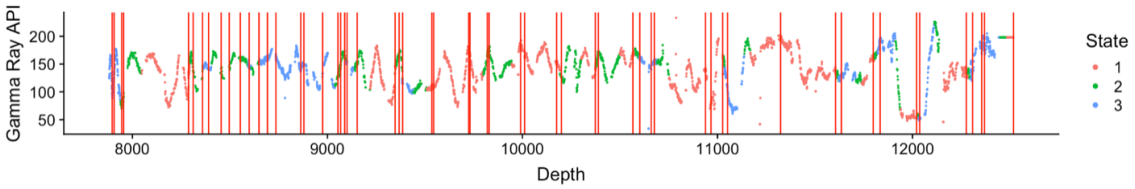

Fig. 47 - A plot showing the gamma ray response along the lateral with the states and change points shown. While there are many fluctuations within the gamma ray response that are not detected by the HMM or change points, there are a number of times that the states/change points do signal the sudden rise or fall in gamma ray values.

The figure shows that there is satisfactory agreement between the change points and the gamma ray response of the formation. However, there are several locations where changes in

gamma ray values do not correspond to change points or changes in HMM-derived states. The overall less-than-satisfactory response of the change point detection algorithm arises from several challenges, as described earlier.

4.3 Summary

In a more complex reservoir, with data from horizontal wells, I have shown two promising approaches to detect changes in multichannel surface drilling signals that correlate with formation properties measured by gamma ray.

These drilling states and change points did not correspond to changes in 3D geomodel-derived rocktypes. In this example, it is likely that the rocktypes derived from the geomodel may not be accurate because of a lack of nearby well control. Additionally, because the rocktypes are derived from porosity-permeability values, the drilling data may not be sensitive to changes in those petrophysical variables but might show better agreement with geomechanical facies. However, this is ongoing work at the time of this writing. The following are conclusions from the analysis presented above.

1. The Hidden Markov Model shows greater promise and better definition of the rocktypes inferred from drilling data. In real-time, the HMM can be used to predict the rocktype at the bit which is promising for geosteering applications. By relying only on bit performance data, I am able to provide early-time estimates of rocktype enabling a real possibility of looking ahead of the bit.
2. While the K-Means model is simple to use, in this specific case study, its performance relative to HMM was poor and the results lacked proper definition of the drilling-derived clusters.

3. Change point detection shows promise in detecting the changes within drilling states that have been shown to correlate with formation properties.
4. The change points and HMM-derived states do not correspond very well to geomodel-derived rocktypes. A re-definition of rocktypes based on mechanical properties is warranted.

Chapter 5: Conclusions and Recommendations

In this study I present two different approaches (a Hidden Markov Model and time-series change point detection) to determine rocktypes/lithologies from surface drilling data. I present two field case studies to demonstrate application of these methods. These case studies are based on data from the Volve North Sea Field and the STACK (Meramec formation). The following conclusions can be drawn from this work:

1. A simple K-Means based clustering approach is likely not feasible to detect rocktypes from surface drilling data. The non-linear relationships between the variables are not captured adequately by K-Means alone.
2. In the Volve North Sea Field, the Hidden Markov Model detects drilling states from surface drilling data that correspond to rocktypes derived from well logs. The change point detection algorithm locates transitions between the different rocktypes. Both methods promising approaches to apply to real-time drilling data to detect lithology at the bit.
3. In the unconventional Meramec play in Oklahoma, both the HMM and change point detection algorithms correspond very well in detecting drilling states. These drilling states are shown to be distinct in terms of their gamma ray response.
4. The geomodel-derived rocktypes do not show a good correspondence with the HMM-derived states or change points. This is likely because the geomodel lacks well control in the area and the spatial distribution of the rocktypes may be associated with high uncertainty. Additionally, the rocktypes were defined using porosity-permeability data that may not be related to rock mechanical data in this play.
5. Both HMM and CPD are viable methods for identifying formation properties in horizontal wells for unconventional formations. In both datasets studied, the methods show promise

in terms of their ability to inform drilling decisions such as altering trajectory in a horizontal well to stay within the target zone, or to alter drilling rate to preserve bit life when detecting a harder formation. With this information, drilling teams can improve efficiency, reduce costs, identify better landing zones and well trajectories, and improve the economics of exploration projects.

Some changes detected by change points or HMM state transitions will be related to alterations made by the surface drilling crew. These changes could be interpreted as changes in lithology, however when applied at the drilling rig in real-time these changes can be disregarded from the analysis. Overall, with knowledge of the drillers decisions made on the rig floor, identification of change points and state transitions can be used to guide drilling decisions to optimize the drilling process.

5.1 Recommendations

Based on the results in the lateral section of the wells in the unconventional Meramec play, I recommend that operators obtain compressional and shear wave transit times as well as bulk density logs so that mechanical properties can be calculated along the horizontal section. It is likely that the mechanical properties control bit performance that can then be used to infer different geomechanical facies from surface drilling data in real-time. This method can also be used to predict the occurrence of zones of increased hardness that can rapidly dull the drill bit. Identifying formations of stronger mechanical properties ahead of the bit can let drillers change drilling mode to preserve bit life and reduce the number of bit trips. In addition to preserving bit life, identification of mechanical properties can aid in hydraulic fracture placement in rocks with more favorable properties for fracturing. Reischman et al. (2017) show that fracture placement in areas

with more favorable mechanical properties along the lateral measured by dipole sonic result in fractures with greater contribution to production. Therefore, tying surface drilling derived states back to mechanical properties by comparing to dipole sonic data can guide operators to select optimal fracturing locations along the lateral.

The methods I have shown in this work can also be used to improve pore pressure predictions from seismic data. Seismic data is often used to derive pre-drill estimates of pore pressure to identify areas of overpressure or under pressure to mitigate kicks or fluid loss. However, this is made difficult because seismic velocities are influenced by changes in lithology, fluid content, and pore pressure that can cause changes in polarity at bed-boundaries (Sayers and Woodward 2001). Therefore, if the drilling team can identify changes in lithology ahead of the bit using HMM, variations in seismic velocity can be attributed to changes in pore pressure with more confidence. Experienced drillers may be accustomed to recognizing these types of changes from surface drilling data alone. However, an automated system creating alerts for changes can help prevent accidents and guide decision making to optimize drilling in harder formations, or adjusting mud weights when detecting pore pressure changes.

Application of this methodology requires real-time streaming capabilities. Although this was not tested in this thesis, I show a promising approach for operational drilling-derived lithology identification. Generalization of this approach to streaming artificial lift data can also help proactively monitor changes in behavior leading to failure and in other real-time anomaly detection functions.

References

Adak, S. 1993. Time-Dependent Spectral Analysis of Nonstationary Time Series. *Journal of the American Statistical Association* **93** (444): 1488—1501. http://dx.doi.org/10.2307/2670062.

Adams, R.P. and MacKay, D.J.C. 2007. Bayesian Online Changepoint Detection. arXiv:0710.3742v1.

Aminikhanghahi, S. and Cook, D.J. 2017. Selective Review of Offline Change Point Detection Methods. *Knowledge and Information Systems.* **51** (2):339-367. http://dx.doi.org/10.1007/s10115-016-0987-z.

Bai, J. 1999. Likelihood Ratio Tests for Multiple Structural Changes. *Journal of Econometrics* **91** (2): 299-323. http://dx.doi.org/10.1016/S0304-4076(98)00079-7.

Bangert, P. 2019. Diagnosing and Predicting Problems with Rod Pumps Using Machine Learning. Presented at the SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, 18-21 March. http://dx.doi.org/10.2118/194993-MS.

Basseville, M., and Nikiforov, I. 1993. *Detection of Abrupt Changes: Theory and Application*, Volume 104. Prentice Hall Englewood Cliffs.

Baum, L.E. 1972. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities III: Proceedings of the 3rd Symposium on Inequalities*, 1–8. Academic Press.

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Springer+Business Media LLC. ISBN-10: 0-387-31073-8.

Bourgoyne Jr., A.T., Chenevert, M.E., Milheim, K.K. et al. 1986. Applied Drilling Engineering. Society of Petroleum Engineers, Richardson.

Chen, W., Shen, Y., and Zhang, Z. 2019. Understand Drilling System Energy Beyond MSE. Presented at the SPE Annual Technical Conference and Exhibition, Calgary, Canada. http://dx.doi.org/10.2118/196050-MS.

Drummond, K.A. 2018, Regional Stratigraphy and Proximal to Distal Variation of Lithology and Porosity Within a Mixed Carbonate-Siliciclastic System, Meramec and Osage Series (Mississippian), Central Oklahoma. M.S. Thesis, University of Oklahoma, Norman, Oklahoma, 145 p.

Dutton, S.P. 1985. Fan-Delta Granite Wash of the Texas Panhandle: Oklahoma City Geological Society Short Course, p. 1-44.

Esmersoy, C., Ramirez, A., Hannan, A. et al. 2013. Guiding Drilling by Look Ahead Using Seismic and LWD Data. Presented at the North Africa Technical Conference & Exhibition, Cairo, Egypt, 15—17 April. SPE 164786. http://dx.doi.org/10.2118/164786.

Equinor. "Volve Data Village." *Equinor Data Portal Beta*. October 22, 2018. https://data.equinor.com/dataset/Volve.

Fisher, R.A. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Human Genetics* **7**(2): 179—188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Gagniuc, P. A. 2017. *Markov Chains: From Theory to Implementation and Experimentation*. USA, NJ: John Wiley & Sons. pp. 1–256. ISBN 978-1-119-38755-8.

Gupta, I., Devegowda, D., Jayaram, V. et al. 2019. Machine Learning Regressors and Their Metrics to Predict Synthetic Sonic and Brittle Zones. Presented at the Unconventional Resources Technology Conference, Denver, Colorado, 22—24 July. http://dx.doi.org/10.15530/urtec-2019-0148.

Gupta, I., Rai, C.S., and Sondergeld, C.H. 2017. Rock Typing in Eagle Ford, Barnett, and Woodford Formations. Presented at the Unconventional Resources Technology Conference, Austin, Texas, 24-26 July. http://dx.doi.org/10.15530/URTEC-2017-2669624.

Haq, B.U. and Schutter, S.R. 2008. A Chronology of Paleozoic Sea-Level Changes. *Science* **322** (5898): 64-68. http://dx.doi.org/DOI: 10.1126/science.1161648.

Hawkins, D.M. 2001. Fitting Multiple Change-Point Models to Data. *Computational Statistics & Data Analysis* **37**: 323-341. https://doi.org/10.1016/S0167-9473(00)00068-2.

Inoue, T., Tanaka, R., and Ishiwata, J. 2019. Attempt of Lithology Prediction from Surface Drilling Data and Machine Learning for Scientific Drilling Programs. Presented at the SPE Europec featured at 81st EAGE Conference and Exhibition, London, United Kingdom, 3—6 June. http://dx.doi.org/10.2118/195444-MS.

Kass, R.E. and Voss, P.W. 1997. *Geometrical Foundations of Asymptotic Inference*. New York, New York: John Wiley & Sons, Inc.

Killick, R., Fearnhead, P., and Eckley, I.A. 2012. Optimal Detection of Changepoints with a Linear Computational Cost. http://dx.doi.org/arXiv:1101/143v3.

Killick, R. 2017. www.nag.co.uk/market/events. *Financial Time Series: Changepoints, Structural Breaks, Segmentations and Other Stories*, (accessed 13 November 2019).

Kim, H. and Siegmund, D. 1989. The Likelihood Ratio Test for a Change-Point in Simple Linear Regression. *Biometrika* **76** (3): 409-423. http://dx.doi.org/10.2307/2336108.

Kullawan, K., Bratvold, R.B., and Nieto, C.M. 2017. Decision-Oriented Geosteering and the Value of Look-Ahead Information: A Case-Based Study. *SPE Journal*. https://doi.org/10.2118/184392-PA.

Lesso, B., Ignova, M., Zeineddine, F. et al. 2011. Testing the Combination of High Frequency Surface and Downhole Drilling Mechanics and Dynamics Data Under a Variety of Drilling Conditions. Presented at the SPE/IADC Drilling Conference and Exhibition, Amsterdam, Netherlands, 1-3 March. http://dx.doi.org/10.2118/140347-MS.

Li, Y. 2019. Prediction of Penetration Rate Ahead of the Bit through Real-Time Updated Machine Learning Models. *SPE Journal*. https://doi.org/10.2118/194105-MS.

