The Pennsylvania State University
The Graduate School
College of Earth and Mineral Sciences

PATTERN MATCHING VIA SEQUENCE ALIGNMENT:
ANALYZING SPATIO-TEMPORAL PATTERNS AND THEIR DISTANCES

A Thesis in
Geography
by
Samuel Stehle

© 2013 Samuel Stehle

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2013
The thesis of Samuel Stehle was reviewed and approved* by the following:

Donna J. Peuquet  
Professor of Geography  
Thesis Advisor

Alexander Klippel  
Professor of Geography

Brent Yarnal  
Professor of Geography  
Associate Department Head  
Graduate Program Officer

*Signatures are on file in the Graduate School
ABSTRACT

In many fields of study, researchers seek patterns to provide insight into processes and phenomena. In particular, time provides a measure to indicate consistency and change. Finding significant temporal patterns thus remains a popular and necessary topic of research in many fields, geography included. This research adds to this scholarship by extending sequence analysis techniques to evaluate the significance of a single established pattern by confirming its existence under new conditions. This effort investigates two questions: How well does the pattern fit in these new conditions? In what ways does the pattern change from the expectation it presents?

By using a variation on the sequence alignment algorithm developed in computational biology this Masters thesis addresses these questions. Sequence alignment uses matrix representations and traversal to compare two sequences by modifying them to make them approximate one another, which can be visually represented. Further, sequence alignment generates a measure of the total modification done that can be used to quantitatively compare sequences of events.

This thesis extends sequence alignment to account for the properties of spatio-temporal sequences extraneous to the original intent of the algorithm. Sequence alignment, developed for analysis of the physical structure of DNA, places a base pair within every possible location within a sequence. In this research, sequence alignment is amended to instead analyze the temporal intervals between successive events. Additionally, unlike the physical structure of DNA in which one base pair can occupy a location in the sequence, events can occupy the same temporal unit. This Masters thesis introduces the Temporal Deviation Distance as a measure of the amount of temporal deviation between an expected pattern of events and empirical data. The output from this operation is color-coded to facilitate visual exploration of those deviations with respect to the times and locations in both sequences that they take place.
This thesis demonstrates the temporally-aware sequence alignment algorithm on a statistically derived pattern of diplomatic events taking place in Yemen between February 2011 to March 2012. This pattern is compared to a series of events in time periods previous to those from which it was derived. I find that patterns change as a function of the temporal distance between those time periods, and thus conform to normal distance decay models. A similar analysis is undertaken that holds constant time rather than space. The pattern derived from events in Yemen is compared to events in the same time period in eight different countries showing that spatial distance decay does not apply and further political and economic differences affect the flow of patterns.
# TABLE OF CONTENTS

List of Figures ........................................................................................................... vii

List of Tables .............................................................................................................. ix

Acknowledgements .................................................................................................... x

Chapter 1 Introduction ............................................................................................... 1
  1.1 Research Questions ........................................................................................... 3
  1.2 Important Distinctions Made in this Research ................................................ 4
  1.3 Geographic Precedence for Analysis of Spatio-Temporal Sequences: Torsten Hägerstrand ................................................................. 5
  1.4 Sequence Analysis, Through Sequence Alignment ............................................ 8
  1.5 Thesis Structure ................................................................................................. 9

Chapter 2 Literature Review ...................................................................................... 10
  2.1 Evolution of Sequence Alignment ..................................................................... 10
      2.1.1 Computer Science .................................................................................. 11
      2.1.2 Sociology ............................................................................................... 13
  2.2 Sequence Alignment in Geography ................................................................... 15
  2.3 Similar Strategies for Pattern Analysis ............................................................. 17
  2.4 Political Pattern Analysis .................................................................................. 18
  2.5 Summary ............................................................................................................ 20

Chapter 3 The Sequence Alignment Procedure ..................................................... 21
  3.1 Matrix Generation ............................................................................................ 22
      3.1.1 Elemental Similarity Matrix ................................................................... 22
      3.1.2 Traversal Matrix ..................................................................................... 24
  3.2 Matrix Traversal ............................................................................................... 26
  3.3 Representing the Aligned Sequences .................................................................. 28
  3.4 Quantifying Sequence Similarity ....................................................................... 29

Chapter 4 Sequence Alignment of Spatio-Temporal Patterns .................................. 31
  4.1 Fundamental Reconsiderations of Sequence Alignment .................................... 31
  4.2 Temporal Matrix Population and Traversal .................................................... 34
      4.2.1 Generating Optimal Values .................................................................... 35
      4.2.2 Generating Sub-Optimal Values ............................................................. 37
      4.2.3 Traversing the Matrix ............................................................................ 38
  4.3 The Temporal Deviation Distance (TDD) ......................................................... 43
      4.3.1 Additional Temporal Descriptors ......................................................... 45
  4.4 Visual Exploration of how the Pattern Deviates .............................................. 46
  4.5 Tools: Java Programming .................................................................................. 47
Chapter 5 Intersecting Sequence Alignment and Temporal Geography through Diplomatic Patterns in Yemen over Time ................................................................. 49

5.1 Generating a Spatio-Temporal Pattern using STempo ........................................... 49
5.2 Comparing the Pattern to Other Temporal Contexts ........................................... 51
5.3 Analysis of Yemen Patterns over Time ................................................................ 53
5.4 Discussion ........................................................................................................ 56

Chapter 6 Intersecting Sequence Alignment and Comparative Politics: Comparing Yemen Patterns over Space and Place ......................................................... 59

6.1 Precedence of a Political Pattern ..................................................................... 60
6.2 Comparing the Pattern to Other Spatial Contexts ........................................... 61
6.3 Results of Spatial Comparison ....................................................................... 63
  6.3.1 Temporal Deviation Distance ................................................................. 63
  6.3.2 Alignment View ..................................................................................... 65
6-4 Discussion ........................................................................................................... 67

Chapter 7 Discussion and Concluding Remarks ....................................................... 69

7.1 Contributions .................................................................................................. 69
7.2 Limitations and Future Work ......................................................................... 71
  7.2.1 Data Quality ......................................................................................... 71
  7.2.2 Temporal Scale .................................................................................... 72
  7.2.3 Algorithm Complexity ......................................................................... 73
7.3 Additional Future Work .................................................................................. 75
  7.3.1 Elemental Similarity ............................................................................ 75
  7.3.2 Spatiality .............................................................................................. 76
  7.3.3 Prediction ............................................................................................. 76
7.4 Conclusion ....................................................................................................... 77

Appendix: Pseudocode .......................................................................................... 78

References ............................................................................................................. 80
LIST OF FIGURES

Figure 3-1. Elemental similarity matrix – represents the binary relationship between each element in both sequences. Example recreated from (Needleman and Wunsch 1970)...

Figure 3-2. Generating a traversal matrix. The darker grey cell is the current cell being filled in. Its value is a function of its value in the elemental similarity matrix and the highest value in the next row or column. Example recreated from (Needleman and Wunsch 1970)...

Figure 3-3. Traversing the traversal matrix. The path demonstrated by arrows is the one-dimensional optimal path through the matrix. Darker grey cells represent matched elements, lighter grey are mismatches or choices that the algorithm must process. Re-created from (Needleman and Wunsch 1970)...

Figure 3-4. Three methods of displaying aligned sequences as explained by (Sankoff and Kruskall 1983). The trace shows which matched elements, the listing explains the edit operations in natural language, and the alignment shows where the sequences were extended to match one another...

Figure 4-1. Generating and traversing the matrix in my extended method.

Figure 5-1. A temporal pattern of diplomatic cooperation generated by STempo. Event types are represented by their numerical codes and event transition is given as an interval of days.

Figure 5-2. (a) Temporal edit distance and (b) alignment representation of expected pattern to Yemen data in 2009-2010. Additional metrics for understanding temporal deviation are included in the distance metric window. In the alignment window, green represents a matched pair of elements, blue represents mismatched elements, and light pink represents intervening events between expected events that are within the expected critical interval. The unexplored sequence is on top, while the expected pattern sequence is on the bottom.

Figure 5-3. (a) Temporal edit distance and (b) alignment representation of expected pattern to Yemen data in 2004. Additional metrics for understanding temporal deviation are included in the distance metric window. In the alignment window, green represents a matched pair of elements, blue represents mismatched elements, and red represents intervening elements that are outside of the expected critical interval. The unexplored sequence is on top, while the expected pattern sequence is on the bottom.

Figure 5-4. (a) Temporal edit distance for Yemen data in 1993-1994. (b) Temporal edit distance for Yemen data between 1986 and 1993. Political and social structure in Yemen took dramatic turns in 1990 (unification of North and South Yemen) and in 1994 (civil war). Additional metrics for understanding temporal deviation are included in the distance metric window.
Figure 5-5. (a) Temporal analysis of expected pattern from 2011-2012 to previous years. (b) Unweighted exponential decay curve. Because distance and similarity are inverses, these curves appear equivalent, and may indicate temporal decay as a concept equivalent to distance decay.

Figure 6-1. Graph of spatial distance to Temporal Deviation Distance. No clear correlation exists in the same way that temporal distance correlated positively with increasing TDD. Other non-spatial factors seem to be bigger indicators of pattern dissimilarity in country-level politics.

Figure 6-2. Alignments of expected pattern to Bahrain (a) and Qatar (b). Both Bahrain and Qatar share the same TDD of 15 days. As the alignment show in the red intervals preceding the ‘engage in negotiation’ (46) event in (a) and following the ‘demonstrate for leadership change’ (1411) in (b), most of the temporal deviation in these countries’ comparisons (as well as the other countries close to Yemen) occur between these two expected events.

Figure 6-3. Alignments of expected pattern to Sudan (a) and Syria (b). Both countries have a similar TDD (Sudan = 1, Syria = 3), but they share the fact that the deviation is negative, indicating a quicker-than-anticipated pattern occurrence. Additionally, the deviation occurs in the same interval between ‘make optimistic comment’ (13) and ‘reject plan/agreement to settle dispute’ (127).
LIST OF TABLES

Table 6-1. Countries compared to Yemen in 2011-2012 as a function of distance ............... 62
ACKNOWLEDGEMENTS

First I would like to thank my advisor, Dr. Donna Peuquet, for her support and constructive feedback during the writing process. She has pushed me when necessary, been receptive and encouraging every time I have sought a new direction, and been a purveyor of confidence amidst harsh realities during my first experience of graduate and professional life. Next I want to thank my committee members, Dr. Alex Klippel and Dr. Deryck Holdsworth, whose comments and edits from the proposal to the edit phases have become fundamental pieces of my work and of this thesis. I would like to thank the many graduate students at Penn State and elsewhere who have given me feedback about this thesis and the work behind it from its concepts to its design and presentation, especially Arielle, who has been patient in constantly critiquing my writing. Most importantly, my family – Laurie, Dave, and Carla, Joyce and Lee – continues to give me support, encouragement, and love from afar, having never seen any part of this finished thesis and knowingly sacrificing their opportunities to see me in doing so.
Chapter 1

Introduction

In generating meaning from the endless events that take place in the world, scientists often rely on finding patterns. Patterns and their recurrence indicate regularity and their sequencing indicates order. These and other properties of patterns play a large part in our ability to explain social and environmental processes. The multifaceted significance of patterns has long interested researchers concerned with the conceptualization of pattern sequences over time and space. Using a spatio-temporal framework for situating patterns, much work has been done in the areas of human and animal movement (Hagerstrand 1970; Laube, Imfeld, and Weibel 2005), weather tracking (Beard, Deese, and Pettigrew 2007), and the spread of disease (Waller et al. 1997) among others, though much understanding of spatio-temporal dynamics evades us (Peuquet 2001). This has been a focus of many within the field of geography; however, these concerns interest others in fields such as public health, development studies, climatology, and behavioral studies.

Geographic examples of pattern interpretation over space and through time are plentiful and indicate physical and social processes at work. Spatio-temporal patterns are perceived in one of two ways: moving entities whose successive locations in space are determined by a factor of previous location and a set of temporal constraints to their range of motion (ubiquitously associated with Hägerstrand’s “time geography” (Hagerstrand 1970)), and more abstract sequences that do not refer to a single moving actor, such as the diffusion of the use of new innovations (Hagerstrand 1967), and the historical becoming of people and places (Pred 1984).
These non-movement based spatio-temporal sequences suggest the recognition of patterns in constant real world processes.

Recent non-path focused pattern analysis research is generally devoted to discovering meaningful patterns in unstructured information. Economists and investors interested in predicting the stock market have been leading the charge in this area (Anderson et al. 2007). Although a common tool for finding and understanding patterns in complex spatio-temporal data, pattern discovery through data mining can only determine patterns as they exist in the analyzed dataset. Pattern matching is a complementary strategy that aims to confirm the existence of known patterns in unexamined data.

Pattern analysis techniques begin at the exploratory pattern identification phase. Recent trends in the collection and social analysis of Big Data (Torrens 2010) and Knowledge Discovery in Databases (Fayyad, Piatesky-Shapiro, and Smyth 1996) aim to separate the signal from the noise within a large dataset. Databases of geographic information present a complex problem for pattern discovery where spatio-temporal organization, projection, interpretation, and summarization are necessary to define the existence of patterns (Miller and Han 2009). Patterns are also theoretically-driven, in such applications as spatially-organized economies (Eaton and Lipsey 1982). These represent two extreme cases; the former a situation where a pattern is derived from one specific collection of information, the latter a situation determined not by observation, but theory. In both cases patterns are generalized by confirming it in new and unexplored information.

The derivation of patterns, both empirically and theoretically, comprise the process of pattern discovery. Pattern matching, by contrast, takes an established pattern and matches it to a dataset that is independent of the one used to derive it. It must also be flexible to account for deviations from the established pattern. An effective pattern matching technique must then be
able to communicate the degree of inexactness. To this end, pattern matching quantifies the “fit” of an expected pattern in the data and provides outputs to visually display the specific deviations.

This Master’s thesis develops and demonstrates a customized technique for matching known spatio-temporal patterns in large, unexplored datasets. Based on the sequence alignment algorithm created in computational biology for DNA sequence matching, this technique includes several extensions for the purpose of utilizing temporal intervals between events in an expected pattern. It generates two outputs that contribute to the goal of pattern matching; a) quantification of the “fit” of the pattern via calculation of the Levenshtein Distance (Levenshtein 1966), and b) visual representation of the aligned sequences after they have been following extended so as to approximate each other’s composition.

I demonstrate pattern matching using a strategy commonly referred to as sequence alignment. By inserting gaps into sequences without changing the order in which sequence elements occur, sequence alignment attempts to measure the similarity between sequences. I demonstrate several extensions to the sequence alignment process aimed at applying the strategy to spatio-temporal patterns and communicating the temporal distortion that results from alignment.

