Time Series Pattern Discovery Techniques for Well-to-Well Log Correlation

Mr. Chanchai Apiwatsakulchai

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Georesources and Petroleum
Engineering
Department of Mining and Petroleum Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2017

กลวิธีการค้นพบแบบรูปอนุกรมเวลาสำหรับการเทียบสัมพันธ์ข้อมูลการหยั่งธรณีระหว่างหลุม

นายชาญชัย อภิวัฒน์สกุลชัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมทรัพยากรธรณีและปิโตรเลียม ภาควิชาวิศวกรรมเหมืองแร่และปิโตรเลียม
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560

Thesis Title	Time Series Pattern Discovery Techniques for Well-to-Well Log Correlation
By	Mr. Chanchai Apiwatsakulchai
Field of Study	
Thesis Advisor	Assistant Professor Suwat Athichanagorn, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

....Dean of the Faculty of Engineering

(Associate Professor Supot Teachavorasinskun, D.Eng.)

THESIS COMMITTEE

....Chairman

(Assistant Professor Jirawat Chewaroungroaj, Ph.D.)

....Thesis Advisor

(Assistant Professor Suwat Athichanagorn, Ph.D.)

....Examiner

(Falan Srisuriyachai, Ph.D.)

....External Examiner

(Sethavidh Gertphol, Ph.D.)

ชาญชัย อภิวัฒน์สกุลชัย : กลวิธีการค้นพบแบบรูปอนุกรมเวลาสำหรับการเทียบสัมพันธ์ข้อมูลการหยั่งธรณีระหว่างหลุม (Time Series Pattern Discovery Techniques for Well-to-Well Log Correlation) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. สุวัฒน์ อธิชนากร, 121 หน้า.

ข้อมูลการหยั่งธรณีเป็นหนึ่งในแหล่งข้อมูลที่มีจำนวนมากที่สุดในบรรดาข้อมูลสำหรับการจัดจำแนกใต้ธรณีทั้งหมดเนื่องจากหลุมที่ได้รับการเจาะแล้วเกือบทุกหลุมจะถูกทำการหยั่งธรณี การจัดจำแนกใต้ธรณีต้องการข้อมูลการหยั่งธรณีที่มีการเทียบสัมพันธ์แล้ว กระบวนการที่ได้รับการพัฒนาในการศึกษานี้ช่วยในการเทียบสัมพันธ์ข้อมูลการหยั่งธรณีระหว่างหลุมโดยใช้การศึกษาสองแบบที่พบได้ทั่วไปคือการจับคู่แบบรูปและการค้นพบแบบรูป

การจับคู่แบบรูปเน้นใช้หลักการค้นหาแบบรูปที่ได้รับความสนใจในข้อมูลการหยั่งธรณี การใช้ระยะทางแบบยุคลิด, ระยะทางแฮมมิง, และระยะทางเลเวนชเตย์นในการวัดความคล้ายร่วมกับการใช้การแทนข้อมูลแบบสัญลักษณ์ Symbolic Aggregate approXimate (SAX) ช่วยในการสร้างการแทนข้อมูลที่มีการลดความละเอียดจากข้อมูลในขณะที่ยังคงลักษณะเฉพาะหลักของข้อมูลอยู่ การวิเคราะห์หลายความละเอียดทำให้ได้คู่เหมือนที่มีความเป็นไปได้มากที่สุดจากความละเอียดของข้อมูลที่แตกต่างกันเนื่องมาจากการตัดแบ่งหลายระดับ ผลการศึกษานี้แสดงให้เห็นว่าการวิเคราะห์หลายความละเอียดระบุคู่เหมือนได้จากช่วงข้อมูลที่ปรากฏบ่อยครั้งที่สุด ผลการศึกษาแสดงให้เห็นว่าการวิเคราะห์หลายความละเอียดช่วยในการค้นหาคู่เหมือน และระยะทางเลเวนชเตย์นมีความสามารถที่สูงกว่าระยะทางแฮมมิงและระยะทางแบบยุคลิดในการค้นหาคู่เหมือนแม้จะมีการผันแปรเฉพาะที่ปรากฎอยู่

การค้นพบแบบรูปเน้นในการค้นหาแบบรูปซ้ำกระจายทั่วข้อมูลการหยั่งธรณีโดยไม่ต้องมีความรู้เกี่ยวกับแบบรูปในข้อมูลการหยั่งธรณีมาก่อน Motif Kymatology (MK) ถูกใช้เป็นกระบวนวิธีฐานในการศึกษานี้ ขั้นตอนที่ถูกแปลงสำหรับใช้ในกระบวนการค้นพบแบบรูปนี้ถูกใช้เพื่อเพิ่มเสถียรภาพในผลของการเทียบสัมพันธ์สุดท้าย

การศึกษาทั้งสองแบบนี้สามารถใช้สร้างกระแสงานแบบผสมผสานเพื่อรองรับการเทียบสัมพันธ์ข้อมูลการหยั่งธรณีและการจัดจำแนกใต้ธรณี

ภาควิชา วิศวกรรมเหมืองแร่และปิโตรเลียม ลายมือชื่อนิสิต ....

สาขาวิชา วิศวกรรมทรัพยากรธรณีและปิโตรเลียม ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

ปีการศึกษา 2560

5871208021 : MAJOR GEORESOURCES AND PETROLEUM ENGINEERING

KEYWORDS: WELL-TO-WELL LOG CORRELATION / SIMILARITY MEASUREMENT / PATTERN MATCHING / PATTERN DISCOVERY

CHANCHAI APIWATSAKULCHAI: Time Series Pattern Discovery Techniques for Well-to-Well Log Correlation. ADVISOR: ASST. PROF. SUWAT ATHICHANAGORN, Ph.D., 121 pp.

Well log data is one of the most abundant data sources for subsurface characterization since almost all drilled wells are logged. Subsurface characterization requires correlated well log information. Processes to help perform well-to-well log correlation have been developed in this study based on the two approaches: pattern matching and pattern discovery.

Pattern matching focuses on finding a known pattern of interest in uncorrelated wells. The use of Euclidean, Hamming, and Levenshtein distances in similarity measurement, Piecewise Aggregate Approximation (PAA), and Symbolic Aggregate approXimation (SAX) data representation helps create reduced-resolution while maintains the main characteristics of the signals. Multi-resolution analysis provides the most probable match from different data resolutions obtained by different discretization levels. The results show that multi-resolution analysis can identify the most probable match from the most frequent data window promoted. The results also show that Levenshtein distance is far superior to Hamming and Euclidean distances in finding the best match even when local variations are present.

Pattern discovery focuses on finding repeating patterns without any prior knowledge of patterns that might exist in the well logs. MK (Motif Kymatology) is used as the base algorithm in this study. For the data sets tried in this study, the proposed algorithm can successfully identify repeating patterns. A modified step is incorporated in order to help increase stability of the final correlated well sections.

In summary, pattern matching and pattern discovery can be formed as an integrated workflow to support well log correlation task and subsurface characterization.

Department: Mining and Petroleum Engineering
Student's Signature ....
Advisor's Signature ....

Field of Study: Georesources and Petroleum Engineering

Academic Year: 2017

ACKNOWLEDGEMENTS

I would like to express my gratitude toward Assistant Professor Suwat Athichanagorn for his consistent, supportive, and sincere guidance and encouragement throughout the course of the study. I would also like to thank Assistant Professor Jirawat Chewaroungroaj, Dr.Falan Srisuriyachai, and Dr.Sethavidh Gertphol, who are members of this thesis committee for their helpful comments and suggestions.

I would like to thank Chevron Thailand Exploration and Production Ltd. for its scholarship scheme provided under Thailand University Partnership Program (UPP) throughout this study. This helps fulfill my wish to obtain an insightful knowledge in upstream petroleum industry.

I would like to thank Dr.Pichit Vardcharragosad, for his viewpoints and suggestions on conducting a research in a graduate-level study while I was a graduate intern under his supervision at PTT Exploration and Production Co., Ltd.

I would like to thank Kansas Geological Survey for making comprehensive oil and gas exploration and production data available and allowing me to use the data to conduct and complete this study.

I would like to express my great appreciation to my classmates and all other faculty members who make some challenging activities become effortless and, more importantly, make my experience in the study become one of my greatest experiences.

I would also like to thank my friends whom I have met throughout my time during my professional life for their encouragement throughout the time of the study.

Lastly, I would like to express my deepest gratitude to my family who have always supported my one of the boldest decisions to take this journey, provided everything needed, and encouraged me towards my accomplishment of this graduate study.

CONTENTS

Page

THAI ABSTRACT iv

ENGLISH ABSTRACT v

ACKNOWLEDGEMENTS vi

CONTENTS vii

List of Figures ix

List of Tables xv

Chapter 1 Introduction 1

1.1 Problem Statement 2

1.2 Objectives 3

1.3 Expected Usefulness 4

1.4 Methodology 4

1.5 Well Log Correlation Approaches 6

1.5.1 Approach A: Pattern Matching 6

1.5.2 Approach B: Pattern Discovery 6

Chapter 2 Literature Review 7

Chapter 3 Theory and Concept 10

3.1 Pattern Matching via Similarity Measurement 10

3.1.1 Data Representation 11

3.2 Pattern (Motif) Discovery using MK Algorithm 23

3.2.1 Intuition behind MK Algorithm 24

3.2.2 MK Algorithm Formal Statements 28

Chapter 4 Methodology 32

Page

4.1 Well Log Data Preparation 32
4.2 Pattern Matching 33
4.3 Pattern Discovery 39
Chapter 5 Pattern Matching........ 48
5.1 Example Correlation 1........ 50
5.1.1 Multi-resolution Analysis 52
5.2 Example Correlation 2........ 61
5.2.1 Multi-resolution Analysis 62
5.3 Example Correlation 3........ 68
5.3.1 Multi-resolution Analysis 69
5.4 Multiple Well-to-Well Log Correlation........ 74
Chapter 6 Pattern Discovery........ 76
6.1 Example correlation 1 76
6.2 Example correlation 2 95
6.3 Multiple Well-to-Well Log Correlation using Pattern Discovery 110
6.4 An Integrated Approach to Well-to-Well log Correlation 114
Chapter 7 Conclusions and Recommendations........ 115
7.1 Conclusions 115
7.1.1 Pattern Matching........ 115
7.1.2 Pattern Discovery 116
7.2 Recommendations........ 117
REFERENCES 118
VITA........ 121

LIST OF FIGURES

Figure 3.1 Original data series 15

Figure 3.2 z-transformed data series........ 15

Figure 3.3 Time series discrete data representation using SAX........ 18

Figure 3.4 Levenshtein distance calculation example........ 21

Figure 3.5 Original-space distances, linear-space distances, and lower bounds 26

Figure 3.6 Step-by-step distance comparisons and best-so-far updates. 27

Figure 3.7 MK Algorithm 31

Figure 4.1 Five smallest distances promoted by conventional NNS 33

Figure 4.2 Approximate signals based on 15 PAA blocks 34

Figure 4.3 Approximate signals based on 25 PAA blocks 34

Figure 4.4 Approximate signals based on 35 PAA blocks 35

Figure 4.5 Approximate signals based on 45 PAA blocks 35

Figure 4.6 Symbolic representation as part of SAX representation on the model pattern........ 36

Figure 4.7 Comparisons of distance measured by Hamming distance and Levenshtein distance........ 36

Figure 4.8 The best match data window promoted by multi-resolution analysis........ 37

Figure 4.9 Pattern Matching process overview workflow 38

Figure 4.10 An artificial well log from a concatenation of Well Y (top: above the breakpoint) and Well X (bottom: below the breakpoint)............................ 40

Figure 4.11 The artificial well log with SD boundaries.. 42

Figure 4.12 First-rank motif pair 44

Figure 4.13 Second-rank motif pair 44

Figure 4.14 Third-rank motif pair 45

Figure 4.15 Final well-to-well log correlation of the section length of 120 indexes (120 data points or 60 ft) using Window 4300 on the artificial well log as the final reference series.. 46

Figure 4.16 Pattern (motif) discovery using MK algorithm.. 47

Figure 5.1 Sliding window of a section of interest... 50

Figure 5.2 Five smallest distances promoted by conventional NNS 51

Figure 5.3 PAA signals on the model pattern using different number of PAA blocks.. 53

Figure 5.4 Selection of number of PAA blocks using Euclidean distance between approximate well log and its original counterpart.. 53

Figure 5.5 Approximate signals based on 15 PAA blocks... 54

Figure 5.6 Approximate signals based on 25 PAA blocks... 54

Figure 5.7 Approximate signals based on 35 PAA blocks... 55

Figure 5.8 Approximate signals based on 45 PAA blocks... 55

Figure 5.9 Well log value (y-axis) discretization using various equi-probable SAX sections 56

Figure 5.10 Comparisons of distance measured by Hamming distance and Levenshtein distance........ 58

Figure 5.11 The best match data window promoted by multi-resolution analysis 59

Figure 5.12 Five smallest distances promoted by conventional NNS 61

Figure 5.13 Selection of number of PAA blocks using Euclidean distance between approximate well log and its original counterpart........ 63

Figure 5.14 Approximate signals based on 20 PAA blocks........ 63

Figure 5.15 Approximate signals based on 25 PAA blocks........ 64

Figure 5.16 Approximate signals based on 30 PAA blocks........ 64

Figure 5.17 Approximate signals based on 35 PAA blocks........ 65

Figure 5.18 Approximate signals based on 40 PAA blocks........ 65

Figure 5.19 Approximate signals of Window 1281 under 20, 25, and 30 PAA blocks over z-transformed signals. 66

Figure 5.20 Approximate signals of Window 8 under 30, 35, and 40 PAA blocks......... 66

Figure 5.21 Comparisons of distance measured by Hamming distance and Levenshtein distance........ 67

Figure 5.22 The best match data window promoted by multi-resolution analysis 68

Figure 5.23 Five smallest distances promoted by conventional NNS 69

Figure 5.24 Selection of number of PAA blocks using Euclidean distance between approximate well log and its original counterpart.. 70

Figure 5.25 Comparisons of distance measured by Hamming distance and Levenshtein distance.. 71

Figure 5.26 The best match data window promoted by multi-resolution analysis 71

Figure 5.27 Multiple well-to-well log correlation using pattern matching approach... 75

Figure 6.1 Sample well logs to be correlated 78

Figure 6.2 An artificial well log of length 5,476 indexes generated by connecting two well logs: Well Y and Well Z with a breakpoint (at dashed line)........ 79

Figure 6.3 Sliding window.. 81

Figure 6.4 Arbitrarily chosen 21 sections based on -5.0 to 5.0 SD span 83

Figure 6.5 21 SD intervals and 14 reference sections each randomly chosen from -5.0 SD to 5.0 SD of the artificial well log mean... 87

Figure 6.6 Reference series of the highest-to-smallest ranks based on standard deviation of distance values.. 89

Figure 6.7 The third-rank motif pair of Window 1168 and Window 3923 with reference series of Window 2112. 91

Figure 6.8 The second-rank motif pair of Window 145 and Window 2887 with reference series of Window 2112. 91

Figure 6.9 The first-rank motif pair of Window 1480 and Window 4235 with reference series of Window 2112. 92

Figure 6.10 The final correlation of section length 120 indexes (60 ft) 93

Figure 6.11 Well logs to be correlated.. 95

Figure 6.12 An artificial well log of length 5,482 indexes generated by connecting two well logs: Well Y and Well Z with a breakpoint (at dashed line). 96

Figure 6.13 Artificial well log with SD boundaries.. 99

Figure 6.14 21 reference sections each randomly chosen from -5.0 SD to 5.0 SD of the artificial well log mean. 100

Figure 6.15 Steps used to generate a set of indexes pointing to lowest-to-highest distances based on the set with largest SD.. 102

Figure 6.16 Reference series of the highest-to-smallest ranks based on standard deviation of distance values.. 103

Figure 6.17 The forth-rank matched pair (motif pair) of Window 1557 and Window 4632 with reference series of Window 3067. 105

Figure 6.18 The third-rank matched pair (motif pair) of Window 1487 and Window 4239 with reference series of Window 3067. 105

Figure 6.19 The second-rank matched pair (motif pair) of Window 1184 and Window 3932 with reference series of Window 3067. 106

Figure 6.20 The first-rank matched pair (motif pair) of Window 124 and Window 2861 with reference series of Window 3067. 106

Figure 6.21 Well log correlation mismatch. 108

Figure 6.22 Corrected well log correlation after a mismatch is excluded.................. 109

Figure 6.23 Well-to-well log correlation using pattern discovery approach 111

Figure 6.24 Well-to-well log correlation using pattern discovery approach 112

Figure 6.25 Multiple well-to-well log correlation using pattern discovery approach with human intervention to align some of the correlated sections. 113

LIST OF TABLES

Table 3.1 A look up table providing list of breakpoints corresponding to number of breakpoints required. 14

Table 3.2 Data range from PAA discretization 16

Table 3.3 Data range from PAA discretization (continued) 17

Table 3.4 An example of Hamming distance calculation on a bit string 20

Table 4.1 Well log data source 32

Table 4.2 Well log value from -5.0 SD to 5.0 SD span in 0.5 SD interval span 43

Table 5.1 Example correlations and the parameters used 73

Table 5.2 Section length and the locations where the section is correlated across Well X, Well Y, and Well Z 75

Table 6.1 Well log value from -5.0 SD to 5.0 SD span in 0.5 SD interval span 86

Table 6.2 Example correlations and the parameters used 94

Table 6.3 Well log value from -5.0 SD to 5.0 SD span in 0.5 SD interval span 98

Table 6.4 Example correlations and the parameters used 107

CHAPTER 1
INTRODUCTION

One of the data acquisition techniques used to obtain data for subsurface characterization is well logging. Well logging is probably the most attractive method to collect basic lithological data and fluid type as the method can provide data within moderate range from the wellbore deep into the formation depending on depth of investigation of logging tools used. Its usefulness is also the time used to obtain required data is relatively small as multiple logging tools for varying properties can run into the wellbore at the same time. According to the benefits aforementioned, well logging is usually performed on multiple wells in order to capture possible deviation of properties across different wells even though their locations are relatively close and their depths are on the same range.

Added values from having multiple well logs are not only for data collection and their interpretation on a well-by-well basis, but it also provides crucial subsurface characterization benefit from a process called *correlation*. Correlation, or more frequently called *well log correlation*, is a process aiming to match a similar portion of interest using the same log type across one or more log responses from different wells. These similar portions obtained from well log correlation process are usually used for many purposes and one crucial purpose is to use them for modelling connections of rock and their properties across depth ranges. The similar portions drawn from multiple wells are generally thought to have certain degree of similarity in terms of properties recorded and their inferences. For example, similar gamma ray log responses (in API degree) behind a potential reservoir section may be used to indicate quality of that potential reservoir rock in this section of interest and the quality may not be on the same degree assuming that lower gamma ray means a better quality of the section of interest since it has lower shale content and is more preferable.

There are efforts trying to perform well log correlation using many techniques apart from conventional well log correlation involving manual process of locating similar sections across multiple wells. The sections may be in the form of log values falling in the same range, having the similar trend with or without *scale variation*[1]. Owing to advancement of data processing in computing realm and the use of digital well log, there has been interests to digitally process well log data and perform well log correlation using different techniques such as extension of moments method and polynomial regression [1], dynamic wave matching [2], signal processing and other related methods (Wavelet Transform, Hilbert Huang Transformation, and Empirical Mode Decomposition and Variational Mode Decomposition) [3-6], artificial intelligence techniques [7, 8]. These techniques listed are not meant to be comprehensive but it is to provide a list of recent advances in well log correlation.

Since well log data is a value on depth scale, different techniques may treat the data differently. For example, polynomial regression is a technique that typically works directly on real domain (value versus depth) while signal processing and related techniques typically works on frequency domain (amplitude versus frequency). Therefore, well log data has to be converted from real domain to frequency domain in order to use signal processing methods.

Ability to process well log data directly on real domain possesses benefits as there is no requirements to change its domain to process, thereby reducing error and bias due to changing domain, and it is simpler to understand with less computation.

1.1 Problem Statement

To date, it may still be able to say that conventional well log correlation is the main practice performed. Conventional well log correlation typically involves approximating a portion of well log response to match another portion of a different well log response. Sometimes, this approximation matching is, unfortunately, arbitrary. At one time, one small deviation from the main trend may be neglected and the

[1] Scale variation in well log is a variation where certain log responses from two wells show some degrees of extension or contraction along the depth axis while major trends of those logs are similar.

match is treated as a satisfactory match, thereby correlating two sections one from each well together. At another time, small deviation is chosen to be included in the correlation. Admittedly, inconsistency of this practice may be masked as an engineer's judgement. Instead of arbitrary inclusion or exclusion of a portion of well log response, it is better to quantify the match before making any judgement on a correlation based on those quantifiable numbers.

Finding interested stratigraphic units that are small sections spread along an ultra-deep well using convention method can be a daunting task and there can be a lot of mistake involved. For example, less similar sections may likely be chosen instead of more similar section since there is no quantifiable figures. In fact, one could actually calculate some basic statistical figures such as mean and standard deviation for all possible sections. Then selection of the sections can be made based on comparisons of the calculated figures. This is by far not the best method that can be done.

Well log interpretation may be done on a certain objective, which may result in looking for a certain pattern or trend in a well log. This means it is possible that some patterns may unintentionally be neglected, thereby overlooking valuable information. This overlooked information may have a significant effect to the reservoir modelling.

Problems aforementioned alone are sufficient to promote more systematic and quantifiable approach to well log correlation task since those problems are crucial to the correlation result. In addition, ability to extract all patterns and make quantifiable match will result in great benefits for any future subsurface study.

1.2 Objectives

As previously mentioned, study objectives are to develop a systematic workflow to meet the requirements as follows:

- To semi-automatically find interested well sections of all lengths using gamma ray log.
- To correlate well sections across multiple wells

1.3 Expected Usefulness

- Data-adaptive pattern discovery method requires little priori knowledge of a pattern. This approach is more beneficial especially when data pool is small, which is a situation when a model-based pattern discovery method is hard to be applied and statistical-based pattern discovery may give high uncertainty.
- Development in time series motif discovery has opened a new way of extracting repeating patterns in a well log. Meaningful repeating patterns are crucial to modern well log analysis.
- Combination of motif discovery techniques can be used on non-linear and non-stationery nature of well log data to extract repeating patterns that may be meaningful and beneficial for modern well log analysis.
- Subsequences extracted from a well log are important to well log analysis as they contain useful information such as lithologic sequence, level of energy as a log response, and etc. These information drawn from the subsequences, in turns, can be used as fundamental elements for classification of local characteristic of accumulation and correlation of similar patterns across spatially-connected wells.

1.4 Methodology

There are, normally, more than one well drilled in a particular area as part of data acquisition for reservoir characterization. One of the challenges is how to correlate similar stratigraphic unit across multiple well logs. Oftentimes, similar stratigraphic unit in one depth interval may not necessarily be correlated with the similar unit in another depth interval. This challenge is, therefore, not entirely easy to tackle. Owing to advances in data mining techniques for pattern discovery, there are methods to help correlating two *similar but different* series together. This study will illustrate how to apply those techniques to tackle well log correlation challenge.

Conventional well log correlation, typically, involves finding 2 characteristic marks on the top and at the bottom one each that covers the interested zone to see if those logs being correlated are on-depth relative to each other [9]. After the marks are found, the real process of conventional well log correlation starts. These 2 steps may be done repetitively for the entire log depth for more than one zone of interest as all zones correlated are crucial for reservoir modelling in the later stages of a reservoir simulation study. In this study, however, it is more preferable to use a more heuristic method to digitally correlate well logs of all possible sections.

In data mining literatures, there are many techniques that can find a pair of sections that are the most similar compared to the rest of the logs. The nature of well log data considered, it is possible to say that time series pattern discovery or, more specifically, motif discovery can be applied to well log correlation problem. Motif discovery involves trying to find recurrent patterns in a time series data. Ability to find recurrent patterns in a time series is beneficial and there has been numerous application in other industries. In this well log correlation problem, if it is possible to identify two sections one from each well as a motif pair, it is also reasonable to correlate those sections from those wells together, provided that those are located on relatively the same depth interval. This process may be referred to as *motif discovery*.

Depending on motif discovery strategy, there may be some cases where similar patterns to a motif are still on the well log data undiscovered. In these cases, it may be necessary to perform a search routine in order to capture all those similar patterns to a particular motif. One method that is used as a search routine is Nearest Neighbor Search. This process may be referred to as *nearest neighbor search*.

In order to perform well log correlation as previously mentioned, there is an absolute requirement to perform similarity measurement. While similarity measurement seems to be a simple process, there are, however, challenges in defining degree of similarity between 2 sections which, in this case, is 2 well logs. Similarity measurement may be done directly on real domain. Alternatively, it may be done on a discretized domain using a certain representation applied to data on real domain before performing similarity on that discretized domain. It is considered to be vital to well log correlation results. This process may be referred to as *similarity measurement*.

The following steps are steps performed in this study. They are intended to give a more concrete example of aforementioned conceptual steps. Since this study involves substantial amount of software programming task, the outlined steps may, at times, be highly dependent on the implemented programs in this study.

1.5 Well Log Correlation Approaches

Well-log correlation may be divided into two approaches based on well-log correlation strategies.

1.5.1 Approach A: Pattern Matching

This approach primarily focuses on finding a known pattern, which may be from a known section having the pattern of interest, in other well logs that have not been correlated such as well logs from recently drilled wells. The main strategy of this approach is to find if there is the pattern of interest in those well logs. It is also used to determine the best match of the pattern to those well logs. In order to identify the best match, similarity measurement is employed in the matching process.

1.5.2 Approach B: Pattern Discovery

This approach focuses on finding repeating patterns across well logs without any prior knowledge of patterns that might be existed in the well logs. As the name implies, the approach aims at extracting inherent patterns from the well logs.

CHAPTER 2
LITERATURE REVIEW

As mentioned in Chapter 1, there are many techniques that can be used to perform well log correlation. This section provides a review of previous works relating to digital processing of well log for correlation.

Lineman et al. [10] presented a system employing Dynamic Time Warping (DTW), a technique used in speech recognition, together with a specific knowledge base of geological information to perform well-to-well log correlation from domain experts commonly found in development of expert system. The system developed would try to first establish a local cost from comparing between two logs. Then representation of the log data to discrete data was performed. Before using DTW to establish warping paths indicating correlations between two well logs, a distance matrix was created using string matching technique where insertion, deletion, or change of a character in a string give a different cost. Apart from the main objective of this work, the authors suggested that an explanatory description or conclusion of a correlation could also be drawn as each discrete value was a representative to some specific rules. Typical to any expert system, the system requires a regular update of its knowledge base from field to field.

Vega [4] developed techniques to correlate well logs using wavelet analysis techniques. The techniques aims at detecting cyclostratigraphic sequences and true boundaries with absent of core data. Cyclostratigraphy was processed by applying wavelet analysis technique in order to assess wavelengths of the strongest cyclicities. Using scaleogram, wave properties such as frequency constituents were able to be extracted. Comparisons of using wavelet analysis, Fourier analysis, and semivariogram showed that Morlet wavelet scaleogram was superior to Fourier analysis and semivariogram in detecting cyclostratigraphy as the method could reveal superimposed cycles of two different frequencies and also the locations of those frequencies on the depth scale while Fourier analysis and semivariogram only identified two superimposed frequencies. True boundary detection was performed on a discretized well log space. Feature extraction was performed at boundary-upper and boundary-lower windows. Training wells with best characterized data were used with the observation wells to calculate total probability from multiplication of boundary-upper probability and boundary-lower probability. Then a boundary was assigned to observation wells at the highest probability. This process was repeated until the training wells had no boundary left. The result was measured in detection performance and it was as high as 90%.

Wavelet analysis method requires a priori which is the wavelet model. Sometimes it requires trial-and-error process in order to select an appropriate wavelet to a particular problem. Many method are aiming at processing non-stationery data such as Empirical Mode Decomposition (EMD). As the name implies, EMD is mostly applied in an empirical manner [5]. Similar to wavelet decomposition, EMD can be used to decompose the original signals into fundamental signals called intrinsic mode. Dragomiretskiy and Zosso [11] developed Variational Mode Decomposition (VMD) that is based on mathematical foundation while it stills maintains data-adaptive wave decomposition technique. Li et al. [12] proposed VMD-based method to perform seismic sedimentary cycle and facies analysis in an unconventional reservoir in Fort Worth, Texas.

Apart from signal processing and their related techniques, there are also machine learning related techniques presented in the literatures. Luthi and Bryant [7] presented correlation technique using back-propagation neural network. The method can be divided as two sub-processes: one for datum correlation and another for marker correlation. The first sub-process was first trained from the key well using tapped delay line (or sliding-window) technique as an input layer. During the training, the network provided output 1 when a boundary was at exactly at the middle node of the input layer. The network itself contained two hidden layers and their number of processing nodes were determined empirically. The second sub-process was first trained with two dataset which are shale volume and vertical distance to datum. Locations at high confidence peaks from output confidence curve were identified as possible boundaries, which were then validated by geologist.

Fischetti et al. [13] presented a method to characterize shale and correlation well logs using density, neutron porosity, and acoustic transition log to calculate M and N parameters, which primarily are used in rock matrix mineralogical identification, as inputs to a competitive neural network. The network consisted of a competitive layer accepting two inputs. The competitive network for shale characterization was used to confirm lateral continuity. The result showed that the network successfully identified correct shale interval as the result agreed with the gamma ray log. Although there was no indication of number of neurons used in the shale characterization case, neurons in the competitive layer that was use in well correlation were as high as 20 neurons in order to allow redundancy and guarantee that all shale clusters were identified.

Cassisi et al. [14] applied time series motif discovery technique proposed by Mueen et al. [15] to investigate recurrent eruptive activity of Mt Etna during 1 January 2011 to 16 November 2011 from seismic volcano monitoring signals. The seismic signal is seismic amplitude time series which were computed using a root-mean square (RMS). The number of data points were 43,103 and the number of references were 10. The study investigated results due to varying parameters such as motif length (50, 100, and 150 – durations of seismo-volcanic phenomena of interests), coefficients used for motif range (2, 3, and 4), and number of desired groups of motifs (maximum at 20). Once the motifs had been extracted, they were undergone similarity measurement using average cross correlation. The study found that exact motif discovery technique using MK algorithm revealed different seismo-volcanic phenomena from differences in discovered RMS trends. For example, one group of RMS trends that exhibited sharp increases and decreases RMS (lasting from seconds to minutes) are earthquakes while another group of RMS trends that exhibited slower changes of RMS (in hours or days) are lava fountains.

To the best of the knowledge, there is no application of exact time series motif discovery technique in well log correlation problem. It may be due to the fact that the technique is still relative new. Therefore, it should be useful to investigate the applicability of exact time series motif discovery to well log problem.

CHAPTER 3
THEORY AND CONCEPT

3.1 Pattern Matching via Similarity Measurement

As previously discussed in Chapter 1, the first approach for well correlation process is pattern matching from a known pattern to any section of a well log of interest. In this approach, the known pattern is slid throughout the well log to be correlated. The best correlation for the pattern may be quantitatively considered as the pair that is the most similar. In order to determine how similar a pair is, similarity measurement techniques are used in this closeness assessment.

In order to measure closeness between two groups, an appropriate comparison technique is required. Depending on definition of closeness being considered, appropriate comparison techniques are typically governed by comparison objectives. Types of the objects also play an important role in choosing comparison technique. Numerical data series, for example, may be compared based on varying means such as direct numeric comparison, discrete representation comparison, descriptive statistical parameter comparison, and model coefficient comparison. Some techniques may provide comparison score which can later be used to quantify how close those two groups being compared are. This process of assessing closeness of two groups may be referred to as similarity measurement, which can be outlined as follows.

1. Once a comparison method is chosen, data preparation is performed according to pre-conditions specified by the method.
2. Data re-representation is typically applied if the method works on different space that that of the data's original space.
3. If more than two datasets are used in the comparison, cross comparison of all dataset may be needed.

4. Resulting comparison scores of each dataset pair are then used. A cutoff score may be defined in order to divide the scores into two groups of datasets that are under and out of a particular cutoff score.

Specific to this study, well log data is considered to be numerical data. Each data point is considered to be independent from its neighboring data points. Therefore, similarity measurement is performed based on its intrinsic value of each data point recorded by a logging tool. Distance measurement is a measurement responsible for generating a comparison score of each similarity measurement. Frequently, distance measurement is not arbitrarily chosen. It is, instead, governed by an algorithm chosen in another process. For example, a pattern discovery process may require a specific distance measurement in order to exploit a special property of the distance measurement. This is also the case of this study as this study employs MK algorithm [15] to find an exact motif pair. The algorithm explicitly uses Euclidean distance due to its specific properties as a distance metric.

Not only is Euclidean distance utilized in this study, but Hamming distance and Levenshtein distance are also used in this study. The following sections provide details about the mentioned distance measurements.

3.1.1 Data Representation

In order to utilize an existing similarity measurement, data to be compared may need to be representative in order to fit the requirement of that particular method. In this study, Hamming distance and Levenshtein distance require that data be character string or discrete form. Therefore, any numerical data is typically represented as character string. Converting one numeric data point to a character is possible. It, however, is uncommon as one benefit of data representation to be achieved is dimensionality reduction, which reduces number of data points to be a set of predefined characters. Although dimensionality reduction results in an approximate (not exact) comparison, it is beneficial as the comparison may be more meaningful due to smearing effect, leaving small changes in data trend to be group with the main trend of the same group which is then re-represented by a character.

There are many data representation techniques presented in literatures. One of which is Symbolic aggregate ApproXimation (SAX). SAX is a symbolic and data-adaptive method representing a long time series with length n to a predefined alphabet string of length w where w is much smaller than n $(w << n)$. Benefits of using SAX are as follows:

- Dimensionality Reduction: Similar to other data representations, SAX provides an approximate series from its original counterpart. Unlike other dimensionality reduction methods such as wavelet-based methods, the generation technique is simpler as it employs Piecewise Aggregate Approximation (PAA).
- Lower Bounding: A valid approximation serves as a representation of its original series while preserving similarity or dissimilarity and main characteristics of the original series on the approximated series. In other words, the dimension of approximated series is lower enough to still maintain majority of its original counterpart. Otherwise, it will not be useful as intrinsic features are changed after dimensional reduction.

