Integrated data-driven techniques for environmental pollution monitoring

Swathi Pakalapati
Louisiana State University and Agricultural and Mechanical College

Follow this and additional works at: https://repository.lsu.edu/gradschool_dissertations

Part of the Chemical Engineering Commons

Recommended Citation
https://repository.lsu.edu/gradschool_dissertations/675

This Dissertation is brought to you for free and open access by the Graduate School at LSU Scholarly Repository. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Scholarly Repository. For more information, please contactgradetd@lsu.edu.
INTEGRATED DATA–DRIVEN TECHNIQUES FOR ENVIRONMENTAL POLLUTION MONITORING

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in
The Cain Department of Chemical Engineering

by
Swathi Pakalapati
B/M.Tech., Indian Institute of Technology Madras, 2004
August 2009
Dedicated to the “hope” – the hope of waking up to a religion, caste, and gender discrimination free society in India
Acknowledgements

I would like to thank my research advisors, Dr. Jose A. Romagnoli and Dr. Ahmet Palazoglu for their patience in guiding me to be a good researcher. Thank you Jose and Ahmet for your support through the course of this work. This work would not have been possible if not for your valuable guidance. I would also like to thank Dr. Scott Beaver for his continued help on the project through my Ph.D. Scott, you have been and will always be an inspiration for me. Many thanks to you for answering my frequent questions long after you have finished working on your dissertation. Thank you for baby walking me and getting me started with my work. I truly appreciate your help in shaping up this dissertation.

I also wish to thank my committee members, Dr. Ralph W. Pike, Dr. Kalliat T. Valsaraj, and Dr. Arthur Penn (Dean’s representative) for giving their time and agreeing to serve on my committee. Thank you all for your efforts in reviewing my research.

I would also like to thank the former and current members of the Process Systems Engineering Laboratory (PSE) at both LSU and UC Davis for sharing their experience and the camaraderie. A special thanks to Robert Willis for always sharing his expertise in computers willingly with me. Thank you Robzy for your time and for listening to my ramblings. I would like to thank the Department of Chemical Engineering staff for always being there to help with administrative issues. I want to thank Dr. Antonella Bodini from the Institute for Applied Mathematics and Information Technologies (IMATI) for answering my questions very promptly. Thank you Antonella, your mails have helped me better understand few concepts really well.

I am grateful to my friends, Sandhya and Lavanya for always being my pillars of support. I can never put my feelings into words and thank you both enough for what you have done for me. Also, I would like to thank Amar, Bharat, Bratati, and Prathyusha for their help in making this journey an enjoyable one. I am truly grateful to my family for giving me the
opportunity to explore and help me succeed in my career and life. Thank you Amma and Nanna for encouraging me to pursue my career. I have learned to face difficult situations from you and be the best I can be. Thank you Suni and Bava, Tillu and Prasanna for being my best friends and for motivating me. Thank you for being a part of my life. Last but not the least, to Lintu for helping me destress with just your laugh.
# Table of Contents

**Dedication**........................................................................................................... ii

**Acknowledgements**.............................................................................................. iii

**Abstract**...................................................................................................................... viii

**Chapter 1 Introduction** ........................................................................................... 1
  1.1 Environmental Pollution Monitoring ......................................................... 1
  1.2 Intelligent Monitoring .................................................................................. 4
  1.3 Research Aims and Objectives ..................................................................... 6
  1.4 Thesis Contributions ..................................................................................... 7
  1.5 Structure of Dissertation ............................................................................. 8

**Chapter 2 Literature Review** ............................................................................... 10
  2.1 Models for Air Quality Analysis .................................................................. 10
  2.2 Models and Methods for Ozone Analysis ..................................................... 15
  2.3 Ozone Analysis and Prediction Studies ......................................................... 17
  2.4 Conclusions .................................................................................................. 19

**Chapter 3 Clustering Algorithms** ....................................................................... 20
  3.1 Introduction .................................................................................................... 20
  3.2 Background .................................................................................................... 21
    3.2.1 Data Handling ......................................................................................... 22
    3.2.2 The Measures of Distance and Similarity ............................................... 22
  3.3 Hierarchical Clustering .................................................................................. 24
    3.3.1 Hierarchical Clustering Background ...................................................... 24
    3.3.2 Agglomerative Hierarchical Clustering ................................................... 26
    3.3.3 Cluster Validation Measures .................................................................. 27
  3.4 Non–Hierarchical Clustering ......................................................................... 28
    3.4.1 Non–Hierarchical Clustering Background .............................................. 28
    3.4.2 Non–Hierarchical $k$–means Clustering .................................................. 29
  3.5 Hierarchical Ensemble of $k$–means Clustering ............................................. 30
  3.6 Dynamic Principal Component Analysis for Time Series Clustering .......... 32
    3.6.1 Dynamic Principal Component Analysis ............................................. 33
    3.6.2 Time Series Clustering .......................................................................... 35
  3.7 Conclusions ..................................................................................................... 36

**Chapter 4 Sequence Analysis** ............................................................................. 37
  4.1 Introduction ..................................................................................................... 37
  4.2 Background ..................................................................................................... 38
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5</td>
<td>Cluster Interpretation</td>
<td>112</td>
</tr>
<tr>
<td>7.5.1</td>
<td>Hourly Wind Clusters</td>
<td>112</td>
</tr>
<tr>
<td>7.5.2</td>
<td>Daily Wind Field Groups</td>
<td>113</td>
</tr>
<tr>
<td>7.6</td>
<td>Ozone Response to Meteorology</td>
<td>119</td>
</tr>
<tr>
<td>7.7</td>
<td>Conclusions</td>
<td>125</td>
</tr>
<tr>
<td>8.1</td>
<td>Background</td>
<td>127</td>
</tr>
<tr>
<td>8.2</td>
<td>HMMs for Associating Surface Wind Field Patterns with Tropospheric Ozone</td>
<td>128</td>
</tr>
<tr>
<td>8.3</td>
<td>HMM Training</td>
<td>128</td>
</tr>
<tr>
<td>8.4</td>
<td>Model Validation</td>
<td>131</td>
</tr>
<tr>
<td>8.5</td>
<td>HMMs for Point—Value Predictions of Daily Maximum Ozone Concentrations</td>
<td>135</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Conditional Distribution of the Daily Maximum Ozone Concentrations</td>
<td>136</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Estimated Model</td>
<td>138</td>
</tr>
<tr>
<td>8.5.3</td>
<td>Goodness of Fit for the Estimated Model</td>
<td>142</td>
</tr>
<tr>
<td>8.5.4</td>
<td>Model Testing and Validation</td>
<td>145</td>
</tr>
<tr>
<td>8.6</td>
<td>Conclusions</td>
<td>151</td>
</tr>
<tr>
<td>9.1</td>
<td>Contributions</td>
<td>152</td>
</tr>
<tr>
<td>9.2</td>
<td>Recommendations</td>
<td>154</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>156</td>
</tr>
<tr>
<td>Appendix</td>
<td>Permission to Use Published Material</td>
<td>165</td>
</tr>
<tr>
<td>Vita</td>
<td></td>
<td>166</td>
</tr>
</tbody>
</table>
Abstract

The adverse health effects of tropospheric ozone around urban zones indicate a substantial risk for many segments of the population. This necessitates the short term forecast in order to take evasive action on days conducive to ozone formation. Therefore it is important to study the ozone formation mechanisms and predict the ozone levels in a geographic region.

Multivariate statistical techniques provide a very effective framework for the classification and monitoring of systems with multiple variables. Cluster analysis, sequence analysis and hidden Markov models (HMMs) are statistical methods which have been used in a wide range of studies to model the data structure. In this dissertation, we propose to formulate, implement and apply a data-driven computational framework for air quality monitoring and forecasting with application to ozone formation. The proposed framework integrates, in a unique way, advanced statistical data processing and analysis tools to investigate ozone formation mechanisms and predict the ozone levels in a geographic region. This dissertation focuses on cluster analysis for identification and classification of underlying mechanisms of a system and HMMs for predicting the occurrence of an extreme event in a system.

The usefulness of the proposed methodology in air quality monitoring is demonstrated by applying it to study the ozone problem in Houston, Texas and Baton Rouge, Louisiana regions. Hierarchical clustering is used to visualize air flow patterns at two time scales relevant for ozone buildup. First, clustering is performed at the hourly time scale to identify surface flow patterns. Then, sequencing is performed at the daily time scale to identify groups of days sharing similar diurnal cycles for the surface flow. Selection of appropriate numbers of air flow patterns allowed inference of regional transport and dispersion patterns for understanding population exposure to ozone. This dissertation proposes to build HMMs for ozone prediction using air quality and meteorological measurements obtained from a network of surface monitors. The case study of the Houston, Texas region for the 2004 and
2005 ozone seasons showed that the results indicate the capability of HMMs as a simpler forecasting tool.
Chapter 1

Introduction

1.1 Environmental Pollution Monitoring

Ozone is a secondary pollutant resulting from the photochemical reactions of volatile organic compounds (VOCs) with oxides of nitrogen ($NO_x$). Elevated concentrations of tropospheric ozone have adverse effects on human health, agriculture, and the environment (Seinfeld, 1998). The adverse health effects of tropospheric ozone around urban zones indicate a substantial risk for many segments of the population (Schlink et al., 2006). Many time series analyses have linked short–term exposure of ozone to premature mortality and morbidity (Bell et al., 2006; Huang et al., 2005; Stieb et al., 1996). Epidemiological studies show that prolonged exposure to elevated ozone concentrations has a greater risk of serious respiratory health effects and mortality than from occasional spikes (Galizia et al., 1999; Jerrett et al., 2009).

In light of the health effects to ground level ozone and other pollutants, the Clean Air Act (CAA) of 1971 requires the federal Environmental Protection Agency (EPA) to set the National Ambient Air Quality Standards (NAAQS) for six criteria pollutants: ozone, respirable particulate matter (PM), sulfur compounds, lead, carbon monoxide (CO), and nitrogen dioxide ($NO_2$). The forecasts from 1978 to 1997 for ozone were based on the 1–hr NAAQS which is 120 ppb. The mixing ratio of ozone in ppb relates the fractional concentration of ozone as the number of ozone molecules per billion air molecules. In 1997, EPA revised the standard to reflect the health studies that indicated risks associated with long term exposure to ozone. The 1997 air quality standard for ozone is based on the 8–hr time averaged ozone mixing ratio and the 8–hr ozone should remain below the threshold of 85 ppb for the attainment of the NAAQS. On March 12, 2008, EPA revised the 8–hr ozone
NAAQS to 75 ppb to protect public health. The methodology to calculate the exceedances of the 2008 ozone standard is the same as the 1997 ozone standard.

The ozone levels that exceed the NAAQS, however, are frequent in many urban regions in the United States. To implement the standards, EPA designates areas as attainment, nonattainment or unclassifiable (due to insufficient data to classify). Then, state governments are required to develop effective regulatory control strategies and attain these standards. State governments typically submit State Implementation Plans (SIPs) that outline pollution reduction strategies and the states are required to attain the standards by a date that EPA establishes as the dead line. The attainment of standards require detailed knowledge of the physical and chemical process affecting ozone accumulation in a region.

Precursor concentrations, meteorology such as wind speed and direction, temperature, solar radiation, cloud cover, humidity and the topography of a region influence the ozone build up on a given day. In the troposphere, photolysis of $NO_2$ results in the formation of ozone as

$$NO_2 \rightleftharpoons NO + O \quad (1.1)$$

$$O + O_2 \rightleftharpoons O_3 \quad (1.2)$$

In photolysis reactions, molecules break down into smaller units through the absorption of sunlight. Thus, the presence of $uv$ light is important for these reactions.

Reactions (1.1) and (1.2) result in relatively low levels of ozone as once formed, the ozone reacts with NO to re–generate $NO_2$ (1.3).

$$O_3 + NO \rightleftharpoons NO_2 + O_2 \quad (1.3)$$

These background mixing ratios for ground level ozone remain at low levels in North America (Vingarzan, 2004). However, in polluted air conditions, high levels of anthropogenic VOCs and $NO_x$ shift the equilibrium of the above reaction to favor ozone formation. The hydroxy (OH), peroxy (RO$_2$), alkoxy (RO) and acyl peroxy radicals (RC(O)O$_2$) in the atmosphere
that result from the volatile organic compounds (VOCs) in an organic/$NO_x$ system contribute to much of the chemistry in the troposphere. These peroxy radicals efficiently convert NO to $NO_2$ as long as $NO_2$ levels are sufficiently high. The competition for the hydroxyl radical determines the behavior of VOCs and $NO_x$ in ozone formation. While increasing VOCs mean more ozone, increasing $NO_x$ may lead to either more or less ozone depending on the prevailing VOC to $NO_x$ ratio. Thus the rate of ozone production is not simply proportional to the amount of $NO_x$ present. At a given level of VOC, there exists a $NO_x$ concentration, at which a maximum amount of ozone is produced. For ratios less than this optimum ratio, $NO_x$ increases lead to ozone decreases. Thus, in most of the tropospheric reactions, except in areas of strong sources of $NO_x$, the availability of $NO_x$ governs the ozone production.

Dennis et al., (1999) employ the measures of ozone concentration and the $O_3/NO_x$ concentrations ratio to assess the sensitivity response of the photochemical system on an ozone response surface. Their study concludes that the $O_3/NO_x$ measure successfully predicts the direction and degree of change in the models while the ozone concentration measure does not. This indicates that reducing emissions alone might not contribute to reducing the ozone concentrations in a region.

Elevated ozone concentrations are associated with warm temperatures and high solar irradiation. Meteorologically, high temperatures are associated with high pressure, stagnant conditions that lead to suppressed vertical mixing and elevated $O_3$ levels. Vertical mixing is the distribution of ozone mixing ratios vertically in the atmosphere. Emissions of biogenic hydrocarbons and evaporative emissions of volatile organic compounds increase with temperature; these increases in emissions are associated with increased ozone. The removal path for ozone depends on the concentration of water vapor; it is most effective in low latitudes at low altitudes where the radiation is intense and the humidity is high.

The lower threshold for 8–hr exceedances results in a wider range of weather conditions that lead to ozone production. Thus, broader range of weather conditions and not just the patterns that lead to extreme events need to be predicted. This implies that additional analyses need to be conducted to better understand the range of weather conditions conducive to ozone formation. The serious health effects associated with the 8–hr standard necessitates the
short term forecast in order to take evasive action such as reducing the emission–producing activities (e.g., driving, lawn mowing) on days conducive to ozone formation by carpooling etc. Therefore, it is important to study the ozone formation mechanisms and predict the ozone levels in a geographic region in relation to the meteorology of the region.

1.2 Intelligent Monitoring

Due to the advances in measuring and information technology, we are going through the era of data explosion, which has led to the related issue of information overloading. Today, in many aspects of our lives, we have highly automated data acquisition systems that monitor events occurring around us. In the specific case of air quality monitoring, surface meteorology and air quality measurements in most cities today are provided by monitoring networks operated by the corresponding environmental agencies. For example, in the Houston area, hourly surface wind speed and direction data are available from a network of 28 monitoring stations. The wind data alone are used to establish the air flow patterns. Temperature data are obtained from a set of 40 surface meteorological monitors. Networks of 20 and 27 Continuous Air Monitoring Stations (CAMS) monitor 8–hr ozone and 1–hr $NO_x$ levels, respectively. The Texas Commission for Environmental Quality (TCEQ) operates the CAMS that measure the air parameters. CAMS are shelters that have the measuring devices to measure the pollutant concentrations in ambient air. These instruments include but not limited to those that measure ozone, oxides of nitrogen, wind speed average, resultant wind speed and direction, outdoor temperature, relative humidity, solar radiation and barometric pressure.

As a consequence of the developments in continuous monitoring, enormous amounts of sensor data are routinely collected in real–time and historical data are available in abundance. The data–driven approach to intelligent monitoring seeks to construct a representation of a given system from a set of measurements that quantify its behavior, without explicitly assuming any a priori knowledge of the underlying phenomena.

However, the data modeling process for information retrieval and presentation, as well as the analysis of the implications directly from these measurements is daunting and remains a major concern. Remarkable characteristics of the data acquired are, in fact, redundancy and
possible insignificance, not to mention the presence of noise that corrupt the measurements. Very often, the amount and quality of the data together with their high–dimensionality can be a limiting factor for the analysis; therefore, the necessity for the availability of efficient and robust methods that:

i) Model the data in order to extract the structures existing in the measurements

ii) Identify the most relevant structures for the scope at hand

iii) Allow for easily interpretable display of data structures when the information is presented to the user for analysis

Cluster analysis, sequence analysis and HMMs are statistical methods which have been used in a wide range of studies (Beaver et al., 2006; Darby, 2005; Elzinga, 2003; Rabiner, 1989; Schliep et al., 2003; Won et al., 2004; Wong et al., 1998; Bellone et al., 2000; Betro et al., 2008) to obtain data–driven statistical models to represent the data structure.

Clustering methods aim at reducing the amount of data by grouping the observations into subsets, or clusters, consisting of similar observations (Everitt et al., 2001). Similarity refers to the property where observations belonging to a cluster should be as much as possible similar to each other, but differ significantly from the observations in the other clusters. In the metric context, similarity is searched among the observations from within–cluster distances and dissimilarity from between cluster distances. Within–cluster distances measure the degree of internal dispersion in the cluster, whereas between–cluster distances measure the separation among clusters.

Clustering algorithms use decisions based on distances to assign labels to a set of independent observations, typically generating classes that are characterized by their mean. Such models though computationally effective due to their simpler algorithms are inappropriate for time series (or sequentially correlated) observations and are not capable of representing physical processes that generate data dynamically.

Sequence analysis (Elzinga, 2003) is an approach to data reduction that is used to unveil the pattern and structure in sequence data. The sequence data is ubiquitous from customer shopping transactions to the biological DNA sequences (Dong et al., 2007). Sequence analysis involves calculating distances or similarities among sequences. This is a descriptive
methodology similar to Optimal Matching (OM) algorithms. OM algorithms seek to find the minimal number of edit operations (inserts, deletes and substitutions) necessary to turn a sequence into another sequence (Mount, 2004). The output of sequence analysis is a dissimilarity matrix between sequences that is similar to correlation coefficients. The output of sequence analysis is not accessed directly but is always combined with data reduction procedures such as cluster analysis.

A HMM is a double stochastic process that can effectively capture the time varying characteristics in a data set (Schliep et al., 2003). A HMM consists of two layers, an underlying hidden layer constructed from the Markov process and the observable layer that is the random realization of certain probability distributions corresponding to the hidden Markov process. HMMs parameterize a group of observations and model their probability density by taking the maximum likelihood approach. Given the parameters of a HMM and an observation, a HMM can effectively determine the class label for the observation by considering its probability of belonging to a group of parameterized observations. HMMs, though rich mathematically are computationally expensive models.

Thus the development of data–driven statistical models is an essential task in carrying out the large scale investigation of ozone pollution given the voluminous amounts of air quality data produced from monitoring networks which record hourly, ground level air quality and meteorological parameters. Statistical methods give insight to ozone build up processes and can be used to forecast categorical ozone levels for developing the air pollution control strategies.

1.3 Research Aims and Objectives

In this dissertation, we propose to formulate, implement and apply a data–driven computational framework for air quality monitoring and forecasting with application to ozone formation. The proposed framework integrates, in a unique way, advanced statistical data processing and analysis tools to investigate ozone formation mechanisms and predict the ozone levels in a geographic region. This aim will be achieved through the following objectives:
• Formulation and implementation of a computational framework for data clustering and classification. This involves research on advanced classification and clustering tools to identify and adapt the most relevant to our purpose.

• Application of the data clustering and classification framework to give insight into capturing ozone formation mechanisms. This involves constructing wind classes that explains the regional meteorology. Determination of groups consisting of days with similar ozone triggering mechanisms and develop models that relate the meteorology to ozone in a geographic region.

• Development and implementation of a methodology, based on HMMs that can then be used to label future observations starting from the results of clustering. The development of an ozone forecast model based on HMMs will have the inherent ability to capture the time series nature of a data set.

• Apply the framework to the Houston, Galveston and Beaumont Port Arthur geographical region and the Baton Rouge, Louisiana region. The Houston nonattainment area is part of a large ozone influence region comprising much of the southeastern United States while Baton Rouge is a major industrial, petrochemical, and port center of the Southern United States that has serious ozone problems.

1.4 Thesis Contributions

The key contributions of this thesis are listed below and are dealt with in detail in the chapters to follow:

• A novel unified framework is proposed by blending advanced data-driven techniques widely used in process monitoring to study the tropospheric ozone problem.

  – Wind classes are constructed to explain the regional meteorology. Cluster analysis is performed at the hourly scale to determine the prevailing meteorological regimes. Groups consisting of days with similar ozone triggering mechanisms are determined.
Cluster analysis is performed at a daily scale to reveal the diurnal air flow patterns that determine the dependency of ozone on meteorology. Models which relate the meteorology to ozone in a geographic region are developed.

- Sequence analysis is proposed, in a unique way, along with cluster analysis to relate regional air flow patterns to determine the ozone build up mechanisms.

- Predictive models are developed for forecasting categorical ozone levels. A combinative method that can be applied for multiple observations is one of the contributions of this dissertation.

- HMMs are constructed starting from the results of cluster analysis for predicting the occurrence of ozone exceedance in a region.

- The daily maximum ozone levels in a region are predicted by training the hidden Markov model on historical daily maximum ozone observation values.

1.5 Structure of Dissertation

The thesis is structured about the key contributions listed above and each chapter is provided with sufficient background information and references in order to maintain lucidity and chapter-wise completeness.

Chapter Two presents a brief literature review of the methods used in the field of air quality both for analyzing the physical system and for predicting the air quality in a region.

Chapter Three has the theory of the traditional clustering algorithms. Topics include the hierarchical, non–hierarchical and the aggregated \( k \)-means clustering methods. The agglomerative hierarchical clustering and the aggregated \( k \)-means methods, described in detail, are used in this dissertation.

Chapter Four presents the theory behind the sequence alignment method which is widely used in sociological research. The quantitative sequencing method and the algorithm are described in detail. Application of this method to the field of air quality is one of the contributions of this dissertation.
Chapter Five consists of the theory related to HMMs, a multivariate statistical method that has been applied in a wide range of fields.

Chapter Six contains the first case study that demonstrates the utility of the methods described in chapters 3 and 4. Cluster analysis is performed at two scales to indicate the scenarios resulting in the ozone exceedances and the non-exceedances. The first scale, the hourly scale is used to determine the prevailing meteorological regimes and the second scale, the daily scale is used to reveal the dependency of ozone on meteorology in Houston, Texas. At the daily scale, a new sequence alignment methodology is used to relate the diurnal airflow patterns to ozone concentrations. Mechanisms for the ozone build up are identified using the cluster analysis and sequence analysis methods.

Chapter Seven contains the case study, applying the methods of chapters 3 and 4 to study the ozone patterns for the 2005 summer ozone season in Baton Rouge, Louisiana area. This case study highlights the ability of the methods to analyze large data sets. To examine differences between exceedance days and non-exceedance days cluster analysis is performed on all days to see whether the exceedance days preferentially appear in certain clusters (or end up in clusters of their own). The different scenarios resulting in high ozone levels are isolated; this is achieved by introducing a new sequencing technique for determining the similarity between the days.

Chapter Eight contains the final case study, applying the HMMs to the ozone exceedance and non-exceedance classes from chapter 6. Categorical forecasting method is developed using the computationally inexpensive statistical HMMs. Results from this method compare well with those of the Texas Commission for Environmental Quality (TCEQ). A point value prediction study to predict the daily maximum ozone levels in the Houston, Texas region. The results indicate the capability HMMs as a simpler forecasting tool.

Finally, Chapter Nine provides a brief summary of the work in this dissertation. Recommendations for future work are presented.
Chapter 2

Literature Review

2.1 Models for Air Quality Analysis

The composition of air pollutants in the last 50 years has undergone changes from sulfur dioxide from the use of fossil fuels to nitrogen oxides, photochemical smog and volatile organic compounds from growing traffic to small particles (Fenger, 2009). Air pollution monitoring and modeling contributed to solving the problems with air pollution known for millennia. Since the inception of the CAA in 1971, levels of these pollutants have been regularly measured in the United States. Atmospheric monitoring networks such as The National Atmospheric Deposition Program (NADP), Clean Air Status and Trends Network (CASTNET) are crucial for quantifying and tracking the response to acid deposition from emissions of sulfur dioxide and nitrogen oxides and ozone measurements, respectively.

The advent of computers permitted the use of multivariate statistical data mining techniques to analyze large data sets made available by the monitoring programs. The development of statistical models is an essential task in carrying out the large scale investigation of air pollution given the voluminous amounts of air quality data produced from monitoring networks which record hourly, ground level air quality and meteorological parameters. The statistical models for the air pollutants aim to investigate the pollutant trends, to obtain the forecasts, to parameterize and test the environmental models that predict future climate change and to increase the scientific understanding of the underlying mechanisms (Thompson et al., 2001). The approaches for air pollution studies can be classified into five
categories (Dong et al., 2009): (1) empirical models, (2) fuzzy logic–based systems, (3) deterministic models, (4) data–driven statistical models, and (5) model–driven statistical learning methods.

Empirical models are developed by field experts and are validated using the data sets from a study region. The method performance depends on the variables considered, the underlying assumptions in developing the model and the geographic location of a study region. Thus, the empirical methods are best applicable ‘locally’ or for specific situations. Cox et al., (1995) fit an empirical model to data from the Houston–Galveston area to model the ambient ozone levels. The model accurately predicts ambient ozone level within the center of the region. However, the model performance depends on the upwind or downwind location of the monitoring site indicating that empirical models are accurate for a limited geographic extent. Jimnez et al., (2002) perform the validation of an ozone empirical model that relates the ratio non–methane organic gases/nitrogen oxides (NMOG/NO\textsubscript{x}) with the ozone levels reached. They conclude that the model predicts ozone peaks reasonably well though it tends to slightly overestimate the mean values of ozone.

Fuzzy logic–based systems are rule–based systems that are developed by experts in a field. Fuzzy logic simplifies the process of taking decisions by simulating the way of reasoning of a human expert in environments characterized by uncertainty and imprecision (Astel, 2007). Sen et al., (2009) identify and predict the maximum ozone concentration episodes for the European side of Istanbul city through approximate reasoning with valid If–Then fuzzy rules.

Deterministic models are Eulerian–based regional chemistry transport models that require several inputs including the data on meteorology, topography and emissions inventory. Such models, discussed by Russell and Dennis (2000), play an important role in understanding the evolution of pollutants in the atmosphere and for guiding policy making to manage air quality. Simulation models use the three–dimensional mesoscale output derived from meteorological models such as the Pennsylvania State University/ National Center for Atmospheric Research mesoscale model (PSU/NCAR MM5) to develop air quality simulation models. These three–dimensional air quality models apply a mathematical model to simulate the transport, diffusion reactions and deposition of ozone. Mao et al., (2006) carry out
numerical experiments to determine the sensitivity of the MM5 model to various planetary boundary layer (PBL) schemes and conclude that the four-dimensional data assimilation (FDDA) in MM5 results in the lack of sensitivity of the model to PBL schemes. Ortega et al., (2004) use two examples of dynamic air quality models in the north of Spain. The first model uses three modules for ozone forecasting. The mesoscale model (MASS) which is the first module provides the initial condition to the second module which is a non-local boundary layer model based on the transient turbulence scheme, while the third module is a photochemical box model (OZIPR) in Eulerian and Lagrangian models and receives necessary information from the two previous modules. The second forecast model, called MM5/ UAM–V, is a grid model that predicts the hourly 3–D ozone concentration field. Both methods give good performance only for specific episodes. However, these 3–D models are extremely difficult to build and there is substantial cost in constructing them.