Loermans, T., Kimour, F., Bradford, C. et al. 2012. Successful Pilot Testing of Integrated Advanced Mud Logging Unit. Presented at the SPWLA 53rd Annual Logging Symposium, Cartagena, Colombia, 16-20 June. SPWLA-2012-184.

MacQueen, J. B. 1967. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press. pp. 281–297.

Melberg, E.V. 2013. https://www.npd.no/en/facts/news/Exploration-drilling-results/2013/159-F-11-and-159-F-11-A (accessed 01 November 2019).

Miller, J. C. 2018. Regional Stratigraphy and Organic Richness of the Mississippian Meramec and Associated Strata, Anadarko Basin, Central Oklahoma. M.S. Thesis, University of Oklahoma, Norman, Oklahoma, 154 p.

Moore, J.C., and Klaus. A. 1998. Explanatory Notes. *Proc. ODP, Init. Repts.,* 171A: College Station, TX (Ocean Drilling Program), 11–15. doi:10.2973/odp.proc.ir.171a.102.

Mottahedeh, R. 2008. Horizontal Well Geosteering: Planning, Monitoring and Geosteering. *Journal of Canadian Petroleum Technology*: 28-32. http://dx.doi.org/10.2118/08-11-28-CS.

Oehm, D. gradientdescending.com. *Hidden Markov Model Example in R with the DepmixS4 Package*, http://gradientdescending.com/hidden-markov-model-example-in-r-with-the-depmixs4-package (accessed 15 November 2019).

Osypov, K., Nichols, D., Woodward, M. et al. 2011. From Quantifying Seismic Uncertainty to Assessing E&P Risks and the Value of Information. 81st SEG Annual International Meeting. San Antonio, Texas, USA. SEG-2011-3683. https://doi.org/10.1190/1.3627966.

Ounsakul, T., Sirirattanachatchawa, T., Pattarachupong, W. et al. 2019. Artificial Lift Selection Using Machine Learning. Presented at the International Petroleum Technology Conference, Beijing, China, 26—28 March. http://dx.doi.org/10.2523/IPTC-19423-MS.

Page, E.S. 1954. Continuous Inspection Schemes. *Biometrika* **41** (1):100-115. https://doi.org/10.1093/biomet/41.1-2.100.

Page, E.S. 1955. A Test for a Change in a Parameter Occurring at an Unknown Point. *Biometrika* **42** (3): 523-527. https://doi.org/10.1093/biomet/42.3-4.523.

Pessier, R.C. and Fear, M.J. 1992. Quantifying Common Drilling Problems with Mechanical Specific Energy and a Bit-Specific Coefficient of Sliding Friction. Presented at the SPE Annual Technical Conference and Exhibition, Washington, District of Columbia, 4-7 October. http://dx.doi.org/10.2118/24584-MS.

Petrushin, V.A. 2000. Hidden Markov Models: Fundamentals and Applications. *Online Symposium for Electronics Engineer*.

Rabiner, L.R. and Juang, B.H. 1986. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine* **3** (1): 4-16.

Reischman, R., Velez, E., Green, A. et al. 2017. Horizontal Cased Hole Evaluation Using a New Pulsed Neutron Spectroscopy Tool and Dipole Sonic. Presented at the Unconventional Resources Technology Conference, Austin, Texas, 24-26 July. http://dx.doi.org/10.15530/URTEC-2017-2689778.

Rensburg, N.J.V. 2019. AI4ESP - Autonomous Well Surveillance for ESP Pumps Using Artificial Intelligence. Presented at the SPE Oil and Gas India Conference and Exhibition, Mumbai, India. http://dx.doi.org/10.2118/194587-MS.

Roberts, S.W. 1966. A Comparison of Some Control Chart Procedures. *Technometrics* **8** (3): 411-430. http://dx.doi.org/10.2307/1266688.

Romanenkova, E., Zaytsev, A., Gruzdev, A. et al. 2019. Real-Time Data-Driven Detection of the Rocktype Alteration During a Directional Drilling. *IEEE Geoscience and Remote Sensing*. http://dx.doi.org/arXiv:11436v1.

Sayers, C.M. and Woodward, M.J. 2001. Enhanced Seismic Pore-Pressure Prediction. Presented at the Offshore Technology Conference, Houston, Texas, 30 April-3 May. http://dx.doi.org/10.4043/13044-MS.

Schuetter, J., Mishra, S., Zhong, Z. et al. 2015. Data Analytics for Production Optimization in Unconventional Reservoirs. Presented at the Unconventional Resources Technology Conference, San Antonio, Texas, 20-22 July. http://dx.doi.org/10.2118/178653-MS/URTeC:2167005.

Sen, A. and Srivastava, M.S. 1975. On Tests for Detecting Change in Mean. *The Annals of Statistics* **3**(1): 98—108. http://dx.doi.org/10.1214/aos/1176343001.

Smith, A.F.M. 1975. A Bayesian Approach to Inference about a Change-Point in a Sequence of Random Variables. *Biometrika* **62** (2): 407—416. http://dx.doi.org/DOI:10.2307/2335381.

Solomon, S.T., Ross, K.C., Burton, R.C. et al. 1994. A Multidisciplined Approach to Designing Targets for Horizontal Wells. *Journal of Petroleum Technology*: 143-149. http://dx.doi.org/10.2118/25506-PA.

Tariq, Z., Elkatatny, S., Mahmoud, M. et al. 2016. A New Artificial Intelligence Based Empirical Correlation to Predict Sonic Travel Time. Presented at the International Petroleum Technology Conference, Bangkok, Thailand, 14-16 November. http://dx.doi.org/10.2523/IPTC-19005-MS.

Teale, R. 1965. The Concept of Specific Energy in Rock Drilling. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts.* **2** (1): 57-73. https://doi.org/10.1016/0148-9062(65)90022-7.

Truong, C., Oudre, L., and Vayatis, N. 2018. Ruptures: Change Point Detection in Python. arXiv:1801.00826.

Truong, C., Oudre, L., and Vayatis, N. 2019. Selective Review of Offline Change Point Detection Methods. *Signal Processing*. http://dx.doi.org/arXiv:1801.00718v2.

Visser, I., Speekenbrink, M. 2010. DepmixS4: An R Package for Hidden Markov Models. J*ournal of Statistical Software*, 36(7), 1-21. URL http://www.jstatsoft.org/v36/i07/.

Whittaker, A.H. 1987. Mud Logging. In *Petroleum Engineering Handbook*. Richardson, Texas: Society of Petroleum Engineers.

Wilson, A. 2013. Wired-Drillpipe Field Trials Reveal Potential Benefits Over Traditional Pipe. *Journal of Petroleum Technology* **65** (6): 105-107. http://dx.doi.org/10.2118/0613-0105-JPT.

Wittman, B. R., 2013. Subsurface Stratigraphy and Characterization of Mississippian (Osagean to Meramecian) Carbonate Reservoirs of the Northern Anadarko Shelf, North-Central Oklahoma. Unpublished Master of Science Thesis, University of Arkansas.

Yu, H., Wang, Z., Rezaee, R. et al. 2016. The Gaussian Process Regression for TOC Estimation Using Wireline Logs in Shale Gas Reservoirs. Presented at the International Petroleum Technology Conference, Bangkok, Thailand, 14-16 November. http://dx.doi.org/10.2523/IPTC-18636-MS.

Appendix A: Additional Results for the Volve Oilfield

Results for two additional wells (15-4 and 15-5) from the Volve data set are presented in this section. The results of these two wells show the utility of the workflow for identifying rock properties and drilling efficiency by HMM states and change points from CPD.

Well 15-4:

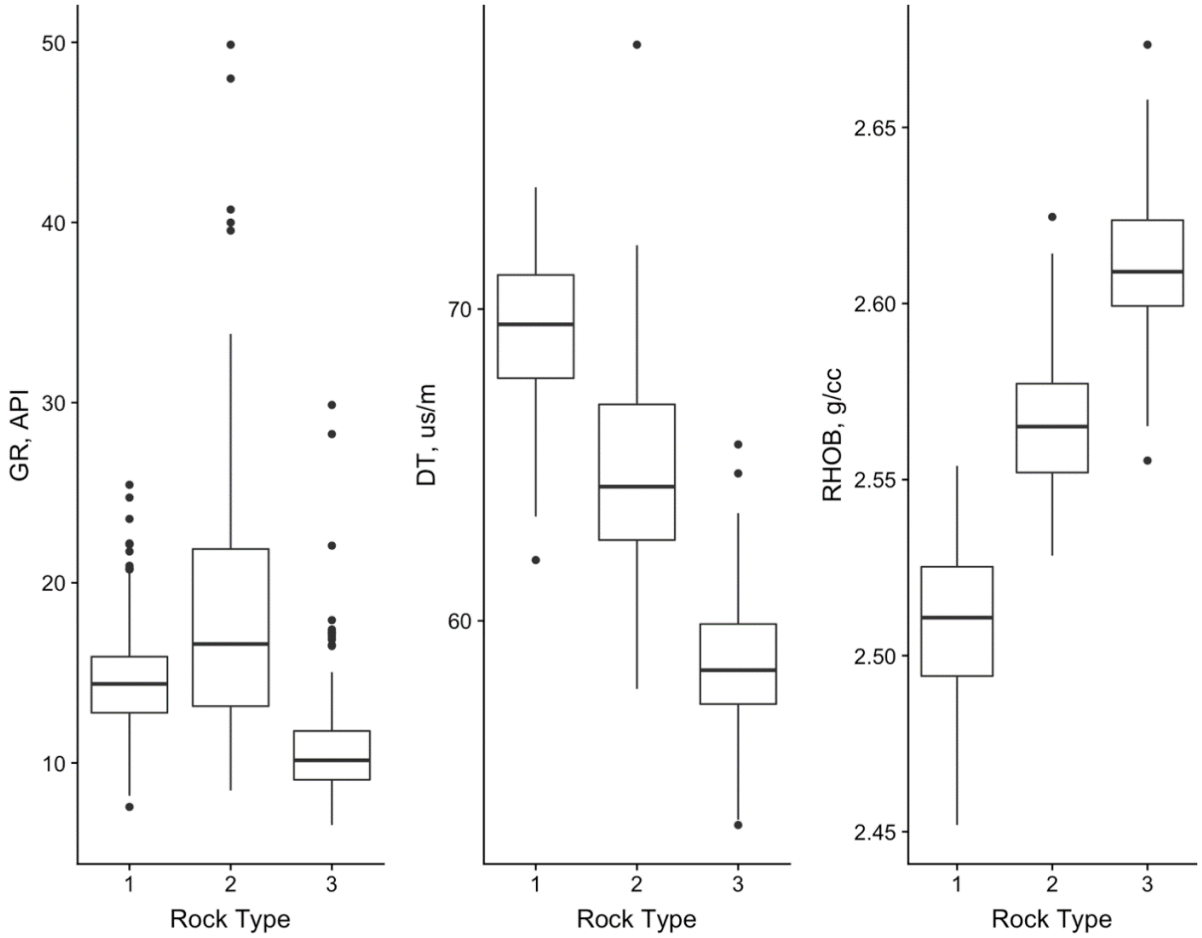

Fig. A1 – Boxplots showing the distribution of log data used for rocktyping from K-Means clustering for well 15-4. The rocktypes display distinct properties from one another that can be used to identify zones of varying reservoir quality. Rocktype 1 is likely a good quality rock as bulk density and gamma ray is low relating to high porosity and low clay.