Steps to create representative data series using SAX are as follows.

1. Perform z-normalization to the original data to convert the data to have mean $(\overline{x}) = 0$ and standard deviation $(SD) = 1$.

Perform PAA on the original data of length n to the approximate data of length w

Given original data series C of length n where $C = \left(c_1, c_2, c_3, \ldots, c_j, \ldots, c_n\right)$ and $j = 1, 2, 3, \ldots, n$. Since $n \in \mathbb{N}$ and $w \in \mathbb{N}$, it is obvious that $\frac{w}{n}$ and $\frac{n}{w} \in \mathbb{R}^{+}$, Approximated series $\overline{C}$ of length w where $\overline{C} = \left(\overline{c}_1, \overline{c}_2, \overline{c}_3, \ldots, \overline{c}_i, \ldots, \overline{c}_w\right)$ and $i = 1, 2, 3, \ldots, w$ can be generated from

$$\overline{c_i} = \begin{cases} \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j, & \frac{n}{w} \in \mathbb{N} \\ \left(1-\left(\left\lceil \frac{n}{w}(i-1) \right\rceil - \frac{n}{w}(i-1)\right)\right) \times c_j\Big|_{j=\left\lfloor \frac{n}{w}(i-1) \right\rfloor+1} + \sum_{j=\left\lfloor \frac{n}{w}(i-1) \right\rfloor+2}^{\left\lceil \frac{n}{w}i \right\rceil-1} c_j + \left(\left\lceil \frac{n}{w}i \right\rceil - \frac{n}{w}i\right) \times c_j\Big|_{j=\left\lceil \frac{n}{w}i \right\rceil}, & i < w \text{ and} \frac{n}{w} \notin \mathbb{N} \\ \left(1-\left(\left\lceil \frac{n}{w}(i-1) \right\rceil - \frac{n}{w}(i-1)\right)\right) \times c_j\Big|_{j=\left\lfloor \frac{n}{w}(i-1) \right\rfloor+1} + \sum_{j=\left\lfloor \frac{n}{w}(i-1) \right\rfloor+2}^{n} c_j, & i = w \text{and} \frac{n}{w} \notin \mathbb{N} \end{cases}$$

where range minimum value is $\frac{n}{w}(i-1)+1$and range maximum value is $\frac{n}{w}i$

This step results in a *block-like* data series of length *w*.

The following trivial example demonstrates how a range is specified when $\frac{n}{w} \notin \mathbb{N}$

Given $n = 33 \text{and } w = 4$

Therefore, $\frac{n}{w} = \frac{33}{4} = 8.25 \notin \mathbb{N} \text{ but} \frac{n}{w} \in \mathbb{R}^+$

ID	Lower Bound	Upper Bound
1	1.00	8.25
2	8.25	16.50
3	16.50	24.75
4	24.75	33.00

2. Perform representative discrete data (or word) c of length *w* generation from the approximate data using Gaussian-distributed, approximately equi-probable symbols with number of symbols *a*, which can be determined from the area under a $\mathrm{N}(0,1)$ Gaussian curve.

Given β_i a breakpoint on a z-normalized y-axis giving an approximately equi-probable region. Therefore, β_i can be listed as follows:

$$\bar{c}_i = \begin{cases} -\infty, & i = 0 \\ \frac{1}{a}, & 0 < i < 1 \end{cases}$$

Lin et al. [16] provided a lookup table for breakpoints under Gaussian distribution with *a* ranging from 3 to 10. Alternatively, breakpoints corresponding to any number of symbol *a* can be found from z-transformed values giving probabilities from $1/a \, \text{to} \, (a-1)/a$

Table 3.1 A look up table providing list of breakpoints corresponding to number of breakpoints required. [16]

a β_i	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

Table 3.1 is easily obtained from normal distribution z-scores which are from an inverse of a normal cumulative density function (cdf) given particular mean and standard deviation with probability ranging from $1/a \, \text{to} \, (a-1)/a$

Example

In order to better understand the whole process of data representation using SAX, the following section provides an example of how a data series is represented into an approximate series.

1. Perform z-normalization to the original data to convert the data to have mean $(\bar{x}) = 0$ and standard deviation $(SD) = 1$

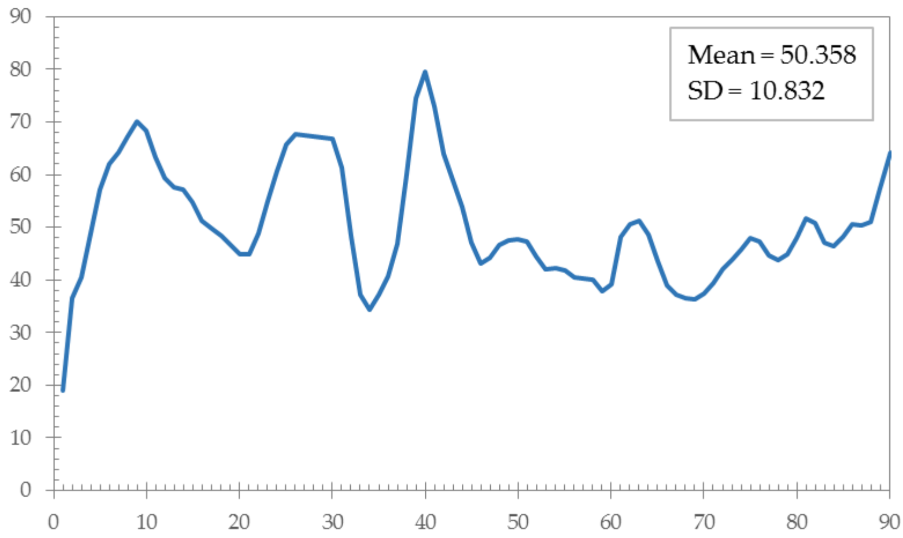


Figure 3.1 Original data series

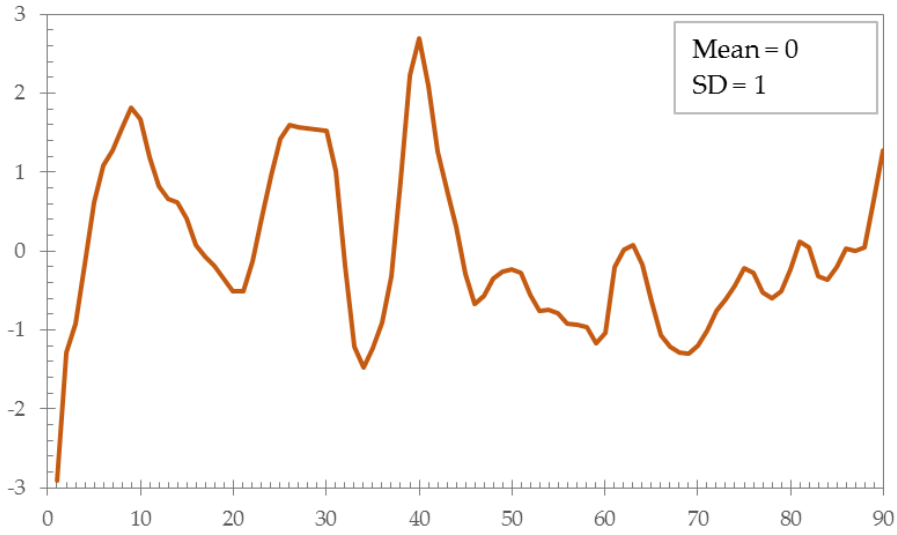


Figure 3.2 z-transformed data series

2. Perform PAA on the original data of length $n = 90$ to the approximate data of length $w = 3$

$$\frac{n}{w} = \frac{90}{3} \in \mathbb{N}$$

Table 3.2 Data range from PAA discretization

ID	Lower Bound	Upper Bound	Average Value
1	1	3	-1.6972
2	4	6	0.5169
3	7	9	1.5505
4	10	12	1.2236
5	13	15	0.5645
6	16	18	-0.0584
7	19	21	-0.4572
8	22	24	0.4163
9	25	27	1.5291
10	28	30	1.5375
11	31	33	-0.1283
12	34	36	-1.1994
13	37	39	0.9319
14	40	42	2.0148
15	43	45	0.2667
16	46	48	-0.5224
17	49	51	-0.2601
18	52	54	-0.6872
19	55	57	-0.8813
20	58	60	-1.0553

Table 3.3 Data range from PAA discretization (continued)

ID	Lower Bound	Upper Bound	Average Value
21	61	63	-0.0376
22	64	66	-0.6228
23	67	69	-1.2602
24	70	72	-0.9857
25	73	75	-0.4178
26	76	78	-0.4649
27	79	81	-0.2065
28	82	84	-0.2128
29	85	87	-0.0569
30	88	90	0.6603

3. Perform representative discrete data (or word) C of length w generation from the approximate data using Gaussian-distributed, approximately equi-probable symbols with number of symbols a , which can be determined from the area under a $\mathrm{N}(0,1)$ Gaussian curve.

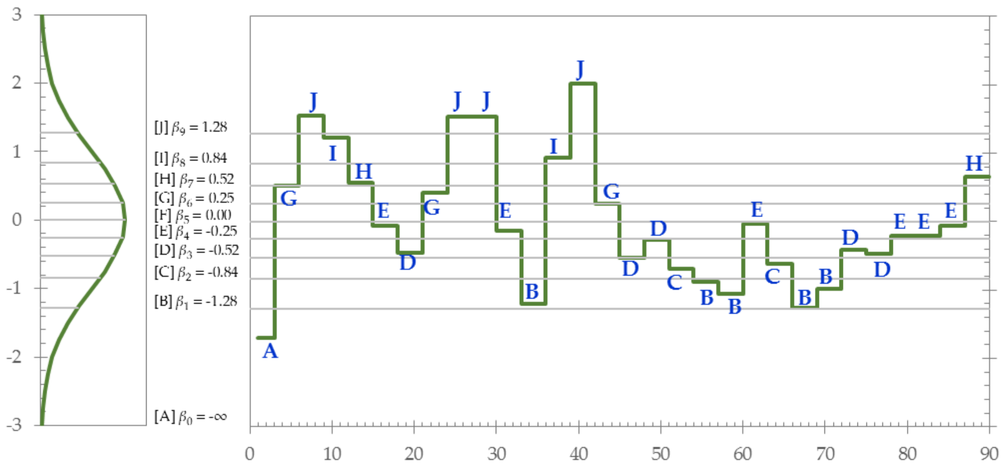


Figure 3.3 Time series discrete data representation using SAX.

Figure 3.3 shows how a z-transformed time series dataset having 90 data points is represented as a character string containing 30 characters. The characters are a combination of A to J representing 10 equi-probable regions. In this example, the final representative discrete data C is AGJIHEDGJJEBIJGDDCBBECBBDDEEEH.

1.1.1 Metric

A metric is a function returning a distance between two instances of a set X where

$$m : X \times X \rightarrow \mathbb{R}$$

For all instance a, b, c, this function must conform to the following criteria [17, 18]

- Non-negativity: $m(a,b) \geq 0$
- Identity: $m(a,b) = 0 \leftrightarrow a = b$
- Symmetry: $m(a,b) = m(b,a)$
- Triangular Inequality: $m(a,b) \leq m(a,c) + m(b,c)$

Metric may be considered as a special distance measurement conforming to the aforementioned properties. In other words, not all distance measurement is a metric. These properties are of benefits to motif discovery. Section 3.2 gives a thorough discussion on how MK algorithm utilizes triangular inequality to help reduce number of distance calculations in most cases.

1.1.2 Distance

A distance is a numerical represented how much 2 given instances are far apart under a particular distance function. A distance is used to indicate similarity (or dissimilarity) of two given instances. There are many distance functions in mathematic literatures. Therefore, only distance functions or calculations used in this study are discussed.

1.1.2.1 Euclidean Distance (ED)

Given

$$x=(x_1,x_2,x_3,...,x_i,\ldots,x_n)\in\mathbb{R}^n$$

and

$$y=(y_1,y_2,y_3,\ldots,y_i,...,y_n)\in\mathbb{R}^n$$

Euclidean distance of n-dimension of x and y is defined as

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Euclidean distance is also a metric as it satisfies all metric criteria.

1.1.2.2 Hamming Distance

From a geometrical model in error detection and error correction codes, Hamming [19] provides a definition of his distance (which is later called Hamming distance) as a metric giving a distance between 2 equal-length bit strings from number of bit difference. For example, Hamming distances between the following data points and 000 are as follows.

Table 3.4 An example of Hamming distance calculation on a bit string

Data	Bit Representation	Hamming Distance
0	000	$d\left(000,000\right)=0$
1	001	$d\left(000,001\right)=1$
2	010	$d\left(000,010\right)=1$
4	100	$d\left(000,100\right)=1$
7	111	$d\left(000,111\right)=3$

The same idea is applicable to character string comparison. For example, Hamming distance between AAAA and AABB is 2. Hamming distance may be divided by string length to normalize its distance to be between 0 and 1. This normalization may be useful for comparison of resulting Hamming distance from different string length.

1.1.2.3 Levenshtein Distance

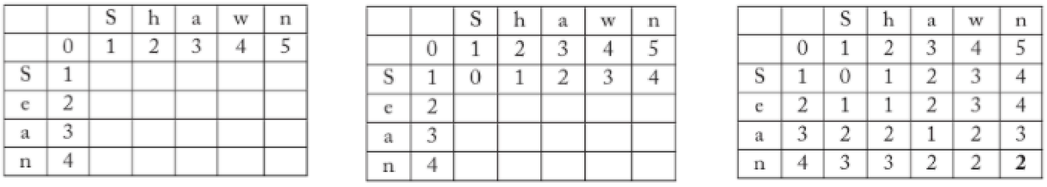
While Hamming distance is similar to substitution distance, Levenshtein distance is similar to edit distance when substitution, insertion, and deletion are also counted. Levenshtein distance may be used to compare between 2 strings having different lengths. Similar to Hamming distance, any change required to convert one string to another string has its associated unit cost (cost is 1). One of the major applications of this distance is in approximation string match when a smaller string is compared with a longer string. Below shows how Levenshtein distance is calculated [20].

Given $x=\left(x_1,x_2,x_3,\ldots,x_n\right)\in\mathbb{S}^n$ and $y=\left(y_1,y_2,y_3,\ldots,y_m\right)\in\mathbb{S}^m$ where $\mathbb{S}$ is a set of symbols, then

$$d\left(x,y\right)=d_{mat}(n,m)$$

$$\text{where}\, d_{mat}(h,k)=\begin{cases} 0, & h=0 \text{ and } k=0 \\ d_{mat}(h,k-1)+1, & h=0 \text{ and } k>0 \\ d_{mat}(h-1,k)+1, & h>0 \text{ and } k=0 \\ d_{mat}(h-1,k-1), & x_h=y_k \\ \min\begin{Bmatrix} d_{mat}(h,k-1)+1 \\ d_{mat}(h-1,k)+1 \\ d_{mat}(h-1,k-1)+1 \end{Bmatrix}, & x_h\neq y_k \end{cases} \quad ; 0\le h\le n \text{ and } 0\le k\le m$$

Therefore, the distance from comparing series x and y can be found at $d(n,m)$



		S	h	a	w	n
	0	1	2	3	4	5
S	1					
e	2					
a	3					
n	4					

		S	h	a	w	n
	0	1	2	3	4	5
S	1	0	1	2	3	4
e	2					
a	3					
n	4					

		S	h	a	w	n
	0	1	2	3	4	5
S	1	0	1	2	3	4
e	2	1	1	2	3	4
a	3	2	2	1	2	3
n	4	3	3	2	2	**2**

(a) Initialization (b) 2nd-row calculation (c) Completion

Figure 3.4 Levenshtein distance calculation example [21]

Referring to Figure 3.4, since S and S are identical in figure (b), the distance is 0. Next comparison is between S and h. Since they are not identical, a cost has to be considered. The cost is $1+\min(0,1,2)$. Therefore, the cost is 1. This calculation is continued until the end of all strings. The real distance of comparing “Shawn” and “Sean” is 2, which is indicated in the lowest right corner of the distance matrix shown in (c).

1.1.3 Nearest Neighbor Search (NNS)

Nearest neighbor search in a typical time series research is a search technique that aims at returning the closest proximity between the subsequence of interest which is used as a *query* and a subsequence extracted from a different series in question. The proximity between the two subsequences typically refer to the use of distance function which gives a numerical distance figure indicating how close the two subsequences are. In a typical meaning of distance, the most similar subsequences are the subsequences having smallest distance compared with the other subsequence pairs.

In this study, NNS is used to find the closest match or matches within a predefined maximum distance of the well log section of interest in other wells. In this search criteria, the search may give more than one match depending on the tightness of the predefined distance. The distance function used in the search is Euclidean distance. In order to be able to search for the matches, a sliding-window method, which slides one data point at a time, is used to form a prospective section of the same length as the length of the interest section, which is the length of window size *w*. The prospective section and the section of interest is then undergone distance calculation. This calculated distance is recorded at index *i* which represents a distance of such comparison between a section from *i* to *i + w* (forming a section of length *w*) with the section of interest of also length *w*. Once all the indexes are associated with distance figures, only sections having the distance figures under the predefined distance cut off are chosen. Since only sections having the distance figures under the predefined distance cut off are chosen, Early abandoning Euclidean distance, a variance of Euclidean distance that stops the distance calculation as soon as the cumulative distance exceeds the square of the predefined distance cut off and the distance is returned as *infinity*, may be used, thereby saving most of the calculation effort in the very much dissimilar sections.

In this study, similarity measurement techniques used in this pattern matching process are varied but it is possible to categorize into 2 groups: exact similarity and approximate similarity. Exact similarity is a similarity that is assessed on the original data series while approximate similarity is a similarity that is assessed on the approximate series. In real application, approximate similarity may serve as a way to smear off some local variations.

3.2 Pattern (Motif) Discovery using MK Algorithm

A motif or a time series motif is a pair of time series or a pair of time series subsequences extracted from longer time series having similar pattern to each other. Motif discovery is, typically, an operation where at least a search strategy is applied on a time series in order to recover intrinsic recurring pairs. Recurring patterns are of interest because they may reveal inherently new and useful information of the time series being studied.

There are many motif discovery techniques which can be categorized by exactness of the similarity measurement technique used in motif discovery algorithms. Simplest exact motif discovery algorithm is brute force algorithm. Brute force motif discovery algorithm involves extracting a subsequence length *s* from the whole time series length *n* $(s < n \text{ or } s \ll n)$. Then the subsequence is compared with a subsequence $s' \; (s' = s < n)$ extracted from original series on a sliding-window basis. Although brute force method gives exact motif discovery, the algorithm is quadratic, making it less applicable to large dataset since the time requires to work as such is intractable. For example, finding a motif pair of length 1,024 from 100,000 objects using brute force method takes 12.7 hours to complete [15]. This is considered to be less attractive approach for a real application.

Apart from approximate motif discovery techniques, Mueen et al. [15] presented a method to perform exact motif discovery with reduced operating time named *MK algorithm* (Motif Kymatology algorithm). The method is reported to reduce the time used to perform a motif pair discovery (of length 1,024 from 100,000 objects) to 12.4 minutes, which account for 61.5 times reduction from conventional brute force method). Therefore, MK algorithm shows a promising performance for applying exact motif discovery on a real application.

The reason underlining a huge reduction of operating time required to find an exact motif pair in a large dataset is due to an application of a pruning technique to avoid unnecessary distance comparison based on triangular inequality property of Euclidean distance metric. Even though this reduced number of comparisons results in reduction in operating time in most cases, its worst case scenario is still the same scenario as that of conventional brute force technique, which is a complexity of $O\left(n^2\right)$ where *n* is number of objects or data points. The worst case scenario only happens when the motif pair distance is larger than any lower bound.

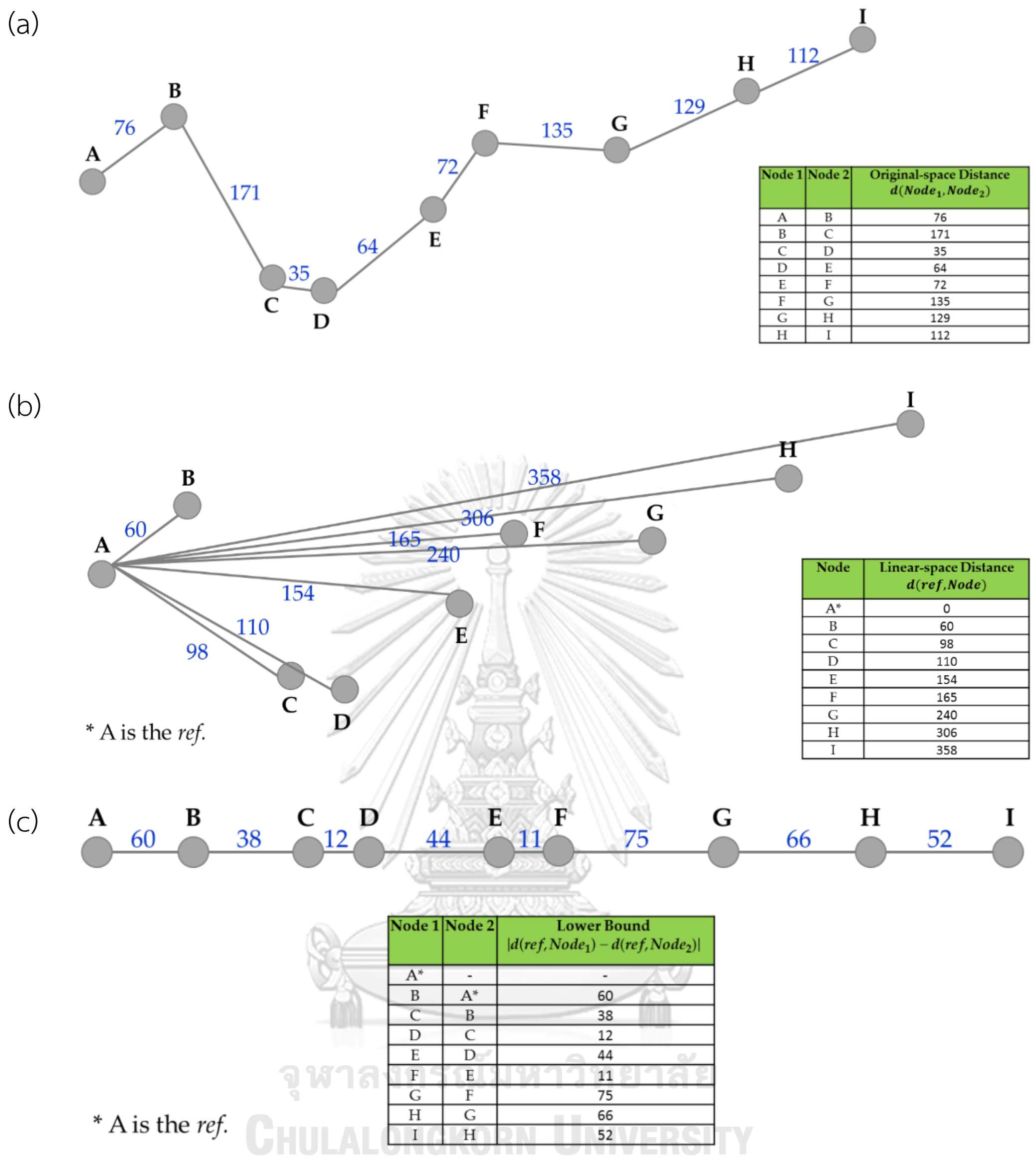
3.2.1 Intuition behind MK Algorithm

As previously mentioned, MK algorithm is able to prune off unnecessary search space from exploitation of triangular inequality property of Euclidean distance metric. Mueen et al. [15] observed that if two objects are close in the original space (small distance), the objects must also be close in the linear ordering space. It must be aware that the reverse is not true; two objects that are close in the arbitrary linear ordering may be very far (large distance) in the original space. Linear ordering space provides a heuristic and useful information for motif discovery.

In order to construct a linear ordering space, a reference point is chosen and distances between other points to the reference point are calculated. As the name of the space implies, the distances to the reference point are considered as distances in the linear space. This distance in linear space is a lower bound for each pair of two adjacent points.

Then a search operation can be started by updating a recorded smallest distance value of the search operation called *best-so-far*. Initially, *best-so-far* is set as infinity in order to reflex the start of the search operation. Starting from the first value of the series, an original-space distance is required to be calculated only if the two adjacent points have the linear-space distance (lower bound) that is less than the current *best-so-far*. Moreover, whenever two adjacent points have its original-space distance between the pair that is less than current *best-so-far*, *best-so-far* is updated with this distance and the distance of the pair on the linear space is also recorded with same distance. If the calculated original-space distance of the pair, however, is not less than current *best-so-far*, only the linear-space distance of the pair is recorded using the pair’s original-space distance previously calculated.

Once the search operation is done, *best-so-far* will already have the smallest original-space distance. Before the pair providing its original-space distance that of *best-so-far* is considered as a motif, another search operation on linear space is required. The pair that is also a motif must be within the same search window when the search window size is of the *best-so-far*. In other words, only the pair having its lower bound lowers than the recorded *best-so-far* is considered a motif pair. Figure 3.5 and Figure 3.6 give a better illustration on how MK algorithm intuition works by using a node-based example. Figure 3.6 shows how *best-so-far* indicating a running, original-space distance is obtained without having to directly calculate original-space distances of all pairs. From example given in the figure, only 4 original-space distance calculation is required.



Node 1	Node 2	Original-space Distance $d(Node_1, Node_2)$
A	B	76
B	C	171
C	D	35
D	E	64
E	F	72
F	G	135
G	H	129
H	I	112

Node	Linear-space Distance $d(ref, Node)$
A*	0
B	60
C	98
D	110
E	154
F	165
G	240
H	306
I	358

* A is the *ref*.

Node 1	Node 2	Lower Bound $\|d(ref, Node_1) - d(ref, Node_2)\|$
A*	-	-
B	A*	60
C	B	38
D	C	12
E	D	44
F	E	11
G	F	75
H	G	66
I	H	52

* A is the *ref*.

Figure 3.5 Original-space distances, linear-space distances, and lower bounds

From Figure 3.5, distances in Figure 3.5(a) are initially unknown to the algorithm as they are calculated only when they are required. The distances in (a) is shown for brevity. Figure (b) is a figure showing how linear space from distances to a reference node is constructed. Figure (c) shows final lower bounds from linear-space distance differences between two adjacent nodes.

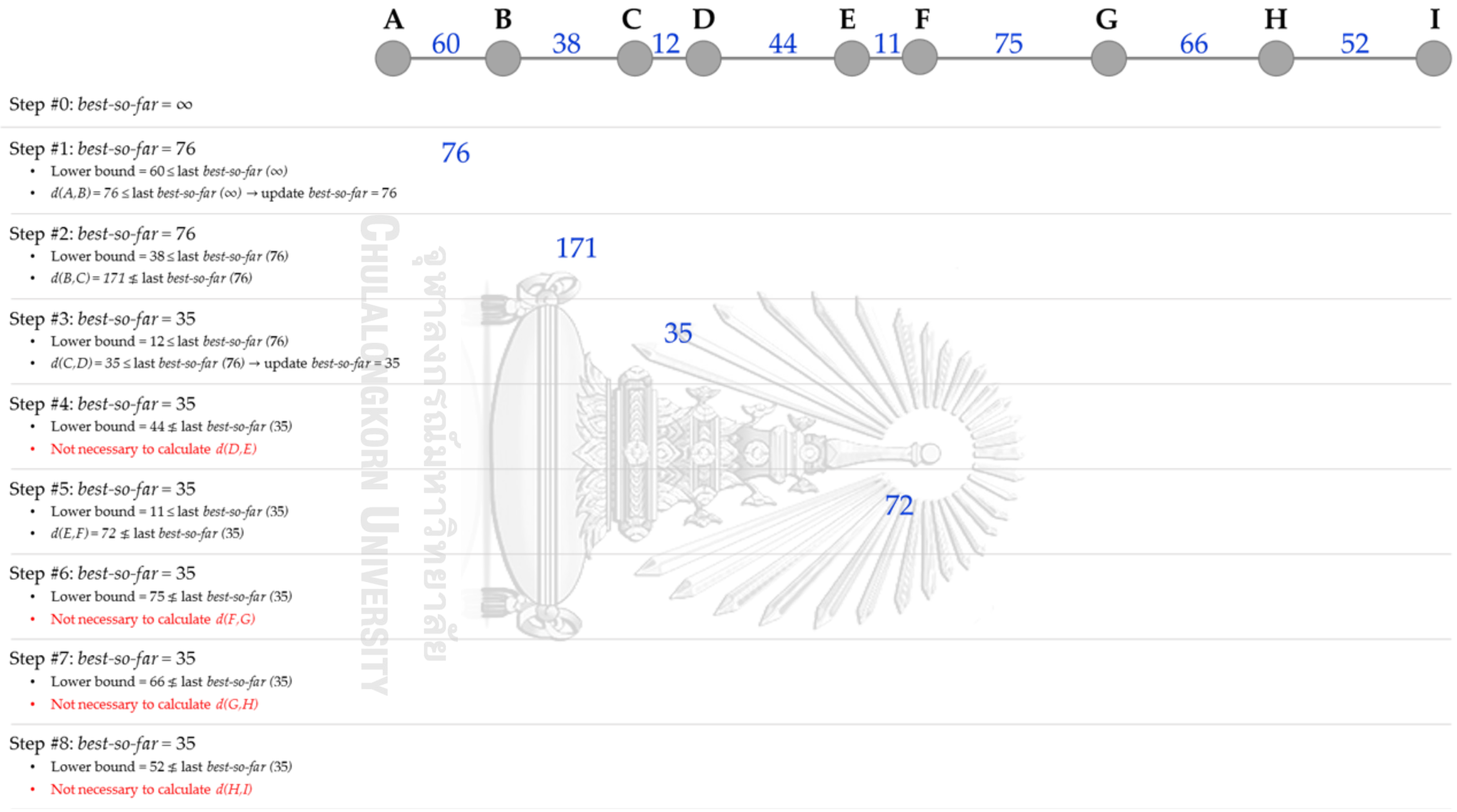

A
B
C
D
E
F
G
H
I
60
38
12
44
11
75
66
52
Step #0: best-so-far = ∞
Step #1: best-so-far = 76
Lower bound = 60 ≤ last best-so-far (∞)
d(A,B) = 76 ≤ last best-so-far (∞) → update best-so-far = 76
76
Step #2: best-so-far = 76
Lower bound = 38 ≤ last best-so-far (76)
d(B,C) = 171 ≰ last best-so-far (76)
171
Step #3: best-so-far = 35
Lower bound = 12 ≤ last best-so-far (76)
d(C,D) = 35 ≤ last best-so-far (76) → update best-so-far = 35
35
Step #4: best-so-far = 35
Lower bound = 44 ≰ last best-so-far (35)
Not necessary to calculate d(D,E)
Step #5: best-so-far = 35
Lower bound = 11 ≤ last best-so-far (35)
d(E,F) = 72 ≰ last best-so-far (35)
72
Step #6: best-so-far = 35
Lower bound = 75 ≰ last best-so-far (35)
Not necessary to calculate d(F,G)
Step #7: best-so-far = 35
Lower bound = 66 ≰ last best-so-far (35)
Not necessary to calculate d(G,H)
Step #8: best-so-far = 35
Lower bound = 52 ≰ last best-so-far (35)
Not necessary to calculate d(H,I)

Figure 3.6 Step-by-step distance comparisons and best-so-far updates.

3.2.2 MK Algorithm Formal Statements

According to intuition behind the algorithm previously discussed, it is essential to find a condition when comparisons can be stopped early while a motif pair can be discovered. Therefore, the condition is a condition where it is certain that further comparisons will give a pair that cannot be a motif. The condition is that when a lower bound is larger than the current *best-so-far*. If a pair falls in this condition, the original space distance of the pair will not necessarily be calculated as the pair cannot be a motif pair. Otherwise, its original-space distance has to be calculated as it is a potential motif. This condition can be shown by using triangular inequality property of a distance metric.

Let *ref* be a reference time series in *x* where $x = (x_1, x_2, x_3, \ldots, x_n)$

Let $\{x_i, x_j\}$ be the pair that its original-space distance is not going to be calculated.

Triangular inequality $d(p,q) \le d(p,r) + d(q,r)$ gives

$$\begin{aligned} d(ref, x_i) &\le d(ref, x_j) + d(x_i, x_j) \\ \left| d(ref, x_i) - d(ref, x_j) \right| &\le d(x_i, x_j) \end{aligned}$$

Therefore, $\left| d(ref, x_i) - d(ref, x_j) \right|$ is essentially the lower bound for $d(x_i, x_j)$ and $d(x_i, x_j)$ can be avoided from calculation if $\left| d(ref, x_i) - d(ref, x_j) \right| >$ *best-so-far* since it is certain that $\{x_i, x_j\}$ is no longer a motif pair. Moreover, a lower bound can easily be calculated as it is a cheap subtraction operation.

Generally, *ref* can be any series whether it is inside or outside of *x.* However, it is preferred to use *ref* from a subsequence of *x* since it would be easier to prevent $\{ref, ref\}$ from happening by assigning the original-space distance to be infinity as the pair is not valid. Up until this point, the trick of using internal reference series only reduces number of distance computations, but it does not reduce the search space as it is unknown that when the stop criteria of the search space would be since distances between the reference series and the other series still spans the entire dataset. In other words, there is still no indexing which the algorithm can be checked if the search operation can be stopped. Sorting indices to linear-ordering space is a better option than sorting the lower bounds. Both methods help guide early search stop but with different complexity.

