SEGMENTATION OF HIGHWAY NETWORKS FOR MAINTENANCE OPERATIONS

FINAL PROJECT REPORT

by

Moo Yeon Kim and Jorge A. Prozzi
The University of Texas at Austin

Sponsorship
Center for Highway Pavement Preservation, Texas Department of Transportation

for

Center for Highway Pavement Preservation
(CHPP)

In cooperation with US Department of Transportation-Research and Innovative Technology Administration (RITA)

July 2021

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Center for Highway Pavement Preservation (CHPP), the U.S. Government and matching sponsor assume no liability for the contents or use thereof.

Technical Report Documentation Page			
1. Report No. CHPP Report-UTA#8-2021	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Segmentation of Highway Networks for Maintenance Operations		5. Report Date June 2021	
		6. Performing Organization Code	
7. Author(s) Moo Yeon Kim and Jorge A. Prozzi		8. Performing Organization Report No.	
9. Performing Organization Name and Address CHPP Center for Highway Pavement Preservation, Tier 1 University Transportation Center Michigan State University, 2857 Jolly Road, Okemos, MI 48864		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No.	
12. Sponsoring Organization Name and Address United States of America Department of Transportation Research and Innovative Technology Administration		13. Type of Report and Period Covered	
		14. Sponsoring Agency Code	
15. Supplementary Notes Report uploaded at http://www.chpp.egr.msu.edu/			
16. Abstract Pavement maintenance and rehabilitation (M&R) is essential for transportation agencies to have a sustainable transportation infrastructure. In maintenance operations, obtaining limits of homogeneous sections is a crucial problem because appropriate segmentation can help yield a more cost-effective M&R plan. To date little attention has been paid to the highway segmentation problem. Only a handful of research studies are available on this topic. Moreover, many of these studies focused on the cumulative difference approach (CDA) method and its modification, which cannot provide statistical inference. Although the segmentation problem is a critical element in pavement maintenance operations, practitioners to date still rely on simple and subjective approaches to obtain manageable sections. Therefore, this research seeks to explain the development of a novel highway segmentation method. The main objective of this study is to propose a Hidden Markov Model-based highway segmentation method using pavement performance data. A secondary objective is to demonstrate the practical application of the proposed method. In this framework, we conducted a thorough literature review regarding research studies focused on the segmentation of highway pavements and we performed an analysis with simulated data and real-world data drawn from TxDOT's most recent pavement condition database, Pavement Analyst (PA). Also, this study explored off-the-shelf tools available for detecting change points and compared the results of implementation with the CDA method. We suggest a segmentation method using Hidden Markov Models (HMMs) as a prospective method for highway management operations. With simulated and real data, the method was tested to demonstrate its benefits as compared to the CDA method. Also, a Bayesian approach to estimate HMM parameters is introduced, which can be a remedy to overcome issues that arose in the maximum likelihood estimation.			
17. Key Words Highway Segmentation; Maintenance and Rehabilitation; Roughness; Rutting; Hidden Markov Models		18. Distribution Statement No restrictions.	
19. Security Classification (of this report) Unclassified.	20. Security Classification (of this page) Unclassified.	21. No. of Pages	22. Price NA

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CHPP: Center of Highway Pavement Preservation
MDOT: Michigan Department of Transportation

Acknowledgments

The authors express their gratitude to the support from the Texas Department of Transportation, in particular to the personnel from the Maintenance Division for their support and interest for this project. We want to acknowledge Dr. Magdy Mikhail, Dr. Jenny Li and Dr. Feng Hong for providing access to data and valuable feedback. We also want to thank Dr. Shannon Crum, Director of TxDOT Research and Technology Implementation Office for providing matching funds through the Texas Pavement Presentation Center.

Executive Summary

The planning and execution of pavement maintenance and rehabilitation (M&R) projects are essential for highway and transportation agencies to manage a sustainable transportation infrastructure system. In maintenance operations, obtaining limits of homogeneous sections is a crucial problem because appropriate segmentation is fundamental for a more efficient and cost-effective M&R plan. To date little attention has been paid to highway segmentation. Only a handful of research studies are available on this topic and there is heavy reliance on empirical approach and personnel experience. Moreover, many of these studies focused on the cumulative difference approach (CDA) method and its modification. This method is deterministic and cannot provide the basis for statistical inferences. Although highway segmentation is a critical element in pavement management and maintenance operations, practitioners, to date, still rely on simple and subjective approaches to obtain manageable sections. Therefore, during this research study, the authors aimed at developing a novel and more objective highway segmentation method. The main objective of this study was to develop and propose a Hidden Markov Model-based highway segmentation method using pavement performance data. A secondary objective was to demonstrate the practical application of the proposed method. In this framework, the authors conducted a thorough literature review reviewing and evaluating previous research studies that focused on the segmentation of highway pavements. Thereafter, the authors performed an analysis with simulated data and real-world data drawn from TxDOT's most recent pavement condition database, Pavement Analyst (PA). In addition, during this research study, off-the-shelf tools available for detecting change points and compared the results of implementation with the CDA method were evaluated and compared. As a result, the authors recommended a segmentation approach using Hidden Markov Models (HMMs) as a prospective method for highway management operations. With simulated and real data, the method was tested to demonstrate its benefits as compared to the CDA method. Finally, a Bayesian-based approach to estimate HMM parameters was introduced, which can be a remedy to overcome issues that arose in the maximum likelihood estimation. The method was used to analyze the results of the application of HMM segmentation to identify M&R project limits with IRI and rut depth differences over time.

Chapter 1. Introduction

1.1 Motivations of Segmentation

Pavement maintenance and rehabilitation (M&R) is essential for transportation agencies to have a sustainable transportation infrastructure. Generally, pavement maintenance consists of various routine and preventive activities such as filling cracks, patching, and applying chip seals, with the primary goal of enhancing the riding experience of drivers and preserving pavement performance by slowing down the damages caused by traffic loads and the environment. Pavement rehabilitation, which includes actions such as overlay and partial to complete reconstruction, focuses on increasing the structural capacity of pavement.

The sheer size of road networks makes M&R a major investment in a transportation system. Accordingly, planning pavement M&R projects is a significant challenge for decision-makers not only because they must plan to spend a massive amount of money but also because they need to determine which road sections need to be treated and when, as well as which treatment should be applied, while taking into account budget limitations. The decision-making process must meet specific goals for maintaining pavement performance while also allocating budgets to maximize cost-effectiveness (Haas et al. 1994).

To increase the efficacy of this complex process, the Texas Department of Transportation (TxDOT) began using the Pavement Management Information System (PMIS) in the early 1990s to support overall pavement management and the related decision-making processes; PMIS was used for storing, retrieving, analyzing and reporting information (Stampley et al. 1995).

PMIS managed data collection in half-mile interval sections. The half-mile section is useful to estimate overall pavement condition and maintenance needs at the network level. At the administrator level, this broad evaluation is helpful for goal setting and budget planning; however, at the DOT district level, the half-mile section data are restrictive, especially for M&R project selection, because projects typically involve longer sections, so it was necessary to aggregate multiple half-mile sections with similar properties to arrive at a workable project length.

In 2016, as TxDOT implemented a new pavement management system—Pavement Analyst (PA)—they began to keep even more detailed pavement performance data, such as roughness and distresses at one-tenth-of-a-mile intervals (Hong et al. 2017). Since the 0.1-mile sections are even shorter than the half-mile segments of PMIS, there remains a need for a method to aggregate the short sections into a manageable segment systemically so that one can plan M&R projects effectively. For that reason, obtaining the limits of homogeneous sections becomes a key problem in pavement management.

Scullion and Smith (1997) presented three options for selecting the limits in such a task:

1. use control sections that were designed and constructed under identical conditions.
2. use the limits proposed by pavement engineers.

3. use the cumulative difference approach (CDA) (AASHTO 1993) to delineate sections based on changes of a particular pavement performance index, such as ride or condition scores (Scullion and Smith 1997).

Although the researchers suggested feasible options to identify the management sections, each approach has its own limitations. The first option involves the control sections, which were created at the time of planning for the construction of a highway. While these sections were most likely homogeneous initially, they are not guaranteed to remain homogeneous after years of being subjected to multiple M&R treatments on different segments at different times within a control section. In practice, M&R projects may be conducted on only partial sections where treatment actions are required due to different distress conditions, traffic loads, or environmental reasons. Also, once some sections within the same control section have had different treatment actions (e.g., chip seal for one section, thick asphalt overlay for another section, and so on), each section might have structural differences such that the rates of performance deterioration differ significantly. Thus, the control sections are not suitable for selecting the limits of homogeneous sections.

In the case of the second option, defining limits using engineering judgment is too demanding for area engineers to cover the large road networks in TxDOT districts. In addition, this process is quite subjective, so that the section limits set by one engineer might not be consistent with those set by another engineer.

As for the last option, although CDA is a straightforward and effective method to divide a highway section into several segments, throughout numerous studies researchers have agreed that CDA has some limitations in finding homogeneous sections. These limitations are discussed in Chapter 2. Since there is no universal method for obtaining homogenous management sections—either in Texas or elsewhere in the U.S.—further investigation is needed to establish a systematic segmentation method with universal criteria to find homogeneous sections appropriately.

Bennett (2004) presented three categories of segmentation methods: fixed-length segmentation, dynamic segmentation, and static segmentation. In the case of fixed-length segmentation, segment limits originate from fixed features and are kept constant over time. The previously mentioned approach of determining segment limits using control sections is one example of fixed-length segmentation. As for the dynamic segmentation, determining the segment's boundaries is based on the homogeneity of pavement sections' attributes, such as ride quality and condition. In this setting, the boundaries vary as the attributes change over time. Static segmentation is similar to dynamic segmentation, except the limits of segments are kept for a specific period to simplify management. For efficient M&R operations, static segmentation based on a three- to five-year span is more advantageous than other segmentation approaches because this approach allows for most of the benefits of dynamic segmentation while maximizing manageability (Bennett 2004).

An appropriate segmentation can yield more cost-effective M&R plans. Figure 1.1 is a good illustration of the advantage of having a proper segmentation. Let us assume a decision-maker uses the median value of the attribute—International Roughness Index (IRI)—to determine which parts of the sections need treatment actions. Then it becomes a task to identify a segment that has a median value that exceeds a predetermined threshold. If one defines a single segment that contains candidate sections as a whole, a maintenance activity must be applied to the whole pavement

section regardless of changes in pavement conditions, i.e., the IRI values across the sections. As shown in the middle of Figure 1.1, because the median of IRI values for the whole section is under the threshold, none of the candidate sections is subject to treatment. This may result in higher costs in the long term since the pavement sections fail to benefit from properly timed maintenance action. Conversely, if one properly identifies two homogeneous segments "A" and "B" as shown in the bottom of Figure 1.1, segment "A" would be treated because it exceeds the median IRI threshold. On the other hand, segment "B" would not be treated because its median is less than the threshold value. In this sense, an appropriate segmentation leads to cost-effective planning by identifying the proper maintenance intervals for those segments demanding maintenance action.

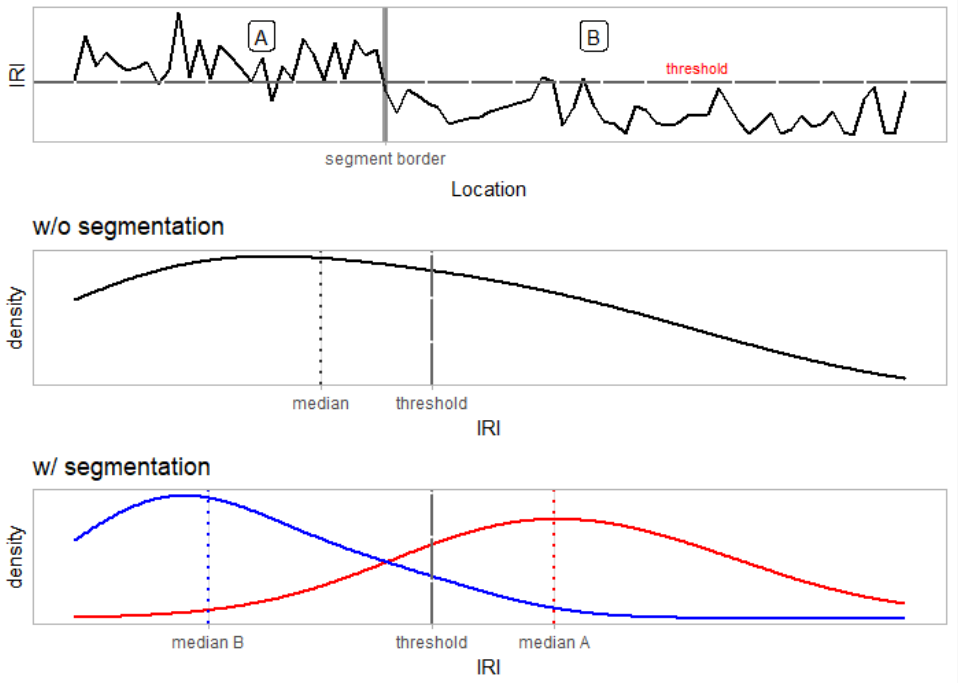

Figure 1.1: The effect of an appropriate segmentation (adapted from Bennett 2004)

As another example, Figure 1.2 demonstrates a situation where a DOT handles discrete data to manage their system as well as take advantage of dynamic segmentation. A DOT might have a project limit based on previous M&R work history on candidate sections. In the case of using the aforementioned fixed-length segmentation or static segmentation strategy, the DOT would keep the historical project limits without choosing to re-evaluate condition ratings for finding updated section limits, as shown in Figure 1.2(a). The historical project limits do not bisect sections with homogeneous pavement conditions, as the pavement sections may experience different traffic loads or have unknown maintenance history. Figure 1.2(b) presents the segment ratings resulting from taking the average of each discrete section's rating. We can observe that some parts in the first segment, which do not require maintenance actions because the overall condition rating does not exceed the threshold, would be treated; however, some sections in the second segment would not be treated even if their conditions indicated the need for treatment. In other words, by not updating segment limits dynamically, a DOT would be failing to apply cost-effective M&R plans.