1.1 Research Questions

The purpose of this Master’s thesis research is to expand upon the sequence alignment method for grounding previously-derived patterns in real events. The sequence alignment technique provides this ability, though requires significant modification for use with spatio-temporal patterns of events. The modification of sequence alignment generates a novel method for finding patterns in events and communicating their fit through quantification and visual depiction of aligned sequences. Spatio-temporal patterns in geographical applications require
specific definitions and assumptions from sequence alignment. To accomplish the objectives presented here, this Masters thesis research asks:

1. How should the sequence alignment algorithm be extended in order to utilize the spatio-temporal properties of events and patterns?
2. How can a modified sequence alignment quantify, visualize, and identify discrete temporal differences between patterns of events?
3. What does an inexact match of spatio-temporal sequences indicate in geographic, historical, and political contexts?

This extended sequence alignment method will complement pattern discovery efforts through its ability to link abstract pattern sequences with real events.

1.2 Important Distinctions Made in this Research

I make an important distinction between sequences and patterns in this research. A sequence is a structure of connected events, which maintains the order of events. Temporal events are linear, and thus succeed each other through time. In contrast, a pattern emerges when a structure of composed connected events is repeated over time, and that structure is seen to be of significance to discernible processes. Further, the repetition of the same events in sequence overtime is what makes patterns significant.

Sequence analysis in geography is based on understanding spatio-temporal processes. Patterns have an implied repetition and reproducibility that means research into spatio-temporal processes often relies upon an identification of patterns to understand the regularity that accompanies many observable social, political, environmental, and economic processes. While
spatial temporal analysis often takes the form of understanding moving entities, patterns are also the result of events in situ. Because sequences of events are not constrained by the continuity of a single moving entity, multiple events that contribute to the formation of a pattern often occur in the same time step.

Event analysis requires very different conceptualizations of spatial and temporal dynamics and methods to understand them than the typical trajectory approach prevalent in geographic research. Torsten Hägerstrand’s influential work on spatio-temporal dynamics (Hagerstrand 1965, 1967) and his later, more well-known work on space-time paths (Hagerstrand 1970) have driven geographic research on space-time, and despite the analytical disparity between event approaches and trajectory approaches, this Masters thesis research is highly motivated by Hägerstrand’s ideas.

1.3 Geographic Precedence for Analysis of Spatio-Temporal Sequences: Torsten Hägerstrand

The temporal emphasis of this work bears significance to Torsten Hägerstrand’s seminal work on time-geography. Hägerstrand’s time-geographic model (Hagerstrand 1970) represents human movement as trajectories in space and time. The basic premises on which this model relies –that humans cannot be in multiple locations at one time, and that time is linear, making it impossible to return to any single point in time – make it extendable to many applications where movement is a factor. Time-geography is aimed at describing the constraints that human actors must consciously and subconsciously navigate when moving from place to place. Hägerstrand describes three types of constraints: 1) capability constraints, which limit human mobility based on the ability of the body, the tools available to extend the body’s capabilities, and bodily requirements, such as the need to regularly eat and sleep 2) coupling constraints, which constrain
movement by the necessity of interaction with other individuals, whose trajectories must in some way coincide with one another to generate information, and 3) authority constraints, which constrain movement through socially or legally defined boundaries. Time-geography has proven to be a necessary component to theoretical advancements of mobility studies for its spatio-temporal representation of movement trajectories.

Several assumptions on which Hagerstrand’s time-geographic theory relies prevent it from being a general process for understanding sequences of geographic events. Most important is time geography’s assumption that an individual actor performs a sequence of events by moving. His time-geographic theory links points of location and time in a trajectory through an individual actor that makes decisions based on previous points and the three constraints. Sequences of events are only connected when a pattern over space emerges. Inevitably, sequences of events, which do not usually detail the movement and behaviors of an individual actor, cannot be traced in the same manner as Hägerstrand’s space-time paths, which are derived from the diary logs, GPS systems, and narratives of the actors that document their paths through space and time. For that reason, a trajectory is not a suitable representation for a sequence of events, and constraints cannot be measured as effecting the transpiring of individual events.

This observation begs the question; what is the comparable analogy to the individual actor that connects events in a pattern? It is the means by which each event triggers the next. It may be the transportation of people who influence the occurrence of events, the communication system that enables ideas to spread, the opinions that are expressed about past events, and potentially many more. Torsten Hägerstrand also considered the spatio-temporal dynamics of more abstract forms of movement, such as the diffusion of technological innovations. He considered the dissemination of successful farming practices between homesteads in a region of Sweden for the spatio-temporal diffusion of those who adopted certain techniques (Hagerstrand 1965, 1967). Through these analyses, Hägerstrand attempted an analysis of motion not of objects
over space and through time, but the spread of information-sharing through public and private channels. These are important models for considering the spatio-temporal dynamics of diffusion beyond that of moving object paths. This Masters thesis research takes a similar route to understanding spatio-temporal sequences and the processes that generate them.

Distance decay, as conceptualized by Hägerstrand, is an analytic concept that measures the movement of actors and objects as well as spatial diffusion. However Hägerstrand examined many other space-time concerns other than movement. The assumptions that guide the spread and adoption of innovations, for example, are unique and cannot be applied to the analysis of sequences of events. Hägerstrand’s models of innovation diffusion rely on specific rules that do not conceptually apply to events and developing patterns of events, such as how and when successive states of the pattern follow one another. Event sequences do not contain a priori understandings about the connections between events as Hägerstrand’s diffusion of innovations does, which means that distance decay as well as temporal expectations regarding the transition between pattern elements cannot be assumed.

Hägerstrand’s work contributes both precedence and contradiction to my efforts. His work has become synonymous with representation of movement through time and space, hence the ubiquitous description of his work as “time geography.” Where a large portion of recent work has found novelty in using Hägerstrand’s time-geographic representation with the increased affordability and accuracy of mobile tracking technologies like Global Positioning Systems (Miller 2005), this thesis extends and challenges his spatio-temporal theories based on movement of people and information by applying them to an analysis of event patterns rather than movement trajectories. By using existing tools and processes alongside Hägerstrand’s time-geographic theoretical precedence, this project advances his framework and also separate methods into the spatio-temporal domain.
1.4 Sequence Analysis, Through Sequence Alignment

This Masters thesis research focuses on one particular method, sequence alignment. Sequence alignment has become a common tool outside of geography to compare sequences, yet holds the potential to be a useful spatio-temporal analysis tool for geographers. Initially developed in biology and computer science, the sequence alignment algorithm compares two sequences with emphasis on the order of elements at similar positions within them. Sequence alignment uses matrix representations to align two sequences as closely as possible by extending portions of each sequence where necessary (Needleman and Wunsch 1970). Each unit of extension that the algorithm performs contributes to the edit distance – a measure of dissimilarity between compared sequences (Levenshtein 1966). Sequential order is preserved through the aligning of sequences, even while the sequences become distorted from their original structure, making this an ideal method for analyzing temporally-organized patterns.

However, not until recently have researchers outside of biology and computer science utilized the approach to compare sequences through time. Sociological applications including individual career paths and analysis of historical social movements helped generate a discussion about temporal sequence comparison. In response to some pushback concerning the suitability of the method for non-DNA-based sequences (Wu 2000), sequence alignment has undergone many modifications to fit new applications, especially with an emphasis on time. Limited research on spatio-temporal movement sequences that have been the focus of geographic research that utilizes sequence alignment have been largely limited to transportation studies (Wilson 1998). However, the utility of sequence alignment for spatio-temporal sequences of events has not been considered.

Sequence alignment began as an attempt to compare sequences of linearly-linked elements like DNA sequences and character strings (Sankoff and Kruskall 1983). Spatio-temporal
patterns lack this structure for two reasons: first, they are composed of events, ordered by temporal intervals, rather than physical structure, creating a sort of uncertain sequence. Second, the intervals between events in a pattern need not be successive as events often occur simultaneously with respect to time. In this Masters thesis, I introduce several extensions to the sequence alignment algorithm developed by Needleman and Wunsch (Needleman and Wunsch 1970) aimed at creating a method that compares patterns comprised of spatio-temporal events. These extensions, implemented in Java, include 1) temporally-expanded matrix calculation capabilities that includes consideration of simultaneous events, interval searches, and reorder-able columns 2) calculation of a modified Levenshtein Distance to account for and communicate temporal pattern deviation, and 3) visual exploration of aligned sequences which emphasizes temporal distortions rather than matched elements. These extensions generate a useful method for analyzing sequences through time and space.

1.5 Thesis Structure

The remainder of this thesis develops as follows: Chapter 2 provides an historical overview of sequence alignment and its applications and advancement over time. Chapter 3 explains the sequence alignment process in more detail, leading into Chapter 4, which describes the modifications necessary to perform sequence alignment on patterns of spatio-temporal events. Chapter 5 diagrams and discusses a case study examining patterns of political transition in Yemen over time. Chapter 6 highlights an additional case study examining the same Yemen pattern as it varies through space. Finally, Chapter 7 concludes with a general discussion and potential future directions for analyzing spatio-temporal sequences using the alignment technique.
Chapter 2

Literature Review

The objective of the research presented in this Masters thesis is to develop and demonstrate a method for grounding spatio-temporal patterns in large, complex collections of real world events. I start with a review that follows the evolution of the sequence alignment algorithm through time. This historical analysis outlines the impacts that the method has had on many fields and the developments necessary to continue impacting sequence analysis. I believe my efforts contribute to this development. I then conclude by reviewing several alternative processes for understanding sequences and patterns whose abilities fall short of the requirements of matching patterns of spatio-temporal events.

2.1 Evolution of Sequence Alignment

Sequence analysis has a rich history of use in many fields. The primary objective of the concept and its derivatives is to assess the similarity between two or more sequences of elements. Sequence comparison methods originated with the goal of matching words and sentences in large texts for decoding and spelling correction. Later, biology sought a method to compare DNA sequences for evolutionary similarity, which spurred the creation of the first algorithm for assessing sequential similarity using an alignment strategy. The sequence alignment algorithm now commonly used within biology and gaining popularity elsewhere was introduced by Needleman and Wunsch (Needleman and Wunsch 1970). Applications of this approach to temporal sequences in sociology have paved the way for its use on data with both temporal and spatial characteristics, and the extensions to the process that are presented here.
2.1.1 Computer Science

Perhaps the most important aspect of the seminal work on sequence alignment came from computer science for the comparison of sequences and for quantifying the similarity of given sequences. Vladimir Levenshtein conceived of an edit distance for the field of natural language processing to compare strings and sequences of strings (Levenshtein 1966). A string is a sequence of characters whose composition gives it its meaning in the form of words or codes. Levenshtein’s measure of similarity is a literal interpretation of the amount of change necessary to make two sequences approximate one another as accurately as possible. This involves the operations of insertion, deletion, and swapping elements in one or both of the compared sequences. Their similarity is then measured by the amount of change performed on both sequences. Quite simply, the more distortion that a sequence requires to approximate the composition of another, the less similar they are. The more intuitive interpretation is the more change done – the more gaps inserted into the sequences – the larger the “distance” between the sequences. Distance here implies an inverse-similarity, or dissimilarity, so interpretation of the edit distance as derived from sequence alignment requires understanding the relationship between the quantitative result of the algorithm and the similarity between compared sequences.

The edit distance has been used in the context of string comparison to detect errors in known stings for tools such as spell-checkers (Wagner and Fischer 1974). The process for automatically detecting and correcting errors in spelling has been recognized as relevant to pattern detection more broadly (Blair 1960). Sequence alignment provides more to this process through its ability to generate a list of similar terms based on the similarity they share with the misspelled word (Wagner 1983). Though its complexity has been difficult to overcome for some tasks (Wang and Jiang 1994), sequence alignment remains a common strategy in computer science. The edit distance, sequence alignment, and their derivatives remain among the most
commonly used methods for solving the problem known as string-to-string correction (Wagner and Fischer 1974; Kukich 1992). The complexity of sequence alignment only becomes an important factor when it is used on larger strings, such as whole documents. The amount of work done to improve the complexity of the algorithms built for the task (Gotoh 1982; Thompson, Higgins, and Gibson 1994; Notredame, Higgins, and Heringa 2000; Farrar 2007) gives a measure of the importance of string correction on large collections of text – such as XML files (Tekli, Chbeir, and Yetongnon 2009).

Although Levenshtein’s edit distance originated for the goal of string comparison in computer science, the strategy was also adopted for analysis of biological sequences, and thus the sequence alignment algorithm for DNA matching was conceived and new language for describing the edit distance established. Needleman and Wunsch (Needleman and Wunsch 1970) introduced the sequence alignment algorithm to compare the base pair sequence of two unique proteins. Comparing two sequences of DNA strands gives the edit distance a new meaning to the operations needed to change one sequence into the other. Insertion and deletion of genetic elements are evolutionary phenomena that can be traced through the use of the edit distance. Similar to the most common error detected by string correction processes, the idea of swapping sequence elements is also utilized in biology. Mutations – changes in the genetic structure of DNA that is noticed as different genetic elements that occur in the same position in two DNA sequences – can be discovered by aligning genetic sequences. Therefore, a tight link is maintained between the biological processes at work on genetic sequences and the operations used to compare them using sequence alignment. This theoretical linkage has provided some concern about the ability for adopting sequence alignment in other fields, especially sociology, where the edit operations do not provide the same fundamental coupling of real-world process and computational technique (Levine 2000).
Nevertheless, sequence alignment’s introduction in biology has given rise to several new uses for and generalizations of the algorithm. The basic structure of displaying evolutionary change is a phylogenetic tree. Phylogenetic trees graphically present species as nodes and relationships between them as edges. Many methods for calculating the structure of phylogenetic trees have been suggested, but evolutionary distance, calculated by the amount of alteration to genetic structure via insertion, deletion, and mutation over time is the most widely used (Fitch and Margoliash 1967). The concept of changing sequences of genes over time is entirely consistent with the edit distance, so it is no surprise there has been significant work merging evolutionary tree creation with edit distance measures (Sellers 1974, 1980; Saitou and Nei 1987) and sequence alignment (Needleman and Wunsch 1970; Kruskall and Sankoff 1983; Murata, Richardson, and Sussman 1985).

2.1.2 Sociology

Though time has played an implicit role in sequence alignment applications through its application to evolutionary distance, temporal sequences had not been analyzed using this method up to this point. The term ‘sequence’ implies order and succession, so it is a wonder that the only sequences compared using alignment and distance strategies through the late 1980s were strings of text and genetic base pairs, and not sequences of temporally-related events. It was at that time that the field of sociology was introduced to the sequence alignment method in the comparison of sequences of dance (Abbott and Forrest 1986). Andrew Abbott has remained the most vocal supporter of the method in the humanities.