Data–driven methods attempt to extract the causal relationship that exists in a data set that cannot be discovered by rule–based or fuzzy logic–based systems. In recent years it has been shown that the data–driven models (including regression models, classification and Regression Tree analysis (CART), sequence alignment methods, clustering techniques and neural networks) successfully complement the physically–based simulation models and allow for quantification of predictions uncertainty. All analyses of atmospheric pollution rely on regression modeling in some form though variation exists in the precise methodology employed ( Comrie, 1997; Zickus et al., 1998; Smith et al., 2000; Krishan et al., 2004).

Regression models consist of two set of variables–response or predicted variables and explanatory or predictor variables and are performed in two modes: Explanatory model and Predictive mode. In the explanatory mode, the model seeks to determine the coefficients of the models that aid in understanding the relationship of meteorology–pollutants while in the predictive mode, the regression models seek to predict a response variable by using multiple explanatory or predictor variables. Regression models are inherently linear, are easy to implement and are based on the assumption of normally distributed data. Air–quality and meteorological data are generally log–normally distributed or have curvilinear relationships between the variables, however. The curvilinear relationships can be incorporated by using
polynomial terms for the predictor variables while known relationships such as log–normal distributions can be pre–specified by transforming the non–linear data by taking a logarithm to improve the model performance. However, regression models do not provide accurate pollutant analyses and predictions in complex non–linear situations and are hence limited in their applicability.

CART is a binary recursive partitioning method that enables to uncover structures in large data sets. For each split, predictor variables such as temperature, wind speed, wind direction, air quality conditions are examined to split the data set in to two groups such that the dissimilarity between groups based on the value of a predictor variable is maximized. In this way the tree is grown exhaustively until no further splitting is possible. Kuebler et al., (2002) describe the use of CART analysis to determine the meteorological conditions leading to high ozone episodes over the Swiss Plateau.

Sequencing aims at identifying similarity relationships between sequences of quantitative or qualitative variables. The field of sequence analysis includes pattern and motif searching, sequence comparison, and secondary structure prediction (Markel et al., 2003). These methods consist of procedures to compare two or more sequences by looking for a series of characteristics that match up in both sequences in order to assess overall similarity. Sequences that are similar have the same function, be it a regulatory role in the case of similar DNA molecules, or a similar biochemical function and three–dimensional structure in the case of proteins (Mount, 2004). Pakalapati et al., (2009) apply sequence analysis method as a data reduction tool to enable the interpretation of ozone scenarios and the relationship between diurnal wind field patterns and ozone.

Cluster analysis is an unsupervised form of multivariate statistics which indicates recurring patterns among a set of observations. The data belonging to a cluster are relatively similar while the clusters themselves characterized by the cluster mean indicate a distinct pattern in the data set. Thus, cluster analysis can be used to extract the features in the data by labeling the data. These features represent the process states in the historical chemical process data. Cluster analysis has been used for spatial classification of ambient air quality
data in the absence of large data sets needed for more sophisticated modeling approaches (Saksena et al., 2002). Clustering algorithms are successfully applied to identify the relationships between ozone and meteorological variables (Beaver et al., 2006, 2008; Davis et al., 1998). Cluster algorithms suffer from low accuracy, sensitivity to the presence of noise and are usually difficult to implement.

Neural Network (NN) systems represent a methodology that offers a more flexible framework than conventional statistical prediction techniques (Heo et al., 2004). Neural networks attempt to simulate human learning and have been used successfully to model complex relationships between variables in an imprecise and non-linear data set (Yi and Prybutok, 1996; Spellman, 1999; Gardner and Dorling, 2000). Hadjiiski et al., (2000) apply neural networks to model and predict ambient $O_3$ concentrations based on a limited number of measured hydrocarbon species, $NO_x$ compounds, temperature, and radiant energy. Their dynamic models adequately predict the $O_3$ concentration 3-hr into the future with a root-mean squared error (RMSE) of 8.21 ppb. The estimation of a neural network model is automated unlike the regression methods which must be re-estimated with every new data set. These models, however, are black-box models that result in equations which do not have satisfactory explanatory capability.

State Space Models and Bayesian networks are the two categories of model-driven methods. These models rely on the assumption of availability of the mathematical model as well as the training data. Model-driven methods are superior to data-driven methods as the exact mathematical model can be established and can be adjusted to improve the accuracy of a model. In a state space model, the observed data are assumed to result from sums of series each with a specified time series structure such as the sum of trend, a seasonal effect, and error (Shumway et al., 2000). These methods find applications in a variety of fields such as economics (Harvey and Pierse, 1984), medicine (Jones, 1984) and tropospheric ozone prediction (Vassiliadis et al., 1998).

Bayesian networks are methods that integrate human expertise with statistical analysis, accept real-time information that they use with the stored knowledge in order to formulate
diagnostic or predictive conclusions. Bayesian networks are used to develop knowledge–based applications in domains that are characterized by inherent uncertainty (Savell et al., 2004). HMMs fall under a subclass of Bayesian networks known as dynamic Bayesian networks which are Bayesian networks for modeling time series data (Ghahramani, 2001). HMMs have been used in the areas such as speech recognition (Rabiner, 1989), process monitoring (Wong et al., 1998), bioinformatics, and genomics (Schliep et al., 2003; Won et al., 2004), precipitation forecast (Bellone et al., 2000), process monitoring (Wong et al., 1998) and particulate matter forecasting (Dong et al., 2009). Gordillo and Conde (2007) applied HMMs to classify spam mails by detecting the words used by spammers. In spite of their high computational costs, these models have many successful applications in practice due to their rich mathematical structure and ease of model interpretation as opposed to the black box data–driven models. In air quality modeling, Dong et al., (2009) develop hidden semi Markov models to predict the particulate matter concentrations (PM2.5) at O’Hare airport in Chicago. Their results indicate that the HMMs provide accurate 1–day ahead predictions of the high PM2.5 concentrations.

2.2 Models and Methods for Ozone Analysis

Numerous photochemical models exist to study the problem of tropospheric ozone. The non–linear photochemical reactions that lead to ozone formation are first modeled by Integrated Empirical Models (Johnson, 1984). Currently, Air Quality models (AQM) that integrate the photochemical models with 3–D meteorological and transport models are used to model the ozone formation. These models are used typically to determine the sensitivity of ozone levels to reductions in the precursors $NO_x$ and VOCs (Dennis et al., 1999; Blanchard et al., 2001). These models though powerful for understanding the ozone formation mechanisms, however, require large amounts of time and resources to develop and can be used to characterize only limited amounts of ozone build up scenarios. Statistical methods, though less powerful than the traditional AQMs and the deterministic models used for air quality analyses, can be implemented on real data for large observation periods.

The statistical models available for ozone studies can be broadly classified as regression–based modeling, extreme–value approaches and space–time models (Thompson et al., 2001).
The regression models that are useful for modeling average behavior, however, are not a good fit for applications that assess effects due to pollutant extremes. Models built on extreme value theory are useful for modeling threshold exceedances. The third category of ozone models—the spatio–temporal models can be used to investigate ozone trends as well as to analyze the ozone extreme value behavior. However, space–time models suffer from increased complexity of modeling and data collection. The choice of models depends on the purpose of analysis and no model is most appropriate for all purposes.

Many methods are available for the analysis and prediction of ozone concentrations. Researchers and forecasters use several methods by balancing the strengths of a particular method with the limitations of another for analysis and prediction of ozone mixing ratios. One of the models described above is used in implementing these methods. This section presents the common methods for ozone forecasting from the available several methods with a brief discussion on their strengths and limitations.

Persistence is a forecasting method based on the assumption that a positive statistical association exists between today’s ozone concentrations and their past or the future values (Wilks, 1995) and the weather patterns. Ozone concentrations in a region depend on the synoptic scale climatology. Synoptic scale patterns have similar characteristics for several days. Thus, ozone concentrations also exhibit similar characteristics for several days and this validates the persistence method assumption. This method is the simplest of all the existing methods and can be used as a starting point for ozone forecasting. Persistence is accurate only if the weather conditions remain similar for several days and this method fails in forecasting the first and the last exceedance days.

Climatology is the study of average and extreme weather conditions in a region (EPA, 1999). This method uses climate tables which are summaries of the occurrence of extreme events in a region for many years to determine the accuracy of a forecast. Thus, this method though simple to develop is not a stand-alone forecasting method but is used to develop bounds on ozone forecasts produced by other methods.

In the Criteria method, threshold criteria of many independent variables such as meteorology or other air quality variables are used to determine the ozone concentrations. Ozone
concentrations are usually associated with high temperatures or specific diurnal wind field patterns. Thus, the occurrence of high temperatures or specific wind speeds is taken as an indicator of high ozone levels in this method. This method can guide in forecasting categorical exceedance or a non–exceedance but cannot determine the exact point–value ozone concentrations in a region. This method is easy to develop and operate but relies on the subjective judgment for choosing the determining variables that affect the ozone values. This method is a simpler version of CART. In CART, which is an objective prediction method, a software selects the variable that describes the largest variance in ozone concentrations for the first data split and the procedure of splitting continues until data in each group is uniform.

Phenomenological or intuition forecasting is a method that involves the analysis and processing of air quality and meteorological information to formulate the ozone prediction. This method requires an accurate understanding of the mechanisms that drive ozone formation along with vast amounts of data and human expertise to develop a consensus between conflicting results from different methods. Since this method involves the judgment of a forecaster, it might involve bias and thus this method should always be complemented by an objective prediction tool.

2.3 Ozone Analysis and Prediction Studies

Many studies are carried out to identify recurring meteorological regimes in a geographical region that aid in the interpretation of the physical mechanisms of ozone formation. These regimes are not known in advance and unsupervised statistical methods can be used to delineate the regimes in a data set. The knowledge of frequency of occurrence of a meteorological regime in a region can guide the selection of representative episodes for AQMs development (EPA, 1996). Rohli et al., (2004) use principal component analysis – a dimensionality reducing method that falls under regression, to establish the relationship between tropospheric synoptic circulation wind patterns and ozone mixing ratios in Louisiana. Their results indicate that synoptic influences play a major role in the formation of ozone in Louisiana. Meteorological studies consider the wind field due to two factors. First, the affect of the wind field on source–receptor relationships determine the air quality in a region. Second, the
wind field determines the dilution rate of the pollutants or the concentration levels of the pollutants in a region. Weber and Kaufmann (1995), Darby (2005), Beaver and Palazoglu (2006) and Pakalapati et al., (2009) perform cluster analysis on hourly wind observations from a network of surface monitors and relate meteorological regimes to ozone levels categorically.

The statistical models developed for ozone prediction can be used for both the point value prediction and for forecasts based on classification. These are computationally less expensive in comparison to the deterministic models and can be done in a short time too. Schlink et al., (2006) assessed 15 point value statistical techniques for operating pollution forecasting and concluded that neural networks and Generalized Additive Models (GAMs) best handle the non-linear associations between atmospheric variables. Prybutok et al., (2000) compared neural networks with the two conventional statistical methods, regression and Box–Jenkins Auto Regressive Integrated Moving Average (ARIMA) for prediction of Houston’s daily maximum ozone concentrations and their results indicate that neural network model is superior to the regression and Box–Jenkins ARIMA models. However, neural networks suffer from convergence to a local optimum. Davis and Speckman (1999) developed a model based on the GAMs approach for predicting the maximum and the 8–hr average ozone in Houston. Though this method appears to do better than the linear, non-linear regression and the neural network models, it has not attained the desired predictive capabilities. This is a testimony to the complex problem of estimating ozone accurately on a continuous scale. Studies on forecast based on classification as an alternative to point value prediction have also been carried out. Ryan (1995) used CART and demonstrated its skill at distinguishing strong and weak ozone cases for the Baltimore area. Ghiaus (2005) reported a 98% probability of detection among a set of 7 ozone classes in a sea breeze regime using linear fuzzy discriminant analysis.

Literature review in the above two sections suggests that cluster analysis (discussed in chapter 3), sequence analysis (discussed in chapter 4) and HMMs (discussed in chapter 5) though employed in a wide range of studies have an innate disadvantage that prevents their wide spread usage. In this dissertation, we propose to develop a framework that combines
the above methods in a unique way. We use categorical pattern recognition methods to identify the recurring meteorological states that lead to high ozone in a region. We considered cluster analysis and sequence analysis as tools for investigating the relationship between wind field patterns and ozone for large data sets. In this dissertation, clustering wind observations is explored and the wind field pattern clusters thus obtained are related to the daily maximum ozone composition levels using sequence alignment methods. We used the agglomerative technique proposed by Beaver and Palazoglu (2006) in this study. Cluster analysis is performed at two scales: one at the hourly scale to determine the prevailing meteorological regimes and the other at a daily scale to reveal diurnal patterns that determine the dependency of ozone on meteorology.

Cluster and sequence analyses, though computationally effective, are good to determine the average spectral patterns and they do not perform well for extreme value analyses. Thus, we propose to develop an ozone forecast model based on model-driven HMMs that have the inherent ability to model extreme values by capturing the time series nature of the data set. Betro et al., (2008) use HMMs to predict the occurrence of rainfall events in Central Sardinia, Italy. We use a similar approach to identify the occurrence of extreme ozone levels based on point-value daily maximum ozone predictions in Houston, Texas region.

2.4 Conclusions

This chapter provides a brief history of the air pollution in the last 50 years along with an overview of the available models for air quality analysis. A summary of methods available in particular for ozone analysis studies and predictions are provided. Finally, a brief literature review was performed summarizing the usage of models described to study the ozone problem in the southeastern United States. The final section of this chapter presents an overview of the methods employed in this dissertation and the rationale for their usage in developing the models for studying the ozone problem is discussed.
Chapter 3

Clustering Algorithms

3.1 Introduction

Any monitoring technique requires some kind of model describing the current operation of the process or system under consideration. In a data-driven approach, such models are obtained by training the samples of historical data that represent the states in the process. Thus, labeling of historical data is critical to the development of models that aid the process monitoring techniques.

Learning from data sets comprises two categories: supervised learning and unsupervised learning. In supervised learning the variables of the study are split into two groups: explanatory variables and one (or more) dependent variables (Wilhelm, 2004). The goal of supervised learning is to build a model of the distribution of class labels in terms of predictor features (Kotsiantis, 2007). The resulting classifier is then used to assign class labels to testing in instances where the values of the predictor features are known but the value of the class label is unknown.

In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables. Unsupervised learning does not require any a priori knowledge about the data that is going to be investigated (Lendl et al., 2005). Unsupervised learning attempts at deriving the hidden structure from the raw data by modeling the probability density of the input data (Wang, 2001).

The development of unsupervised statistical models is an essential task in carrying out the large scale investigation of ozone pollution given the voluminous amounts of air quality
data produced from monitoring networks which record hourly, ground level air quality and meteorological parameters. Statistical methods give insight to ozone build up processes and can be used to forecast categorical ozone levels for developing the air pollution control strategies.

In this chapter, the unsupervised clustering approach for data classification and typical clustering algorithms are introduced. In section 3.2 a background on clustering and clustering approaches along with a short description of the measures of similarities and data handling measures are provided. Sections 3.3 through 3.5 provide the hierarchical, non–hierarchical and hierarchical ensemble of $k$–means aggregation clustering methods. Section 3.6 provides a brief theory of the time series clustering algorithms and finally, in section 3.7 a summary about the methods introduced in this chapter is given.

### 3.2 Background

Cluster analysis (Everitt et al., 2001) is a class of statistical methods that seeks to partition a set of $N$ observations into distinct groups. Each observation corresponds to a particular sampling interval (distinct period in time) for which corresponding measurements are available on the same set of $p$ variables. The objective of clustering this data is to seek natural groupings so that the properties of these clusters can be determined. Five basic types of clustering methods have been identified: i) hierarchical techniques, ii) optimization techniques, iii) density–seeking techniques, iv) clumping techniques, v) others. The choice between them depends primarily on the aim, type of data and the availability of the method.

Hierarchical clustering and optimization methods are the most popular. In the optimization techniques, clusters are formed by the optimization of a clustering criterion. The clusters are mutually exclusive, thus forming a partition of the set of entities. Optimization techniques usually require large amounts of computer time and consequently cannot be recommended for use with very large data sets. In hierarchical clustering, classes themselves are classified into groups, the process being repeated at different levels to form a dendrogram (formally to be introduced in section 3.3.2). Hierarchical techniques are most suitable and do have the considerable advantage over optimization techniques in requiring far less computer time, consequently they may be used with larger data sets and is the one used in our work.
These techniques may be subdivided further into *agglomerative* methods which proceed by a series of successive fusion into groups, and *divisive* methods which partition the set of entities into finer partitions.

### 3.2.1 Data Handling

As with most multivariate statistics, proper scaling of the data is critical to successful and meaningful modeling. Because traditional clustering algorithms rely on distance calculations, which are sensitive to the scaling of the variables, each of the $p$ variables should be normalized to an equal level of variability so they are weighed equally in any distance calculations. Such scaling can be achieved by auto scaling the data i.e. subtracting the mean and then dividing by the standard deviation for each variable. The statistical calculations cannot handle any missing measurements. Missing data are prevalent in real world data sets due to factors such as failed or missing sensors. Any missing observation in the data set must be imputed or the entity represented by that measurement must be removed from the analysis altogether.

As the imputation of data is specific to the nature of the data in question, any imputation of data that might be performed will be described in the case studies.

### 3.2.2 The Measures of Distance and Similarity

Similarity, dissimilarity and distance are three closely related terms used to discuss the comparability of objects. Distance is the most precisely defined measure and is governed by three conditions:

1. $D_{ij} \geq 0; D_{ij} = 0$ if $x_i = x_j$
2. $D_{ij} = D_{ji}$
3. $D_{ia} + D_{ja} \geq D_{ij}$

where, $D_{ij}$ denotes the distance between two objects $i$ and $j$. A similarity coefficient measures the relationship between two individuals given the values of a set of $p$ variables common to both. In general, similarity coefficients take values in the range 0 to 1. There exist several measures of similarity and distance. Euclidean, Mahalanobis, Correlation are few of the distance measures for measuring the distance between entities. The distance
method to be employed depends on the goal of the analysis and the observations data at hand. The element $D_{ij}^{\text{Euclid}}$ is the Euclidean distance between observations $i$ and $j$.

$$D_{ij}^{\text{Euclid}} = \sqrt{\sum_{k=1}^{p} (X_{ih} - X_{jh})^2}. \quad (3.1)$$

Euclidean distance used on raw data is badly affected by changing the scale of a variable. Hence, variables are frequently standardized before employing Euclidean distance by dividing the raw data by the standard deviation of $p^{\text{th}}$ variable. The Euclidean distance calculated from the standardized variables preserves relative distances. The Mahalanobis distance is similar to the Euclidean distance but each term in the summation is multiplied by different weights. Variables with large correlations to other variables are assigned small weights while the variables capturing independent features are assigned larger weights. This reduces the effect of similar variables from biasing the cluster solution.

$$D_{ij}^{\text{Mahab}} = (X_i - X_j)'S^{-1}(X_i - X_j) \quad (3.2)$$

Here, $S$ is the pooled within groups variance-covariance matrix and $X_i$ and $X_j$ are the vectors of observations $i$ and $j$. The correlation metric or the Pearson correlation distance is one minus the sample correlation between two points or vectors of data. The element $D_{ij}^{\text{Corr}}$ is the correlation distance between two observations $i$ and $j$ and is defined as

$$D_{ij}^{\text{Corr}} = 1 - \frac{1}{p - 1} \sum_{h=1}^{p} \frac{(X_{ih} - \bar{X}_i)}{S_{X_i}} \quad (3.3)$$

Here, $\bar{X}_i$ and $S_{X_i}$ are the observation vector $i$ mean, and standard deviation respectively. When the correlation between variables on which the entities are measured is zero, Mahalanobis measure is similar to the Euclidean measure. Different measures of similarity result in different results and the choice of correct measure to use depends on the structure of the data. In this research, the Euclidean distance and Correlation metric were explored as the dissimilarity measures for hierarchical clustering.
The measure of dissimilarity between two groups is referred to as the *linkage* method: the nearest neighbor or single linkage measures the distance between two groups as the distance between their closest members, one in each group; the furthest neighbor or complete linkage method gives the distance between two groups as the distance between their most remote pair of individuals and is exactly the opposite of the single linkage methods; A linkage method such as average linkage quantifies the decision making process to compare the clusters in each stage. The distance \( Z(r, s) \) between any two clusters \( r \) and \( s \) is the average of all distances for \( i \subset r \) and \( j \subset s \).

\[
Z(r, s) = \frac{1}{N_r N_s} \sum_{i \subset r} \sum_{j \subset s} D_{ij}
\]  

(3.4)

Here, \( N_r \) and \( N_s \) denote the respective cluster sizes (i.e. the number of observations currently assigned to each cluster). The *Ward linkage* (Ward, 1963) method considers every possible pair of clusters at each step in the analysis and the two clusters that result in minimum increase in the sum of error squares are combined. Ward’s method, however, assumes that all clusters are realized with equal frequency and is inappropriate for most systems. The choice of a linkage method depends on the nature of the data sets being studied and no single method is better than the other. Average linkage was explored as the dissimilarity measure between groups for hierarchical clustering.

### 3.3 Hierarchical Clustering

#### 3.3.1 Hierarchical Clustering Background

In a hierarchical classification, the data are not partitioned into classes in one step. Rather they are first separated into a few broad classes, each of which is further divided into smaller classes, and each of these further partitioned, and so on until terminal classes are generated which are not further subdivided. Essentially, hierarchical techniques are subdivided in to agglomerative methods which proceed by a series of successive fusions of the \( N \) entities into groups and divisive methods which partition the set of \( N \) entities successfully in to finer partitions. Both types of hierarchical technique are attempts to find the most efficient step at each stage in the progressive subdivision or synthesis of the data. Since all agglomerative
methods ultimately reduce the data into a single cluster containing all the entities and the divisive techniques split the data into \( N \) groups each containing a single entity, the stage at which the procedure has to be stopped thus determining the number of clusters is a subjective part of the methodology. Hierarchical methods do not allow for reallocation of entities. Hence if a mistake is made initially, it is carried through and no chance for rectification is allowed at a later stage.

Cluster analysis begins not with the data matrix \( \mathbf{X} \) but with an \( N \times N \) symmetric distance matrix \( \mathbf{D} \) that indicates the degree of dissimilarity between two individual observations. To demonstrate clustering and the resolution of the cluster solution, the scatter plot for an example data set as shown in Figure 3.1 is considered. The scatter plot consists of 11 objects (labeled A to K) for which 2 variables \((x_1\) and \(x_2)\) are measured. Using the Euclidean distance of Equation 3.1, the dissimilarities between the various pairs of objects are calculated. For example, objects G and F are separated by a very small distance and are concluded as relatively similar, while objects G and E have a large distance between them and are concluded relatively dissimilar. The distance between objects I and G is somewhere in between these previous distances, indicating an intermediate degree of association.

---

**Figure 3.1:** Scatter plot of 11 objects represented by 2 variables.
3.3.2 Agglomerative Hierarchical Clustering

The agglomerative hierarchical classification consists of a series of partitions that run from \( N \) clusters each containing a single observation to a single cluster containing all the observations. Differences between methods arise because of the different ways of defining distance between an entity and a group containing several entities or between two groups of entities (linkage method). Hierarchic classifications are represented graphically by a diagram known as a dendrogram. The dendrogram illustrates hierarchical nesting of the clusters and their distances for merging made at each successive stage of the analysis. Figure 3.2 shows a dendrogram for the clustering of the data of Figure 3.1 using an agglomerative hierarchical algorithm with Euclidean distance and average linkage. The \( x \)-axis of the dendrogram contains \( N \) sorted labels corresponding to the \( N \) singleton clusters formed on initialization while the \( y \)-axis indicates the average distance at which clusters are merged.

Clusters are selected from the dendrogram using a “top-down” approach. Across the bottom, the \( N \) clustered sampling intervals appear in a non–unique order that allows the dendrogram to be drawn without any overlapping branches. Upside–down, U–shaped “branches” contain all sampling intervals appearing directly underneath them; member sampling intervals are indicated by tracing connecting dendrogram lines downward (and never upward) in the hierarchy. Clusters are nested in the hierarchy such that multiple, smaller, more highly specific clusters are completely contained within a larger, more general cluster. The difference in
the $y$–axis values between a pair of nested branches indicates the gain of specificity (loss of
generality) upon splitting the sub cluster from its parent. A relatively large $y$–axis distance
indicates the presence of a statistically distinct cluster. A set of $k$ clusters are selected from
the dendrogram by considering successive sets of increasingly smaller candidate clusters,
starting from the top of the hierarchy and moving downward.

For example, in Figure 3.2 the 11 objects can be split into 2 groups represented by the
2 main branches of the dendrogram: objects C, E, D, A, and B form one group while F, G,
H, I, J, and K for a second cluster. Clearly these are the 2 largest clusters of data in the
scatter plot of Figure 3.1. Near the bottom of the dendrogram exist a large number of small
clusters that are finely resolved and only represent a small portion of the original data set
at a high level of detail. In the example, the dendrogram of Figure 3.2 indicates the pair of
objects A and B as one small cluster of closely spaced points in the scatter plot of Figure
3.1. Thus, hierarchical clustering is merely a tool that allows the user to quickly generate a
dendrogram describing clusters at various levels of resolution.

3.3.3 Cluster Validation Measures

Once the clusters are obtained, validity of the clusters is determined in one of the many
possible ways: several clustering techniques based on different assumptions could be used on
the same data set and only clusters produced by all or majority of methods accepted; the
data could be randomly divided into two and each half clustered independently. Membership
assignment in the partitioned samples should be similar to that of the entire sample, if the
clusters are stable; Deletion of a small number of variables from the analysis in most analysis
should not greatly alter the clusters found and this could be explored as one of the validation
methods.

The Cophenetic correlation coefficient $C$ assays the degree of agreement between the	abular elements of $D$ and their graphical representation as a dendrogram. It is a measure
of how well a dendrogram represents the multidimensional relationships within the input
data. The Cophenetic correlation coefficient is given by

\[ C = \frac{\sum_{i<j}(D_{ij} - \mu_D)(Z_{ij} - \mu_Z)}{\sqrt{\sum_{i<j}(D_{ij} - \mu_D)^2 \sum_{i<j}(Z_{ij} - \mu_Z)^2}}. \tag{3.5} \]

Here, \( Z_{ij} \) is the average distance at which objects \( i \) and \( j \) are first placed in to the same cluster and \( D_{ij} \) is the y–axis value of the dendrogram. \( \mu_D \) and \( \mu_Z \) are the mean values of \( D_{ij} \) and \( Z_{ij} \) respectively. \( C \) measures the correlation between pairwise distances \( D_{ij} \) and pairwise cluster merging distances \( Z_{ij} \). The value of \( C \) ranges from zero to unity where the values approaching unity indicate that the dendrogram does not greatly distort the original structure in the input data. Prior to selecting clusters from the dendrogram, Cophenetic correlation should be used to check the validity of the dendrogram. The above methods allow for validating a cluster solution. The first method is generic while the second method is particular to the hierarchical solution. These measures are sufficient but the goodness of a cluster solution can be determined only by interpreting the results to check for physically consistent and meaningful solutions.