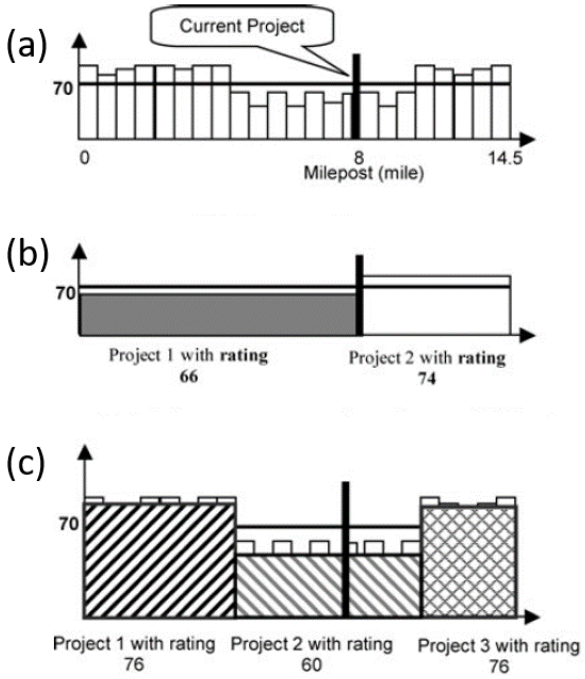

Figure 1.2: The advantage of a dynamic segmentation (adapted from Yang et al. 2009)

Under a better segmentation scheme, the two segments would become three segments with homogeneous pavement conditions, as shown in Figure 1.2(c). The resulting segmentation makes it possible to be cost-effective by planning M&R activities only on sections where treatment is actually needed (Yang et al. 2009).

1.2 Hidden Markov Models

The hidden Markov models (HMM) (Baum et al. 1970) comprise a statistical model that has been significantly successful in modeling time series or spatial data since they were developed. HHM applications are very common, especially in automatic speech recognition (Rabiner 1989), but their use has expanded to other areas because HMMs are a relatively simple yet very powerful means to handle the dynamic behavior of data. HMMs have been applied to a wide range of research fields, such as diverse forms of automated recognition, bioinformatics, environment, finance, biophysics, and ecology (Zucchini et al. 2017).

HMM applications have been used in multiple areas of transportation engineering, such as in traffic prediction (Qi and Ishak 2014; Wang et al. 2015), driving behavior prediction (Li et al. 2016; Zou and Levinson 2006), autonomous driving (Kaplan et al. 2010; Song et al. 2016), and vehicle recognition (Miller et al. 2015). In the pavement management area are a few studies (Kobayashi 2010; Kobayashi et al. 2012; Lethanh and Adey 2012, 2013) related to the prediction of pavement deterioration. However, no study has been found that used HMM for the purpose of highway segmentation.

A Markov chain is a stochastic process that models transitions from one state to another according to a certain probability. HMMs are an expansion of the Markov chain model. Ordinary Markov chain models and HMMs are different in that the state sequence to corresponding observations is known for Markov chain models, whereas, for HMM, the states are unobservable, which means literally hidden, as its name reflects. The observations of HMM are generated from a probabilistic distribution of the hidden states, but the underlying state sequence is a hidden stochastic process (Rabiner 1989). Figure 1.3 illustrates the HMM structure. The state sequence Y_i, which is hidden, follows a transition probability, and each state generates observation X_i from an emission probability. More details will be discussed in Chapter 3.

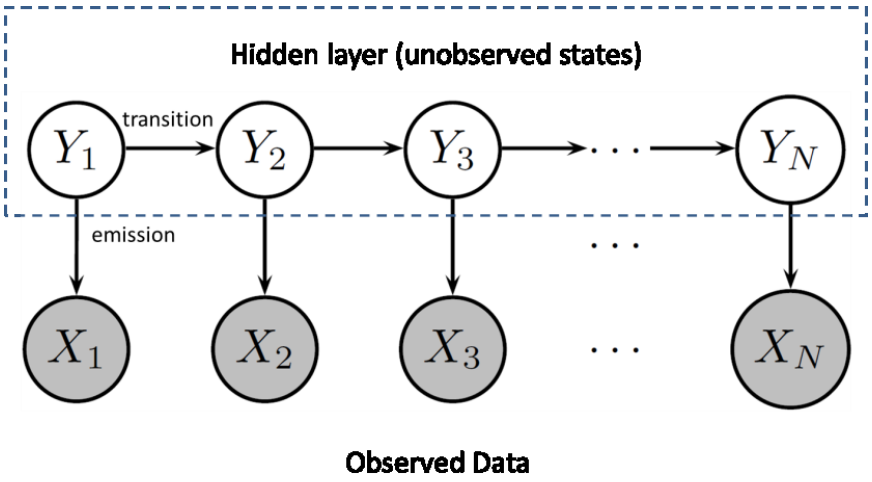

Figure 1.3: Illustration of HMM structure

HMMs are more flexible because they lift the restriction of the observable state sequence; furthermore, they are more realistic because the observations are not assumed to be independent but conditionally independent. Based on an observation sequence, HMMs can provide a solution for the most probable sequence of states, which can be used as a method for segmentation.

1.3 Objectives

To date little attention has been paid to the highway segmentation problem. Only a handful of research studies are available on this topic. Moreover, many of these studies focused on the CDA method and its modification, which cannot provide statistical inferences. Although the segmentation problem is a critical element in pavement maintenance operations, practitioners to date still rely on simple and subjective approaches to obtain manageable sections. Therefore, this research seeks to explain the development of a novel highway segmentation method. The main objective of this study is to propose an HMM-based highway segmentation method using pavement performance data. A secondary objective is to demonstrate the practical application of the proposed method. In this framework, we conducted a thorough literature review and analysis with simulated data and real-world data drawn from TxDOT's most recent pavement condition database, Pavement Analyst (PA). It is hoped that the suggested methodology makes an essential contribution to the field of pavement asset management.

The remainder of this document is organized as follows. Chapter 2 presents the result of a literature review regarding research studies focused on the segmentation of highway pavements. Also, we explore off-the-shelf tools available for detecting change points and compared the results of implementation with the CDA method. In Chapter 3, we suggest a segmentation method using HMMs as a prospective method for highway management operations. With simulated and real data, the method was tested to demonstrate its benefits as compared to the CDA method. In Chapter 4, a Bayesian approach to estimate HMM parameters is introduced. The proposed method can be a remedy to overcome issues that arose in the maximum likelihood estimation presented in the previous chapter. Chapter 5 analyzes the results of the application of HMM segmentation to identify M&R project limits with IRI and rut depth differences over time. Chapter 6 summarizes and concludes this study along with the suggestion of future work.

Chapter 2. Literature Review

2.1 Segmentation Methods for Pavement Management

Through the literature review on segmentation methods for application in pavement management, only a relatively small body of literature was found that is concerned with pavement segmentation studies for developing an approach to identify homogeneous segments. This section presents some of the findings from the literature review.

2.1.1 Cumulative Difference Approach

The cumulative difference approach (CDA) seems to be the most popular method for highway segmentation because the method was introduced in the American Association of State Highway and Transportation Officials (AASHTO) Pavement Design Guide (AASHTO 1993) that is used worldwide. The AASHTO guide states that the CDA method is straightforward and powerful (AASHTO 1993).

Figure 2.1 illustrates the overall concept of the CDA method. In the top portion of Figure 2.1 are three unique constant values, r_1, r_2, and r_3, with three intervals 0 to x_1, x_1 to x_2, and x_2 to x_3, respectively. The cumulative area at x, the solid line in the middle figure, can be calculated as the following integral:

$$A_x = \int_0^{x_1} r_1 \, dx + \int_{x_1}^{x} r_2 \, dx$$

Also, the cumulative area of the average project response, $\overline{A_x}$, depicted by a dashed line in the middle portion of the figure, can be calculated by using total project length L_P and total cumulative area A_T as follows:

$$\overline{r} = \frac{\int_0^{x_1} r_1 \, dx + \int_{x_1}^{x} r_2 \, dx}{L_P} = \frac{A_T}{L_P}$$

Then, the cumulative difference variable Z_x is determined. In the third portion of the figure, Z_x is illustrated as the difference between cumulative areas at x.

$$Z_x = A_x - \overline{A_x}$$

When Z_x is plotted over the length of a project, as illustrated in the bottom part of the figure, the boundary location can be determined at the location where the slope of Z_x changes, for example, from negative to positive or vice versa (AASHTO 1993).

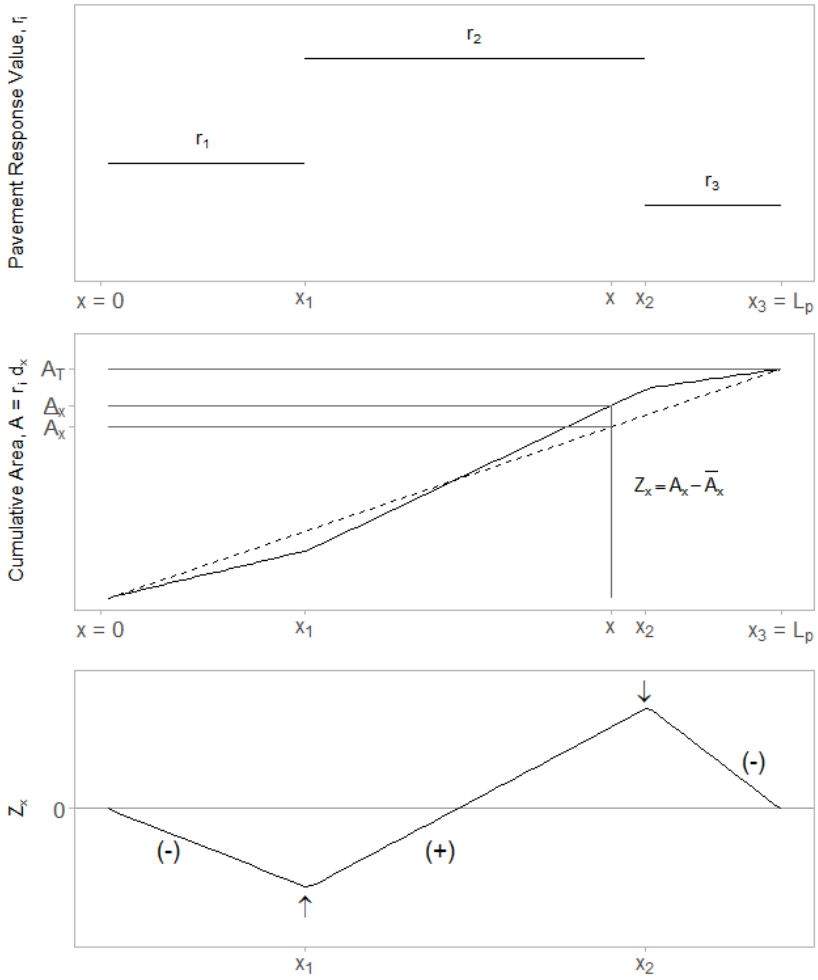

Figure 2.1: Concept of cumulative difference approach (adapted from AASHTO 1993)

Although the CDA method has the advantages of simplicity and applicability for segmentation, various research studies have identified and discussed the limitations of the CDA and suggested modified procedures for CDA or new methods. For example, Misra and Das (2003) presented the limitations of CDA as follows. In the case when more than one homogeneous sections with different mean levels consecutively exist above or below the mean horizontal line, the CDA fails to delineate those sections because the sign of Z_x does not change. Also, they mentioned that CDA has no control over the number of homogenous sections, and minimum section length cannot be chosen by the user. They suggested the classification and regression trees (CART) algorithm (Breiman et al. 1983) as an alternative method.

Divinsky et al. (1997) recommended a modification of the CDA procedure by taking into account statistically homogeneous scatter characteristics, such as standard deviation and range, to overcome the limitation of significant sensitivity to changes in the mean levels of segments.

Ping et al. (1999) introduced a procedure for automated segmentation of pavement rut data using CDA. They developed a program that performs the iterative process with resulting segments from

the previous pass to obtain sub-segments until the program produces the same segments as the previous pass. In the program, two user-specified constraints, such as a minimum segment length and a minimum difference in mean, are incorporated (Ping et al. 1999). Kennedy et al. (2000) conducted the CDA on IRI data with a similar procedure as the study of Ping et al. and found that the use of such constraints helps produce an adequate number of segments. Cafiso and Di Graziano (2012) also developed a similar iterative procedure referred to as MINSSE (Cafiso and Di Graziano 2012) and they compared performances of the CDA to that of the iterative method using two constraints as in the aforementioned study. The study concluded that the CDA method fails to identify some significant changes, but the proposed method detects a majority of change-points successfully.

Thomas (2005) argued that the CDA is primarily a graphical method to detect the homogeneous sections, and it is not suitable for narrowly spaced measurements. Also, the CDA always suggests at least two segments unless all measurements in a given series are identical. Thomas introduced a Bayesian approach that will be presented in the next section.