Abbott has contributed to a theoretical and methodological basis for the sequence alignment method in the humanities and utilized the procedure in several social projects with a temporal focus. In an analysis of career paths of musicians, sequence alignment was used to
compare sequences whose consecutive elements are noted on the basis of time (Abbott and Hrycak 1990). Abbott’s review of sequence analysis introduced sequence alignment and its potential for sociological applications with a focus on temporal sequences under the heading of “optimal matching” (Abbott 1995), and soon thereafter analyses of career paths and life histories using this method became more common (see Halpin and Chan 1998; Blair-Loy 1999).

Sequence alignment has experienced some criticism concerning its suitability for certain goals. Much of the criticism and defense of sequence alignment for non-biological applications has come from the area of sociology. In perhaps the most detailed review of sequence alignment and optimal matching in sociology, Wu (2000) responded to Abbott and Tsay’s (2000) chronicle of the subject in an annual sociological review. In this response, Wu questioned the ability of sequence alignment to represent social phenomena (specifically noting hypothetical studies of job histories and pregnancy patterns because its roots in biological science and the availability of alternative tools developed within the social sciences for the task such as event history models. Wu was also concerned with Abbott’s claim that sequence alignment strategies can approach an understanding of not just units, but context (Abbott 1995). Levine also questioned the nature of sequence alignment as a model of social processes (2000), and a similar question was raised by Elzinga (2003). Abbott responded (Abbott 2000) to the critiques of Levine and Wu in a way that assuaged a good deal of the fear of foreign methods in sociology. A quick search of Google Scholar shows a four-fold increase of published works on the topic of sequence alignment in sociology following the issue of Sociological Methods and Research in which Abbott and Tsay, Levine, and Wu exchanged words. Abbott hypothesized a growing reliance on alignment strategies that would not come about for another decade or more (Abbott 2000), but a growing comfort with the method is being shown in sociology. Geographic applications suggest similar concerns, though the work in sociology paved the way for Geography’s subsequent adoption of the method for spatio-temporal analysis.
2.2 Sequence Alignment in Geography

Led by Abbott’s advances, researchers inside and outside of sociology have utilized sequence alignment for projects that focus on time. Spinoffs of the sequence alignment algorithm were created for temporal analysis, such as the dynamic time warping method to align sequences consisting of different time spans and intervals between successive events. Time warping works by interpolating values between key matched elements in order to extend similar sequences when they exist over different spans of time. This method has become crucial in applications such as speech recognition (Rabiner, Rosenberg, and Levinson 1978) and video sequence matching (Veeraraghavan, Roy-Chowdhury, and Chellappa 2005), where similar sequences of events occur over different spans of time. In geography, dynamic time warping has been demonstrated on animal population totals through time (Berndt and Clifford 1994) and rainfall events (McIntosh and Yuan 2005). This strategy would not only be impossible by interpolation of qualitative event sequences, but would generate unrealistic results by incorporating hypothetical events. Nevertheless, dynamic time warping has proven to be an important method for spatio-temporal sequence analysis and one of geography’s limited explicitly sequence-oriented methods.

Berndt and Clifford’s initial examination of dynamic time warping provides insight into theorizing sequence alignment for spatio-temporal applications (1994). Their use of the warping process includes several constraints based on assumptions of temporal sequences that are meant to decrease the complexity of the algorithm. While the restrictiveness of their constraints are partly a function of the computing power available to them at the time, the discrete numerical sequences in their application of predator and prey populations in yearly increments accommodated the constraints. Event sequences that cannot be reduced to quantitative totals, contain multiple events per time step, and have varying temporal intervals between successive events violate Berndt and Clifford’s assumptions and necessitate a more flexible analysis process.
Nonetheless, Berndt and Clifford make a theoretical leap in pattern matching by defining a pattern “template” and seeking to locate that template in unexplored data. By aligning the template with other data, Berndt and Clifford (1994, 364) are able to quantify the amount of warping and establish a “degree of fit.” While the time warping method is not perfect for the event sequences I am examining here, the method shares many concepts with sequence alignment, providing insight into sequential analysis with explicit reference to temporal sequences.

Geographical analyses also add the spatial dimension to the concept of a sequence. Geographic research into sequences usually falls into one of two categories: spatial diffusion over time – for example contagions like AIDS (Shannon and Pyle 1989) – or movement trajectory analysis. Torsten Hägerstrand has become known for his work on both, beginning early in his career with the diffusion of innovations (Hagerstrand 1967, 1965), and transitioning later to trajectory modeling (Hagerstrand 1970). Critical advances to integrating spatial and temporal dynamics were made in efforts to assess the similarity of trajectory patterns, such as human travel paths (Vanhulsel et al. 2010), hurricane paths (Dodge, Laube, and Weibel 2012), and soccer player movement (Laube 2005). Analysis of movement paths dominates geographic research on spatio-temporal sequences, and the advances made by several of these studies in pattern similarity are important ones that can be applied to event sequences. Similarly, sequence alignment’s use of the edit distance provides the capability of assessing the similarity of spatio-temporal sequences.

Sequence alignment has on relatively few occasions been used to analyze sequences in both space and time. Geographical examples stem largely from literature in transportation and mobility studies at the individual scale. Shoval and Isaacson (2007) introduced the method to geographic literature by deriving groups of similar tourist trajectories based on the visit order of points of interest. Jawad, Kersting, and Andrienko (2011) used sequence alignment on simplified traffic trajectories, showing it to be as effective of state-of-the-art data mining techniques.
Fabrikant et al (2008) and Çöltekin, Fabrikant, and Lacayo (2010) examined sequences of eye movement recordings to assess digital time series information layout and visual analytics strategies. Mavoa et al (2011) assessed the accuracy of sequence alignment in comparison to diary logs in a study of the mobility of children. Tuggener, Çöltekin, and Fabrikant (2012) used sequence alignment to understand long-term mobility patterns as they relate to indicators of social inequality. These projects were able to incorporate spatial variability by viewing location not just as an attribute of each event in a time step, but as the variation between events in successive time steps. In this way, sequences of movement – whether an individual’s location in space or eye gazes on a computer screen – differed based on the locations, duration of time spent at those locations, and order of locations “visited.” Geographers have also used sequence alignment strategies in the analysis of map sketching (Huynh et al. 2008) and geovisual tool creation (Griffin 2009).

2.3 Similar Strategies for Pattern Analysis

Hazard models are one of the most common statistical applications for temporal event analysis. Hazard models, for which Cox Regression (Cox 1972) is the most well-known, model the likelihood of an event occurrence given a list of risks. One example is breast cancer mortality as a result of various demographic factors, time elapsed before seeking diagnosis, and disease characteristics (Prentice and Gloeckler 1978). Event history models are used in sociology to measure transition between events with a focus on the temporal duration of transition (Yamaguchi 1991). These models are useful for estimating event-to-event transitions and tipping points, but are unable to compare full sequences of many events, making pattern analysis impossible.
Time series analyses and Markov models attempt to analyze events as step-by-step processes that maintain sequential structure. Time series models attempt to model events in time as a result of the dependencies on earlier events in a sequence (Box, Jenkins, and Reinsel 1994). Markov models are also based on probabilities of transition from event to event, but with the assumption that previous events in sequence do not affect future distributions (Stewman 1976). Time series models approach an analysis of sequences as both full sequence and individual event transitions and the focus on the temporal aspect of event sequences is a desirable property in analysis of geographic patterns. However, even these models assume a causal link between events in a sequence. Causal links are not known a priori for a pattern found in unconnected events, so this assumption makes time series methods impossible for matching patterns.

Beyond statistical analysis of patterns in spatio-temporal information, geographers have been increasingly relying on geovisual analytics. Visual exploration of large, complex datasets can help discover unexpected patterns and connections between events (Keim 2002). Geovisualization has also been used in a spatio-temporal context to analyze temporal intervals (Weaver et al. 2007), and movement trajectories (Murray et al. 2011). Pattern matching that requires direct comparison of two large sequences is not efficient using visual methods, especially for the goal of quantifying the match between those sequences. Although visualization would make pattern matching difficult and subjective, visual exploration techniques can be leveraged to inspect sequences and their resulting alignment.

2.4 Political Pattern Analysis

This thesis draws on much of the same intentions and methods associated with comparative political and policy diffusion research, though at an elementary and potentially naïve level. The comparative politics methodology aims to derive an understanding of difference
between places by acknowledging what separates them as political entities. Part of what that entails is making inferences about the political structure of a country from observable events and patterns (Landman 2003). Peters (1998) also emphasizes the role of inference in comparative politics by asserting that comparative politics is an analysis of the differences between regimes as a function of how they respond to events of significance. Experimentation as a method to discover these interactions is impossible, which is why inference about a combination of reactions to events and geographic and political difference are necessary factors that the method of comparison must help to illuminate.

Pattern detection techniques, such as those that are aimed at finding significant patterns in big data, facilitate the step of inference about what causes an event to occur after another, albeit in a singular way. Typically, a pattern detection technique for finding relevant patterns in political data uses event data that is specific to a small number of states or even a single state. Such a strategy ensures patterns are specific to the study area, but ignores the bigger picture of similar patterns that occur elsewhere. Integrating a comparative political approach to the analysis of meaningful patterns of events and their indications for the places at which those events take place would require comparable differences between patterns and place-specific characteristics to which the variation of patterns can be attributed. The technique introduced in this Masters thesis facilitates comparative analyses of places by comparing the patterns of events that occur from one place to one another and quantifying their differences. Following Landman (2003) and Peters (1998), inferences can then be made about the political function of a state in relation to another through the deviations in their compared patterns.

Political science often uses a theory of policy diffusion to explain differences among countries’ reactions to significant events. Similar to Hägerstrand’s work on innovation diffusion (Hagerstrand 1967), researchers have identified several ways that policy diffuses through space and time in traceable patterns (Shipan and Volden 2008). Also similar to Hägerstrand’s findings,
geographical factors largely affect the observable patterns of policy diffusion. Distance decay, where more spatially proximal policy-makers experience similar patterns of diffusion, is a visible factor. Social factors also influence policy diffusion, such as the ability to learn from previous adopters, economic ties, desire to imitate previous adopters, and susceptibility to coercion (Shipan and Volden 2008). While Hägerstrand’s studies did not examine the motivations for implementing innovations, Shipan and Volden do not consider how information about potential policy changes are obtained by those who implement them. A consideration of the spatiality, temporality, and sociality of both approaches is necessary for understanding diffusion. This Masters thesis takes a middle ground, examining the ways that information is disseminated in order to create spatial and temporal patterns, and the ways that those patterns indicate fundamental differences in the political structure of the places that exhibit them.

### 2.5 Summary

The first part of this review was an overview of sequence alignment as it evolved through time and across disciplinary bounds. Sequence alignment has a history of use in many fields, which is proof to its flexibility and power. By observing how the strategy has evolved over time, the modifications that I introduce in this Masters thesis are put into an historical precedence and context. Geography has only recently been inserted into sequence alignment’s history, though the field stands to benefit greatly from sequence alignment’s potential. Other methods for analyzing patterns, especially those within Geography for spatio-temporal analysis, have different capabilities, which warrants the introduction of methods like sequence alignment for comparing spatio-temporal patterns. Other fields possess methods for analyzing patterns, such as comparative politics in political science, which I also use in conjunction with sequence alignment to understand a particular pattern of diplomatic cooperation in Yemen and other countries.
Chapter 3

The Sequence Alignment Procedure

Given the simplicity of sequence alignment’s goal, the algorithm presents a surprisingly complex computational problem. Gaps must be placed at strategic locations within a sequence to both approximate the composition of the other sequence, as well as minimize the potential for adding further gaps in the same sequence. The Needleman and Wunsch algorithm (Needleman and Wunsch 1970) anticipates the effects of any manipulation and aligns sequences to minimize the ultimate Levenshtein Distance (Levenshtein 1966) between them. The algorithm uses a two-dimensional similarity matrix of size $m \times n$ (where $m$ and $n$ are the sizes of each sequence) with each cell $i$ containing a numerical representation of the strength that the combination of the two corresponding elements at indices $m_i$, $n_i$ would have in an alignment of the sequences. A higher cell value indicates that aligning the two corresponding elements contributes greatly to the sequences’ similarity. The algorithm then traverses the matrix to maximize the sum of “collected” cell values along the way (Needleman and Wunsch 1970). Traversing the matrix yields a single optimal combination of elements at corresponding positions from both sequences and ensures that every element from both sequences is included in the final alignment. Gaps are placed in a sequence when the traversal indicates that the sequence must be extended to accommodate for intervening elements in the other sequence. In this chapter, I demonstrate the steps taken by sequence alignment to compare two sequences. The process is demonstrated using two fabricated sequences of characters as given by Needleman and Wunsch (1970): “ABCNJRQCLCRPM” is compared to “AJCJRCKCRBP.”
3.1 Matrix Generation

The sequence alignment process begins by creating an $m \times n$ matrix that compares each element in both sequences to each other. This step serves the necessary functions for comparing sequences as both larger structures and a composition of elements that occur in different orders and times. First, the two-dimensional structure compares individual elements in both sequences not only at the same index within the sequence, but each element in each sequence is compared against every element in the other sequence. Second, while examining the individual elements does not take the full sequence into account, the matrix remains organized with index 1 of each sequence located at the top-left cell $(0, 0)$, increasing to the bottom and right $(m, n)$, so the order both sequences is maintained. Combining both elemental comparison and sequence composition gives sequence alignment an advantage over methods that only utilize one or the other to compare linear and sequential information.