3.4 Non–Hierarchical Clustering

3.4.1 Non–Hierarchical Clustering Background

These techniques differ from hierarchical methods in that they admit relocation of the entities, thus allowing for the possibility that a poor initial partition might be corrected at a later stage. They also differ from the above methods in that their solutions do not necessarily portray hierarchical relationship among the entities. Contrary to the hierarchical procedures, to perform the non–hierarchical clustering the desired number of clusters has to be pre–defined.

These methods seek a partition of the data which optimizes some predefined numerical measure. They begin by finding \( k \) points in the \( p \) dimensional space which act as initial estimates of the cluster center. These \( k \) points could be the initial \( k \) cluster mean vectors or a trial value of \( k \) larger than thought is necessary and sets up cluster centers regularly spaced at intervals of one standard deviation on each variable. The parameter \( k \) controls the
resolution of the solution, as increasing $k$ generates solutions with smaller and more compact clusters. The number of groups is then reduced until a criterion based on the residual sum of squares is satisfied. Entities are allocated to the cluster to whose center they are nearest to and the estimate of the center is updated after the addition of each entity to the cluster or only after all the entities have been allocated. Once an initial classification has been found by one of the methods mentioned above, a search is made for entities whose reallocation to some other group will cause an improvement in the particular clustering criterion being considered. The procedure is continued until no further move of a single entity produces a better criterion value. A local optimum of the criterion is thus reached and the solution may be accepted or an attempt may be made to improve it by repeating the procedure using a different starting configuration.

Non–hierarchical algorithms give different solution for each random initialization. Thus, for a fixed $k$, the algorithm is capable of yielding different solutions as the objective function reaches a local optimal value of the available solution space of the cluster analysis. Non–hierarchical methods also suffer from the inability to determine the number of clusters $k$ directly. A method to determine the value of $k$ is given in section 3.5. Simulated annealing (Kirkpatrick et al., 1983), the $k$–means algorithm of MacQueen, (1967) are two of the many available advanced non–hierarchical algorithms. In this research, $k$–means algorithm was explored.

3.4.2 Non–hierarchical $k$–means Clustering

The $k$–means algorithm starts with $k$ initial means for each cluster. All of the $N$ objects are then compared with the prototype mean by means of the Euclidean distance and are assigned to the closest cluster. After all the days are assigned, the mean is recalculated for each cluster. For a fixed $k$, a distinct solution is produced for each random initialization of the algorithm. To determine the accurate solution that gives the most optimal partitioning between observations out of these distinct solutions, this algorithm aims at minimizing the objective function – mean squared errors between each observation and the cluster mean for each solution. Thus, the algorithm seeks to minimize the objective function $J$: the sum of squared Euclidean distances between the scaled $\tilde{u}_{tp}$ and their cluster mean $\tilde{u}_p$. 

29
\[ J = \sum_{c=1}^{k} \sum_{t \in c} (\tilde{u}_{tp} - \bar{u}_{cp})^2 \]  \hspace{1cm} (3.6)

with

\[ \bar{u}_{cp} = \frac{1}{n_c} \sum_{t \in c} \tilde{u}_{tp} \]

and scalings

\[ \tilde{u}_{tp} = \frac{u_{tp}}{p_p}; p_p = \frac{1}{Nh} \sum_{t=1}^{N} u_{tp} \]

Subscript \( c \) is an index over the \( k \) total clusters, \( t \in c \) indicates the set of observations \( t \) assigned to cluster \( c \), and \( n_c \) is the number of observations assigned to cluster \( c \). The observations for each variable \( p \) are scaled using the mean value \( p_p \) for that same variable; this ensures each variable is weighted equally in the analysis.

### 3.5 Hierarchical Ensemble of \( k \)–means Clustering

Hierarchical and optimization methods (non–hierarchical) are the two broad classes of clustering. Both have innate disadvantages that preclude their widespread use. Hierarchical methods use heuristic set of irreversible decisions to pair existing clusters and this simplicity aspect of the algorithm results in its poor performance. They are, however, easy to implement and provide the solution in a single diagram known as dendrogram. The dendrogram shows the hierarchical merging of clusters. Non–hierarchical solutions attempt to optimize an objective function representing the goodness of fit of the observations. These methods, however, require that the number of clusters be specified in advance and the optimization routine mostly converges to a local optimum, thus resulting in a different solution for each random initialization.

Hybrid hierarchical–\( k \)–means algorithm, proposed by Beaver et al., 2006, is a method that hybridizes the traditional clustering procedures. This method aggregates an ensemble of non–hierarchical solutions into a single final hierarchical solution. This method gives the solution as a dendrogram and the solution is invariant to the random initializations of the individual non–hierarchical runs. This algorithm aggregates ensemble of \( k \)–means solutions into a single, final partitioning. The solution from the single run of the \( k \)–means algorithm
is stored in the $N \times k$ binary matrix $B^{(m)}$, where element $B^{(m)}_{ir}$ is 1 if object $i$ is assigned to cluster $r$ in run $m$ and 0 otherwise. An ensemble of $n$ runs is concatenated into a single matrix $B$ where the columns are the matrices $B^{(m)}$ for the $n$ individual solutions.

$$B = [B^{(1)}, B^{(2)}, \ldots, B^{(n)}]$$ (3.7)

The aggregated distance matrix $D^{Avg}$ is then calculated as

$$D^{Avg} = 1 - \frac{1}{n}BB^T$$ (3.8)

where, $1$ is a matrix of ones. The second term gives the fraction of $n$ runs for which the episode days are placed in the same cluster. Thus, $D^{Avg}$ is a measure of dissimilarity or is a measure of distance and hence can be input into the traditional hierarchical clustering. The dendrogram produced contains branches and it remains to the discretion of the user to select clusters ranging from 1 to $N$ based on the desired resolution. This method can be repeated for a range of $k$ values and by concatenating the obtained solutions with each $k$. However, the solutions with larger $k$ have little correspondence and hence $k_{max}$—the largest $k$ with meaningful $k$–means solutions has to be determined.

To determine $k_{max}$, non–hierarchical algorithm is performed with incrementally larger trial values of $k_{max}$ starting from 2 and an aggregated distance matrix $D^{Avg}(k'_{max})$ is calculated using Equation (3.7) and Equation (3.8) for all the $n$ runs of $k$ from 2 to $k'_{max}$. A sum of squared measure denoted $\triangle SSE(k'_{max})$ is used to assay the discrepancy between distance matrices generated by the values.

$$\triangle SSE(k'_{max}) = \sum_{i<j} \{D^{Avg}_{ij}(k'_{max} + 1) - D^{Avg}_{ij}(k'_{max})\}^2.$$ (3.9)

The parameter $k_{max}$ is the smallest trial $k'_{max}$ value to produce a converged distance matrix. Thus, the aggregation algorithm is capable of determining the proper range of the number of clusters solving the typical problem of non–hierarchical methods involving the determination of $k$. This measure of distance with the optimal $k_{max}$ value is input into the
traditional hierarchical clustering to produce an intuitive graphical result from which an appropriate partitioning can be selected at a desired level of resolution.

The hybrid hierarchical–$k$–means clustering method uses the aggregation process to calculate a dissimilarity matrix which is input into a traditional hierarchical clustering algorithm. The hierarchical algorithms produce a dendrogram describing solutions with a ranging number of clusters. Hence, the aggregation process is performed for a range of $k$ – number of clusters. The aggregate distance matrix is the average result of an optimization algorithm and is not a simple heuristic like the Euclidean dissimilarity measure. Hence, the aggregated measure better compares the clustered observations than the traditional hierarchical dissimilarity measures. Thus, the hybrid hierarchical–$k$–means clustering method retains the multi–resolution and graphical aspects of the hierarchical clustering while benefiting from the optimization performed by the non–hierarchical algorithm.

3.6 Dynamic Principal Component Analysis for Time Series Clustering

Cluster analysis algorithms are applicable for only independent observations. Many real world processes, however, are dynamic and result in observations that are correlated in time. To apply cluster analysis to a time series of measurements, a time series clustering algorithm based on dynamic Principal Component Analysis (DPCA) and moving window approach is proposed by Beaver et al., (2006). Empirical Orthogonal Function (EOF), more commonly known as Principal Component Analysis (PCA) is a multivariate statistical method that aids in data reduction by modeling correlation structure in a data set. A moving window defined by its length $L$ and moving rate of $R$ is used to generate $N$ windows of time series data $X_i$ ($L \times p$) from the original time series $X$. This method is depicted in Figure 3.3.

*Autoscaling* is a standard operation used on data to center and normalize the data prior to applying the PCA. A PCA is the rotation of the coordinate system for the original observations; it must be centered at the origin. If the data does not have zero mean, the first PC that points from the origin to the dominant direction of variability of data will point to the data and cannot capture any meaningful relations between variables. Data
are centered by removing the mean from each variable. Centering of the data matrix $X$ is ensured by centering the $N X_i$ matrices. DPCA, like any time series model requires that the time series observations be detrended. This is achieved by subtracting the mean or any trend from the observations. The purpose of detrending is to reduce the correlations between the observations and not to remove the trend effects. A description of PCA and time series clustering is given below.

### 3.6.1 Dynamic Principal Component Analysis

A PCA models the correlation structure between a set of $p$ correlated observations and transforms them into a new set of $q$ orthogonal components called the Principal Components (PCs). A PC is a linear combination of the original variables and usually a small number of PCs adequately capture the variations in a data set. PCA is an eigenvalue decomposition a matrix ($X_i$) into a complementary approximation ($\hat{X}_i$) and an error matrix ($E$) using a loading matrix $P_r$.

$$X_i = \hat{X}_i + E = X_i P_r P_r^T + E$$ (3.10)

The PCA loading matrix $P_r$ ($p \times q$) is obtained by retaining $q$ eigenvectors corresponding to the first $q$ eigenvalues of covariance matrix $X_i^T X_i$ sorted in a decreasing order:

$$P_r = [P_1, P_2, \ldots, P_q]$$
The number of PCs, \( q \), is chosen to exceed a threshold limit of variability (e.g. 90%). The eigenvalues can be used to calculate the amount of variability in the original data set retained by the \( q \) PCs by using Equation 3.11.

\[
\text{Captured \% variance} = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \times 100 \quad (3.11)
\]

Once a \( P_r \) is estimated, a matrix \( X_i \) can be projected into this model to determine the goodness of fit between the PCA model and the observations. This is determined by calculating the residual error matrix \( E(L \times p) \).

\[
E = X_i(I - P_r P_r^T) \quad (3.12)
\]

The loss of information \( Q(i, r) \) due to the projection of \( X_i \) on to \( P_r \) is given as

\[
Q(i, r) = \sum_{j=1}^{L} \sum_{k=1}^{p} E_{jk}^2 \quad (3.13)
\]

Traditional PCA as described above accounts for multivariate relationships but is not intended for time series or sequentially correlated observations. Dynamic PCA (DPCA) was proposed by Ku et al., (1995) as an extension to the traditional PCA for modeling correlated observations. A DPCA augments the data matrix by including the lagged variables. Each \( X_i \), after centering, is augmented by concatenating \( M \) lagged copies of the variables to form a matrix \( \tilde{X}_i \) with \( \tilde{p} = p(M+1) \) columns.

\[
\tilde{X}_i = [X_i, CX_i, C^2X_i, ..., C^M X_i] \quad (3.14)
\]

where, \( C \) is the back shift operator that shifts each observation in the data matrix \( X_i \) backwards in time by \( k \) samples.

The traditional PCA can then be performed on this matrix \( \tilde{X}_i \) to determine the DPCA loading matrix \( P_r \). To determine \( M \), partial autocorrelation function (PACF) described in Shumway and Stoffer (2000) is used. The PACF, with a range of [-1, 1], at a lag \( \tau \) indicates
the correlation of a variable at time \( t \) with itself at time \((t-\tau)\) by circumventing the effects of intermediate lags from \((t-1)\) to \((t-\tau+1)\). The number of lags \( M \) is best determined based on the statistical estimate of the time scale for the observations being modeled since large \( M \) results in computational burden.

### 3.6.2 Time Series Clustering

The first step in this method is to *window* the detrended time series observations into small subsets of continuous time series. This results in two dimensional matrices of time series data, \( X_i(L \times p) \), as opposed to the one dimensional feature vectors used in the traditional clustering methods described above. To resolve this inconsistency, each window of time series data is unfolded to form a feature vector acceptable for the clustering algorithm. For example, each \( L \times p \) is unfolded to form a one-dimensional feature vector \( 1 \times L \times p \). This unfolded time series data can then be input into the clustering algorithms to determine the patterns based on the standard distance based dissimilarity concepts. It is to be noted that cluster analysis of unfolded data is not applicable for time series data in general. It is applicable only for data sets such as hourly wind observations where windows of time series data can be compared logically (Cape et al., 2000).

The clustering algorithm of time series observations termed \( k \)–PCA is analogous to the \( k \)–means algorithm. The prototypes for clusters in the case of \( k \)–PCA method are PCA models \( P_r \) while in the \( k \)–means algorithm the prototypes are the mean vectors. A brief description of methods to obtain \( P_r \) is given in section 3.6.1. The \( k \)–PCA algorithm is initialized by choosing \( k \) and \( k \) non–overlapping windows \( X_i \) to determine the cluster prototypes \( P_r \). The algorithm then partitions \( N \) windows of time series into \( k \) clusters by optimizing an objective function (Equation 3.15) describing the goodness of partitioning. The algorithm considers each window of observations as a single entity. Thus, a window of data \( X_i \) assigned to a cluster \( r \) has all the observations in the window belonging to the cluster \( r \). At any iteration of the algorithm, all the \( N_r \) windows of time series assigned to a cluster \( r \) are stacked to form a single supra matrix \( \tilde{X}_r(LN_r \times p) \). This supra matrix is used to update the prototype PCA \( P_r \) at any iteration. Each of the \( X_i \) windows of observations belonging to a cluster \( r \) are then projected onto each of the \( k \) prototypes \( P_r \). The loss of information by projecting
a window $X_i$ of observations onto $P_r$ is given by Equation [3.13]. The quality of the solution at the end of iteration is determined by computing the objective function $J$—the total loss across all the windows in the new configuration.

$$J = \sum_{i=1}^{N} \min_{r} Q(i, r)$$  \hspace{1cm} (3.15)

The algorithm converges to a stable solution and terminates when no cluster reassignments are possible. The $k$–PCA method, like any non–hierarchical clustering methods produces a different solution for each random initialization of the cluster prototypes. Also, it gets trapped in the local minimum of the solution space. Thus, solutions for each random initialization can be aggregated in a hierarchical fashion as described above in section [3.5] to obtain a final solution.

### 3.7 Conclusions

This chapter has introduced a class of data–driven methods, clustering algorithms. The traditional clustering algorithms are described in detail. The strengths and limitations of each method are discussed. It is noted that no single clustering method or choice of the similarity or distance metric yields the best result for a particular data set and hence cluster analysis has to be viewed as an exploratory analysis tool. A methodology that combines the positive aspects of the hierarchical and $k$–means algorithm, known as the hierarchical aggregation of $k$–means algorithm is described in detail. An algorithm to address the dynamic and correlated nature of time series observations is described.
Chapter 4

Sequence Analysis

4.1 Introduction

The field of data mining involves finding patterns and knowledge from data sets. There exist several data mining algorithms, of which sequence analysis is one of the methods. Mining data from sequences of data is known as sequence data mining or sequence analysis. Sequence analysis is the problem of determining similarity or distance metrics between categorical time series observations (Elzinga, 2003). Sequence analysis methods are independently and parallelly discovered by biologists and computer scientists. Computer scientists make the relationship between two sequences explicit by a list of edit operations while biologists align the sequences to determine the similarity. Historically, sequence analysis has been used extensively in a wide variety of fields such as molecular biology, genome informatics, sociological research and economics (Abbott, 1995).

In this chapter, sequence analysis methods are introduced. Section 4.2 provides a background on sequence analysis methods while section 4.3 details the methods available for finding the similarity between sequences. Section 4.4 elaborates the non–aligning similarity index along with an illustration and description of the algorithm. Section 4.5 discusses the combinatorial methods to count the subsequences used for developing similarity metrics. Finally, section 4.6 gives the conclusions drawn from the description of the theory in this chapter.
4.2 Background

Sequence analysis is an approach to data reduction that is used to unveil the pattern and structure in data. This approach is useful for discovering information in sequences. A sequence is an ordered set of tokens or symbols. The elements of sequences are known as events or tokens or symbols and are drawn from all possible universe of events. The events of a sequence can be unique or can repeat. A sequence in which events do not repeat is a nonrecurrent sequence. These sequences are sampled from the universe without replacement. The maximum length of a recurrent sequence is the size of the universe. A sequence in which events repeat - where the sequences sample the universe with replacements, is a recurrent sequence.

Sequence analysis involves calculating distances or similarities among sequences. A family of algorithms known as optimal matching (OM) is commonly used to determine the distance or similarity between sequences. An OM algorithm seeks to find the minimal number of edit operations (inserts, deletes and substitutions) necessary to turn a sequence into another sequence (Elzinga, 2003). These algorithms result in a perfectly aligned pair and are known as the string–to–string correction problem. OM is by definition an optimization problem that seeks to identify an optimal solution by considering all possible combinations of edit operations to find similarity between two sequences. This problem is solved by means of dynamic programming methods.

Sequence analysis is similar to OM algorithms and is a descriptive technique–an approach to data reduction that may or may not lead to identifying the pattern in sequences. OM algorithms are based on manipulations of time. In sequence analysis, however, the time of occurrence of events is central to the process. Insert or delete operations warp the timing of a process in order to identify the subsequences or tuples while substituting an event preserves the timing but the event in itself is approximated by another event. Warping time destroys the temporal relations that exist between sequences that share the same time scale. Thus, sequence analysis is a combination of operations to match subsequences of two sequences when the flow of time is normal. The output of sequence analysis is a dissimilarity matrix between sequences that is similar to correlation coefficients. The output of sequence analysis
is not accessed directly but is always combined with data reduction procedures such as cluster analysis.

4.3 Methods

Sequence analysis methods are classified into two main categories: character alignment methods and the non-aligning distance functions over sequence similarity methods. Measures based on sequence distance functions such as the Hamming (Euclidean) sequence similarity fall under the non-alignment methods.

4.3.1 Character Alignment Methods

The methods for the alignment of pairs of sequences can be put into three broad categories: (1) Dot matrix analysis (2) The dynamic programming (DP) algorithm (3) Word or k-tuple methods.

A dot matrix analysis, first described by Gibbs and McIntyre (1970), is primarily a method for comparing two sequences to look for possible similarity of characters between sequences. In the dot matrix method of pair wise sequence comparison, a sequence (x) is listed along the x-axis and the second sequence (y) is listed along the y-axis. The first character in sequence y is compared to the entire sequence x and a dot is placed in any column where the character in sequence x is the same. The procedure is repeated for all the characters of y until all the possible matches between the two sequences are listed. Any possible sequence similarity is displayed as diagonals on the matrix. The major advantages of this method are that it lists all possible matches between sequences and it can readily reveal the presence of direct and inverted repeats within a sequence which are difficult to find by other more automated methods while the major limitation is that most of the time this method does not show an actual similarity.

The dynamic programming method, first used by Needleman and Wunsch (1970), generates similarity between two sequences by starting at the ends of the two sequences and then attempts to match all possible pairs of characters between sequences by following a score scheme for matches, mismatches and gaps. This method provides the very best or optimal similarity between sequences. The dynamic programming method can be very slow as it
involves large number of computational steps, however. The complexity of computations
and the computer memory requirements in this approach increase as the square or cube of
the sequence lengths. Thus, this method is inappropriate for long sequences. Also, optimal
methods always yield the optimal solution even though the solution might not have a phys-
ical meaning. On the other hand, less optimal local solutions sometimes yield a good result.
Thus, it is a mistake to look only at the optimal sequence alignment solution.

The word or $k$–tuple methods find similarity between two sequences quickly by searching
for identical short stretch of sequences called words or $k$–tuples where $k$ is the length of
the short stretch. These methods then join the tuples to find the similarity between pairs of
sequences. These empirical methods that use rules of thumb to find similarity indices between
sequences are fast enough to be suitable for finding similarity between large numbers of long
sequences.

4.3.2 Non–Aligning Distance Functions over Sequences

A way of comparing sequences is by calculating the distance functions between sequences to
measure the similarity or dissimilarity. A sequence distance function bound between 0 and
1 satisfies the following four properties (Dong et al., 2007).

- $d(x,y) > 0$ for sequences $x$ and $y$ such that $x \neq y$,
- $d(x,x) = 0$ for all sequences $x$,
- $d(x,y) = d(y,x)$ for all sequences $x$ and $y$,
- $d(x,y) \leq d(x,z) + d(z,y)$ for all sequences $x$, $y$ and $z$.

A character counting method counts the number of a character type and creates a vector
of frequencies that add up to the sequence length. This method does not consider the sequential
arrangement of categorical observations and involves only the count of the observations.
Another non–alignment method known as the Hamming similarity is based on the count of
matches between characters at the corresponding positions in two sequences. This implies
that the sequences must be of the same length. A symbol in a sequence, however in reality,
can correspond to a symbol in a different position in the other sequence. Another approach
to finding similarity between sequences is a feature based metric. The first step in a feature based approach of distance calculation is to extract the features between sequences. Then, the distances between sequences are calculated as the distances between feature vectors of the sequences. In this dissertation, a similarity metric based on the character counting non-alignment methods, proposed by Elzinga (2003) and detailed in sections 4.4 and 4.5.2 are used to quantify the similarity between pairs of sequences.

4.4 Sequence Similarity Metric by Non–Alignment

Dijkstra and Taris (1995) define four principles to construct the similarity index, used to quantify the similarity among sequences. These four principles are:

1. Sequences with no common tokens are dissimilar maximally
2. Sequences with same tokens in the same order are maximally similar
3. Sequences with more tokens in common are more similar
4. Sequences with more common order among the common tokens are more similar

Using these principles, the sequence Similarity Index $S_{xy}$ is computed between all pairs of sequences $x$ and $y$. The relative degree of similarity between pairs of sequences $x$ and $y$ is determined using the concept of precedence $P$. In mathematics, an order is a binary relation that satisfies the properties of irreflexivity, asymmetry and transitivity and is defined on a Cartesian product set. In the theory of sequence analysis, however, the concept of order as defined in mathematics does not apply. Elzinga (2003) defines an interpretation for the order in a sequence as ‘precedence’. Token $a$ precedes $b$ if in a sequence $x$, $a$ is encountered prior to $b$ when reading the sequence from left to right. This is represented as $aPb$. However, $aPb$ does not imply that $a$ and $b$ are consecutive elements. This approach does not assume any property of the precedences or anything about a spatial representation of the sequences. It just uses all of the information on precedences within sequences. The binary precedence relation is reflexive, symmetric and intransitive and these properties become significant for sequences with repetitive indices. The precedence relation is reflexive as a sequence with a token $a$ appearing in more than one position will have a preceding $a$ i.e. $aPa$. A sequence of
the form \( a, b, c, d, a \) has \( aPb, bPc, aPc \), but also has \( cPa \). Thus, the relation is intransitive. Also, there is \( aPc \) and \( cPa \). This enables the symmetry to the precedence relation \( P \). Thus, this index discriminates sequences accurately as it can quantify sequence of symbols by considering the above properties. \( S_{xy} \) is normalized to range on \([0, 1]\), with values of unity indicating maximum similarity between sequences \( x \) and \( y \) \((S_{xx} \) is defined as unity).

\[
S_{x,y} = \frac{1}{L} \sum_{l=1}^{L} \frac{m_{xy}(l)}{\sqrt{m_{xx}(l) m_{yy}(l)}}
\]

(4.1)

Here, \( L \) is the minimum of the sequence lengths of \( x \) and \( y \), \( m_{xy}(l) \) is the number of matching ordered \( l \)-tuples (subsequences with lengths of exactly \( l \) hours) to appear in both sequences \( x \) and \( y \) with disregard to their positions in \( x \) and \( y \) at which the matching sequences appear. Quantities \( m_{xx}(l) \) and \( m_{yy}(l) \) are defined similarly to indicate the numbers of repeated \( l \)-tuples occurring within \( x \) and \( y \), respectively. The complexity of listing a \( l \)-tuple in a sequence and matching this to other \( l \)-tuples from other sequences is roughly exponential to the sequence length. Hence, a computationally efficient, iterative method for computing \( S_{x,y} \) is given by Elzinga (2003) and is discussed below.

### 4.4.1 Algorithm

Given two sequences \( x = \{x_1, x_2, \ldots\} \) and \( y = \{y_1, y_2, \ldots\} \) of lengths \( l_x \) and \( l_y \) respectively, the algorithm begins first by listing all the pairs of matching elements from \( x \) and \( y \) in a matrix \( W \) with two columns and at most \((l_x \times l_y)\) rows. Matrix \( W \) is defined as \([w(k, 1) = i \& w(k, 2) = j] \iff x_i = y_j\).

An \( l_x \times l_y \) matrix \( H_1 \) is then defined such that \( h_1(i, j) = 1 \iff x_i = y_j \) and 0 otherwise. The number of matches when each element of \( x \) is compared with those of \( y \) is obtained as the sum of all the elements of \( H_1 \).

\[
m_{x,y}(1) = \sum_{i,j} h_1(i,j)
\]

(4.2)
Next, a matrix $V_1$ is constructed from $H_1$ such that

$$v_1(i, j) = \sum_{a > i, b > j} h_1(a, b) \quad (4.3)$$

The elements of $V_1$ contain the matches when matching the elements of $x$ with those of $y$ disregarding the first $i$ elements of $x$ and the first $j$ elements of $y$. A new matrix $H_2$ is defined from $V_1$ as $h_2(i, j) = v_1(i, j) \Leftrightarrow (i, j) \in Z$ and 0 otherwise. The number of matching 2-tuples from $x$ and $y$ are obtained as

$$m_{x,y}(2) = \sum_{i,j} h_2(i, j) \quad (4.4)$$

Matrix $V_2$ is defined from $H_2$ similar to $V_1$ from $H_1$ and consists of the number of matching 2-tuples among sequences $x$ and $y$ disregarding the first $i$ elements of $x$ and the first $j$ elements of $y$. This procedure is repeated for all the tuples ranging from 2 to $n$ where $n$ is minimum of the length of the sequences of $x$ and $y$. Once all the matching tuples are obtained, similarity index between sequences $x$ and $y$ is calculated by using Equation (4.1).