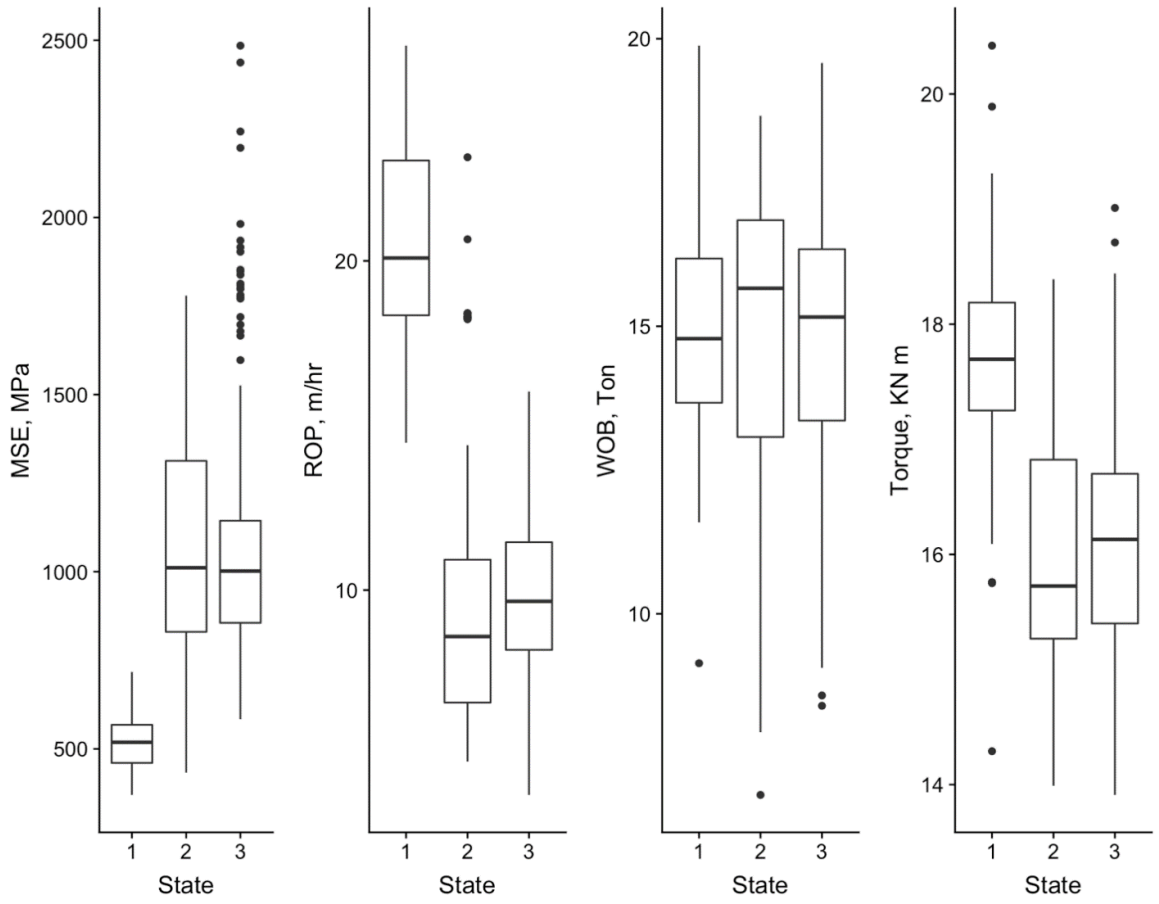

Fig. A2 – Boxplots showing the distribution of data for each HMM state derived from surface drilling data for well 15-4. The HMM states are able to define states with varying properties that can be monitored during drilling for identification of formation properties and drilling efficiency.

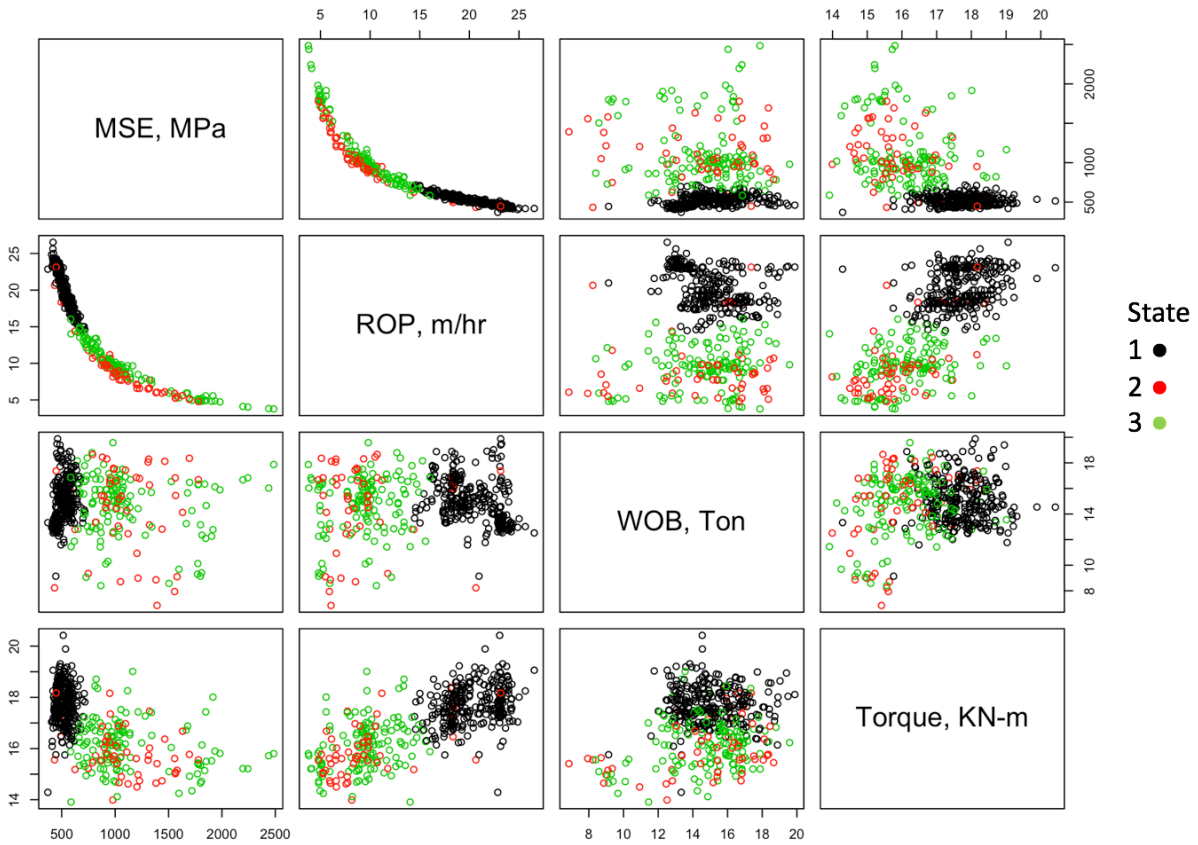

Fig. A3 – A scatterplot matrix of the surface drilling variables colored by HMM state. The HMM states are able to create good distinctions for states as shown in the MSE versus ROP plot. The states are able to capture different portions of the curves with little overlap.

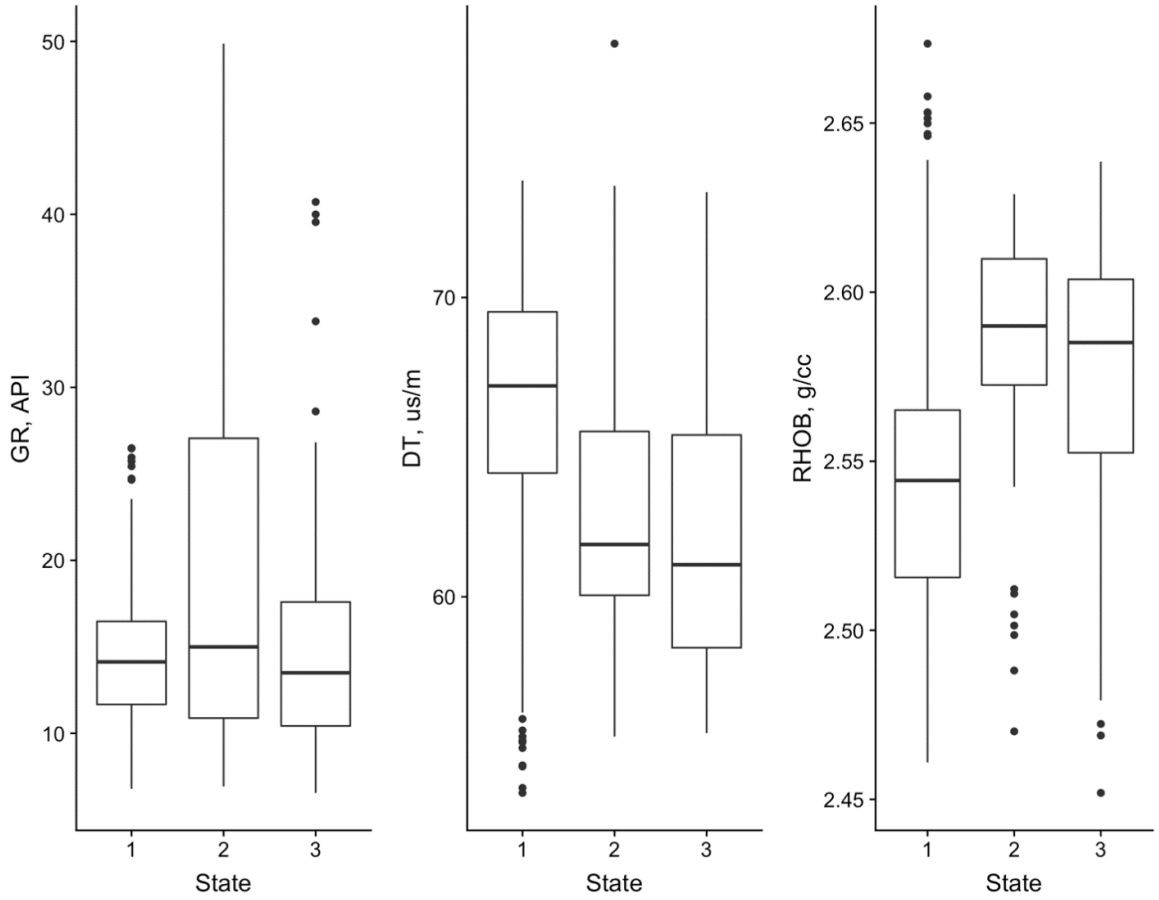

Fig. A4 – Boxplots showing the distribution of log data for each of the surface drilling derived HMM states for well 15-4. The states are able to create good distinction for RHOB and DT which shows how HMM states can be used to monitor formation properties.

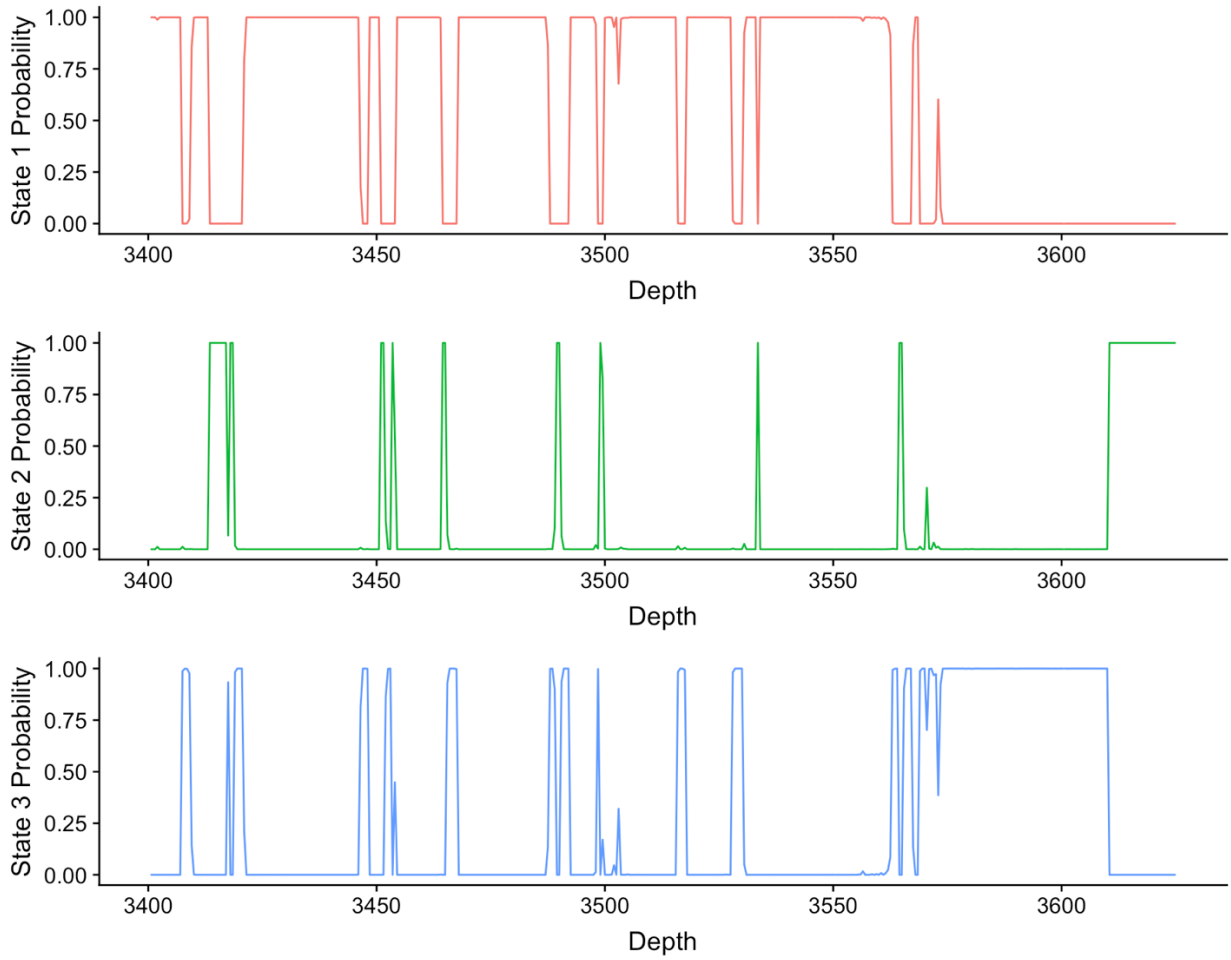

Fig. A5 – Plots showing the posterior probabilities of the HMM states from surface drilling data versus depth for well 15-4. The predicted state is taken to be the state with the greatest probability at each depth interval.