Lemma 3.1

Given x_I a set of indices to linear-ordering distances
Let *offset* be a positive integer
If $x_{I(j+offset)} - x_{I(j)} >$ *best-so-far* for all $1 \le j \le n - offset$ and $offset > 0$ then
$x_{I(j+w)} - x_{I(j)} >$ *best-so-far* for all $1 \le j \le n - w$ and $w > offset$

Lemma 3.1 indicates that if a lower bound (a difference between 2 distances in linear ordering space) referenced by $x_{I(j+offset)}$ and $x_{I(j)}$ is greater than a running minimum distance in original space *best-so-far*, a lower bound referenced by $x_{I(j+w)}$ and $x_{I(j)}$ is also greater than *best-so-far* since *w* is greater than *offset.* This is due to the fact that x_I is a set providing indices to smallest-to-biggest distances between one series and another reference series. Therefore, a lower bound at *w > offset* is certain to be greater than *best-so-far* if the lower bound at *offset* is already greater than *best-so-far*.

Lemma 3.2

If $offset = 1, 2, 3, \ldots n-1$ and $j = 1, 2, 3, \ldots, n - offset$ Then $\left\{x_{I(j)}, x_{I(j+offset)}\right\}$ generates all the possible pairs.

Lemma 3.2 states exactness of the algorithm [22] as it ensures search operation to be performed in all possible pairs. Interested readers may find proofs to the lemmas in the original literature [22].

In a very large time series, multiple references may help tighten lower bound. However, not all references are equivalently good as references. When multiple references are utilized, only the biggest lower bound is used for search operation as the biggest lower bound is likely to give the earliest early search stop and comparison rejection. To choose the biggest lower bound, the linear-ordering distance set with largest standard deviation is chosen as the larger the deviation is the larger the lower bound. ***gap*** is used to allow indices between the prospective motif pair to be far apart. For example, sometimes it is unlikely that the prospective motif pair to be closed more than a certain number of data points. In this case, that data points are represented by a ***gap*** . Figure 3.7 shows a flowchart of MK algorithm.

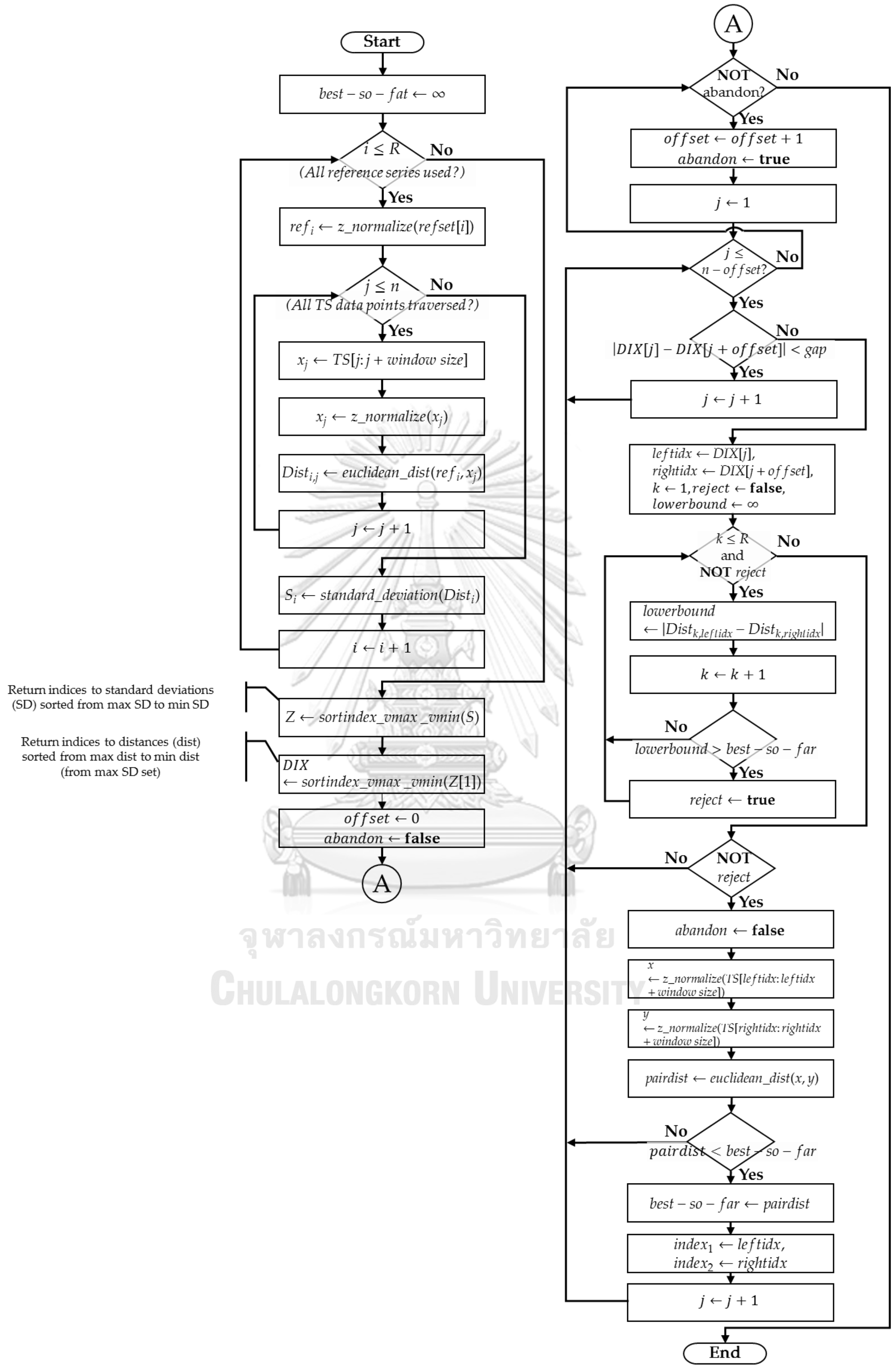

Start
best − so − fat ← ∞
i ≤ R
(All reference series used?)
No
Yes
ref_i ← z_normalize(refset[i])
j ≤ n
(All TS data points traversed?)
No
Yes
x_j ← TS[j:j + window size]
x_j ← z_normalize(x_j)
Dist_i,j ← euclidean_dist(ref_i, x_j)
j ← j + 1
S_i ← standard_deviation(Dist_i)
i ← i + 1
Return indices to standard deviations (SD) sorted from max SD to min SD
Z ← sortindex_vmax_vmin(S)
Return indices to distances (dist) sorted from max dist to min dist (from max SD set)
DIX ← sortindex_vmax_vmin(Z[1])
offset ← 0
abandon ← false
A
A
NOT abandon?
No
Yes
offset ← offset + 1
abandon ← true
j ← 1
j ≤ n − offset?
No
Yes
|DIX[j] − DIX[j + offset]| < gap
No
Yes
j ← j + 1
leftidx ← DIX[j],
rightidx ← DIX[j + offset],
k ← 1, reject ← false,
lowerbound ← ∞
k ≤ R and NOT reject
No
Yes
lowerbound ← |Dist_k,leftidx − Dist_k,rightidx|
k ← k + 1
No
lowerbound > best − so − far
Yes
reject ← true
No
NOT reject
Yes
abandon ← false
x ← z_normalize(TS[leftidx: leftidx + window size])
y ← z_normalize(TS[rightidx: rightidx + window size])
pairdist ← euclidean_dist(x, y)
No
pairdist < best − so − far
Yes
best − so − far ← pairdist
index_1 ← leftidx,
index_2 ← rightidx
j ← j + 1
End

Figure 3.7 MK Algorithm (based on [15])

CHAPTER 4

METHODOLOGY

In this study, there are two approaches to well log correlation being presented. Pattern matching is the technique to be used when a portion of well log signal around section of interest in known and there is a need to find the similar section(s) in other nearby wells while Pattern discovery is the technique to be used when only section length is known. The detailed methodology is as follows.

4.1 Well Log Data Preparation

Unless specified otherwise, there are three gamma ray logs used throughout this study. The well logs were obtained from three wells from SPIVEY - GRABS – BASIL field, Kingman County, Kansas, USA. The source of this material is the Kansas Geological Survey website at http://www.kgs.ku.edu/. All Rights Reserved.

Table 4.1 Well log data source

Well Name	Well Depth From – To (ft)	Logging Step (ft/data point)	Usage in this study	
			Depth From - To (ft)	Purpose
TJADEN A #6-13	302.0 – 4,411.5	0.5	3,000.0 - 4,370.0 (2,741 data points)	Well Y, Sections extracted as model patterns
TJADEN A #7-13	350.0 – 4,409.5	0.5	3,000.0 - 4,367.0 (2,735 data points)	Well Z
#3 TJADEN "C"	0.0 – 4,408.5	0.5	3,000.0 - 4,370.0 (2,741 data points)	Well X

Unless specified otherwise, all data windows used are *zero-based* data windows. This means the first index is referred to as index 0 (*not* index 1). Therefore, data window whose first index is at index 0 is also referred to as data window 0 or Window 0.

4.2 Pattern Matching

The following steps are used in for performing well-to-well log correlation using pattern matching approach.

1 Given a model pattern, find top five other similar patterns (data windows) of the same length from the different than the well the model pattern is chosen using Nearest Neighbor Search (NNS). Figure 4.1 shows that there are five data windows promoted by NNS given the model pattern of length 130 indexes shown in the upper left section of the figure. The five data windows

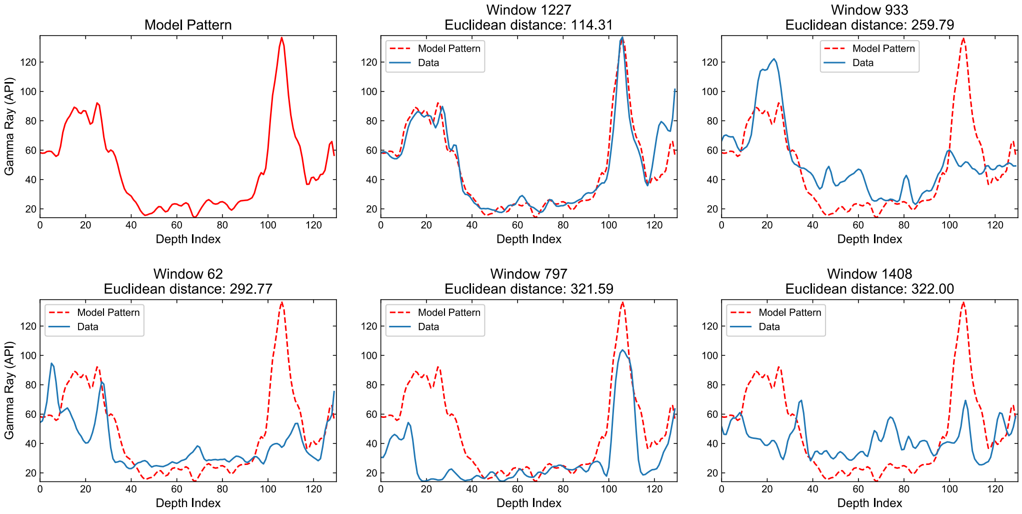


Figure 4.1 Five smallest distances promoted by conventional NNS

2 Perform multi-resolution analysis and summarize the best correlation result in a probabilistic manner. Typically, maximum number of PAA blocks does not exceed *1/2 × number of data points.* Therefore, number of PAA blocks for any section length (window length or window size) having 120 and 130 data points (indexes) are from 5 to 50 blocks with an increment of five PAA blocks each interval (5, 10, 15, 20, 25, 30, 35, 40, 45, and 50) and number of PAA blocks for

any section length having 200, 220, and 225 data points (indexes) are from 10 to 100 with an increment of 10 PAA blocks each interval (10, 20, 30, 40, 50, 60, 70, 80, 90, and 100). Number of SAX sections are fixed at 4, 6, 8, and 10 sections.

1) Perform Piecewise Aggregate Approximation (PAA). This is a discretization over depth (x-axis). Figure 4.2 to 4.5 provides examples of approximate signals generated by PAA using number of PAA blocks of 15, 25, 35, and 45 respectively. Since this model pattern and all data windows are of length 130, multi-resolution analysis is performed based on PAA of 5 – 50 blocks with increment of five blocks each interval.

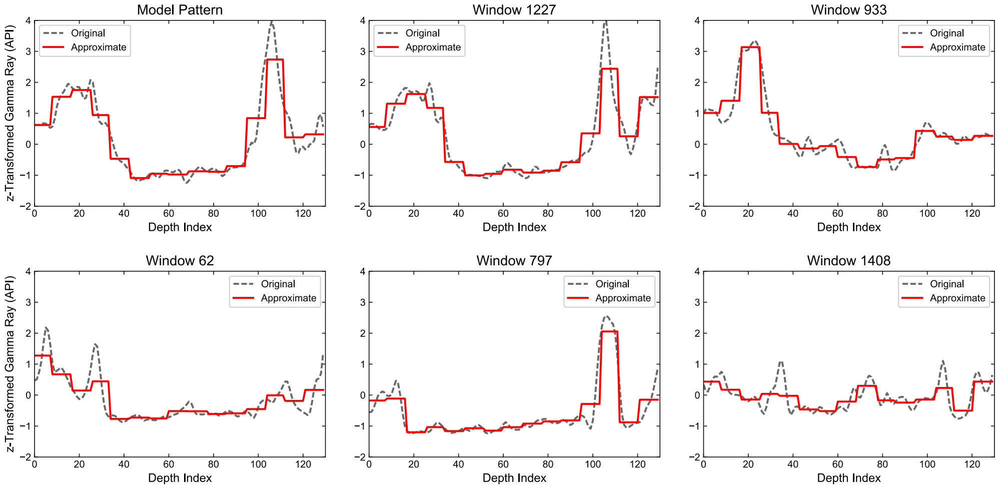


Figure 4.2 Approximate signals based on 15 PAA blocks

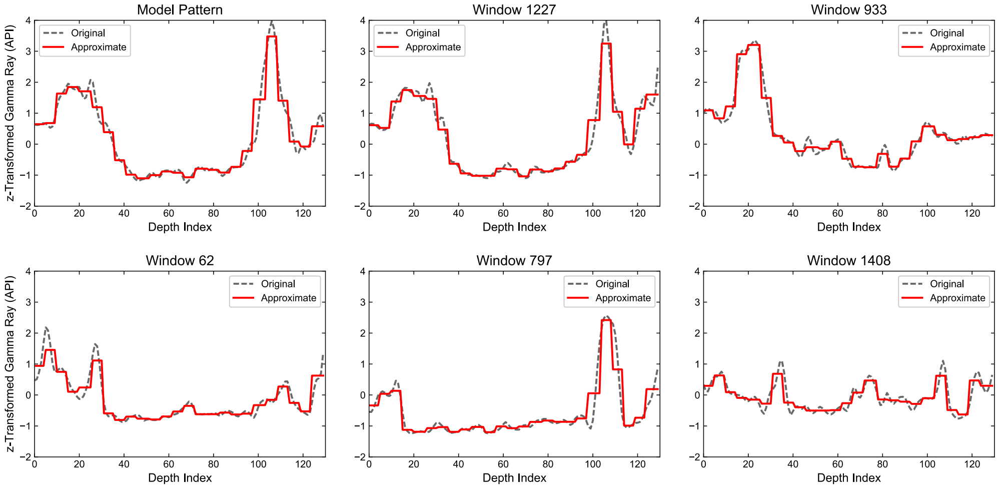


Figure 4.3 Approximate signals based on 25 PAA blocks

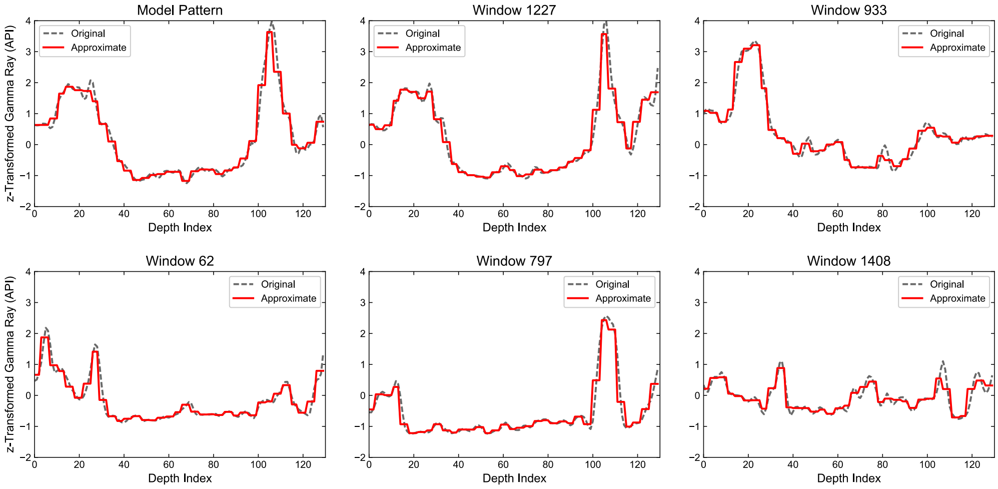


Figure 4.4 Approximate signals based on 35 PAA blocks

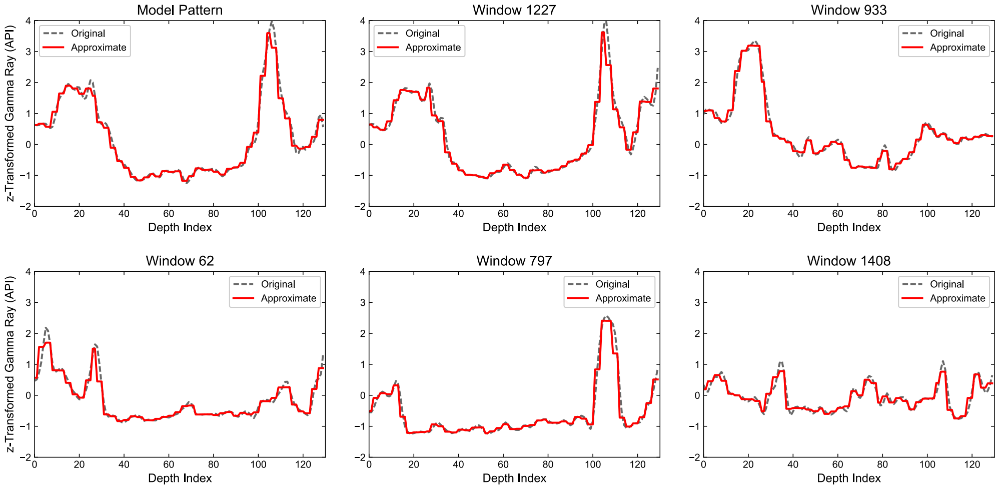


Figure 4.5 Approximate signals based on 45 PAA blocks

2) Perform symbolic representation by discretizing log value (y-axis) with 4, 6, 8, and 10 SAX sections where each section has equal interval probability. For example, each section in four SAX sections has 0.25 probability while each section in 10 SAX sections has 0.10 probability. Figure 4.6 provides an example of symbolic representation over the model pattern. Note than symbols A to D (A, B, C, and D) are used to represent four SAX sections while symbols A to J (A, B, C, D, E, F, G, H, I, and J) are used to represent 10 SAX sections. This representation is performed for all number of PAA blocks.

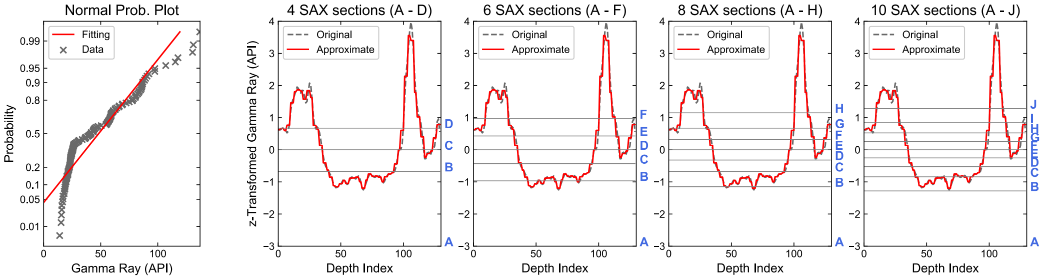


Figure 4.6 Symbolic representation as part of SAX representation on the model pattern.

3) Perform discrete series similarity measurement using Hamming distance and Levenshtein distance. For each distance measurement method used, there are 40 measurement for each data window and there are 200 measurement for all five data windows. Therefore, there are 400 measurement for each model pattern. Figure 4.7 provides an example of the results of all similarity measurement.

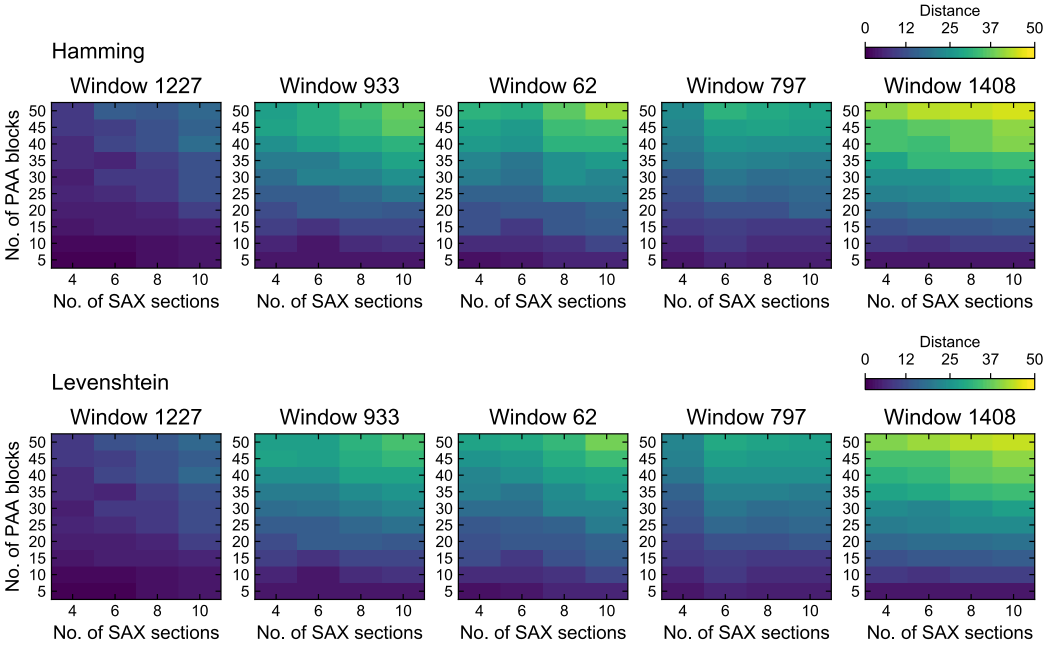


Figure 4.7 Comparisons of distance measured by Hamming distance and Levenshtein distance

4) Perform distance ranking for each resolution (each combination of number of PAA blocks and number of SAX sections). The first rank of the distance ranking is considered as the best match to the given model

pattern. Figure 4.8 shows two color maps representing the best match given by any particular resolution. Data windows previously promoted by conventional NNS are represented by different colors.

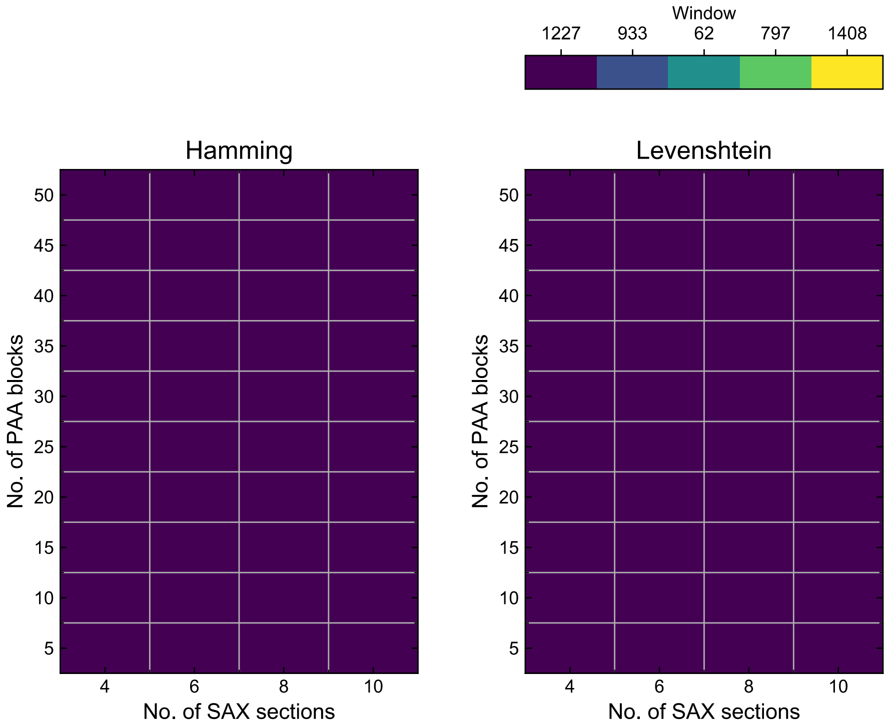


Figure 4.8 The best match data window promoted by multi-resolution analysis

3 Observe and make necessary judgement on the results. For example, Figure 4.8 suggests that Window 1227 is the best match under multi-resolution analysis from both Hamming distance and Levenshtein distance given the fact that its probability of being the best match (first-rank distance) of 1.000 (100.0%).

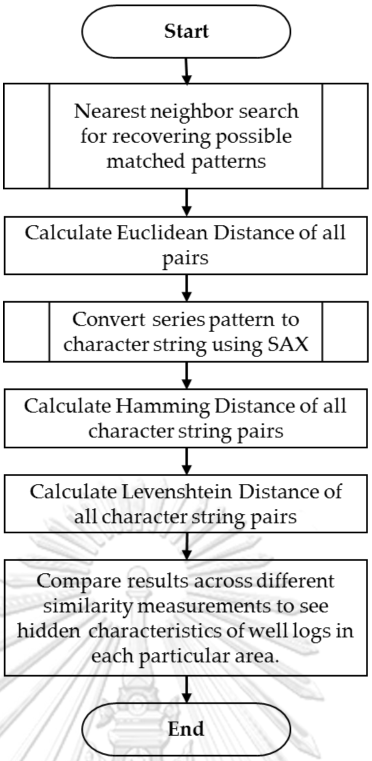

Start
Nearest neighbor search for recovering possible matched patterns
Calculate Euclidean Distance of all pairs
Convert series pattern to character string using SAX
Calculate Hamming Distance of all character string pairs
Calculate Levenshtein Distance of all character string pairs
Compare results across different similarity measurements to see hidden characteristics of well logs in each particular area.
End

Figure 4.9 Pattern Matching process overview workflow

4.3 Pattern Discovery

1 Select two well logs of the same log type to be correlated and connect them together to form an artificial well log. Note the breakpoint which is the last index of the first well log. Figure 4.10 shows an artificial which is created by concatenating Well Y (2,741 indexes) and Well X (2,741 indexes) with index 2,741 served as a breakpoint between the two well logs.

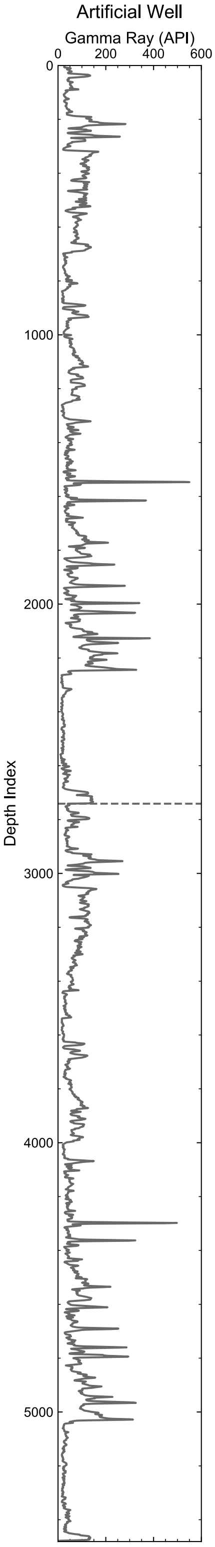


Figure 4.10 An artificial well log from a concatenation of Well Y (top: above the breakpoint) and Well X (bottom: below the breakpoint)

2 Unless specified otherwise, standard deviation (SD) spans of -5.0 to 5.0 with incremental of 0.5 each SD interval are chosen in this study based on the gamma ray responses of the well logs chosen. These SD intervals will be used in heuristic reference series selection to cover the entire log value response. This is a modified step from the original MK provided by Mueen et al. [15].

3 provides the list of all SD intervals and their value range. The table also includes key statistical values of the artificial well log. In this case, the artificial well log is in the range of -1.5 SD to the maximum possible value at 5.0 SD. Figure 4.11 shows the artificial well log signal with SD interval boundaries. As can be seen in the grey-highlighted area in the figure, some portions of the signals are discarded as they are outside of the predefined SD intervals.

4 Specify a section length for correlation. In this demonstration, a section length of 120 indexes is used.

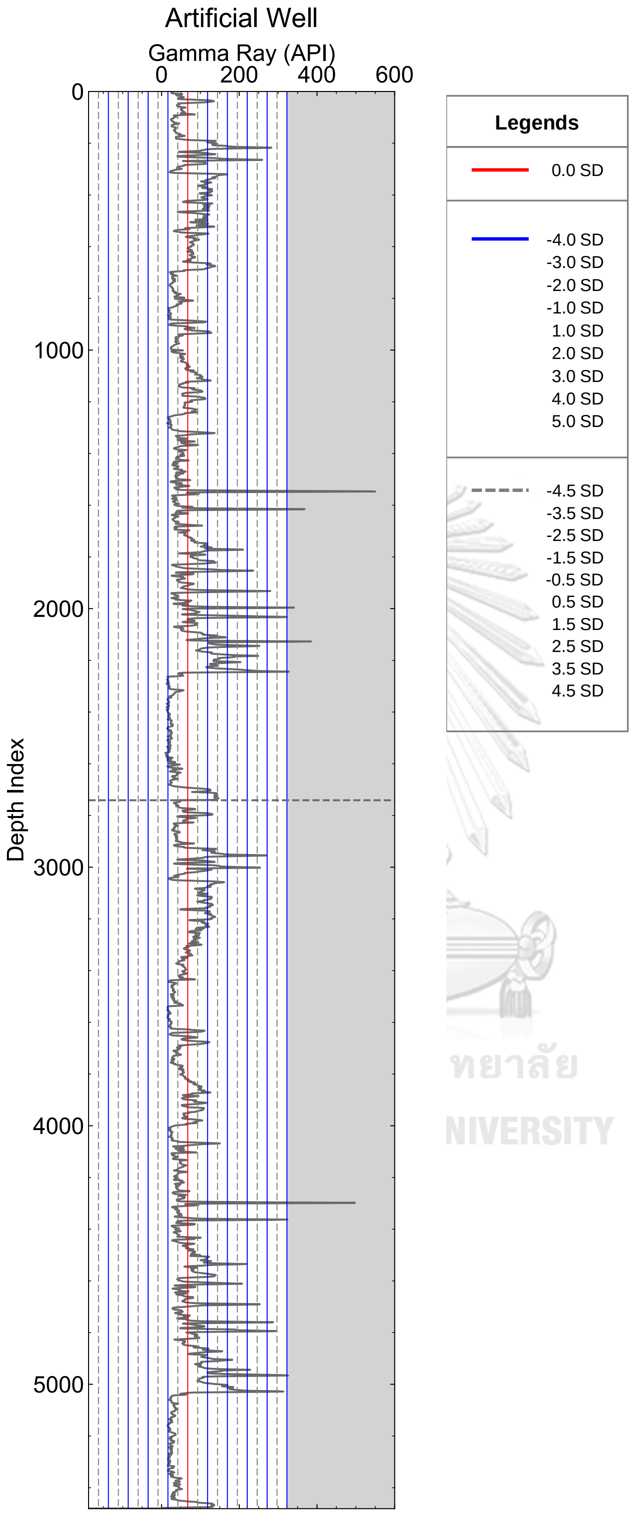


Figure 4.11 The artificial well log with SD boundaries

Table 4.2 Well log value from -5.0 SD to 5.0 SD span in 0.5 SD interval span

Standard Deviation		Gamma Ray (API)		
Min	Max	Min	Max	
-5.0	-4.5	-187.87	-162.34	No data available
-4.5	-4.0	-162.34	-136.82	
-4.0	-3.5	-136.82	-111.29	
-3.5	-3.0	-111.29	-85.77	
-3.0	-2.5	-85.77	-60.24	
-2.5	-2.0	-60.24	-34.71	
-2.0	-1.5	-34.71	-9.19	
-1.5	-1.0	-9.19	16.34	Data available
-1.0	-0.5	16.34	41.86	
-0.5	0.0	41.86	67.39	
0.0	0.5	67.39	92.91	
0.5	1.0	92.91	118.44	
1.0	1.5	118.44	143.96	
1.5	2.0	143.96	169.49	
2.0	2.5	169.49	195.01	
2.5	3.0	195.01	220.54	
3.0	3.5	220.54	246.06	
3.5	4.0	246.06	271.59	
4.0	4.5	271.59	297.11	
4.5	5.0	297.11	322.64	
5.0	5.5	322.64	348.16	

Mean	67.39	API
SD	51.05	API
Max	549.99	API
Min	5.09	API

5 Once motif pairs (correlation results) are available, observe and make necessary judgement on the results. As shown in Figure 4.12 to 4.14, first to third-rank motif pairs are given as examples. The first-rank motif pair is the pair of Window 98 and Window 2836. The second-rank motif pair is the pair of Window 1205

and Window 3953. The third-rank motif pair is the pair of Window 2212 and Window 4997. If all data windows were rebased to the indexes in Well Y and Well X, the first pair, for example, would be the pair of Window 98 on Well Y and Window 95 on Well X. The reference series for these motif pairs is Window 4300 in the artificial well log and it is one of the 21 possible reference series drawn from samples in the SD spans from -5.0 to 5.0. Figure 4.15 shows the final well-to-well log correlation given these motif pairs. Notice that all data windows are rebased to their respective wells: either Well Y or Well X.