This method is illustrated below. Let $x = \{1 \ 2 \ 3 \ 2 \ 3 \ 4 \ 5\}$ and $y = \{1 \ 2 \ 3 \ 4 \ 5\}$. Thus, $l_x = 7$ and $l_y = 5$ and $L = \min(l_x, l_y) = 5$. From these sequences, as described above we have

$$W^T = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 2 & 3 & 2 & 3 & 4 & 5 \end{bmatrix};$$

$$H_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$
Thus, $m_{x,y}(1) = 7$. $V_1$ is then constructed as

$$V_1 = \begin{bmatrix}
6 & 5 & 4 & 3 & 2 & 1 & 0 \\
4 & 4 & 3 & 3 & 2 & 1 & 0 \\
2 & 2 & 2 & 2 & 2 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix};$$

$H_2$ is obtained from $V_1$ as

$$H_2 = \begin{bmatrix}
6 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 4 & 0 & 3 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 2 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix};$$

and $m_{x,y}(2) = 18$. This method is continued until no matching subsequences can be found in between the sequences. In this illustrative case, only matching tuples up to 5 can be constructed as $V_5 = 0$. The above algorithm can be generalized as:

1. Create a matrix $H_n$ from the matrix $V_{n-1}$ by

   $$h_n(i, j) = v_{n-1}(i, j) \Leftrightarrow (i, j) \in W$$

2. Compute $m_{x,y}(l)$ as

   $$m_{x,y}(l) = \sum_{i,j} h_l(i, j)$$

3. Create $V_l$ from $H_l$ as

   $$v_l(i, j) = \sum_{a\beta, b\beta} h_l(a, b)$$

4. Iterate the procedure until $m_{x,y}(l) = 0$. 

44
4.5 Algorithms for Counting Common Subsequences

A sequence that can be obtained by deleting some symbols from another sequence is referred to as the subsequence of the original sequence. A subsequence that is common to both the sequences is a common subsequence. A common subsequence of the maximum possible length is the longest common subsequence (LCS). The order of the common tokens from both the sequences must be the same although the subsequence may not be the contiguous part of the sequence. There exists a wide variety of dynamic programming algorithms to solve the problem of counting the number of different subsequences, the number of different common subsequences, the number of longest common subsequences, the number of different subsequences with a specific length. In this dissertation, we explore the metrics based on the number of common subsequences (NCS) and the length of the longest common subsequence (LLCS).

4.5.1 Number of Common Subsequences

Elzinga et al., (2007) discuss the theorem and algorithm to compute the number of common subsequences. The algorithm is given below (Figure 4.1) along with an example parallelly. The algorithm begins with first enumerating the character position in a sequence along the length of the sequence. If a character is repeated, the latest position of the character replaces the listing for the character. This, for a sequence \( x = \{bbac\} \) and \( y = \{bca\} \) translates to \( l_x(i, \sigma) \) where \( \sigma \) is a character in the sequence:

Table 4.1: An example to illustrate the computation of number of common subsequences: part 1

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>i=0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>j=0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Then, the number of common subsequences is computed (Table 4.2) following the algorithm of Figure 4.1 as follows:

In the above example, an example to show the usage of algorithm in Figure 4.1 is shown by using it to compute the element \( N(2,1) \). This computation needs the elements \( N(1,0) \)
for $i = 0:l_x$
    $N(i, 0) = 1$;
end
for $j = 1:l_y$
    $N(0, j) = 1$;
end

for $i = 1:l_x$
    for $j = 1:l_y$
        $N(i, j)= N(i-1, j)$
        if $l_x (j, x_i) > 0$
            $N(i, j)= N(i, j)+ N(i-1, l_y (j, x_i)-1)$
        end
        if $l_x (i-1, x_i) > 0$
            $N(i, j)= N(i, j) - N(l_x (i-1, x_i)-1, l_y (j, x_i)-1)$
        end
    end
end

Figure 4.1: Algorithm to compute the number of common subsequences.
Table 4.2: An example to illustrate the computation of number of common subsequences:
part 2

<table>
<thead>
<tr>
<th>(N(i, j))</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

and \(N(1, 1)\). These are assumed to be known for the purpose of this calculation. Thus, from Table 4.2, \(N(1, 0) = 1\) and \(N(1,1) = 2\).

1. \(N(2,1) = N(1,1) = 2\)

2. \(l_y(j,x_i) = l_y(1,b) = 1 > 0\) thus, \(N(2,1) = N(2,1) + N(1,1-1) = 2 + 1 = 3\)

3. \(l_x(1-1,x_i) = l_x(1,b) = 1 > 0\) thus, \(N(2,1) = N(2,1) - N(1-1, 1-1) = 3 - 1 = 2\)

Thus, \(N(2,1)\) is calculated from the above algorithm as 2. Thus, a total of 6 subsequences can be obtained from the above sequences \(x\) and \(y\). Once the number of common subsequences is determined, the distance metric based on the NCS can be computed as

\[
d_{NCS}(x, y) = NCS(x, x) + NCS(y, y) - 2NCS(x, y) \tag{4.5}
\]

4.5.2 Length of Longest Common Subsequence Metric

Given two sequences \(x= \{abacbb\} y = \{bcabbc\}\), the longest common subsequence is \(u = \{babb\}\). Thus, the length of the common subsequence is 4 and \(l_x = 7\) and \(l_y = 6\). The length of the LCS is unique even if the common subsequence is not. The distance metric based on the LLCS can be computed as

\[
d_{LLCS}(x, y) = LLCS(x, x) + LLCS(y, y) - 2LLCS(x, y) \tag{4.6}
\]

The LLCS is bound between 0 and minimum of \(l_x l_y\) and the complexity to compute the LLCS us proportional to \(l_x l_y\). The algorithm to compute the LLCS as written in an unpublished work by Greenberg (2003) is given below. Let \(L(i, j)\) be the length and \(D(i, j)\) be the number of distinct LCSs in sequences \(x\) and \(y\). \(L(i, j)\) is given as:

47
for $j = 0: l_y$
  for $i = 0: l_x$
    if $i = 0$ or $j = 0$
      $D(i, j) = 1$;
    else
      $D(i, j) = 0$;
    end
    if $x_i = y_j$
      $D(i, j) = D(i-1, j-1)$;
    elseif $L[i-1, j] = L(i, j)$
      $D(i, j) = D(i, j) + D(i-1, j)$;
    elseif $L[i, j-1] = L(i, j)$
      $D(i, j) = D(i, j) + D(i, j-1)$;
    else
      $D(i, j) = D(i, j) - D(i-1, j-1)$
    end
  end
end

Figure 4.2: Algorithm to compute the length of the longest common subsequence.

\[ L(i, j) = 0 \quad \text{if} \quad i = 0 \quad \text{or} \quad j = 0 \] (4.7)

\[ = L(i - 1, j - 1) + 1 \quad \text{if} \quad i, j > 0 \quad \text{and} \quad x_i = y_j \]

\[ = \max \{L(i - 1, j), L(i, j - 1)\} \]

The algorithm to compute the LLCS is given in Figure 4.2.

4.6 Conclusions

This chapter introduced the concept of sequence analysis. These methods though popular in computer science, biological sciences and social sciences are not widely used in the field of air pollution. The theory behind the sequence similarity metrics and the combinatorial sequence algorithms are given in detail in this chapter along with illustrative examples.
Chapter 5

Hidden Markov Models

5.1 Introduction

Discrete-valued time series though common in practice have no well known methods for their analysis. When observations in a data set are categorical in nature or are quantitative but very small, models that consider the discrete nature of the data set are beneficial. This chapter examines a class of discrete-valued time series models known as hidden Markov models. These models describe not only the univariate discrete-valued time series but also multivariate discrete-valued series and real valued time series. This chapter presents the theory behind these models and highlights their scope and applicability to real world problems such as ozone value prediction. These models are structurally simple and are sufficiently versatile to apply to many real world problems. Many statistical questions regarding the HMMs such as the theoretically justified model selection techniques, however, remain unanswered.

In section 5.2, a background on HMMs along with model selection criteria is provided. Section 5.3 gives a description of the model, the elements of the HMM, the problems in HMM along with the solution. Section 5.4 provides the framework for multiple observation training; section 5.5 provides the statistical measures for verification of forecasts and section 5.6 describes an overview of the implementation of multivariate or multiple observations HMM training. Section 5.7 gives the summary and conclusions of this chapter.
5.2 Hidden Markov Models Background

Probabilistic fields are widely used to simulate and predict the behavior of stochastic systems in numerous fields of study. HMMs were developed and described in a series of papers by a number of statisticians including Baum and Petrie (1966), Baum and Eagon (1967), Baum and Sell (1968), Baum et al., (1970) and Baum (1972). HMMs are a finite and a doubly stochastic probabilistic approach to study a discrete time series (Rabiner and Juang, 1986). The theory and the application of HMMs to real world problems such as speech recognition are covered extensively in a tutorial by Rabiner (1989).

A HMM is a double stochastic model where observations are modeled conditioned on the small number of discrete states, with Markovian transitions between the states. Each state is associated with an independent probability distribution called the Emission Probability Distribution. When each time a state is visited it emits a symbol or observation depending on its state distribution density function. A starting state is chosen according to an Initial Probability distribution function while the transition between states is given by a set of probabilities called the transition probability distribution function. The states visited are, however, not visible and thus the model is named ‘hidden’ Markov model.

HMMs are based on three main assumptions. (1) The Markovian assumption which states that the probability of the next state depends on the current state i.e. the model does not remember the past states other than the current state. (2) The Stationary assumption which states that the transition probabilities are independent of the time at which the transition actually occurs. (3) The Independence assumption in which the symbol emitted by a state is independent of the symbols emitted by the state previously.

The accuracy of learning the parameters of the model from the observations plays a crucial role in simulation and prediction. Several approaches exist for training the HMMs from observations. The Baum–Welch algorithm, a dynamic programming algorithm of the Expectation–Maximization, is the most widely used method for HMM parameter learning.
This algorithm performs a re-estimation of initial guess parameters of the model to maximize the likelihood of a given sequence of observations in the model. The above procedure is repeated until the probability of observation of the training data given by the new model shows no improvement over the previous model.

One of the limitations of Baum–Welch algorithms is that they suffer from the strong dependence on the choice of initial parameters for the model. For example, there is no definitive way to determine the number of states of a HMM. Choosing a smaller number of states results in faster training time and analysis but could result in an unsatisfactory classification. On the contrary, choosing a large number of states slows the training and may also give unsatisfactory classification due to an overfitting phenomenon. For each random initialization of the model, the algorithm may converge to sub-optimal local maxima. Thus, several iterations of the algorithm with different initial choice for the models are required to find an optimal solution. The Bayesian Inference Criterion (BIC) is used to address the problem of distinguishing competing models. The BIC (Schwarz, 1978) takes the form

\[ BIC = 2(ll) - k \log(T) \] (5.1)

where \( ll \) is the log-likelihood, \( k \) is the number of model parameters and \( T \) is the number of observations used for training the model. The model that corresponds to minimum BIC value is selected.

### 5.3 Model Description

A brief overview of the model is given below using the classic urn and ball example (Rabiner, 1989). Consider \( N \) urns with a large number of colored balls. These balls are assumed to have \( T \) distinct colors (Figure 5.1). Observation sequences are obtained by selecting an urn and a ball from the urn based on probabilistic distributions of the urns and the balls. The ball is replaced in the urn after recording its color as the observation and the procedure is repeated to generate a finite observation sequence of colors. This sequence of colors forms the observable output of the model. The objective of this model is to explain the observed
sequences of balls with out any prior knowledge of the urns (hidden layer) generating the sequences.

Figure 5.1: An urn and ball model to illustrate a hidden Markov model with $N$ states

5.3.1 Elements of HMM

HMMs are characterized by five parameters:

- $N$, the number of hidden states of the HMM. The states are denoted as $S = \{s_1, s_2, ..., s_N\}$ and the state at time $t$ is $q_t$.

- $T$, the number of observation symbols denoted as $V = \{v_1, v_2, ..., v_T\}$ corresponding to the output of a process. In the above example, these were the colors of the balls.

- The state transition probability matrix $A = \{a_{ij}\}$

where,

$$a_{ij} = P[q_{t+1} = s_j \mid q_t = s_i].$$  \hspace{1cm} (5.2)

i.e the probability of transiting from a state $s_i$ at time $t$ to the state $s_j$ at $t+1$.

- The probability distribution of observations in state $i$

$$b_i(k) = P[v_k \text{ at } t \mid q_t = s_i] \hspace{1cm} 1 \leq i \leq N \hspace{1cm} (5.3)$$

$$1 \leq k \leq T$$
Figure 5.2: Graphical representation of a homogeneous hidden Markov model.

- The initial state distribution $\gamma = \{\gamma_i\}$ i.e. the probability of being in state $i$ at time $t=1$.

$$\pi_i = P[q_1 | s_i]. \quad (5.4)$$

Thus, given the number of states and the number of symbols, a HMM $\lambda$, can be uniquely described by $\gamma$, $A$, and $B$ as

$$\lambda = (\gamma, A, B). \quad (5.5)$$

The graphical model for an HMM is shown in Figure 5.2.

If the transitions between states depend not just on the previous state but on additional variables i.e.

$$a_{ij} = P[q_{t+1} = s_j | X_t, q_t = s_i]. \quad (5.6)$$

the model then is a non–homogeneous HMM. A graphical interpretation of a non–homogeneous model is given in Figure 5.3.

5.3.2 The Three Basic Problems of HMMs

The HMM $\lambda$ can be used to compute the probability of observation of a sequence of output symbols, $O = \{o_1, o_2, ..., o_T\}$; to determine the most likely state sequence, given the model and a set of observations $O$ and to adjust the parameters $\gamma$, $A$, and $B$ in order to maximize $P(O | \lambda)$. These three problems can be stated as:

53
Figure 5.3: Graphical representation of a non–homogeneous hidden Markov model.

1. Evaluation Problem: Given a model and a sequence of observations, to compute the probability of an observation sequence being produced by the model.

2. Optimal State Sequence Problem: This problem attempts at uncovering the hidden part of the model. Given an observation sequence and a model, objective of this problem is to find the “correct”state sequence that best explains the observations.

3. Model Training Problem: To optimally adjust the model parameters so as to best describe the occurrence of a given observation sequence.

5.3.3 Solutions to the Three Basic Problems of HMMs

5.3.3.1 Evaluation Problem: The Forward–Backward Procedure

The first two of the three problems are solved in the test phase while the third problem is solved in the training phase. The training problem determines the set of model parameters that best explain the observed data. The Baum–Welch re–estimation algorithm, also referred to as the forward–backward algorithm (Baum, 1972) that uses maximum likelihood estimation approach is used to solve both the training and the testing problems. The forward–backward algorithm consists of the computation of the ‘forward probabilities’ and the ‘backward probabilities’. These probabilities are defined as follows– for all states $i$ of
the Markov chain and for all $t$ from 1 to $T$:  

$$
\alpha_t(i) = P(o_1 o_2 ... o_t, q_t = S_i \mid \lambda). \quad (5.7)
$$

and  

$$
\beta_t(i) = P(o_{t+1} o_{t+2} ... o_T, q_t = S_i \mid \lambda). \quad (5.8)
$$

$\alpha_t(i)$ is the probability of observing a partial observation sequence $O$ until time $t$ and state $S_i$ at time $t$. The forward algorithm performs the following recursive computation:

1. For $i = 1, 2, ..., N$

$$
\alpha_1(i) = \gamma_i b_i(o_1) \quad (5.9)
$$

2. For $j = 1, 2, ..., N$ and for $t = 1, 2, ..., T$

$$
\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \quad (5.10)
$$

The first step of the forward algorithm is the initialization step and it computes the initial state distribution for each state for an observation of unit length. The second step inductively computes the state distribution for an observation sequence of length $t > 1$ from the state distribution vector of its $(t-1)$ length predecessor.

$\beta_t(i)$ is the probability of observation sequence from $t+1$ to the end of the sequence given the model and the state at instant $t$ as $i$. The backward algorithm can be solved inductively as:

1. For $i = 1, 2, ..., N$

$$
\beta_T(i) = 1 \quad (5.11)
$$

2. For $j = 1, 2, ..., N$ and for $t = 1, 2, ..., T$

$$
\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \quad (5.12)
$$

The probability of observing a sequence $O = \{o_1, o_2, ..., o_t\}$ given the model $\lambda$ can then
be obtained as:

\[ P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i), \forall t \]  \hspace{1cm} (5.13)

or

\[ P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \] \hspace{1cm} (5.14)

This is the sum of the terminal forward variables \( \alpha_T(i) \)

### 5.3.3.2 Model Training: Updating Parameters and Probability Calculations

The Baum–Welch algorithm iteratively updates the parameters of a HMM to produce a new model that has a higher probability of generating a given observation sequence. The iterative procedure terminates when no more significant improvement in probability can be obtained.

The parameter updating equations of the algorithm are defined in terms of joint events and state variables. Joint event is the probability of being in state at time \( t \) and state at time \( t+1 \) and is defined as

\[ \xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda) = \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \frac{P(O \mid \lambda)}{P(O \mid \lambda)} \] \hspace{1cm} (5.15)

while the probability of being in state at time \( t \) is defined as the state variable and is computed as:

\[ \phi_t(i) = P(q_t = S_i \mid O, \lambda) = \sum_{j=1}^{N} \xi_t(i, j) \] \hspace{1cm} (5.16)

For \( i = 1,2,...,N \); \( j = 1,2,...,N \) and \( k = 1,2,...,T \) the parameter updating equations are:

\[ \bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \phi_t(i)} \] \hspace{1cm} (5.17)

\[ \bar{b}_j(k) = \frac{\sum_{t=1}^{T} \rho_t = v_k \phi_t(j)}{\sum_{t=1}^{T} \phi_t(j)} \]

\[ \bar{\gamma}_i = \phi_1(i) \]
5.3.3.3 Optimal State Sequence – the Viterbi Algorithm

Given an observation sequence \( O = o_1, o_2, ..., o_T \) the objective of the Viterbi algorithm is to find the best state sequence \( Q = q_1, q_2, ..., q_T \). The variable \( \delta_t(j) \) stores the probability of observing the partial sequence \( o_1, o_2, ..., o_t \) using the most likely path ending in state \( i \) at time \( t \) given the model \( \lambda \). A variable \( \Psi_j(t) \) corresponding to this variable stores the argument that leads to the most probable path. The complete procedure for finding the best state sequences is as follows:

1. For \( i = 1, 2, ..., N \)

\[
\delta_1(i) = \pi_i b_i(o_1) \\
\Psi_1(i) = 0
\] (5.18)

2. For \( j = 1, 2, ..., N ; t = 2, 3, ..., T \)

\[
\delta_t(j) = \max\left[ \delta_{t-1}(j) a_{ij} \right] b_j(O_t) \\
\Psi_t(j) = \arg\max\left[ \delta_{t-1}(j) a_{ij} \right]
\] (5.19)

3. \( P^* = \max[\delta_T(j)] \) (5.20)

\( q_T^* = \arg\max[\delta_T(j)] \)

4. For \( t = T-1, T-2, ..., 1 \)

\[
q_t^* = \Psi_{t+1}(q_{t+1}^*)
\] (5.21)

The third step in the algorithm is the termination step while the fourth step of the algorithm keeps track of the state sequence path. The implementation of the Viterbi algorithm is similar to the forward–backward algorithm, except that Viterbi algorithm uses maximization over the preceding states where the forward-backward algorithm uses a summing procedure.
5.4 Multiple Observations Training

HMMs were initially limited to obtaining the parameters of the model by training a single observation sequence. This method has been extended to training from multiple observation sequences. The initial multiple sequence methods assumed the multiple sequences to be statistically independent. Several HMM estimations obtained from training the individual sequences are combined to a single HMM. For the case of set of multiple observation sequences, Li et al., (2000) present a formal treatment in which the multiple observation probability is expressed as a combination of individual observation probabilities. Their method uses each set of observation sequence to iteratively update a single HMM parameter set. They generalize the auxiliary function of Baum to derive the training equations that maximize the convergence of the training process. Given a set of observation sequences \( O = \{O^{(1)}, O^{(2)}, \ldots, O^{(K)}\} \), where,

\[
O^{(k)} = o_1^{(k)} o_2^{(k)} \ldots o_T^{(k)}, \quad 1 \leq k \leq K
\]

are individual observation sequences, the multiple observation probability is expressed as:

\[
P(O \mid \lambda) = \prod_{k=1}^{K} P(O^{(k)} \mid \lambda) \tag{5.22}
\]

and the updating parameters for \( i = 1,2,\ldots,N; j = 1,2,\ldots,N \) and \( k = 1,2,\ldots,K \) are

\[
\bar{a}_{mn} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \xi_t^{(k)}(m,n)}{\sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \phi_t^{(k)}(m)}, \quad 1 \leq m \leq N, 1 \leq n \leq N \tag{5.23}
\]

\[
\bar{b}_n(m) = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \phi_t^{(k)}(n)}{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \phi_t^{(k)}(n)}, \quad 1 \leq n \leq N, 1 \leq m \leq M \tag{5.24}
\]

\[
\bar{\gamma}_n = \frac{1}{K} \sum_{k=1}^{K} \phi_t^{(k)}(n), \quad 1 \leq n \leq N \tag{5.25}
\]
5.5 HMMs for Forecasting

A HMM trained on a set of observations can be used in different ways. A number of HMMs can be trained on different categories of observations and given a new sequence, the probability of the new observation belonging to the HMMs can be determined. Thus, HMMs can be used to determine the overall system classification. This is termed the categorical forecasting. A point–value prediction on the other hand involves the use of HMMs to determine the actual values of an observation.

5.5.1 Verification Statistics for Categorical Forecast

To evaluate a two category forecast model, a contingency table that shows the frequency of occurrence of observed and forecasted events is constructed. A contingency table for two categories consists of elements $e$, $f$, $g$, $h$, that represent the number of correctly forecasted non exceedances, false alarms, missed exceedances, and the correctly forecasted exceedances, respectively (Figure 5.4). A perfect forecast implies zero missed events and zero false alarms (i.e., $f=0$ and $g=0$ in Figure 5.4).

Using the values in the contingency table, verification statistics which evaluate the performance of a forecast model are calculated. These statistics consist of Accuracy ($A$) – the percent of correctly predicted exceedance and non–exceedances; False Alarm Rate ($FAR$) – the percent of times a forecast high ozone exceedance is not observed; Critical Success Index ($CSI$) – how well the ozone exceedances are predicted; Probability of Detection ($POD$) –
ability to predict ozone exceedance events, Frequency of Hits (FOH); Frequency of Misses (FOM); Detection Failure Ratio (DFR); Frequency of Correct Null (FOCN); Probability of Null event (PON); Probability of False Detection (POFD). The equations for computing these measures are: 

\[
A = \frac{(e+h)}{(e+f+g+h)} \times 100; \quad \text{FAR} = \frac{f}{(f+h)} \times 100; \quad \text{CSI} = \frac{h}{(f+g+h)} \times 100; \quad \text{POD} = \frac{h}{(g+h)} \times 100; \quad \text{DFR} = \frac{g}{(e+g)}; \quad \text{PON} = \frac{e}{(e+f)}; \quad \text{FOH} = 1-\text{FAR}; \quad \text{FOM} = 1-\text{POD}; \quad \text{FOCN} = 1-\text{DFR}; \quad \text{POFD} = 1-\text{PON}.
\]

An interpretation for certain aspects of the 2 × 2 contingency table is given in Figure 5.5 (Doswell et al., 1990; Ghiaus., 2005). The dotted line represents a perfect model with no false alarms and no missed exceedance events. The solid lines are the regression lines of observations upon forecasts and forecasts upon observations. The distance between the dashed and solid lines indicate the performance of the model: the closer the solid lines to the dashed line, the better the model performance.

An alternative measure for forecast verification is the two alternative forced choice (P_{2AFC}) test. A P_{2AFC} test for dichotomous forecasts (Mason et al., 2009) determines the skill of a forecast to discriminate between observations based on the proportion of correctly picked observations (Equation 5.26). The denominator is the total number of forecast observation pairs and is the product of the events and nonevents. The numerator is the number of
correctly discriminated observations plus half the number of observations that cannot be
discriminated in either of the two categories. For a correct skill forecast, the $P_{2AFC}$ score
exceeds 50% with better forecasts close to 100%.

$$P_{2AFC} = \frac{eh + 0.5(fh + ge)}{(e + f)(g + h)}$$  \hspace{1cm} (5.26)

5.6 Implementation of HMMs

In this dissertation, two toolboxes are used for developing the HMMs. The first one (Mur-
phy, 1998) is a set of MATLAB algorithms to solve the HMM modeling problem. This
toolbox is used to develop the HMMs for classifying a test observation sequence into one of
several classes. The second one is a C++ implementation of the algorithms developed by
Kirshner (2005). This toolbox is used to train the HMMs for use in point-value predictions
of observations.

5.6.1 MATLAB Implementation of HMM Toolbox

This toolbox consists of routines that support inference and learning for HMMs with both
discrete, Gaussian and mixtures of Gaussian output. The toolbox consists of three direc-
tories - HMM, KPMstats, and KPMtools. These directories contain MATLAB implementa-
tion of the Baum–Welch learning algorithm and the supporting statistical software for
HMMs, miscellaneous statistics functions and miscellaneous matlab functions respectively.
The MATLAB scripts used explicitly in this dissertation are given below:

- dhmm_em.m: to learn the parameters of an HMM with discrete outputs

- dhmm_logprob.m: to evaluate the log-likelihood of a trained model given discrete input
test data

- mhmm_em.m: to compute the parameters of an HMM with (mixtures of) Gaussians
  output

- mhmm_logprob.m: to compute the log-likelihood of a data set using a (mixtures of)
  Gaussians HMM
The algorithm begins by making a random guess for the initial parameters:

\[
\begin{align*}
\text{prior1} &= \text{normalise}(\text{rand}(Q,1)); \\
\text{transmat1} &= \text{mk\_stochastic}(\text{rand}(Q,Q)); \\
\text{obsmat1} &= \text{mk\_stochastic}(\text{rand}(Q,O));
\end{align*}
\]

The function 'normalise' makes the sum of the entries of an array equal to 1. $Q$ is the number of states. 'mk\_stochastic' is a function to ensure that the argument is a stochastic matrix, i.e., the sum over the last dimension is 1. 'O' is the number of discrete symbols in the data.

The guess parameters are then improved as:

\[
[\text{LL}, \text{prior2}, \text{transmat2}, \text{obsmat2}] = \text{dhmm\_em}(\text{data}, \text{prior1}, \text{transmat1}, \text{obsmat1}, '\text{max\_iter}', 5);
\]

'max\_iter' is the number of maximum EM iterations to be performed to be able to reach the convergence threshold of 1e-4 (default). The data can be a vector or a matrix of observations.

The log-likelihood of the above trained model given test data proceeds as follows:

\[
\text{loglik} = \text{dhmm\_logprob}(\text{testdata}, \text{prior2}, \text{transmat2}, \text{obsmat2})
\]

The function 'dhmm\_logprob' uses the forward-backward algorithm to compute the log-likelihood of the data.

---

Figure 5.6: A demonstration script for implementing the discrete HMM using MATLAB Toolbox.

A demo of the HMM program with discrete output is given in the script dhmm\_em\_demo.m; it gives an example of how to learn an HMM with discrete outputs (Figure 5.6) while the script mhmm\_em\_demo.m provides a demonstration (Figure 5.7) of the mixture of Gaussian output program. These scripts are reproduced below:

### 5.6.2 Multivariate Nonhomogeneous Hidden Markov Model (MVN–HMM) Toolbox

The MVN–HMM toolbox (Kirshner, 2005) is a collection of algorithms to model vectors of data by means of HMMs. The toolbox is developed in C++ and can be implemented on the windows platform. This toolbox consists of algorithms used to determine the parameters of the HMM using Baum–Welch algorithms, to obtain the most likely state sequence.
The algorithm begins by making a random guess for the initial parameters:

\[
\text{prior0} = \text{normalise(rand}(Q,1)));
\text{transmat0} = \text{mk\_stochastic(rand}(Q,Q)));
\text{[mu0, Sigma0]} = \text{mixgauss\_init}(Q\times M, \text{data, cov\_typeouden})\);
\text{Sigma0} = \text{reshape(Sigma0, [O O Q M])};
\text{Mixmat0} = \text{mk\_stochastic(rand}(Q, M)));
\text{mu0} = \text{reshape(mu0, [O Q M])};
\]

The function 'mixgauss_init' estimates the initial parameters for a mixture of \( M \) Gaussians. The function 'reshape' returns an \( N \) dimensional array with the same elements as an \( m \)-by-\( n \) matrix \( X \) but reshaped to have the size \( m \)-by-\( n \)-by-\( P \). \text{cov\_type} is the parameter to specify the covariance matrix type. Options for this parameter include: 'diagonol', 'symmetric' and 'full'. \text{mu0, Sigma0} is the mean and the covariance of the Gaussian.