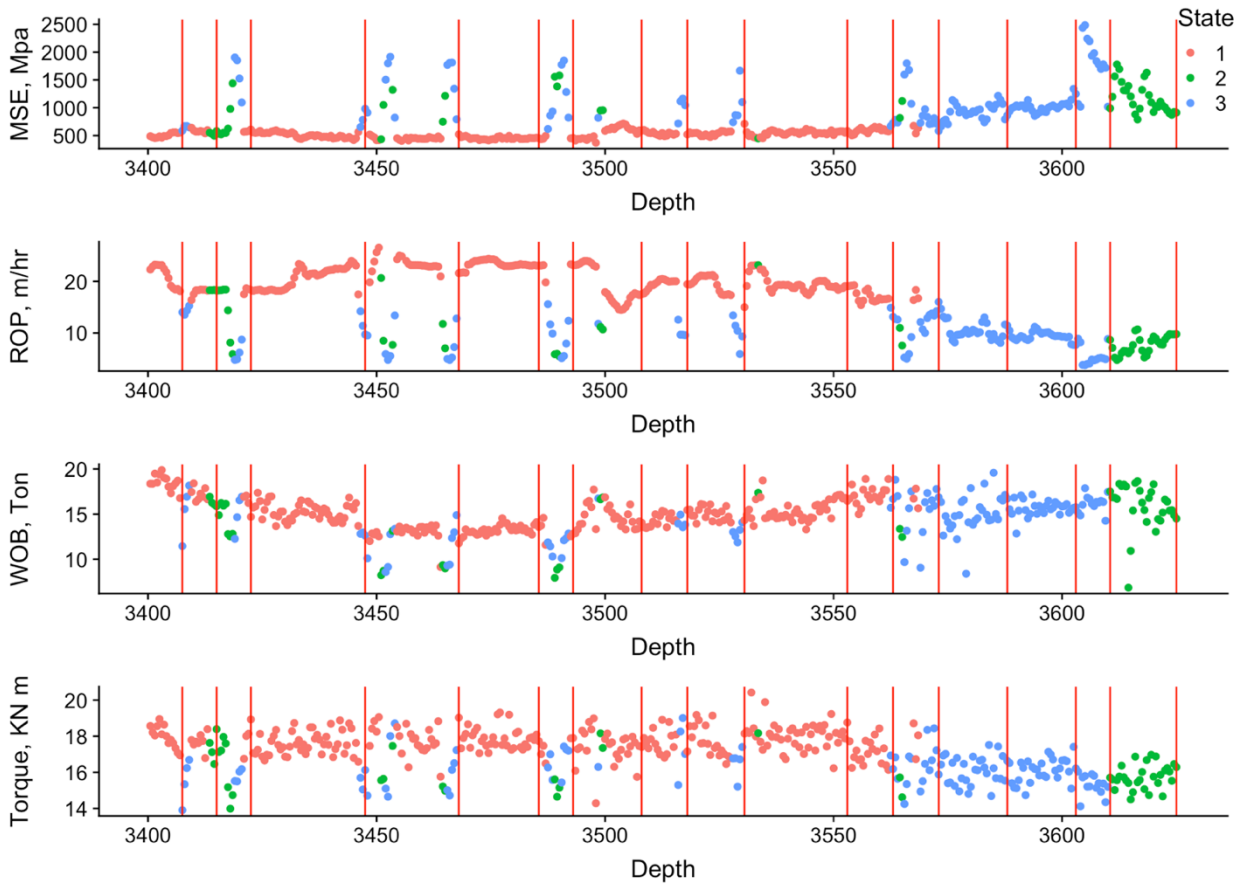

Fig. A6 – Plots showing the surface drilling data from well 15-4 colored by the HMM derived states with change points from CPD shown in red vertical lines. The change points are able to detect transitions in the HMM states with good accuracy.

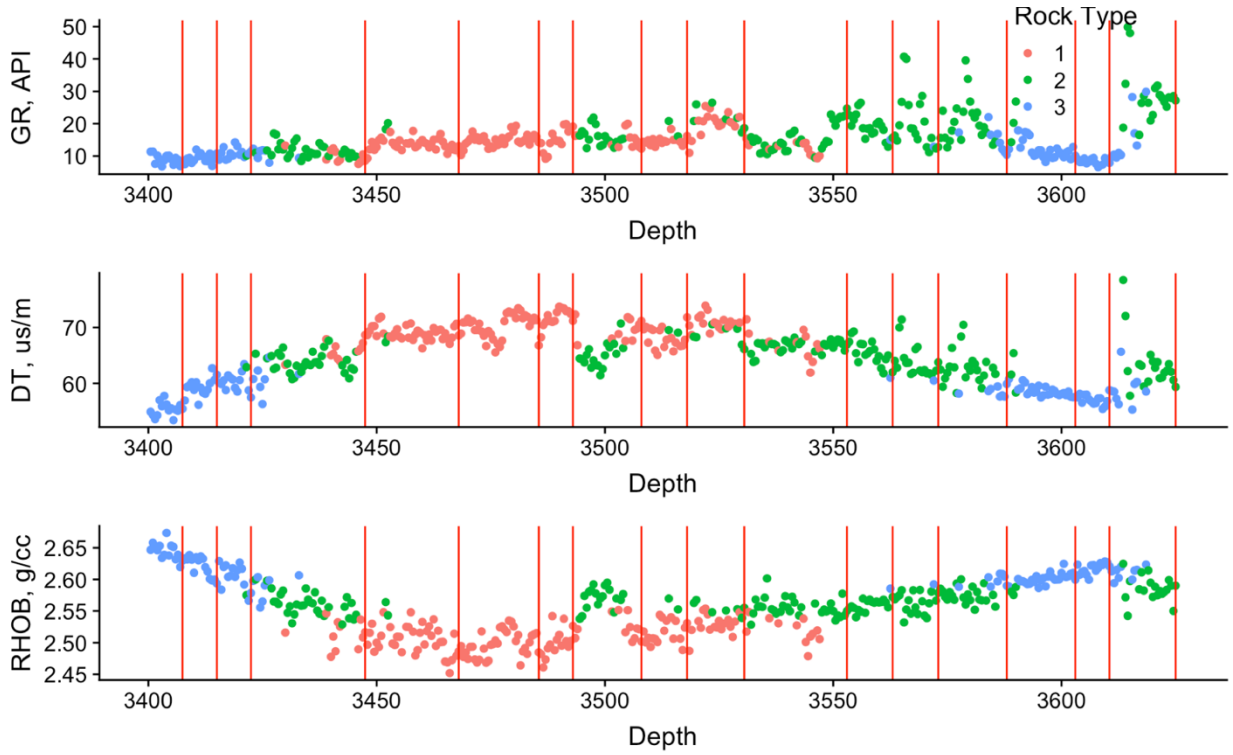

Fig. A7 – Plots showing the log data colored by rocktype with the change points shown in red vertical lines for well 15-4. In many areas the change points are able to accurately identify the transition between rocktypes from surface drilling data.

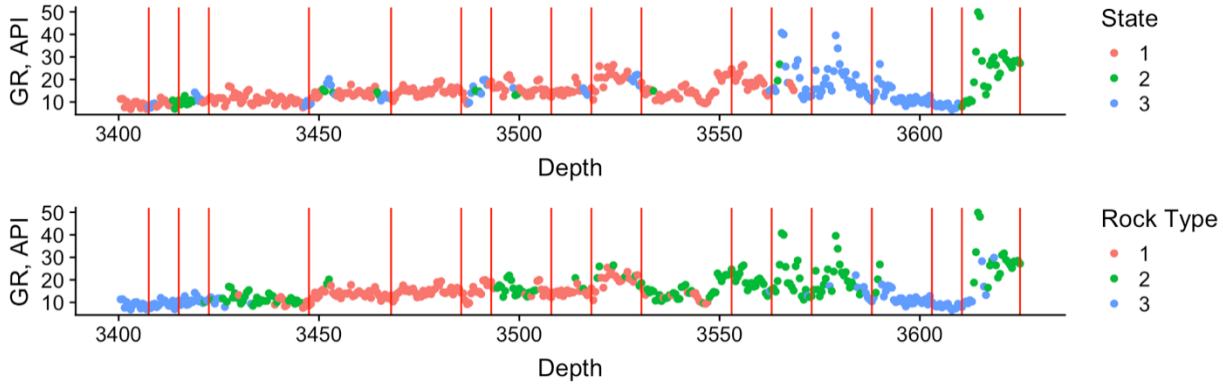

Fig. A8 – Plots showing the surface drilling HMM states, log derived rocktypes, and change points for well 15-4. The change points from surface drilling data are able to identify transitions in both HMM states and rocktypes showing that CPD can be used to monitor transitions in formation properties and drilling efficiency.

Well 15-5:

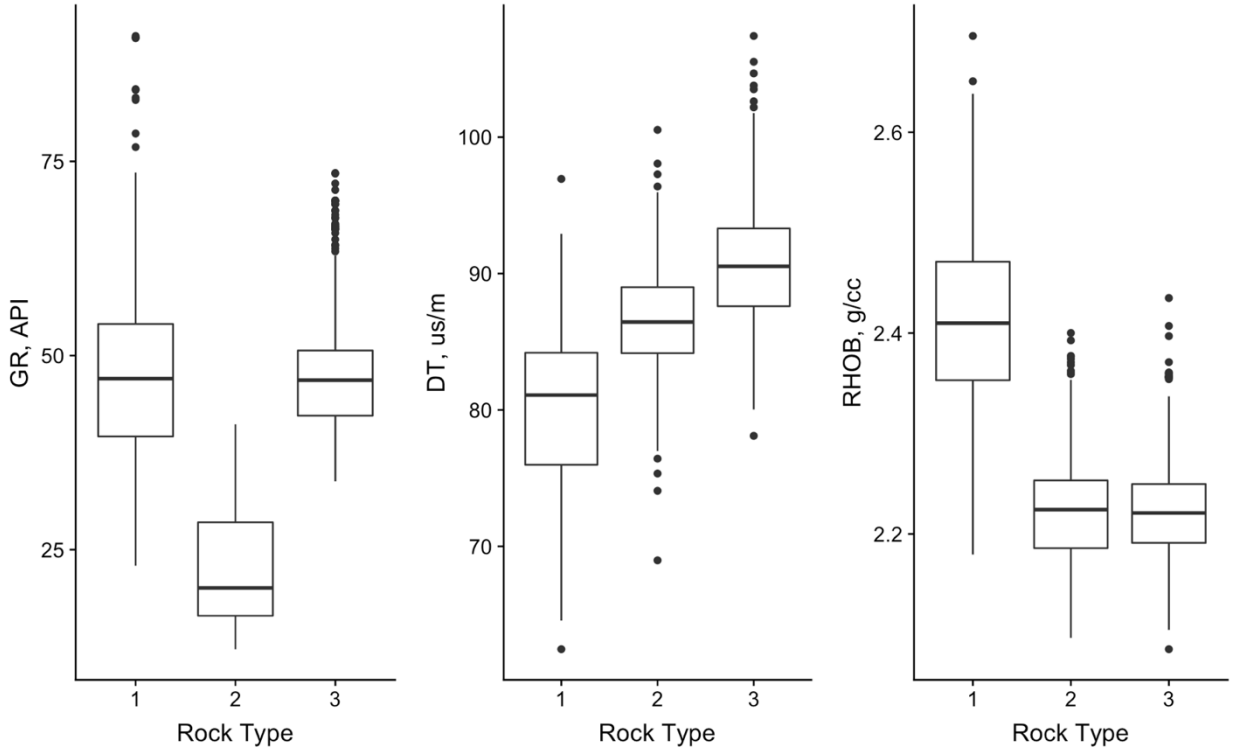

Fig. A9 – Boxplots showing the distribution of log data used for K-Means derived rocktypes for well 15-5. The rocktyping is able to derive rocktypes with distinct formation properties that can be used to identify reservoir quality.