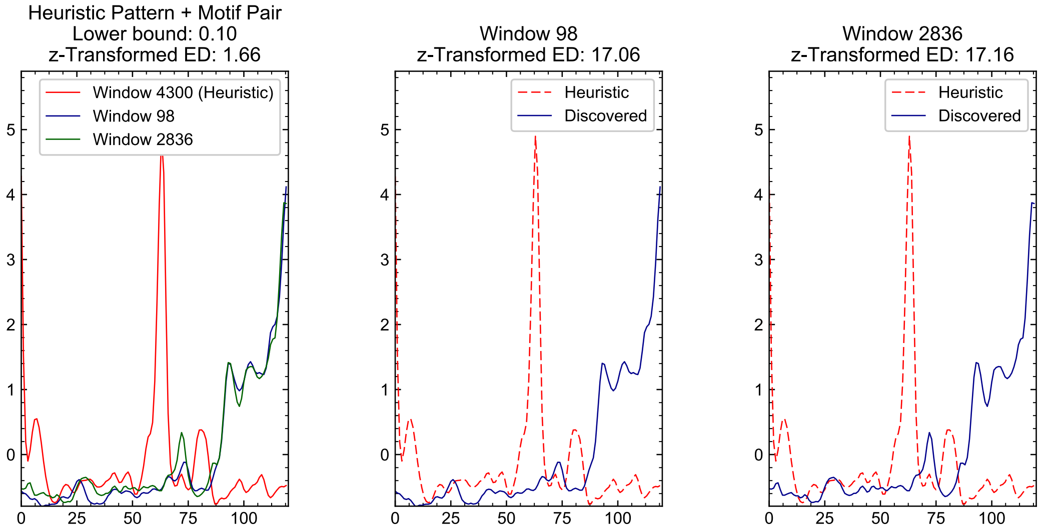


Figure 4.12 First-rank motif pair

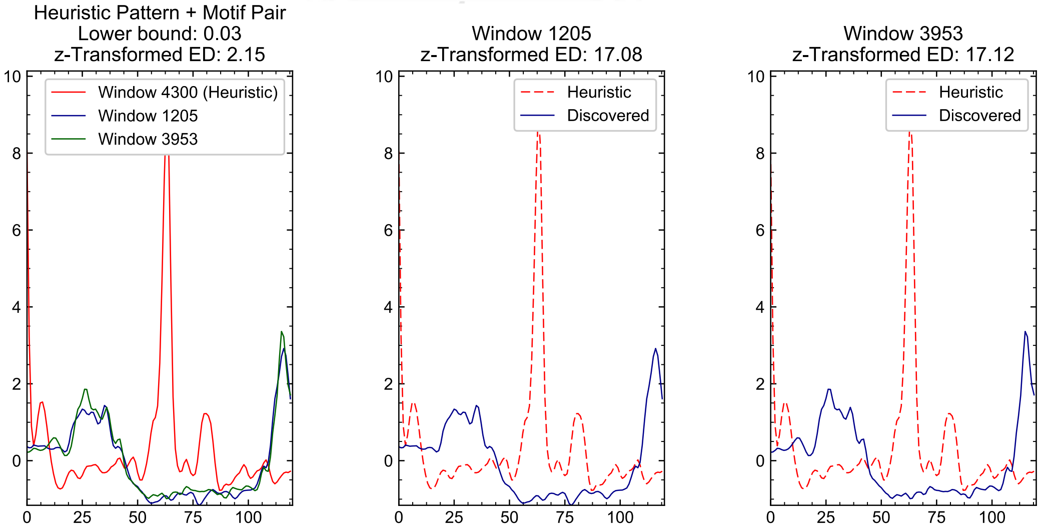


Figure 4.13 Second-rank motif pair

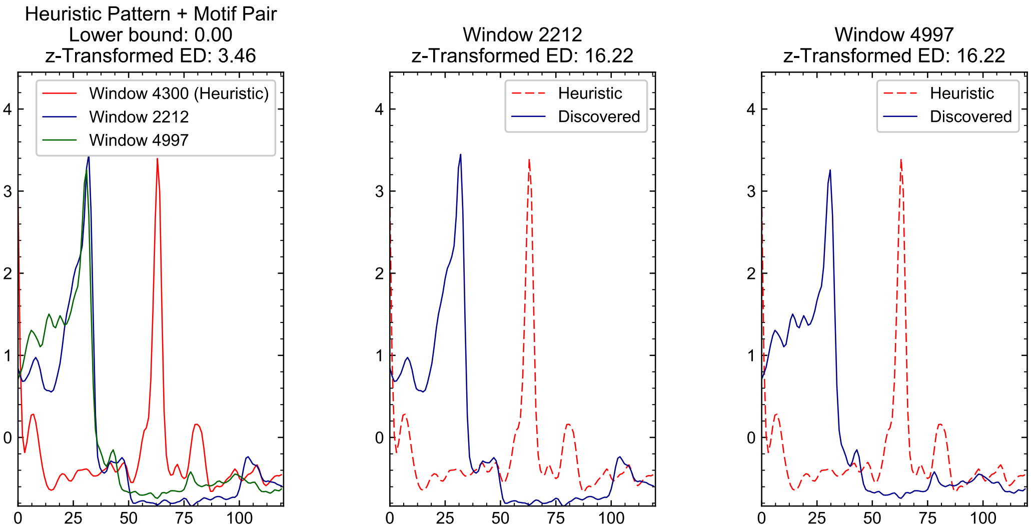


Figure 4.14 Third-rank motif pair

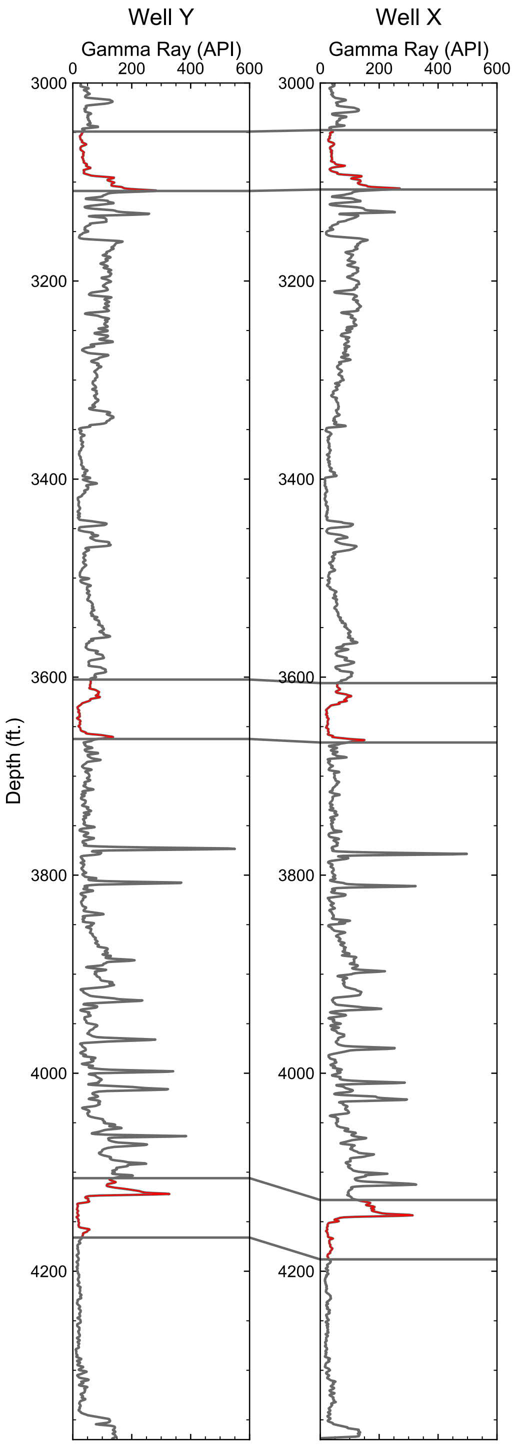


Figure 4.15 Final well-to-well log correlation of the section length of 120 indexes (120 data points or 60 ft) using Window 4300 on the artificial well log as the final reference series

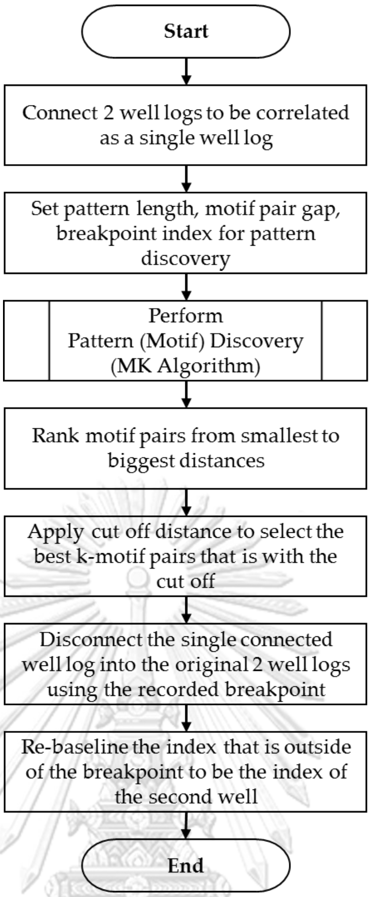

Start
Connect 2 well logs to be correlated as a single well log
Set pattern length, motif pair gap, breakpoint index for pattern discovery
Perform Pattern (Motif) Discovery (MK Algorithm)
Rank motif pairs from smallest to biggest distances
Apply cut off distance to select the best k-motif pairs that is with the cut off
Disconnect the single connected well log into the original 2 well logs using the recorded breakpoint
Re-baseline the index that is outside of the breakpoint to be the index of the second well
End

Figure 4.16 Pattern (motif) discovery using MK algorithm

Figure 4.16 provides a summary of steps for pattern discovery technique when a distance cutoff (maximum allowable distance) is used. In this case, the number of motif pairs from MK will depend on the pairs' distance values and the distance cutoff specified before starting the pattern discovery process.

CHAPTER 5

PATTERN MATCHING

Conventional well log correlation typically is a challenging task requiring experienced engineers, geologists, and petro-physicists and sometimes deep understanding of local geological characteristics. With advanced techniques in data processing, digital well log correlation is still a challenging task.

In this study, pattern matching-based digital well log correlation techniques revolve around correlating known patterns of interest from one well to other nearby wells. Digital well log data in this study is a data series whose data points are responses recorded by a logging tool with associated locations in the well where the responses are recorded. To make sure that digital well log correlation is valid, logging frequency is assumed to be the same for all sections to be correlated. To correlate a section of interest, sliding window-based technique is used on top of similarity measurement. Sliding window-based technique ensures that all possible candidate sections are included in similarity measurement where closeness between the section of interest and candidate section is measured as a quantifiable result called distance. As shown in Figure 5.1, a section of interest is being slid over a well log. While the section is being slid, similarity measurement is also performed to quantitatively assess closeness of the section of interest with the data window at which the window is. The section length is constant throughout the sliding operation and the step size is one index at a time. For example, a window of length 120 indexes at depth index 500 is called Window 500 and it covers from depth index 500 to depth index 620. The window will be called Window 950 when it is slid for another 450 indexes. Unless stated otherwise, the step size equals one and it also means the window is moved one data point at a time given one index is one data point. After similarity measurement is performed over all candidate sections, the best match having the smallest distance is chosen for correlation. The approach outlined is essentially a search technique named NNS. Therefore, it may be able to say that pattern matching-based digital well log correlation is NNS-based well log correlation. It has been known that the details incorporated in each variety of NNS are what make application-led NNS interesting. Owing to simplicity and robustness of the algorithm and applicability to well log correlation, conventional NNS and its cutting-edge variances are chosen as the backbone of pattern-matching-based digital well log correlation. This chapter shows how different distance measurements play a major role in robustness of well log correlation, especially when local variations are present. To demonstrate the applicability of NNS and its variances in well log correlation, three example correlations using real well logs are given.

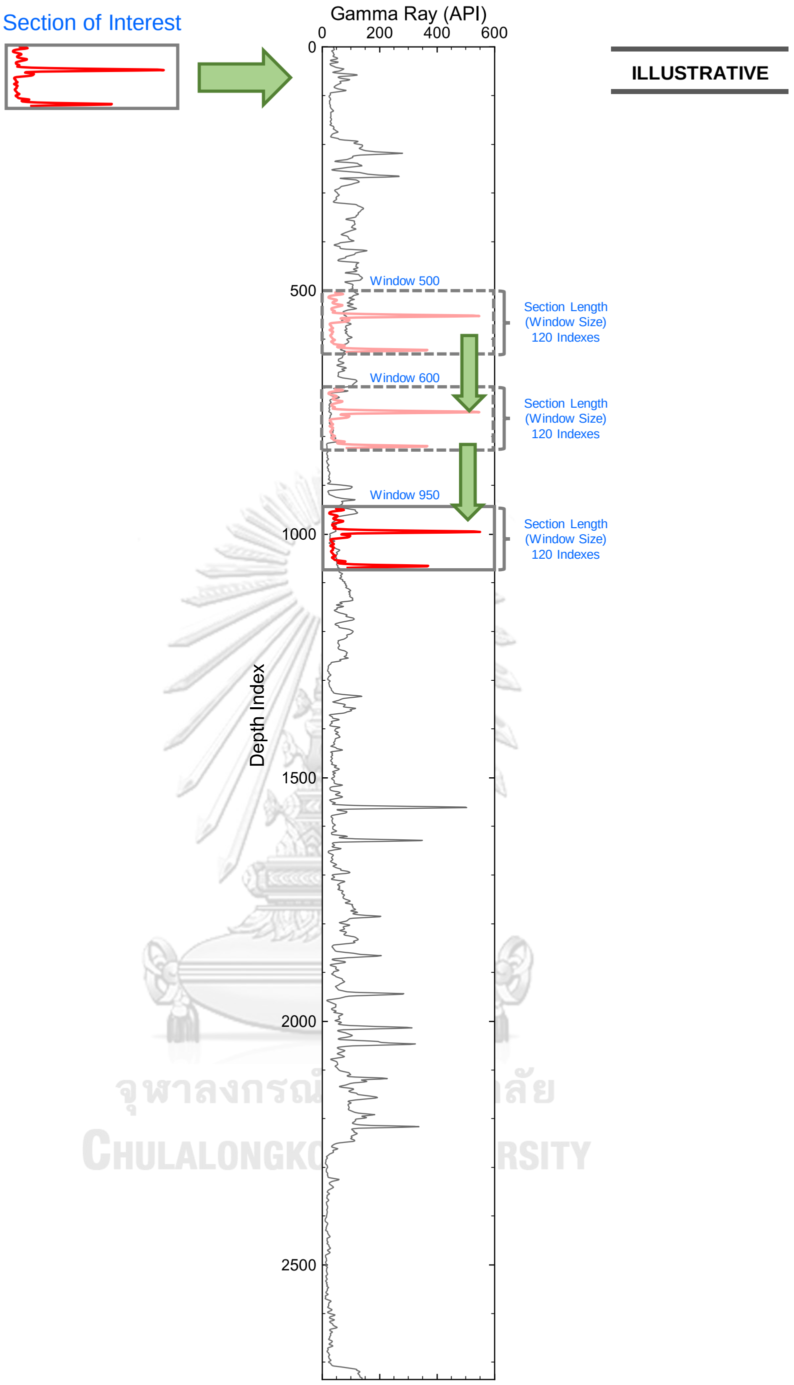


Figure 5.1 Sliding window of a section of interest.

5.1 Example Correlation 1

This example discusses how conventional NNS can be used in well log correlation. To start performing well log correlation using pattern matching technique, there must be a section of interest. This section is then used as a model pattern to find the most similar section in other wells using a sliding window-based technique. The model pattern is slid throughout other wells where well log correlation is needed.

While the model pattern is being slid along the well log to be correlated, a distance measurement between the model pattern and well log data within that window is performed. In conventional NNS, Euclidean distance is used as the distance measurement. Euclidean distance is considered suitable for absolute distance measurement as it treats values being used in the measurement as continuous values because even slightest difference between two numerical values can still be quantified. Top five data windows having smallest distances are chosen for discussion once the model pattern is slid to the end of the well log to be correlated. Figure 5.1 shows that five smallest distances of length 120 indexes (the same length as the model pattern) based on Euclidean distance are promoted as potential sections to be correlated.

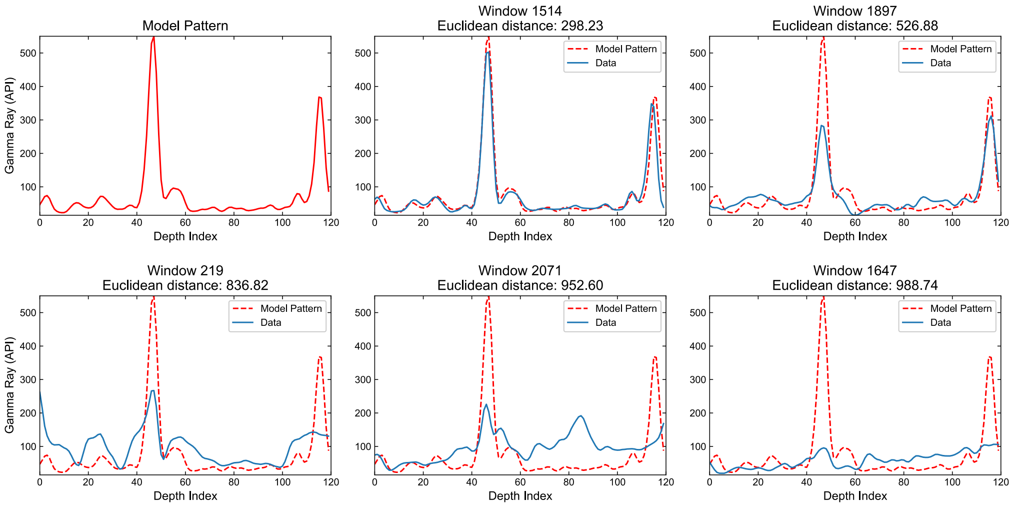


Figure 5.2 Five smallest distances promoted by conventional NNS

Figure 5.2 also shows that Window 1514 is the best match to the given model pattern because it has the smallest distance (298.23) among all data windows. In this example, the best match can be easily identified even without a distance measurement as it almost perfectly matches the model pattern. This case also shows that difference in data magnitude or difference of well log response at a certain depth contributes to distance measured. Window 1897, the second best match, gives a good example as its distance (526.88) is much higher than that of Window 1514 even though its overall data trends relatively match the model pattern. This is due to the fact that there are regions where large magnitude differences are present in the calculation and all the differences are equally treated under Euclidean distance.

While continuous-based approach already gives a good correlation to the model pattern, it is also of interest to see if multi-resolution analysis based on discrete-based approach will give any additional information on the correlation.

5.1.1 Multi-resolution Analysis

To perform multi-resolution analysis, basically original well logs have to be discretized (in both depth axis and data value axis) and be represented as discrete well logs, which have details reduced from the original well logs. Discrete well logs are well logs represented by series of predefined set of symbols. Then distance measurement is performed using a method that is capable of providing distance figure when at least two discrete series are being compared. SAX representation with Hamming distance and Levenshtein distance as distance measurement methods is chosen in this study.

The first step in SAX representation is to perform Piecewise Aggregate Approximation (PAA) to turn real-valued high frequency data of original well log data into real-valued data with fixed number of blocks by performing discretization on well log's depth axis. The resulting well log signal is called approximate signal. Block values are average data values of those original data points falling into the same blocks. For example, each PAA block is an average of 8, 4.8, 3.5 (approx.), and 2.7 (approx.) data points given the desired number of PAA blocks are 15, 25, 35, and 45 blocks over an original well log signal of 120 indexes as shown in Figure 5.3. It can be seen from Figure 5.3 that increased number of PAA blocks gives PAA signals with more details which capture peaks and troughs from the original signal. Too small number of PAA blocks may result in too-coarse approximate well log signal while too large number of PAA blocks may result in too much details kept in approximate well log signal.

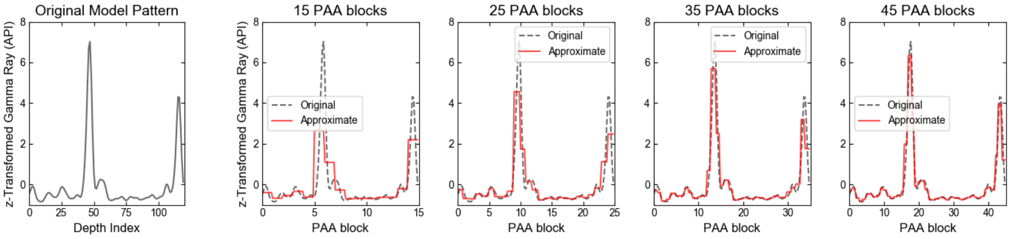


Figure 5.3 PAA signals on the model pattern using different number of PAA blocks

A balance point is typically required for generating an approximate well log signal from a high-frequency well log signal. A good balance point must provide the smallest distance (between the approximate well log signal and its original counterpart) as number of PAA blocks increases. One technique that can be used to determine an appropriate number of PAA blocks is to calculate the distance between original well log and its approximate version using a distance measurement method such as Euclidean distance. The most appropriate number of PAA blocks is when the distance between the two series is just about to increase as it is the point where the approximate well log is closest to its original counterpart. The method is shown in Figure 5.4.

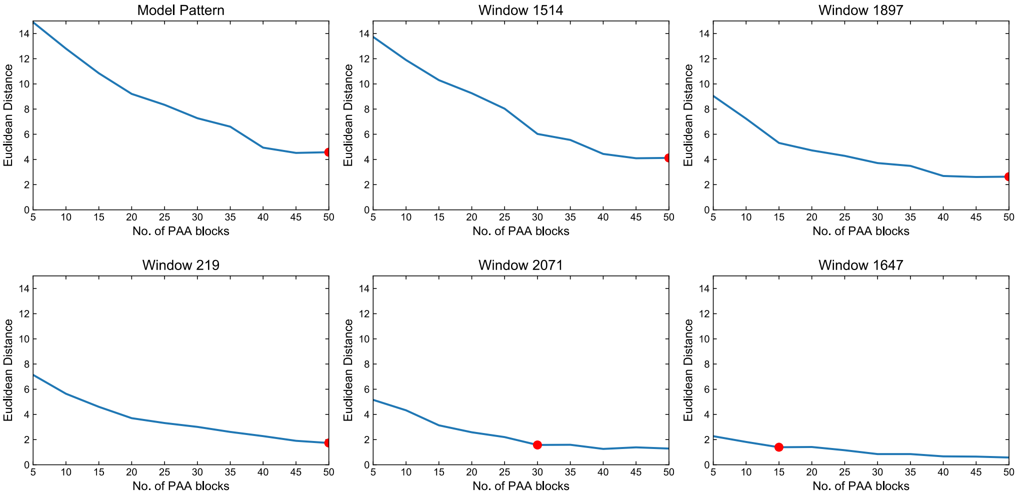


Figure 5.4 Selection of number of PAA blocks using Euclidean distance between approximate well log and its original counterpartFigure 5.4 shows that different data windows may have different balanced numbers of PAA blocks. While having higher number of PAA blocks means the discrete signals have higher resolution and are closer to their original counterparts as can be seen from reducing distance trends, it

is commonly found in practice that the optimal number of PAA blocks (shown with red dots) may not be the same in all data windows. To prevent bias in choosing just one particular number of PAA blocks, it is better to use multi-resolution analysis, i.e., analysis based on different numbers of PAA blocks. In this example, 10 sets of number of PAA blocks starting from 5 PAA blocks up to 50 PAA blocks, with increment of 5 PAA blocks, were used in multi-resolution analysis. Approximate signals based on 15, 25, 35, and 45 PAA blocks representing coarse resolution to fine resolution are shown as examples in Figure 5.5 to 5.8.

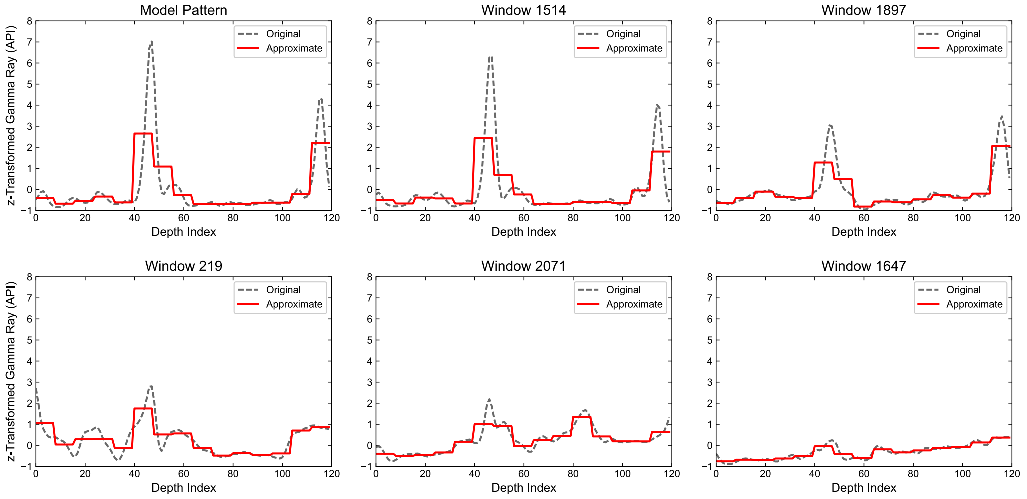


Figure 5.5 Approximate signals based on 15 PAA blocks

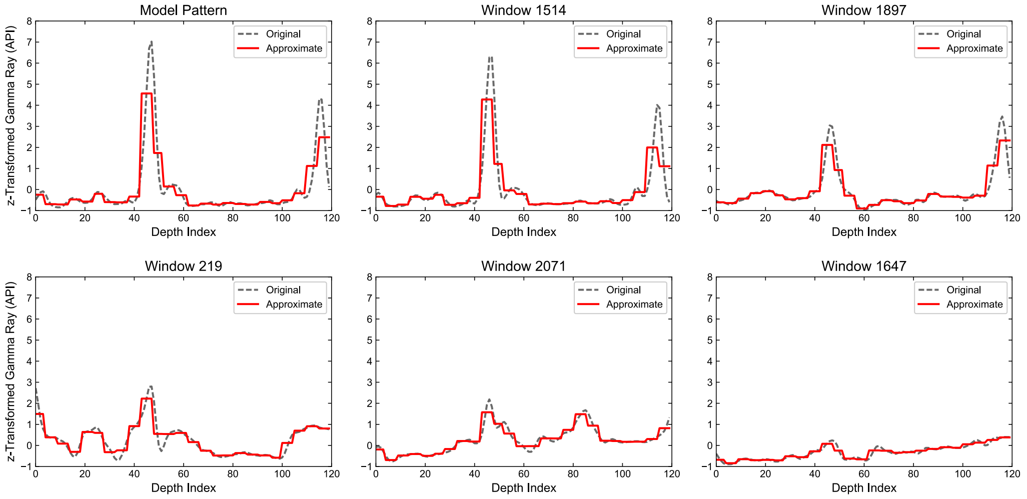


Figure 5.6 Approximate signals based on 25 PAA blocks

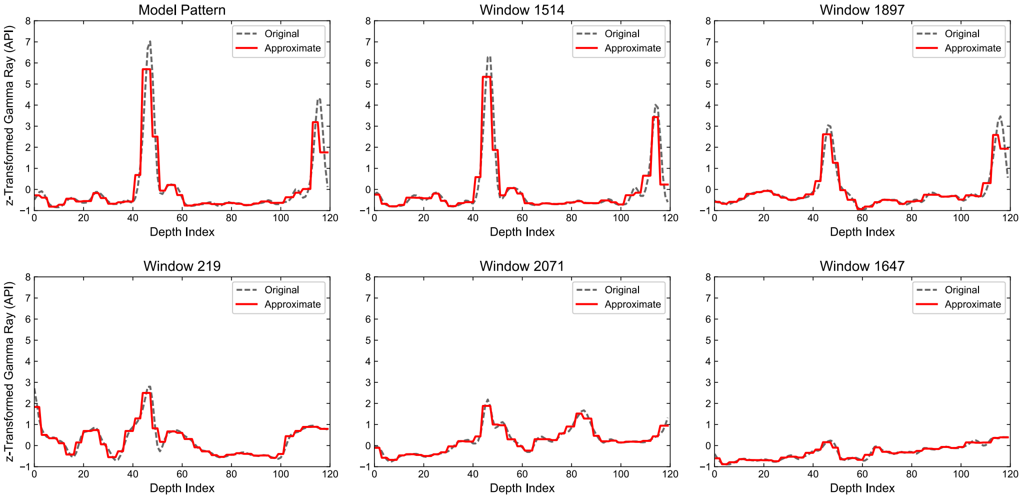


Figure 5.7 Approximate signals based on 35 PAA blocks

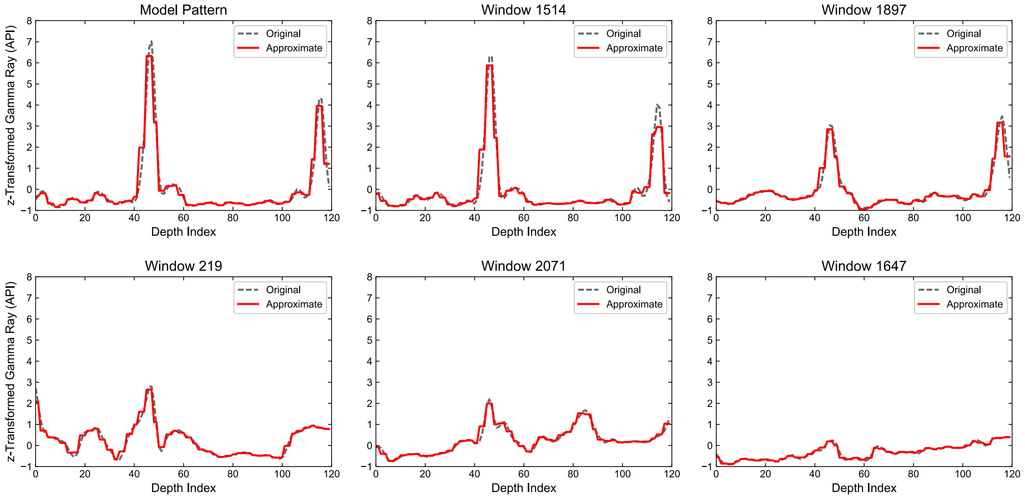


Figure 5.8 Approximate signals based on 45 PAA blocks

Figure 5.5 to 5.8 clearly show that different data windows need different number of PAA blocks in order for the approximate signals to cover not only their main trend but also characteristic peaks and troughs which later result in reducing distance trend as approximate signals are becoming closer to their original counterparts.

Given a specific number of PAA blocks, SAX representation can then be performed. While PAA is used to discretize depth (x-axis) into equal-size intervals, SAX equally discretizes the value axis (y-axis) based probability under normal distribution assumption, i.e., each interval of the y-axis has the same probability. This means that well log response must be assumed to be normally distributed. To make this method work practically, it is assumed that normal distribution is drawn from model pattern of each pattern matching task. Then SAX equally discretized sections are established based on number of equi-probable sections needed for model pattern. Finally, section boundaries drawn from SAX discretized model pattern will be used in all other data windows. Completing PAA discretization on the depth axis and SAX discretization on the value axis turn a real-valued continuous well log response to a discrete well log response. Figure 5.9 provides examples of a mode pattern under varying SAX discretization over an approximate well log of 120 PAA blocks. On the left most of the figure, it shows normal probability plot of the original model pattern. The fitting line is the line of best fit of the normal probability plot. The four remaining plots on the right of the figure provides what SAX discretization on the value (y-axis) looks like. As typically seen in normal distribution, the value ranges for each SAX interval are smaller near the mean value and larger far away from the mean value. Each value interval, however, is equally probable. The symbols on the right of each plots are the symbols used as representative values for each SAX section. For example, each SAX section accounts for probability of 0.250 when the value axis is discretized into four SAX sections represented by symbols A, B, C, and D. In this study, 4, 6, 8, and 10 SAX sections covering equal probabilities of 0.250, 0.167 (approx.), 0.125, and 0.100 are chosen.

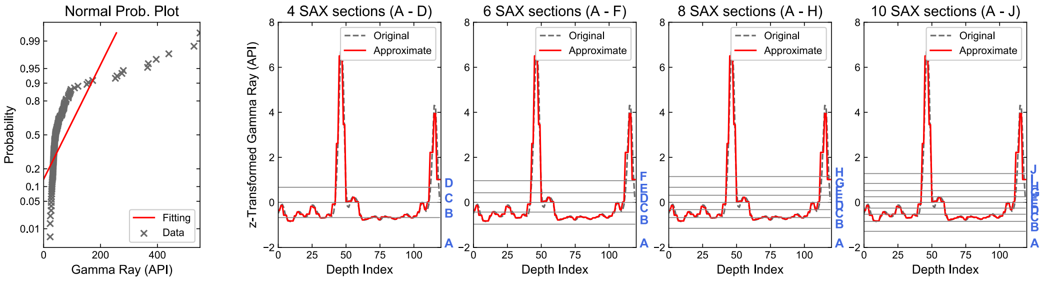


Figure 5.9 Well log value (y-axis) discretization using various equi-probable SAX sections

Based on an assumption that well log data are normally distributed, SAX can be used. The range covered by a SAX symbol is dictated by probability number where each SAX section covers an equal probability of well log value on the range. While it is designed based on normal distribution assumption, conventional SAX is also applicable to well log data that may not be perfectly normally distributed. However, accuracy of identified first rank match may vary also with distance measurement used.

In this study, Hamming distance is used for providing distance value based on same-position mismatch comparisons. Distance values are honored as there are mismatches of symbols used to represent well log values drawn from two (or more) well logs in the same depth interval (at the same PAA block). The method checks if there are substitutions of one symbol on the model pattern to another symbol on any given data window. This is beneficial in finding if there are changes in properties at the same depth interval given different well locations (from different well logs) and types of the well logs. Levenshtein distance not only captures possible substitutions of symbols but it also captures possible properties shifting of the same symbols in different depth intervals. Properties shifting is a challenging digital well log correlation problem. This is also beneficial in well log correlation as it helps digital well log correlation to be more accurate as properties shifting results in high Euclidean distance and Hamming distance, but may not result in high Levenshtein distance.