Next, the parameters are improved as:

\[
\text{[LL, prior1, transmat1, mu1, Sigma1, mixmat1]} = \ldots
\text{mhmm\_em(data, prior0, transmat0, mu0, Sigma0, mixmat0, 'max\_iter', 2)};
\]

The log-likelihood of a trained model given test data is evaluated as follows:

\[
\text{loglik} = \text{mhmm\_logprob(testdata, prior2, transmat2, obsmat2)}
\]

Figure 5.7: A demonstration script for implementing the continuous HMM using MATLAB Toolbox.
using Viterbi algorithm, to estimate the log–likelihood of data as well as to compute simple
statistics such as mean of the data. The code after compilation can be run by adding the
directory of the compiled code to path. The command line to run the program is \texttt{mvnhmm<parameter file>}.

The first step to use this toolbox is to specify a \texttt{parameter} file. The \texttt{parameter} file consists of all the options for the parameters with which the program will run. There exist four types of parameters with various options for each parameter.

1. The mode in which the program is to be run: MVN–HMM can be run in the \texttt{action}, \texttt{xval_type} and \texttt{examples_out} modes.

The \texttt{action} mode gives the task the program has to perform. The type of action can be \texttt{learn} (to learn the HMM parameters), \texttt{viterbi} (to find the most optimal hidden state sequence path given the observations and the HMM model parameters), \texttt{ll} (to determine the log–likelihood of a data given the HMM model parameters), \texttt{simulation} (to simulate data given the HMM parameters) and analysis (to calculate the statistics of data in the pre–processing step. This option is independent of the HMM model).

\texttt{xval_type} parameter is the method for cross–validation. Cross–validation is a method to determine how the model performs if some training sequences are left out. The mean of the sum of squares error after cross–validation by leaving out few observation sequences should be low. This mode consists of two options: \texttt{none} in which there is no cross–validation i.e. all the input data sequences are used in the entire model runs; \texttt{leave_n_out} cross–validation requires the number of sequences to be left out as the input. \texttt{examples_out} is a positive integer that specifies the number of sequences to leave out in the \texttt{xval_type} cross–validation. The parameter \texttt{examples_out} is bound between 1 and half the total number of sequences.

2. Model related parameters: \texttt{num_states}, \texttt{model_type}, \texttt{emission}.

\texttt{num_states} is a positive integer and is the number of hidden states for the model. \texttt{model_type} is the type of the HMM. The probable types of HMM are: \texttt{hmm}– which is the homogeneous hidden Markov model and \texttt{nhmm}–the non-homogeneous HMM.
In a nhmm, the transitions between states vary with input variables such as time. 

**emission** specifies the probability distribution type for a hidden state and this is assumed to be the same for all the hidden states. Few of the distribution types are independent bernoulli, gaussian, independent delta–exponential, independent delta–gamma.

**independent** $<\text{num}_1> \text{ bernoulli } <\text{num}_2>$ is a conditionally independent Bernoulli distribution on $\text{num}_1$ variables with $\text{num}_2$ outcomes. **gaussian** $<\text{num}>$ is the normal distribution with one parameter $\text{num}$ which is the number of data variable components. **independent** $<\text{num}_1> \text{ delta–exponential}$ distribution is a conditionally independent distribution on $\text{num}_1$ variables with a mixture of point mass at zero (Dirac Delta function) and $\text{num}_2$–1 exponential functions on $\text{num}_1$ variables. **independent** $<\text{num}_1> \text{ delta–gamma}$ distribution is a conditionally independent distribution on $\text{num}_1$ variables with a mixture of point mass at zero (Dirac Delta function) and $\text{num}_2$–1 gamma distribution functions on $\text{num}_1$ variables.

3. Input and Output related parameters:

**Data** $<\text{filename}>$ is an input file that contains the data to be analyzed.

**num_data_sequences** specifies the number of vectors of data sequences with equal lengths. **data_sequence_length** is the length of the vectors of data sequences.

**model_filename** is the output from HMM learn that contains the parameters of the model from training the observation sequences. **output** is the location of the file for the results of the program. **num_discrete_data_components** and **num_real_data_components** are the number of discrete and real valued components in a data vector respectively.

4. Specifics and Auxiliaries: These are **num_restarts** which is the random number of initializations for EM used during the estimation of the parameters of the HMM. This option is used only with the **learn** mode; **em_precision** that determines the sensitivity limit for the EM optimization algorithm. The algorithm terminates if the difference in the log–likelihood falls below the threshold; **filling** is the option used for all evaluations
related to the hole filling analysis. Hole filling attempts to fill the missing entries such as the cumulative log probability of the left out data (‘log_p’), probabilities of hidden states for an out of sample data (‘hidden−states’).

The above options and modes are specified in a parameter file, depending on the desired action. The parameter file is then run in a command line to run the program. A template to specify the parameter file for a data set that contains 10 sequences of length 20 each with 3 real valued components is given in a script file in Figure 5.8.

5.7 Conclusions

In this chapter, the general form of HMM is introduced theoretically. The assumptions of the model along with its limitations are described. Then a brief description of the three basic problems of the HMM with solutions to each problem and extension to multiple sequences is given. To illustrate the implementation of a HMM for multiple observations, a description of an available toolbox for multiple sequence observations is provided. The choice of a particular distribution type for the emission probabilities, the model type and specification of the number of states depend on the data to be analyzed and will be described in the case studies.
Name of the parameter file: hmm_learn_gaussian
# Action: specify the action from the above options
action learn

# Type of the model: hmm or nhmm
model_type hmm

# Number of hidden states
num_states 3

# Emission distribution: Gaussian distribution with 2 components
Emission
gaussian 2

# Data file
INPUT FILE PATH/ INPUT DATA FILE NAME

# Number of real-valued vector components for the data
num_real_data_components 3

# Number of data sequences
num_data_sequences 10

# Length of each sequence
data_sequence_length 20

# Output file
OUTPUT FILE PATH/ OUTPUT DATA FILE NAME

# Number of random initializations
num_restarts 10

The command line to run this template is
mvnhmm <hmm_learn_gaussian>

Figure 5.8: A template for implementing the HMM training using MVN–HMM Toolbox.
Chapter 6

Case Study: Wind Patterns and Their Correlation with Ozone in Houston, Texas

6.1 Description of the Problem

A number of counties around Houston, Texas are in moderate non–attainment of the National Ambient Air Quality Standard (NAAQS) for 8–hr ozone (85 ppb). Though not considered in NAAQS, large population exposure to elevated ozone levels makes Houston a priority area. Results from the TexAQS 2000 field program indicate progress has been made in understanding the Houston ozone problem (Daum et al., 2002; Senff et al., 2002; Banta et al., 2002; Ryerson et al., 2003). Incomplete understanding of Houston ozone formation processes, however, impedes development of optimal regulatory strategies. The goal of this case study is to apply the cluster analysis techniques to study the problem of tropospheric ozone in Houston, Texas region.

Two main factors are attributed to the complexity and uniqueness of the ozone problem (TCEQ/TAD, 2002). First, intense (and variable) industrial VOC emissions from the Houston Ship Channel allow for increased (and variable) precursor reactivity, generally favoring increased rates of local ozone formation (Daum et al., 2003; Kleinman et al., 2002). Second, due to its proximity to both the Gulf of Mexico and Galveston Bay, the Houston region experiences a variety of sea breeze circulations (Darby et al., 2002). Thus, Houston experiences highly variable transport and dispersion of locally formed ozone and its precursors (Banta et al., 2005). Understanding the ozone response under all representative meteorological scenarios is required for robust evaluation of control strategy efficacy. Identification of
recurring surface air flow patterns for ozone exceedance days can aid this modeling process—vast numbers of historical observations can be reduced into meaningful episode classes.

NAAQS exceedances around Houston typically occur under favorable (and often recurring) emissions and/or meteorological scenarios. Intense, intermittent industrial emission events are known to trigger elevated Houston ozone levels; the timing (hour of day) for such releases greatly impacts whether or not an exceedance occurs (Nam et al., 2006). Regardless of whether or not industrial releases occur, meteorological processes strongly influence the spatial distribution of ozone. Elevated ozone levels worldwide are associated with anticyclonic conditions. These cells of high upper-atmospheric pressure result in reduced wind speeds (or stagnation), clear skies (increasing solar flux), elevated temperatures, and subsidence; all of these are conducive to ozone formation and/or buildup. Indeed, Davis et al., (1998) show that some of the highest Houston ozone levels result under the influences of migrating anticyclones.

Given favorable synoptic settings, meso scale flows including land breeze/sea breeze cycles contribute to the complexity of the Houston ozone problem. Sea breeze flows occur from water to land, perpendicular to the coast; they are typical during and after periods of high solar influx to the ground level, especially when the large-scale pressure gradient is weak. Bao et al., (2005) show the interaction between the large-scale flow and the sea breeze is essential for the development of ozone exceedances. Using wind back-trajectory analysis, Berkowitz et al., (2005) associate rapid Houston ozone buildup with the repeated passage (or stagnation) of an air parcel over a source area. Zhang et al., (2007) further discuss the transport and mixing of pollutants over Houston emission sources. Both the shifting horizontal wind directions and the shallow depth of the sea breeze flow (Miller et al., 2003) inhibit the dispersion of pollutants.

Statistical investigation of historical data can identify recurring meteorological characteristics associated with ozone exceedances. Cluster analysis (Everitt et al., 2001) is attractive because it requires no a priori knowledge of how the patterns manifest themselves in the data. Weber and Kaufmann (1995) perform cluster analysis on hourly $u$ and $v$ wind components from a network of surface monitors. Clustering these routine surface data readily indicates
a number of 1–hr wind field patterns; each day can further be described by its sequence of 24 hourly cluster labels. Darby (2005) performs a similar cluster analysis of hourly surface wind measurements for 27 summer days in Houston, 15 of which are exceedances of the previous 1–hr NAAQS for ozone (120 ppb). 8–hr sequences of these hourly flow patterns for each day corroborate the diurnal patterns identified previously in an intensive Doppler lidar field campaign; together, these labels help identify various air flow scenarios for Houston exceedances.

Grouping days having similar sequences of hourly flow patterns is naturally appropriate for explaining ozone levels which peak during the afternoon of each day. Here, we perform clustering of hourly wind field measurements like the above authors; then, a quantitative algorithm is used to group days sharing similar 18–hr sequences (0000–1700 CST) of the identified cluster labels. The study considers the entire 2004 ozone season (1 April to 31 October) for the Houston, Galveston, and Beaumont–Port Arthur, Texas domain. Clustering is performed for the entire ozone season, 5136 consecutive hourly samples, to identify representative surface flow patterns. Then, an automated sequencing method (Elzinga, 2003) quantitatively determines groups of days sharing similar 18–hr sequences of cluster labels. (8–hr sequences as considered qualitatively by Darby (2005) do not provide a sufficient basis of comparison between the days). The computationally intensive sequencing calculations are performed only for 32 exceedance days during the 2004 ozone season; the goal of the study is to identify different classes of diurnal air flows under which exceedances occur. Spatially distributed ozone and \( NO_x \) level measurements are then compared between the identified diurnal flow groups to infer transport and dispersion patterns.

In this chapter, clustering algorithms (Chapter 3) and Sequence analysis (Chapter 4) are used sequentially to determine the meteorological patterns that lead to high ozone mixing ratios in Houston, Texas. Section 6.2 gives a short synopsis of the study region and the data used for this study. Section 6.3 shows the application of the methods to the data. First, wind filed time series clustering (section 6.3.1) is applied using dynamic \( k\)-PCA described in section 3.6.1. This is followed by a section 6.3.2 that demonstrates the application of clustering at the hourly and daily time scales. Section 6.4 describes the physical significance
of the results obtained from cluster analysis. Then, a short analysis is performed in section 6.5 to determine the characteristics of ozone exceedances and non–exceedances under the 1–hr standard and the 8–hr standard proposed by EPA in 1997. Finally, a short description of the results found from this analysis are presented in section 6.6.

6.2 Description of Study Domain and Data

The Houston, Galveston and Beaumont–Port Arthur non–attainment area is part of a large ozone influence region comprising much of the southeastern United States. The study domain (Figure 6.1) is on the northwest coast of the Gulf of Mexico. Houston lies on the flat, low–lying Gulf Coastal Plain. Regional high humidity and temperatures (typically over 30° C) occur during the warm season, along with scattered and possibly sudden precipitation events. The Gulf Coast meteorology provides for a wide variety of surface flow patterns. Frequently, multiple, sequential air flow patterns are realized in a single day as the dominant wind direction shifts or even reverses. The Houston Ship Channel, a major source of precursor emissions, lies to the east and southeast of downtown Houston. This case study focuses on the period 1 April through 31 October of the year 2004, during which 32 8–hr ozone NAAQS exceedance days occur in the study domain. Henceforth, an “exceedance day” indicates that maximum 8–hr ozone level for at least 1 monitor in Figure 6.1 exceeds 84 ppb. Ultimately, each exceedance day will be assigned to a group with other days sharing similar 18–hr (0000–1700 CST) evolution of the surface wind field.

Surface meteorology and air quality measurements are provided by two separate monitoring networks operated by the Texas Commission for Environmental Quality (TCEQ). Hourly surface wind speed and direction data are available from a network of 28 monitoring stations (Figure 6.1). The wind data alone are used to establish the air flow patterns. Temperature data are obtained from a set of 40 surface meteorological monitors. Networks of 20 and 27 Continuous Air Monitoring Stations (CAMS) monitor 8–hr ozone and 1–hr \( NO_x \) levels, respectively. Positions of the monitoring stations are shown in Figure 6.1 and stations discussed explicitly in the text are listed in Table 6.1.

The statistical calculations cannot handle any missing wind measurements from any station at any hour of the day during the study period. The 28 meteorological stations
Figure 6.1: Map of Houston study domain showing 20 air quality and 28 surface meteorology monitors. Water is shown as dark gray and major urban areas are shown as light gray. Contour lines are shown every 20 m. Stations discussed explicitly in the text are labeled and described in Table 6.1.

Table 6.1: Selected surface monitors shown in Figure 6.1.

<table>
<thead>
<tr>
<th>Location</th>
<th>Index for Figure 6.1</th>
<th>Latitude (°N)</th>
<th>Longitude (°E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galveston</td>
<td>G</td>
<td>29.26</td>
<td>94.85</td>
</tr>
<tr>
<td>Houston, downtown</td>
<td>H</td>
<td>29.73</td>
<td>95.31</td>
</tr>
<tr>
<td>La Porte, south of Ship Channel</td>
<td>L</td>
<td>29.66</td>
<td>95.12</td>
</tr>
<tr>
<td>Texas City</td>
<td>T</td>
<td>29.38</td>
<td>94.93</td>
</tr>
</tbody>
</table>

are chosen from a larger set of 65 to maximize the temporal coverage for the study period while retaining adequate spatial coverage of the study domain. In total, approximately 5% of the observations from these 28 monitors are missing. Any 1–hr gaps for each monitor are interpolated linearly in time using observations from that same monitor. Larger gaps in the records are imputed using the method of Schneider (2001). The $u_p$ and $v_p$ wind components from all stations $p$ are considered as independent variables, and Expectation–Maximization
is applied to estimate the missing values. All subsequent analysis is performed on these imputed data containing no missing values.

6.3 Application of Methods

Cluster analysis is explored as the data–driven method to study the Houston, Texas region. The purpose of this analysis is to determine the wind patterns that affect the ozone formation in this region. For this, a time series clustering based on $k$–PCA is explored to determine the synoptic scale wind patterns. The results from this cluster analysis show that Houston, Texas region is influenced by land/sea breeze cycle that happen on a meso scale. Thus, an alternative clustering methodology based on clustering at two time scales is explored to study the meso scale climatology that affects the ozone mixing ratios. The application of the methods and the results from the analysis are described below.

6.3.1 Houston Wind Field Time Series Clustering

The dynamic $k$–PCA time series clustering as described in section 3.6 is applied to the hourly wind data sets from the years 2001 through 2004. Time series clustering can easily handle large data sets and thus a larger data comprising four years of data is used for this case study. To detrend the time series, mean diurnal cycle is removed from each of the 56 variables. The number of lags is then estimated for the imputed and detrended observations by means of PACF. The PACF is calculated for lags up to 30 to determine the dependency on previous day values. A lag value of 2 is chosen for $M$ to balance the trade–off between model complexity and accuracy. A proper value for the window length $L$ is determined by trial and error. The desired time scale for this analysis is synoptic scale, which is defined as a scale longer than 24 hr. Thus, window lengths of 24, 48 and 72 hr were tested with a constant $R$ (window sliding rate) of 12 hr. Window length of 48 was chosen to be analyzed as it isolated the patterns best. Thus, using $L = 48$ and $R = 12$, the observations are windowed as illustrated in Figure 3.3 to form subsets of imputed, scaled, detrended and lagged (starting at 0000 and 1200 CST) observations.

The original time series comprises of 20544 hourly observations (5136 hours corresponding to 214 days per year) and 56 variables ($u$ and $v$ wind components). This upon windowing
reduces to \( N = 1699 \) windows of observations with dimensions \( 48 \times 168 \). A DPCA model is formed for each window of time series observations to determine the number of principal components (PCs) and 35 PCs are noted to capture the maximum variability over the \( N \) windows. Next, to apply the hierarchical aggregation of \( k \)-means ensembles, the parameter \( k_{\text{max}} \) is determined. Trial values of \( k'_{\text{max}} \) ranging from 2 to 20 are used to generate 40 randomly initialized runs and the corresponding distance matrices are calculated. Figure 6.2 shows the sum of square errors for the trial values and it is observed that the \( \Delta \text{SSE} \) decreases with increasing \( k'_{\text{max}} \). A \( k_{\text{max}} \) value of 20 is chosen and hierarchical clustering with average linkage is used to produce the dendrogram in Figure 6.3.

![Figure 6.2](image)

Figure 6.2: The \( \Delta \text{SSE}(k'_{\text{max}}) \) comparing aggregated distance matrices \( D(k'_{\text{max}}) \) and \( D(k'_{\text{max}}+1) \).

The Cophenetic correlation coefficient is 0.84 indicating that the dendrogram provides an accurate representation of the distance matrix. Four main branches of the dendrogram are selected as the final cluster solution. For time series data, the final solution can be validated by trimming the windows. Patterns of shorter duration and high frequency events are trimmed as synoptic scale is the chosen scale. A solution with a small number of trimmed windows is consistent with the identification of synoptic patterns. Only 56 of the 1699 windows are trimmed, thus validating the cluster solution. Also, to check the number of PCs, prototype loading matrices are computed for supramatrices formed by collecting the
untrimmed windows assigned to a cluster $r$. Then, each supra matrix is projected on to the DPCA prototype. The percentage of variance of each cluster captured by the corresponding DPCA prototype is 90%. This confirms the choice of the number of PCs. The trimmed membership for the windows is then scaled down to the days assigned to each cluster as shown in Figure 6.4. A day belongs to a cluster only if all the 24 hour observations of the day belong to the trimmed cluster.

To interpret the clusters in terms of the physical mechanisms, cluster averaged wind vectors (Figures 6.5 through 6.7) are calculated at 0700, 1200 and 1800 CST for each station and are plotted on the geo–spatial coordinates.

The ultimate goal of this study was to detect synoptic patterns associated with air quality. Thus, average ozone compositions (Figure 6.8) were computed for the hours belonging to the wind clusters at each station and are plotted on the geo–spatial coordinates. These average values are calculated for only hours corresponding to the exceedance days only.
Figure 6.4: Cluster membership for summer days in the study period of 2001–2004; the position of asterisk on $y$–axis indicates the cluster membership for a day. Tick marks on the $y$–axis from bottom to top correspond to clusters 1–4 respectively. Vertical lines indicate the days that exceed the 1997 ozone NAAQS of 85 ppb.

Little or no correspondence was found between the average wind and ozone plots. Normalized Mutual Criterion (NMI) is computed to validate this observation (Figure 6.9). A NMI is a symmetric metric that quantifies the statistical information shared between a pair of clusters. It takes values ranging from 0 to 1 with values close to one indicating the best correspondence between clusters. The NMI for the wind and ozone clusters was found to be 0.06 indicating that there is no correspondence between the wind patterns and the ozone exceedance days.

Weather maps at the 500–hPa and 850–hPa pressure level are obtained from the National Centers for Environmental Prediction reanalysis data made available by the National Oceanic and Atmospheric Administration-Cooperative Institute for Research in Environmental Sciences Climate Diagnostics Center (http://www.cdc.noaa.gov/). The average daily weather maps (Figures 6.10 and 6.11) at these two different pressure levels are computed for 1800 UTC. At the 500–hPa level, not much distinction was found between the clusters. At the 850–hPa level, nearer the surface, clusters are distinguishable.
Figure 6.5: Daily wind velocity ($ms^{-1}$) at 0700 CST at each station, averaged among days assigned to each cluster. The stations correspond to those in Figure 6.1.

These weather maps show the importance of low altitude land/sea breeze cycle for Houston area, as opposed to synoptic conditions driven by upper atmosphere. Cluster #1 has lower pressure than others over Houston while cluster #2, compared with #3, has high pressure center “bending” over Houston, causing westerly deflection of the sea breeze by high pressure over the urban heat sink. Cluster #4 has high pressure center over Houston, responsible for mesoscale flow reversal. Thus, it is observed that in Houston, Texas synoptic climatology has little role and the mesoscale patterns indicated by the presence of land/sea breeze cycle could be the possible driving airflow patterns that lead to high ozone.

6.3.2 Hierarchical Clustering at Hourly and Daily Scales

To determine the mesoscale effects of wind field on ozone, the hierarchical clustering algorithm described in section 3.3 will be applied successively to matrices $D^h$ and $D^d$ quantifying
Figure 6.6: Daily wind velocity ($ms^{-1}$) at 1200 CST at each station, averaged among days assigned to each cluster. The stations correspond to those in Figure 6.1.

the degree of dissimilarity between hourly and daily sampling intervals, respectively. Methods for calculating the matrices are explained below. The following terminology is henceforth used to distinguish between the results of the clusterings performed at the hourly and daily time scales. Hierarchical clustering at the hourly time scale is said to identify “clusters” of hourly surface flow patterns; hierarchical clustering at the daily time scale is said to identify “groups” of days with similar diurnal cycles for the surface flow.

6.3.2.1 $k$–means Ensemble Aggregation for Hourly Flow Patterns

Hierarchical clustering is first performed at the hourly time scale to identify a set of hourly surface flow patterns. Matrix $D^h$ indicates the degree of dissimilarity between the surface wind fields for each pair of $N^h$ total hours in the study period. The dissimilarity calculations assume a Cartesian coordinate system—hourly surface wind speed and direction data are transformed into vector components $u_{ts}$ (westerly) and $v_{ts}$ (southerly). Subscript $t$ is the
Figure 6.7: Daily wind velocity \( (ms^{-1}) \) at 1800 CST at each station, averaged among days assigned to each cluster. The stations correspond to those in Figure 6.1.

hourly sample number starting at 1 at the beginning of the observation period (1 April 2004 0000 CST), and \( s \) is an index referencing the \( p \) wind monitors used in the study.

Matrix \( D^h \) is not calculated directly from the wind field measurements. A robust estimate for \( D^h \) is determined as the average degree of dissociation between pairs of hourly observation periods over a large ensemble of individual runs of the \( k \)–means clustering algorithm (Beaver and Palazoglu, 2006). The wind components for each station \( s \) are scaled using the mean wind speed \( s_s \) for that same station; this ensures each station is weighted equally in the analysis while preserving the directional relationships between the raw and scaled data.

The \( k \)–means algorithm is iterative and requires the user to provide initial estimates for \( k \), \( u_{ts} \), and \( v_{ts} \); it nearly always converges to a local minimum of the solution space. A
Figure 6.8: Daily maximum 8-hr averaged ozone (ppb) averaged among days assigned to clusters at each station. The stations correspond to those in Figure 6.1.

A large number of randomly initialized $k$-means runs are performed to ensure convergence of the dissimilarity matrix $D^h$. Its elements are calculated as the aggregate proportions of $k$-means runs for which pairs of hourly observations are assigned the same cluster label.

To determine the parameter $k_{\text{max}}$, trial values of $k'_{\text{max}}$ are tested incrementally starting from 2. For each value of $k'_{\text{max}}$, 100 runs of the $k$-means algorithm are performed. For higher values of $k'_{\text{max}}$, one or more of the clusters become empty occasionally and that particular solution is discarded. For each value of $k'_{\text{max}}$, an aggregated distance matrix is calculated using all the runs with $k \leq k'_{\text{max}}$. Sum of the squared errors is computed and it quantifies the difference between the $D^{\text{Avg}} (k'_{\text{max}})$ generated using successively incremented $k'_{\text{max}}$. The plot (Figure 6.12) of $\Delta$ SSE ($k'_{\text{max}}$) and $k'_{\text{max}}$ indicates convergence of the distance matrix for $k'_{\text{max}}$ of 9. However a $k_{\text{max}}$ value of 10 has been used as aggregating runs with larger $k$. 

80
Figure 6.9: Normalized mutual information (NMI) for the wind clusters and ozone clusters obtained from time series clustering.

Figure 6.10: Daily 500–hPa weather maps for clusters 1 (Top Left), 2 (Top Right), 3 (Bottom Left), and 4 (Bottom Right), averaged among days assigned to each cluster. The Houston study region of Figure 6.1 is highlighted in a rectangular box.
Figure 6.11: Daily 850–hPa weather maps for clusters 1 (Top Left), 2 (Top Right), 3 (Bottom Left), and 4 (Bottom Right), averaged among days assigned to each cluster. The Houston study region of Figure 6.1 is highlighted in a rectangular box.

will have a negligible effect on the aggregated distance matrix and on the connectivity of the resulting dendrogram.

Figure 6.12: \( \Delta SSE(k'_{max}) \) comparing aggregated distance matrices \( D^{Avg}(k'_{max}) \) and \( D^{Avg}(k'_{max}+1) \).
The converged distance matrix obtained using a $k_{\text{max}}$ of 10 and 900 individual $k$-means runs are subjected to the hierarchical clustering of section 3.3 to generate the dendrogram in Figure 6.13. 10 clusters of hourly wind observations are selected such that each represents a distinct surface flow pattern. The mean wind field associated with a cluster is calculated as the average of the $u_{tp}$ and $v_{tp}$ wind components at each station $p$ for each hourly sample $t$ assigned to that cluster (Figure 6.14). Clusters #1 – #4 have winds predominantly from the south to southeast. Cluster #5 represents stagnant conditions. Cluster #6 has winds from the northwest; #7 from the northeast; #8 from the northwest (offshore flow); and #9 and #10 from the east. In addition to stagnant pattern #5, clusters #6 and #10 have lower than average wind speeds; patterns #1 and #2 have the highest wind speeds present as strong onshore flow. Many of the cluster patterns indicate wind directions along Galveston Bay and Sabine Lake that deviate from the regional flow pattern; these monitors capture local bay and lake breeze circulations which may not be present on all days.