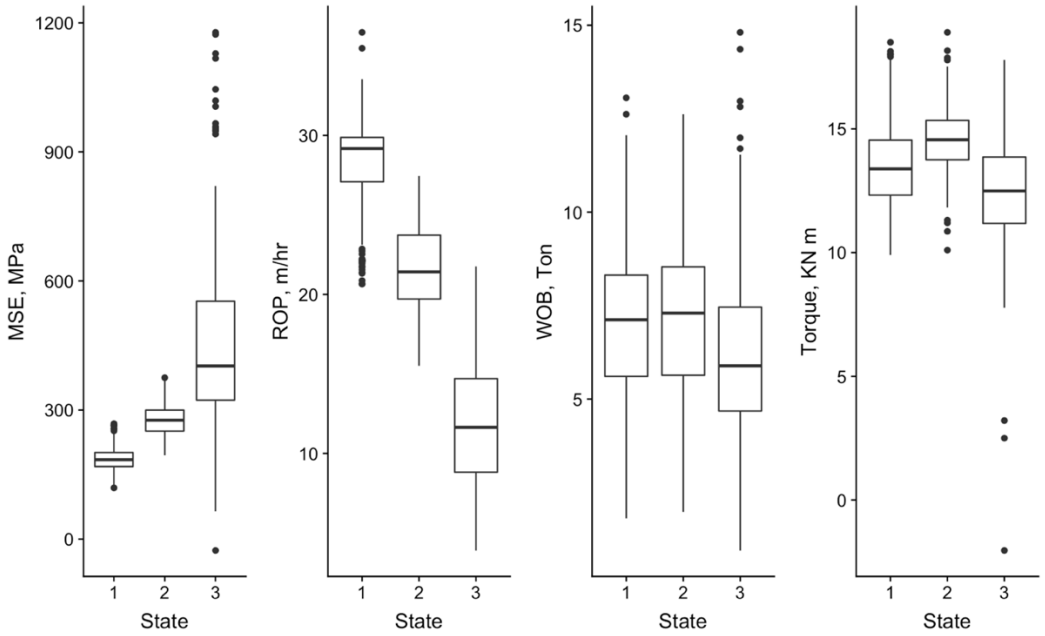

Fig. A10 – Boxplots showing the distribution of surface drilling data for the HMM derived states from well 15-5.

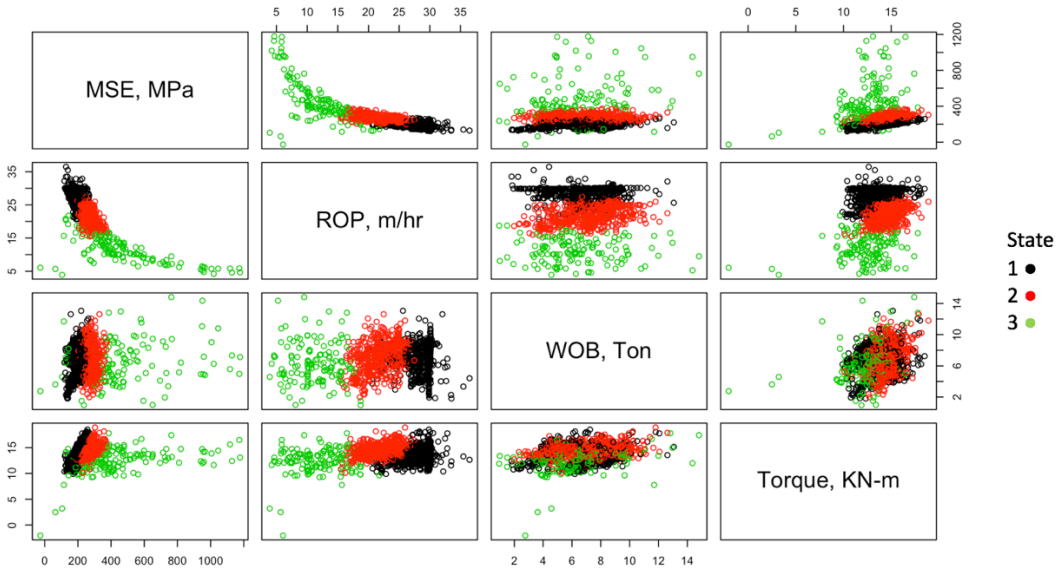

Fig. A11 – A scatterplot matrix showing the surface drilling variables colored by HMM state for well 15-5. The HMM-derived states are able to create distinct groups of surface drilling data. In the plot of MSE versus ROP the HMM states define different regions of the curve.

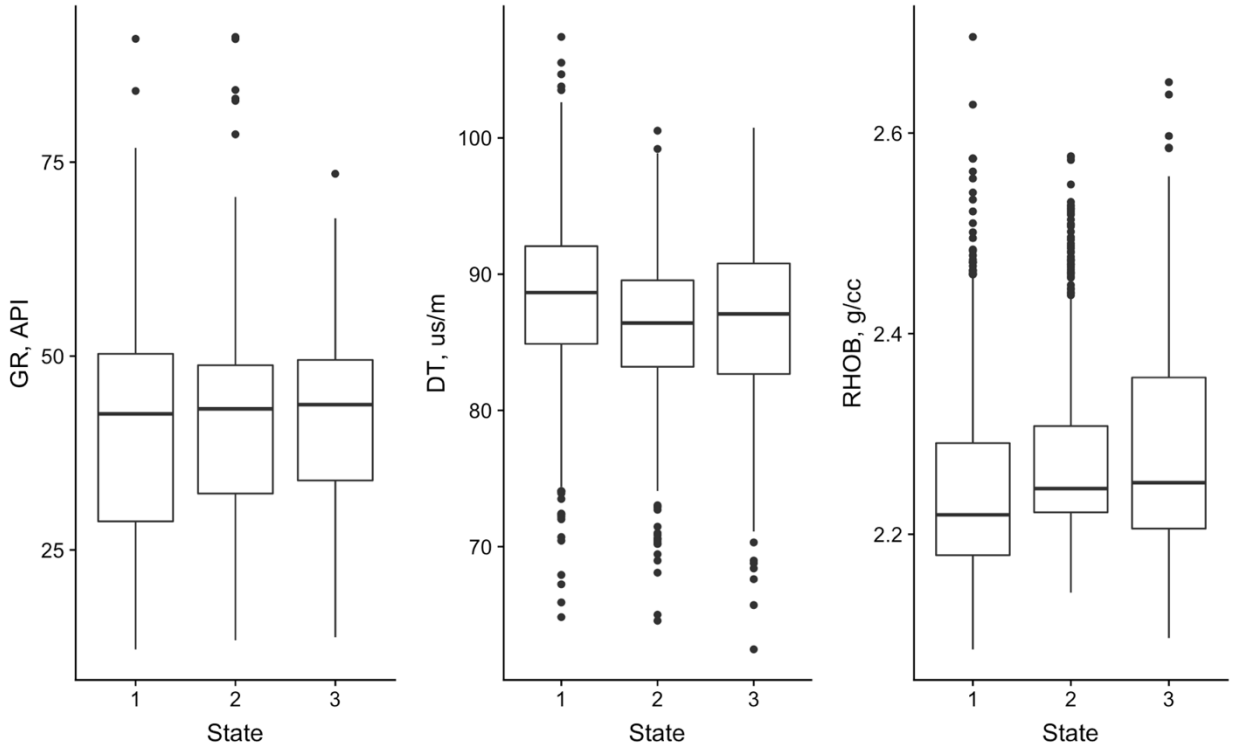

Fig. A12 – Boxplots showing the distribution of log data for the surface drilling derived HMM states for well 15-5. The HMM states are able to create groups with distinct formation properties, particularly in RHOB and DT. These states can be monitored for formation properties in real-time from surface drilling data.

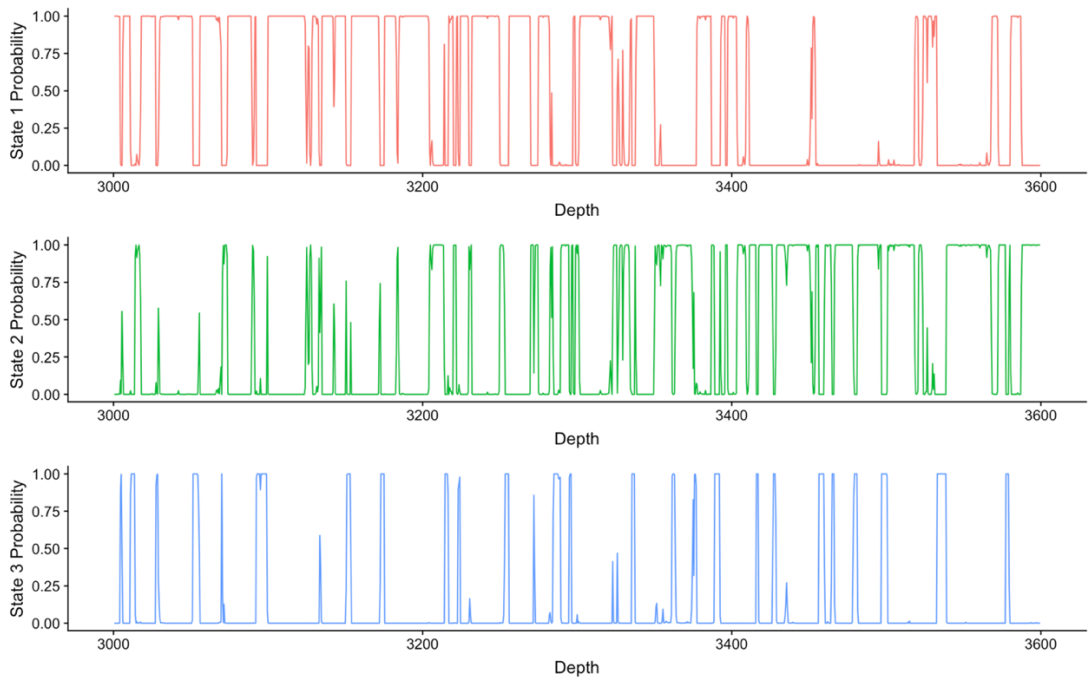

Fig. A13 – A plot of the HMM state probabilities from surface drilling derived states for well 15-5. The rapid transitions show that the HMM had more trouble classifying states into distinct groups than in the other examples. The predicted state is the state with the maximum probability at each depth interval.

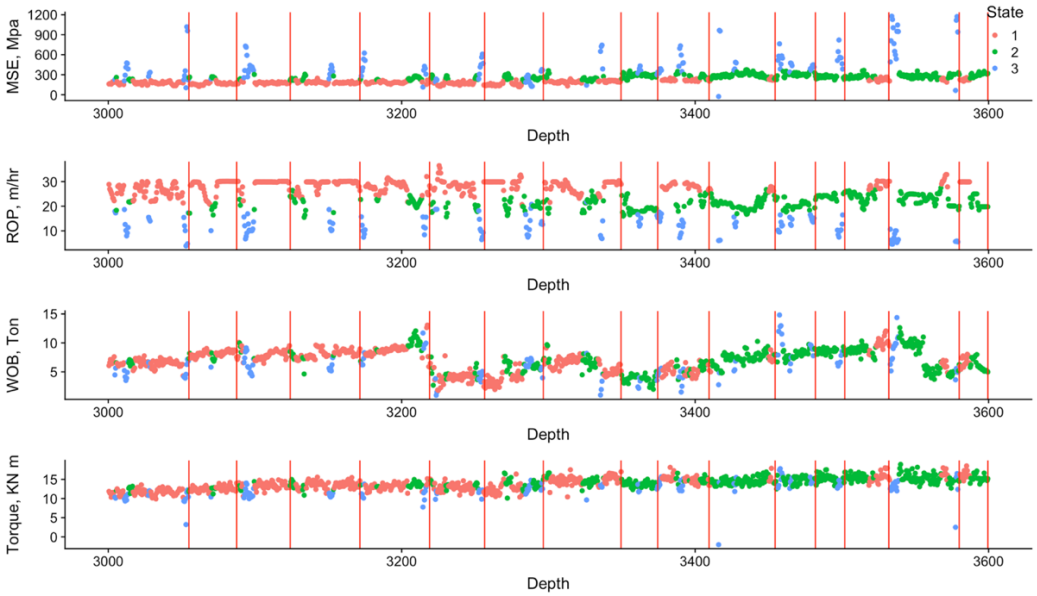

Fig. A14 – Plots showing the surface drilling data colored by HMM state with change points shown in red vertical lines for well 15-5. The change points are able to detect transitions in the HMM states. A few transitions are missed however.