Similarity measurement in terms of Hamming and Levenshtein distances at different number of PAA blocks and SAX sections is used to observe how converted well logs into discrete series are valid in well log correlation task and how Hamming and Levenshtein distances help decide the best match to a given model pattern. In this example correlation, similarity measurement using Hamming and Levenshtein distances is performed against all data windows (1514, 1897, 219, 2071, and 1647) using all combinations of number of PAA blocks and number of SAX sections mentioned. Distance values are then plotted in color maps as shown in Figure 5.10. Notice that possible distance in these discrete-series similarity measurement is ranging from 0, identical to the model pattern, to the maximum number of PAA blocks, complete dissimilar to the model pattern. For example, the distance values can be between 0 and 30 when 30 PAA blocks are used.

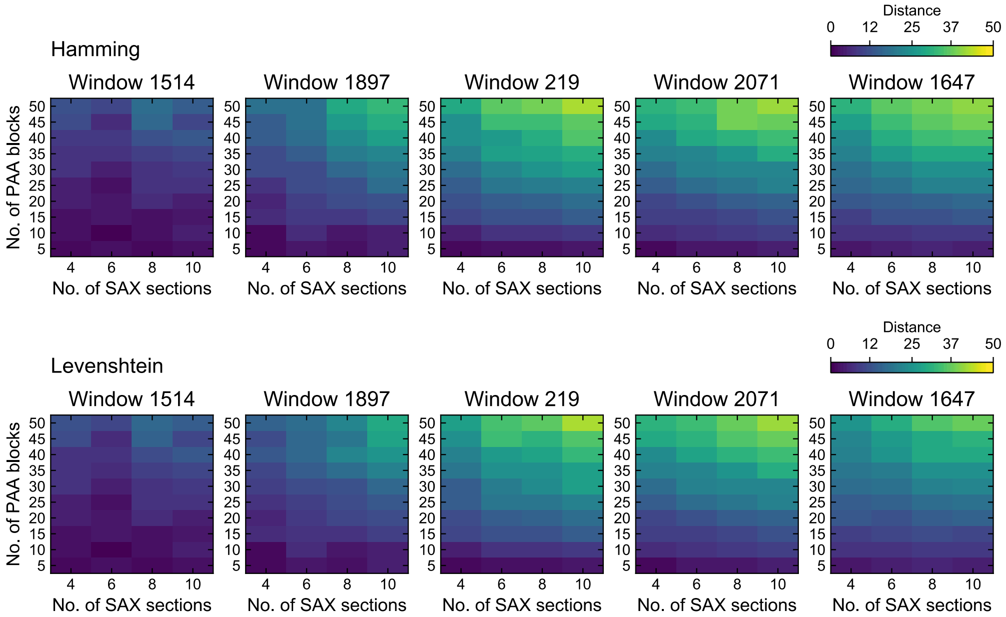


Figure 5.10 Comparisons of distance measured by Hamming distance and Levenshtein distance

Figure 5.10 shows similarity measurement in terms of Hamming and Levenshtein distances at different number of PAA blocks and SAX sections of the five data windows previously chosen by continuous-based well log correlation approach. The darker color represents smaller distance than that of the brighter color regardless of distance measurement methods used. From this figure, it confirms that Window 1514 is the best match to the given model pattern as it has the smallest Hamming and Levenshtein distances for different resolutions (different PAA blocks and SAX sections). An important point that can also be seen in the figure is that the higher the resolution of discrete signals (more number of PAA blocks and SAX sections) the higher the chance those signals pickup details from the original counterparts, thereby resulting in higher distance figures. While very small number of PAA blocks and SAX sections provides small distance values, it may not give enough details in the comparisons, thereby making less sense to use if the well log signals show significant degree of fluctuation or have local variations. This is due to the fact that too small number of discretized PAA blocks and SAX sections creates a too-coarse approximation, which in turn results in small distance as approximate and discrete signals are gathered around mean values, taking almost no peak and trough in the represented signals.

To finally suggest the best match under this multi-resolution analysis, only the smallest distance for each combination of different PAA blocks and SAX sections is promoted. As shown in Figure 5.11, the best matches are promoted independently for each combination of PAA blocks and SAX sections. Each data window is assigned a color to use in the resulting color maps. When one of the data windows is promoted as the best match given any particular combination of PAA and SAX, its representative color is painted in the color maps. The final best match based on multi-resolution analysis is considered in a probabilistic manner, by calculating number of that data window promoted ass the best match over total combinations.

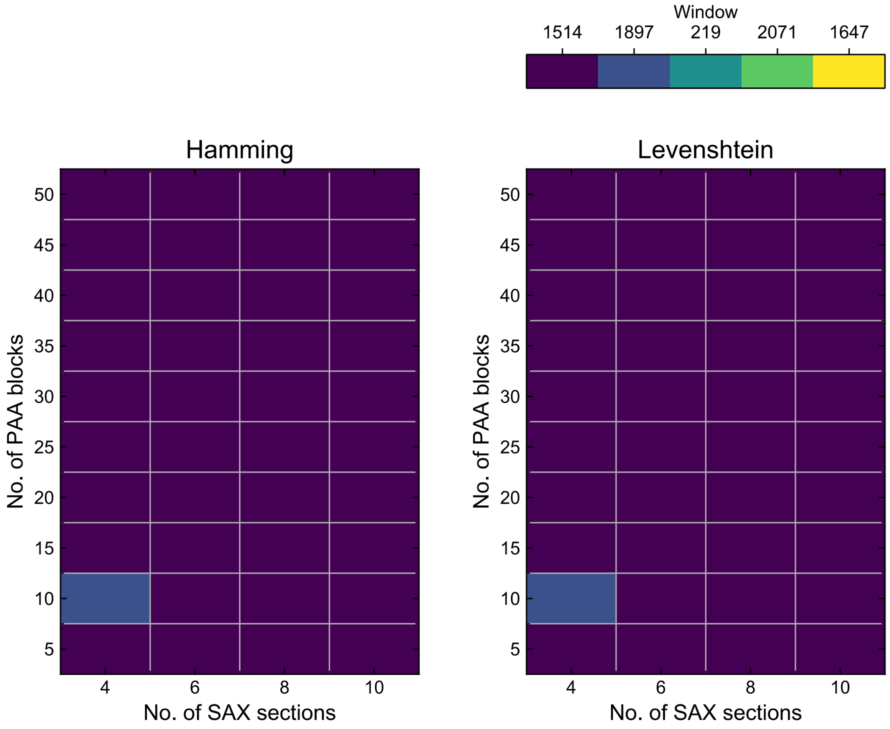


Figure 5.11 The best match data window promoted by multi-resolution analysis

Figure 5.11 shows that Window 1514 is most likely the best match to the model pattern given its probability of being the first rank, given a pair of number of PAA blocks and number of SAX sections used, of 0.975, outpacing the probability of Window 1897 to be the best match as its probability is only 0.025. To be promoted as the best match, the data window must have the smallest distances measured by both Levenshtein distance and Hamming distance. As can be implied from the results, multi-resolution not only helps confirm which data window provides the best match to the given model pattern, but also provide a probabilistic view of how best the best match is to be correlated to the model pattern. If both distance measurement methods cannot promote only a single best match, human interpretation may be needed and the distance values may be used as an aid for selecting the best possible matches.

5.2 Example Correlation 2

This example discusses how slight local variations affects well log correlation performed by NNS and how multi-resolution analysis can unveil the correct correlation. The model pattern of length 120 indexes used by this example is the same pattern used in 5.1. However, the pattern will be used to find another different best match from a different nearby well.

As outlined in section 5.1, NNS is first performed to find top five data windows giving the five smallest distances given the model pattern. Once NNS is complete, Window 1903, Window 1510, Window 214, Window 1281, and Window 8 are the five smallest data windows promoted by the search with distances to the model pattern of 578.36, 643.48, 923.47, 961.95, and 975.69 respectively.

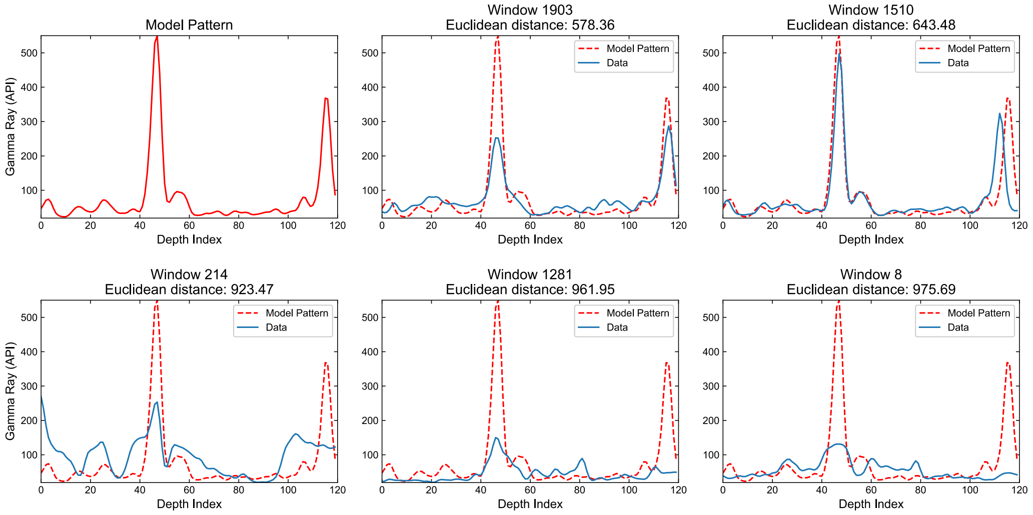


Figure 5.12 Five smallest distances promoted by conventional NNS

Under visual inspection, it could be said that Window 1903 and Window 1510 are arguably the best matches given some specific considerations. For example, Window 1903 could be chosen as the best match given the fact that the two gamma ray peaks are at almost the same location even though the two peaks do not have similar gamma ray levels and Window 1510 contains deviations in gamma ray starting from depth index around 105 onwards. Both better match in Window 1903 and deviations found in Window 1510 could be used to justify and select Window 1903 as the best match. However, one might say that Window 1510 is the best match to the given model pattern due to the fact that the first 60 depth indexes of the data window provides almost a perfect match to the model pattern and the two gamma ray peaks from the data window are relatively at the same or similar levels even though there are deviations in depth axis from depth around 105 onwards and its overall distance to the model pattern is higher than that of Window 1903.

To find true best match to the model pattern in this example, a more robust method is required. It is typical in nature that the sections correlated may contain some variations and it is challenging to justify which section is the best match in such situation by using only conventional NNS with Euclidean distance alone. Multi-resolution approach can help unveil true best match in such situation.

5.2.1 Multi-resolution Analysis

To continue with multi-resolution analysis, the model pattern and all the five smallest-distance data windows promoted by NNS undergo data representation to convert them from the original series first into block-like series and then further convert into discrete series by using SAX representation.

The first step of SAX representation is to perform PAA which gives block-like series based on originally continuous series and chosen PAA blocks. As can be seen from Figure 5.13, it is unable to find a common number of PAA blocks to convert the original series into approximate series since Window 1903, Window 1510, and Window 214 are more appropriate for PAA at the maximum number of PAA blocks at 50 blocks while Window 1281 is more appropriate at 20 PAA blocks and Window 8 is more appropriate at 30 PAA blocks. This variety in number of PAA blocks suggests that comprehensive multi-resolution using multiple number of PAA blocks is required. Therefore, PAA blocks of 5 to 50 with incremental of five PAA blocks are used in the multi-resolution analysis. Approximate signals based on 20, 25, 30, 35, and 40 PAA blocks representing coarse resolution to fine resolution are shown as examples in Figure 5.14 to 5.18.

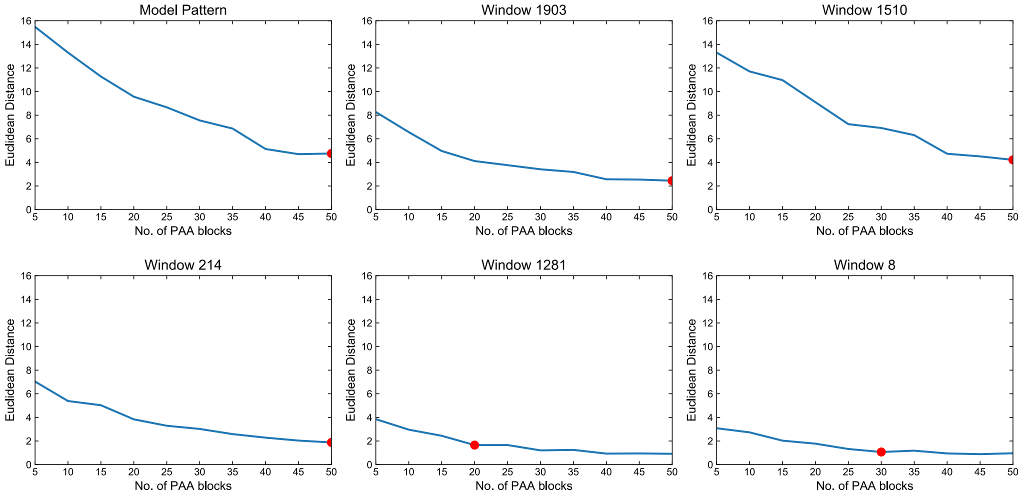


Figure 5.13 Selection of number of PAA blocks using Euclidean distance between approximate well log and its original counterpart

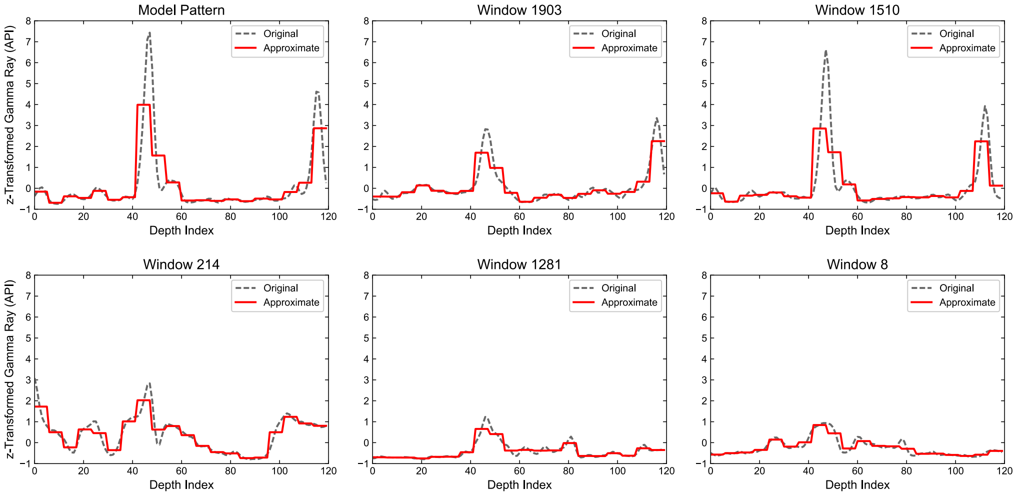


Figure 5.14 Approximate signals based on 20 PAA blocks

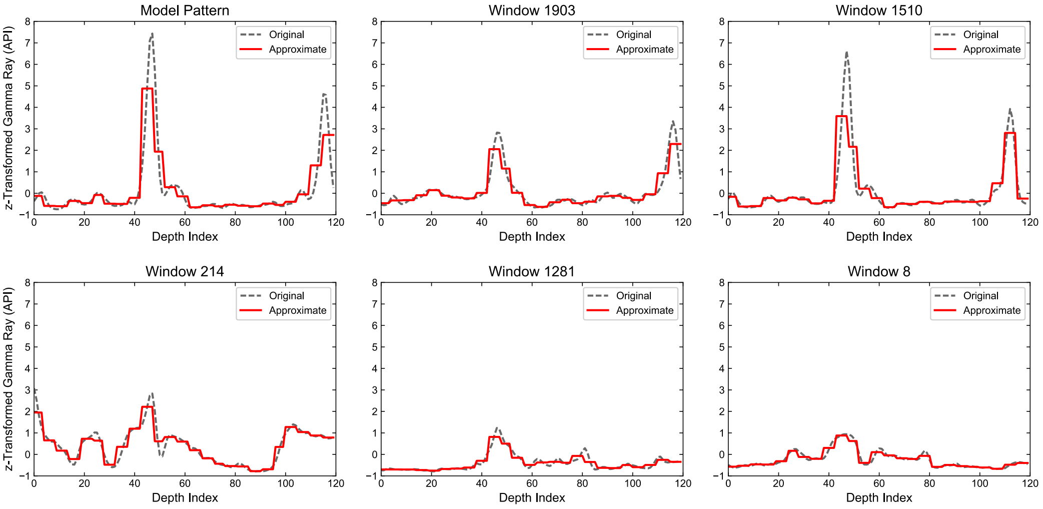


Figure 5.15 Approximate signals based on 25 PAA blocks

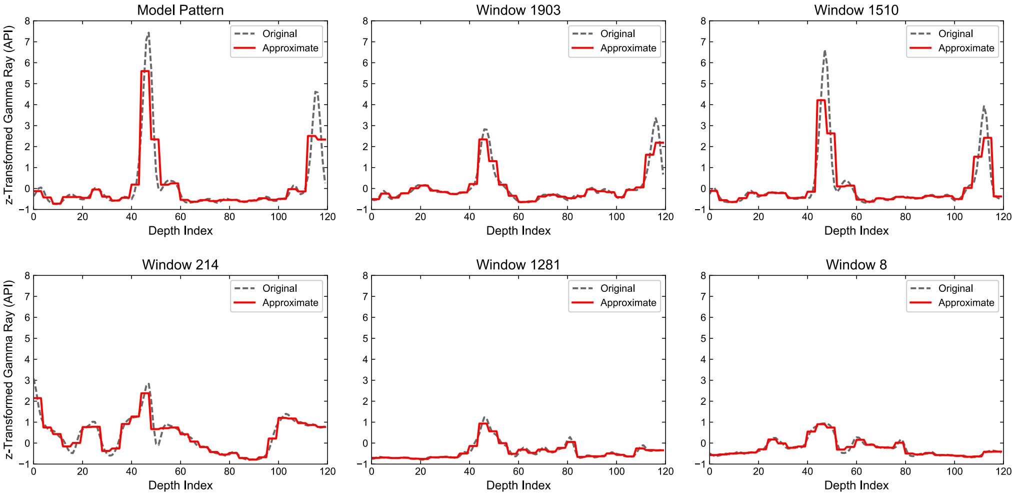


Figure 5.16 Approximate signals based on 30 PAA blocks

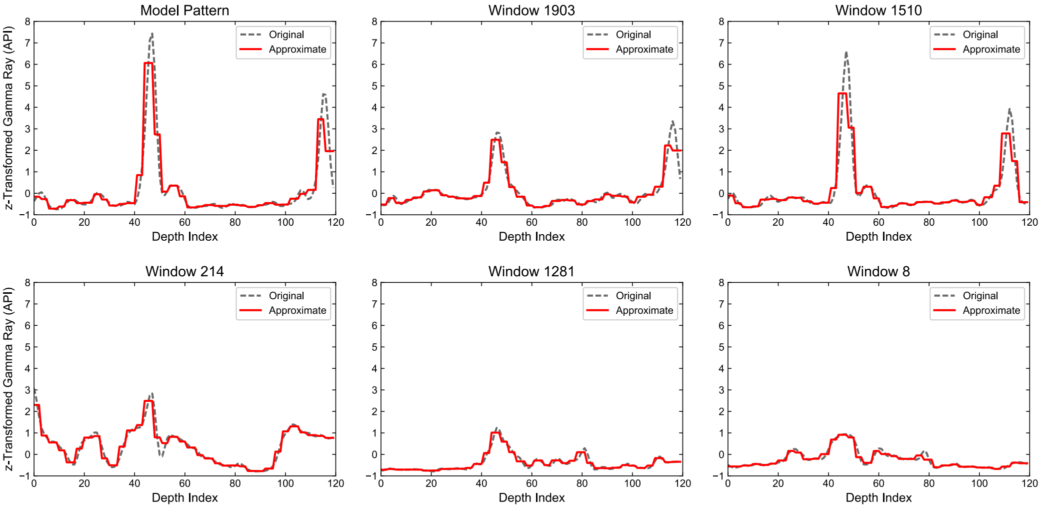


Figure 5.17 Approximate signals based on 35 PAA blocks

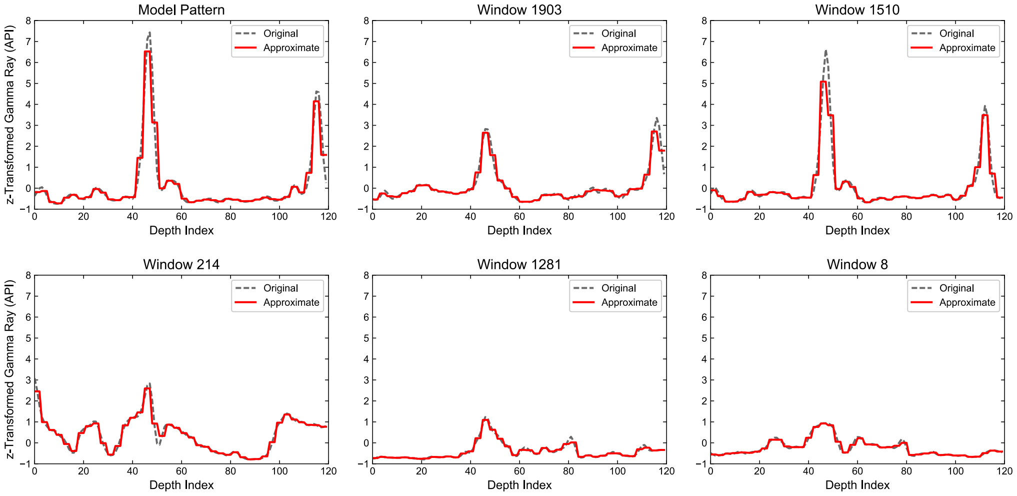


Figure 5.18 Approximate signals based on 40 PAA blocks

Increasing number of PAA blocks typically results in a closer approximation to its original counterpart. Logically, this increase in number of PAA blocks should result in the decrease in Euclidean distance calculated between approximate signal and its original signal. However, it is frequently observed that the higher number of PAA blocks do not always guarantee a closer approximation, especially when number of PAA blocks are relatively small compared to the total number of data points making the original signals. This is due to the fact that number of data points used in the numerical average in each PAA block may result in higher than that of the portion of original signal, thereby contributing to higher distance overall. This behavior is often seen in locations with fast and drastic change in the trends of the well log signals as shown in Figure 5.19 for approximate signals of Window 1281 under 20, 25, and 30 PAA blocks and Figure 5.20 for approximate signals of Window 8 under 30, 35, and 40 PAA blocks. The figures are highlighted with example locations where increased number of PAA blocks result in poorer approximation, signified by increased overall distance values.

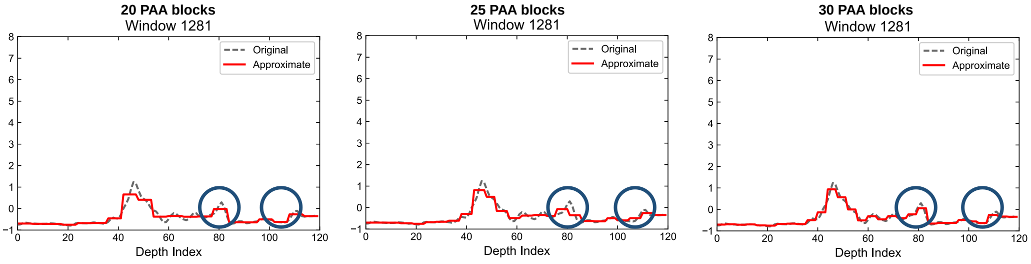


Figure 5.19 Approximate signals of Window 1281 under 20, 25, and 30 PAA blocks over z-transformed signals.

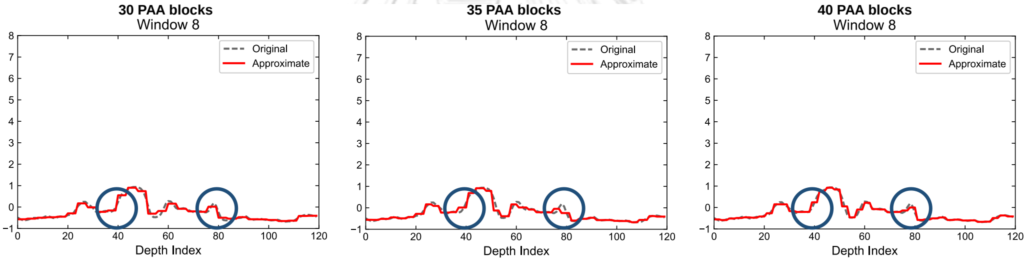


Figure 5.20 Approximate signals of Window 8 under 30, 35, and 40 PAA blocks over z-transformed signals.

Number of SAX sections are 4, 6, 8, and 10 as previously chosen from 5.1. This study found that these four sets of SAX sections are enough to be used for differentiating well log correlation under local variations. These 10 sets of number of PAA blocks, discretizing depth (x-axis) into equal-size intervals, and 4 sets of number of SAX sections, discretizing the log values (y-axis) into equal-probability intervals, account for 40 resolutions for each data window. This results in 200 resolutions for Hamming distance and 200 resolutions for Levenshtein distance, totaling 400 resolutions in the analysis. Figure 5.21 shows color maps of distance values calculated by Hamming distance and Levenshtein distance of all the 400 resolutions aforementioned. The figure also suggests that only one set of SAX sections is not sufficient for multi-resolution analysis. As can be seen in Window 1510's color maps, 10 SAX sections do not always give the closet approximation to the original signals since the overall distance values are higher than that of 8 SAX sections signified by the brighter color on the resolutions with 10 SAX sections than that of the resolutions with 8 SAX sections.

Once distance values of all resolutions are available, the data window having smallest distance value in each resolution combination is then promoted as the best match for that particular resolution, which results in Figure 5.22 where each resolution has only one best match to the model pattern. To finally select the overall best match, a probabilistic score is calculated. From the figure, Hamming distance promotes Window 1510 as the best match to the model pattern with probability of 0.875 while Levenshtein distance promotes the same data window with probability of 0.900. This whole process does not involve indicating the location of the model pattern at Window 1500 but multi-resolution analysis can still uncover the correct best match pair of Window 1510 in a nearby well.

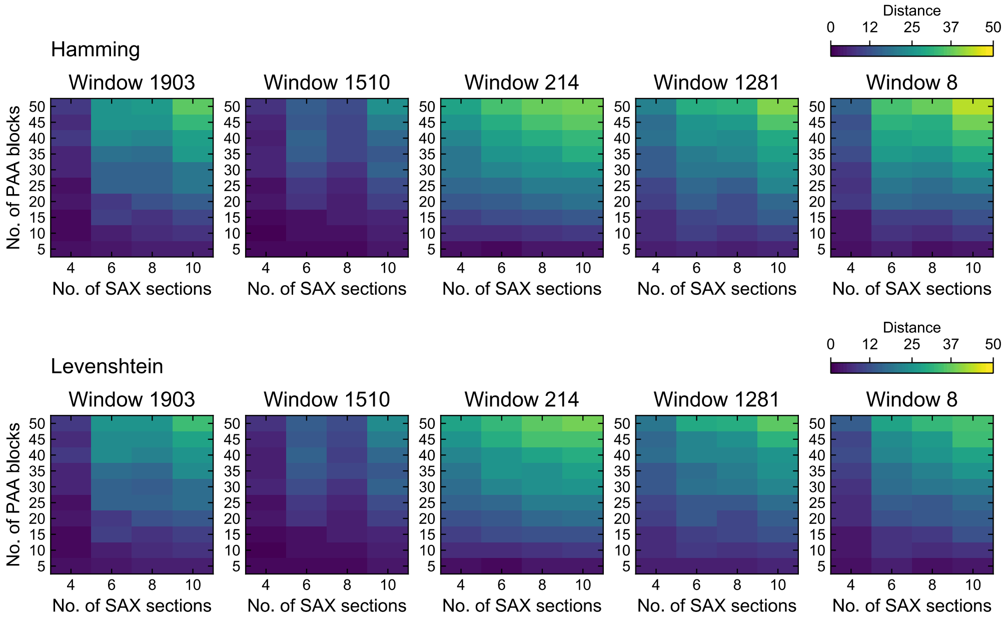


Figure 5.21 Comparisons of distance measured by Hamming distance and Levenshtein distance

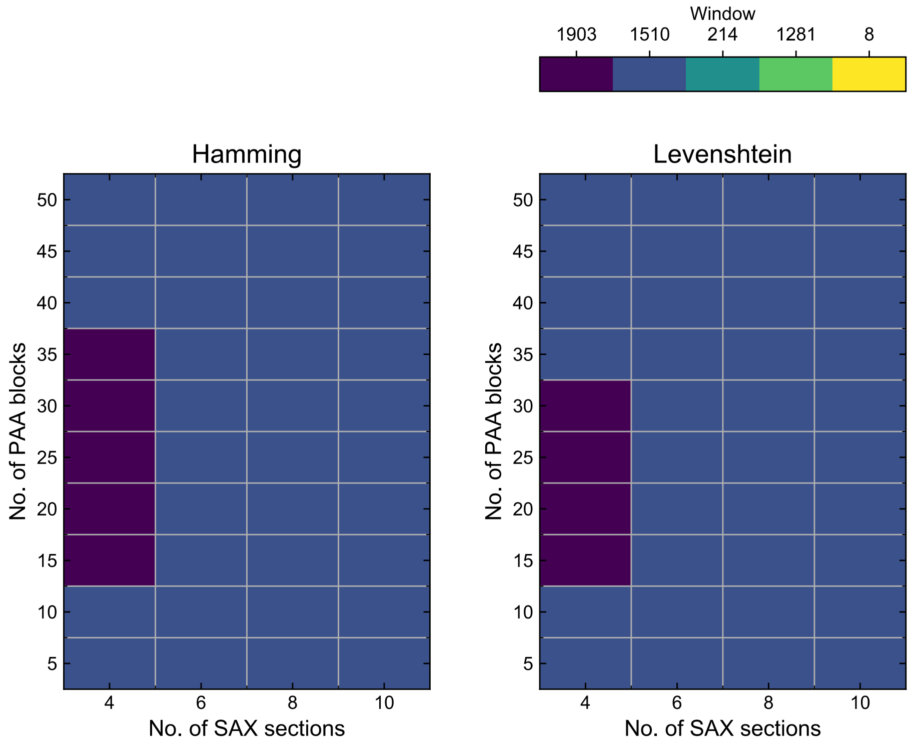


Figure 5.22 The best match data window promoted by multi-resolution analysis

5.3 Example Correlation 3

This example provides an example correlation using a longer section length than that of the first two example correlations. The model pattern used in this example is of length 200 indexes.

To start the process, NNS is first performed in order to find the five smallest distance under Euclidean distance given the model pattern. Figure 5.23 shows the five data windows (Window 901, Window 1290, Window 0, Window 698, and Window 2527) having smallest Euclidean distance from conventional NNS. Based solely on visual inspection, all data windows share some portions of them with relatively good match to the given model pattern shown on the upper left of the figure. For example, Window 2527 shows a relatively good match during depth index of 175 to 200 while it has the highest distance of all the five data windows. Window 1290 (the second rank) and Window 901 (the first rank) share a relatively good match during depth index of 25 to

50 but Window 901 has the lower overall distance. Window 0 shows a relatively good match during depth index 80 onwards but there are differences in the magnitudes of the log signals. While Window 698 seems to have the poorest match under visual inspection, it is however ranked four as mostly the data window follows the peaks and troughs of the model pattern, albeit much difference in the log signal magnitudes. Such patterns would be interesting under multi-resolution analysis.

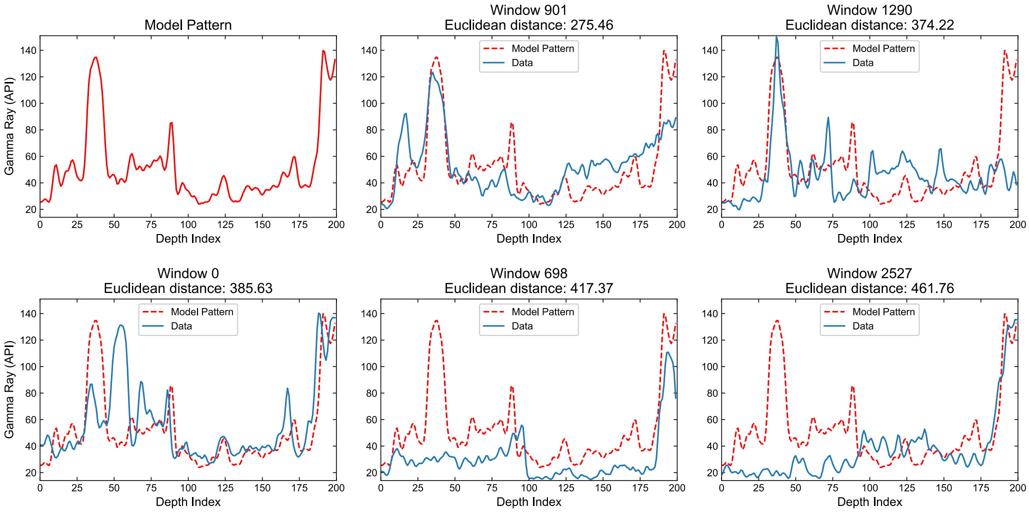


Figure 5.23 Five smallest distances promoted by conventional NNS

5.3.1 Multi-resolution Analysis

As previously mentioned, all the data windows promoted by conventional NNS share some similarity to the model pattern. This makes an additional analysis necessary as it may help finalize the best match. Since the model section is of length 200 indexes, the set of PAA blocks starts at 10 PAA blocks to 100 PAA blocks with block incremental of 10 as outlined in Chapter 4.

To start the multi-resolution analysis, an analysis of the suitable number of PAA blocks for discretization of the section depth (x-axis) is performed. Figure 5.24 shows different optimal number of PAA blocks in the model pattern and all other five data windows. The model pattern and Window 901 share the same optimal number of PAA blocks at 60 blocks while both Window 698 and Window 2527 share the same optimal number of PAA blocks at 50 blocks. Window 0 shows that its optimal number of PAA blocks is at 80 blocks while Window 1290 shows that 100 blocks are the optimal PAA blocks as the distances between the approximate signals and their counterpart are monotonically decreased throughout the entire selection of number of PAA blocks. Such varying number of PAA blocks make multi-resolution analysis necessary as it is impossible to choose only single number of PAA blocks without a bias.