2.1.2 A Bayesian Approach

Thomas (2003) presented a method to detect a change in the mean, in the variance, and/or in the autocorrelation of a series using a Bayesian approach that allows communicating the existence and possible location of a change-point in terms of probabilities. The author emphasized that the method requires no prior knowledge and distributional assumption. Later, Thomas (2005) introduced a Box-Cox transformation to meet the normality assumption of observations, and a heuristic algorithm to detect multiple change-points to overcome the limitation of at most one change-point algorithm, which means the algorithm can detect none or only one change-point at a time. These two studies were based on his dissertation (Thomas 2001). Detailed statistical proofs are presented in the thesis, so here we introduce the basic concept of his Bayesian approach.

A general Bayesian approach to determine change-point is introduced in Thomas's thesis as follows. A sequence of random variables, $x_1, \dots, x_n$, is divided into subsequences $x_1, \dots, x_r$; $x_{r+1}, \dots, x_n$ by a change-point r, where $1 \leq r < n$. M_0 indicates the model with no change in underlying parameters and its joint density can be expressed as $p(x_1, \dots, x_n|M_0)$. Meanwhile, a model with change in one or more of the parameters at r is denoted as M_r and its joint density is $p(x_1, \dots, x_n|M_r)$. Using Bayes theorem, the posterior probability of a model is the following:

$$p(M_r|x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n|M_r)p(M_r)}{\sum_{all\ r\prime} p\ (x_1, \dots, x_n|M_{r\prime})p(M_{r\prime})} \propto p(x_1, \dots, x_n|M_r)p(M_r)$$

Under this framework, two model comparisons are interesting. One is comparing the models $M_1, \dots, M_{(n-1)}$ to specify that change in parameters occurs at $r = 1, \dots, n-1$. Another is comparing some or all models that have a change at r with M_0 to test whether a change occurs at all. Two models, where change-point at r and s, respectively, can be compared by following a Bayes factor. In general, Bayes factors provide a way of quantifying the evidence in favor of a null hypothesis based on the data.

$$\frac{\frac{p(M_r|x_1,\ldots,x_n)}{p(M_s|x_1,\ldots,x_n)}}{\frac{p(M_r)}{p(M_s)}} = \frac{p(x_1,\ldots,x_n|M_r)}{p(x_1,\ldots,x_n|M_s)} = B_{rs}$$

Comparing the hypothesis of no change versus a change in the series can be done using the following:

$$\frac{1-p(M_0|x_1,\ldots,x_n)}{p(M_0|x_1,\ldots,x_n)} \Big/ \frac{1-p(M_0)}{p(M0)} = \sum_{r=1}^{n-1} B_{r0} \frac{p(M_r)}{1-p(M_0)}$$

A Bayes factor, under the assumption that the numerator and the denominator are identical, can be interpreted using the guidelines given in Table 2.1 for interpreting Bayes factors.

Table 2.1: Guidelines for interpreting Bayes factors (adapted from Jeffreys 1998)

Bayes factor	Interpretation
>100	Decisive evidence for H_A
30–100	Very strong evidence for H_A
10–30	Strong evidence for H_A
3–10	Substantial evidence for H_A
1–3	Anecdotal evidence for H_A
1	No evidence

Although the Bayesian approach proposed is a statistically rigorous method that offers the segmentation results, an iterative process should be involved to obtain multiple change-points. Because the method identifies at most one change-point per iteration, it does not lead to the optimal solution for segmentation.

2.1.3 Fuzzy C-Mean Clustering

Yang et al. (2009) developed a spatial clustering algorithm using Fuzzy C-Mean clustering (FCM). The algorithm minimizes the pavement condition rating variation in each project while taking into account minimum length, costs, barrier, and pavement surface type of a project. In order to accomplish the goal, the algorithm uses two objective functions. One is for minimizing rating variation and the other is for minimizing costs for projects. Together with constraints, the optimal result can be achieved. The optimization process is repeated for the range of cluster numbers. Among the multiple results of optimal number of clusters, the best segmentation would be selected based on the cost objective function.

The left side of Figure 2.2 shows the partitions found when applying the FCM procedure to SR 10 in Georgia, with five, six, seven, and eight clusters. The right side of the figure shows the associated cost for each segmentation. From this, we see that the best segmentation case is when

the number of clusters is equal to six, based on the minimizing cost criterion while satisfying all constraints.

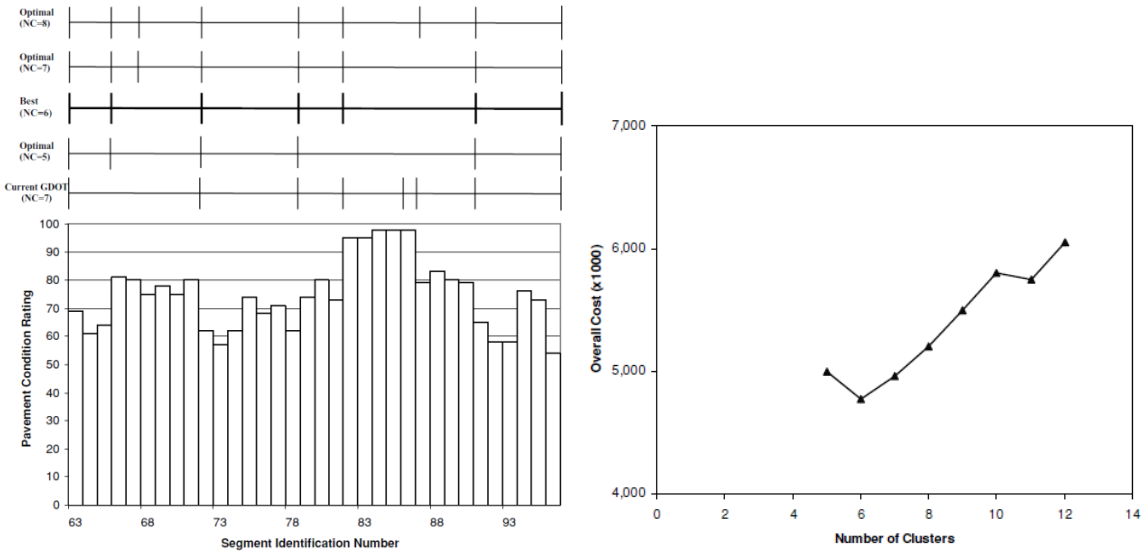

Figure 2.2: An example of optimal solutions for FCM algorithm (reprinted from Yang et al. 2009)

This method offers a way to cluster sections with the optimization scheme that includes construction costs. Thus, it provides a robust solution for combining unit length of individual PMIS sections into multiple homogenous segments. However, the proposed algorithm does not result in a global optimization, and cannot use multiple pavement performance ratings at the same time. Also, the method does not offer any statistical inference.

2.1.4 Wavelet Transform

An algorithm based on wavelet transforms for automated segmentation was presented by Cuhadar (Cuhadar et al. 2002). The properties of wavelet transform, such as de-noising and singularity detection, were used to delineate sections with respect to the pavement condition data. The original data is transformed into a smoother waveform by using de-noising, and then, singularity detection was applied on the smoothed data. The algorithm results in the pavement condition data being grouped into regions that have similar characteristics (Cuhadar et al. 2002). Boroujerdian et al. (2014) also used wavelet theorem for the dynamic segmentation of highway sections. In the study, based on the wavelet theory, the length of high crash road segments was identified by converting accident data to the road response signal. Wavelet transformation seems to outperform the CDA because this approach overcomes the sensitivity to small variability in the data by de-noising. However, the method of singularity detection seems able to identify only sudden changes in the level of univariate data.

2.1.5 CART

Misra and Das (2003) proposed a method using CART (Breiman et al. 1983). The algorithm produces a binary tree through the exhaustive search to find the point that minimizes the sum of

squared errors (SSE). By recursive binary splitting, the original tree is produced as shown in Figure 2.3(a). Once the original tree is produced, based on a set of constraints, such as a minimum section length and a number of sections, the tree is reduced by merging divided sections in the original tree. In Figure 2.3(b), the resulting sub-tree is illustrated, indicating eight delineated sections.

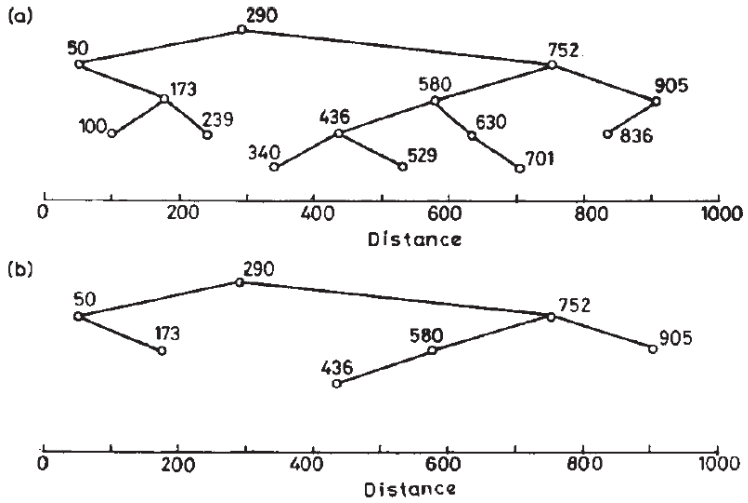

Figure 2.3: Example of the CART result: (a) original tree; (b) sub-tree (reprinted from Misra and Das 2003)

The proposed algorithm provides a simple and fast solution for segmentation without any assumption on the distribution of data. The limitations of CDA are overcome by constraining the minimum section length and choosing the number of segments. Nonetheless, this approach also does not produce the optimal solution because it uses recursive binary trees as approximations.

2.1.6 MINSSE

Cafiso and Di Graziano (2012) introduced the minimum sum of squared error (MINSSE) method. The method relies on finding the minimal SSE of partitions as follows:

$$SSE_k = \sum_{j}^{k+1} \sum_{i \in S_j} (x_i - \overline{x}_{S_j})^2$$

where, k is the number of segments. S_j, a set of element in j^{th} segment.

Once change-points are determined by minimizing the SSE under a given minimum segment length, t-tests are conducted to check if adjacent segments meet the criterion of a minimum difference and those segments are combined if the test fails. The authors compared the MINSSE with the CDA and the Bayesian approach (Thomas 2003) and concluded that their method resulted in similar segmentation to the Bayesian approach, although the method is less complex than the Bayesian approach. Figure 2.4 shows a comparative graph of these three approaches: CDA, Bayesian, and MINSSE using rut data.

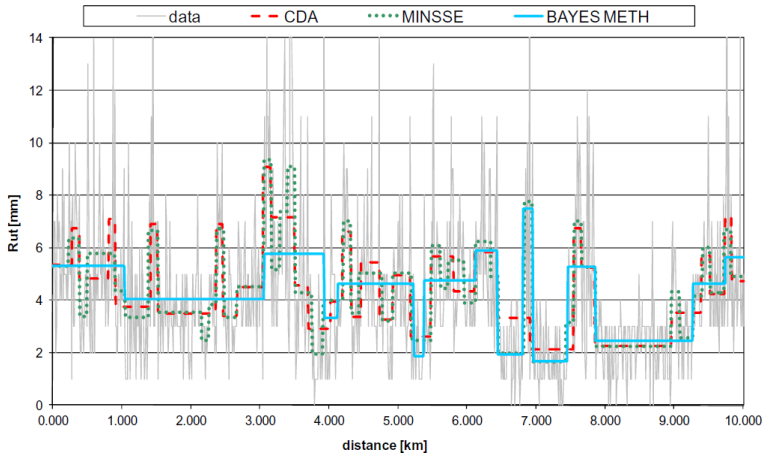

Figure 2.4: Comparison of different segmentation method: CDA, Bayesian, and MINSSE (reprinted from Cafiso and Di Graziano 2012)

2.1.7 Summary and Discussion

The majority of the existing studies on segmentation focused on establishing highway segmentation based on pavement performance data, such as IRI, skid, and rut. Although several safety-related studies for highway segmentation have been conducted, only one of those studies, which is based on wavelet theorem, was mentioned because the analyses on crash count data have different characteristics than analyses on performance data.

There seems to be general agreement on the limitations of the CDA. Some studies modified the CDA to improve it, while other studies suggested alternative methods that showed better performance. The developed methods commonly adopted constraints, such as minimum section length, minimum difference, and number of segments, to overcome the limitations of the original CDA.

Most studies have focused on delineating segments based on the difference of mean levels between segments. Only the Bayesian approach by Thomas (Thomas 2001) has attempted to develop a method that takes into account variance and autocorrelation. In addition, no studies have explored multivariate data. For example, no method can perform segmentation based on rut and skid data simultaneously. Thus, it would be of interest to develop a method that can address some of the aforementioned gaps.

2.2 Change-Point Detection Methods

A change-point analysis is a process to identify the location where a change of the statistical properties of a sequence of time or spatial observation occurs. Various change-point detection techniques have been developed and applied to the fields of finance, bioinformatics, and image processing. This section explored two change-point methods readily available in R (R Core Team 2018) and applied them to real pavement performance data to evaluate the feasibility in pavement management practice. Those methods are the Pruned Exact Linear Time (PELT) algorithm (Killick

et al. 2012), and a Bayesian change-point method by Barry and Hartigan (1993). These two test methods were tested and their performance was compared to that of the CDA algorithm.