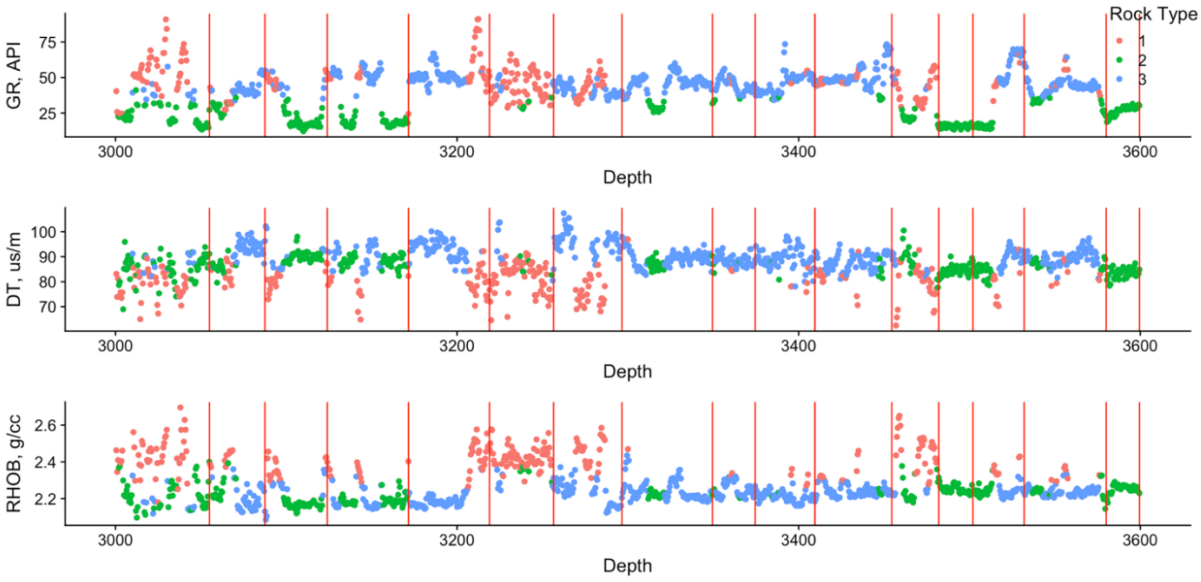

Fig. A15 – Plots showing the log data colored by K-Means derived rocktype with change points from surface drilling data shown as red vertical lines for well 15-5. The change points are able to identify the transition from one rocktype to another from surface drilling data.

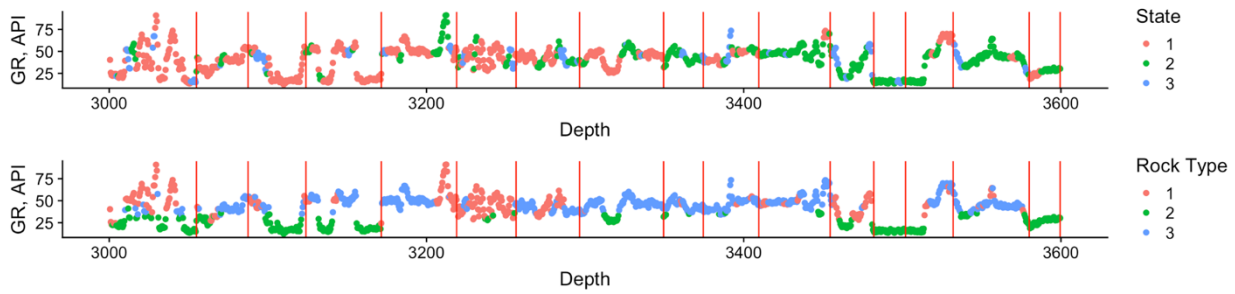

Fig. A16 – Plots showing gamma ray colored by HMM states and K-Means rocktypes for well 15-5. The change points are able to identify transitions within states and rocktypes with fairly good accuracy. This shows that HMM states and change points are able to identify transitions in formation properties and drilling efficiency that can be monitored from surface drilling data.

The results from the two wells presented support the conclusions made in the main body of the thesis. Both Hidden Markov Models and Change Point Detection methods applied to multichannel surface drilling data show promise in terms of their ability to provide information about changes in formation properties. Therefore, both methods can be applied for near real-time lithology prediction for geosteering purposes.

Appendix B: Additional Results for the Oklahoma Unconventional Reservoir

This section provides additional results related to the application of Change Point Detection (CPD) to surface drilling data. The surface drilling data is smoothed using a 110 ft. moving average as in Chapter 4. Here, I will again compare the occurrence of change points with transitions between HMM states and geomodel-derived rocktypes.

Well 2:

The input parameters of the sliding window CPD algorithm are window size = 110 ft and a linear Penalty = 0.5.

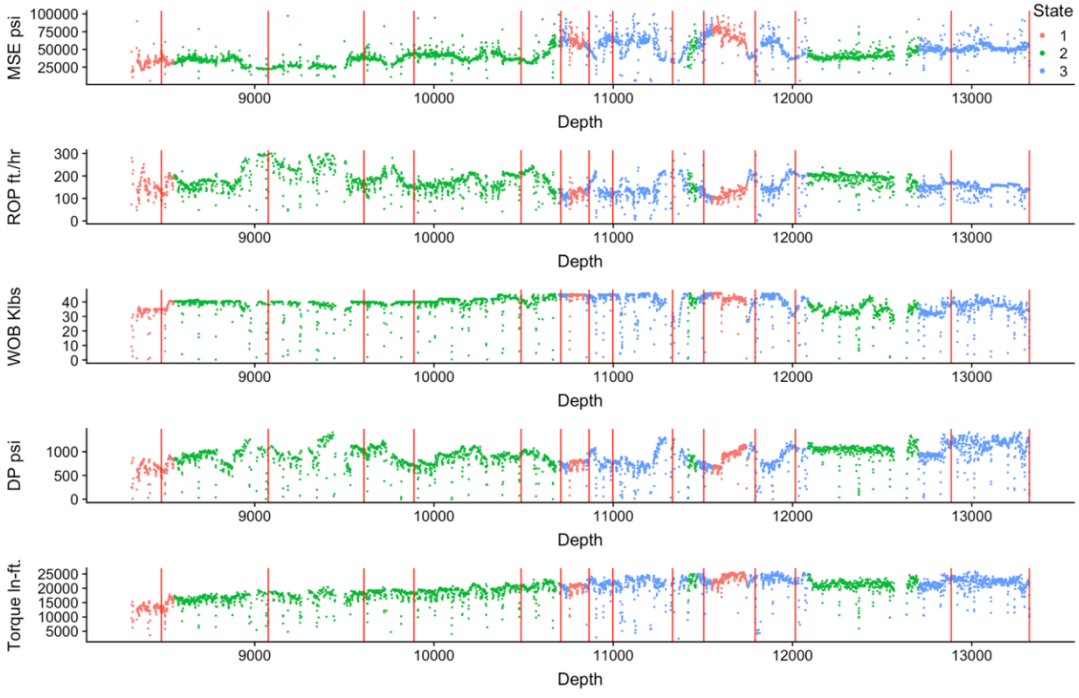

Fig. B1 – Surface drilling data for Well 2 colored by HMM state with change points shown as red vertical lines. Change points are aligned with some transitions in HMM state transitions. The CPD algorithm appears to be overly sensitive with the parameters chosen in this case.

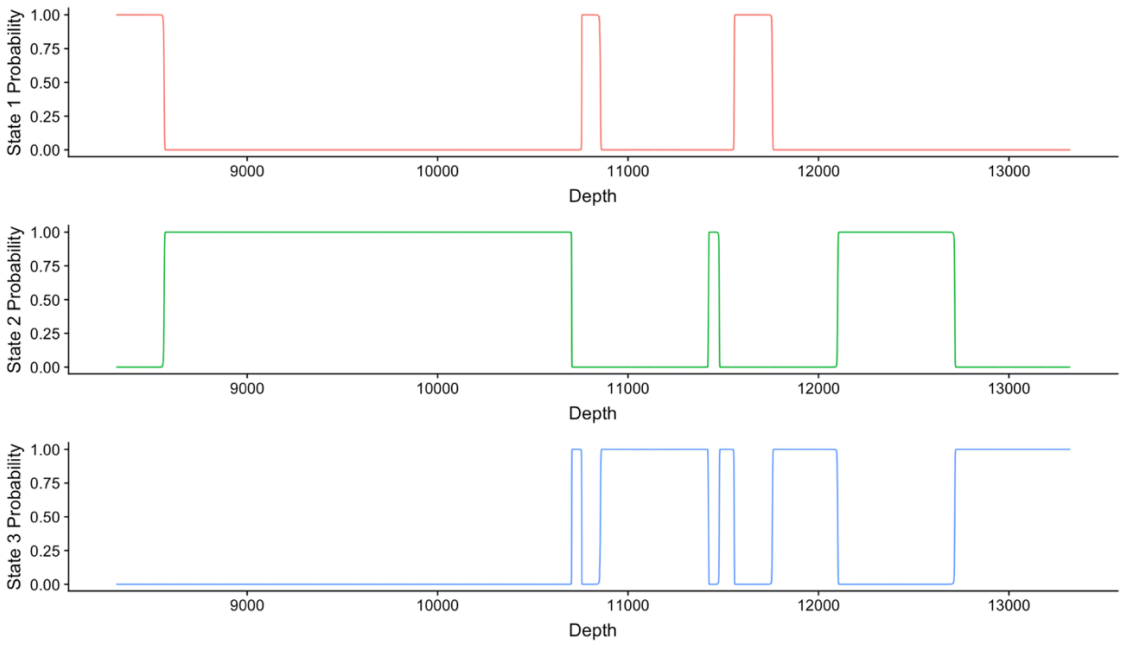

Fig. B2 – Plots showing the HMM state probabilities for each of the prediction intervals. For most of the series, the HMM predicts state 2 to be the highest probability. There are fewer extreme variations in the drilling data in this example resulting in fewer state transitions and change points.

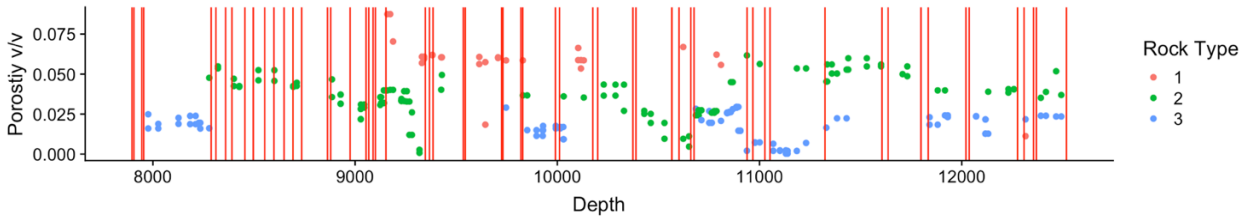

Fig. B3 – A plot showing the geomodel derived porosity colored by rocktype with the change points shown as red vertical lines. Similar to Well 1 in Chapter 4, the change points do not align well with the rocktypes. This is only being provided for informational purposes. The geomodel derived rocktypes are undergoing refinement at the time of this writing.