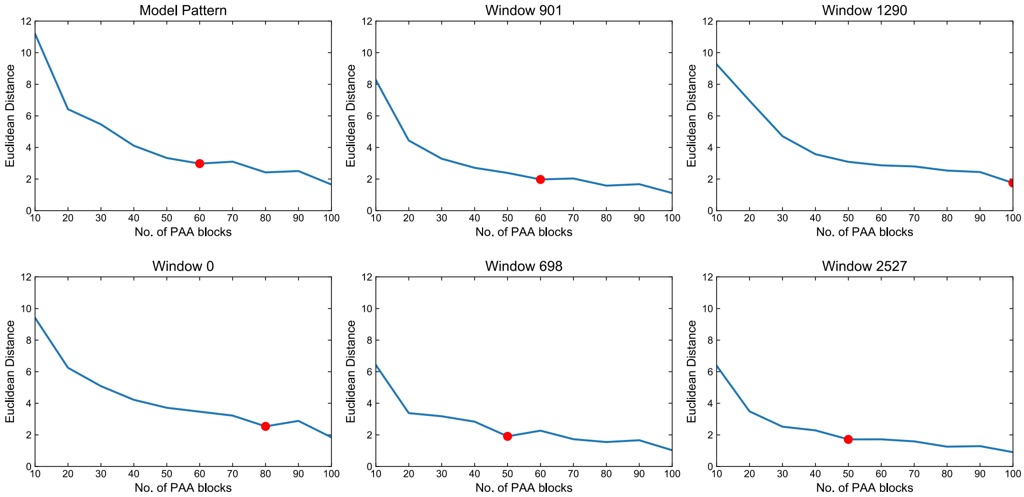


Figure 5.24 Selection of number of PAA blocks using Euclidean distance between approximate well log and its original counterpart

Once similarity measurement using Hamming and Levenshtein distances are complete, Figure 5.25 is plotted showing color maps of all resolutions. Based solely on visual inspection, Window 0 seems to have the darkest color maps overall, signifying that Window 0 could potentially be the best match. The interesting thing is that Window 0 generated by Levenshtein distance measurement seems to have a bit darker color than that of generated by Hamming distance measurement.

To finalize the best match, all data windows are ranked based on Hamming and Levenshtein distances. Figure 5.26 shows color maps of the best match under varying resolution under Hamming and Levenshtein distances. Hamming distance suggests that Window 0 be the best match to the model pattern with its probability of being the best match of 0.700 while Levenshtein distance suggests that Window 0 be the best match with higher probability at 0.850. Interestingly, Window 901 only has probability of being the best match only at 0.300 under Hamming distance and only at 0.150 under Levenshtein distance. No other data windows are promoted in this multi-resolution analysis, albeit similarity in the well log signals aforementioned.

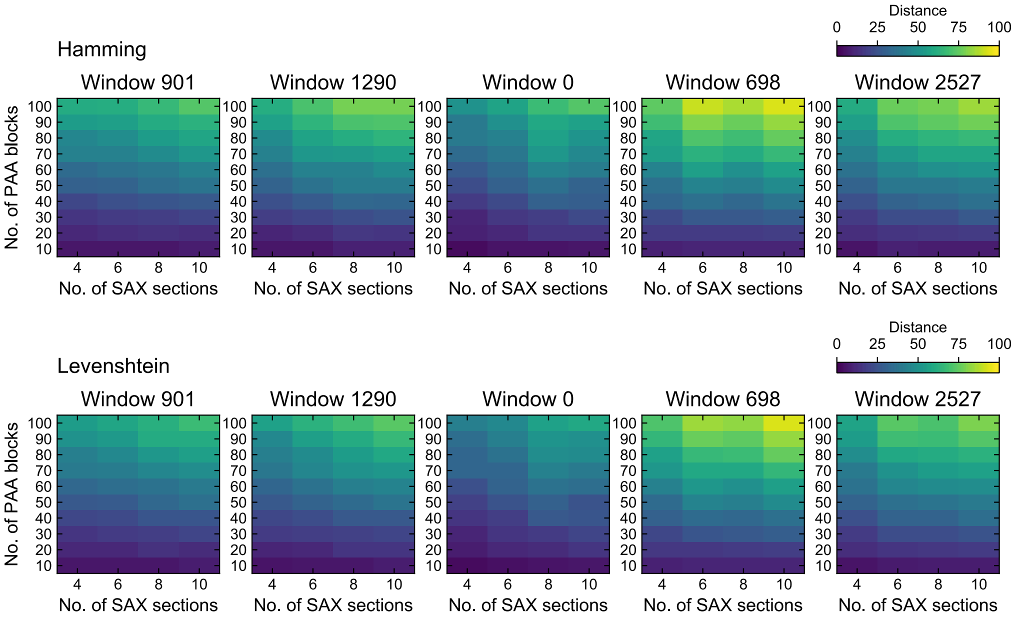


Figure 5.25 Comparisons of distance measured by Hamming distance and Levenshtein distance

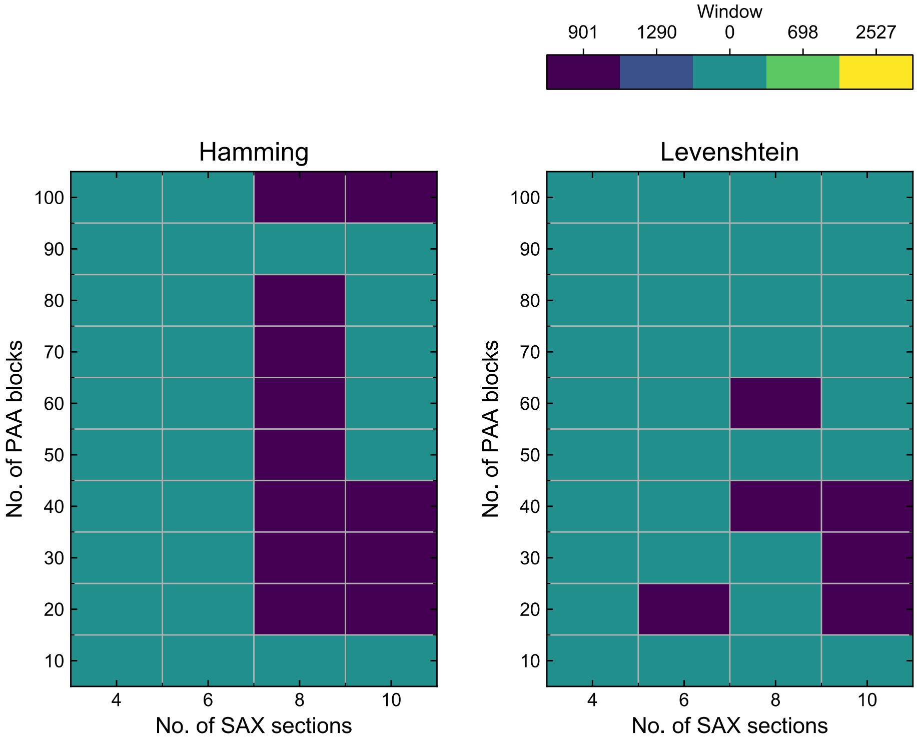


Figure 5.26 The best match data window promoted by multi-resolution analysis

As can be seen on all three example correlations, discrete series with multi-resolution helps identify best matches given different model patterns while Euclidean distance fails to identify the correct best matches. Table 5.1 provides a summary on parameters used in multi-resolution analysis and the final results based on the analysis of the three example correlations. All of the correlated sections are kept as they are valid sections.

Table 5.1 Example correlations and the parameters used

Example	Model Pattern Well	Model Pattern Window	Window Size (indexes)	Number of PAA blocks	Number of SAX Sections	Well Correlated	Data Window Correlated	Status
1	Well Y	1500	120	5 – 50 (incremental of 5 indexes)	4, 6, 8, and 10	Well Z	1514	Kept
2	Well Y	1500	120	5 – 50 (incremental of 5 indexes)	4, 6, 8, and 10	Well X	1510	Kept
3	Well Y	0	200	10 – 100 (incremental of 10 indexes)	4, 6, 8, and 10	Well X	0	Kept

5.4 Multiple Well-to-Well Log Correlation

As shown in three example correlations, pattern matching approach shows a promising result on well-to-well log correlation even when local variations exist. This example shows that multiple well-to-well log correlations can be performed using pattern matching approach, given all model patterns for each correlated sections are known. To perform multiple well-to-well log correlation, each well log pair is undergone well-to-well log correlation using known model patterns until all sections are correlated. Then the same model patterns are used to correlate another set of well-to-well log correlation. The approach works under an assumption that the model well whose model patterns are extracted must be the center of the other nearby wells to be correlated.

Figure 5.27 shows an example of multiple well-to-well log correlation using pattern matching approach. From the figure, patterns from Well Y are used as model patterns to correlate Well X and Well Z. Model pattern P has its length of 200 indexes. Model pattern Q has its length of 225 indexes. Model pattern R has its length of 130 indexes. Model pattern S has its length of 120 indexes and Model pattern T has its length of 220 indexes.

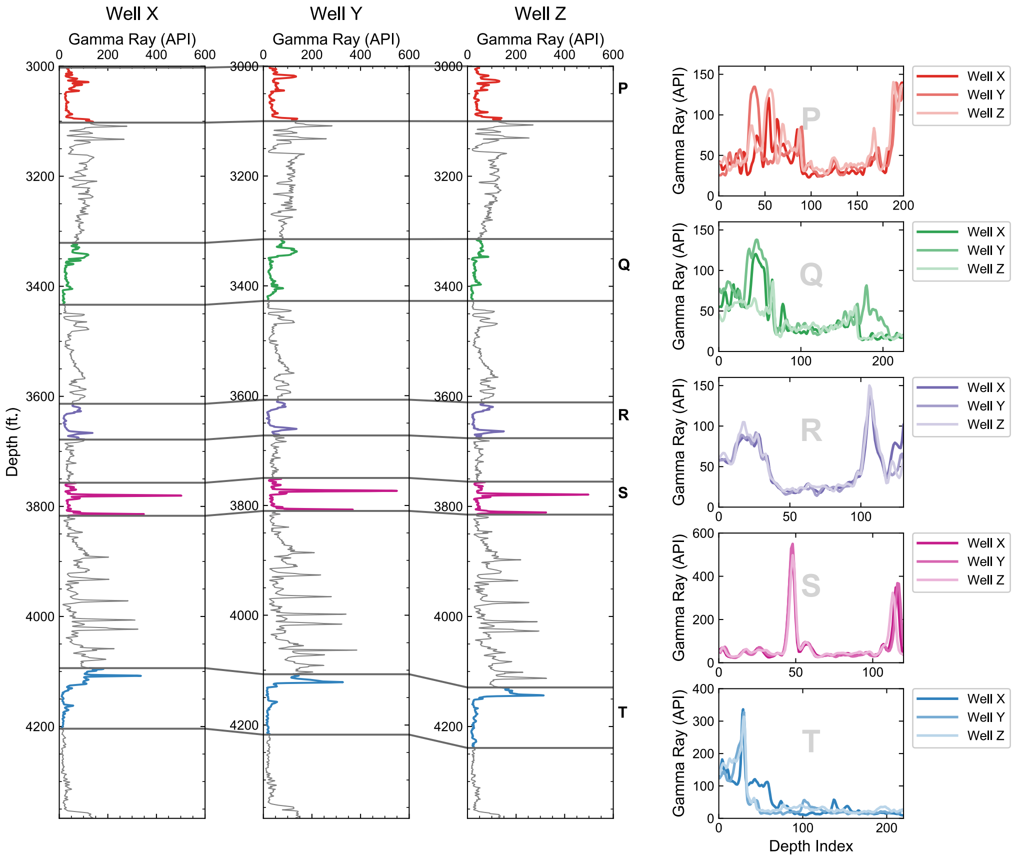


Figure 5.27 Multiple well-to-well log correlation using pattern matching approach

Table 5.2 Section length and the locations where the section is correlated across Well X, Well Y, and Well Z

Section	Section Length (indexes)	Well X Index	Well Y Index	Well Z Index	Status
P	200	5	0	0	Kept
Q	225	642	630	628	Kept
R	130	1227	1215	1222	Kept
S	120	1514	1500	1510	Kept
T	220	2188	2215	2258	Kept

CHAPTER 6

PATTERN DISCOVERY

Well log correlation still mostly relies on trying to find the best match for a known section drawn from a well-studied well log. This results in well log correlation shortcomings because well log correlation might not be started when there is no known section. Pattern discovery technique is used to allow well log correlation to start early when only a section thickness (well log section length) is known with certainty. One of the pattern discovery techniques that is almost readily applicable for well log correlation is MK algorithm.

The ability to find a pair of best match in a long series makes MK algorithm (MK) useful for well log correlation task. Given a section length, MK automatically returns the best-match pair having smallest Euclidean distance. Similar to Pattern Matching techniques, this smallest-distance pair can be correlated because the pair is considered the most similar sections of all other sections in the entire well logs. This chapter discusses how MK can be used in digital well log correlation without having to know any shape of well log pattern.

6.1 Example correlation 1

To use MK for well log correlation across any two wells, an artificial single well log must be generated from the well logs to be correlated. Connecting two well logs together would invalidate well log depth for each well log being connected because depth of a well log is monotonically increased along its depth scale. Therefore, well log depth has to be regenerated. The easiest way to regenerate well log depth for this artificial well is to use depth index. Typical depth index starts at *zero* at the first data point and *number of data points – 1* is the final depth index for the very final data point in the generated artificial well log. For example, the artificial well log will have 5,476 indexes given one well log of 2,741 data points and the other well log of 2,735 data points. Figure 6.1 and Figure 6.2 illustrate what original well logs are and how an artificially generated well log would be.

Without modification made in MK would have allowed any two well log sections to *self-correlate* on the same well. Therefore, MK must be slightly modified so that any best-match pair belongs to the same well is discarded. This slightly modified MK takes a *breakpoint*, which is the final depth index of the first well, into account and makes sure that both sections (forming best-match pair) are on different side of this breakpoint.

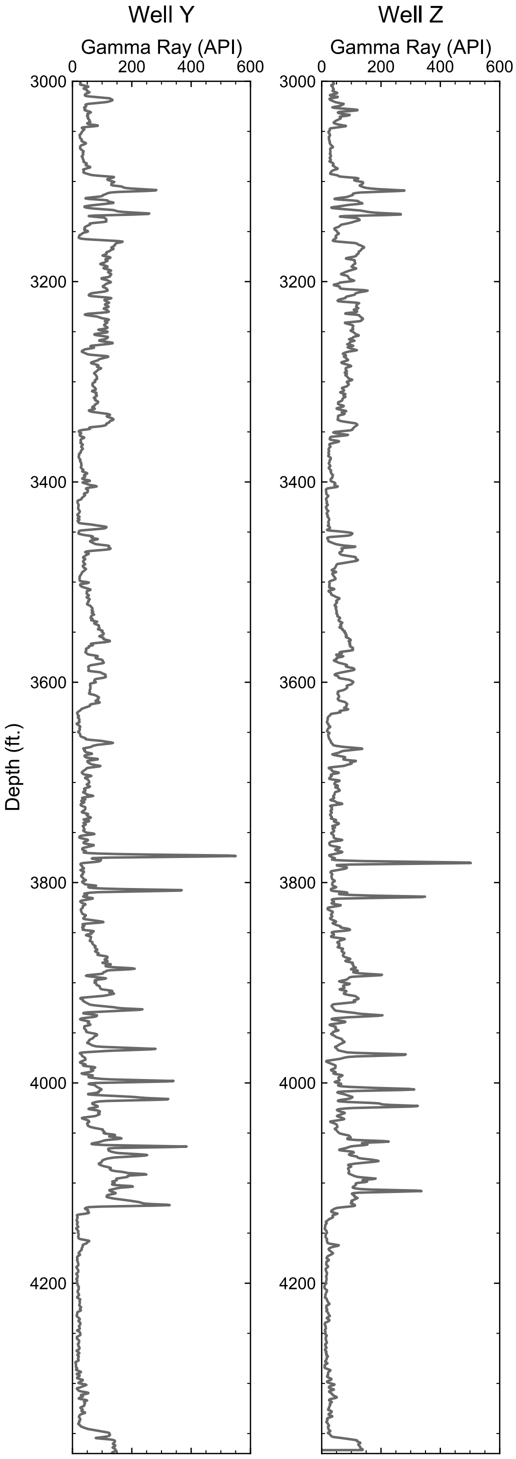


Figure 6.1 Sample well logs to be correlated

Figure 6.1 shows two original gamma ray well logs from Well Y and Well Z. Well Y was logged from depth 3,000.0 ft. to 4370.0 ft. (total depth indexes of 2,741 indexes) and Well Z was logged from depth 3,000.0 ft. to 4,367.0 ft. (total depth indexes of 2,735 indexes)

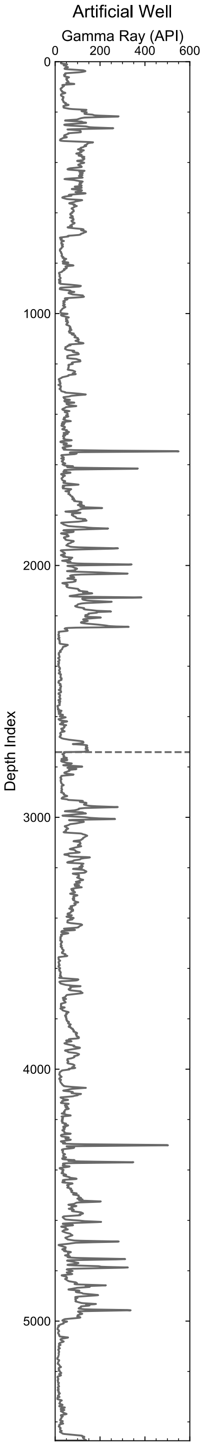


Figure 6.2 An artificial well log of length 5,476 indexes generated by connecting two well logs: Well Y and Well Z with a breakpoint (at dashed line).

Figure 6.2 shows that the artificial well log with 5,476 data points is generated from a connection of Well Y, having 2,741 data points, and Well Z, having 2,735 data points. In this case, 2,741 is the breakpoint (indicated by a dashed line in the middle of the generated artificial well log) between Well Y and Well Z. It is noticeable that Figure 6.2 uses depth index instead of real well depths from the two original well logs.

After the generation of artificial well log, the well log is then fed into MK with a section length (window size) and other parameters. MK will search for the best match given the section length. Given a section length, MK starts by randomly choosing a section to be used as a reference section. This reference section is then slid across the entire artificial well log and each comparison's distance (Euclidean distance) is measured. Therefore, this reference section is functionally analogous to a model pattern, which is manually chosen based on human knowledge, in pattern matching approach. More than one reference section is better at tightening distance search space, which could help prune off unnecessary comparisons and speed up the process to get the best-match pair.

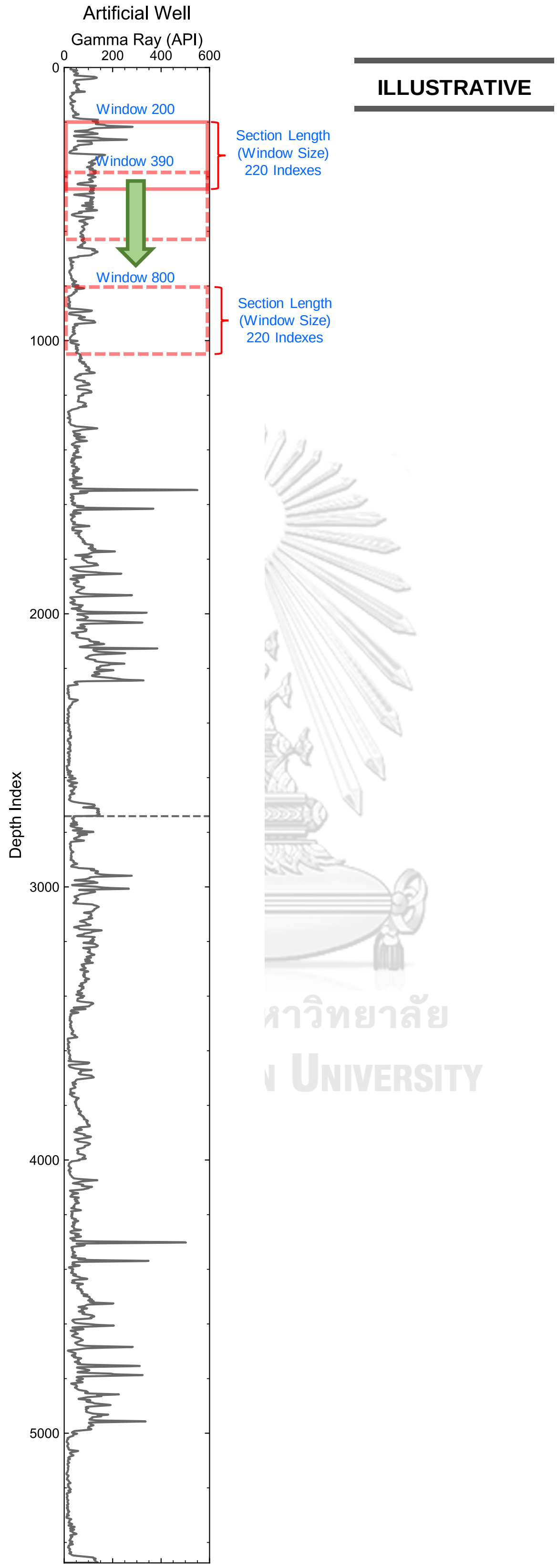


Figure 6.3 Sliding window

To make sure that multiple reference sections span across all available well log values, this study modifies how random reference sections are chosen. The reference sections are randomly chosen from each value range within 0.5 standard deviation (SD) span. SD is the SD of artificial well log. For example, there will be 21 reference sections randomly chosen within each 0.5 SD interval ranging from -5.0 SD to 5.0 SD (-5.0 SD, -4.5 SD, -4.0 SD, ..., 4.0 SD, 4.5 SD, and 5.0 SD) given one reference series is picked from each range. These minimum and maximum boundaries can be arbitrarily chosen based on number of reference series required.

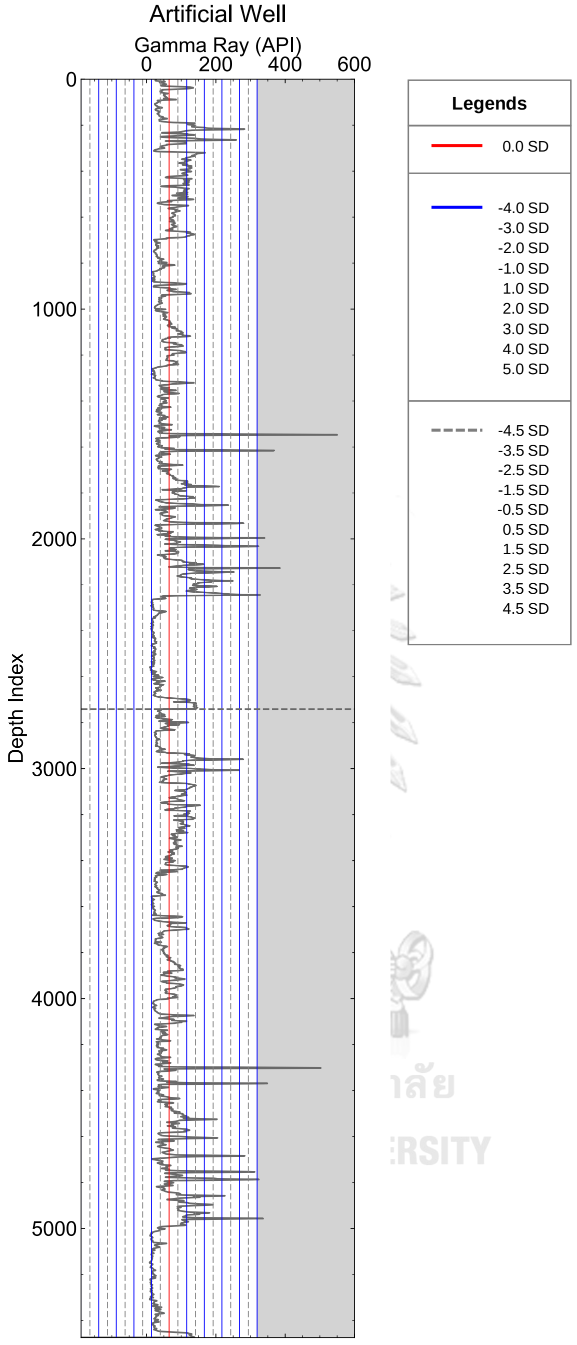


Figure 6.4 Arbitrarily chosen 21 sections based on -5.0 to 5.0 SD span

Figure 6.5 shows data windows that are randomly chosen reference series within -5.0 SD to 5.0 SD of the artificial log data values. For example, the data value at index 2571 is in -1.5 SD to -1.0 SD and is randomly chosen to be the candidate for this SD interval. The reference series from index 2571 is called Window 2571 which covers data from index 2571 to 2691 given the chosen section length or window size of 120. This is also the case for Window 4956 that is generated from index 4956, randomly chosen from the bucket 5.0 SD to 5.5 SD, to index 5076 given the pre-defined section length of 120 indexes. Notice that there are no reference series chosen on intervals of -5.0 SD to -2.0 SD due to the fact that there is no data value available in those ranges. This makes the remaining 14 intervals (14 data windows: one data window from each interval) useable as reference series. There are five indexes (262, 347, 792, 2029, and 2112) from Well Y and nine indexes (2234, 2320, 2571, 3957, 4299, 4609, 4754, 4755, and 4956) from Well Z.

Table 6.1 provides the complete list of data value ranges and key statistical parameters. It can be seen that the last SD interval does not cover all the data values from 344.74 API to 549.99 API. This is relatively still safe as there are only 12 data points out of total 5,476 data points. In addition, very rare data values do not likely give a good reference series which can be used to extract many pairs.

Table 6.1 Well log value from -5.0 SD to 5.0 SD span in 0.5 SD interval span

Standard Deviation		Gamma Ray (API)		
Min	Max	Min	Max	
-5.0	-4.5	-188.80	-163.41	No data available
-4.5	-4.0	-163.41	-138.01	
-4.0	-3.5	-138.01	-112.62	
-3.5	-3.0	-112.62	-87.23	
-3.0	-2.5	-87.23	-61.83	
-2.5	-2.0	-61.83	-36.44	
-2.0	-1.5	-36.44	-11.05	
-1.5	-1.0	-11.05	14.35	Data available
-1.0	-0.5	14.35	39.74	
-0.5	0.0	39.74	65.14	
0.0	0.5	65.14	90.53	
0.5	1.0	90.53	115.92	
1.0	1.5	115.92	141.32	
1.5	2.0	141.32	166.71	
2.0	2.5	166.71	192.10	
2.5	3.0	192.10	217.50	
3.0	3.5	217.50	242.89	
3.5	4.0	242.89	268.28	
4.0	4.5	268.28	293.68	
4.5	5.0	293.68	319.07	
5.0	5.5	319.07	344.47	

Mean	65.14	API
SD	50.79	API
Max	549.99	API
Min	9.05	API

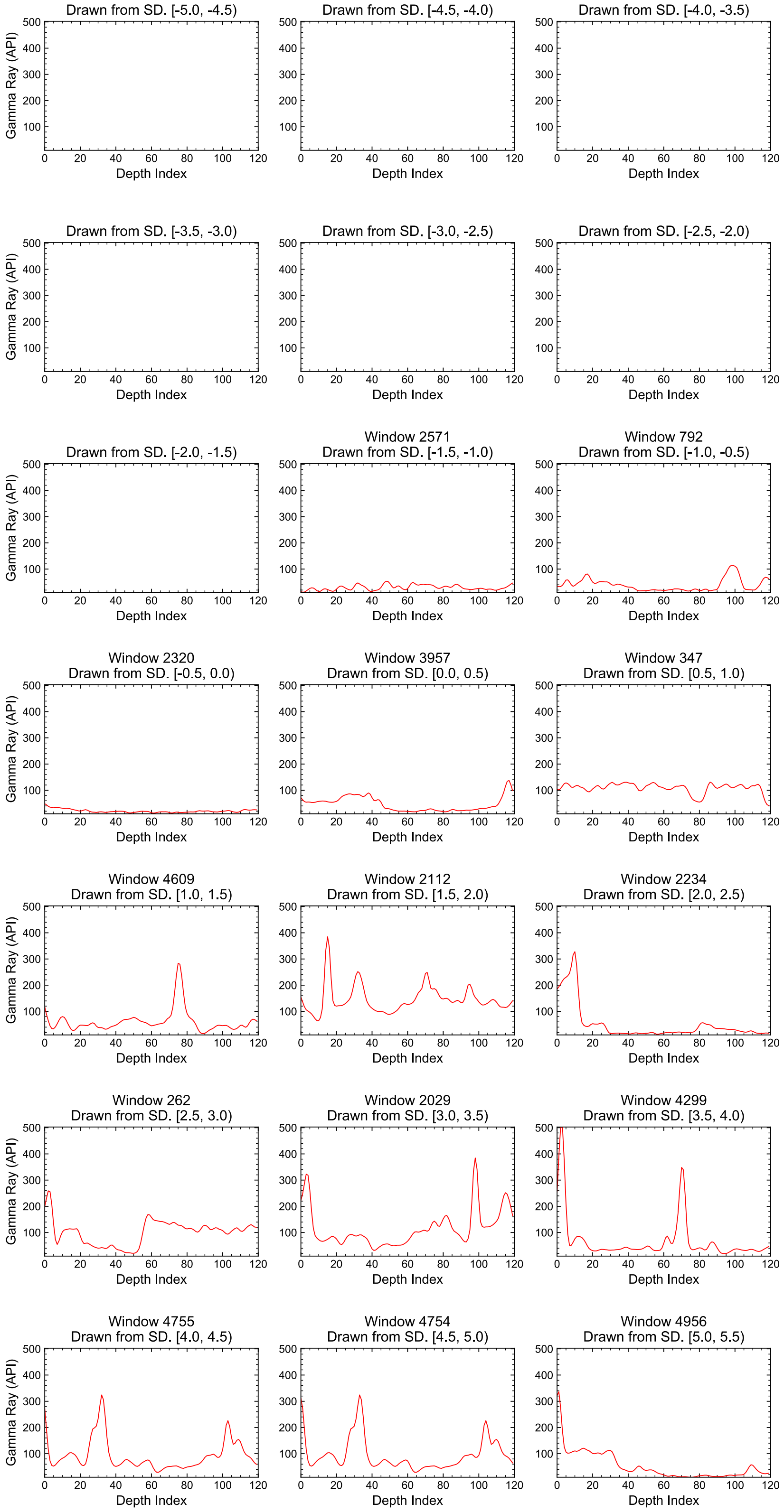


Figure 6.5 21 SD intervals and 14 reference sections each randomly chosen from -5.0 SD to 5.0 SD of the artificial well log mean.

Once reference sections are available, those sections is then used in distance measurement. In this case, Euclidean distance is used for measuring distance between each reference series and each data window being slid for the entire artificial well log and distances between the reference series and each data window is collected. Given 14 chosen reference series, there are also 14 sets of distance values and each set contains 5,356 distance values (5,476 – 120 = 5,356).

Then those 14 sets of distance values between the reference series and the other data windows are ranked by their standard deviation of the distance values of the sets. The set with highest standard deviation of the set’s distance values is the first rank and the set with lowest standard deviation of the set’s distance values is instead the last rank. As shown in Figure 6.6, Window 4956 has its standard deviation of distance values of 2.95 which is the highest value in all reference series and Window 2112 has its standard deviation of distance values of 1.32 which is the lowest value in all reference series.

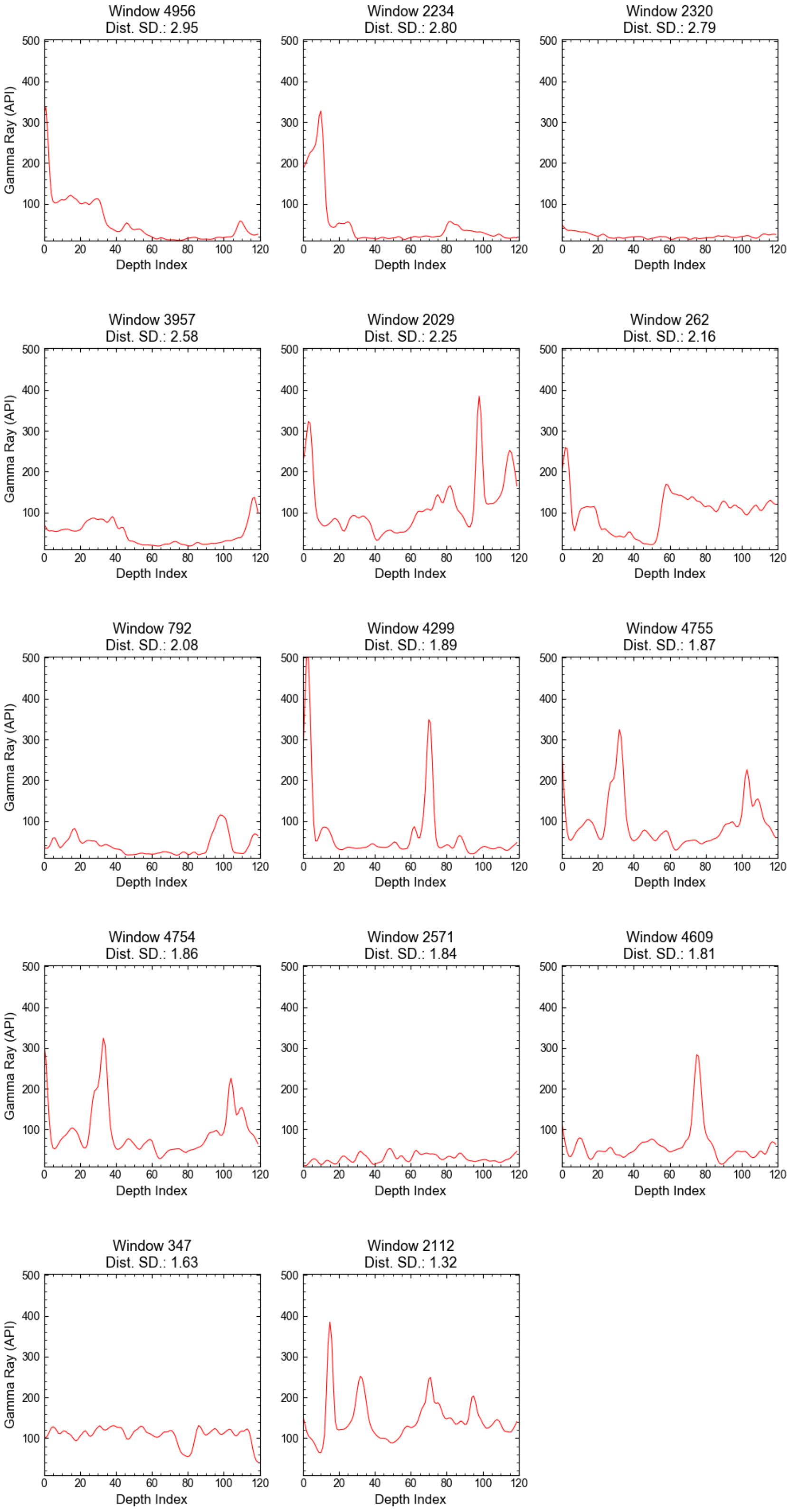


Figure 6.6 Reference series of the highest-to-smallest ranks based on standard deviation of distance values.