2.2.1 PELT Algorithm

PELT is an algorithm for multiple change-point detections based on the maximum likelihood framework (Killick et al. 2012). The likelihood framework for a single change-point detection is a hypothesis test where the null hypothesis and the alternative hypothesis correspond to no change-point and a single change-point, respectively. A test statistic of the likelihood ratio is constructed by the ratio between the maximum log-likelihoods under the null and alternative hypotheses.

$$\lambda = 2\left[\max_{\tau_1}\{\log p(y_{1:\tau_1}|\hat{\theta}_1) + \log p(y_{(\tau_1+1):n}|\hat{\theta}_2\} - \log p(y_{1:n}|\hat{\theta})\right]$$

where $y_{1:n} = (y_1, \dots, y_n)$ is a sequential data; $\tau \in \{1, \dots, n-1\}$ is a change-point. We reject the null hypothesis if the test statistic is greater than a threshold ($\lambda > c$), then we can estimate a change-point, $\hat{\tau}_1$, that maximizes the maximum log-likelihood under the alternative hypothesis. As an extension of the single change-point detection, there can be m change-points, $\tau_{1:m} = (\tau_1, \dots, \tau_m)$. The multiple change-point problem consists of searching a combination of $\tau_{1:m}$ that maximizes the maximum log-likelihood of $\tau_{1:m}$. By adopting a cost function C and a penalty $\beta f(m)$ to prevent over fitting, minimizing the following objective function solves the problem:

$$\sum_{i=1}^{m+1}\left[C(y_{(\tau_{i-1}+1):\tau_i})\right] + \beta f(m)$$

The PELT algorithm provides an exact solution of the segmentation problem efficiently by using dynamic programming and pruning so that computational time increases linearly as data grows.

The approach mentioned above has been implemented in the "changepoint" package (Killick et al. 2016) that also offers other searching algorithms such as binary segmentation and segment neighborhood. These methods are available both for changes in mean and variance by using the assumption of either independent normal distribution or nonparametric cumulative sum (Killick and Eckley 2014).

2.2.2 Bayesian Change-Point Method

The "bcp" package is the implementation of the Bayesian change-point (BCP) procedure by Barry and Hartigan (1993). A probability distribution is provided in the Bayesian procedure instead of specific locations of change-points.

Barry and Hartigan (1993) used the assumption that the observations are independent $N(\mu_i, \sigma^2)$. The assumption could be relaxed by assuming that the observations in different blocks are mutually independent, given partitions and parameters. Thus, the prior distribution of μ_{ij} is drawn from $N(\mu_0, \frac{\sigma^2}{i-j})$, where the block begins at $i+1$ and ends at j. There is a partition $\rho = (U_1, \dots, U_n)$, where $U_i = 1$ if a change occurs at point $i+1$, if not $U_i = 0$. By using the Monte Carlo Markov

Chain (MCMC) method, U_i is drawn from the conditional distribution of U_i given the data and the current partition, and the odds of a change-point at a position can be estimated as follows.

$$\frac{p_i}{1-p_i} = \frac{P(U_i = 1|X, U_j, j \neq i)}{P(U_i = 0|X, U_j, j \neq i)}$$

Although dynamic programming can be used to solve the problem exactly, the solution is not practically feasible due to the complexity of the problem $n(O^3)$. The MCMC implementation of the algorithm reduces the complexity to the order of $n(O^2)$ and provides the posterior distributions of the change-points and the means. The package can be applied to both univariate and multivariate change-point analysis (Erdman and Emerson 2007).

2.3 Case Study: Pavement Performance Data in Texas

2.3.1 Data and Software

The test data used in this analysis was selected from in-service highways in Texas. We considered the length of highway sections and the availability of attribute data extracted from TxDOT's Pavement Management Information System (PMIS). Among available pavement attributes, such as condition score, distress score, ride score, and IRI, distress score was selected for the analysis. The distress score is an index that ranges from 0 to 100 and it is calculated by normalizing the multiplication of present utility values for individual distresses, including cracks, rutting, and patching. Each data point represents a half-mile-long section whose spatial coordinates are given in terms of Texas Reference Markers (TRMs).

The software program used to analyze the data was R (R Core Team 2018). As for the CDA, a CTR team member built his own code based on the original description in AASHTO's design guide (AASHTO 1993) instead of using the modified versions of the CDA proposed by other studies, which involve iterative procedures to automate and enhance the method. The PELT and BCP algorithms were implemented by off-the-shelf packages in R: 'changepoint' (Killick et al. 2016) and 'bcp' (Erdman and Emerson 2007), respectively. In the case of these algorithms, the CTR team conducted a number of trials to be able to control some parameters until the best possible outcomes could be obtained.

2.3.2 Analysis Results

In Figure 2.5, we present one of the analysis results using the three different methods—CDA, PELT, and BCP—with the test data set of a state highway. The y-axis represents the distress score, and the x-axis represents the TRM. The vertical lines indicate segment borders with numbering. Those border lines were horizontally offset to the center of two neighboring points to identify which data points belong to a segment clearly. As a result, CDA, PELT, and BCP yielded 8, 18, and 12 change-points, respectively. In other words, 9, 19, and 13 homogeneous segments were identified. Some change-points were identical across the methods at some locations where noticeable changes occurred; however, each method shows the noteworthy differences in their properties.

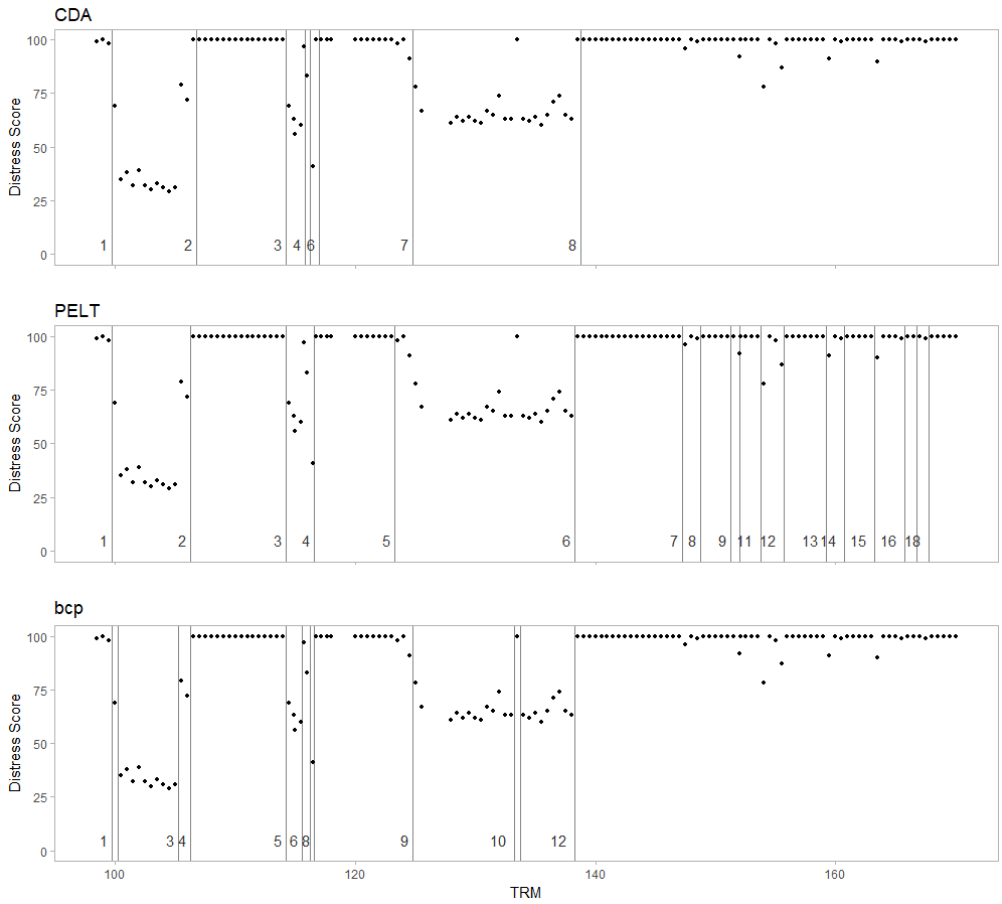

Figure 2.5: Segmentation results with respect to the distress score of three different methods—CDA, PELT, and BCP

As for the CDA, the overall quality of the segmentation result seemed fair. This approach reasonably located the sudden changes in the data despite the fact that no pre- or post-process was carried out. Nonetheless the CDA revealed several limitations. In Figure 2.5, we can observe that the CDA identified a single data point as a segment between border number 4 and 5. When it comes to the project length, a half-mile section is too short to ensure cost-efficiency for an M&R project. For instance, if a decision-maker set the criteria of the minimum length of a project as one mile, considering mobilization costs, a single data point is not worth accepting as an isolated segment. In addition, the segment delineated by border numbers 5 and 6 consists of two data points significantly different each other. Even if the segment exceeds the minimum length criterion, the result suffers from non-homogeneity with respect to the distress score. In contrast, the CDA ignored small variabilities at several locations after border number 8, which is advantageous in terms of securing the minimum length by disregarding noise-like data. Note that there is a remedy for obtaining more consistent results via constraining the minimum length of a segment during post-processing, which is proposed in the modified versions of the CDA. However, we tested only the original algorithm during this study. In addition, we observed absurd borders, such as number 2, 6, and 8, which make the corresponding segments contain a data point with unacceptably high distress score. This issue might be tolerable when each data point indicates a relatively small scale. For instance, if each data point represented 30 ft. (9 m) then a few erroneous points within a

segment would be acceptable as noise at network-level analysis. However, in the case of our test data, each point represents a half-mile (0.8 km) section. Therefore, small errors that occur in the analysis may lead to inefficient M&R planning. The results demonstrate that the CDA failed to produce sensible results for network-level application without further modified processes.

The PELT algorithm also provided reasonable segmentation results. The algorithm requires at least two data points in a segment so that a single point segment did not occur. Furthermore, the outcome did not show any data point belonging to an awkward segment as the result of the CDA. However, as shown in Figure 2.5, border numbers from 7 to 18 show that the algorithm is sensitive to small variability and produced several segments that are not practically meaningful, although the result was obtained after applying a penalty and adjusting parameters to control sensitivity and to prevent overfitting. This can be attributed to the fact that the PELT algorithm can detect changes in variances as well as in means.

As for the BCP, the overall quality of the segmentation was the most promising among the three algorithms, since the resulting segments seemed relatively homogeneous. However, a few single-point segments arose due to the lack of a feature to constrain the minimum length of a segment. The locations of border numbers 4 and 12 were very similar to border numbers 2 and 8 of the CDA result; however, the BCP algorithm did not fail to allocate points around the border to the proper segment, unlike the CDA result.

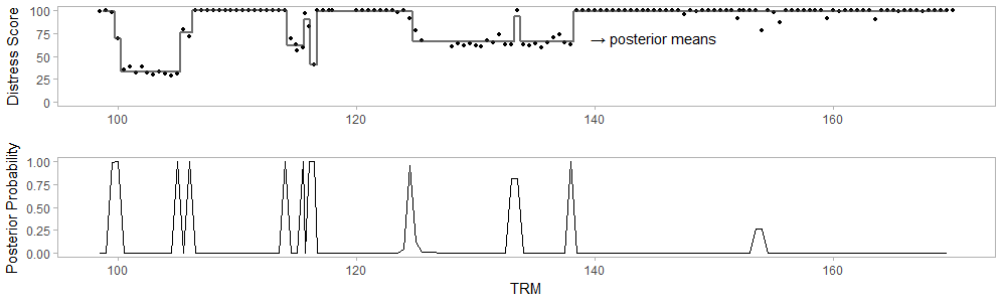

Figure 2.6: Properties of the BCP method: showing posterior probability

Furthermore, thanks to the Bayesian property, the posterior probability of being a change-point can be estimated as shown in Figure 2.6. Therefore, one could capture the uncertainty of the segmentation result. For example, the last bump at around TRM 155 in the posterior probability plot in Figure 2.6 indicates that the probability of being a change-point is less than 50%. Accordingly, one can potentially judge whether or not a change occurs based on the probability. Moreover, posterior means and variances can also be estimated. As Figure 2.6 shows, the posterior means for each segment are indicated by a solid line.

2.4 Conclusions

The analyses reported in this chapter were performed to investigate the current segmentation techniques, thus providing a direction to develop an improved method. Two main aspects were addressed to achieve the objective: a literature review was conducted and a case study was performed.

The majority of the segmentation methods delineate segments by identifying one change-point at once and repeating the algorithm to detect more changes using the divided segments from the previous run. Even though additional adjustments are suggested as a remedy, this type of approach does not result in optimal multiple change-points. Although finding optimal solutions increases computational time, the power of modern computer systems and efficient optimization algorithms make it possible to obtain the global optima. Therefore, the developed methods should have the ability to identify the multiple homogeneous segments at once without losing optimality.

Most studies have focused on delineating segments based on different mean levels of segments. Few studies have attempted to develop a method that takes into account changes in variance and autocorrelation. Pavement performance data might be autocorrelated by its very nature. That is, each observation is not independent but correlates to the adjacent one, so that the performance measure of current section has something to do with that of the next section. Thus, it would be of interest to develop a method that can take into account such correlation. In addition, few studies have explored multivariate data for pavement segmentation. Thus, it would be also interesting to develop a method that, for example, can conduct segmentation based on rut and skid data simultaneously.

Throughout the case study of the CDA and two off-the-shelf packages in R, we evaluated and compared the qualitative performance of each method. Although the overall performances of three methods presented in this study seemed reasonable, the two change-point algorithms produced more reasonable results than the CDA. One benefit of a change-point analysis is that it controls the variability. Also, the change-point algorithms have features to prevent overfitting. Although the PELT and BCP could be implemented conveniently using the packages, establishing a tweaking process—by adjusting the penalty and parameters to obtain desirable segmentation results—is a challenging and subjective task. For both algorithms, the results are highly sensitive to those adjustable user inputs, but visual inspections after multiple runs are the only practical way to optimize the input values.