3.1.1 Elemental Similarity Matrix

Matrix generation begins with a direct comparison of every element across both sequences. Generally, this constitutes a binary relationship: are the elements equivalent or aren’t they? In a numerical representation as the matrix requires, a false, nonequivalent comparison generates a $0$ or a blank value, while a true, equivalent comparison generates a $1$ (see figure 3-1). While this stage only considers individual matched elements rather than full sequences, visual clues can indicate the emergence of a sequential structure. A pathway leading from the top-left cell in the matrix in figure 3-1 to the bottom-right may be traceable to the naked eye or to those who can identify negative regression at a quick glance. This pathway suggests that the order of the full sequence is important in identifying the sequences’ similarity.
Figure 3-1. Elemental similarity matrix – represents the binary relationship between each element in both sequences. Example recreated from (Needleman and Wunsch 1970)

The matrix representing the similarity between each element in both sequences (subsequently I will refer to as the ‘elemental similarity matrix’) contains the potential for a more nuanced understanding of the elements’ relationships to one another. The numerical representation of each element pair \((m_i, n_i)\) can be explored as more than a binary same or different (Dodge, Laube, and Weibel 2012). This detailed similarity representation makes little sense in the application of sequence alignment to biological sequences, as the base pair elements that are examined do not have important similarities beyond the binary. To illustrate an example, I will use the sequences compared in Figure 3-1. In the top-left cell, an “A” is compared to another “A” and a similarity of 1 is achieved. One cell to the right, a “B” is compared to an “A” and because they are not equivalent, the cell is left blank and counted as 0. However, in the English alphabet, a “B” is very close to an “A,” so the cell could be given a value indicating a less
than perfect comparison but also non-equivalence in the form of a range between 0 and 1. The theoretical concept is simple, though it is subjective in implementation and is not something I have attempted in this Masters thesis research.

With an elemental similarity matrix completed, sequence alignment proceeds by preparing the matrix for the traversal step. This constitutes placing restrictions on which cells can be traversed from the current cell. The rule on which this step relies is simple – a traversal must always be going forward, that is it cannot go “backwards” from an element to a predecessor in the sequence, nor can it remain at the same index within either sequence. In the two-dimensional matrix, this means that the traversal must always be proceeding to a cell that lies at a diagonal to the current one, with increasing row and column indices. A perfect traversal goes between directly diagonal cells, though it is not necessary. The goal of the traversal step is to match the most elements while adhering to the simple rule, though the elemental similarity matrix is not detailed enough to determine the optimal traversal – one that passes the most cells with a 1 in them and deviates as little from a perfect diagonal trajectory – without attempting every possible pathway through the matrix, which would yield an algorithm with complexity $O((mn)!)$). Such a complex algorithm would be unacceptable for real-world application, necessitating that sequence alignment dynamically traverse the matrix just once.

3.1.2 Traversal Matrix

Sequence alignment simplifies the matrix traversal step by constructing a second matrix from the elemental similarity matrix for indicating the most optimal pathway through the values in its cells. Preparing the matrix for traversal involves taking the simple traversal rule that states a traversal must always move forward, and working backwards in the similarity matrix, starting at the bottom-right cell. Each cell is examined one-by-one. The current cell is given a value
representing the highest value that could be reached by beginning a traversal at that cell and applying the rule. Smith and Waterman (Smith and Waterman 1981) and Sankoff and Kruskall (Sankoff and Kruskall 1983) demonstrate that the value for each cell \( M_{i,j} \) is given by the maximum of three possibilities: the cell directly diagonal, the largest cell in the next row, and the largest cell in the next column. These three options are shown formulaically in equations 3-1, 3-2, and 3-3, respectively. The value in the similarity matrix at the same location \( S_{i,j} \) is then added to that maximum value.

\[
M_{i+1,j+1}
\]

Equation (3-1) Option 1: the cell directly diagonal to the current cell, represented by an increase in both row \( i \) and column \( j \).

\[
\max_{k \geq 2} \{ M_{i+1,j+k} \}
\]

Equation (3-2) Option 2: finding the cell with the largest value in the next row. The current cell is at row \( i \) and column \( j \). \( k \) is the number of columns beyond the current one to check.

\[
\max_{l \geq 2} \{ M_{i+l,j+1} \}
\]

Equation (3-3) Option 3: finding the cell with the largest value in the next column. The current cell is at row \( i \) and column \( j \). \( l \) is the number of rows beyond the current one to check.

The process for filling-in the matrix uses a recursive algorithm, as the values that were previously computed are used again to calculate each cell in turn, as well as building on cell values that already exist in the elemental similarity matrix. Needleman and Wunsch (1970) demonstrate the process of creating what I will, for the rest of this thesis call a ‘traversal matrix,’ diagrammatically, and I have re-created their representation in figure 3-2.
3.2 Matrix Traversal

The traversal matrix serves an important role in the sequence alignment process because it combines the information of each full sequence as well as the comparison of each individual element to one another. Because of the recursive and additive calculation of cells in the matrix and the reversed order in which cells are populated, cell values increase to the top-left, where matrix traversal begins. If the first elements in both sequences are equivalent, then the highest value will be contained within that cell, and that is where the traversal begins. Traversal also may begin at any cell in the first row or column if the highest value is not in the cell at \((0, 0)\). From that point, matrix traversal proceeds to the next highest cell in the next row or column until the end of the matrix is reached by way of a cell in the final row or column. In the event of a tie,
matrix traversal ultimately gives priority to the cell immediately diagonal to the current one, then prioritizes the next closest cell – or one that “skips” the least amount of rows or columns.

Needleman and Wunsch (Needleman and Wunsch 1970) demonstrate the process as displayed in figure 3-3. Note that the traversal generally follows the shape of the pattern observable in the elemental similarity matrix in figure 3-1.

Figure 3-3. Traversing the traversal matrix. The path demonstrated by arrows is the one-dimensional optimal path through the matrix. Darker grey cells represent matched elements, lighter grey are mismatches or choices that the algorithm must process. Re-created from (Needleman and Wunsch 1970)

Figure 3-3 illuminates the possible situations that sequence alignment algorithm encounters when traversing the matrix. First is the optimal matching of elements at back-to-back indices, represented by a diagonal step from a cell at index $M_{i,j}$ to the one at $M_{i+1,j+1}$. In an alignment of two equivalent sequences, the entire traversal would be composed of optimally matched cells from the top-left corner to the bottom-right. The example in figure 3-3 only has one
such situation – the matching of “C” and “R” at consecutive indices near the end of both sequences. The second situation appears to the traversal process as an optimal match, but actually represents a mismatching of elements at corresponding indices in the sequences. Mismatches are common and expected when comparing non-equivalent sequences, and are represented in the traversal by the first of two cells with the same value in order. Because in the matrix-creation step the mismatched cell does not contain a value from the elemental similarity matrix, it must contain only the value from the next highest cell. In the third situation, matrix traversal adjusts to sub-optimal matches that do not occur on the diagonal between cells. Figure 3-3 contains several of these situations, which are indicated by the skipping of either one or more column or one or more rows between a cell and the next highest value. Skipped columns or rows are important because they represent intervening elements between an expected matched pair. Finally, figure 3-3 shows a situation where a choice of two possibilities is presented to the traversal process. The choices represent nearly identical choices, where the only difference is in which sequence an element is skipped, so except in situations where this is significant, the match chosen by the algorithm makes no difference and it will match whichever possibility it processes first.

3.3 Representing the Aligned Sequences

An attempt to understand the similarity of compared sequences using only the traversed matrix shown in figure 3-3 would prove insufficient. Sankoff and Kruskal (Sankoff and Kruskal 1983) identify three forms of presenting the sequences after the traversal process is completed: a trace which emphasizes the matched elements in the original sequences by connecting them with traces, a listing which describes the edit operations performed on one of the sequences to transform it into the other in natural language, and an alignment or matching which emphasizes
modification to the sequences by aligning matched elements vertically with one another. An example of each is presented in figure 3-4 using the traversal generated in figure 3-4.

(1) Trace

```
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>N</th>
<th>J</th>
<th>R</th>
<th>Q</th>
<th>C</th>
<th>L</th>
<th>C</th>
<th>R</th>
<th>P</th>
<th>M</th>
</tr>
</thead>
</table>
A | J | C | J | N | R | C | K | C | R | B | P |
```

(2) Listing

```
ABCNJRQCLCRPM
AJCNJRQCLCRPM
AJCJRQCLCRPM
AJCJNRQCLCRPM
AJCJNRCLCRPM
AJCJNRCKCRPM
AJCJNRCKCRBPM
AJCJNRCKCRBP
```

(3) Alignment

```
A | B | C | N | J | - | R | Q | C | L | C | R | - | P | M |
A | J | C | J | N | R | - | C | K | C | R | B | P |
```

Figure 3-4. Three methods of displaying aligned sequences as explained by (Sankoff and Kruskall 1983). The trace shows which matched elements, the listing explains the edit operations in natural language, and the alignment shows where the sequences were extended to match one another.

### 3.4 Quantifying Sequence Similarity

Sankoff and Kruskall’s (1983) alignment representation is the most common among commercial software for sequence alignment. Through this representation the concept of a gap is
introduced. Gaps are placed in an alignment to signify that it has been extended to accommodate an intervening element from the other sequences. The alignment representation can also be used to demonstrate the Levenshtein edit distance (Levenshtein 1966). The edit distance is interpreted as the amount of modification done to both sequences in order to transform one into the other. Because mismatched elements can be weighted differently than skipped elements, Sankoff and Kruskall’s ‘listing’ is not sufficient. The computation method is somewhat flexible, but in essence, the number of mismatched elements and inserted gaps (represented by a dash in the alignment display) denoting skipped elements are summed, resulting in a total number of weighted operations separating the two sequences. The particular weighting scheme is heavily application-dependent (Lesnard 2010), so for ease of demonstration, the unweighted edit distance of the example demonstrated above is 7.
Chapter 4

Sequence Alignment of Spatio-Temporal Patterns

Although sequence alignment is a flexible process in many ways, temporal sequences have unique properties that require some modifications to the process. As Levine (2000) has explained, the analogy that compares proteins and base pairs to events in time and space (in Levine’s case, individuals’ states of employment) is not well developed for the purpose of aligning sequences of them. In this chapter, I present several modifications to sequence alignment that bridge the analogy between DNA sequences and spatio-temporal event sequences in an effort to compare them and quantify their differences. First, I reconceive some of the assumptions that sequence alignment makes in order to use the process for pattern matching. Next, I introduce a series of modifications aimed at utilizing the temporal properties of spatio-temporal pattern sequences. These involve new ways of assembling, reordering, and traversing matrices to account for temporal intervals between events. Next, I describe a re-considered means of calculating the edit distance between sequences with a temporal unit that I call the “Temporal Deviation Distance.” Finally, I discuss the importance of some minor changes to the ways that aligned sequences are displayed, which emphasize the specific deviations between the sequences.

4.1 Fundamental Reconsiderations of Sequence Alignment

Despite the progression that many have helped facilitate over the last four decades to expand the potential of sequence alignment, additional reconsideration is necessary to expand the process for matching patterns of spatio-temporal events. These reconsiderations are less based in the code of the algorithm and more necessary to the understanding of the process as fundamental
assumptions. First is the relative size between the compared sequences. Sequence alignment assumes that sequences are of approximately the same length, while pattern matching seeks to find instances of a short (two to several elements) pattern within a large unexplored sequence of temporally-ordered events. The Levenshtein Edit Distance undergoes one significant calculation change to satisfy this requirement, and it will be introduced later. Sequence alignment still finds the differences between the compared sequences, though in pattern matching, the calculations are confined to the bounds of the shorter pattern.

The major differences between the original implementation of sequence alignment used for biological sequence comparison and the application of this modified strategy for comparison of spatio-temporal event sequences stem from the assumptions that the algorithms make about the structure of the sequences themselves. Biological sequences abide by different physical laws than event sequences because biological sequences are comprised of elements with physical structure. No two base pair combination elements can occupy the same space within a biological sequence. Sequence alignment acknowledges this restriction through the generation of the matrix structures, where base pair elements are evenly spaced in the indices of the two-dimensional matrix representation. Just as the same space within a DNA sequence cannot be occupied by two different base pairs, each column in the matrix is represented by one element from the sequence.

When dealing with event sequences, however, this restriction is loosened. Events often occur simultaneously, a fact evidenced here at the daily scale by events that occur at the same date. By extension, each time step represented in the data may contain many events. Because simultaneous events occur at the same basic unit of time scale (in this case, the day), they can be re-arranged without altering the necessary sequential property of ordering. For the purpose of sequence alignment, this presents two possibilities for representing temporal event sequences in a matrix format: generate columns that represent each time step in the data, or make each column represent an individual event in the data.
Since time is the primary driver in the organization of a spatio-temporal sequence in this study, the inclination is to make the columns in the similarity and traversal matrices represent each successive time step present in the data. Functionally, this strategy ensures that each event that occurs at the same time is given equal priority in the order of the sequence. Analytically, however, this strategy does not fit with the capabilities or intentions of sequence alignment. Particularly, formulating a matrix for this strategy presents a non-trivial and dysfunctional task. Prioritizing time over events generates a column representation that must be flexible enough to handle a time step containing zero events, or as many as hundreds. A two-dimensional matrix representation is unable to accomplish the task without extension into \( n \) dimensions, \( n \) being the maximum number of events in a single time step. A many-dimensional matrix organized in this manner becomes sparse and a burden on computation time and storage very quickly due to the variability of events per time step. Alternatively, a three-dimensional matrix with the third dimension representing the number of events occurring on that day would not be sparse, though the high variability contained therein would create issues for the traversal process.

The alternate method that I adopt for this extended sequence alignment technique prioritizes events in the necessary matrix-generation process while maintaining each event’s crucial temporal component. I organize the matrices such that the indices of the x and y axes are not successive steps in time, but of events in sequence. This strategy overcomes the issue of generating a very sparse matrix and enables detailed consideration of event variability within the basic unit of time. Simultaneous events, or events with a 0-day interval between them, prove that this sub-unit variability is a necessary component to aligning spatio-temporal sequences.

In a pattern match as opposed to the biological application of sequence alignment, the pattern sequence and the unexplored sequence have separate properties which require that the matrix be generated consistently. In a biological application, the sequences are reversible on the x and y axes when assembling the similarity and traversal matrices. To maintain a timeline-like
representation based on the unexplored sequence, I put the unexplored sequence on the x axis, where each element is given by a column in the matrix, and the pattern sequence on the y axis, where each row is an element in the pattern. Associated with each column is the time stamp of the event that it represents in the unexplored sequence. Since the pattern is a template of what is expected, rows do not contain time stamps, they contain the temporal intervals between the expected event signified by that row and the next one in the pattern. Therefore, the final row in the matrix does not have a reference to any temporal interval. This temporally-aware matrix organization contains the necessary information to align the two sequences, though the process for doing so also requires some modification to accommodate the new matrix.

### 4.2 Temporal Matrix Population and Traversal

The original method for aligning sequences based on a perfect diagonal traversal from the top-left of the matrix to the bottom-right is not compatible with the temporal irregularities of this new matrix. Here, an optimal alignment is redefined to be based on the expected temporal interval between specific events, rather than simply the adjacent events in the sequence. This yields the potential for multiple optimal alignments as multiple events of the expected type may occur within the interval. Conceptually, extending the matrix assembly and traversal steps to accommodate multiple potential optimal alignments is a simple extension (with a means for resolving ties, which is actually accomplished internal to the algorithm), though in practice, it takes some explanation.