Figure 6.13: Dendrogram for hierarchical clustering of dissimilarity matrix $D^h$ for 5136 hourly wind field observations. Individual hours are not shown across bottom of dendrogram due to space limitations. Vertical lines indicate 10 clusters (#1–10) selected from the hierarchy.
Diurnal distribution plots (Figure 6.14) indicate that many of the clusters are preferentially realized during certain phases of the diurnal cycle. Clusters #1, #2, and #9 tend to occur during the daylight hours and into the evening. These 3 clusters exhibit onshore flows with different directions and timings—#1 and #9 are more prevalent during the morning than #2. The daytime, regional-scale shoreward flows for these 3 clusters may capture Gulf Breeze influences. Clusters #3 and #5 occur during nighttime and into the morning hours. Cluster #6 occurs most frequently during the early morning, while #4 and #10 are infrequent and generally appear in the evening. Clusters #7 and #8 capture daytime flows opposing the Gulf Breeze patterns (#1, #2, and #9); the former pair of patterns is less biased toward the daylight hours than the three Gulf Breeze patterns. The clustering algorithm does not consider the measurement sampling times; the emergence of these diurnal biases in the hourly cluster labels validates their correspondence to real wind field patterns.

Figure 6.14: Hourly wind field patterns and diurnal distributions for 10 clusters of Figure 6.13. Cluster number and symbol are located in lower right corner of each map. Length of arrow indicates mean 1–hr wind speed as indicated on scale.

To capture effects of the diurnal flow field relevant to ozone buildup, the clusters are deliberately selected to have different levels of specificity. The overnight clusters #3 and #5 together account for 37% of the 5136 hours in the study period. Only 2 clusters are needed
to capture the bulk of the nighttime conditions; subtle variations in the nighttime flows do not strongly affect the daily maximum ozone levels resulting later during the afternoon due to rapid photochemical production. The typical afternoon Gulf Breeze clusters #1, #2, and #9, on the other hand, capture more subtle variations in the wind field (southerly, southeasterly and easterly respectively); subtle differences in the timing associated with these southerly flows are important for explaining daily maximum ozone levels. Other morning and/or daytime clusters such as #6, #7 and #8 are relatively infrequent; they are retained as distinct flow patterns, however, because they are relevant for explaining variability in ozone levels. The daytime flow patterns are purposely selected to be more highly specific (less general) than the nighttime patterns. Increased specificity (splitting of clusters) for the nighttime flow patterns would contribute little to explaining the daytime ozone levels. Increased generality (merging of clusters) for the daytime flow patterns, however, would adversely affect their subsequent sequencing to isolate diurnal flow patterns favoring elevated ozone levels.

The 10 clusters vary widely in their proportions of hours which fall on an exceedance day (any hour from 0000–2300 CST). Hourly sampling intervals for clusters #1–#10 occur on exceedance days with frequencies 2%, 2%, 8%, 25%, 28%, 40%, 18%, 9%, 19%, and 23%, respectively. Hourly patterns #5 and #6, with the lowest wind speeds, occur most frequently on exceedance days—roughly 40% and 30% of the time, respectively. Clusters #4, #7, #9, and #10 are less indicative of episodic conditions, with approximately 20% of these hours falling on exceedance days. Patterns #1 and #2, with the highest wind speeds and flows arriving from the clean Gulf marine layer, occur on exceedance days just 2% of the time. The presence of these strong southerly flows generally indicates that an exceedance will not occur. Patterns #3 and #8 are also moderately disassociated with exceedance days. Differences in their levels of association with exceedance days further validate the physical correspondence of the hourly cluster labels to real wind field patterns; still, the hourly patterns alone do not suffice to isolate episodic conditions or differentiate between exceedance scenarios.
6.3.2.1.1 Cluster Validation

The validity of the wind clusters is obtained by deleting a small number of observations and performing cluster analysis on the remaining observations. The data set is split in such a way that the first 80 percent of the surface hourly wind observations fall into data set one and the rest of 20 percent belong to data set two. Data set one consists of 489, 514, 925, 42, 652, 340, 544, 208, 318 and 72 hourly observations belonging to hourly wind clusters #1–#10 from Figure 6.13. Observations belonging to data set two are deleted and cluster analysis is performed on data set one. Figure 6.15 shows the dendrogram obtained after performing the hierarchical cluster analysis on data set one.

Figure 6.15: Dendrogram for hierarchical clustering of dissimilarity matrix \( D^h \) for 4109 hourly wind field observations. Individual hours are not shown across bottom of dendrogram due to space limitations. Vertical lines indicate the 9 clusters (A–I) selected from the hierarchy.

Ten clusters labeled A to J are selected from this dendrogram. The memberships of the cluster labels A – J obtained are compared to the cluster labels 1–10. Table 6.2 indicates the membership assignment of the cluster labels for the two sets of analysis. 98%, 86%, 71%, 79%, 76%, 82%, 62% 95%, 60%, 88% of observations belonging to cluster #1–#10 are
assigned to clusters H, I, F, G, A, B, C, E, C, D respectively. Thus, deletion of small number of observations does not alter the cluster assignment greatly establishing the stability of the cluster analysis.

Table 6.2: Cross-validation by comparing the occurrence of a cluster from figure 6.13 in figure 6.15.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>498</td>
<td>29</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>67</td>
<td>277</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>339</td>
<td>0</td>
<td>190</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>112</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>205</td>
<td>197</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>4</td>
<td>655</td>
<td>0</td>
<td>75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>55</td>
<td>247</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>H</td>
<td>479</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>441</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>489</td>
<td>514</td>
<td>925</td>
<td>42</td>
<td>652</td>
<td>340</td>
<td>544</td>
<td>208</td>
<td>318</td>
<td>72</td>
</tr>
<tr>
<td>Percentage</td>
<td>98</td>
<td>86</td>
<td>71</td>
<td>79</td>
<td>76</td>
<td>82</td>
<td>62</td>
<td>95</td>
<td>60</td>
<td>88</td>
</tr>
</tbody>
</table>

6.3.2.2 Daily Wind Field Groups

Hierarchical clustering is next performed at the daily time scale to identify groups of days having similar diurnal cycles for the surface wind field. Sequences of chronologically ordered hourly wind field pattern labels $l_t$ obtained from hourly wind field clustering are used to calculate $D^d$. For each day $d$, labeled serially from 1 to $N_d$ days, the 0000–1700 CST sequence of 18 hourly wind field pattern labels are stacked into the vector $\lambda_d = [l_{24d-23}, l_{24d-22},...,l_{24d-6}]$. This transformation between vector $\lambda_d$ and scalar $l_t$ merely relates each consecutively numbered day $d$ to its 18 sequential 0000–1700 CST hourly cluster labels, such that $N_d = N_h/24$.

Using the principles of sequence analysis, the sequence similarity index $S_{ij}$ is computed between all pairs of days $i$ and $j$. The relative degree of similarity between pairs of 18–hr sequences $\lambda_i$ and $\lambda_j$ is determined using the concept of precedence. Days are not required to exhibit the same surface flow patterns at the same hour of the day to be considered as similar; rather, days are regarded as similar if they share the same temporal order in which their hourly cluster labels are realized. A computationally efficient, iterative method for computing $S_{ij}$ given by Elzinga (2003) is summarized in section 4.4.1. Combinatorial optimization of the discrete sequences of cluster labels is considerably less complex than that for the continuous
wind field measurements. Also, Equation 6.1 scales poorly with sequence length (here 18 hours) and is in general computationally burdensome. Thus, robust estimation of $S_{ij}$ does not necessitate an ensemble averaging technique, and $D^d$ is calculated directly.

$$D^d_{ij} = 1 - S_{ij} \quad (6.1)$$

Matrix $D^d$ is then input to the hierarchical clustering algorithm, and the user selects $g$ groups of days from the resulting dendrogram. Input data consist of 18-hr sequences of hourly cluster labels $\lambda_d$ for each day $d$; the output is a set of $g$ different labels to which the days are assigned. Each day is assigned to exactly 1 group: days assigned to the same group share similar diurnal cycles for the surface wind field. The daily sequencing calculations are strongly dependent on the hourly wind field patterns defined by the previous clustering at the hourly time scale; as such, the hourly flow patterns must be selected judiciously to capture physically relevant events at appropriate levels of specificity.

Hierarchical clustering is used to generate the dendrogram in Figure 6.16. Four main groups (E1–E4) are selected from the hierarchy having 7, 12, 10, and 3 days. The sequences of hourly wind field patterns associated with these 4 groups of daily air flow patterns are indicated in Figure 6.17. Exceedance group E1 has stagnant wind patterns that shift to southerly winds later in the day. Days assigned to E2 have northwesterly flows in the morning followed by either southerly or stagnation flow. Group E3 has an easterly flow component for most hours of the day, with winds gradually shifting from northeasterly to southeasterly during the afternoon. E4 experiences northwesterly wind field patterns in the morning that shift to northeasterly and easterly flows in the afternoon.

Figure 6.17 indicates the general directionality and timing of the directional shifts for each exceedance day. Physical interpretation of the diurnal patterns, however, requires investigation of actual hourly wind measurement. Figure 6.18 shows time series of hourly wind speed and direction at 3 selected locations transecting the Houston area. Texas City and LaPorte are sited along bodies of water and may experience local thermally driven
Figure 6.16: Dendrogram for hierarchical clustering based on Sequence Similarity Index for 32 exceedance days. Vertical lines indicate 4 groups (E1–E4) selected from the hierarchy. Circulations; the monitor in downtown Houston is farther inland and may be affected by the stronger local flows penetrating from the coastline. For a given monitor, the diurnal wind fields are relatively consistent within each group, yet distinct differences appear between the exceedance groups. Most exceedance days experience shifts toward southerly flow directions, however the timing of these shifts and associated wind speeds differentiate the exceedance group patterns.

The dendrogram of Figure 6.16 can be interpreted at 2 levels of resolution, starting at the lowest possible resolution (top of the hierarchy). The left (E1 & E2) and the right (E3 & E4) main branches of the hierarchy are distinguished by the predominant flow direction, excluding the common southerly flows. The left branch indicates 19 days with westerly flows, whereas the right branch indicates 13 days with easterly flows. Each of these 2 main branches was bisected to yield the 4–group solution: these subclusters are distinguished by the timing of the southerly shifts and the afternoon wind speeds, as observed at Texas City where the diurnal wind field is least variable. E1 & E4 experience sudden changes in wind direction, whereas E2 and E3 have slowly shifting winds throughout the day. Selection of
additional exceedance groups was explored; increasing the specificity of the groups, however, does not appear to distinguish meaningful modes of variability in the diurnal wind field.

### 6.3.2.3 Non–Exceedance Day Groups

In a procedure analogous to that performed for the 32 exceedance days, the sequence analysis is used to generate a dendrogram describing the similarity relationships between the 182 nonexceedance days (Figure 6.19). The nonexceedance days are both larger in number and exhibit a wider variety of meteorological conditions than the exceedance days. As such, 8 main groups are selected (N1–N8) in addition to the lone outlier 26 June. Sizes for the main groups are 49, 24, 15, 30, 15, 8, 25 and 15 days.

The sequences of hourly wind field patterns associated with these 8 groups of daily meteorological patterns are shown from Figure 6.20 to Figure 6.27. Groups N3 to N7 experience nearly constant flow patterns throughout the day–predominant flows are southerly or southeasterly flows. N2, on the other hand, typically experience shifts in direction at some time during the day. N2 experiences southerly morning winds that follow up through the
Figure 6.18: Wind speed and direction at 3 locations: Texas City, LaPorte, and downtown Houston. Each column corresponds to an exceedance group (E1-E4), and each row indicates 1-h wind speed or direction at a specific monitor. Hourly time series are superimposed for 0000-2300 CST of all days assigned to each exceedance group. The single outlier day 15 June 2004 (group E2) is plotted with dots indicating each hourly observation.

day. Days of N2 that experience easterly or northeasterly flows in the mornings shift in direction to southerly flow during the course of the day. N8 typically experiences a shift from northeasterly to southeasterly flows. N1 is distinguished from the other nonexceedance groups by the presence of various flow patterns in the mornings of various days with a shift to Gulf Breeze activity in the late afternoon and on a few days following a brief period of stagnation.

6.4 Discussion on Diurnal Air Flow Patterns and Air Quality Response

The air quality responses for the identified exceedance and non-exceedance day groups are assayed using the group–averaged ozone and $NO_x$ levels at each monitoring station. The ozone response is shown as the mean daily maximum 8-hr ozone levels, while $NO_x$ levels are characterized by the average peak NO level occurring before noon of each day.
Figure 6.19: Dendrogram for hierarchical clustering based on Sequence Similarity Index for 182 nonexceedance days. Individual days are not shown across bottom of dendrogram due to space limitations. Vertical lines indicate 8 groups (N1–N8) selected from the hierarchy. An outlier day appearing between N4 and N5 remains unassigned.

Figure 6.20: Sequences of hourly wind field patterns for nonexceedance days group N1 from 2004 ozone season indicated in Figure 6.19.
Figure 6.21: Sequences of hourly wind field patterns for nonexceedance days group N2 from 2004 ozone season indicated in Figure 6.19.

Figure 6.22: Sequences of hourly wind field patterns for nonexceedance days group N3 from 2004 ozone season indicated in Figure 6.19.
Figure 6.23: Sequences of hourly wind field patterns for nonexceedance days group N4 from 2004 ozone season indicated in Figure 6.19.

Figure 6.24: Sequences of hourly wind field patterns for nonexceedance days group N5 from 2004 ozone season indicated in Figure 6.19.
Figure 6.25: Sequences of hourly wind field patterns for nonexceedance days group N6 from 2004 ozone season indicated in Figure 6.19.

Figure 6.26: Sequences of hourly wind field patterns for nonexceedance days group N7 from 2004 ozone season indicated in Figure 6.19.
Figure 6.27: Sequences of hourly wind field patterns for nonexceedance days group N8 from 2004 ozone season indicated in Figure 6.19

6.4.1 Exceedance Groups

Group E1 has the highest ozone levels (Figure 6.28) on the northern edges of Houston and Beaumont. Ozone levels are higher than average throughout Beaumont and Port Arthur; 5 of these 7 days have exceedances in here. This group also experiences high $NO_x$ (Figure 6.29) levels to the northwest of Houston and near the city center. The near-stagnant, northerly, morning winds followed by moderate southerly flows allow pollutant buildup towards the northern edge of the study domain.

E2 captures source–receptor relationships in which Galveston ozone levels are much higher than average. E2 contains 6 of the 9 Galveston exceedances for 2004; each other group contains a single Galveston exceedance day. A gradient exists across Houston such that ozone levels decrease from east to west. Most days (10 of 12) for E2 experience exceedances nearby the Ship Channel and/or downtown Houston. Though below the exceedance threshold, Beaumont ozone levels are also highest toward its southern side. $NO_x$ levels are elevated near downtown Houston. This pattern experiences more northerly flows than the other groups; ozone levels are elevated south of the major source areas.
E3 experiences easterly flow with ozone levels increasing from east to west across both Beaumont and Houston. Most days from the other groups exhibit peak $NO_x$ levels near downtown Houston; group E3, however, often exhibits peak morning NO levels northwest of Houston. E3 has the lowest NO levels near the Ship Channel and ozone levels increase from east to west across Houston. Exceedances are largely confined to downtown Houston and northwest of the city center. Beaumont and Port Arthur ozone levels are also elevated, but usually not to the exceedance level. The easterly winds present throughout these days allow ozone buildup toward the west sides of Houston, the Ship Channel, and Beaumont-Port Arthur.

E4 experiences exceedances and elevated $NO_x$ levels in downtown Houston alone. This pattern is similar to E3, however air quality is generally improved. With only 3 days, this smallest exceedance group may not capture a representative exceedance scenario.
The daytime shifting of wind directions towards southerly capture the bay breeze and Gulf Breeze influences around Houston. E1 has a sudden counterclockwise shift in wind direction before noon from westerly to southerly, at which point the wind speed picks up. This sudden shift in wind direction implies a thermally induced, local bay breeze flow; likely the largescale influence on the local surface winds is relatively weak. Morning winds are northwesterly in LaPorte and Texas City, but more southerly in Houston; discrepancies in these flow directions may indicate a local land breeze forming along the bodies of water for E1. The other groups exhibit more consistent wind measurements at the inland and coastal monitors. Relatively high afternoon wind speeds (\(\sim 3 \text{ m/s}\)) along the bodies of water for E1 ventilate the coastal locations—pollutant buildup occurs to the north.

For E2 the winds are constantly shifting clockwise throughout the day; winds from nearly all directions are experienced over these 24–hr periods. Wind speeds are relatively constant throughout most of the day for E2. Several days experience increasing winds after 1800 CST; these evening gusts, however, are likely irrelevant for explaining these exceedances. The gradual shifting of the winds for E2 toward southerly afternoon flows appears to capture regional Gulf Breeze influences; large-scale conditions likely influence these ozone episodes more than for E1. Relatively low afternoon wind speeds compared to the other groups allows ozone to build up near major source areas.

Like E2, E3 also has gradual, clockwise shifting directions throughout the day; unlike E2, however, wind speeds for E3 tend to pick up as directions transition toward southerly. E3 appears to have an even stronger large-scale influence than E2, as the wind directions at the inland Houston monitors are quite similar to those nearby bodies of water. The persistent easterly flows for E3 (and also E4) are most relevant to explaining the spatial distribution of pollutants for these days.

Beyond examining surface wind data, temperature measurements were also used to characterize the exceedance groups. No consistent temperature response was found. Temperatures were variable within each group, and ozone levels do not correlate with local temperature. The lack of temperature responses for the exceedance groups indicates that they do not
truly capture meteorological regimes; still, the consistent pollutant spatial distributions observed for each group confirm that they do capture meaningful transport and dispersion patterns. This analysis has focused on regional surface conditions, however upper–atmospheric and large–scale meteorological conditions also have significant impacts on Houston ozone levels. For 2004, all but the single exceedance day 15 June occur under strongly anticyclonic settings. On 15 June, a cyclone is present directly over Houston—the daily temperature range is small, indicating cloudy conditions, and minimal precipitation is observed. This outlier day experiences 8–hr ozone level of 89 ppb at the LaPorte monitor just south of the Ship Channel, however the next–highest ozone level is only 62 ppb. As indicated in Figure 6.18, this outlier day has a different diurnal wind field than other days in group E2 (or any group)—afternoon winds shift to westerly for several hours. This exceedance occurring under cyclonic conditions is likely anomalous; the localized ozone buildup near the Ship Channel suggests that this exceedance may result from an industrial emissions event.

The sequencing was also performed using 24–hr sequences of cluster labels on 0000–2300 CST to provide a larger basis of comparison between each day (24 versus 18 cluster labels). The resulting diurnal flow patterns, however, were less clear and the response of pollutant spatial distribution was weaker. Inclusion of the 1800–2300 CST cluster labels in the sequencing degraded the performance of the algorithm to identify physically meaningful diurnal flow patterns affecting ozone levels. Thus, it is concluded that the 1800–2300 CST wind field is not directly relevant to describing pollution events occurring previously in the day. Use of 24–hr sequences starting 1800 CST of the previous day (1800–1700 CST) did not appear to improve the results either. Thus, air flow patterns during the previous evening may not be relevant to Houston ozone exceedances due to the rapid daytime ozone formation.

6.4.2 Non–Exceedance Day Groups

Group–averaged plots are formed for ozone level (Figure 6.30) and peak morning NO (Figure 6.31) at each monitor for the 8 main nonexceedance groups (N1–N8). There is an outlier day between N4 and N5 that is not assigned to any group; Because of the spurious conditions associated with such nonexceedance outliers, they will not be considered further.
Of the nonexceedance groups, N1 and N3 have relatively low NO levels in downtown Houston and very low ozone levels (though ozone levels are lower than for any of the exceedance days). These groups experience a pocket of high surface pressure over the Gulf of Mexico, driving strong marine winds toward Houston throughout the day. Despite the fact that the daytime flows arrive from offshore, little sea breeze development is present; the
cooling effect of the marine layer prevents the land–sea temperature gradient from becoming large enough to induce a sea breeze. Thus, the sea breeze recirculation mechanism does not contribute to ozone buildup around Houston for these conditions.

N2, N4, N6 and N8 have strong southerly flow components for most hours of the day, however, indicating that marine ventilation suppresses ozone buildup. N4 has south to southeasterly winds throughout the day, resulting in peak ozone levels northwest of Houston. These set of nonexceedance clusters exhibit moderate ozone levels of the nonexceedance groups. The 18–hr sequences of wind field clusters for N2 are southerly winds or easterly winds in the morning that shift to southerly in the afternoon. The southerly winds explain the high levels of NO in the northern regions of Houston indicating that marine ventilation suppresses ozone buildup leading to moderate ozone levels.

Ozone levels for N5 and N7 are the highest of all the nonexceedance groups. This pair of groups have ozone conditions similar to those of E3 and E4 respectively i.e. the ozone levels are highest in the interiors of Houston. These groups typically exhibits a ridge of high pressure extending into the northern Atlantic; N7 typically exhibits a coastal trough; however the trough is deeper than for E4 and is often associated with a deep polar low pressure center over the Great Lakes. Conditions for N5 and N7 compared to their respective exceedance groups (E3 and E4) have slightly higher wind speeds because of the reduced intensity of their anticyclonic systems.

Thus, ozone levels for N2, N4, N6 and N8 are higher than for N1 and N3 (with strong ventilation and no sea breeze development) but lower than for N5 and N7 (with moderate sea breeze development).

6.5 Characteristics of Ozone Exceedances under the 1997 EPA Standard

EPA has revised the standard for ozone concentrations from 1–hr standard of 125 ppb to 8–hr running average of 84 ppb in 1997. This is because longer exposures even to low concentrations of ozone are more detrimental to human health than the one hour exposures.
Effective 31 October, 2008, the Houston–Galveston–Brazoria (HGB) area of the study domain was reclassified as severe nonattainment from previously classified status of moderate nonattainment area under the 1997 ozone eight-hour standard. The attainment date is set as expeditiously as practicable but no later than June 15, 2019. This case study aims at studying the characteristics of tropospheric exceedances under the 1hr and the 8hr standard in the study domain.

### 6.5.1 Exceedance Frequencies

Daily maximum ozone concentrations are analyzed for the summer seasons of 2001 through 2004 at the 20 ozone monitoring stations of the study domain. The 8–hr running average ozone standard results in 145 exceedances while the 1–hr standard results in 78 exceedances only. Monthly frequencies of exceedances under the two standards are shown in Figure 6.32. The month of August has the most number of exceedances under the one hour standard while under the eight–hour standard the exceedances continue to increase and reach the maximum in the month of September. The number of exceedances under the old and the new standard for the ozone monitoring sites are shown in Table 6.3. The exceedances increase in the range of 40–100 across the sites under the new standard.

![Figure 6.32: Frequency of occurrence of exceedances under the 1–hr and the 8–hr ozone standard during the years 2001 through 2004 in Houston, Texas.](image-url)
Table 6.3: Number of 1–hr and 8–hr exceedances during the years 2001 through 2004 in Houston, Texas.

<table>
<thead>
<tr>
<th>Station</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>125 ppb</td>
<td>3</td>
<td>16</td>
<td>13</td>
<td>8</td>
<td>11</td>
<td>17</td>
<td>23</td>
<td>13</td>
<td>11</td>
<td>14</td>
<td>20</td>
<td>17</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>84 ppb</td>
<td>30</td>
<td>41</td>
<td>40</td>
<td>24</td>
<td>10</td>
<td>46</td>
<td>66</td>
<td>26</td>
<td>25</td>
<td>27</td>
<td>35</td>
<td>28</td>
<td>46</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>34</td>
<td>14</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

6.5.2 Relationships between 1–hr and 8–hr Ozone Concentrations

Regression analysis is used to identify the relationships between peak one hour and peak eight–hour concentrations (Figure 6.33) at each site. This analysis is performed to relate the new standard to the more familiar old standard easily. The plot indicates that one hour concentrations of the range of 90–100 ppb produce an eight–hour exceedance under the eight hour standard.

![Figure 6.33: Relationship between the 1hr and 8hr ozone concentration levels in ppb during the years 2001 through 2004 in Houston, Texas.](image)

6.6 Conclusions

This case study demonstrates the utility of clustering algorithms and sequence algorithms. A final summary of the case study is provided below.

- Time series clustering of wind field observations was performed. The results from this study indicate the presence of mesoscale patterns and no synoptic scale climatology in Houston, Texas.
• Statistical identification of hourly and daily surface flow patterns is performed for an extended record of surface winds around Houston. Recurring meso scale meteorological scenarios have a strong influence on regional ozone levels.

• Exceedances tend to occur under anticyclonic conditions in which strong sea breeze activity contributes to the recirculation of pollutants around Houston. Resulting transport and dispersion patterns under such conditions are a superposition of the prevailing synoptic motions and shifting flows associated with the sea breeze. These different meteorological scenarios favor ozone exceedances that are spatially localized in distinct portions of the study domain. Temperature does not appear to directly correlate with observed ozone levels; mesoscale transport and dispersion mechanisms appear to dominate Houston ozone buildup processes. Anomalous meteorological conditions can occasionally result in exceedances as well.

• Nonexceedance days exhibit a wider variety of meteorological conditions than the exceedance days and generally lack strong sea breeze development.

• The case study of characterizing tropospheric ozone exceedances in the Houston area under the EPA–proposed eight–hour running averaged ozone concentrations of 85 ppb identifies the features of the tropospheric ozone of the study domain with respect to the one–hour standard of 120 ppb. Results from the analysis indicate that the frequency of exceedances increase under the new eight–hour ozone standard in comparison to the exceedances of the one–hour standard.
Chapter 7

Case Study: Sequence Analysis for Relating Wind Patterns with Ozone in Baton Rouge, Louisiana

7.1 Introduction

In this chapter, cluster analysis and sequence analysis are applied to determine the transport and dispersion patterns in Baton Rouge, Louisiana. In the previous study in chapter 6, these methods are applied separately on the 32 exceedances and the 182 non-exceedances from the summer ozone season of 2004 in Houston, Texas region separately. This analysis was performed separately on two groups of days to determine the patterns leading to exceedance and non-exceedance events, respectively. In this case study, all the days from the summer ozone season of 2005 from Baton Rouge, Louisiana are analyzed. The goal of this study is to determine if the exceedances preferentially realize within a few clusters or end up in individual clusters of their own.

Section 7.2 provides a background on the studies conducted till date in the study region and discusses their merits. Section 7.3 gives a brief description of the study domain and the data used in this case study. Section 7.4 provides the methodology used to apply the clustering and sequencing algorithms on the data. Section 7.5 discusses the clusters and groups obtained from the cluster and sequence analysis. Section 7.6 provides a physical interpretation for the relationship between the clusters obtained and the ozone levels in the region. Finally, section 7.7 summarizes the results obtained from this analysis.
7.2 Background

Tropospheric ozone is one of the criteria pollutants with increasing concerns in many urban areas in the United States. Baton Rouge, Louisiana is one such metropolitan area with ozone levels exceeding the established federal standards. Baton Rouge experienced an upward trend in ozone concentrations from 1995 to 2001 with a significant drop in 2003 followed by a slow recovery (Klasinc et al., 2008). This trend is due to steady increase in $NO_x$ concentrations during this period (Sather, 2003). Effective June 2003, Baton Rouge area was bumped up from the serious category to a severe ozone non-attainment category. Exceedances of the ozone standard in this region occur between April and October and are associated with weak or stagnant synoptic winds, weak pressure gradients, surface anticyclonic activity and advection from the southeastern Texas urban–industrial corridor (Rohli et al., 2004).