The PELT algorithm can detect multiple changes in mean and variance but cannot employ multivariate data. As opposed to the PELT, the BCP does not offer variance change detection, but it can handle multivariate data analysis. Due to the nature of the Bayesian approach, the BCP algorithm gives not the location of change-points but the posterior probability of change-points at locations. This property is beneficial in terms of diagnosing the uncertainty of segmentations. However, a post-process is necessary to obtain change-point locations using a threshold value with respect to the posterior probability. This process introduces additional subjectivity to the segmentation results. Also, a potential problem of the BCP is that it would not produce identical results every time it runs because the segmentation results are obtained by the MCMC method. For gaining more consistent and rigorous results over multiple runs of the algorithm, proper MCMC settings are required, including an initialization, a burning, and the number of iterations.

In conclusion, this chapter's comparison should be treated with caution because the case study was conducted using a limited number of data points. Nevertheless, all aspects of the analyses point to the conclusion that each approach has limitations and an improved method for handling those limitations is needed. We suggest the following desirable properties of a segmentation method based on these findings:

- Detect multiple change-points simultaneously;
- Provide optimal or near-optimal solution;
- Detect changes in mean, variance, or autocorrelation;
- Adjust sensitivity in terms of a change in parameters;
- Control the minimum length of a segment;
- Provide a measure of uncertainty; and
- Handle multivariate data.

In Chapter 3, Hidden Markov Models will be introduced as an alternative method for the segmentation that meets the aforementioned desired characteristics.

Chapter 3. Segmentation Method using Hidden Markov Models

This chapter introduces hidden Markov models (HMM) in the context of highway segmentation using performance measures. First of all, this chapter will define the essential components of HMM for pavement management application. This is followed by a review of the three problems to take in applying HMM to actual data. Finally, HMMs are illustrated by examples using simulated and real data, along with a comparison to the results of the CDA method.

3.1 Hidden Markov Models

A Markov model is a stochastic process used to solve problems in which condition states are observable. This process sometimes is too restrictive to model realistic scenarios in which states are not directly observable. For example, when modeling pavement performance, we cannot directly measure finite states of performance (i.e., good, fair, and poor); we can only estimate the performance through observable measures, such as roughness and different types of distresses. HMMs can be useful as an expansion of Markov models to obtain more flexibility. HMMs make it possible for each hidden state to generate observations (e.g., different distress types) according to a different probability distribution for each distress that depends only on a state. These probability distribution functions are referred to as *emission probability*. For example, we can think of a pavement's current performance as a state, and each state produces roughness and distresses following different probability distributions. Figure 3.1 illustrates HMM for an application to pavement segmentation.

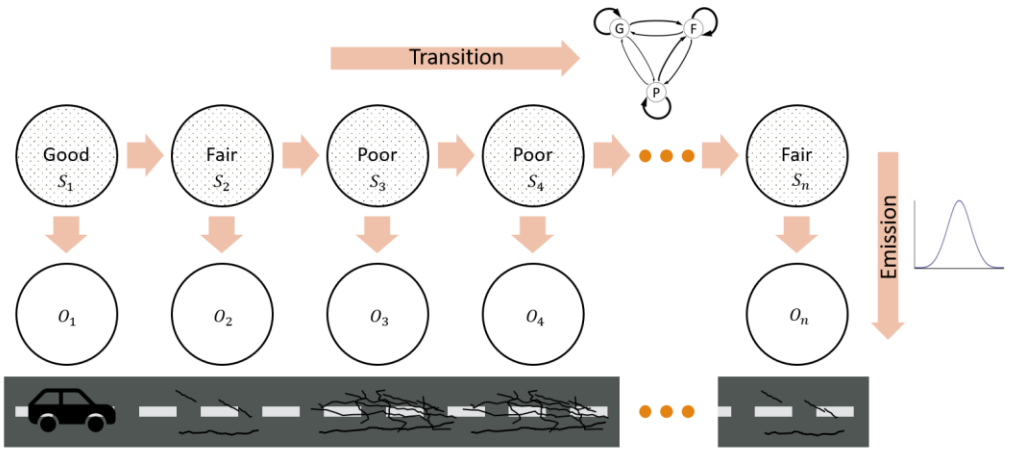

Figure 3.1: HMM with three hidden states and continuous observations

Rabiner (1989) referred to HMM as a doubly stochastic model, as one underlying stochastic process is not observable while the other process produces the sequence of observations. In Figure 3.1, S is a hidden state and O is an observation at a pavement section. Such observations could be indications of roughness or different distress types. The model makes two major assumptions.

First, the current state depends entirely on the previous state, which is a property of a first order Markov chain. In Figure 3.1, the changes between the three hidden states, such as good, fair, and poor, occur by a transition probability following a Markov chain as follows:

$$A = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} 0.867 & 0.098 & 0.035 \\ 0.058 & 0.925 & 0.017 \\ 0.039 & 0.031 & 0.930 \end{pmatrix}$$

where, A is a transition probability matrix for three states. An example transition matrix is presented on the right side. Each element indicates the probability of transition from one state to another. For instance, when we assume that state 1 is good, state 2 is fair, and state 3 is poor condition, p_{11} is the probability of transition from good state to good, p_{12} is the probability of transition from good to fair, and p_{13} is that from good to poor. The sum of $p_{11}, p_{12}, and\ p_{13}$ should equal one to meet the Markov chain rule. In the example, if the current section's state is good, the next section's state is likely good, too, because the relatively low probability that a proximal section would transition from one state to another.

Second, the observations are independent of each other, and depend only on the current state. An observation value (for example, roughness) can be drawn only from an emission distribution that depends on the hidden states, such that observations are spatially independent of previous observations and states. Therefore, a good state might evince less roughness than a fair or poor state, and the resulting roughness at the current section is independent of the roughness at the neighboring sections. This conditional independence of observations allows HMM to provide a more realistic model while retaining a relatively simple structure.

3.1.1 Definitions of HMM Elements

This section defines some essential elements of HMM as they are used in this study. Notations were adopted primarily from Rabiner (1989), with the exception of the substitution of time notation, t, with spatial notation, x, since spatial sequences (rather than time series) are of interest in this study.

Note that the following definitions are based on the HMM with a finite number of discrete observations. HMM with continuous observations will be discussed later in this chapter.

Number of hidden states in the model: N

Hidden states: $S = (S_1, \ldots, S_N)$; state at section x: q_x

Observation sequence: $O = (O_1, \ldots O_X)$

Initial state probability: $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = S_i),\ 1 \leq i \leq N$

Transition probability: $A = \{a_{ij}\}$, where $a_{ij} = P(q_{x+1} = S_j | q_x = S_i),\ 1 \leq i, j \leq N$

Emission probability: $B = \{b_j(O_x)\}$, where $b_j(O_x) = P(O_x | q_x = S_j),\ 1 \leq j \leq N$

Parameter vector of HMM: $\lambda = (\pi, A, B)$

3.1.2 Three Problems of HMM

Applying HMM to actual data involves three fundamental problems: evaluation, decoding, and learning.

Evaluation: The first problem is computing the probability of the observed sequence produced by the model, using the forward and backward algorithm (Baum and Eagon 1967; Baum and Sell 1968). The outcome of this step enables evaluation of the model given the presence of several competing models.

Decoding: The second problem is finding the most probable path, given a set of observations, and thus reveal the hidden part of the model. One approach to completing this step is using the Viterbi algorithm (Viterbi 1967), as it results in the single best state sequence based on an observation sequence.

Learning: The third problem is optimizing the model parameters for the given observed sequence. A maximum likelihood approach can help complete this step, using a special extension of the expectation-maximization (EM) algorithm (Dempster et al. 1977), called the Baum-Welch algorithm (Baum et al. 1970).

Forward Algorithm

To solve the first HMM problem (evaluation), we need to calculate the probability of an observation sequence, O, given the model parameters, λ, i.e., $P(O|\lambda)$. Consider a fixed state sequence $Q = (q_1, \ldots, q_X)$; then the probability of the observation sequence O for the state sequence Q is

$$P(O|Q,\lambda) = \prod_{x=1}^{X} P(O_x|q_x,\lambda) = b_{q_1}(O_1) \cdot b_{q_2}(O_2) \cdots b_{q_X}(O_X)$$

The probability of a state sequence Q can be written as

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{X-1} q_X}$$

Our target probability, $P(O|\lambda)$, is obtained by the summation over all possible state sequences q of the joint probability of O and Q as follows:

$$\begin{aligned} P(O|\lambda) &= \sum_{all\ Q} P(O|Q,\lambda) P(Q|\lambda) \\ &= \sum_{q_1,\ldots,q_X} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{X-1} q_X} b_{q_X}(O_X) \end{aligned}$$

The calculation of the above probability is not feasible because there is a significant number of $all\ Q$. When the length of a sequence is X, then calculations are needed on the order of $2XN^X$. A dynamic programming algorithm, the forward-backward procedure (Baum and Eagon 1967; Baum and Sell 1968), provides an efficient means to obtain the solution.

Consider the joint probability of the partial observation sequence from the beginning to a specific section x and state S_i at section x as the forward variable:

$$\alpha_x(i) = P(O_1 \cdots O_x, q_x = S_i|\lambda)$$

The above variables can be expressed inductively as follows:

Initialization:

$$\alpha_1(j) = \pi_j b_j(O_1), \quad 1 \le j \le N$$

Induction:

$$\alpha_x(j) = \sum_{i=1}^{N} \alpha_{x-1}(i) a_{ij}\, b_j(O_x), \quad 1 \le j \le N,\ 1 \le x \le X$$

Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_X(i)$$

The output needed for the evaluation procedure can be obtained by recursive calculations of the forward variable (i.e., use of the forward algorithm), as shown in the above steps. In this dynamic programming algorithm, a forward variable in the previous iteration is used for the next iteration until the algorithm terminates with the solution.

Viterbi Algorithm

To solve the decoding problem, the Viterbi algorithm (Viterbi 1967) is used to calculate the most probable state sequence Q, given a sequence of observations and the corresponding model parameter λ.

To find the single best state sequence, we need to define the quantity $\delta_x(j)$—that is, the highest probability (the best score) along a single path that accounts for the first x observations and the end state, j, given the model λ.

$$\delta_x(j) = \max_{q_1,\cdots,q_{x-1}} P(q_1 \cdots q_{x-1}, O_1, , O_x, q_x = S_j|\lambda)$$

We can compute this Viterbi path probability recursively by using the previous probability; the backtracking array, ψ, is obtained as

$$\delta_x(j) = \max_{i=1}^{N} \delta_{x-1}(i) a_{ij} b_j(O_x)$$

$$\psi_x(j) = \operatorname{argmax}_{i=1}^{N} \delta_{x-1}(i) a_{ij} b_j(O_x)$$

The Viterbi algorithm is very similar to the forward algorithm. Unlike the summation that is used in the forward algorithm, max is used in the Viterbi algorithm. The algorithm terminates with the best probability P^* and the state at the end of section X, q_X^* for backtracking all path recursively. Thus, we obtain the most probable state sequence as a result.

$$The\ best\ score:\ P^* = \max_{i=1}^{N} \delta_X(i)$$

$$The\ start\ of\ tracking:\ q_X^* = \operatorname{argmax}_{i=1}^{N} \delta_X(i)$$

$$The\ back\ tracking:\ q_x^* = \psi_{x+1}(q_{x+1}^*),\ x = X-1, X-2, \cdots, 1$$

Baum-Welch Algorithm

The third HMM application problem (learning) involves estimating the model parameters, including transition and emission probabilities. The standard algorithm for the solution is the forward-backward algorithm, also called the Baum-Welch algorithm (Baum et al. 1970). This is a special case of the EM algorithm (Dempster et al. 1977).

Firstly, we need to introduce the backward variable, which is the probability of the partial observation sequence from a section $x+1$ to the end, given state S_i at section x and the model parameters λ:

$$\beta_x(i) = P(O_{x+1} \cdots O_X | q_x = S_i, \lambda)$$

Similar to the forward algorithm, the backward algorithm is computed inductively as follows:

Initialization:

$$\beta_X(i) = 1, \quad 1 \le i \le N$$

Induction:

$$\beta_x(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{x+1}) \beta_{x+1}(j), \quad 1 \le i \le N,\ 1 \le x \le X$$

Termination:

$$P(O|\lambda) = \sum_{j=1}^{N} \pi_j b_j(O_1)\beta_1(j)$$

By using both forward and backward probabilities, we now can begin to see how to estimate the transition probability as:

$$\hat{a}_{ij} = \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i}$$

The numerator can be obtained by summation over all section x to the joint probability of being in state i at section x and state j at section $x+1$, given the observation sequence. The joint probability is obtained as:

$$\begin{aligned} \xi_x(i,j) &= P(q_x = S_i, q_{x+1} = S_j|O,\lambda) \\ &= \frac{\alpha_x(i)a_{ij}b_j(O_{x+1})\beta_{x+1}(j)}{\sum_i^N \sum_j^N \alpha_x(i)a_{ij}b_j(O_{x+1})\beta_{x+1}(j)} \end{aligned}$$

The denominator of $\hat{a}_{ij}$ can be calculated by summation over all transitions out of state i as:

$$\hat{a}_{ij} = \frac{\sum_{x=1}^{X-1} \xi_x(i,j)}{\sum_{x=1}^{X-1} \sum_{k=1}^{N} \xi_x(i,k)}$$

Similarly, we can estimate the emission probability (when observations are of discrete nature) as:

$$\hat{b}_j(v_k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j}$$

The numerator is the sum of the following probability over all section length x in which the observation is v_k.

$$\begin{aligned} \gamma_x(j) &= P(q_x = S_j|O,\lambda) \\ &= \frac{P(q_x = S_j, O|\lambda)}{P(O|\lambda)} \\ &= \frac{\alpha_x(j)\beta_x(j)}{P(O|\lambda)} \end{aligned}$$

As for the denominator, we need to sum $gamma_x(j)$ over all sections x.

$$\widehat{b_j}(v_k) = \frac{\sum_{x=1\ s.t. O_x=k}^{X} \gamma_x(j)}{\sum_{x=1}^{X} \gamma_x(j)}$$

In the EM algorithm, we first initialize the transition and emission probabilities, and then calculate $\gamma_x(j)$ and $\xi_x(i,j)$ using the forward and backward probabilities in E step. Next, the HMM parameters are estimated as an M step. These EM steps are iterated to update the parameters until they converge in terms of log-likelihood.