Using the formulas diagrammed in the previous chapter as a starting point, three new formulas are generated to take temporal intervals into account. Optimal matches are redefined to be contained in the particular intervals between events, which may include searches that skip and reorder columns. Sub-optimal matches expand the search beyond optimal matches both before
and after expected temporal intervals and pattern elements. Finally, matrix traversal uses temporal
intervals in a similar way while prioritizing optimal matches and sub-optimal column matches to
preserve the expectation of the pattern.

4.2.1 Generating Optimal Values

Since the extended matrix infrastructure which provides rows \((i)\) for elements of a pattern
and columns \((j)\) for events in an unexplored sequence can be built simply, each cell contained
within given indices \((i, j)\) can be used to compare event timing to particular temporal intervals.
Because every cell is related to a single pattern element at index \(i\) of the pattern sequence and a
single event at index \(j\) of the unexplored sequence, it contains both an expected temporal interval
and a specific time stamp. Remember from the previous chapter that the value of each cell in the
traversal matrix is a function of the largest of three things: the directly diagonal cell, the highest
value in the next column, and the highest value in the next row. Since the diagonal cell contains
the optimal match from any given cell, to extend the process to temporal sequences the search for
an optimal match includes any cell considered optimal based on the temporal interval of the
current cell and the dates of the columns in the matrix. In the Y dimension, an optimal solution
remains the same as in the original version, where the next row \((index i + 1)\) is the only option
that does not involve skipping an expected pattern element. Therefore, the formula from equation
3-1 of the original process is extended to include a list of optimal columns \((P_i)\) and is shown in
equation 4-1.
Equation (4-1) Option 1 in temporally-extended method. Based on equation 3-1, where \(i\) is the index of the current cell’s row. \(P_i\) is a list of the optimal columns for the current cell given the interval of the pattern element at index \(i\).

This extended process of finding optimal columns presents a couple of potential peculiarities that are worth mentioning. First is the obvious question of how to resolve ties when presented with multiple potential “optimal” solutions. A tied value indicates that two (or more) events occurred when they were expected, and if they occur on the same day, it does not matter in a temporal sense which is selected to represent the pattern. In reality, one of the tied events many have more significance to the pattern, but since detailed consideration of particular events is not viable at this stage of the procedure, this question is not addressed. A tie may also occur on separate days within the temporal interval, but both options still represent equivalent optimal options. If taking one of the two tied values over the other in the traversal stage meant that a future event no longer fit in its expected interval, the values would not have been tied in the first place. By default, the algorithm selects the first instance of a tied value to refer to, searching columns in increasing time away from the current cell’s.

Another possibility is that an optimal column lies at an index less than that of the current cell when simultaneous events occur. At first glance, this violates the order of the sequence. At the matrix-generation stage where the maximum reachable value is the only item of interest, the algorithm is only extended to include a search of optimal columns with indices less than the current one and sequential order is maintained. Traversal also maintains order by re-ordering simultaneous events. Additionally, optimal columns may be non-adjacent to the current cell’s column. This is likely given temporal intervals of two or more days. This option does not pose a problem to this stage of the process, as the value is the only item of interest for generating a traversal matrix.
4.2.2 Generating Sub-Optimal Values

The flexibility of sequence alignment is in its ability to match events when they occur outside of the expected temporal interval. This is called a sub-optimal match because a match is found within a reasonable distance from the temporal scope of the optimal columns or a section of the pattern is matched although one or more elements could not be found. This step in the generation of the traversal matrix is nearly identical to equation 3-2 in the original matrix generation process. The first part of this step finds instances of matched events outside of the expected temporal interval for the current pattern element by searching all columns in the next row with indices greater than the current cell’s. Incorporating the multiple potential optimal columns from the previous step, this search for sub-optimal matches involves simply searching columns with indices beyond that of the optimal ones. Additionally, a sub-optimal match may occur at a lesser interval than the expected one. To illustrate this fact, a temporal interval of two or more days between expected events would involve searching optimal columns that are not adjacent to the current cell because all columns within one day lie between the current cell and the events occurring on the second day. Therefore, a sub-optimal match may occur following or preceding the given temporal interval. The extended second step is computed using equation 4-3.

\[
\max_{k \geq 1} \{ M_{i+1, P_i, \pm k} \}
\]

Equation (4-2). Option 2: finding sub-optimal columns, where \(i\) is the index of the current row, \(P_i\) is a list of the optimal columns given the temporal interval of \(i\), and \(k\) is the number of columns beyond the optimal set of columns.

The third part of the search for sub-optimal matches is done in a similar way as just described. This part searches for matches in sub-optimal rows, which assumes the opposite of step 2 – that the anticipated temporal interval remains intact, though the expected event did not
occur. A sub-optimal row match indicates an intact pattern but for the lack of a single expected event. The extended step is computed as displayed in equation 4-3.

\[
\max_{l \geq 2} \{M_{i+l, P_i}\}
\]

Equation (4-3). Option 3: finding sub-optimal rows given the current row index \(i\), and the optimal columns \(P_i\). \(l\) is the number of rows to search beyond \(i\).

Finally, these three options are combined into a single equation representing the process of generating a traversal matrix displayed in equation 4-4. The largest value from the three steps is chosen for the value which the current cell takes. To that value, the value from the similarity matrix at the same row and column index \((S_{i,j})\) is added for a cumulative total of matches reachable from that cell.

\[
H_{i,j} = \max \{ \max_{P_i} \{M_{i+1, P_i}\}, \max_{k \geq 1} \{M_{i+1, P_i \pm k}\}, \max_{l \geq 2} \{M_{i+l, P_i}\}\} + S_{i,j}
\]

Equation (4-4) Complete computation of each cell in the matrix at index row = \(i\) and column = \(j\). Equation is a composite of equations 4-1, 4-2, and 4-3.

4.2.3 Traversing the Matrix

Finally, the alignment process concludes by traversing the generated matrix. Traversal ensures the consideration of each element in both sequences in the final alignment, resolves ties where multiple cells contain the same value, defines the location and nature of any modification done to either of the sequences, and establishes the basis for calculating the edit distance between them. More aspects of sequence alignment rely on this simple repeated step of finding the largest value from a list of numbers than any other operation. Most importantly, every step of the
traversal process includes the identification of an element from the pattern and an event from the unexplored sequence which match together and form part of the aligned result.

The traversal step always begins at the top-left corner and proceeds by increasing row and column index, which leads towards the bottom-right corner. Along the way, it passes through individual cells with the goal of “collecting” the highest values. The traversal accomplishes this goal by simply finding the highest cell value available to it by following the same rules that were used to set up the matrix in the first place. Though this may seem redundant, traversal is a necessary step that generates the optimal alignment of the sequences given the values calculated in the traversal matrix. Most notably, the traversal step is able to resolve the ties that are generated in the matrix by prioritizing certain groups of cells.

Matrix traversal generally begins at the top-left corner at \( M_{0,0} \) where the cell represents the first elements in both compared sequences. This top-left cell represents the optimal location to begin the traversal as it represents the first cells of both the pattern sequence and the unexplored sequence, and is the first priority choice given to the traversal process. The optimal choice for the traversal step coincides with the first option of the matrix generation process, which searches all of the optimal cells for this highest value. The traversal step searches those same cells for the highest value contained within them. But in addition to remembering the value of that cell, the traversal process remembers the location of that cell in the matrix by storing the indices of its row and column. If that cell contains the largest value reachable from the current cell being traversed, then it becomes the next current cell in the traversal.

The second priority is to find the highest value in a sub-optimal column. In this case, the algorithm compares each cell’s value in the list of sub-optimal columns (identified the same way as they are in the matrix generation phase) to the stored highest value from the optimal cells. Only if a sub-optimal value is higher does it take precedence over the largest optimal value. This is an efficient way of resolving ties and ensures that optimal cell choices receive the highest priority,
even when an equal choice is presented through a sub-optimal match. However, if the sub-optimal value is, in fact, larger, then it would indicate that the most optimal alignment occurs by matching events outside of the expected temporal interval. In the case of a higher value, the location of that cell is stored and the third step occurs with the traversal searching sub-optimal rows for an even larger value. Traversal follows the same guidelines from the matrix-generation step three to determine which cells to search for sub-optimal rows. A higher value in one of these cells indicates the partial continuation of a pattern following the absence of an expected event.

Though not as common as the previous two situations, the absence of an expected pattern element is significant and an important situation for the traversal to recognize. Figure 4-1 diagrams the alignment process from matrix-generation through traversal with a partially-completed matrix.
Figure 4-1. Generating and traversing the matrix in my extended method.

(a) Matrix organization with columns as events that have dates and rows as expected pattern elements that have temporal intervals between them. In this case, intervals are in days.
(b) Similarity matrix with all cells where row and column are equal are given a value of one.
(c) Filling in the matrix cell by cell. The current cell is in grey.
(d) Checking optimal columns for highest value. Current cell is in light grey, optimal cells are in dark grey.
(e) Checking sub-optimal columns indicating match outside of expected temporal interval. Sub-optimal cells are in dark grey.
(f) Checking sub-optimal rows indicating partial match that skipped an expected element of the pattern. Sub-optimal rows are in dark grey.
(g) Traversing the completed matrix. Chosen values in the expected temporal interval are in dark grey, and matches outside of the expected interval are in light grey.
4.3 The Temporal Deviation Distance (TDD)

In an effort to make this extended sequence alignment method more useful in a standalone way, the traditional edit distance generated by the original algorithm is modified slightly. The traditional edit distance measures the amount of modification done to both sequences to make them essentially equal. As discussed earlier, gaps and mismatched elements are often weighted differently in this calculation. This method of quantifying the dissimilarity between two sequences is impossible to interpret due to the lack of a unit attached to it. The original biological application of sequence alignment avoids this impossibility by comparing the edit distance between two sequences against the results from other alignments to develop a phylogenetic tree of sequences and their relationships to one another. Sequence alignment can describe the distance between a pattern and a sequence of unexplored events in a standalone way without the comparison of other sequences. To accomplish this, the Levenshtein edit distance is modified for temporal properties of events and pattern, and in so doing, its unit is changed to measure time. I call it the Temporal Deviation Distance (TDD).

One necessary change should be immediately noticeable from this re-conceptualization of the alignment of a pattern sequence with a large, unexplored sequence. Sequence alignment necessarily inserts gaps where elements of one sequence do not match the other, so the number of inserted gaps is, in part, a function of the size difference between the compared sequences. In a comparison of DNA sequences that are assumed to be the same size, this is entirely realistic. However, my application works with the assumption that patterns are only a fraction of the size of the unexplored sequences to which they are compared. Instead of changing the way that the matrix is defined or traversed, I introduce the concept of a “zero-weight gap” that is part of the alignment, but does not contribute to the TDD and is de-emphasized in the final alignment display. Assuming an expected pattern is most likely to occur beginning at some event in the
middle of the date range present in the unexplored event sequence, the large number of gaps necessarily inserted between the first event in the unexplored data and the matched instance of the first pattern element as well as between the matched instance of the last pattern element and the last event in the unexplored sequence only indicate the temporal location in the unexplored data where the expected pattern occurred – not deviations from the pattern. Since they do not actually modify the expectation of the pattern, I deemphasize and disregard those gaps as analytically unimportant, hence the assignment of zero weight.

Similarly, not all gaps inserted between expected pattern elements should count against the TDD or the sequences’ alignment. Since what is analytically important in this comparison of pattern to unexplored events is the differences between the expectation of a pattern and the true event occurrences, the situations where events occur in the same sequence and temporal interval in which they are expected should be ignored by the TDD calculation and de-emphasized in the alignment. In practice, a zero-weight gap is assigned where events “intervene” between matched elements of the pattern within the expected temporal interval. Ignoring those instances where zero-weight gaps belong, the algorithm is left only considering intervening events between matched pattern elements that occur outside of their expected temporal interval and mismatched events indicating only a partially matched pattern.

Most importantly for demonstrating the standalone Temporal Deviation Distance, sequence alignment requires a shift in focus from events to time. Levenshtein’s edit distance is a measure of the number of elements (with weights, so not a literal interpretation) that differ between the compared sequences. As the name suggests, I introduce the Temporal Deviation Distance as a measure of time by which the compared pattern and unexplored event sequence differ. Calculating the TDD requires examining the event information and keeping track of the dates of the events that were marked as intervening at a temporal unit beyond the expected
interval between pattern elements. The amount of time represented in those events are summed and added to the TDD calculation.

The nature of the TDD as a distance measure means that there is no distinction between matches that occur after the maximum expected temporal unit in the interval and ones that occur before the minimum unit in the interval. Any deviation in the positive or negative direction additively increases the TDD. As important as it is to be able to tell the difference between a pattern that transpires more slowly than it is expected to and one that transpires more quickly, making positive and negative distances runs the risk of generating a TDD of zero, which would indicate a pattern that occurred perfectly as expected, when in reality, the pattern occurred very differently. For now, the TDD remains a non-negative measure of distance, while the alignment view introduced in the next section is used to determine slower or quicker pattern deviations.

### 4.3.1 Additional Temporal Descriptors

In addition to the Temporal Deviation Distance as a standalone quantification of an expected pattern, two descriptive measures of the temporal properties of the pattern sequence and its relationship to the unexplored sequence and the TDD are also computed. Again, the goal is to describe the dynamics of a given pattern with respect to its time, rather than its number of events. As indicated, a pattern match involves the comparison of a relative short pattern sequence to a very large unexplored event sequence. But this pattern matching algorithm deals with time, so comparing a five-element pattern to a 500,000-event sequence is very different from comparing a pattern that spans a maximum of two weeks to a sequence with events that span one month. The former situation is one in which a match is less likely to be found because of the small window of time. Hence, an important measure of the significance of match is the ratio of time contained in
the pattern to the time contained in the unexplored sequence. The smaller the ratio, the more likely the pattern will be found simply given the potential searchable area.

Once the TDD has been calculated, it is also important to know the significance of the result. This is accomplished by generating a percent error with respect to the maximum time span of the pattern sequence and establishing a ratio of TDD with its temporal unit to maximum time span of the pattern. The larger the ratio, the greater the likelihood that the deviation from the pattern is attributable to chance and that the matched events are random rather than part of the pattern. When used together, the TDD and both ratios give a good indication of the significance of a matched pattern.