The next step is to search for the closest pair newly drawn from artificial well log with an aid of the promoted reference series. To illustrate how well log section search and comparison are performed, Figure 6.7 to Figure 6.9 are used. Assume that current best-so-far was a distance value D higher than 2.20 before the pair shown in Figure 6.7 was discovered. In the process of finding well log section pair to be correlated using MK, Window1168 and Window 3923 are used to calculate lower bounds with all other 14 data windows (shown in Figure 6.6), starting from lower bound calculation with Window 4956 (having the highest standard deviation of distance values), Window 2234, ..., Window 347, and Window 2112 (having the lowest standard deviation of distance values). Window 1168 and Window 3923 both passed lower bound testing for all 14 reference series (lower bounds with all reference series are less than the then best-so-far D). This makes both data windows require a confirmation if a distance between themselves is also less D. After the confirmation calculation, it turns out that the distance between Window 1168 and Window 2112 is smaller than D. Therefore, the pair is kept and best-so-far is updated with 2.20, the distance between Window 1168 and Window 3923. After some discards on data windows failing lower bound tests with best-so-far 2.20, the search and comparison continues for Window 145 and Window 2887 pair. The lower bound for this pair is 0.13 which is lower than current best-so-far (2.20), making additional check required. The distance between both data windows is 2.10 which is also smaller current best-so-far. Therefore, current best-so-far is updated with 2.10 and the data window pair is kept. After some more discards on data windows failing lower bound tests with best-so-far 2.10, the process continues for Window 1480 and Window 4235 pair. It turns out that the pair is valid and best-so-far is updated with its distance of 1.53. All other potential pairs either fail lower bound tests or have their distances larger than or even equal to the best-so-far of 1.53. Therefore, MK ends with the first-rank pair of the best-so-far of 1.53 given the 14 reference series (shown in Figure 6.6) and section length of 120 indexes.

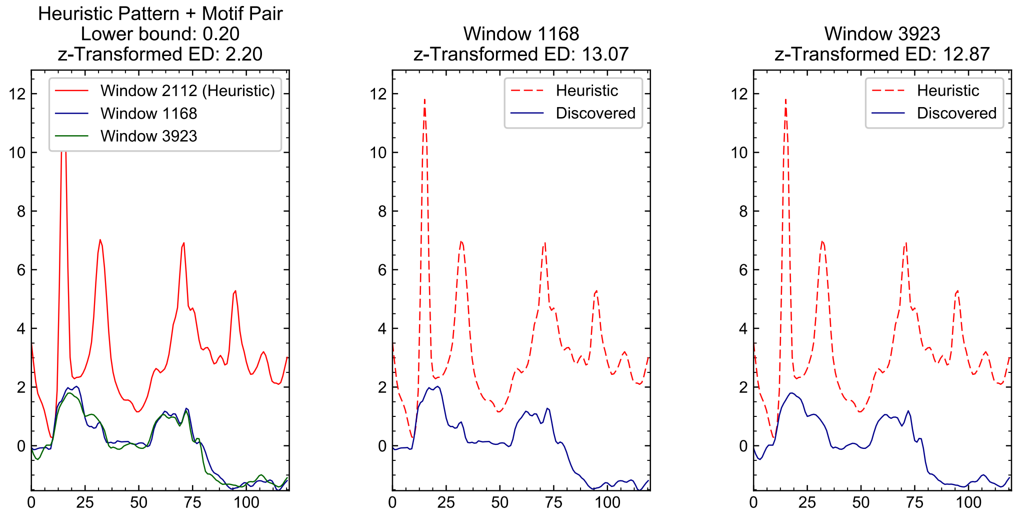


Figure 6.7 The third-rank motif pair of Window 1168 and Window 3923 with reference series of Window 2112.

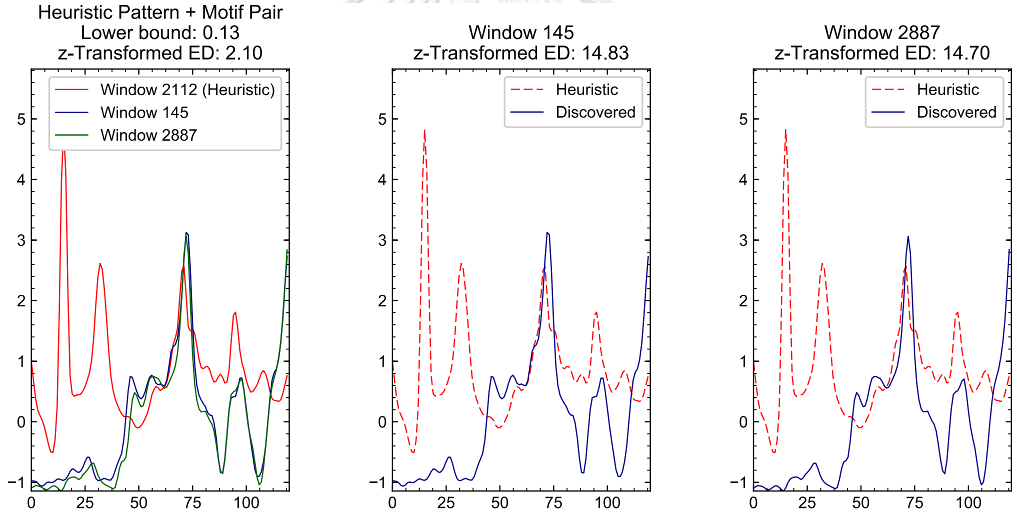


Figure 6.8 The second-rank motif pair of Window 145 and Window 2887 with reference series of Window 2112.

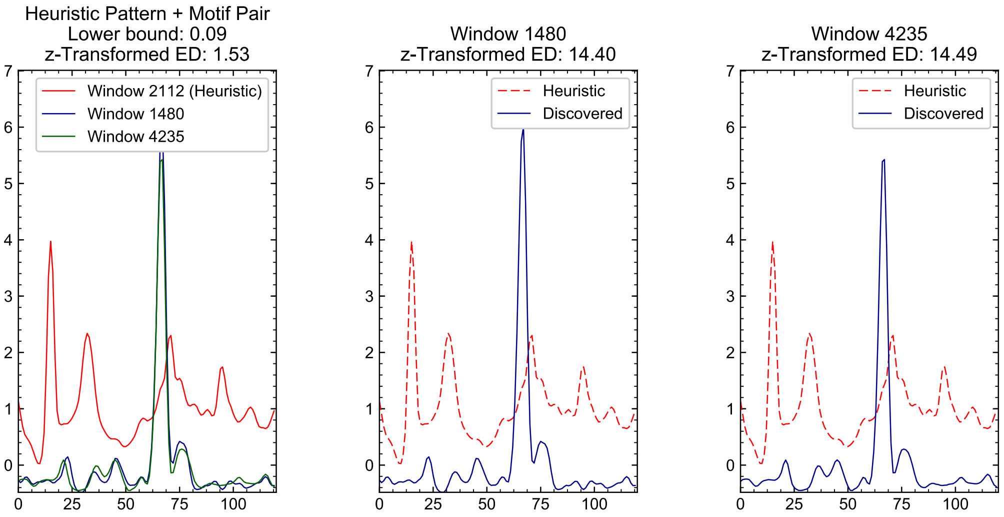


Figure 6.9 The first-rank motif pair of Window 1480 and Window 4235 with reference series of Window 2112.

The resulting well log correlation of this section length of 120 indexes is as shown in Figure 6.10.

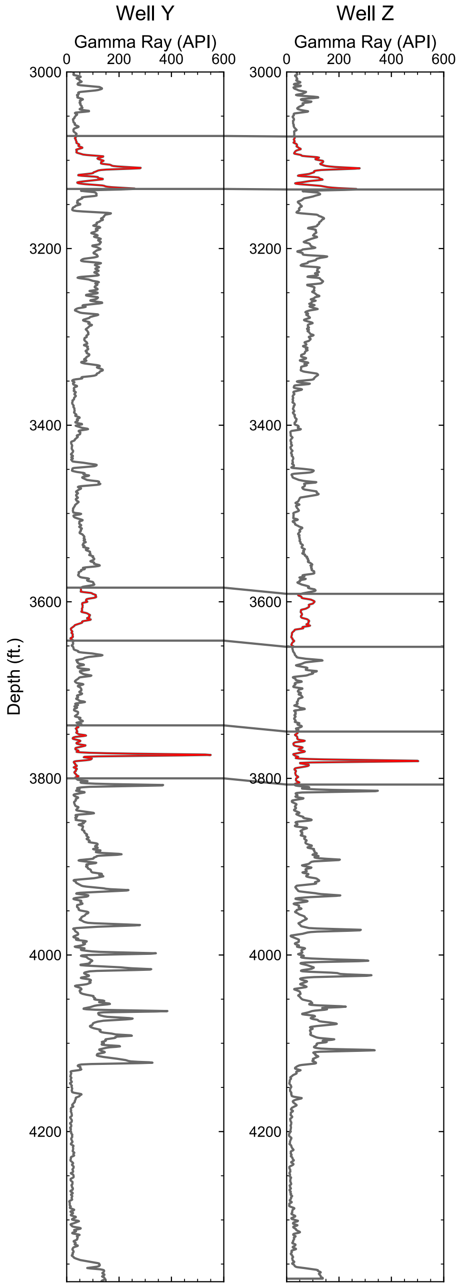


Figure 6.10 The final correlation of section length 120 indexes (60 ft)

Figure 6.10 shows locations of the motif pair in their own well log. Since the breakpoint between the two well logs is 2741, any data window index that is greater than the break point can be converted to local index of Well Z. For example, Window

4235 on artificial well log is Window 1494 Well Z's well log. Table 6.2 shows the resulting correlations of this well log signals

Table 6.2 Example correlations and the parameters used

Well 1	**Well 2**	**Section Length**	**Section**	**Well 1 Index**	**Well 2 Index**	**Pair Distance**
Well Y	Well Z	120	1	1480	1494	1.53
			2	145	146	2.10
			3	1168	1182	2.20

6.2 Example correlation 2

In this example, a new well called Well X is going to be correlated with Well Y. As outlined in the example correlation 1, the two wells must be connected together to form one artificial well prior to using MK.

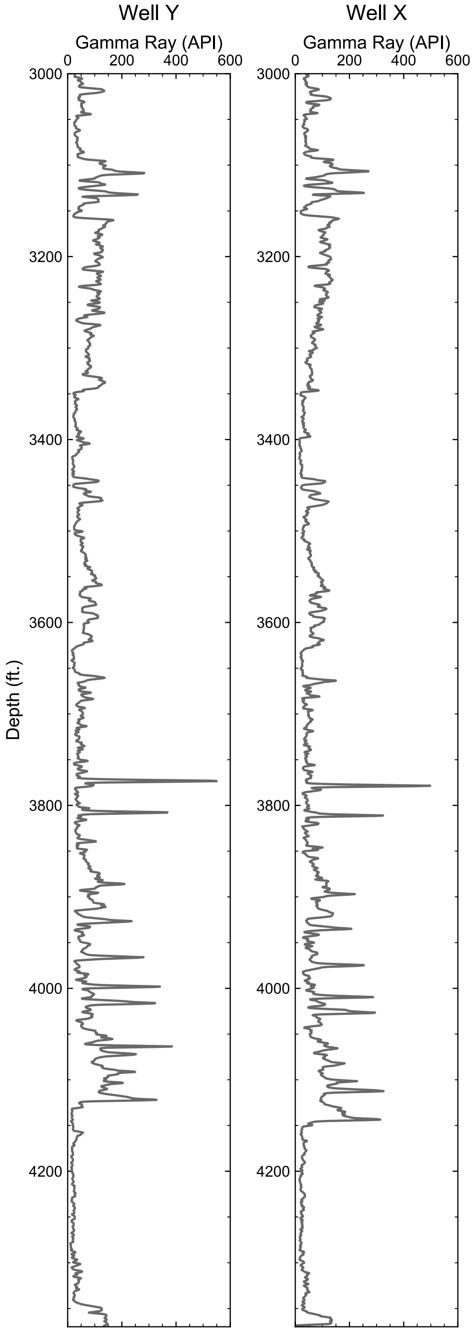


Figure 6.11 Well logs to be correlated

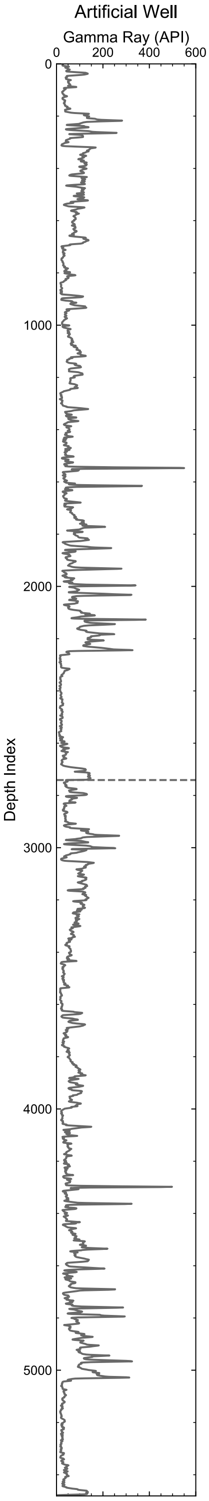


Figure 6.12 An artificial well log of length 5,482 indexes generated by connecting two well logs: Well Y and Well Z with a breakpoint (at dashed line).

Figure 6.12 shows that the artificial well log with 5,482 data points is generated from a connection of Well Y, having 2,741 data points, and Well X, having 2,741 data points. In this case, 2,741 is the breakpoint (indicated by a dashed line in the middle of the generated artificial well log) between Well Y and Well X. It is noticeable that depth index is also used in this artificial well log.

Once the artificial well log is generated, heuristic reference series selection can be performed. The selection is done by randomly choosing one index in each 0.5 standard deviation (SD) interval of the entire -5.0 to 5.0 SD span, thereby having 21 reference series in total. Although SD spanning from -5.0 to 5.0 is also used as they were in the example correlation 1, the value ranges in each standard deviation interval are different due to the fact that there are slight different in mean and SD of this artificial well lo

shows key statistical parameters and value ranges. As it is the case in example correlation 1, values from 348.16 to 549.99 is the range that is out of the SD span. The range, however, only has 10 data points so this should be as the data points only account for less than 0.2% of the total data points making the artificial well log. As indicated in the table, only intervals from -1.5 SD are used to get indexes and thus reference series as there is no data point when the value is below -1.5 SD.

Table 6.3 Well log value from -5.0 SD to 5.0 SD span in 0.5 SD interval span

Standard Deviation		Gamma Ray (API)		
Min	Max	Min	Max	
-5.0	-4.5	-187.87	-162.34	No data available
-4.5	-4.0	-162.34	-136.82	
-4.0	-3.5	-136.82	-111.29	
-3.5	-3.0	-111.29	-85.77	
-3.0	-2.5	-85.77	-60.24	
-2.5	-2.0	-60.24	-34.71	
-2.0	-1.5	-34.71	-9.19	
-1.5	-1.0	-9.19	16.34	Data available
-1.0	-0.5	16.34	41.86	
-0.5	0.0	41.86	67.39	
0.0	0.5	67.39	92.91	
0.5	1.0	92.91	118.44	
1.0	1.5	118.44	143.96	
1.5	2.0	143.96	169.49	
2.0	2.5	169.49	195.01	
2.5	3.0	195.01	220.54	
3.0	3.5	220.54	246.06	
3.5	4.0	246.06	271.59	
4.0	4.5	271.59	297.11	
4.5	5.0	297.11	322.64	
5.0	5.5	322.64	348.16	

Mean	67.39	API
SD	51.05	API
Max	549.99	API
Min	5.09	API

As depicted in Figure 6.13, most of the signal is within the SD intervals of -1.5 SD to 5.0 SD while several signal peaks are well over 5.0 SD. Once the intervals' boundaries are available, the next step is to randomly choose indexes in the intervals.

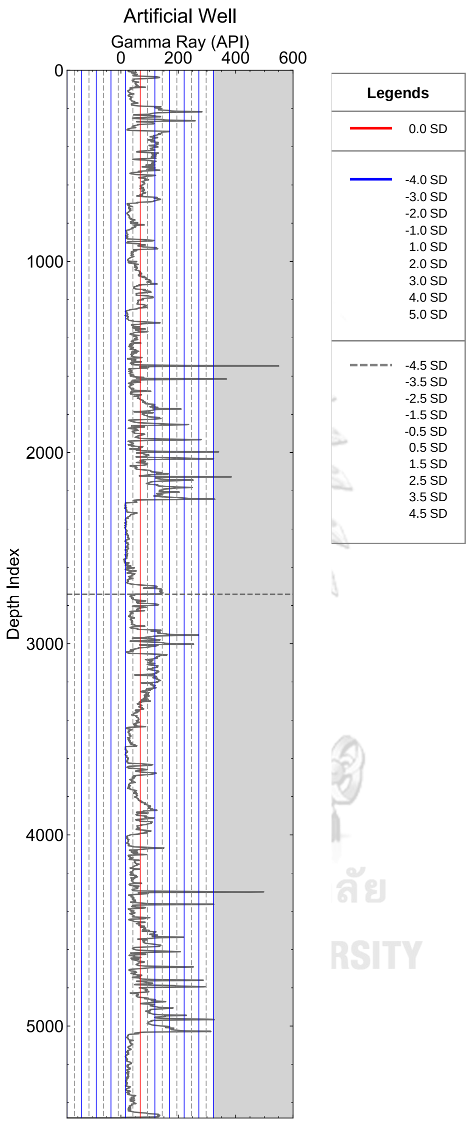


Figure 6.13 Artificial well log with SD boundaries

As suggested by Figure 6.14, the SD intervals that are less than -1.5 SD do not have any data point. Therefore, there is no index, and thus data window, chosen for those intervals. From the figure, there are 14 chosen indexes from Well Y (212 and 2080) and Well X (2243, 2264, 2441, 3067, 4300, 4363, 4495, 4505, 4536, 4537, 4944, and 5026), given that the breakpoint between these two well logs is 2174 and section length (window size) of 120.

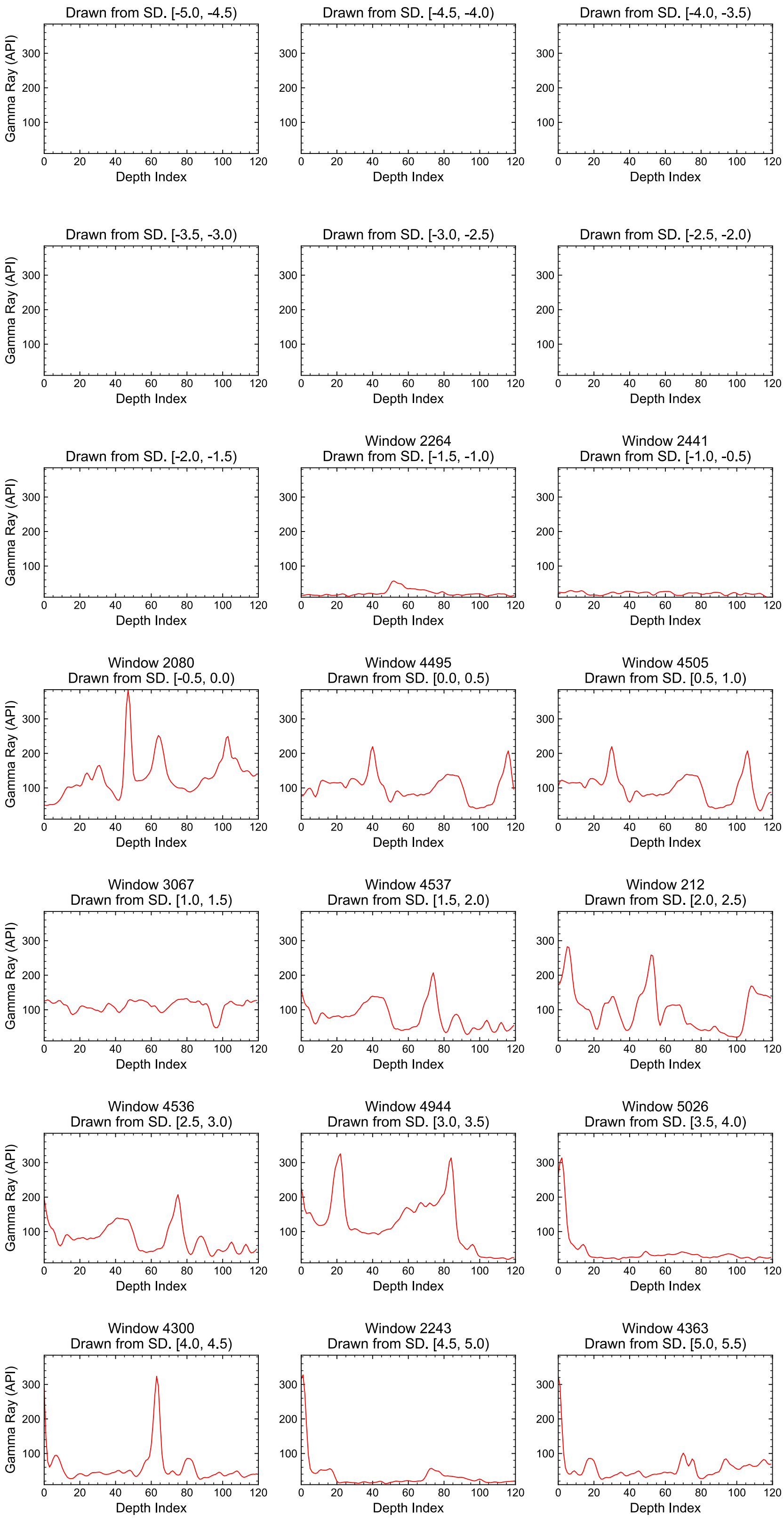


Figure 6.14 21 reference sections each randomly chosen from -5.0 SD to 5.0 SD of the artificial well log mean.

When reference series are ready, the next step is really the first step that uses MK. 14 reference series are fetched to MK in order for MK to calculate distance between each reference series and all other remaining data windows. For example, MK calculates distances between the reference series Window 212, covering index 212 to index 332, and other data windows such as Window 1 (the first data window), covering index 1 to index 120. The distance between reference series and the sliding window of the same window will be assigned with distance *infinity* instead of its real distance *zero* to prevent them from being promoted as a matched pair. There will be 14 sets of 5,362 (or 5,482 – 120) elements of the distance values as a result of this calculation of distances between the reference series and all other remaining data windows. Once the 14 sets of distance values are ready, MK then prepares information necessary to search for matched pairs. First, MK ranks the 14 sets of distance values from highest to lower standard deviation of the distance values. Given the distance-value dataset with highest standard deviation, then MK ranks the indexes from the index providing the smallest distance between the reference series and all other data windows to the index providing the largest distance between the reference series and all other data. To make these two steps more intuitive, Figure 6.15 outlines the steps necessary to generate a set of indexes from lowest to highest distances with the reference series which has the highest SD. of the distance values as aforementioned.

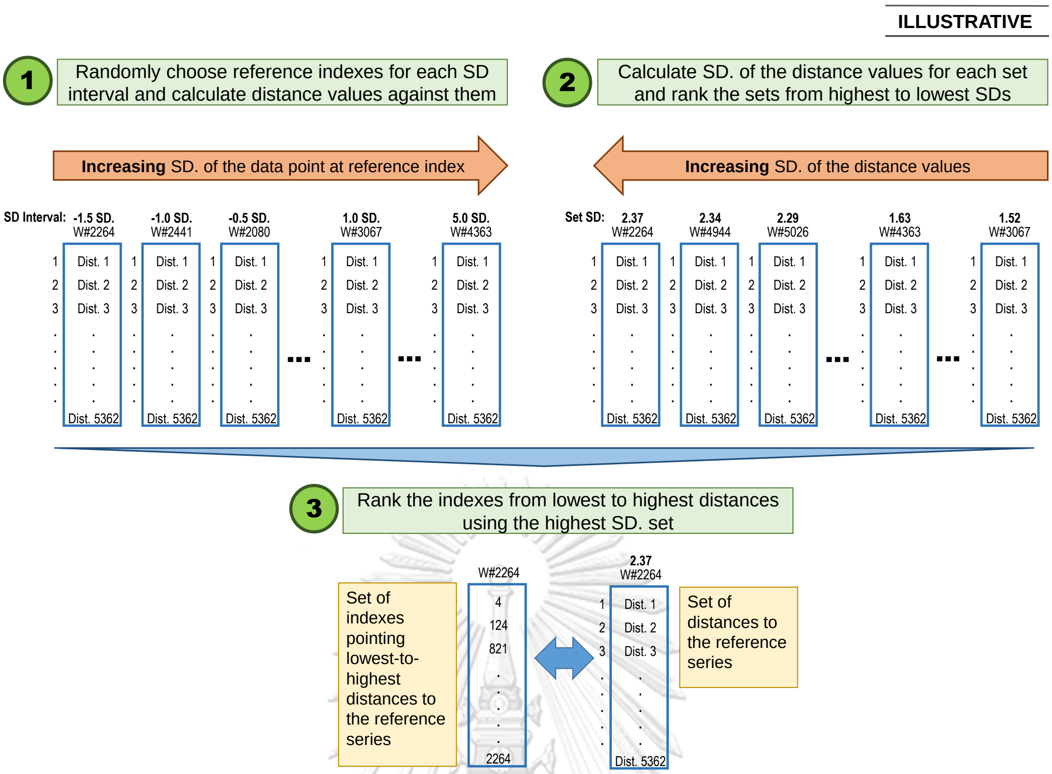
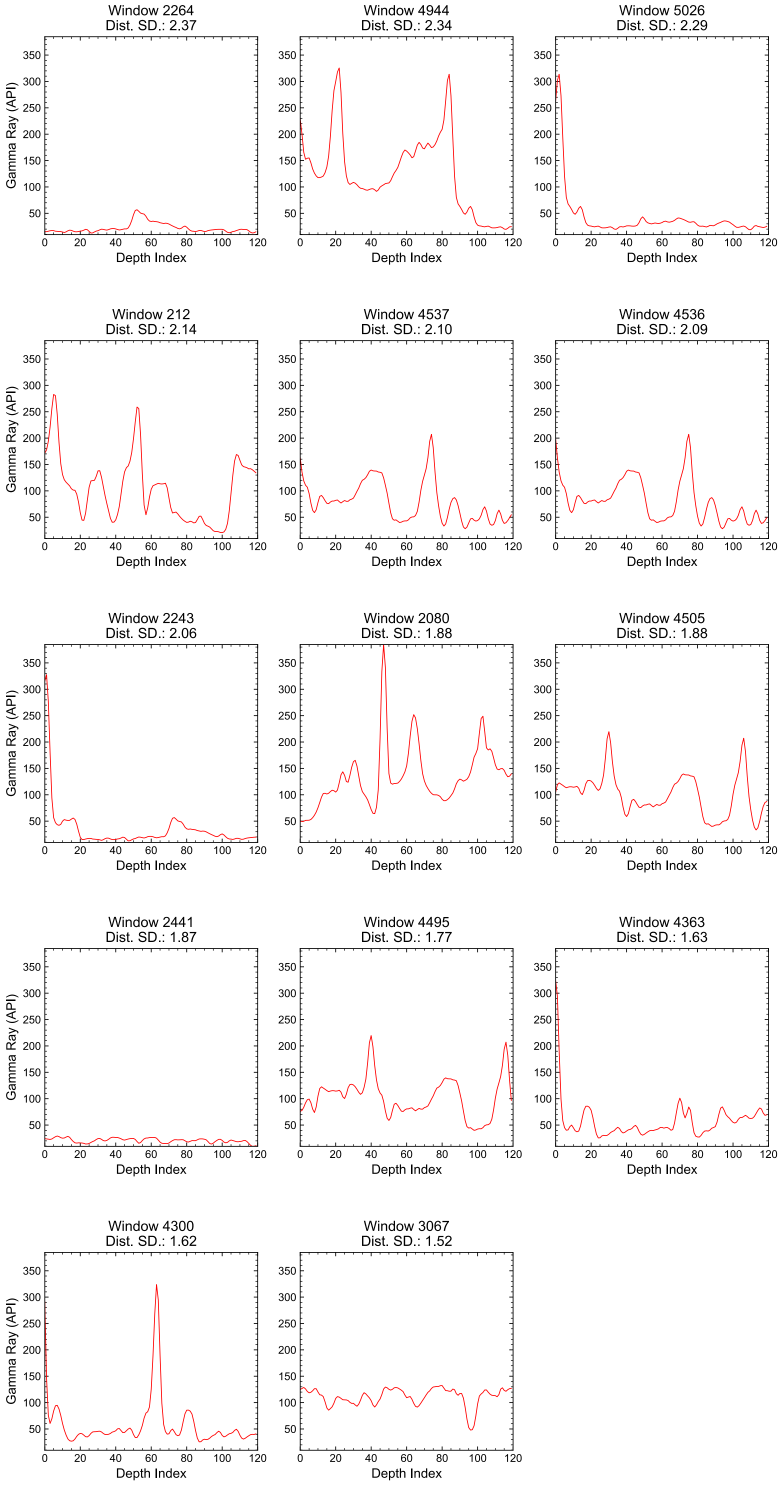

ILLUSTRATIVE
1
Randomly choose reference indexes for each SD interval and calculate distance values against them
2
Calculate SD. of the distance values for each set and rank the sets from highest to lowest SDs
Increasing SD. of the data point at reference index
Increasing SD. of the distance values
SD Interval:
-1.5 SD. W#2264
-1.0 SD. W#2441
-0.5 SD. W#2080
1.0 SD. W#3067
5.0 SD. W#4363
Set SD:
2.37 W#2264
2.34 W#4944
2.29 W#5026
1.63 W#4363
1.52 W#3067
Dist. 1
Dist. 2
Dist. 3
Dist. 5362
3
Rank the indexes from lowest to highest distances using the highest SD. set
W#2264
4
124
821
2264
2.37 W#2264
Set of indexes pointing lowest-to-highest distances to the reference series
Set of distances to the reference series

Figure 6.15 Steps used to generate a set of indexes pointing to lowest-to-highest distances based on the set with largest SD.

Figure 6.16 shows the ranks of 14 reference series. As aforementioned, the ranks are from calculating the standard deviations of the distance values in which each reference series is used in distance calculation of all other data windows and sorting the standard deviations from highest to lowest number. For example, distance values calculated based on reference series from Window 4397 are mostly varied while distance values from Window 2112 are much less varied. Therefore, the standard deviation of the distance values from Window 4397 is much higher than that of from Window 2112. The plots in this figure simply show the plots of the reference series.



Window 2264 Dist. SD.: 2.37

Window 4944 Dist. SD.: 2.34

Window 5026 Dist. SD.: 2.29

Window 212 Dist. SD.: 2.14

Window 4537 Dist. SD.: 2.10

Window 4536 Dist. SD.: 2.09

Window 2243 Dist. SD.: 2.06

Window 2080 Dist. SD.: 1.88

Window 4505 Dist. SD.: 1.88

Window 2441 Dist. SD.: 1.87

Window 4495 Dist. SD.: 1.77

Window 4363 Dist. SD.: 1.63

Window 4300 Dist. SD.: 1.62

Window 3067 Dist. SD.: 1.52

Figure 6.16 Reference series of the highest-to-smallest ranks based on standard deviation of distance values.

The next step is to search matched pairs based on triangular inequality concepts after the set of indexes to data windows, having lowest to highest distances, is available. Starting from setting best-so-far with a starting distance value, each subsequence search introduces a prospective pair to check if the pair's lower bound is smaller than the current best-so-far. If the pair has its lower bound really lower than the current best-so-far, then the pair has to undergo further test whether the pair's distance is also not greater than that of the current best-so-far. The pair can be called a matched pair (or motif pair) if the pair passes the two criteria mentioned.