Note that this maximum likelihood approach converges only to a local maximum rather than the global maximum. The initialization of the EM procedure plays a critical role in dealing with this limitation.

3.1.3 Continuous Observation HMM

The previous section presented methods to solve the three problems to effective HMM application, based on the scenario that observations comprise finite discrete values. To apply HMM to the task of pavement segmentation, we have to deal with observations with continuous values. Therefore, we need to consider a different form of the emission probability.

Typically, an emission probability of this setting is assumed to be a Gaussian distribution. A general form of the emission probability is

$$b_j(O) = N(\mu_j, U_j), \quad 1 \le j \le N$$

where, μ_j is mean vector for the j^{th} state; U_j is covariance matrix for the j^{th} state.

In the EM procedure, parameter reestimation can be calculated thusly:

$$\hat{\mu}_j = \frac{\sum_{x=1}^{X} \gamma_x(j) \cdot O_x}{\sum_{x=1}^{X} \gamma_x(j)}$$

$$\widehat{U}_j = \frac{\sum_{x=1}^{X} \gamma_x(j) \cdot (O_x - \mu_j)(O_x - \mu_j)'}{\sum_{x=1}^{X} \gamma_x(j)}$$

where $\gamma_x(j)$, the probability of being in state j at section x, given observation O_x

$$\gamma_x(j) = \frac{\alpha_x(j)\beta_x(j)}{\sum_{j=1}^{N} \alpha_x(j)\beta_x(j)}$$

Hence, in order to estimate HMM with L number of observable variables, the following L-dimensional multivariate Gaussian distribution can be an emission probability:

$$b_j(O) = \frac{1}{(2\pi)^{L/2}|U_j|^{1/2}} \exp\left(-\frac{1}{2}(O-\mu_j)'\, U_j^{-1}\, (O-\mu_j)\right)$$

3.2 Simulation Study

The HMM method for segmentation was analyzed using simulated data, and the performance assessed by comparing it to the CDA method (AASHTO 1993). For HMM implementation, the MATLAB toolbox capable of analyzing Gaussian HMM, written by Kevin Murphy (Murphy 2005), was used along with R for reporting results. For the CDA method, we used the same code written in R based on AASHTO (1993), as demonstrated in the previous chapter.

A segmentation procedure using HMM was conducted as follows. Firstly, with a known number of states, the HMM parameters were estimated by the learning procedure as described in Section 3.1.2, using the EM algorithm—also known as Baum-Welch algorithm (Baum et al. 1970). Once the EM procedure converges, as a next step, the most probable state sequence is calculated by the Viterbi algorithm (A. Viterbi 1967) using the estimated parameters.

3.2.1 Local Variance

To evaluate the performance of HMM with respect to a local variance, which is the variability within a state, we generated two data sets through a three-state HMM with state-dependent Gaussian distributions in which variances vary from low to high. In total, a thousand data points were generated per each data set. The states of two data sets were randomly switched across the data by the same transition probability, but different Gaussian distributions determined local variances.

Figure 3.2 shows the results of CDA and HMM using the low local variance data. The mean levels of each state are obviously distinctive despite a small amount of noise within a segment. In the CDA result (top of Figure 3.2), a red vertical line indicates a segment border where state transition potentially happens.

On the bottom of Figure 3.2, the HMM result is shown with mean levels and two standard deviations from the mean of each state as a line and shade area, respectively. This is one of the benefits of using the HMM method. It produces not only the boundaries of segments but also extra information regarding the distribution of a segment.

As a result, both methods seemed to reasonably delineate the underlying states. However, the CDA result did miss several change-points where state transitions happened between the highest and medium levels. No such error occurred in the HMM results for the low local variance case.

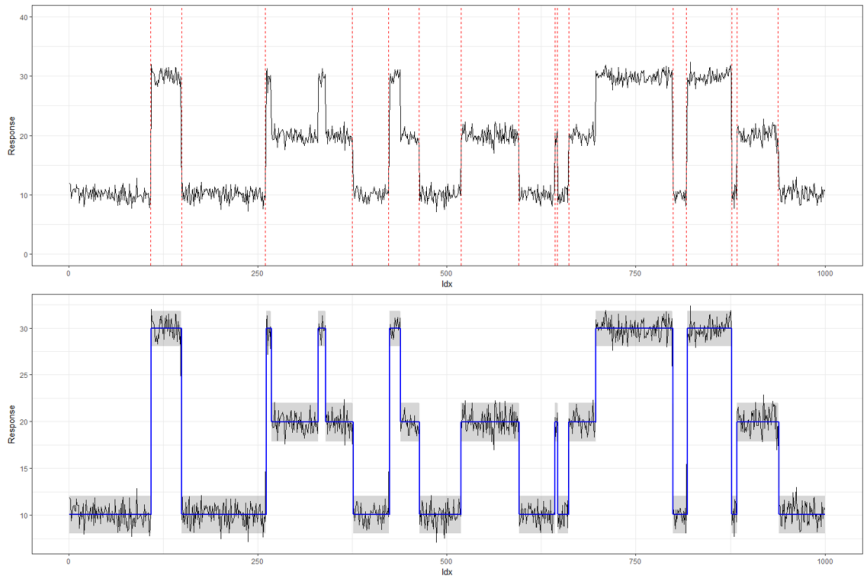

Figure 3.2: Segmentation results of the low local variance data from CDA (top) and HMM (bottom)

Figure 3.3 shows the results of the two methods tested using other data. This data set has the same state transition as the low local variance data but has higher local variances within a segment. Nonetheless, the changes in mean levels are still readily apparent; the result from the CDA method demonstrates that the method is so sensitive to the local variance that an excessive number of segments were detected incorrectly. In contrast, the HMM approach correctly conducted the segmentation without any significant error.

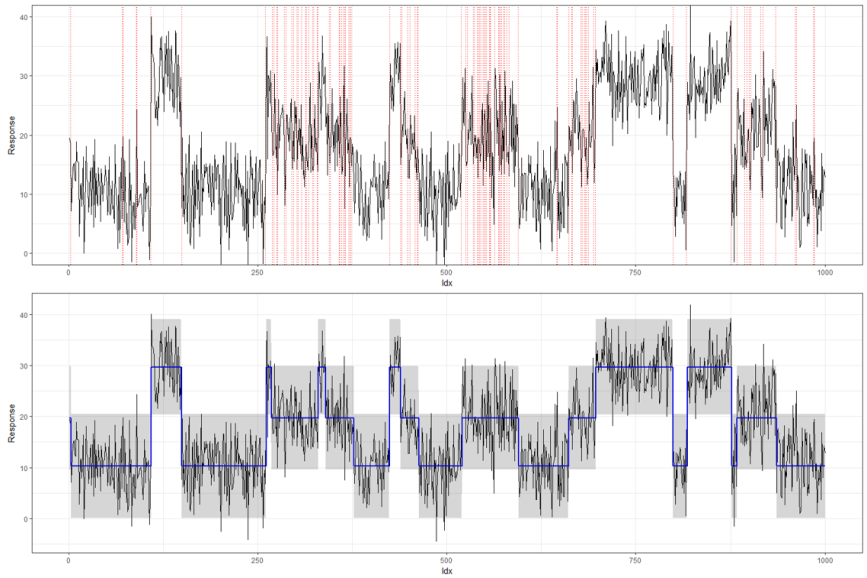

Figure 3.3: Segmentation results of the high local variance data from CDA (top) and HMM (bottom)

In the previous chapter, it was mentioned that one limitation of the CDA method is the high sensitivity to a local variance; the analysis with the simulated data verifies this limitation. Even in

such a trivial case with low local variance, the CDA method could not identify the true segments, missing some change-points. In the case of the high local variance, the opposite result arose. Too many segments were detected by the CDA method. In sum, the CDA method does not have a control to adjust this sensitivity. The only remedy is an iterative procedure to split or merge the resulting segments until the best fit is found. On the other hand, the HMM approach provided correct segmentation regardless of the local variance level in the data.

3.2.2 Variance Detection

Another comparison between the CDA and the HMM approaches was conducted to confirm that the HMM method has the ability to detect variance changes between states while the CDA does not. Another thousand synthetic data points were generated from three Gaussian distributions with identical means and different variances.

As shown in the top of Figure 3.4, since the CDA method is not capable of detecting variance changes at all, that method produced a significant number of change-points.

Meanwhile, the HMM could perfectly find the true segments, as the bottom of Figure 3.4 illustrates. As expected, the HMM method could provide the estimated means and variances of states.

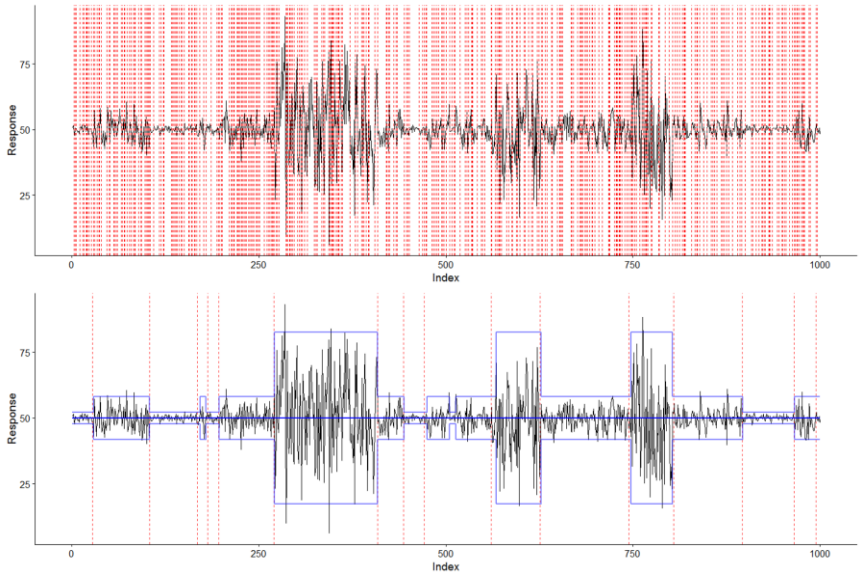

Figure 3.4: Segmentation results of the variance-change data from CDA (top) and HMM (bottom)

There is no remedy for the CDA method to detect variance changes because it takes into account only the cumulative changes in values. However, HMM can detect variance change as well as the mean level changes. Hence, one can identify the underlying states differentiated in terms of mean and variance of observations.

3.2.3 Multivariate Analysis

Using a multivariate Gaussian distribution as an emission probability of HMM, we can conduct a segmentation that involves multiple variables at once. This can be a solution for the analysis to obtain highway segments by using different measures—for example, roughness, rut depth, and cracking all at once.

As shown in Figure 3.5, three series of data were generated. Two data series were created using the same hidden state sequence: one with relatively high means and variables and another with relatively low means and variables. A series of random noise was created to test whether HMM is rigorous against noise.

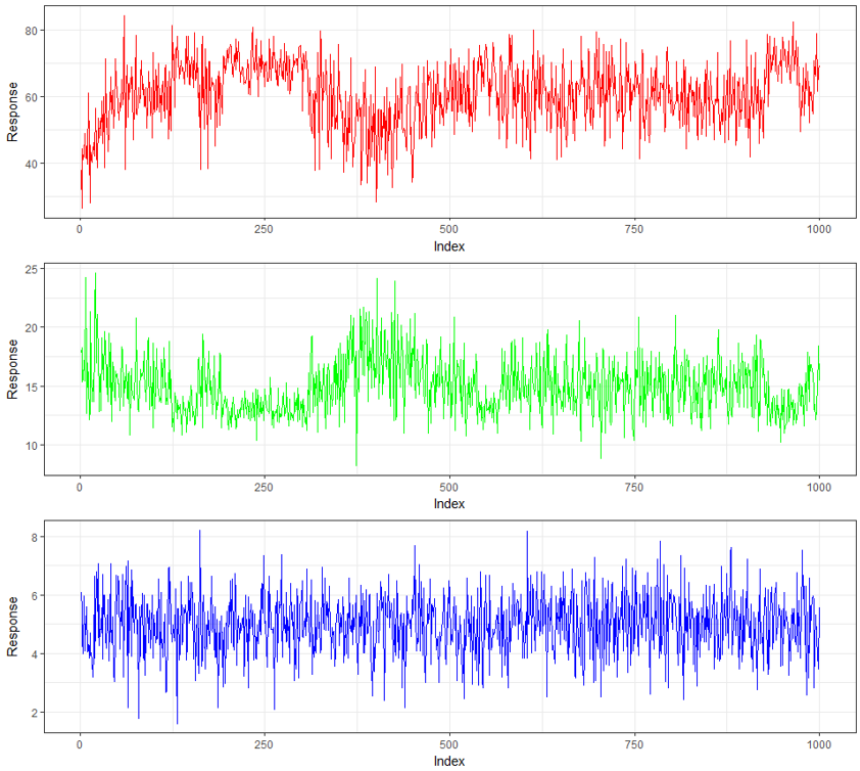

Figure 3.5: Simulated data with three states for testing multivariate case: high mean and variance (top), low mean and variance (middle), random noise (bottom)

Figure 3.6 displays the data on the same scale and the result from the HMM with state-dependent multivariate Gaussian. The segment borders are not easily located with a visual inspection. By using three data series together (although one series with random noise would not help at all), it could be possible to detect the change-points identical to the ground truth, with the exception of one state at the end of data.