4.4 Visual Exploration of how the Pattern Deviates

I make one final extension to the original sequence alignment procedure meant to gain a more nuanced understanding of the nature of a pattern’s match to unexplored events. The Temporal Deviation Distance is a simple measure for generating knowledge of the pattern quickly, but the next step is to find out where deviations occur and if they are a result of mismatches, slower than expected events, or quicker than expected events. For this goal, previous applications of sequence alignment in common software programs – like the CLUSTAL series (Higgins and Sharp 1988), CLUSTALG generalized for social science applications (Wilson, Harvey, and Thompson 1999), and MUSCLE (Edgar 2004) – utilize the alignment representation of displaying results. These programs use color to delineate each of the element types when they match strongly in multiple places. Such a strategy works for these examples because biological applications have a limited number of event types, and because these applications are most interested in the most significant matches and locations of them.
Though my pattern matching process also utilizes color to show the alignment results, my strategy varies slightly. The number of event types present in my data is much too large to assign a unique color to each one, and even just coloring the event types that comprise the pattern makes little sense. There is no need to distinguish between event types, so instead, I emphasize the lack of a match when it is expected by using color and hue. I use green to delineate matches when they occur where they are expected to, blue to indicate mismatches, red to indicate gaps between matches that occur outside of the expected temporal interval, and a subtle, lighter red to show zero-weight gaps that occur between expected matches. The color scheme and hue that I have adopted serve to highlight the most important aspects of the pattern matching results so that aligned sequences can be efficiently explored for their particular deviations.

4.5 Tools: Java Programming

I have accomplished all of the modifications described above in a custom environment using the Java programming language. Besides being a language in which I have extensive working experience, Java has capabilities that are conducive to this project as a whole. First of all, Java is object-oriented, which is entirely consistent with our way of thinking about events as entities with attributes and relationships (Shipley 2008). Second, Java’s built-in concept of abstraction aids in the understanding of how patterns, events, and time work together but have separate functions. Abstraction allows a Java user to create objects that extend and build on the functions of other objects, and is a property that I have taken advantage of. For one example, a zero-weight gap is a special form, which serves the same purpose of a weighted gap, but has the additional property of not counting against the computed Temporal Deviation Distance. Gaps themselves have the same function as elements in both patterns and unexplored sequences because they are placed alongside them in an alignment and must be comparable for the purpose
of matching. See the appendix for natural-language pseudocode representation of the entire process, in detail.

Another advantage of using Java to create this program is its ubiquity in geographic and other programs of its type. Java’s versatility has been leveraged by many and has become one of the most common programming languages for creating visualization tools from scratch. Therefore, Java has a well-developed network of users who offer advice and assistance in online forums and other venues. I have utilized that network frequently throughout the process of creating this tool. Finally, because Java is such a popular tool for creating custom analysis techniques, there is a vast network of potential applications and partnerships for sequence alignment. The GeoVISTA Lab at Penn State possesses an array of geographic visualization tools created in Java which provide a significant resource for generating inputting patterns, or that could benefit from the pattern validation capability that this technique provides. Of these, the most notable is the GeoVIZ Toolkit (Hardisty and Robinson 2011) and STempo (Peuquet 2012). Although this pattern matching method is designed to be used as a standalone application, it would see a significant advantage by working with other tools that similarly analyze patterns in spatio-temporal events.
Chapter 5

Intersecting Sequence Alignment and Temporal Geography through Diplomatic Patterns in Yemen over Time

This chapter demonstrates the extended sequence alignment method on real world data. It starts by describing a spatio-temporal pattern as generated by the STempo program. It then introduces four additional datasets to which the generated pattern is compared for an analysis of pattern deviation over time. The majority of the chapter examines each of the four datasets for its Temporal Deviation Distance and visual alignment. Finally, the chapter concludes by discussing the results from those comparisons by paralleling them with distance decay with respect to time.

5.1 Generating a Spatio-Temporal Pattern using STempo

Although the significance of a specific pattern may be left for interpretation, in general patterns are defined by the properties of sequence and recurrence. Here, I examine patterns that exhibit sequencing through the succession of events through time, and recurrence through the statistically significant relationship that each event has with proceeding events and successive events in the pattern. I use the STempo visualization environment for spatio-temporal pattern analysis to discover patterns in independent event data (Peuquet 2012). STempo uses a method called T-_pattern analysis (Magnusson 2000) to discover chains of events that typically occur within a moving time window of one another, which generates patterns of significant temporal relationships among types of events. The user selects a significance level, which restricts or expands the interval of time to define a significant event relationship. T- Pattern analysis derives a sequence of events with temporal intervals between them that indicate the expected amount of
time before the next significant event. One such pattern that has been found to be consistent in the domain of diplomatic negotiation in Yemen during the Arab Spring is displayed in figure 5-1 and will be used in this chapter to illustrate how this modified sequence alignment procedure can be used to understand the changes a pattern undergoes over time.

Figure 5-1. A temporal pattern of diplomatic cooperation generated by STempo. Event types are represented by their numerical codes and event transition is given as an interval of days.

STempo finds patterns as they appear in a dataset focused in a spatial and temporal context. Generating a pattern from data that does not have a specific temporal or spatial focus makes the pattern too general to be used to facilitate any significant conclusion-making. For this analysis, I have selected a dataset consisting of political, diplomatic, cultural, military, civil, and natural events specific to Yemen during the span of time from February 2011 to March 2012 during the height of the Arab Spring. The data’s location of Yemen is defined in one of two ways: one of the actors involved in generating or receiving the event is related to Yemen (such as the Yemen government, citizens, or a rebel group operating in Yemen) or the location assigned to the event given information from its source (where available) falls within a rectangular bounding box approximating the border of Yemen. The data is coded to a predefined dictionary of event types using a combination of the CAMEO event coding framework (Gerner et al. 2002) and hand-
coding from RSS news feeds. STempo finds many patterns in this dataset over this timeframe, though I have selected one of the more intuitive patterns representing a traceable process of understandable events. The pattern is one indicating turmoil, negotiation, and failure of those negotiations. An accusation of human rights abuses (event code 1122) is followed within one day by a demonstration for a change in leadership (1411), followed within another day by engagement in negotiation (46), followed between three and eight days by an optimistic comment being made (13), concluding four to six days later with rejection of the plan that would have settled the dispute (127). The pattern makes logical sense, but STempo’s T-Pattern analysis shows that the events that compose the pattern transpire in this temporal sequence several times concerning Yemen during the timeframe of the Arab Spring, satisfying the recurrence property associated with my definition of a pattern.

5.2 Comparing the Pattern to Other Temporal Contexts

The statistical certainty of this discovered pattern does not contribute enough understanding to describe the transition of Yemen as a political entity through time. The dataset from which the pattern was derived consists of events that took place over the course of the 13 months from February 2011 to March 2012, and restricted to events with some relation to Yemen. The next step described here is to compare this known pattern to other data in order to answer geographic questions of distance and similarity, and political questions of difference based on event transitions. The remainder of this chapter contains an analysis of this Yemen pattern as it changes over time while keeping the Yemen location constant. Chapter 6 takes the opposite approach, keeping time constant and comparing the Yemen pattern to events in other locations.

I have selected four periods of political conflict in Yemen with increasing temporal distance from the events that generated the above pattern identified in 2011-2012: the rise of Al
 Qaeda in 2009-2010, increased rebellion and related conflict in 2004, civil war between northern and southern forces from late 1993 through 1994, and a period of small skirmishes from 1986 to 1993 surrounding the unification of North and South Yemen in 1990. The data for these time periods was extracted from the GDELT dataset (Leetaru and Schrodt 2013). GDELT is a unique politically-oriented event data set that contains detailed event information on a global scale from 1979-2012. The extracted events during those time frames are reduced to only the necessary event attributes of date, event code, and latitude and longitude location of the event.

Comparing patterns between two separately-generated datasets provides a few distinct advantages. In this case, both STempo’s Yemen data and GDELT are at least partially coded using the CAMEO framework (Gerner et al. 2002) (STempo includes some custom codes for economic and religious events and involved extensive hand-coding), which ensures that the event type designations in both sets of data are interoperable and directly comparable. Both datasets were also created using different news sources (and different numbers of them, which significantly changes the likelihood that any given pattern of events will appear in the data), ensuring a level of out-of-sample confirmation to complement STempo’s in-sample recursion detection. Although GDELT’s spatial and temporal coverage is many times greater than that of STempo’s, I have used simple Python scripts to subset into a useable single dataset for Yemen for each of the four years that I have compared.

Although the time range of 2011-2012 examined here is only a fraction of the political history of Yemen, these small slices of its modern history are enough to reveal the changing political dynamic of the country. In a stationary world, the observed pattern of accuse of human rights violation, demonstrate for leadership change, engage in negotiation, make optimistic comment, reject plan would transpire similarly regardless of the time frame or location examined. In this case, comparing the pattern to four separate times reveals some historical trends to the
changes experienced by this expected pattern over time, and therefore, about the changing political atmosphere of Yemen.

5.3 Analysis of Yemen Patterns over Time

Predictably, the least amount of temporal deviation from the expected pattern occurs in the dataset closest in time to that from which the pattern was derived. In the 2009-2010 data, the events in the expected pattern occurred in total six days slower than the expected time span of the pattern (Figure 5-2 (a) shows the output from the modified sequence alignment program). Looking closer at the specific events, a more detailed understanding of this pattern can be found. Figure 5-2 (b) shows the resulting alignment of the unexplored event sequence and the expected pattern sequence, and shows that one of the expected events did not event occur in sequence with the others. The expected temporal interval of two days preceding and following that missing event (a maximum of one day was expected between event type 1122 [demonstrate human rights abuses] and 1411 [demonstrate for leadership change], and another one day between 1411 and 46 [engage in negotiation]) is added to the total temporal deviation, which means that between two additional events in the pattern (absent from the figure due to display size restrictions) four additional days took place than what was expected for the pattern to transpire.
Figure 5-2. (a) Temporal edit distance and (b) alignment representation of expected pattern to Yemen data in 2009-2010. Additional metrics for understanding temporal deviation are included in the distance metric window. In the alignment window, green represents a matched pair of elements, blue represents mismatched elements, and light pink represents intervening events between expected events that are within the expected critical interval. The unexplored sequence is on top, while the expected pattern sequence is on the bottom.

Next, I compared the expected pattern to a period of rebellion and clashes in Yemen in 2004. As anticipated, the expected pattern deviates slightly more than that of the time frame of 2009-2010, which initially suggests a nonstationary process changing the trajectory of the observed pattern, though the deviation is not dramatically different from what was expected or the deviation that occurred in the 2009-2010 dataset. In Figure 5-3 (b), the same phenomena of the missing ‘demonstration for leadership change’ event occurs, though in addition, the time between the remaining events is increased by seven days – three more than in the previous example – and the interval between the missing 1411 event and the 46 event increased several days beyond what was expected.
Figure 5-3. (a) Temporal edit distance and (b) alignment representation of expected pattern to Yemen data in 2004. Additional metrics for understanding temporal deviation are included in the distance metric window. In the alignment window, green represents a matched pair of elements, blue represents mismatched elements, and red represents intervening elements that are outside of the expected critical interval. The unexplored sequence is on top, while the expected pattern sequence is on the bottom.

The final two examples continue to reinforce the observation of a changing pattern through time at a given locational context. Although the time periods of 1986-1993 and late 1993-1994 are not separated by much, the temporal deviation between the compared pattern and the two datasets is striking (despite the vast different in temporal range in each of these datasets, the number of events that they contain is approximately equal. The events-per-day ratio does have an effect on the potential for matching patterns). Twenty-two days is a reasonable amount of deviation for events occurring in 1994 considering the trend in deviation exhibited by the 2009-2010 data and the 2004 data, but an additional 16 days were needed to complete the pattern in the events from 1986-1993. This would seem to indicate a large and sudden shift in political and social organization, most likely attributable to the unification of North and South Yemen in 1990 and subsequent civil war. Adding to the already large temporal deviation, the pattern matching algorithm was unable to process the event data from these years because not only was the
demonstration for leadership change event not present in the match to these datasets, neither was
the first event in the pattern – accuse of human rights abuses. Even with an incomplete pattern
input to the algorithm, the temporal deviation is extreme, which may indicate an additional
disruptive even between the years 1994 and 2004 to cause such an unexpected change in the
expected pattern.

Figure 5-4. (a) Temporal edit distance for Yemen data in 1993-1994. (b) Temporal edit distance
for Yemen data between 1986 and 1993. Political and social structure in Yemen took dramatic
turns in 1990 (unification of North and South Yemen) and in 1994 (civil war). Additional metrics
for understanding temporal deviation are included in the distance metric window.

5.4 Discussion

This modified sequence alignment procedure generates results that should not be
surprising for an analysis of a country’s change over time. Yemen and its former separate
components have undergone much political, structural, and social change over the course of the
last 30 years. I have attempted to capture periods of rapid change in the country in which similar
patterns of transition would be expected to occur, and quantify their differences based on the
change observed between successive instances of a particular pattern. Not surprisingly, that
expected pattern changes dramatically over time. What I did not expect was for the deviations in
the expected pattern to nearly perfectly follow an inversed exponential decay curve with respect
to elapsed time. The amount of deviation from the expected pattern as the temporal reference
point gets further away from that from which the pattern was discovered, the deviation between the observed pattern and the expected pattern exponentially rises. Pattern deviation is the inverse of pattern similarity, so an exponential growth curve of deviation is the same as an exponential decay curve of similarity. Compare the graphs in figure 5-5, where 5-5 (a) is a graph of the above four examples of different temporal distances, and 5-5 (b) is an unweighted exponential decay curve.

![Graphs](image)

**Figure 5-5.** (a) Temporal analysis of expected pattern from 2011-2012 to previous years. (b) Unweighted exponential decay curve. Because distance and similarity are inverses, these curves appear equivalent, and may indicate temporal decay as a concept equivalent to distance decay.

To bring this temporal study back to a more geographically relevant context, I want to focus on the utilization of distance as a way to understand deviation. Temporal distance drives the difference observed between instances of a pattern, just as spatial distance is a determining factor in the similarity of locations. Not only is Tobler’s First Law of Geography (Tobler 1970) at work here in reference to the temporal variable, but the comparison can be taken one step further and applied to spatial distance decay. The equivalence between the graphs in figure 5-5 shows that when an exponential decay function is applied to spatial variables – creating a function of
distance decay – time and space can be seen to have the same effect on the similarity of certain phenomena, in this case the observance of a similar pattern.