Many researchers studied the tropospheric ozone and air quality in Louisiana. Rohli et al., (2003) investigated quantitative models to predict ozone mixing ratios in Baton Rouge. Muller et al., (1985) employed typical thermal and wind properties to characterize a synoptic weather-type system for Shreveport, Louisiana to determine the air quality potential in terms of mixing heights and horizontal dispersion. Rohli et al., (2004) examine the relationship between $O_3$ concentrations in Louisiana and surface and low synoptic circulation and conclude that local conditions along with synoptic influences are important to determine the observed ozone behavior in Louisiana. Specific and closer inspection of the distinguishing meteorological patterns related to high ozone mixing ratios, however, has not been investigated. This study aims at identifying the relationship between tropospheric ozone concentrations and meso scale weather patterns in Baton Rouge, Louisiana.

This case study considers the 2005 ozone season of 1 May through 30 September comprising of 153 days for the Baton Rouge area. Similar to the previous study in Chapter 6, this case study uses a similar approach for the Baton Rouge data, Louisiana region. Cluster analysis is first applied to hourly wind measurements measured across 7 monitoring stations around Baton Rouge, Louisiana to determine the number of surface flow patterns, and then a quantitative sequencing technique is applied to generate groups of days sharing similar 24 hr sequences of hourly flow patterns (0000–2300 CST) corresponding to the 153 days of 2005.
In this study, we use the correlation distance metric in the first stage of clustering hourly surface wind field data and a metric based on the longest length of common subsequence for the second stage of clustering at the daily time scale for the 153 days.

### 7.3 Study Region and Data

This case study focuses on the period 1 May through 30 September of the year 2005, during which 25 8-hr ozone NAAQS exceedance days occurred in the study domain. Any day with maximum 8-hr ozone level for at least 1 monitor in Figure 7.1 exceeding 84 ppb is an “exceedance day”. The data, surface meteorological and air quality measurements, analyzed in this study are obtained from the Louisiana Department of Environmental Quality (LDEQ) website (http://www.deq.louisiana.gov/portal/tabid/2420/Default.aspx). The air analysis section of LDEQ maintains a statewide monitoring network to measure, analyze and validate the ambient air pollutant concentrations and meteorological data. Hourly surface wind speed and direction data are available from a network of 7 monitoring stations (Figure 7.1). These stations are located along the industrial corridor of Louisiana. The industrial area of the lower Mississippi River of South Louisiana has been termed the Industrial Corridor due to the high concentration of oil refining and petrochemical facilities in this region (Sai et al., 2004).

The data includes 8-hr ozone and 1-hr $NO_x$ measurements from networks of 7 ambient air monitoring stations. The raw wind speed and direction data are transformed into $u$ (northerly) and $v$ (easterly) components. This data for the 153 days (with 24 hours each) and 7 wind monitors (with 2 wind components each) are stacked chronologically into a matrix $X$ of dimension $3672 \times 14$. To handle missing operations, 1-hr gaps for a monitoring station are linearly interpolated using observations from the same monitor. Any larger gaps, if present, are imputed by the method of Schneider (2001). This method assumes a multivariate normal distribution among the monitoring stations and uses Expectation–Maximization to fill the missing values. The wind components for each station are scaled using the mean wind speed for that same station to ensure each station is weighted equally in the analysis.
Figure 7.1: Map of Baton Rouge study domain showing 7 air quality and surface meteorology monitors. Water is shown as dark grey and the station labels are given in the legend.

while preserving the directional relationships between the raw and scaled data. This scaled and imputed data matrix \( \mathbf{X} \) will be input to the clustering algorithm to generate the hourly wind field patterns that are ultimately sequenced to form the daily meteorological labels.

Additionally, weather data at the 500–hPa geopotential height are obtained from the National Centers for Environmental Prediction reanalysis data made available by the National Oceanic and Atmospheric Administration–Cooperative Institute for Research in Environmental Sciences Climate Diagnostics Center (http://www.cdc.noaa.gov/). The data consists of individual observations of pressure level at the 500–hPa height measured 4 times daily at regular intervals. This data is used to determine the relationship between synoptic scale weather patterns and ozone mixing ratios in the study region. The spatial domain is bound by 15° – 45 North° and 255°– 300° East. This resulted in the extraction of 1460 data points for the year 2005. The study region is bounded such that the pressure field is centered on Louisiana.
7.4 Application of Methods

7.4.1 Cluster Analysis

The application of aggregated $k$–means algorithm of section 3.5 begins not with the $u$ and $v$ components of the hourly averaged continuous wind field measurements $X$ but with the distance matrix $D$. The goal of this clustering is to label the hours and determine the meteorological regimes that affect regional air quality. Matrix $D$ is obtained by computing the correlation distances among the observations.

The first step in the implementation of aggregated $k$–means algorithm is to determine the parameter $k_{\text{max}}$. For this, trial values of $k'_{\text{max}}$ are tested incrementally starting from 2. For each value of $k'_{\text{max}}$, 200 runs of the $k$–means algorithm are performed. For higher values of $k'_{\text{max}}$, one or more of the clusters become empty occasionally and that particular solution is discarded. For each value of $k'_{\text{max}}$, an aggregated distance matrix is calculated using all the runs with $k \leq k'_{\text{max}}$. Sum of the squared errors is computed and it quantifies the difference between the $D_{\text{Avg}}(k'_{\text{max}})$ generated using successively incremented $k'_{\text{max}}$. The plot (Figure 7.2) of $\Delta \text{SSE}(k'_{\text{max}})$ and $k'_{\text{max}}$ indicates the convergence of the distance matrix for $k'_{\text{max}}$ of 9. However, $k_{\text{max}}$ value of 10 has been used as aggregating runs while larger $k$ will have a negligible effect on the aggregated distance matrix and on the connectivity of the resulting dendrogram.

The converged distance matrix obtained using a $k_{\text{max}}$ of 10 and 1800 individual $k$–means runs are subjected to the traditional hierarchical clustering to produce the aggregated dendrogram (Figure 7.3). The Cophenetic correlation coefficient between this dendrogram and the aggregated distance matrix $D_{\text{Avg}}$ is 0.93 indicating that the dendrogram accurately depicts the relationships described by the aggregated distance matrix. It is at the discretion of the user to choose a set of clusters from the hierarchy. Seven well–separated clusters have been selected from this dendrogram as the final aggregated $k$–means solution such that each cluster represents a distinct surface flow pattern.
Figure 7.2: The $\Delta\text{SSE}(k'_{\text{max}})$ comparing aggregated distance matrices $D(k'_{\text{max}})$ and $D(k'_{\text{max}}+1)$.

Figure 7.3: Dendrogram for hierarchical clustering of dissimilarity matrix $D^h$ for 3672 hourly wind field observations. Individual hours are not shown across bottom of dendrogram due to space limitations. Vertical lines indicate the 7 clusters (1–7) selected from the hierarchy.
7.4.2 Sequence Analysis

To determine the dependence of ozone on meteorology, sequence analysis is performed on the cluster labels obtained from the cluster analysis of hourly wind measurements for 153 summer ozone days. This clustering uses similarity index based on the length of the longest common subsequence as the distance metric since the objective of this part of the study is to cluster predominant wind directions which are categorical variables. The similarity indices, $S \ (153 \times 153)$ are calculated for the 24 hr sequences (0000–2300 CST) of diurnal wind field labels corresponding to the summer ozone days. These are then grouped by hierarchical clustering using average linkage with the distance matrix as $1-S$, where 1 is the identity matrix, to determine the groups of exceedance days having similar sequence of wind field directions. The dendrogram (Figure [7.4]) obtained has a Cophenetic correlation coefficient of 0.72, which is an adequate representation of the distance matrix. Seven well separated groups are selected from this dendrogram leaving out two outliers.
Figure 7.5: Hourly wind field patterns and diurnal distributions for 7 clusters of Figure 7.3. Cluster number and symbol are located in lower right corner of each map. Length of arrow indicates mean 1-hr wind speed as indicated on scale.

7.5 Cluster Interpretation

7.5.1 Hourly Wind Clusters

To interpret the clusters obtained, the prevailing wind conditions for each of the seven clusters and at each monitoring site are calculated as the mean of the wind measurements and these wind vectors are plotted in geospatial coordinates (Figure 7.5). The direction of the wind is indicated by the direction of the arrow heads while the length of the vector represents the magnitude of the wind speed. Figure 7.5 includes the frequency of occurrence of the cluster through the day. Cluster #1 and #3 have winds predominantly from the north–northeast and northeast direction; cluster #2 has winds from the northwest; #4 from the southeast. Cluster #5 represents stagnant conditions. Cluster #6 and #7 have winds from the southwest; Clusters #6 and #7 differ in terms of the strength of the winds. Cluster #6 has low wind speeds in comparison to cluster #7. These clusters are significantly different in capturing different wind speed and direction patterns.
The diurnal distribution plots indicate that many of the clusters are preferentially realized during certain phases of the diurnal cycle. Clusters #1, #3, #5 and #6 tend to occur during the night time and into early morning. Clusters #2 and #4 occur during the daylight hours and into the mid afternoon and evening, respectively. Cluster #7 occurs during the evening and into the night. The clustering algorithm does not consider the measurement sampling times; the emergence of these diurnal biases in the hourly cluster labels validates their correspondence to real wind field patterns. The morning clusters account for 52% while the night time clusters account for 48% of the 3672 hours in the study period. This is done deliberately to capture the effects of the diurnal fields relevant to ozone build up. Advection of pollutants in the night time from the petrochemical corridor of southeastern Texas is claimed to be one of the reasons of ozone problems in Baton Rouge, Louisiana. Thus, the night time clusters are selected to have similar level of specificity to the daytime clusters.

The 7 clusters vary widely in their proportions of hours which fall on an exceedance day (any hour from 0000–2300 CST). Hourly sampling intervals for clusters #1–#7 occur on exceedance days with frequencies 15%, 20%, 14%, 33%, 9%, 5%, and 6%, respectively within the cluster. This indicates that episodic conditions for clusters #6 and #7, with approximately 11% of the total 600 exceedance hours are minimal. These clusters have wind directions from the southwest and carry pollutants from the petrochemical corridor of Texas. Patterns #2 and #4, with flows arriving from the northwest and southeast, occur on exceedance days 53% of the time while the remaining night time clusters occur 47% of the time. Differences in their levels of association with exceedance days further validate the physical correspondence of the hourly cluster labels to real wind field patterns; still, the hourly patterns alone do not suffice to isolate episodic conditions or differentiate between exceedance scenarios.

### 7.5.2 Daily Wind Field Groups

Seven main groups (G1–G7) are selected from the dendrogram of Figure 7.4 having 26, 32, 20, 13, 16, 36 and 8 days. The sequences of hourly wind field patterns associated with these groups of daily air flow patterns are indicated in Figures 7.6 through 7.9. These figures indicate the general directionality and timing of the directional shifts for each day. Groups
Figure 7.6: Sequences of hourly wind field patterns for the first 2 groups of days from 2005 ozone season of Figure 7.4. Symbols correspond to cluster patterns in Figure 7.5.

G1 and G2 have southeasterly winds throughout the day with either stagnant winds or winds from the northeast on very few days. G3 has northeasterly winds in the morning that shift to southeasterly on all days with afternoon stagnation on few days. G4 has stagnant winds during the morning hours that shift to northwest direction. G5 has winds from the northeast in the mornings on all days that shift to northwest on very few days. G6 has northwesterly winds till late in the afternoon after which they shift to southwest direction or be stagnant. G7 has winds from the southwest in the morning that shift to northeast in the afternoon and another shift to southeast in the evenings. The two outlier days (09/13/05 and 09/15/05) have winds from the southwest through the days (Figure 9). Most of the groups have minimal shifts in direction through the day. This suggests the influence of synoptic forcing on ozone build up. Thus, we explore the weather maps also to determine the relationship between tropospheric ozone and synoptic weather patterns.
Figure 7.7: Sequences of hourly wind field patterns for days from groups 3 through 5 of the 2005 ozone season of Figure 7.4. Symbols correspond to cluster patterns in Figure 7.5.
Figure 7.8: Sequences of hourly wind field patterns for days from groups 6 and 7 of the 2005 ozone season of Figure 7.4. Symbols correspond to cluster patterns in Figure 7.5.

Figure 7.9: Sequences of hourly wind field patterns for the 2 outlier days from 2005 ozone season of Figure 7.4. Symbols correspond to cluster patterns in Figure 7.5.
Elevated ozone levels worldwide are associated with anticyclonic conditions. These cells of high upper-atmospheric pressure result in reduced wind speeds (or stagnation), clear skies (increasing solar flux), elevated temperatures, and subsidence; all of these are conducive to ozone formation and/or buildup. In this section, we investigate the influence of synoptic circulation patterns on ozone buildup. Figures 7.10 to 7.11 indicate the weather maps for the 7 ozone groups. G1 and G2 have the crest of a high pressure ridge formed in Caribbean basin that drifts to the west of the study domain. The pressure gradients are very weak through the study domain though G2 has slightly stronger pressure gradient than G1. This puts the study domain in the back side of the high synoptic pattern.
Figure 7.11: Daily 500–hPa weather maps for Groups G5 (Top Left), G6 (Top Right) and G7 (Bottom Left), averaged among days assigned to each group.
Groups G3 through G6 are characterized by two fronts; a ridge over the southern Great Plains with downstream troughs on the Pacific and upstream troughs on the Atlantic coast of the United States and a ridge in the Caribbean basin. The ridge in the Texas region dominates the pressure pattern, however, as the ridge across Louisiana drifts towards northeast. G5 shows a front to the south of the Texas–Mexico border. The pressure gradients are strong for this group. The strong pressure gradients indicate strong winds.

G7 has an anticyclone over peninsular Florida. This results in the advection of air from Gulf of Mexico. On the outlier day of 09/13/05, there is an anticyclone (Figure 7.12) over the western Atlantic coast. The state of Louisiana is in the back side of the high surface pressure synoptic with very weak pressure gradient and drifting from south to southeast. A back of the high pattern is observed when the study region is to the west of the ridge axis. The outlier day 09/15/05 is also on the back of the high synoptic of western Atlantic (Figure 7.13). This day, however, has another anticyclone in the Gulf that advects clean air into the study region.

7.6 Ozone Response to Meteorology

To determine the response of ozone to meteorology, prototype mean vectors are calculated from the 8–hr averaged ozone (Figure 7.14) and NO$_x$ (Figure 7.15) observations for the 7 groups of 153 summer ozone days at each of the monitoring station and are plotted on the geospatial coordinates.

1. G1 has low levels of ozone in the southern sites of the study domain with medium levels to the northwestern and northern portion of the industrial corridor at Grosstete and Baker. G2 has low levels of ozone through the study domain excepting at Baker where high levels of ozone are observed. The wind directions are predominantly southeast. Thus, the surface anticyclone causes the ozone and the precursors to drift from the southeastern sites to northwestward resulting in relatively high ozone levels at Grosstete and Baker.

2. G3 has high ozone levels at Baker and Capitol and low levels through the rest of the monitoring stations. G4 has high ozone levels at Captiol, Baker and Port–Allen and
Figure 7.12: 500-hPa weather maps at 0000 CST (Top Left), 0600 CST (Top Right), 1200 CST (Bottom Left) and 1800 CST (Bottom Right) on the outlier day of 09/13/05.
Figure 7.13: 500–hPa weather maps at 0000 CST (Top Left), 0600 CST (Top Right), 1200 CST (Bottom Left) and 1800 CST (Bottom Right) on the outlier day of 09/15/05.
Figure 7.14: Group-averaged daily maximum 8-hr ozone levels (ppb) at 7 monitors for 7 groups of summer ozone days. Empty triangles indicate values below scale minimum of 40 ppb.

Figure 7.15: Group-averaged daily maximum 8-hr $NO_X$ levels (ppb) at 7 monitors for 7 groups of summer ozone days. Empty triangles indicate values below scale minimum of 12 ppb.
low levels on the western side of the industrial corridor. G5 has the highest ozone levels recorded in the southern region of the study domain at Carville and low levels at all the other stations. The north–northeast winds observed for these days transport pollutants from the northern portion of the study domain to the south. This results in higher ozone levels found in the southern portion of the study domain. G6 has medium levels of ozone through the study domain. The groups G3 to G6 seem to have more of synoptic forcing than the meso scale patterns. Thus, for these groups the winds flow towards northeast resulting in higher ozone levels on the north and northeastern portions of the study domain. These groups contain days on which hurricane Katrina has affected the study domain (08/27/05 through 08/29/05). The trough or cold weather pattern observed for these groups produces clouds and precipitation that reduce the photochemistry. The stronger winds act to reduce the ozone concentrations by dispersing the pollutants and the precursors. The very high $NO_x$ levels observed in the northern portion of the study domain indicate the emissions from automobiles. A significant increase in automobile traffic from New Orleans could be the reason for an increased levels of precursors in the northern portion of the study domain.

3. G7 has high ozone levels at Baker and low to medium levels of ozone at all of the remaining sites. G7 has winds from the southwest predominantly and flow from the Gulf of Mexico. The clean air advected from the Gulf displaces ozone northward from the industrial corridor resulting in dilution of pollutants similar to the weather patterns of G1 and G2. This results in low levels of ozone in the southern portion of the study domain and high levels northward. G7 has low levels on the western portion and higher levels on the eastern side of the study domain. This pattern indicates the role of advection of pollutants from the southeast Texas petrochemical corridor by the southwest winds. The days (08/30/05 and 08/31/05) in this group are days during which hurricane Katrina has passed the study domain. These days experienced cold and warm fronts associated with the low pressure. These days are associated with high levels of $NO_x$ on the eastern side of the study domain.
Figure 7.16: Group-averaged daily maximum 8–hr ozone (ppb) and NO\textsubscript{X} levels (ppb) at 7 monitors for 7 groups of summer ozone days. Empty triangles indicate values below scale minimum of 40 ppb and 12 ppb, respectively.

This is indicative of the increase in emissions due to the increase in automobile traffic in the study region on these days.

NO\textsubscript{X} levels are elevated near the eastern side of the study domain for all the groups. This is consistent with higher ozone levels generally observed on the eastern part for all the groups. The consistency of the pollutant spatial distributions indicates that the groups capture significant transport and dispersion patterns. The first outlier of 09/13/05 has very high ozone levels (Figure 7.16) of the order of 90 ppb recorded while the second outlier day has relatively low ozone levels. NO\textsubscript{X} levels for these days are consistent with the observed ozone concentrations. The intense ridge in the Atlantic is usually associated with very high ozone levels in this region (Rohli et al., 2004). Thus, this outlier day falls in the synoptic pattern with very high ozone levels. On the second outlier day, the ridge moves away from the study region towards the northeastern United States through the day. This pattern is associated with advection of clean air from the Gulf of Mexico that dilutes pollutants resulting in lower levels of ozone and NO\textsubscript{X} on this day in Baton Rouge, Louisiana.

Selection of additional clusters (wind clusters more than 7) and groups (G>7) were explored. This did not result in significantly different dispersion patterns for the study period. Cluster analysis at the hourly scale was explored using Euclidean distance metric instead of the Correlation metric. The results obtained, however, were misleading. The dendrogram obtained from this cluster analysis and the corresponding cluster averaged wind
Directions are shown in Figures 7.17 and 7.18, respectively. In this analysis, #4 shows a near-stagnant pattern while #5 is not stagnant though it is very close to #4. On the other hand, #10 has large merging distance but is a stagnant cluster with very low wind speeds. Thus, this analysis indicates that no single metric or clustering methodology is perfect for a data set.

7.7 Conclusions

Clustering has been applied at two different time scales to study the effect of meteorology on ozone. The first stage of clustering is carried out for the surface wind observations on hourly scale with agglomerative clustering algorithm. This algorithm uses a correlation distance metric to determine the dissimilarity among observations to obtain the meteorological regimes. The second stage of clustering (grouping) is carried out for the 24–hr wind labels corresponding to the 153 days of the study period. This clustering is performed using the LLCS metric and hierarchical clustering to determine the dissimilarity among the sequence of symbols of observations. This two-stage approach of clustering which is performed on a
Figure 7.18: Hourly wind field patterns and diurnal distributions for 10 clusters of Figure 7.17. Cluster number and symbol are located in lower right corner of each map. Length of arrow indicates mean 1-hr wind speed as indicated on scale.

Large scale of data emphasizes the effect of both the local and synoptic scale driven climatology on ozone formation in Baton Rouge, Louisiana. The main results from this case study can be summarized as:

- Seven diurnal wind field patterns are isolated using the cluster analysis on hourly surface wind observations.
- The clustering at daily scale resulted in seven different groups that capture distinct ozone formation mechanisms in the study region.
- The predominant directions and the weather maps indicate that both the meso scale and synoptic scale meteorology are important to isolate the role of meteorology on ozone in this region.
Chapter 8

Case Study: Forecasting Ozone Levels in Houston, Texas Region

8.1 Background

Pollution forecasts are issued by air quality agencies for public notification purposes to help the public avoid or minimize exposure to unhealthy air. There exist several methods for predicting the ozone concentrations and the air quality. No single method, however, is very accurate and several methods have to be used in conjunction to produce a more accurate forecast. A brief review of the methods available for air quality forecasting in general and ozone concentrations forecasting in particular is given in chapter 2.

In this chapter, we explore the applicability of HMMs for ozone forecasting. Two separate case studies are performed to envision the capability of a HMM for ozone forecasting. The first case study (section 8.2) uses HMM to classify days into two categories: exceedance and non-exceedance. The study domain is the same Houston, Texas region examined in chapter 6. A description of the study region is given in section 6.2. Hourly wind speed, direction and ozone data are obtained from the monitoring stations shown in Figure 6.1. The second case study involves the development of HMMs to predict one day ahead forecasts of daily maximum concentrations in Houston, Texas. The study domain for this case study is a subsection of the monitoring stations shown in Figure 6.1. These stations are discussed below.
8.2 HMMs for Associating Surface Wind Field Patterns with Tropospheric Ozone

Chapter 6 uses the hierarchical clustering method to determine the diurnal air flow patterns relevant for ozone build up in the Houston, Galveston and Beaumont–Port Arthur, Texas region. Clustering is performed at the hourly time scale on the hourly surface wind measurements to identify the surface patterns affecting the study domain. A second clustering is applied at the daily time scale by taking the input from the output of the first clustering at the hourly scale. It has been found that ten different wind regimes contribute to four groups of exceedances and eight groups of non–exceedances. This study in chapter 6 has been done for the 32 exceedance days and for the 182 non–exceedance days of 2004. These hourly wind field patterns and the diurnal air flow patterns capture the distinct ozone exceedance and non–exceedances scenarios for the summer ozone months of 2004.

The present case study on HMM classification considers the entire 2004 and 2005 ozone seasons (1 April to 31 October) for the Houston, Galveston, and Beaumont–Port Arthur, Texas domain. HMMs are developed for the different wind clusters and diurnal flow pattern groups. The data set used to develop the HMM for the wind clusters consist of the continuous hourly surface wind measurements corresponding to the clusters of hourly labels. Thus, these HMMs are “continuous HMMs” while the HMMs developed for the diurnal flow pattern groups consist of sequences of categorical symbols. Thus, these are referred to as “categorical HMMs”. The HMMs thus developed are then used to label non–modeled (future) days of the summer months of 2005 (214 in total) to aid in the ozone forecasts. The goal of the study is to envision application of HMMs to meteorological forecasting for classifying air quality as belonging to one of two categories: ozone exceedance or non–exceedance. The 10 wind cluster patterns consist of 637, 576, 1119, 49, 760, 428, 731, 299, 457 and 80 hours of the total 5136 hourly observations, respectively.

8.3 HMM Training

Data set for the continuous HMM training derives from the hourly surface wind speed and direction data available from 28 monitoring stations. The hourly wind data is transformed into westerly and southerly vector components. Thus the data set for continuous HMM
training consists of 5136 rows and 56 columns of hourly wind field observations. The training data for a cluster $c$ of size $N_c$ has a matrix of $N_c$ rows and $s$ columns of hourly observations. For each HMM, the Bayes Information Criterion (BIC), is used as a model selection criterion to determine the number of states.

The training procedure begins with the initialization step where the parameters of the model, $\gamma$ and $A$ are initialized randomly. To estimate the emissions distribution, initially the conditional independent model with Gaussian distributions is considered. The model for most of the randomly initialized runs failed the maximization of the likelihood. Next, mixtures of multivariate Gaussian emissions are considered. Each state for each continuous HMM is modeled as a mixture of Gaussians (Figure 8.1).

The best homogeneous multivariate Gaussian in terms of the BIC model selection for each cluster is selected to implement the validation procedure. The parameters of HMM are reestimated using the reestimation formulas of the Baum–Welch algorithm until the probability of observation of the training data given by the parameters of the new model show no improvement over the parameters of the previous model.

The 10 wind cluster patterns consist of 637, 576, 1119, 49, 760, 428, 731, 299, 457 and 80 hours of the total 5136 hourly observations, respectively. The training data set for the continuous HMMs consist of the observations corresponding to the 10 classes of wind patterns. For example, the training data for cluster 1 consists of 637 hours of observations measured at the 28 sites (637 $\times$ 56 with $u$ and $v$ wind components of data at each station).
For each cluster, experiments are performed for the number of hidden states from 2 through 10 and for each case with a random set of values for the HMM parameters $\gamma, A$. The HMMs investigated in this case study uses continuous Gaussian mixtures to model the output observation distributions for each state. The model parameters are continuously updated until no improvement can be obtained for the given set of training data. The BIC selection (Table 8.1) allows 5, 4, 5, 2, 2, 8, 4, 4, 4 and 2 states respectively for the ten continuous HMMs.

Next, HMMs are developed for the ozone exceedance and non–exceedance classes. The training data for discrete HMM consist of sequences of wind cluster labels belonging to that ozone group. The training data for the discrete HMMs consists of the 0000–1700 CST sequence of 18 hr cluster labels for each day. For an ozone group of size $n$ days, the training data set for the discrete HMM consists of a matrix of observations of size $n \times 18$. The group size for the ozone exceedance classes obtained from sequencing is 7, 12, 10 and 3 days while the non–exceedance classes have 49, 24, 15, 30, 15, 8, 25 and 15 days. Thus, for instance, training data for group #1 of the exceedance groups consists of 7 sequences of 18–hr length. Discrete HMMs are developed by considering random initializations for the parameters and a multinomial distribution for the emission probable densities of the discrete symbols 1 through 10. These 10 symbols correspond to the 10 wind field patterns. Several randomly initialized trial runs are performed to arrive at the optimal solution. The BIC (Table 8.2) allows selection of states for discrete HMMs 1 through 12 (4 exceedances and 8 non–exceedances).
The exceedance groups have 2 states each for all the groups and non-exceedance groups have 3, 4, 2, 3, 2, 3, and 2 states respectively. The initial model parameters are updated until no improvement can be found in the log likelihood of the training observations.