Figure 3.6: Segmentation results of the multivariate data: true segmentation (top) and HMM segmentation (bottom)

3.3 Application to Real Pavement Data

As an example of a real-world application for the HMM segmentation, following are the results of an analysis of one of the highways in the Austin District of TxDOT, State Highway 27 (SH0027 K). Using IRI values from 2017, which were recorded in a 0.1-mile sections across the highway, univariate analysis using CDA and HMM were conducted as previously done with the simulated data.

3.3.1 Comparison with the CDA Method

Firstly, Figure 3.7 shows the results obtained from the CDA method when used on the real data. The solid black line indicates IRI measurements along the x-axis—distance from origin (or DFO, as shown in the figure)—in a mile. The vertical red lines are segment borders resulting from the CDA segmentation.

Some resulting segments were reasonably identified, especially when sudden changes in IRI arose. Nevertheless, as noted in the simulation study covered in the previous sections, the CDA method is prone to deliver a significant number of segments due to its sensitivity to the local variance. Therefore, Figure 3.7 identifies many segment borders that have little practical use because of insufficient length.

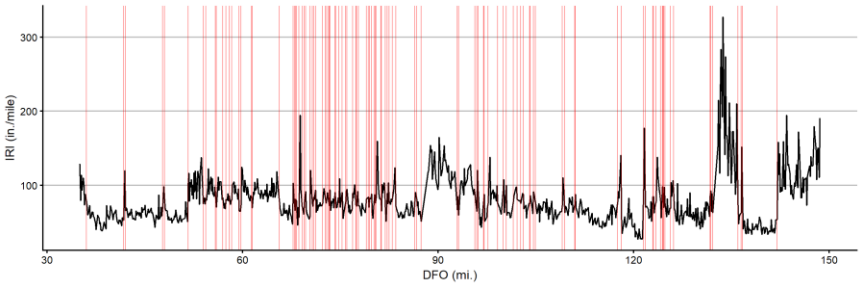

Figure 3.7: Result of segmentation on the real data from CDA

Even though the CDA method does not yield means and variances of each segment detected, for comparison purposes, the means of segments were calculated from the CDA result. In Figure 3.8, the two approaches are compared, and the red and blue lines indicate the mean of CDA and HMM segment, respectively.

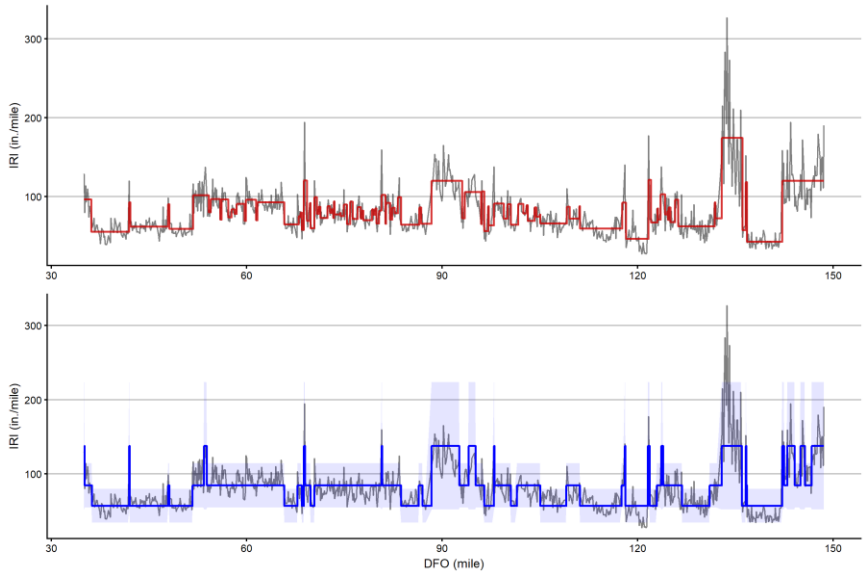

Figure 3.8: Comparison of the segmentation results of the real data: CDA (top) and HMM (bottom)

Qualitatively speaking, the CDA results are not showing consistency—in some regions with high variability, overfitting seems to occur, but in other regions underfitting occurs as well. Hence, evaluating the overall quality of the segmentation is difficult.

The bottom of Figure 3.8 presents the HMM segmentation result. Three states were defined for the model, allowing observation of three different means and the variances dependent on the states. The general quality of the HMM segmentation seems better compared to that of the CDA segmentation, although abrupt jumps in IRI value negatively affect the results, producing many short segments.

As the HMM approach provides the estimation of the distribution parameters, we can quantitatively evaluate the segmentation result. The three states are normally distributed by the assumption of the model's emission probability, as Figure 3.9 shows. State 3 has the highest mean

and variance and state 2 has the lowest mean and variance, as Figure 3.8 shows. Each state's parameters and frequency are presented in Table 3.1.

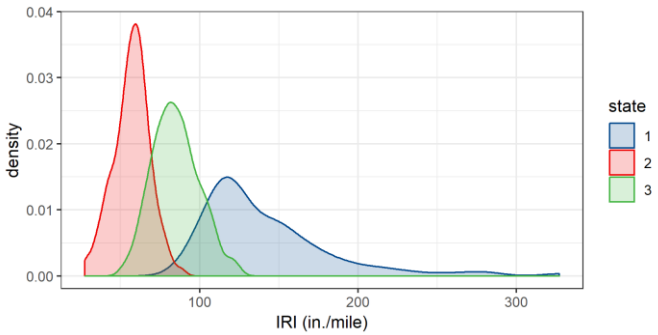

Figure 3.9: Distribution of each state from the HMM segmentation

Table 3.1: The estimated parameters of each state from the HMM segmentation

State	Freq.	IRI_Mean	IRI_SD
1	112	137.7	42.9
2	359	56.9	11.4
3	370	84.5	14.9

3.3.2 Limitations of the HMM Method

Number of States

One of the limitations of the HMM method is that the number of states should be predetermined to learn the parameters. In a real-world situation, we typically do not know the number of hidden states. Figure 3.10 demonstrates the segmentation results based on five states and ten states.

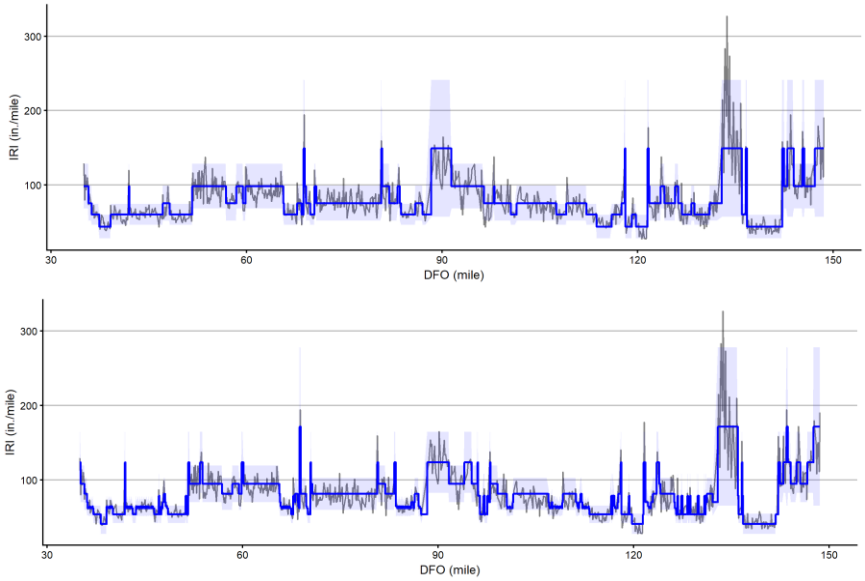

Figure 3.10: HMM segmentation results with 5 states (top) and 10 states (bottom)

As the number of states increases, the model's accuracy increases, fitting the data better. In other words, the likelihood becomes higher as the number of states increases. When it comes to the maximum likelihood strategy, a model with a higher likelihood can be considered better. However, as shown in the case of 10 states in Figure 3.10, an inadequately large number of states might cause overfitting so that we cannot obtain the practically meaningful length of segments as a result. To avoid the problem, we need to select a proper number of states by criteria. The methods will be discussed in Chapter 5.

Local Maximum

The EM algorithm causes another critical limitation of the HMM method in the estimation procedure of model parameters. The EM method cannot guarantee the global maximum in terms of log-likelihood. The process obtains only a local maximum, and the accuracy of the results is dependent upon the initialization of parameters.

Figure 3.11 presents one of the segmentation results from the same three states HMM. The top figure oddly seems to indicate that only two states exist even though three states were predetermined. The bottom figure shows the reason why this happened. The distributions of states 2 and 3 overlap each other as if they are one. This result is from one of the local maximums in the iterations of the EM algorithm. Therefore, the log-likelihood of this solution is less than that of the model presented in Figure 3.8.

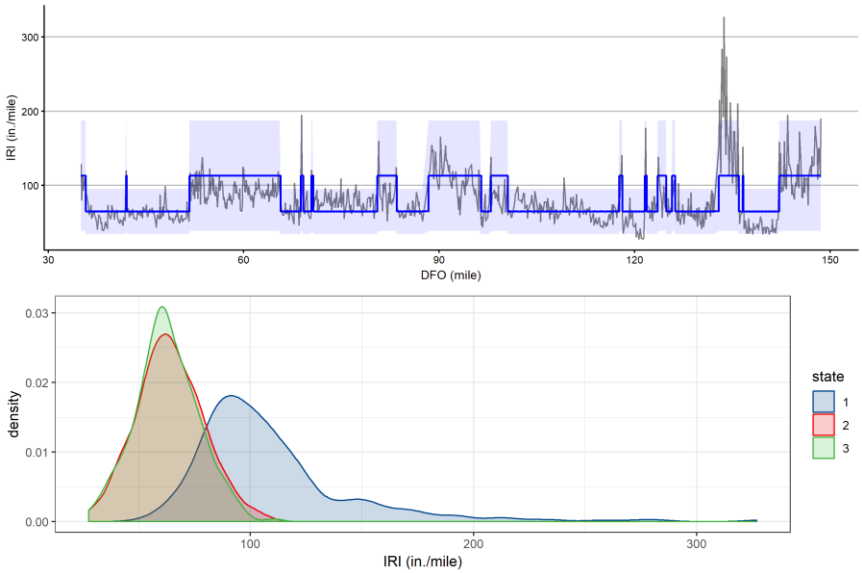

Figure 3.11: HMM segmentation with a local maximum issue: segmentation result (top); state distribution (bottom)

To resolve the issue, we need to estimate the HMM parameter multiple times until it reaches near-global maximum. Again, Chapter 5 will provide more detail on this issue.

3.4 Summary and Discussion

This chapter introduced use of HMM for the task of highway segmentation. HMM consists of two layers. The hidden layer is a sequence of unobserved states that explain another layer, the observation sequence.

Highway segmentation using HMM can be achieved by estimating HMM parameters, such as transition probability and emission probability, through the EM algorithm. Then the Viterbi algorithm can be used to obtain the most probable state sequence, which can identify segments.

A comparison between the CDA and the HMM approaches to both simulated and real data sets found that the HMM method is more advantageous, as the HMM approach is rigorous to a local variance, can detect variance changes as well as mean changes, and enables multivariate analysis.

It cannot be concluded that CDA is always worse than HMM because CDA might present more advantages when used with other data sets. However, there is no control for adjusting the behavior of the CDA method in finding segments. Hence, HMM can generally produce better results regardless of data.

Some limitations of using the HMM method were identified when applying HMM to real pavement data. One such limitation is the problem of remaining stuck in a local maximum during the EM estimation. To overcome this limitation and also to prevent short-length segments, which are not practical, we can use an alternative approach to estimate HMM parameters using a Bayesian framework, as will be presented in the next chapter.

Chapter 4. Hidden Markov Models with a Bayesian Inference

4.1 Introduction

Chapter 3 presented a method using the Maximum Likelihood Estimation (MLE) approach for the estimation of HMM parameters. This chapter will discuss a different approach using Bayesian inference to estimate HMM parameters.

In the MLE approach, a set of parameters $\hat{\lambda}$ is estimated by maximizing the likelihood L, which is

$$L(\lambda) = p(O|\lambda)$$

$$\hat{\lambda} = argmax_{\lambda} L(\lambda)$$

The EM algorithm can be used to find the HMM parameters that maximize the likelihood function, which offers a point estimate. Meanwhile, a Bayesian approach uses Bayes theorem, which factorizes the posterior distribution into the likelihood and the prior as follows:

$$p(\lambda|O) \propto p(O|\lambda)p(\lambda)$$

Intuitively, by using a known parameter (prior) before observations are made, the Bayes theorem makes it possible to update the parameter to a new distribution based on the observations (posterior). Because the posterior distribution often cannot be solved analytically, sampling using a Markov Chain Monte Carlo (MCMC) algorithm can be employed to estimate the posterior distribution. Unlike the MLE approach, which provides a point estimate, a Bayesian estimator does offer probabilities of the parameters. Also, the existence of prior distribution can be beneficial, as it offers preferences for each parameter.