Figure 6.17 to 6.20 show four matched pairs returned from MK using the 14 reference series and the section length of 120. As it is the case for example correlation, it is assumed that Window 1557 – Window 4632 pair's best-so-far must be D > 5.74 before this pair is being considered. When this pair comes in, MK checks whether the lower bound is less than best-so-far. Since lower bound (1.23) of the pair is less than best-so-far (D), an additional check, which is to confirm if the distance between the pair is less than current best-so-far, must be performed. It becomes apparent that the pair can be selected as a matched pair since its distance is 5.74, which is smaller than D. The next matched pair (Window 1487 – 4239) is in the third rank. The current best-so-far is 5.74, which is the distance of the previous pair. This third-rank pair is considered a matched pair because it also fulfills two criteria (its lower bound of 0.12 and its distance of 2.92 are both less than best-so-far of 5.74) as is the Window 1557-Window 4632 pair. The latter two matched pairs (Window 1184 – Window 3932 and Window 124 – Window 2861) also follow pass the two criteria (their lower bounds and distances are less than best-so-far of 2.92 and 2.39 respectively) in the same manner. At the end of this whole process, the best-so-far gets updated by the distance between Window 124 – Window 2861, which is 1.62.

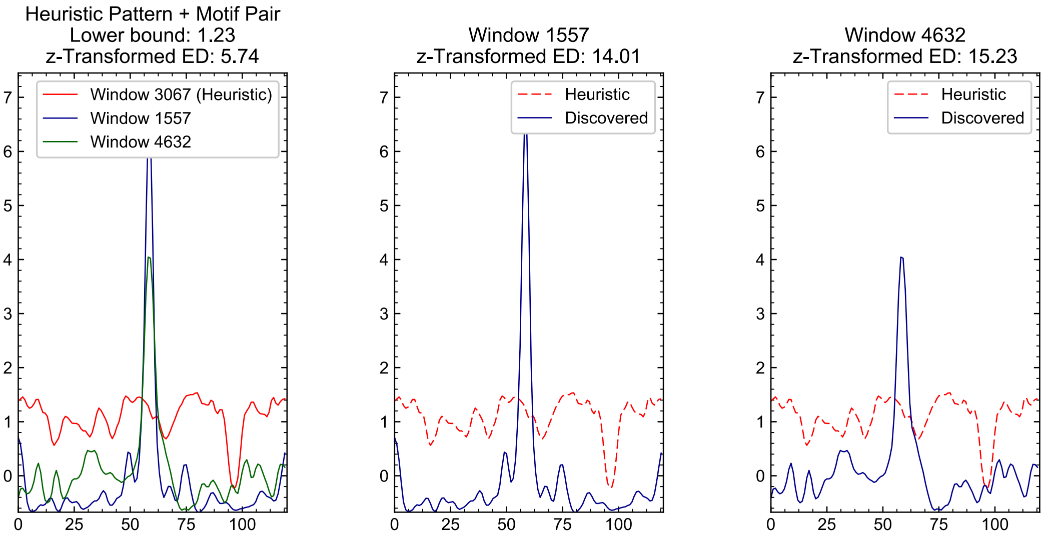


Figure 6.17 The forth-rank matched pair (motif pair) of Window 1557 and Window 4632 with reference series of Window 3067.

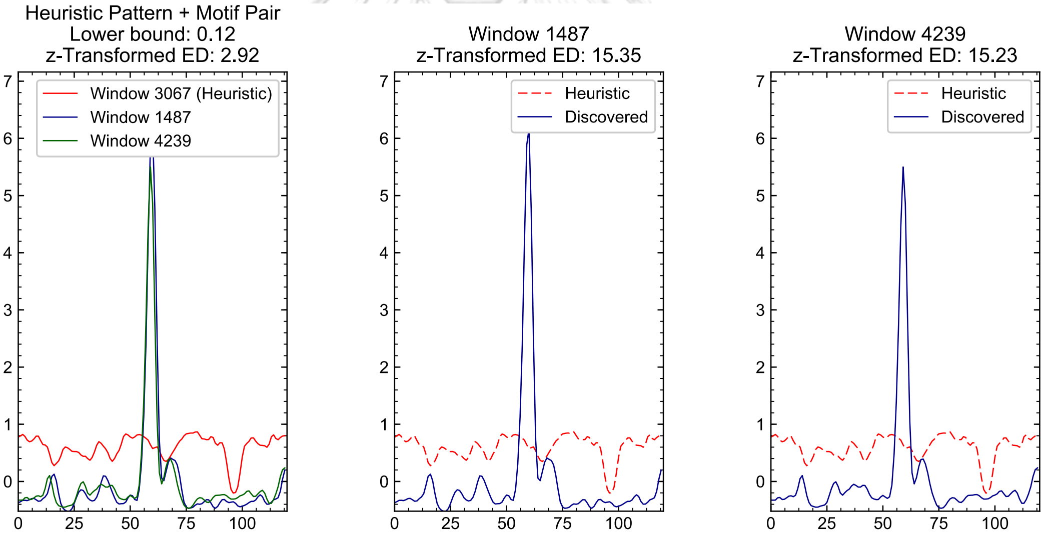


Figure 6.18 The third-rank matched pair (motif pair) of Window 1487 and Window 4239 with reference series of Window 3067.

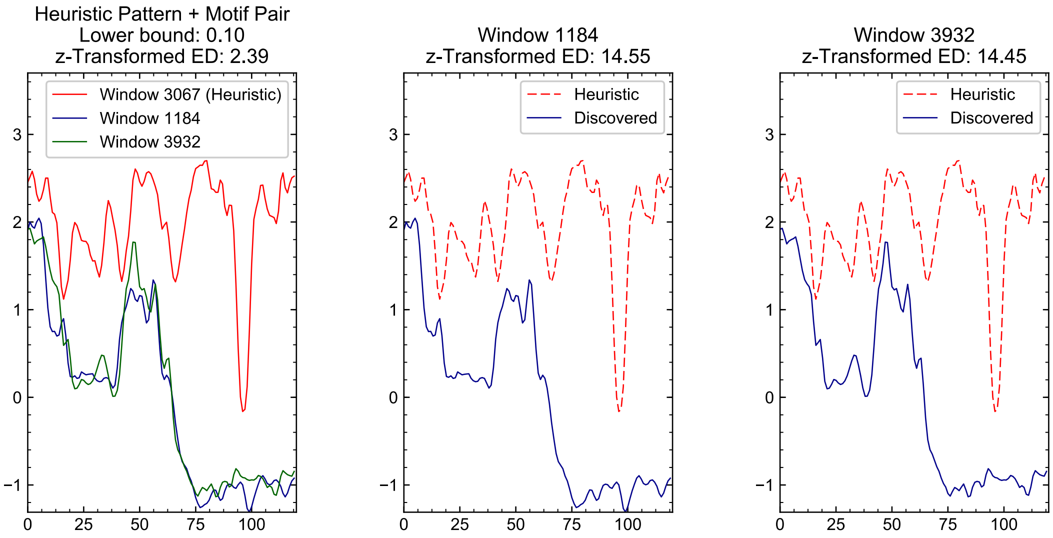


Figure 6.19 The second-rank matched pair (motif pair) of Window 1184 and Window 3932 with reference series of Window 3067.

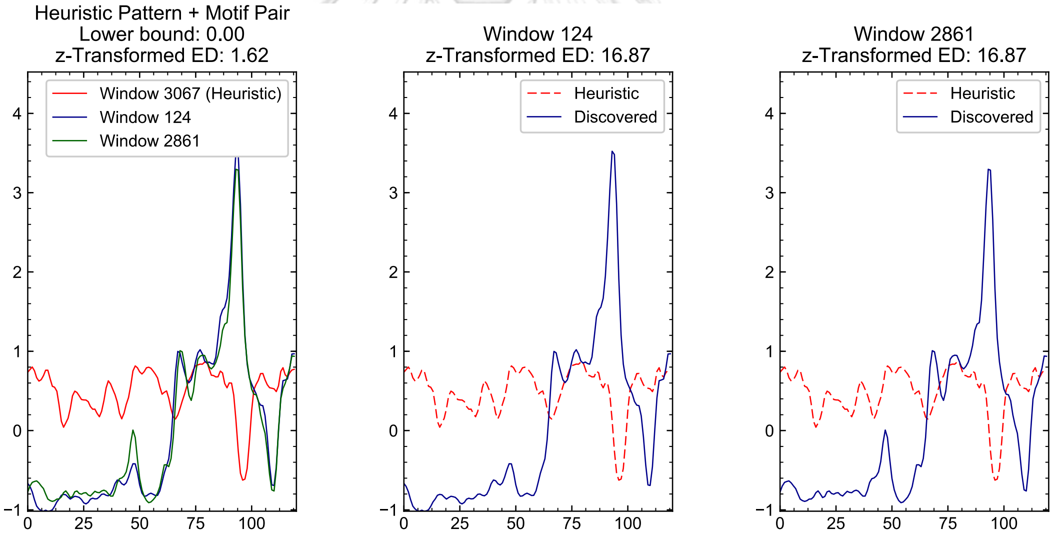


Figure 6.20 The first-rank matched pair (motif pair) of Window 124 and Window 2861 with reference series of Window 3067.

While MK is applicable to well log correlation task, human interpretation is sometimes still required, especially when there is a section mismatch when locations on the depth scale is considered. As shown in Figure 6.21, there is a mismatch between one section at 3,778.5 ft on Well Y and the other section at 3,945.5 ft on Well X. These two sections are the sections from the forth-rank matched pair shown in Figure 6.17.

This is due to the fact that MK perform Z-transformation on all the data series in all distance calculation as part of a technique to perform scale invariance, considering only similar shapes of the signals while the signals share similarity in Z-transformed space, comparisons. In this case, a human interpretation may be needed to rule out these kind of mismatch. The corrected well log correlation after a human intervention is shown in Figure 6.22. Table 6.4 shows that locations where the wells are correlated. Section 4 is discarded due to the fact that the correlation is not correct since both sections from each well should have been closer over the depth. This process requires human effort as MK does not have the information.

Table 6.4 Example correlations and the parameters used

Well 1	Well 2	Section Length	Section	Well 1 Index	Well 2 Index	Pair Distance	Status
Well Y	Well X	120	1	120	124	1.62	Kept
			2	1184	1191	2.39	Kept
			3	1487	1498	2.92	Kept
			4	1557	1891	5.74	Discarded

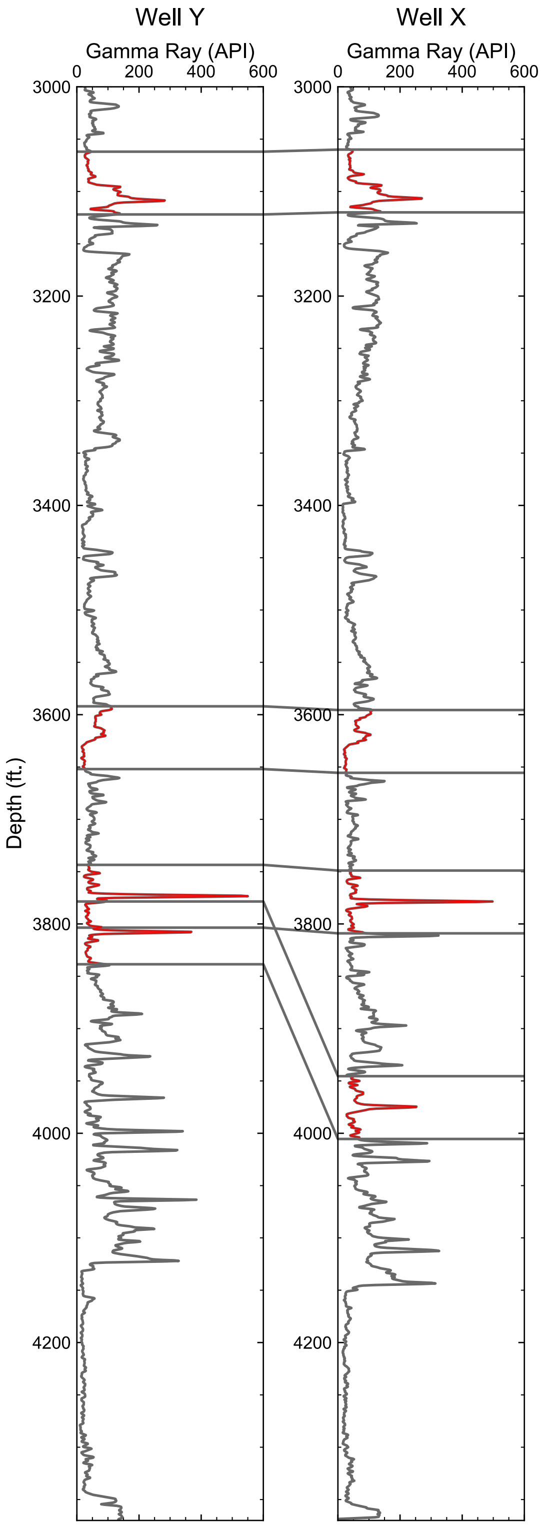


Figure 6.21 Well log correlation mismatch.

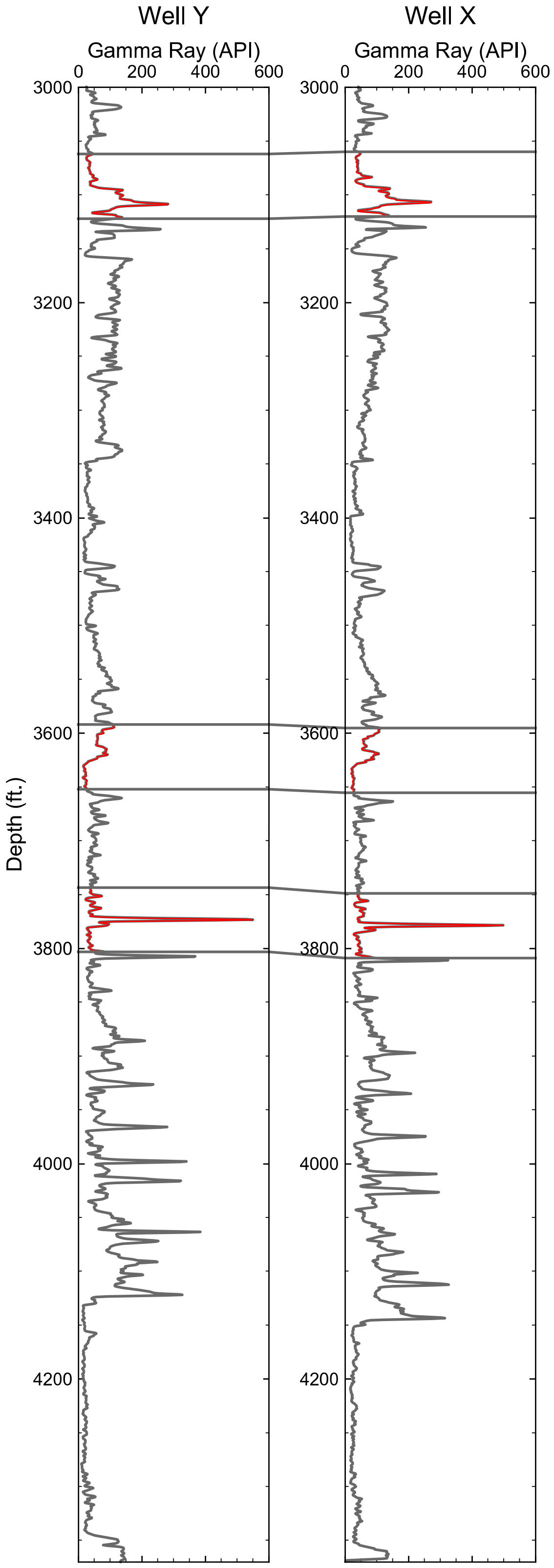


Figure 6.22 Corrected well log correlation after a mismatch is excluded

6.3 Multiple Well-to-Well Log Correlation using Pattern Discovery

As can be seen from the two previous examples, pattern discovery can be applied in well-to-well log correlation task given section lengths are known. This example shows how multiple well-to-well log correlation can be achieved using pattern discovery approach. Figure 6.23 and Figure 6.24 are well-to-well log correlations both performed using Well Y as a shared well in order to correlate across three wells from Well X to Well Y (shared well) and then Well Z.

The challenge in multiple well-to-well log correlation using pattern discovery approach is that there is no guaranteed that the motif pairs from the nearby wells will correlate with the shared well at the same locations to make perfect joining across multiple wells. As shown in Figure 6.23 and Figure 6.24, Well X and Well Z do not join Well Y at the same indexes. Therefore, human intervention to manually adjust the joining locations on Well Y are required. Depending on each data window, joining location may be assigned based on matched peaks and troughs across multiple wells. Figure 6.25 shows the result after human intervention to manually align the correlated section from well-to-well log correlation between Well Y and Well Z and well-to-well log correlation between Well Y and Well X. Several previously correlated sections are also removed as they are not available in all well logs to be correlated.

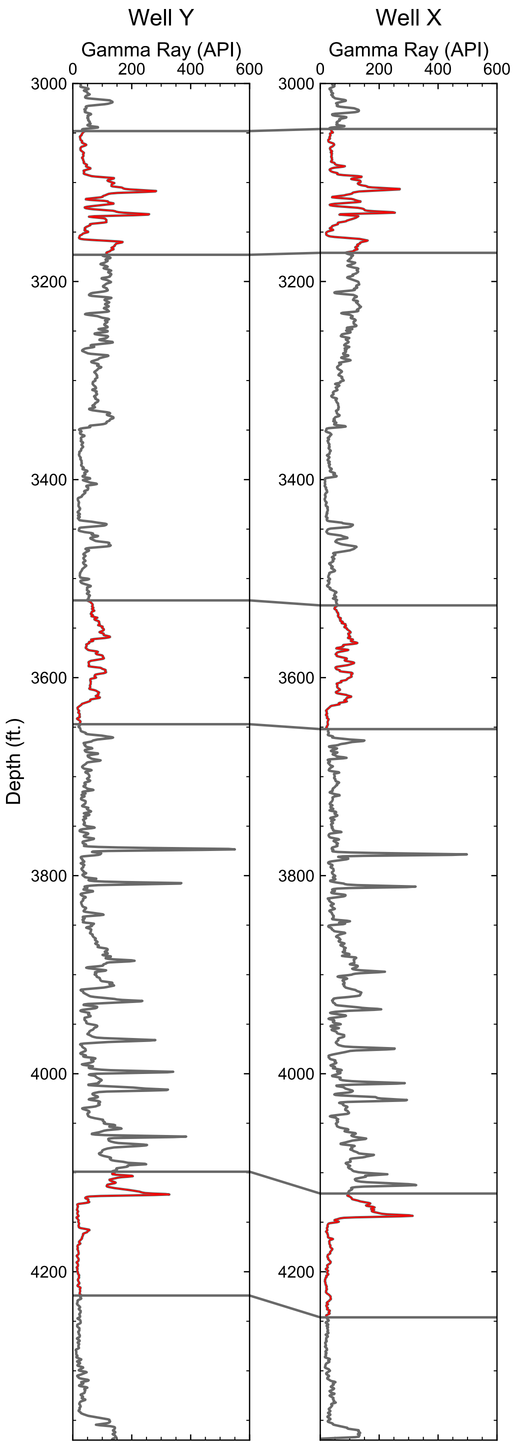


Figure 6.23 Well-to-well log correlation using pattern discovery approach (left section)

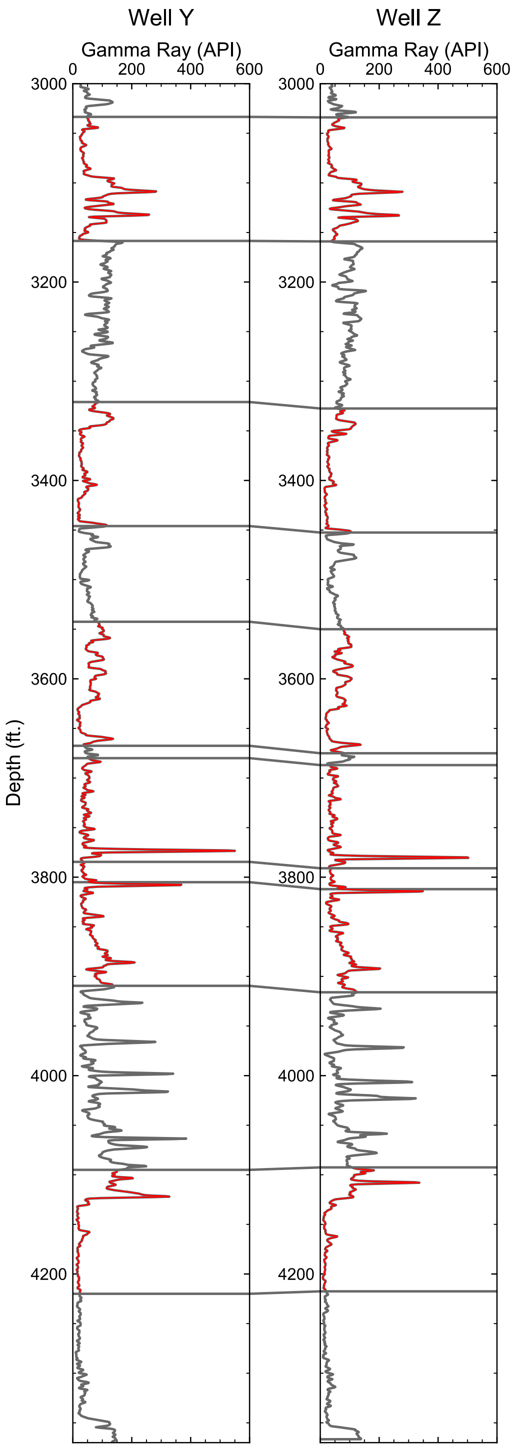


Figure 6.24 Well-to-well log correlation using pattern discovery approach (right section)

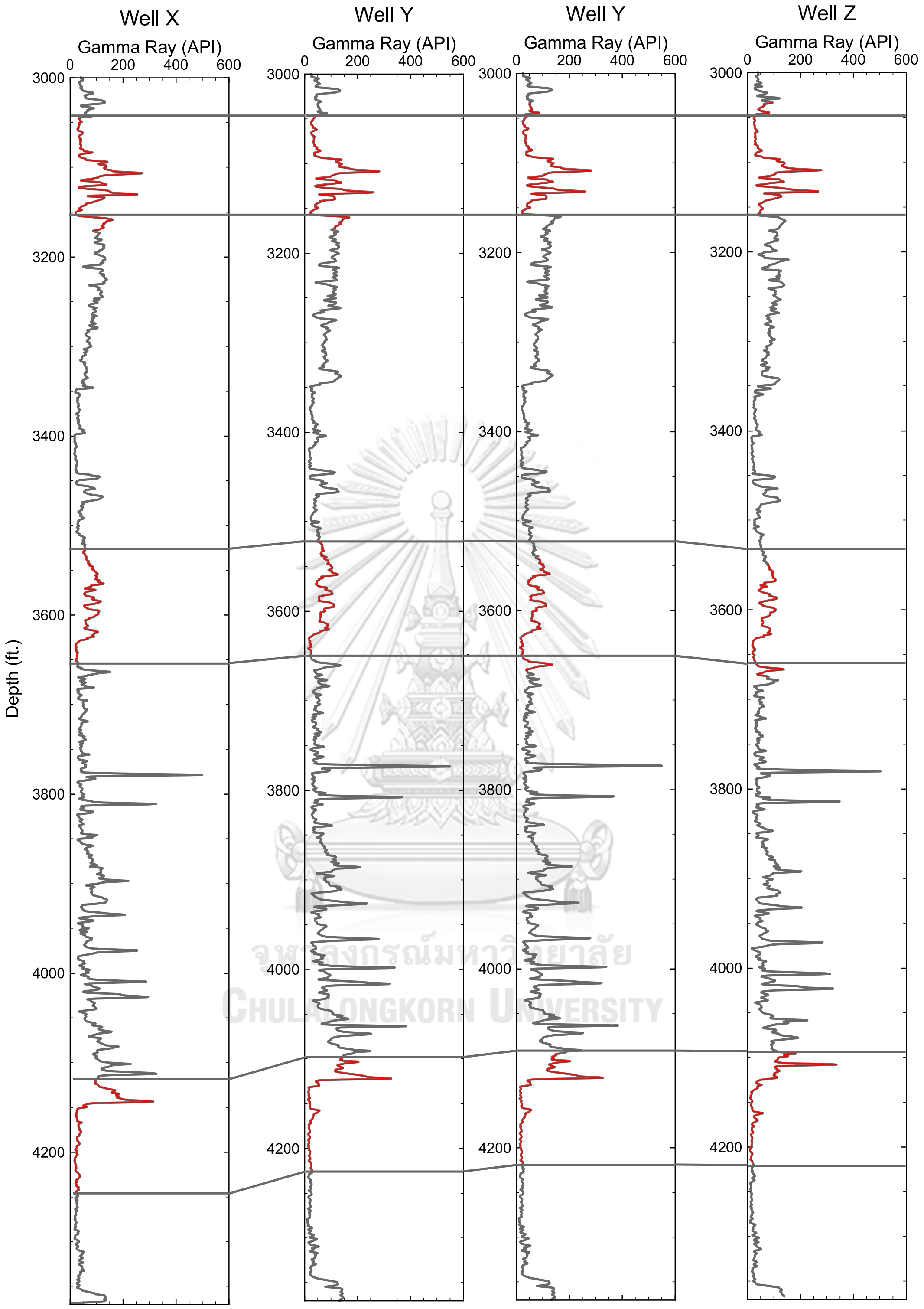


Figure 6.25 Multiple well-to-well log correlation using pattern discovery approach with human intervention to align some of the correlated sections.

6.4 An Integrated Approach to Well-to-Well log Correlation

In real field application, it is possible to jointly use both pattern discovery and pattern matching to form a comprehensive and integrated well-to-well log correlation. Pattern discovery is suitable when only minimal information about the field is known. In the early stages of field development, only section length is known as the information may come from other data sources such as seismic survey. In the latter stages of field development, pattern matching may be more suitable as it can easily be used to correlate multiple wells (by connecting multiple wells as a single well and using a pattern of interest with NNS or multi-resolution analysis). The steps outlined below may be used as a guideline for integrated well-to-well log correlation.

1 When the first two well logs are available, use pattern discovery to find inherent patterns shared by the two wells.

2 Repeat step 1 using different section lengths (window size) in order to cover patterns of varying lengths

3 Once well logs from different wells become available, use pattern matching to check whether the patterns extracted from step 1 and 2 (pattern discovery) are available in the other wells nearby the first two wells. Changes in probability of best match can also be analyzed as the well logs from the wells that are far from the first two wells become available.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

In this chapter, conclusions and recommendations on how time series pattern matching and pattern discovery can be used in well-to-well log correlation are discussed.

7.1 Conclusions

- This study presents the processes applying pattern matching and pattern discovery approaches of time series data mining to perform well-to-well log correlation. The processes are aimed to be used to propose possible matches under specific conditions identified by each approach. Full automation of the processes still requires further continued study and development.
- A sliding-window technique is used to extract well log data for a given length (or window size). Each time the window moves by one index to ensure that all data window is included in finding a match.

7.1.1 Pattern Matching

- Euclidean distance can be used to provide an absolute distance between any two well log signals. This distance can be used to exactly determine how good a matching is. Euclidean distance, however, does not provide a very good matching result, especially when variations (such as signal shifting and insertion/deletion of some small portions) between the two signals are prominent.
- Different number of PAA blocks may give different best-match results. The higher number of PAA blocks may capture too much detail from the original signals while the lower number of PAA blocks may not sufficiently capture detail from the original signals. Both scenarios can lead to pattern mismatch due to improper level of details.

- Different numbers of SAX sections, which is the discretization of data values at discretized depths into different symbols (groups), may also give different best-match results. The higher number of SAX sections introduces variety of symbols, which emphasizes too much on insignificant differences, i.e., identifying small difference as different symbols. The lower number of SAX sections combines data having a wide range of values in the same group, resulting in inability to distinguish differences in data values. Both scenarios can lead to pattern mismatch due to improper identification of log value boundaries.
- Multi-resolution analysis allows us to find the most probable match from different levels of discretization. From different levels of PAA blocks and SAX sections, the data window that is best matched for the most frequent is identified as the most probable match.
- Levenshtein distance is far superior to Hamming distance in similarity measurement when there are local variations in well log signals because its comparison technique can offset the variations (to some certain extents) and thus the distance value is less affected by the variations. Hamming distance is, however, superior to Euclidean distance when local variations are in the well log signals since Euclidean distance is inflexible to this kind of situations.
- Based on the matching results under multi-resolution analysis, SAX can still be used to generate discrete series even though the well log signals do not follow perfect normal distribution.

7.1.2 Pattern Discovery

- Drawing reference series from each well log interval (SD interval) helps increase stability of the motif pairs discovered by MK.
- Depending on reference series used, MK can distinguish sections that should be correlated based on distance values of any two well log signals being correlated without prior knowledge of the signals.

- MK can help uncover similar sections under Euclidean distance across two different well log signals which can later be used as a guideline to perform multiple well-to-well log correlation.

7.2 Recommendations

- A further study is needed to tackle stability of the resulting well log correlation generated by MK because of randomness of the reference series used. Currently, it is required that the algorithm be executed several times to obtain the results when they are relatively stable.
- A further study is needed to observe the performance of MK algorithm in well-to-well log correlation using other distance metrics such as Hamming and Levenshtein distances.
- A further study is needed to generate well-to-well log correlation which is readily applicable for multiple well-to-well correlation by calculating global distance of all well logs given the same or very close sections.

REFERENCES

[1] J. H. Doveton, R. R. Charpentier, E. P. Metzger, and Kansas Geological Survey., *Lithofacies analysis of the Simpson Group in south-central Kansas* (Petrophysical series, no. no 5). Lawrence, Kansas: Kansas Geological Survey, 1990, pp. iii, 34 p.

[2] J. H. Fang, H. C. Chen, A. W. Shultz, and W. Mahmoud, "Computer-Aided Well Log Correlation," *American Association of Petroleum Geologies,* vol. 76, no. 3, p. 11, 1992.

[3] A. Chakraborty and D. Okaya, "Frequency-time decomposition of seismic data using wavelet-based methods," *Geophysics,* vol. 60, no. 6, pp. 1906-1916, 1995.

[4] N. R. Vega, "Reservoir Characterization Using Wavelet Transforms," Ph.D., Petroleum Engineering Ph.D. Dissertation, Texas A&M University, 2003.

[5] N. E. Huang and Z. Wu, "A review on Hilbert-Huang transform: method and its applications to geophysical studies," *Reviews of Geophysics,* vol. 46, no. 2, 2008.

[6] Y. Wang, "Seismic time-frequency spectral decomposition by matching pursuit," *Geophysics,* vol. 72, no. 1, pp. V13-V20, 2007.

[7] S. M. Luthi and I. D. Bryant, "Well-log correlation using a back-propagation neural network," *Mathematical Geology,* journal article vol. 29, no. 3, pp. 413-425, 1997.

[8] J.-S. Lim, "Reservoir properties determination using fuzzy logic and neural networks from well data in offshore Korea," *Journal of Petroleum Science and Engineering,* vol. 49, no. 3–4, pp. 182-192, December, 15, 2005.

[9] Z. Bassiouni, *Theory, Measurement, and Interpretation of Well Logs* (SPE textbook series, no. 4). 1994, pp. ix, 372 pages.

[10] D. J. Lineman, J. D. Mendelson, and M. N. Toksoz, "Well-to-well log correlation using knowledge-based systems and dynamic depth warping," in "Earth

Resources Laboratory Industry Consortia Annual Report," 1987, Available: http://hdl.handle.net/1721.1/75091.

[11] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing,* vol. 62, no. 3, pp. 531-544, 2014.

[12] F. Li, T. Zhao, Y. Zhang, and K. J. Marfurt, "VMD based sedimentary cycle division for unconventional facies analysis," presented at the Unconventional Resources Technology Conference, San Antonio, Texas, August, 1-3, 2016, 2016. Available: http://library.seg.org/doi/abs/10.15530/urtec-2016-2455478

[13] A. I. Fischetti, A. C. Fischetti, and A. Andrade, "Shale characterization and well correlation by competitive neural networks," presented at the 9th Intl Congress of the Brazilian Geophysical Society & EXPOGEF, Brazil, September,11-14, 2005, 2005. Available: http://library.seg.org/doi/abs/10.1190/sbgf2005-014

[14] C. Cassisi*, et al.*, "Motif discovery on seismic amplitude time series: the case study of Mt Etna 2011 eruptive activity," *Pure and Applied Geophysics,* journal article vol. 170, no. 4, pp. 529-545, 2013.

[15] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover, "Exact Discovery of Time Series Motifs," in *Proceedings of the 2009 SIAM International Conference on Data Mining*, 2009, pp. 473-484.

[16] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," *Data Mining and Knowledge Discovery,* journal article vol. 15, no. 2, pp. 107-144, 2007.

[17] J. D. Kelleher, B. MacNamee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Aata Analytics: Algorithms, Worked Examples, and Case Studies*. Cambridge, MA: The MIT Press, 2015, pp. xxii, 595 pages.

[18] W. A. Sutherland, *Introduction to Metric and Topological Spaces*, 2nd ed. (Oxford mathematics). Oxford ; New York: Oxford University Press, 2009, pp. xi, 206 p.

[19] R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell System Technical Journal,* vol. 29, no. 2, pp. 147-160, 1950.

[20] A. Apostolico and Z. Galil, *Pattern Matching Algorithms*. New York: Oxford University Press, 1997, p. 377 p.

[21] F. Naumann and M. Herschel, *An Introduction to Duplicate Detection*. Morgan and Claypool Publishers, 2010, p. 92.

[22] A. A. Mueen, "Exact Primitives for Time Series Data Mining," Ph.D., Computer Science, UC Riverside, 2012.

VITA

Chanchai Apiwatsakulchai completed his first degree in Computer Science (Second-class honor) from Kasetsart University, Bangkok, Thailand. He then joined professional environments working as an Information Technology (IT) consultant for IT consulting and system integration firms. He provided recommendations on how IT can help shape clients to be more competitive in highly competitive business settings through the use of appropriate technologies to well-recognized companies from varying industries in Thailand.

He has a growing interest in utilizing data to help businesses make better decisions. His special interest is also in upstream petroleum sector where he sees opportunities where modern data-driven decision making can help business operations in the sector to be more robust.

He currently works as a technology consultant for a global management consulting and professional services company. His activities involve providing recommendations on business transformation through the use of technology capability and his specialization is in resources industries, which cover businesses in energy, utilities, chemicals, and natural resources sectors.