### 8.4 Model Validation

The validation procedure for the HMMs consists of classifying the 5136 hours of data corresponding to the 214 days of 2005 among the 10 continuous HMMs. Each hourly observation of the surface wind data \((1 \times 56)\) is input to each continuous HMM to generate the probability of the hour belonging to the model. Classification is then realized by choosing the class represented by a model with the maximum likelihood of an observation being realized from that HMM. The classification procedure of the testing phase of the HMM is given in Figure 8.2. Next, the 0000–1700 CST labels belonging to the 18hr of a day for the 5136 hours obtained from the continuous HMMs are stacked in to vectors each with a sequence of 18 observations. These 214 sequences \((214 \times 18)\) are then input individually \((1 \times 18)\) to the discrete HMMs and the class label is then determined for each day as the one with the maximum probability. The future non-modeled day is labeled as exceedance if the sequence of observations from the day has the maximum log-likelihood realized from any of the four exceedance HMMs and is a non-exceedance if the log-likelihood corresponds to any of the eight non-exceedance HMMs. To evaluate the model performance, the classifications to the two categories realized are compared to the observed ozone class labels and to the forecast
Figure 8.2: Classification of a test sequence using $N_i$ HMMs. The test sequence belongs to the model with the highest probability among the $N_i$ models.

statistics issued by TCEQ. The above procedure of predicting the ozone class starting from wind measurements is shown in flow chart of Figure 8.3.

In the present study, forecast is issued for two categories: whether an Ozone Action Day (OAD) occurs; or not. An OAD is defined as a day when the ozone levels are forecast to be equal to or greater than the daily maximum of 85 ppb, which is the 1997 ozone NAAQS, at any of the monitoring sites within the study domain. To evaluate the model statistically, verification statistics defined in section 5.5.1 are computed. The model performance is evaluated using archived meteorological data for the days ranging from 1 April to 31 October of 2005 (214 in total). To verify and validate the model, statistical measures are computed for the forecasting program based on our current approach of HMMs and for the ozone forecast program of the Texas Commission for Environmental Quality (TCEQ) as shown in Figure 8.4.

Use of actual archived wind data gives our method an unfair advantage over TCEQ models as the TCEQ forecast models use simulated wind fields which we did not have access to. The current modeling method, however, is a proof of concept to demonstrate its feasibility for ozone forecasting and an actual forecasting could be performed in a similar way with simulated data. TCEQ uses a variety of information such as the numerical
Figure 8.3: Flow chart for ozone prediction.

Figure 8.4: Contingency tables for two category forecast using HMMs and TCEQ models. The skill parameters calculated for each model are shown below the respective table.
ozone model output from the air quality forecast guidance of NOAA/NWS for east Texas (http://www.nws.noaa.gov/aq/sectors/easttexas.php#tabs), output from the East Texas Air Quality Forecasting System of the University of Houston (http://www.imaqs.uh.edu/aqfmain.htm), and the output from statistical based models to make the ozone forecasts in Texas (Lambeth, 2009).

A two–category forecast, a classification of 8–hr ozone concentrations at or above 84 ppb and below 84 ppb, is evaluated for the 214 day period. The accuracy (A) of the TCEQ model is higher than that for the HMM as seen from Figure 8.4. FAR for both the programs are high; this is due to incorrectly predicting one non–exceedance event as an exceedance event. The FAR for TCEQ model is, however, very high and is almost one and half times the FAR of the HMM model. The CSI measures the forecaster’s ability to predict the high ozone events, while excluding the large occurrence of correctly forecasted low ozone days. HMM has a low CSI of 25%, meaning that only 25% of the high ozone events were forecasted correctly while TCEQ has a CSI of 49% and hence does a better job of predicting the ozone events. HMM has a measure for the probability of detection equal to 39% while the TCEQ model has a POD of 89%. HMMs have higher level of difficulty in predicting the ozone exceedance when it actually does happen. TCEQ model has a higher level of detecting the exceedance. The TCEQ model, however, has a very high FAR indicating a poor performance in spite of the high POD. Figure 8.5 shows the graphical interpretation of the contingency table for both the models.

The regression line of forecasts upon observations for the TCEQ models is higher given the high POD when compared to the regression line of forecasts upon observations of the HMM. The regression lines of observations upon forecasts are closer to the dashed line for the HMM than that of the TCEQ model. This is due to the high FAR of the TCEQ model. Overall, TCEQ has a higher accuracy due to the high number of correctly forecast exceedance events. The $P_{2AFC}$ score for the HMMs is 62 percent while for the TCEQ models, it is 84 percent. TCEQ uses models that are computationally intensive and HMMs which are simpler statistical methods have the accuracy levels (76%) which are almost the same as
that of the TCEQ models (81%). Thus, HMMs that can be implemented relatively easily attain a level of comparable accuracy as that of TCEQ models.

The predictive ability of the HMMs could be improved by incorporating data sets spanning multiple years. Several wind regimes can be determined from multiple year wind observations. This in turn increases the ability of the model to predict the forecast. One of the drawbacks of the HMMs is that it converges to a local minimum and the choice of the initial conditions affects the performance of the model. The model has been experimented with numerous cases of number of states and random initializations for the model parameters. There exists the possibility that the HMMs have converged to a sub–optimal local maxima. Thus, HMMs could be improved by starting with a different set of initial conditions and the number of states.

### 8.5 HMMs for Point–Value Predictions of Daily Maximum Ozone Concentrations

The second case study aims at applying the HMMs to daily maximum ozone observations from the summer months (April to October) of 2001 through 2004. These models are then used to predict the daily maximum ozone concentrations for the summer months of April to October from the year 2005. The study domain for this case study is a subset of the previous
Figure 8.6: Map of Houston study region with the subset of stations from figure [6.1] used in this study.

study, shown in Figure [6.1] Daily maximum ozone levels from 7 monitoring stations (stations 1, 2, 3, 7, 11, 13 and 14 from Figure [6.1]) are used for developing the HMMs and are shown in Figure 8.6. The HMMs of this case study are referred to as point–value HMMs. This case study demonstrates the applicability of HMMs for forecasting the point value ozone mixing ratios. The MVN–HMM toolbox (section [5.6.2]) developed by Kirshner (2005) is used extensively to develop and validate the models. Thus, the aim of this case study is to investigate the capability of hidden Markov models to determine the daily maximum ozone levels in Houston, Texas region.

8.5.1 Conditional Distribution of the Daily Maximum Ozone Concentrations

In the air quality literature, various probability models have been proposed for fitting distributions to air quality data (El–Shaarawi et al., 2002). Models include, but are not limited to, the log–normal distribution, gamma distribution and the Weibull distribution. For this case study we have explored numerous distributions to find the model that best fits the data. In the HMM framework, models include:
1. The conditional independence models with gamma, mixtures of gamma, log–Normal, gaussian distributions and mixtures of Weibull emissions. The data from the 20 ozone monitoring stations are used for developing the models. Models are considered with the number of states ranging from 2 to 10. In general, the BIC values are found to decrease as far as the number of states increases. Very often, the MVN–HMM toolbox fails the maximization of the likelihood. These models are developed based on the hypothesis of conditional independence at the different stations. The stations, however, are highly correlated and so these models resulted in a poor fit for the data. This suggested a reduction in the number of stations. The stations that are spatially very close are not included in further analysis. The models developed on the reduced data set of 13 (stations A, B, C, E, F, H, I, K, M, N, P, Q, and S from Figure 6.1) stations also did not yield a good fit. The data set is further reduced to 9 (stations A, B, C, G, K, M, N, P and S from Figure 6.1) stations and mixtures of Weibull emissions are fitted to the data. The algorithms meet the maximization of the likelihood. The best HMM model obtained with 3 states and 1 Weibull component, however, gave a very poor data fit.

2. Multivariate Gaussian emissions: Initially, models are developed for the entire 20 stations. In this case, there exist no issues in maximizing the likelihood and the 3–state model has the minimum BIC value. The conditional distributions obtained, however, assign a positive mass to negative values. As an alternative data is transformed using logarithmic and box–cox transformations. Multivariate Gaussian models are developed for the transformed data. This resulted in reduced data variability and, in turn, an increased computational burden. Then, multivariate Gaussian emissions are fit to the set of 9 stations. The best homogeneous model, in this case, has 4 hidden states. As a first approximation, the problem of a positive mass to negative values is ignored. This approximation is validated by the simulations which generated the negative values in very few cases. The main observed quantities (marginal means and variances) are reproduced very closely. The model, however, is not completely adequate in reproducing the frequency of days with ozone concentrations that exceed the threshold of 84 ppb.
Table 8.3: BIC values for model selection of the point–value HMMs.

<table>
<thead>
<tr>
<th>States</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>4.30E+04</td>
<td>4.26E+04</td>
<td>4.24E+04</td>
<td>4.24E+04</td>
<td>4.25E+04</td>
<td>4.26E+04</td>
</tr>
</tbody>
</table>

3. The non–homogeneous HMM with two different set of input variables: the 18 series of temperatures and $NO_x$ values and a set of 5 linear combinations of these variables, obtained by principal components analysis is considered. The model selections (comparison of BIC values) for the non–homogeneous models are not advantageous over the homogeneous models.

The homogeneous 4–state multivariate Gaussian model for the 9 stations is selected as the best model for preliminary investigation to implement the validation procedure with an out–of–sample daily maximum ozone concentration data from the year 2005. There are 7 stations in 2005 that have data coincident with the 9 stations from 2001 through 2004 that are used to develop the models. The MVN–HMM toolbox is not capable of handling the missing data when the multivariate Gaussian emission is considered. Therefore, the multivariate Gaussian model used in this case study is developed using the data from 7 (Figure 8.6) stations (stations A, B, C, G, K, M and N from Figure 6.1). These 7 stations have complete data records for the years 2001 through 2005.

### 8.5.2 Estimated Model

HMM is applied to the daily maximum ozone concentrations from 7 stations for the summer months of April 1 to October 31 for the years 2001 through 2004. Thus, the data used for estimating the model parameters consist of 856 days of observations measured at 7 stations. The MVN–HMM toolbox is used to estimate the parameters of the model: transition matrix, means and covariances at the 7 stations for the emission distributions and the stationary initial state distributions. The BIC (Table 8.3) allowed us to select 4 states for the model. The conditional distributions of the ozone concentrations for each state (Figure 8.7) and the Box plots (Figure 8.8) are used to interpret the estimated hidden states. The Box plots compare the Viterbi sequence and the observed data at each station and are grouped by the state.
In Figure 8.7, the plot in blue represents station 7 (Beaumont). The mean values for state 1 are around 40 ppb for all the stations; around 30 ppb for state 2; around 80 ppb for all stations excepting for station 7 in state 3; around 50 ppb for state 4. This indicates that states 1 through 4 represent “low”, “very low”, “high” and “medium” levels of ozone concentrations. The Box plots (Figure 8.8) too reveal the same information that at each station, state 1 has low values; state 2 represents the lowest ozone concentrations; state 3 has the highest values and state 4 has the medium values of ozone levels.

To further explore this observation, the extreme ozone events in the training period captured by the viterbi sequence of states are considered. The representation of the Viterbi sequence by year is represented graphically in Figure 8.9.

The Viterbi sequence represents the most likely sequence of states associated with the data. There are 125 extreme events in the period 2001–2004. Out of these, 110 days are realized in state 3; 14 days in state 4 and the remaining 1 day in state 1. This indicates that most of the extreme events are realized in state 3 and thus state 3 conditionally represents a high ozone state.
Figure 8.8: Box plots of the daily values grouped according to the classification of the day in the Viterbi sequence.
Figure 8.9: Graphical representation of the Viterbi sequence year by year: black = state 1, white = state 4.

The plot in Figure 8.7 shows that conditionally to state 3 (high ozone values) Beaumont has the lowest probability of recording an extreme value. An explanation for this phenomenon is found from the geospatial plots of the ozone levels observed in Houston and Beaumont (Figures 6.28 and 6.30). These plots are obtained from the mean daily maximum 8-hr ozone levels on the exceedance and non-exceedance days of 2004 and for 20 stations. The patterns, however, can be considered to be representative of the region for the years 2001 through 2004 though some new patterns arise if larger data sets are considered. Houston and Beaumont experience different ozone exceedance levels for consistent wind measurements. Beaumont and Port Arthur ozone levels are elevated but usually not to the exceedance levels. This could be explained by the low levels of the ozone precursors ($NO_x$) recorded at the Beaumont stations in comparison to Houston (Figures 6.29 and 6.31). The conditional distributions at state 2 have tails with very low ozone levels at all the stations. This state can be associated with the non-exceedance groups since none of the 125 exceedance days are realized conditionally in this state. This interpretation of the states altogether seems to be consistent with the actual ozone concentrations observed and the corresponding precursor concentrations.
8.5.3 Goodness of Fit for the Estimated Model

To further validate the model, the goodness of fit is checked for the model. The HMMs reproduce the observed frequencies of the daily means and the standard deviations very closely (Figures 8.10 and 8.11). The observation density (Figure 8.12) of ozone concentrations is calculated by assuming the stationary distribution to be the initial distribution for the hidden Markov chain so that the observable process is stationary. These plots indicate that the model accurately fits the observed data.

Thus, the unconditional density of the observation for a day $d$ and station $s$ is

$$f^*_d(u) = \sum_{j=1}^{4} \frac{\exp[-(u - \mu_s^j)^2/2\sigma_{s,s}^{j}]}{\sqrt{2\pi\sigma_{s,s}^{j}}}.$$  \hspace{1cm} (8.1)

where, $u$, $\gamma_s^j$, $\mu_s^j$ and $\sigma_{s,s}^{j}$ are the ozone concentrations ranging from 0 to infinity, stationary distribution, estimated mean and standard deviations at state $j$ and station $s$, respectively. $s$ has values from 1 to 7. The probability of the occurrence of an exceedance is estimated using the model parameters as

$$p^s = \int_{u=84}^{\infty} x^*_d(u) \, du.$$  \hspace{1cm} (8.2)
Figure 8.11: Comparison of the observed and the estimated standard deviations.

The estimated log–odds ratios are calculated to compare the stations pairwise with respect to extreme events, as

\[ \log \left( \frac{p^s / (1 - p^s)}{p^j / (1 - p^j)} \right) \]

Equation 8.3

Here, \( s, j \) range from 1 to 6. This is then compared to the observed log–odds ratios of the extreme events. The plot (Figure 8.13) of the log–odds ratios for the observed and estimated instances correctly estimates the probability at all stations excepting for comparisons involving Beaumont.

The model, unconditionally, overestimates the probability of an extreme event at Beaumont while it correctly estimates the probability at the other stations. There are 30, 41, 40, 66, 35, 46, and 4 exceedances respectively at each station from 1 through 7. The large numbers of stations that are situated around Houston have high ozone values in comparison to the lone outlier–Beaumont. This is consistent with the low conditional probability of Beaumont at state 3 which captures the high ozone events. This causes the model to learn more from stations in Houston than from Beaumont and thus the model over estimates the exceedance events in Beaumont.
Figure 8.12: Histogram of data together with the estimated density of observations at each station.
Figure 8.13: Comparison of the log-odds ratios for the estimated and the empirical distributions.

8.5.4 Model Testing and Validation

The validation procedure consists of applying the model parameters estimated using the data from 2001 through 2004 to predict the daily maximum ozone data from 2005. The 2005 data is an out-of-sample season data and consists of 27 exceedances measured among the 7 stations. The validation procedure is based on the predictive distribution:

\[
P(X_{d+1}^s \leq x | X_1, \ldots, X_d).
\]  

(8.4)

where \(X_{d+1}^s\) is the daily maximum ozone concentration at station \(s\) and day \(d+1\); vector \(X_d\) denotes the vector of daily maximum ozone concentrations on day \(d\) measured for the whole network of \(s\) stations.

The above equation, based on the HMM assumptions can be expressed as:

\[
\sum_{i=1}^{4} P(X_{d+1}^s \leq x | q_{d+1} = i) \sum_{j=1}^{4} p(q_{d+1} = i | q_d = j) p(q_d = j | X_1, \ldots, X_d). 
\]

(8.5)
The first term in the equation is the conditional distribution from Equation 8.1, second term is the transition probability matrix estimated from the model; and the third term is the probability of being in a state \( j \) on day \( d \). This is obtained numerically by filtering procedure as:

\[
p(q_d = j|X_1, ..., X_d) = \frac{\alpha_d(j)\beta_d(j)}{\sum_{j=1}^{4} \alpha_d(j)\beta_d(j)}. \tag{8.6}
\]

Filtering involves the prediction of the conditional distribution of a state given the observations. The quantities \( \alpha_1(j), \alpha_2(j), ..., \alpha_d(j) \) and \( \beta_1(j), \beta_2(j), ..., \beta_d(j) \) are computed successively for each state \( j \) to determine the conditional distribution of the state.

The state sequence obtained from filtering is shown in Figure 8.14. The daily maximum concentrations that exceed the threshold value of 84 ppb are found mostly in state 3 (16 out of the total 27 exceedances while state 2 has the least number of exceedances (1 out of 27). This validates our observation that state 3 conditionally corresponds to high ozone levels. Once the predictive distribution of the observations is determined, the observed ozone concentrations are compared with the 95% mass of the predictive distribution i.e. the quantiles of the order 0.025 and 0.975 of the distribution are computed numerically and are compared with real data. Figures 8.15 to 8.21 show the 1-day ahead interval forecasts of daily maximum ozone concentrations from 1 April 2005 to 31 October 2005 (214 days) for each station. The plot in red is the moving average of 20 observations at a time. The difference between the forecast mean (green) and the moving average is computed and if this difference is higher than a set value of 10 ppb, an exceedance notice is issued.

The plots for all the stations can predict the daily maximum ozone mixing ratios reasonably well, although they miss certain observations. In the plots, there are certain intervals that do not contain the observed values after the prediction was made. The predicted intervals capture 97%, 99%, 98% 99%, 98%, 99% and 98% of the observed concentrations at each station respectively. The errors found are due to the lower bound of the predictive distribution i.e. the observed ozone levels on the days that do not fall in the prediction interval are lower than the lower bound of the predicted interval. Thus, this analysis shows that the prediction interval covers the observed daily maximum concentration values accurately.
Figure 8.14: Estimated state sequence for the daily maximum concentrations of 2005 obtained using the Viterbi algorithm.

Figure 8.15: Interval forecasts 1–day ahead expressed as 95% prediction intervals compared with the daily maximum ozone concentrations from April 1 to October 31, 2005 at station 1.

O₃ [Ppb]
Figure 8.16: Interval forecasts 1–day ahead expressed as 95% prediction intervals compared with the daily maximum ozone concentrations from April 1 to October 31, 2005 at station 2.

Figure 8.17: Interval forecasts 1–day ahead expressed as 95% prediction intervals compared with the daily maximum ozone concentrations from April 1 to October 31, 2005 at station 3.
Figure 8.18: Interval forecasts 1–day ahead expressed as 95% prediction intervals compared with the daily maximum ozone concentrations from April 1 to October 31, 2005 at station 4.

Figure 8.19: Interval forecasts 1–day ahead expressed as 95% prediction intervals compared with the daily maximum ozone concentrations from April 1 to October 31, 2005 at station 5.
Figure 8.20: Interval forecasts 1–day ahead expressed as 95% prediction intervals compared with the daily maximum ozone concentrations from April 1 to October 31, 2005 at station 6.

Figure 8.21: Interval forecasts 1–day ahead expressed as 95% prediction intervals compared with the daily maximum ozone concentrations from April 1 to October 31, 2005 at station 7.
8.6 Conclusions

This case study demonstrates the utility of HMMs presented in chapter 5. A final summary of the case study is provided below:

- Statistical HMMs are developed for hourly surface winds and for ozone exceedance and non-exceedance classes around Houston. The results show the potential of HMM for categorical ozone prediction.

- These models compare well with the models currently being employed by TCEQ for ozone forecasts. While these results are interesting, the utility of HMM model is limited by its inability to attain a high probability of detection.

- HMM model could be improved by exploring the problem of choosing the initial model. Future study includes exploring the initial model problem and extension of the forecast methodology to include three or more categories.

- The daily maximum ozone concentrations observed for the year 2005 at each station are covered by the predictive distribution from the HMMs. The HMMs predict the daily maximum ozone concentrations accurately.
Chapter 9

Conclusion

9.1 Contributions

The work described in this dissertation has made substantial contributions to the engineering community in the field of environmental pollution monitoring. Three significant contributions have been made in the area of data–driven techniques based classification and monitoring. Advanced data–driven techniques widely used in process monitoring are extended to the field of environmental monitoring to study the tropospheric ozone problem.

The main contribution of this work is the novel methodology for data analysis and pollution monitoring. This methodology enables identification of transport and dispersion patterns that significantly enhance ozone production. Multivariate statistical methods– cluster analysis, sequence analysis and hidden Markov models are applied to three case studies from the field of air quality.

- **Cluster analysis for system classification**: The first case study of Chapter applies a hierarchical aggregation scheme to the traditional $k$–means cluster analysis. This study demonstrates the applicability of similarity metric based on sequencing algorithms which are widely used in the field of bioinformatics to determine the ozone buildup mechanisms. Cluster analysis has been applied by several researchers. The method employed in this case study is nearly automatic and allows for the analysis of large data sets with minimal subjective input from the user. The automation is made possible by the use of sequence analysis. The resulting clusters indicate sets of exceedance days that share a common pattern. These patterns result from the
differences in the prevailing meteorological conditions in the Houston, Texas region. Four distinct mechanisms that lead to high ozone formation and 8 mechanisms for low levels of ozone affecting the study region are isolated. The results are useful for an approximate assessment of the ozone exposure and to determine the representative conditions for developing the AQMs for estimating ozone exposure.

- **Sequence analysis for data reduction**: The second case study of Chapter 7 highlights the applicability of statistical methods to large data sets. A metric based on dynamic programming algorithms is explored in this case study. This is the first known example of application of this metric in the field of air quality. Cluster analysis is applied to observed data from a network of monitoring stations to reveal the synoptic patterns in Baton Rouge, Louisiana. This study aids in visualizing the mesoscale as well as synoptic patterns that contribute to ozone formation in this region. The analysis reveals seven mechanisms that lead to different levels of ozone. These mechanisms can be summarized as three distinctly different synoptic weather patterns. The first mechanism is driven by a ridge in the Caribbean basin while the second mechanism is driven by two ridges of high pressure patterns in the Gulf of Mexico. The third mechanism is driven by an anticyclone in the peninsular Florida. These mechanisms have varying levels of ozone observed in the study region. Few days are driven by local conditions than the synoptic patterns. This indicates the influence of both mesoscale and synoptic weather conditions in this region.

- **Hidden Markov models for time series prediction**: The third case study of Chapter 8 highlights the applicability of hidden Markov models to forecast ozone levels. This study demonstrates the usefulness of the proposed novel combinative method based on criteria forecasting. The proposed methodology can be applied for multiple observations and is built on the results obtained from Chapter 6. This procedure of developing hidden Markov models for predicting the occurrence of an ozone exceedance starting from the wind clusters of diurnal field patterns identified from cluster analysis is one of the contributions of this dissertation. The motivation for this method comes from an analogous problem in speech recognition called the isolated
word recognition problem. The results for categorical forecasting obtained from this analysis are comparable to the rigorous models employed by TCEQ. This demonstrates the capability of the HMMs for categorical forecasting. The HMMs are also used to predict daily maximum ozone concentrations. The results from this analysis capture the observed values of ozone within the bounds of the predictive distribution.

9.2 Recommendations

The methodologies proposed in this work are generic in nature and have tremendous application in different areas of engineering. Apart from the environmental field, the proposed framework has significant applications in myriads of areas with time series measurements such as bioinformatics and social sciences. To ensure that the statistical models are in parsimony with a specific data set, modifications are required, however. Some issues that could be investigated are given below:

1. In the field of air quality, the methodology could be extended to analyze particulate matter (PM) measurements. Ozone and PM measurements could be analyzed simultaneously. This helps in identifying the mechanisms that are common to both that could lead to development of common strategy for the abatement of these pollutants.

2. The sequence methods could be improved by considering the combinatorial representations of the sequences. There exist numerous dynamic programming algorithms to determine the similarity between sequences in literature. These could be extended to the field of air quality. These algorithms aid in data reduction and thus a better and improved representation of patterns for better understanding of air quality phenomena could be obtained. Combinatory algorithms quantify the number of subsequences present in a sequence, thus resulting in a new metric for measuring distances. These metrics then could be used as the distance metric in clustering algorithms.

3. The models developed for the categorical HMM prediction use mixtures of Gaussians for each state. The model complexity could be reduced by tying all the clusters to share single mixtures of Gaussian for all the states. This method is known as tied mixture
hidden Markov model (TM–HMM) method. TM–HMMs provide balance the conflict between detailed modeling and robustness to insufficient training data. Thus, the method of TM–HMM could be extended to this study to reduce the model complexity.

4. The application of clustering methods to BTR area indicates that both mesoscale and synoptic scale patterns are important to ozone formation. To determine the affect of synoptic climatology more accurately, time series clustering based on dynamic PCA could be explored. The application of HMMs to predict the daily maximum concentrations could be improved by incorporating the wind filed data. One of the factors in ozone formation is dispersion of pollutants and including wind data into model development could lead to improved model accuracy. Similar to the HMM classification framework for developing categorical forecast, cluster analysis could be explored to effectively assign a new data set to the clusters developed.


EPA, 1996: Guidance on Use of Modeled Results to Demonstrate Attainment of the Ozone NAAQS. EPA-454/B-95-007, USEPA Office of Air Quality Planning and Standards, Research Triangle Park, NC.


Lambeth, B., 2009: Personal communication.


TCEQ/ TAD., 2002: Conceptual model for ozone formation in the Houston–Galveston Area. Appendix A.


Appendix

Permission to Use Published Material

from  Hanwant Singh <hanwant.b.singh@gmail.com>
to    Swathi Pakalapati <spakal1@tigers.lsu.edu>
date  Tue, Mar 24, 2009 at 11:06 AM
subject Re: Permission for the reprint of article
mailed-by gmail.com
signed-by gmail.com

You do not need any permission to reprint your own work. There is no problem here. Hanwant Singh

from  Swathi Pakalapati <spakal1@tigers.lsu.edu>
to    Hanwant.B.Singh@nasa.gov,
Hanwant.B.Singh@gmail.com

date  Mon, Mar 23, 2009 at 6:11 PM
subject Permission for the reprint of article
mailed-by tigers.lsu.edu

Dear Mr. Singh:

Please reply with your permission to reprint the following Atmospheric Environment Journal article that I intend to include as a part of my PhD dissertation. I will include an appropriate credit line referencing the original article.


Thank you. I look forward to an e-mail from you.

Sincerely,
Swathi Pakalapati
Doctoral Student,
Chemical Engineering Department,
Louisiana State University,
Baton Rouge, LA 70803
(510) 735 1909
Vita

Swathi Pakalapati was born in June, 1982, in Palvancha, India. She completed her high school studies and college from Vijayawada, India. She graduated from Indian Institute of Technology Madras, India, in May 2004 earning both Bachelor and Master of Technology degree in the chemical engineering discipline as part of the 5 year integrated dual degree program. In August 2004, she was admitted to Louisiana State University, Chemical Engineering Department. There she worked under Dr. Jose.A.Romagnoli and completed her research work successfully in August 2009. Swathi is expected to receive a doctorate degree in chemical engineering in the summer of 2009. She hopes to develop and pursue her career in research oriented industry and eventually come back to academia.