Time’s effect on the constitution of a pattern is intuitive. As Hägerstrand determined, both public and private forms of communication influence the spatio-temporal dynamics of the movement of information among individuals (Hagerstrand 1967). The technological means for personal and communication alone have changed dramatically in the time period examined in this chapter, though the use of electronic communication devices in Yemen is low due to low average income (The World Factbook 2013). Add the influence of a more accessible and diversified public media following unification, and Hägerstrand’s public and private sources appear to be increasing in influence while pattern deviations decrease. Individual events, such as unification in 1990, also exhibit the potential for instantaneous change in the political and social structure of the country. While this pattern appears to exist in varying degrees with respect to changing time over the course of Yemen’s modern history, the next chapter attempts to investigate changing patterns as a function of location.
Chapter 6

Intersecting Sequence Alignment and Comparative Politics: Comparing Yemen Patterns over Space and Place

This chapter returns to the area of study known as comparative politics. The goal of comparative politics is to understand the way that the political system of a country or political entity works through examination of its similarity to others’ systems. Because it is functionally impossible to do a comparative analysis empirically based on perfectly comparable conditions, the strategy involves making inference about reaction, decision-making, and cause and effect (Landman 2003). Generally, a strategy of comparison begins with an understanding of the expectation of one system, and that expectation is compared against the relatively unknown state of another system. The expectation of some similarity is an important assumption that makes this strategy possible. The extended sequence alignment procedure that I present here can contribute to the process of political comparison by emphasizing expected events and patterns of them as a function of their similarities to events in a political system of interest.

A comparative political approach to understanding a political system relies heavily on observed patterns to correlate events and situations. The interrelatedness of events that comprise a pattern through time is an important indicator of the processes that generate that pattern. Introduce temporal specificity to the pattern in the form of event timing and expected intervals between successive events and a single pattern represents a strong correlative relationship about the way that a political system functions and evolves.

Where a specific, observed pattern says important things about the way that a given system functions, comparative politics attempts to understand political structure through comparison with other, perhaps less-understood examples. Comparative politics is not defined by
any single political analysis approach, but is generally viewed as a broad heading under which researchers explore international and domestic politics based primarily on observable and explainable differences (and similarities) between multiple states. I believe a comparative approach to analyzing patterns is necessary for both a validation of a pattern that has is assumed to exist given a specific context, and as a tool to explore the differences between places as a result of different political patterns. This chapter explores political systems in several countries based on the distance between an expected pattern and the observed events that presumably generate similar pattern instances.

6.1 Precedence of a Political Pattern

For this spatial political comparison, I use the same pattern of diplomatic cooperation that was presented in the previous chapter. Refer to figure 5-1 for a graphic representation of the pattern. To quickly review, this pattern was extracted from a dataset of events that occurred during the time-span of February 2011 through March 2012 that either took place in Yemen or that in which a Yemeni actor was involved. The pattern was mined using the STempo spatio-temporal pattern discovery environment (Peuquet 2012). The fact that this pattern was found to be significant by STempo means that it occurred as defined several times in the course of the events contained in the dataset – which provides a form of in-sample pattern validation as well as evidence of recurrence.

Because pattern detection procedures like STempo rely on unstructured sets of independent events for discovering patterns, the patterns that arise are often determined statistically instead of logically. The pattern I use in this demonstration, however, provides a good example for making inferences about the ways that political systems compare. The diplomatic nature of the negotiation process present in the pattern suggests that this pattern indicates a
significant international component to the political events occurring in Yemen during this time frame. In a temporal sense, deviations from this pattern’s expected set of intervals would be important indicators of the international cooperation capabilities of the examined state with respect to Yemen. Additionally, the lack of an expected event in its expected location in a pattern indicates more deviation in the form of a similar substituted event or a completely skipped event. The compounding of factors leads to a complex situation of deviating conditions from that which was expected. Quantity of deviation is made known through the Temporal Deviation Distance, and the location of deviation is examined using the alignment view in my extended process.

6.2 Comparing the Pattern to Other Spatial Contexts

To demonstrate the opposite situation examined in the previous chapter, this chapter compares Yemen’s pattern in February 2011 through March 2012 to data defined by different spatial locations at the same span of time. Where the Temporal Deviation Distance proved useful in the previous example for comparing pattern decay over increasing temporal distance, this analysis uses the same technique while substituting temporal distance for spatial distance. 2011-2012 was a period of intense conflict in Yemen, which contributes fairly strongly to the composition of the observed diplomatic pattern, so countries that experienced similar events during that time were chosen for comparison of their event patterns. Based on my perception of events during the time of the Arab Spring, I chose Libya, Sudan, and Syria as the countries with the most similar time scale and intensity of internal conflict in which I predict to find similar pattern trajectories. Additionally, Yemen’s most spatially similar countries of Oman, Saudi Arabia, The United Arab Emirates, Qatar, and Bahrain are also compared against the expected pattern. Table 6-1 shows each of these countries as a function of the Euclidean distance from their centroid to Yemen’s.
<table>
<thead>
<tr>
<th>Country Name</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bahrain</td>
<td>738</td>
</tr>
<tr>
<td>Libya</td>
<td>2144</td>
</tr>
<tr>
<td>Oman</td>
<td>636</td>
</tr>
<tr>
<td>Qatar</td>
<td>699</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>617</td>
</tr>
<tr>
<td>Sudan</td>
<td>1239</td>
</tr>
<tr>
<td>Syria</td>
<td>1455</td>
</tr>
<tr>
<td>UAE</td>
<td>644</td>
</tr>
</tbody>
</table>

Table 6-1. Countries compared to Yemen in 2011-2012 as a function of distance. Distance is measured in miles from centroid to Yemen’s centroid.

GDELT was used in a similar way to extract the necessary data as presented in the previous chapter. The GDELT dataset’s (Leetaru and Schrodt 2013) global coverage provides the biggest advantage, as subsetting the full data by time and by country is a simple task through scripting in Python. Because of the disparity in the amount of political activity between my chosen countries, files range from 3100 events in Oman to 345,000 events in Syria. The concentration of events per day greatly influences the likelihood that pattern matching will find matching events and patterns, though when that concentration is a factor of the events that occurred, and not of a disparity in the number of sources used to generate those events, then the number of events reflects the reality of occurrences in those places.

Predicting the outcome of this spatial analysis is a difficult task. A geographer not familiar with the area might be apt to assume that Yemen’s closer neighbors are more likely to experience similar patterns of diplomacy. However, those countries did not experience the same
intensity of political conflict as Yemen or the other analyzed countries, nor do they exhibit very similar social organizations. Economically, these countries also present two distinct groups of low-income (Yemen and other conflict-ridden countries during 2011-2012) and mid- to high-income (oil-rich Arabian Peninsula countries) countries. Other factors, such as type of government, ethnic division, and urban-rural population vary widely between these chosen countries (The World Factbook 2013) and produce very different political and social environments that affect the trajectory of patterns.

6.3 Results of Spatial Comparison

6.3.1 Temporal Deviation Distance

For this spatial comparison, the Temporal Deviation Distance alone is not sufficient to explain the deviation from Yemen’s pattern to the events in seven other countries. More accurately, there is no correlation between the TDD and the spatial distance between Yemen and its neighbors. Figure 6-1 shows a graph comparing spatial distance with the TDD of the seven compared countries. The figure shows two clusters of observations – indicated in quadrants II and IV from the axes inset in red onto the graph – one that includes Yemen’s nearest spatial neighbors (quadrant II) with a significantly higher TDD, and the remaining countries with the most similar political atmosphere to Yemen despite their larger Euclidean distance (quadrant IV).
Spatial comparison seems to indicate that more is at play in understanding country-scale political similarity with respect to the patterns that occur than spatial distance. This should not be entirely unexpected, though potentially specific to a comparison originating from a pattern in Yemen. Yemen’s is an isolated country in that its nearest neighbors are quite different in their political and social structures, while more similar countries such as the ones observed to have the least Temporal Deviation Distance have much more in common with Yemen. Later in this chapter, I will investigate this observation using current comparative politics literature. Saideman (2012) similarly observes a non-homogeneous spatial pattern of regime change during the Arab Spring, suggesting that spatial diffusion of violent demonstration in the sense of distance-decay was not a process at work during this time period. These TDD results do not facilitate any assumptions of policy diffusion, which should not be surprising given the differing levels of political and social context in Yemen compared to its neighbors. However, this should not be seen as indicating any general process because of the particular differences between Yemen and its neighbors. Much of that difference can be attributed oil export (Hess 2013)

Figure 6-1. Graph of spatial distance to Temporal Deviation Distance. No clear correlation exists in the same way that temporal distance correlated positively with increasing TDD. Other non-spatial factors seem to be bigger indicators of pattern dissimilarity in country-level politics.
6.3.2 Alignment View

With the Temporal Deviation Distance and spatial Euclidean Distance providing little evidence to explain the observed trends in pattern similarity, the alignment view provides additional evidence of the deviations specific to each comparison. In this example, the alignment view examines countries from each of the two clusters identified in figure 6-1 to find similarities in the ways their events match, and correlate them with the political and social differences between those countries.

Among the comparisons between Yemen and its closest neighbors on the Arabian Peninsula (which share the largest TDDs with Yemen), a couple of common characteristics are obvious. The most noticeable deviation that occurs among these countries’ pattern matches is that the large deviations occur in a more clustered manner, rather than spacing evenly among all of the elements of the pattern. My observation has been that this is the case in most pattern matches, but with TDD values around 15 and as much as 42 days, those clusters of additional days beyond the expected temporal interval become more significant.

There are many potential explanations for this clustering tendency, and it depends on the types of events between which clusters of deviations occur. In the case of Yemen’s neighbors, a cluster of deviation consistently occurs between the ‘demonstrate for leadership change’ (1411) event and the ‘engage in negotiation’ (46) event. See the alignment examples from Bahrain and Qatar in figure 6-2. The nature of what is likely to take place between these particular events and the extended time taken for them to transpire could be interpreted as providing indication of a more stable political situation than that of Yemen, Syria, Sudan, and Libya. For instance, additional days between these events could be the result of less intense demonstrations, quickly-quelled demonstrations, a leadership more able to endure demonstrations, or simply unrelated events given their large temporal separation. Positive identification of the reasons behind this
observed temporal deviation without a comprehensive examination of every event preceding, intervening between, and following these expected events, would be time consuming and prohibitively difficult. Therefore, inferences are necessary correlating the observed TDD with political differences between countries.

Figure 6-2. Alignments of expected pattern to Bahrain (a) and Qatar (b). Both Bahrain and Qatar share the same TDD of 15 days. As the alignment show in the red intervals preceding the ‘engage in negotiation’ (46) event in (a) and following the ‘demonstrate for leadership change’ (1411) in (b), most of the temporal deviation in these countries’ comparisons (as well as the other countries close to Yemen) occur between these two expected events.

With reference to the countries that share more similar characteristics with Yemen, the alignment view shows similar trends that are common among all three compared countries. Like the previous example, temporal deviations tend to cluster between two events, rather than space evenly throughout the pattern, though in a less significant manner since the deviations are so small. What is significant is that the more politically-similar countries to Yemen contain a higher likelihood of negative deviation – or instances of a pattern occurring more quickly than anticipated. Libya experienced a perfect match of the expected pattern, but examples of the alignment view for Sudan’s TDD of 3 (though a subtraction of three from the expected time) and Syria’s TDD of 1 (again, subtracted) in figure 6-3 show the events around which the deviation
occurs. Notice that the deviation occurs in the same interval in both comparisons – between the ‘make optimistic comment’ (13) and ‘reject plan/agreement to settle dispute’ (127) events.

Figure 6-3. Alignments of expected pattern to Sudan (a) and Syria (b). Both countries have a similar TDD (Sudan = 1, Syria = 3), but they share the fact that the deviation is negative, indicating a quicker-than-anticipated pattern occurrence. Additionally, the deviation occurs in the same interval between ‘make optimistic comment’ (13) and ‘reject plan/agreement to settle dispute’ (127).

6-4 Discussion

This chapter examined the difference between countries with respect to how well an expected pattern fits in a dataset of observed events. Despite the small sample in each group of countries representing similar and dissimilar pattern matches, the TDD and alignment view show that members of each group share common characteristics of pattern matches. Perusing the comparative politics literature also shows that within those two groups of similar TDD and matching characteristics, countries share very similar structures when it comes to several political and economic factors. Several authors identify the strength of an authoritarian regime as an important correlate to a country’s likelihood of experiencing violent rebellion on par with the events in many Middle Eastern countries during the Arab Spring (Saideman 2012). Among the countries listed as having similarly weak authority over their citizens and military (the primary
force behind successful uprisings), Dalacoura (2012) identifies Yemen with Libya and Syria, confirming part of a cluster detected using the TDD (Sudan is a unique case because it is the only country not in the Middle East and therefore outside of the scope of Dalacoura’s study). Interestingly, Dalacoura places Bahrain in the same category, though she does qualify that classification on the fact that the government of Bahrain quickly and quietly repressed the growing opposition movement.

In the alternate situation, oil serves as the primary bond between countries that exhibit greater pattern deviation from the expected pattern observed in Yemen. Hess (2013) identifies Saudi Arabia, Qatar, The United Arab Emirates, and Bahrain among several countries defined by their high oil export value and similar lack of threat from a rebellious citizen force. Hess (2013, 9) sums this argument up nicely: “Clearly, an economy that is highly dependent upon oil exports offers major benefits to an autocratic ruler” such as those in place throughout the countries of the Middle East. With the clear disparity in oil production and economic prosperity resulting from it present between the groups identified in this spatial analysis, the Temporal Deviation Distance reinforces the separate conclusions of Hess, Dalacoura, and Saideman in a quantitative way.
Chapter 7

Discussion and Concluding Remarks

Having explained the history of sequence alignment, how it is performed, the modifications that I have made to the original method, and demonstrated its use on events in different temporal and spatial contexts, I would like to provide some summary comments. First, I will explain what I think are the primary contributions – intellectually and in the real world – then I will describe some of the limitations of this method. Finally I will outline some future work and conclude the thesis.

7.1 Contributions

The primary contribution in extending this sequence alignment method for spatio-temporal analysis is in the quantification of a pattern’s fit. Despite our ability to recognize and describe patterns, large datasets make the task prohibitively complex without the help of an automatic process. This thesis introduces modifications to the sequence alignment algorithm to accommodate large, temporal datasets and extends Levenshtein’s edit distance to describe the temporal deviation between an expected pattern and a sequence of unexplored events, creating the Temporal Deviation Distance. Temporal units allow this method to be used in a standalone fashion, where any alignment does not need to be compared with others to get a sense of the degree to which observed events fit an expected pattern.