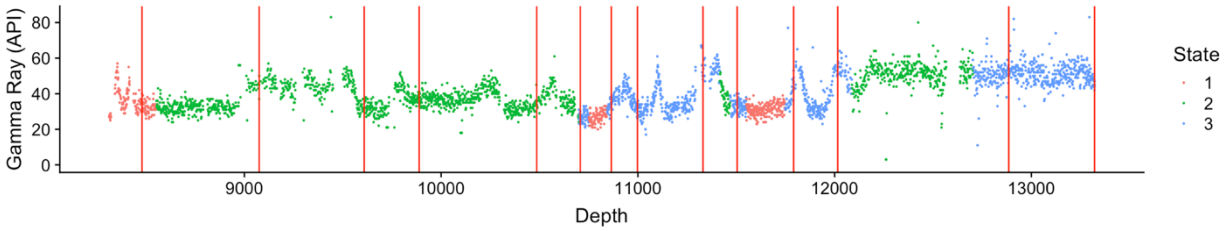

Fig. B4 – A plot showing the MWD measured gamma ray colored by HMM states with change points shown. In some cases, the change points and HMM states are able to detect sudden changes in the gamma ray measured by the MWD tool.

Well 3

The input parameters of the sliding window CPD algorithm are window size = 110 ft and a linear Penalty = 1.0.

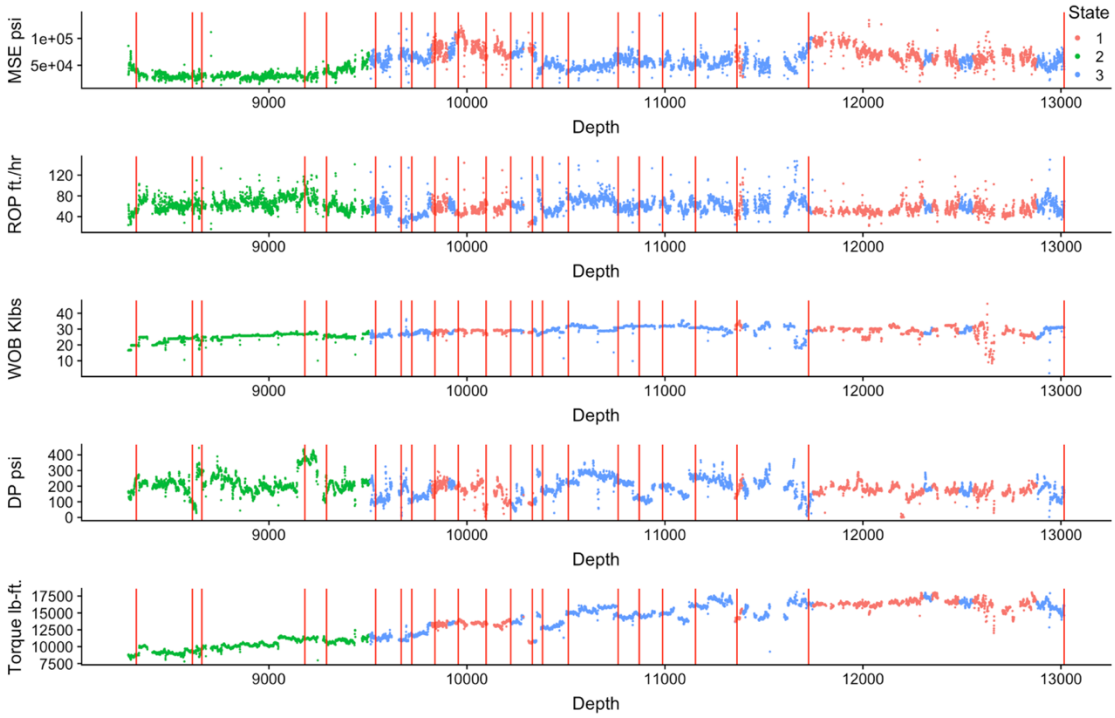

Fig. B5 – Plots showing the surface drilling data for well 3 colored by HMM state with change points shown. In this example, there are fewer state transitions; however, the CPD algorithm appears to be overly sensitive.

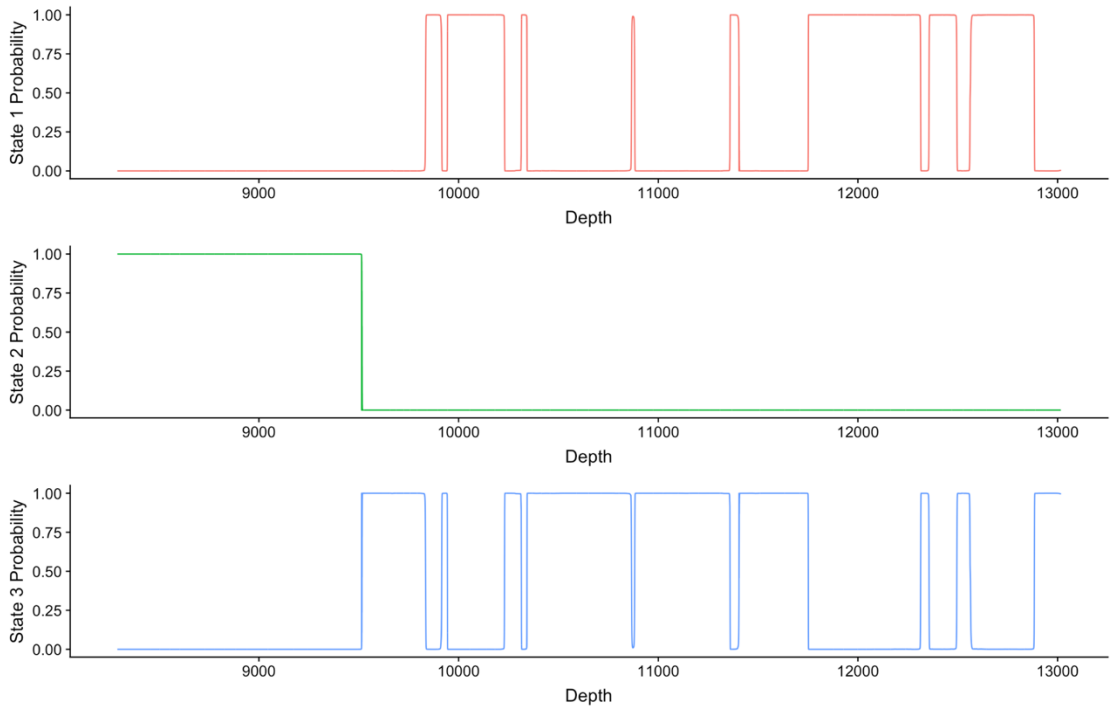

Fig. B6 – Plots showing the HMM state probabilities for each prediction depth. The smoothed data reduces variation in the signal resulting in very few HMM state transitions.

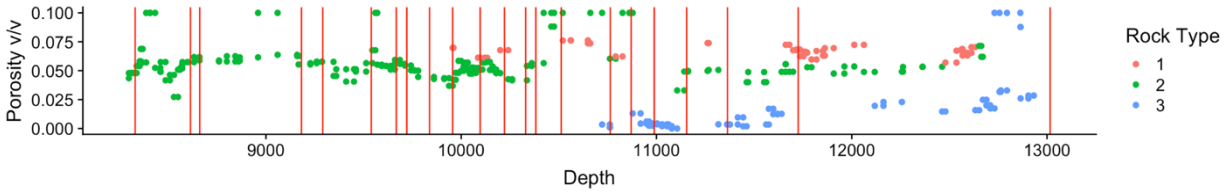

Fig. B7 – A plot showing the geomodel derived porosity for Well 3 colored by rocktype with the detected change points shown in red vertical lines. I see little to no correspondence of change points with rocktype transitions in this case. I have discussed the reasons earlier in Chapter 4.

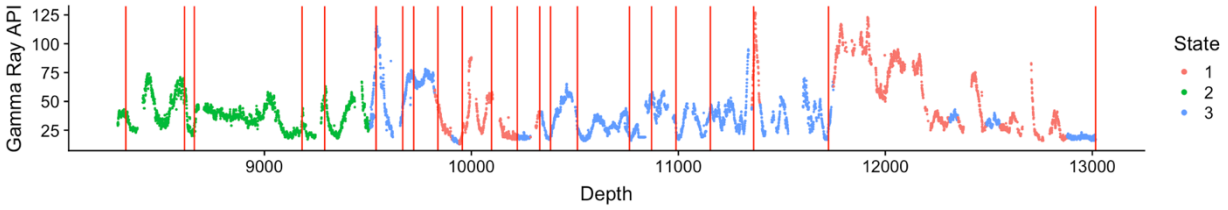

Fig. B8 – A plot showing the MWD measured gamma ray for Well 3 colored by HMM state with the change points shown. The HMM states correspond to large transitions in the gamma ray values while the change points do satisfactorily except in the lower section of the well where no change points are detected.

For Well 3, it appears that using a 110 ft. moving average to smooth the drilling signal impairs the utility of the HMM. I repeat the analyses with the raw, un-smoothed data and present the results below.

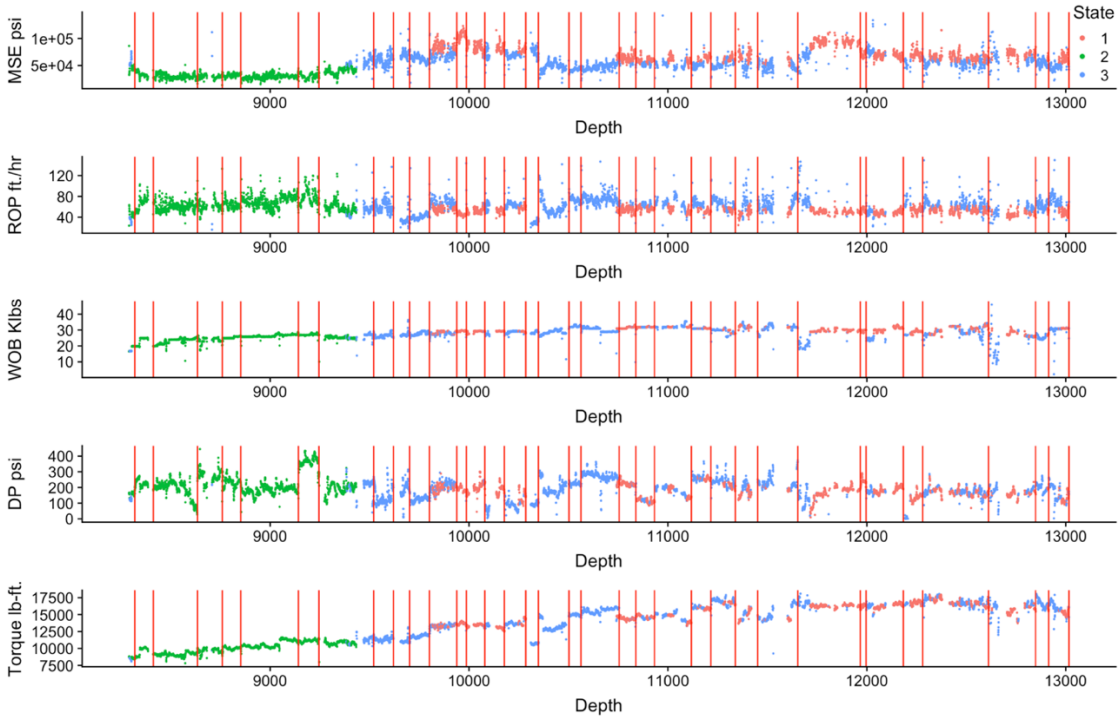

Fig. B9 – Plots of the surface drilling data for Well 3 with the HMM states and change points derived on the raw non-smoothed data. By using the non-smoothed data, more state transitions are identified in the signal, many of which occur at the change points.

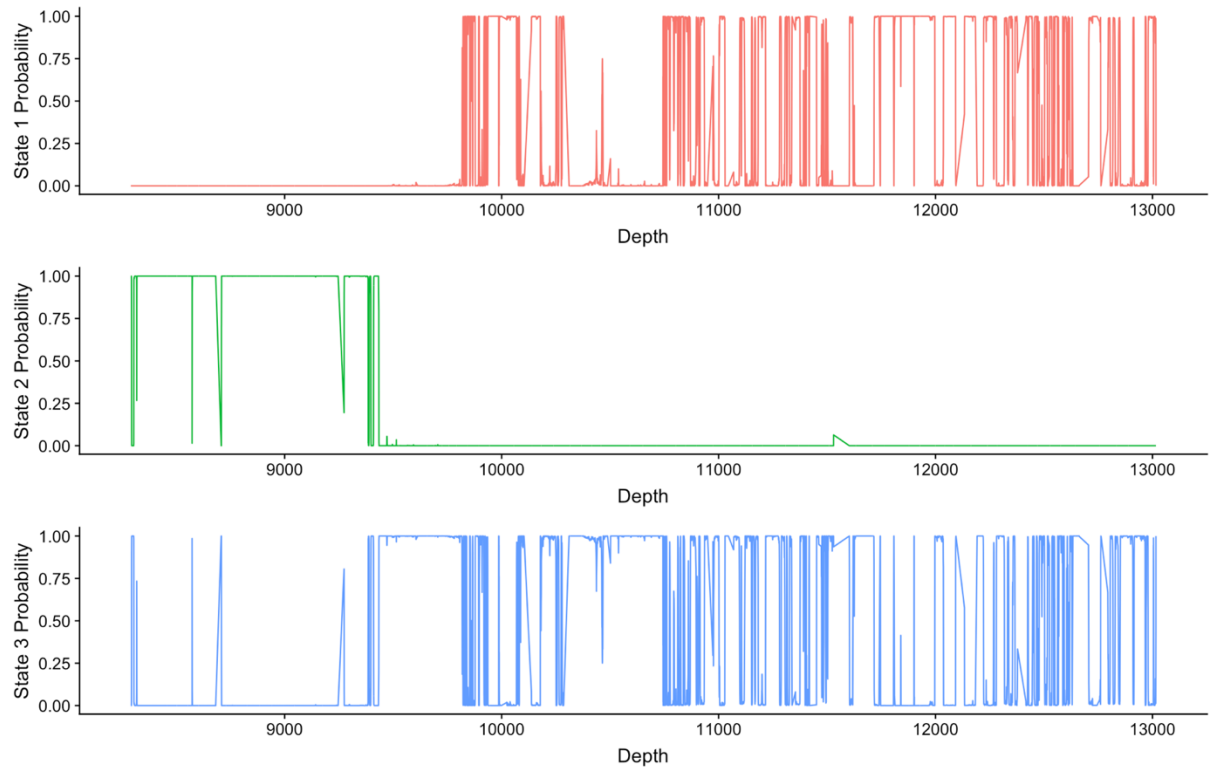

Fig. B10 – Plots showing the HMM state probabilities at each prediction depth for well 3 on the non-smoothed data. This shows that there are extremely rapid transitions between drilling states 1 and 2 for the last half of the well bore. This shows that some form of smoothing is likely required to obtain confident state predictions without rapid fluctuations which reduce the utility of identified states.

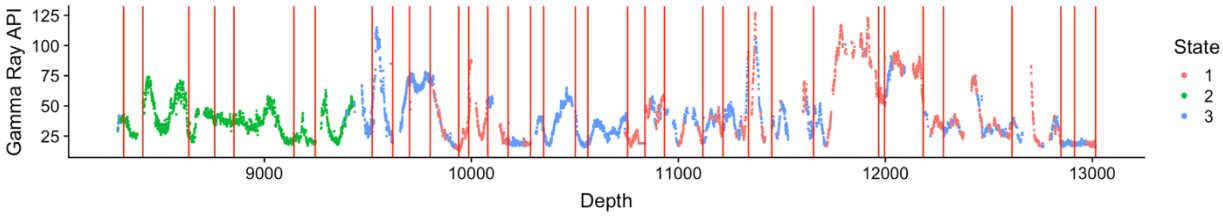

Fig. B11 – Plot showing the MWD measured gamma ray colored by HMM state with the change points shown. Both HMM states and change points are derived using the raw non-smoothed signal.

Shown above are the results for Well 3 without any smoothing of the surface drilling signal. While some of the additional state transitions better align with change points than in the smoothed signal, there are large amounts of rapid fluctuations between HMM states that reduce the utility of

state identification. If the states are transitioning so rapidly that there is not enough time for sufficient drilling to have occurred, it is unlikely the transitions will be related to significant changes in formation properties. Therefore, some level of smoothing is likely to be required for this analysis and can act as a tuning parameter to identify the lowest length moving average required to remove excess noise while retaining as much data structure as possible.

Appendix C: State Prediction Ahead of the Bit with HMM

HMM can be used in a predictive fashion using the derived transition probabilities. During drilling, an HMM can be built and fit on the data that is recorded up to the location of the bit. This includes estimation of the transition probabilities by the Forward-Backward algorithm. These probabilities can be used to identify the possibility of transitioning to another state at the future time step. In **Fig. C1** I show the resulting state classification on a portion of the data from well 15/9.

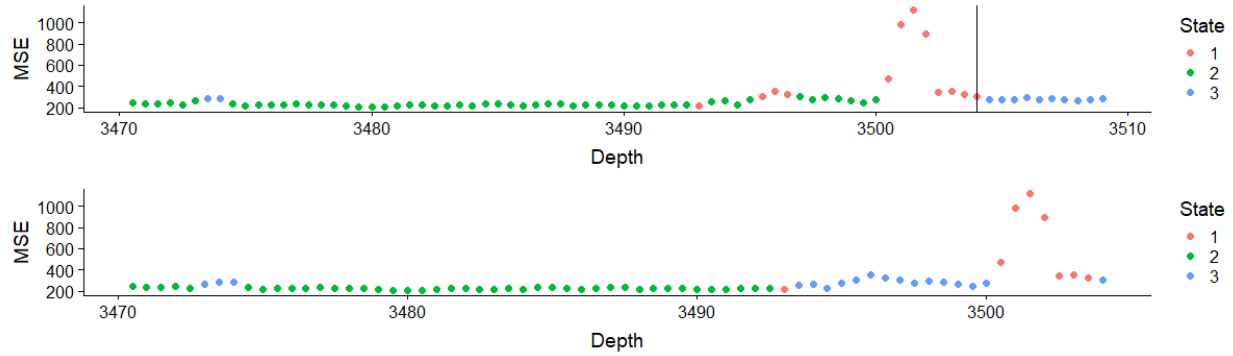

Fig. C1 – A plot of the resulting state classification for well 15/9 showing a portion of the well. The top figure shows the same predicted states from Chapter 3 shown earlier, for just a portion of the well. The bottom figure shows the resulting state prediction up to the black vertical line shown in the top figure.

In the bottom figure, states are predicted using only a portion of the data up to the black line shown on the top figure. This would be representative of predicting lithology during the drilling process for depths beyond the black line. Because an HMM can only predict the next state from the current state (and not any other future states because of the Markovian property), I only show the state prediction for the immediately following depth interval. At the location of the black line, the HMM predicts a transition to state 3 from state 1 which is the correct state ahead of the bit location. The transition probabilities for transitioning from state 1 to state 3 at this location is

21.3% which is the greatest transition probability for the HMM for transitioning to a different state for all states.

This HMM with parameters estimated up to the black line can be used to make predictions on new observations (surface drilling data) that come in from the real-time signal. **Fig. C2** shows the prediction of states for new data coming in past the black line.

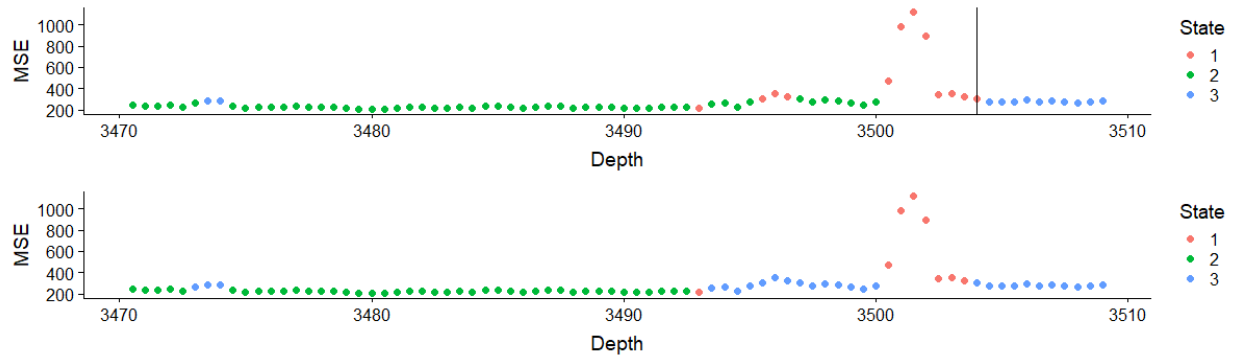

Fig. C2 – A plot showing the predicted states using an HMM fit up to the black line in the top figure. This example is representative of observations of surface drilling parameters recorded in real-time with the HMM making predictions for each observation past the black line. Predicted states past the black line match the states for the true predictions.

In this case, the predicted state (for intervals beyond the black line) perfectly matches the true state in the topmost figure. However, because the HMM was only trained with limited data, there is some mismatch in the preceding intervals which get refined as more data becomes available. In **Fig. C3** predictions for 150 ft. past the black line are shown and compared with the true predicted states. In **Fig. C4** the HMM is updated and refit using the observations recorded up to 150 ft. past the black line and compared with the true predicted states.

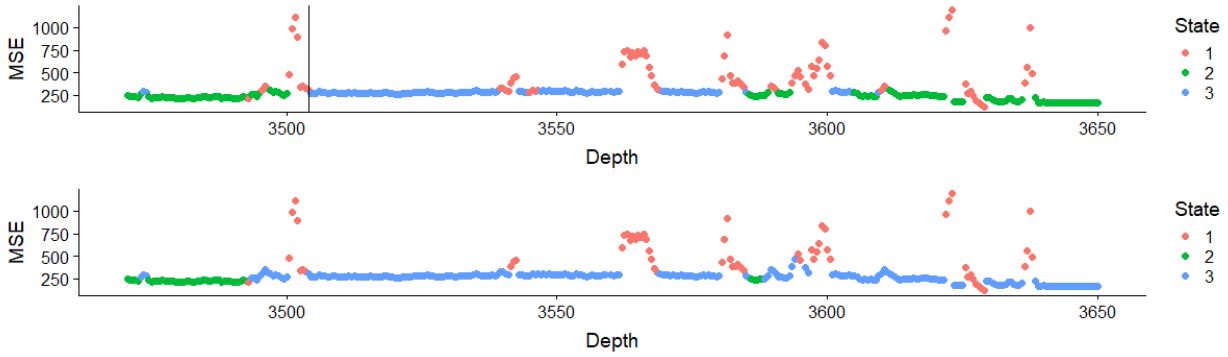

Fig. C3 – Predicted states for 150 ft. ahead of the bit (black line). Because of the Markovian property, states closer to the black line (up to which training data was available) are predicted accurately, but farther away, the prediction accuracy drops.

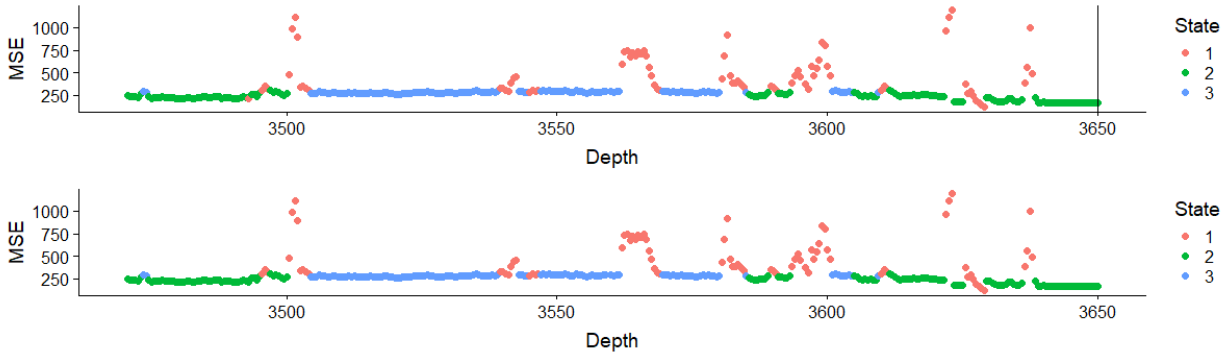

Fig. C4 – The topmost figure shows the HMM states using all the data from 3450ft to 3700 ft. The bottom figure shows the predicted states using data up to 3650 ft. It is evident that with more data, the HMM does a better job at predicting states.

The two figures, **Fig. C3** and **Fig. C4** show the utility of fitting and refitting an HMM as additional observations become available. For a short distance ahead of the previous fit, there is likely to be a good match between the true and predicted states; however, as more observations are made it is beneficial to periodically refit the HMM using the additional observations.