The confirmatory nature of this technique contributes to knowledge discovery, data mining, Big Data, and theory-based spatio-temporal analysis. Since confirmation is necessary for patterns known, or believed to exist through pre-existing knowledge, as well as patterns discovered
through data mining, this method serves a useful purpose for the general goal of confirming and validating expected patterns. As we are increasingly able to represent human dynamics with publically-generated and publically-available social data – a growing trend across academia known as Big Data – pattern recognition techniques, such as the one introduced in this thesis, are increasingly necessary to sort the meaningful from what amounts to mostly irrelevant information. While patterns are understood differently in both the theory-driven and data-driven approaches, this method serves as a means to further understand those patterns in a uniform way.

Using a broad definition of a sequence, sequence alignment has been used in several fields. Geography began to adopt the method in a limited sense in the last five years to analyze movement sequences. Sociology had previously made a successful transition from rigid DNA sequence comparison to comparison of temporally-organized sequences. The extensions that I have presented here take sequence alignment further toward a goal of general usability.

This Masters thesis research also contributes to Geography’s emphasis on spatio-temporal phenomena. While Hägerstrand’s on spatio-temporal dynamics work spanning four decades, his “time-geography” says very little about processes of events through time and space. Since Hägerstrand’s seminal piece introducing time geography (Hagerstrand 1970), quantitative spatio-temporal research has focused primarily on tracking, grouping, and analyzing movement, which represents only a fraction of topics that deal with spatio-temporal variables such as the event patterns explored here.

In the real world, this extended sequence alignment method has proven to be a useful tool for comparing patterns of events. The Temporal Deviation Distance provides a simple measure of not just the similarity between two patterns, but how much time had elapsed between an expected series of events of different types and what actually occurred. Using the TDD to compare multiple different sets of events as they change through time and space generates a unique historical/spatial analysis of change in reference to an expected event pattern. I have
demonstrated that differences in the TDD can be correlated with other factors, in this case political ones such that a quantitative comparative political analysis can be performed. As demonstrated in chapters five and six, the novelty of using the alignment view to further understand the observed differences indicated by the TDD. The alignment view shows where deviations occur in the course of an unexplored sequence of events, which indicates the particular events that did not occur as expected. The specific locations and concentration of deviations within a sequence themselves indicate patterns. Clusters of deviations in the same location show something very different than deviations that occur more regularly throughout the sequence. I have shown that this is the case by comparing multiple countries with varying political structures, which exhibit very different deviation patterns after aligning them with an expected pattern originating from events in Yemen.

7.2 Limitations and Future Work

Although the modifications that I have introduced in this chapter are necessary for creating a tool that can compare patterns of spatio-temporal events, there a couple of limitations to this approach that are worth discussing. The first topic deals with the treatment of events and how they are classified. The second topic covers issues with temporal representation and scale. The final issue that requires discussion is computational complexity.

7.2.1 Data Quality

Obviously, the way that events are classified into types makes a big difference on the content of a pattern. This creates more of a problem for pattern discovery techniques than pattern matching, but is an issue that still effects the conclusions that can be made using my extended
procedure. While pattern confirmation remains unconcerned with the method used to classify events, it does require that pattern elements and unexplored events be classified using the same coding scheme for the purpose of matching equivalence of two event designations. While limiting the applications for which this extended method can be used, the need to consistently classify events also leaves the usefulness of this method at the mercy of the data generation process and the pattern definition phase. That includes how each event is classified, how free of spelling inconsistencies the unique identifiers are, and how detailed the data is.

As mentioned previously, this technique is very sensitive to the ratio of number of events to days contained in the datasets that it compares. Similar to one of the common concerns about big data, this limitation proves that, given enough data, any pattern one wishes to find can be found. Inevitably, some types of events are more common than others, so whenever an expected part of a pattern is common, many potential pattern matches can be found. However, to assume that every compared set of data contains the same concentration of events and the same types of events would not be accurate. The countries, time frames, or other contexts represented by compared datasets are expected to experience different types of events at different frequencies, which generates the deviations that this method seeks to explore.

### 7.2.2 Temporal Scale

Another limitation of this method concerns the temporal scale of events and patterns. This method assumes that all events occur at the same scale, or similarly, that all events that occur at the same scale are the only events of interest. In reality, both interpretations are not true. Because political and diplomatic events derived from RSS news feeds contain a single date, though often to not have any indication of time within the daily scale, the day is the lowest temporal scale possible for representing the entire datasets compared here. Similarly, events that
span intervals of time greater than a day are difficult to detect because reporting on those events is also inconsistent such that a multi-day event cannot be deciphered even by associating multiple news stories with the same event.

Event data with multiple temporal scales contained within it would be difficult for this method to handle. The patterns that this method is built to handle are defined by the transition intervals between expected events, which are much less clear with multi-temporal scale events. There is precedence for thinking about how the many potential relationships between events would be characterized in a pattern, and Allen’s (1983) work on the subject would be most useful for resolving this issue. Allen defines a series of potential relationships between two events with respect to when they occur in comparison to one another, however, those relationships do not consider the absolute (vs. relative) time at which events begin and end. Therefore, a pure interpretation of Allen would render the Temporal Deviation Distance useless. However, the sequence alignment algorithm has been modified here such that it that allows for simultaneous events, so it is conceivable that it could be extended further to include events “contained by,” “overlapping,” “meeting,” “starting,” and “finishing” one another in addition to its current capability (Allen 1983).

### 7.2.3 Algorithm Complexity

Much work has been done by other researchers to improve the computational complexity of the sequence alignment process. Improvements have been made such that multiple sequence alignment (aligning more than two sequences at once) can be accomplished in better than \( O(L^m) \) time where \( L \) is the length of the largest sequence and \( m \) is the number of sequences (Edgar 2004). For ease of programming, however, this modified method implements Needleman and Wunsch’s original algorithm without use any of the improvements that were introduced later.
The original algorithm, as described by Needlemen and Wunsch, is accomplished with $O(M^2N)$ complexity, where $M$ and $N$ are the sizes of the two compared sequences. In addition, the temporal interval searches and multiple optimal column representation extensions presented here significantly increase the amount of time and comparisons needed to align two sequences. In practice, the number of comparisons needed to complete a pattern match is very variable, based on the length of temporal intervals between each pattern element, the number of events per unit of time, and the length of both the pattern sequence and the unexplored sequence. The alignment view, which requires writing both sequences in their entirety to the display, also requires significant memory allocation. All of these factors add significant processing time to the pattern matching procedure.

While this method did not implement any existing means for improving sequence alignment’s efficiency for the sake of maintaining conceptual simplicity, in the future, this pattern matching process will benefit from improvement to its computational complexity. Many have implemented their own improvements to the sequence alignment algorithm to speed up the process, which can be drawn from in this process. One of the first improvements at the fundamental level was to use one-dimensional arrays and recursion (Gotoh 1982), which would be the simplest way to decrease computation time of the process introduced in this thesis. Farrar implement sequence alignment such that it can be accomplished by dividing the instructions to parallel processors (2007). Many others have made improvements in the area of sequence alignment’s complexity. Much of the complexity of this extended procedure is a result of my own simplification of Java code, which is why the biggest improvement that I can make to my algorithm is by incorporating more efficient programming techniques, such as more efficient drawing to the display, and using non-standard data types that contain their own efficient methods for doing much of the processes that I have built myself.
7.3 Additional Future Work

While I have attempted to address several limitations to this extended sequence alignment process and how I might use future work to remove them, there are a few other improvements that I would like to accomplish in the future.

7.3.1 Elemental Similarity

The first such improvement is an enhancement to the way that sequence alignment represents the similarity between compared elements. As briefly mentioned in chapter three, elemental similarity can be represented in more degrees of detail than the ‘same’ or ‘different’ binary that I use. In the matrix generation phase of the procedure, a “1” is placed in a cell that represents equivalent event types in its corresponding row and column and a “0” anywhere else. Many applications of sequence alignment, however, use a range of numerical values to represent partial similarity between types of elements (Dodge, Laube, and Weibel 2012). Sequence alignment could utilize this more nuanced comparison of events to help manage the issues of inconsistent event-coding between datasets as well as improve upon mismatched scenarios.

While incorporating elemental similarity would not present an extensive change to the code, the theory behind it is lacking. An spatio-temporal event extension that utilized elemental similarity would require a semantic network connecting individual event types to one another with varying degrees of similarity. To claim that event type “A” is more similar to event type “B” than it is to event type “C” presents a theoretically tricky situation. Political science presents a viable field of study where a sufficient understanding of the events that I use here could originate to construct such a semantic. I anticipate pursuing the avenue of political science theory to improve my work here in a theoretical way, for the goal of representing elemental similarity.
7.3.2 Spatiality

While the modifications introduced here emphasize the temporal aspect of events and patterns, sequence alignment can also consider the space in which events occur. It would be simple to say that events have a location, given by latitude and longitude coordinates, and exhibit spatial patterns as events transpire over time. In that case, sequence alignment could be further extended to search expected spatial intervals for matching events in a way that nearly parallels that of a temporal interval search. Varying degrees of event match would therefore be very important, and the TDD would become the Spatio-Temporal Deviation Distance. However, the spatiality of events may not be so straightforward. Many events exist at a spatial scale beyond that of a single latitude and longitude location, and patterns may not follow any traceable spatial sequence as they proceed through time. I consider the spatial scale of events and how it fits into the temporal organization of patterns to be an issue in need of further thought, as I am unsure if the added complexity of spatiality provides increasingly accurate portrayal of a pattern.

7.3.3 Prediction

My ultimate hope of continuing this research into spatio-temporal patterns is to facilitate predictive analysis. Prediction comes in two forms as it relates to this work: anticipation of how well an expected pattern will match a new context, and forecasting individual events that are part of a pattern. Both cases require a deeper theoretical foundation in the drivers of particular patterns. Predicting the trajectory of an entire pattern requires foreknowledge of the difference between two times and places how that difference will affect the flow of the pattern. While both chapters five and six show how time geography and comparative politics, respectively, provide a backdrop for highlighting how and why patterns change, those analyses represent a simplified and
insufficient means for being able to quantify the differences that generate the changes in a way that provides for accurate prediction. One way that I expect pattern matching can be used immediately to facilitate prediction of a single event in time is by aligning a partial pattern. For instance, using the example pattern of diplomatic cooperation from Yemen examined in chapters five and six, the first four elements (‘accuse of human rights abuses,’ ‘demonstrate for leadership change,’ ‘engage in negotiation,’ and make optimistic comment’) represent a partial pattern that, when aligned, provide hints about the presence of the fifth element: the ‘reject plan/agreement to settle dispute’ event. Many variables exist that make this a very uncertain process, but the Temporal Deviation Distance for the partial pattern partially indicates what to expect for the timing of the final event.

7.4 Conclusion

I conclude this thesis by saying that this extended sequence alignment procedure contributes a number of important features to current efforts in spatio-temporal analysis. The extended sequence alignment process that I introduce here builds on and contributes to several bodies of literature, including the progress of sequence alignment, temporal geography, pattern analysis, historical analysis, and comparative politics. Although the process of aligning sequences is simple in concept, the modifications that I have made to adapt it to temporal sequences significantly increases its complexity. I have demonstrated the new procedure on multiple time periods in Yemen to shown that the Temporal Deviation Distance parallels a pattern of distance decay, and on multiple spatial contexts to show that political patterns seem to be dependent on the political structure of a state. In the future, this method can be used to match additional spatio-temporal data, consisting of political events or otherwise, to contribute to a process of predicting the time and place of important events.
Appendix

Pseudocode

Algorithm: Similarity Matrix Generation
Input: PatternSequence, UnexploredSequence, empty matrix
begin
1: for (index row)
2:   for (index column)
3:     if (PatternSequence at row and UnexploredSequence at column are =)
4:       then set matrix cell at index (row, column) = 1
5:     else
6:       then set matrix cell at index (row, column) = 0
7:   end for
8: end for
end
Output: full matrix

Algorithm: Traversal Matrix Generation
Input: PatternSequence, UnexploredSequence, SimilarityMatrix, empty matrix
begin
1: for (index row = bottom row)
2:   for (index column = last column)
3:     /* max is the maximum reachable value from this cell*/
4:     set max = 0
5:     get optimal columns
6:     for (index optimal column)
7:       if (cell at (row+1, optimal column) > max)
8:         then set max = cell
9:     end for
10: end for
11: get similarity value = SimilarityMatrix at index (row, column)
12: set matrix cell at (row, column) = max + similarity value
13: end for
end
Output: full traversal matrix

Algorithm: Traversing the Matrix
Input: PatternSequence, UnexploredSequence, TraversalMatrix
begin
1: set max to 0
2: for (cell in first row and column)
3:   if (cell > max)
4:     set max to this cell
5:     remember cell’s index
6: end for
7: for (each cell skipped in a row or column)
8:   add a gap to correspond to the skipped value
9: if (elements in both sequences corresponding to the cell are =)
10: set color to matched color
11: else
12: set color to mismatched color
13: set current cell to stored cell’s index
end

Input: current cell
begin recursive process
1: set max to 0
2: for (each cell in current cell’s optimal columns (given row’s interval and column’s date)
3: if (cell > max)
4: set max to this cell
5: remember cell’s index
6: end for
7: for (each cell in the sub-optimal columns in the previous row (current index +1)
8: if (cell > max)
9: set max to this cell
10: remember cell’s index
11: end for
12: for (each cell in optimal cols in sub-optimal rows)
13: if (cell > max)
14: set max to this cell
15: remember cell’s index
16: end for
17: if (max was in optimal column)
18: if (matched col index < current cell index)
19: reorder matrix so that matched col is > current cell
20: for (each intervening element)
21: insert zero-weight gap
22: end for
23: if (elements are =)
24: set color to matched color
25: else
26: set color to mismatched color
27: else if (max was in sub-optimal column)
28: for (each intervening element outside of optimal cells)
29: insert weighted gap
30: count number of days, not the same as number of cells
31: add number of days to TDD
32: end for
33: if (matched elements are =)
34: set color to matched color
35: else
36: set color to mismatched color
37: else if (max was in sub-optimal row)
38: for (each skipped row)
39: insert weighted gap
40: count day intervals between preceding and next elements
41: add number of days to TDD
42: end for
43: if (elements are =)
44: set color to matched color
45: else
46: set color to mismatched color
47: set current cell to remembered index
48: repeat recursive procedure with new current cell
end
Output: TDD, aligned pattern and unexplored


Laube, P. 2005. Spatio-Temporal Data Mining - Coping with the Increasing Availability of Motion Data in Geography. The 17th Annual Colloquium of the Spatial Information Research Center, Dunedin, New Zealand, November 24-25
Mavoa, S., M. Oliver, K. Witten, and H. M. Badland. 2011. Linking GPS and Travel Diary Data Using Sequence Alignment in a Study of Children's Independent Mobility. international Journal of Health Geographies 10:64.