The primary goal of this chapter is to suggest a Bayesian approach to estimate HMM parameters. Firstly, the proposed method will be explained, followed by examples using simulated and real data. Then some characteristics of the introduced method that can be beneficial to the highway segmentation application will be discussed.

4.2 Method

4.2.1 Markov Chain Monte Carlo (MCMC)

MCMC is a sampling method driven by computers. The MCMC method makes it possible to characterize a distribution through random sampling from distributions without knowing the mathematical properties of them. MCMC comprises two components: the Monte Carlo approach and the Markov chain. A Monte Carlo approach is beneficial when random sampling from a distribution is easy enough to allow estimation of the properties of a distribution without analytical solutions. A Markov chain plays a role in a sampling sequence in that the next random samples are drawn from the current random samples. Further, the next samples depend only on the current ones—especially for a 1^{st} order Markov chain (van Ravenzwaaij et al. 2018).

4.2.2 Gibbs Sampler

The Gibbs sampler (Geman and Geman 1984) is an MCMC algorithm. This technique generates random variables from a distribution indirectly without calculating the density (Casella and George 1992). The Gibbs sampler can be illustrated through this example application: obtaining a marginal density of a joint density $f(x, y_1, \ldots, y_p)$:

$$f(x) = \int \cdots \int f(x, y_1, \ldots, y_p) dy_1 \cdots dy_p$$

When integrations are difficult or infeasible, the Gibbs sampler allows us to generate samples $X_1, \ldots, X_m \sim f(x)$ without calculating $f(x)$. Thus, with a sufficient number of samplings, we can estimate the desired density.

For instance, when there is a two-variable case with a pair of random variables (X, Y), the Gibbs sampler draws samples from the conditional distributions $f(x|y)$ and $f(y|x)$ iteratively in the following manner:

$$Y_0, X_0, Y_1, X_1, \ldots Y_k, X_k$$

$$X_j \sim f(x|Y_j = y_j)$$

$$Y_{j+1} \sim f(y|X_j = x_j)$$

The Gibbs sampler draws samples for each parameter from the conditional distribution of the parameter. Thus, when we know the full conditional distributions, sampling using the Gibbs sampler is feasible.

In the case of the HMM estimation, the Gibbs sampler can be divided into a few categories. Firstly, two sampling methods can be used for a state sequence: the pointwise and the blocked sampler. The pointwise sampler resamples a single state q_x at a time, while the blocked sampler resamples a whole state sequence, $q_1, \ldots, q_X$, at a time by implementing dynamic programming such as the forward-backward algorithm. Secondly are the explicit and the collapsed sampler. The explicit sampler samples the HMM parameters explicitly together with states, while a collapsed sampler resamples only the states by integrating out the HMM parameters (Gao and Johnson 2008).

In the course of implementing the Gibbs sampler, each pair of the Gibbs sampler categories, including a pointwise-explicit sampler, a pointwise-collapsed sampler, a blocked-explicit sampler, and a blocked-collapsed sampler, were examined by applying the methods to simulated and real data. The blocked-explicit sampler was selected for use in this study since the blocked-explicit sampler converges faster than a pointwise one. The Gao and Johnson study (2008) reported that the blocked-explicit sampler leads to faster convergence than the pointwise and collapsed ones. In the next section, a method to implement the blocked sampler, the Forward-Filtering Backward-Sampling algorithm, will be introduced.

4.2.3 Forward-Filtering Backward-Sampling (FFBS) Algorithm

This study's Gibbs sampling procedure implemented the FFBS algorithm (Chib 1996; Frühwirth-Schnatter 1994). This blocked sampling method helps MCMC mixes more rapidly than its competitor, the pointwise sampler (Scott 2002). FFBS outputs an independent posterior sample of the state sequence, given a sequence of observations and parameters. The algorithm utilizes the induction of the forward variable and backward variable, α and β, from the forward-backward algorithm introduced in Chapter 3.

The FFBS will be briefly described using the same definition of the HMM elements in Chapter 3. The description of the study from Rao and Teh (2013) is adopted as a reference.

Define $\alpha_x(i) = p(O_1 \cdots O_x, q_x = S_i)$; then we have the following recursion,

$$\alpha_x(j) = \sum_{i=1}^{N} \alpha_{x-1}(i) L_x(i) a_{ij}$$

where, $L_x(i) = p(O_x | q_x = i)$, a likelihood of an observation at a section, is given a state.

Throughout the forward pass for $x = 1 \rightarrow X$, we obtain a vector:

$$\beta_X(i) = L_X(i)\alpha_X(i) \propto p(q_X = i|O)$$

Hence, q_X can be sampled from β_X. Next, by the backward pass for $x = T - 1 \rightarrow 1$,

$$p(q_x = i|q_{x+1} = j, O) \propto \beta_x(i) = \alpha_x(i) a_{ij} L_x(i)$$

we can sample a state sequence $(q_1, \ldots, q_{X-1})$ sussecively from $\beta_x(i)$.

4.2.4 MCMC Procedure for HMM Segmentation

In this study, the Gibbs sampler with a blocked-explicit approach is implemented by alternating between sampling a state sequence and sampling HMM parameters from their respective full conditional distributions. The sequence of the Gibbs sampler is as follows:

1) Update a state sequence: sample a state sequence based on observations and parameters using FFBS

2) Update parameters: sample parameters based on observation and the state sequence previously sampled

As explained in the previous section, we first obtain a whole state sequence through the FFBS algorithm with initial HMM parameters and an observation sequence. Simply, a state sequence can be drawn from a conditional distribution, given an observation and the HMM parameters as:

$$q_1, \ldots, q_X \sim p(q_1, \ldots, q_X | O, \lambda)$$

Secondly, to update the HMM parameters, we need to draw sample parameters from a conditional distribution, given a state sequence and an observation as follows:

$$\lambda \sim p(\lambda|q_1, \ldots, q_X, O)$$

Here, two HMM parameters, the transition parameter A and the emission parameter B, should be sampled separately:

$$A \sim p(A|q_1, \ldots, q_X, O, B)$$

$$B \sim p(B|q_1, \ldots, q_X, O, A)$$

In order to draw samples from each parameter's conditional distribution, conjugate priors must be used. A conjugate prior for the likelihood function allows the posterior distribution to be the same distribution family as the prior. Therefore, with a known likelihood distribution, we can obtain the posterior distribution by choosing a proper conjugate prior. Then, sampling from the posterior distribution becomes possible. Table 4.1 shows the distributions of likelihood and conjugate priors to each HMM parameter.

Table 4.1: Distribution of likelihood and conjugate prior for each HMM parameter

Parameter	Likelihood	Conjugate Prior
Transition Parameter A	Multinomial	Dirichlet
Emission Parameter B	Normal (known variance)	Normal
Emission Parameter B	Normal (known mean)	Inverse Gamma

As for the transition parameter, the likelihood function follows a multinomial distribution, for which conjugate prior is a Dirichlet distribution. In the simple form, the transition parameter and its conjugate distribution can be expressed as follows:

$$q_x|q_{x-1} = q \sim Multi(A)$$

$$A|\alpha \sim Dir(\alpha)$$

where, a hyperparameter α is a fixed uniform Dirichlet prior that controls the sparsity of the state-to-state transition probabilities.

Now, the posterior transition parameter can be drawn from a Dirichlet distribution as:

$$A \sim Dir(\alpha + c)$$

where, c is the state-to-state transition counts—for instance, counting all leaving state $q_{x-1} = S_j$ to the current state $q_x = S_i$.

For the emission parameter, since a normal emission probability is assumed as the likelihood, we need to estimate a mean and variance. Here, the mean and variance can be sampled separately from a normal distribution with known variance and a normal distribution with known mean. Each case has different conjugate priors, a normal distribution, and an inverse gamma distribution, respectively. Therefore, the updated mean and variance from the corresponding posterior distributions can be drawn alternately from the normal distribution and inverse gamma distribution as follows:

$$B_\mu \sim N\left(\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right)$$

where, μ_0 and σ_0^2 are hyperparameters of the normal prior, n is the number of a specific state, x is the observation corresponding to a state.

$$B_{\sigma^2} \sim \Gamma^{-1}\left(\alpha+\frac{n}{2}, \beta+\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}\right)$$

where, α and β are hyperparameters of the inverse gamma prior.

All in all, the Gibbs sampler is conducted with the following order iteratively:

1) update a state sequence $q_1, \ldots, q_X$
2) update the transition probability A
3) update the mean of the emission probability B_μ
4) update the variance of the emission probability B_{σ^2}
5) iterate 1) through 4)

4.3 Examples

The HMM estimation using the aforementioned Gibbs sampler approach was implemented by writing a Python code. Meanwhile, data generation, processing, and result presentation were done using R. The example results from generated and real data are presented in this section.

4.3.1 Generated Data

Often, a numerical simulation is used to verify stochastic models. In order to compare the result from the EM estimation, the same generated data, for which variances are different while means are identical, were used. Figure 4.1 shows the generated data and the true segment limits represented by red vertical lines. These limits are considered the original segments for comparison with the estimated result of the HMM model. The Gibbs sampler proposed in the previous section was applied to the data.

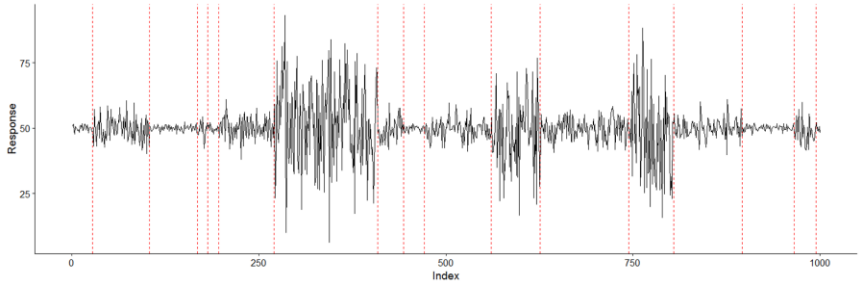

Figure 4.1: Simulated data with different variances for three states

Figure 4.2 presents the results of the MCMC procedure. The left portion of Figure 4.2 illustrates the sampling outputs of variances for each state. It can be observed that the variances rapidly converge after a relatively small number of iterations.

In the right portion of Figure 4.2, the histogram of variances is plotted. As a result of using the Bayesian approach, we can obtain a distribution of parameters instead of a point estimate. In the original data, each state has standard deviation values 1, 4, and 16 for states 1, 2, and 3, respectively. The MCMC estimation of the standard deviations were 1.05, 4.04, and 16.10, which are very close to the ground truth.

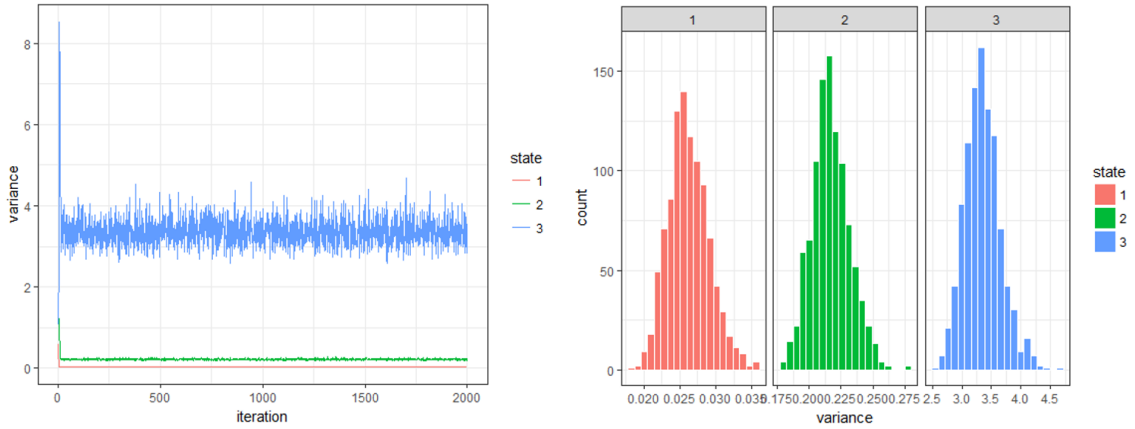

Figure 4.2: Variance changes over MCMC iterations (left); histogram of the estimated variances from the Gibbs sampler (right)

Additionally, the Bayesian approach enabled to capture of the segmentation uncertainty. Therefore, we can also evaluate the segmentation result in terms of probabilities. Figure 4.3 shows how the MCMC method captures the uncertainty. A red line indicates two standard deviations from the mean. For each iteration of the MCMC procedure, a red line was added and superimposed onto the previous lines to generate the figure. Thus, we can visually confirm which segment is more probable than others by examining the thickness of the red lines.

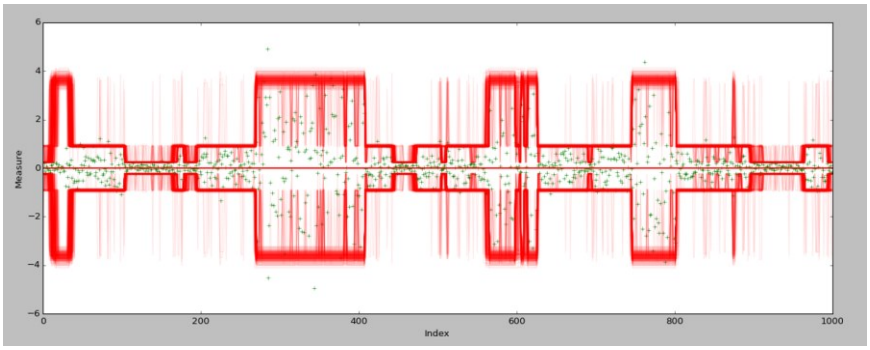

Figure 4.3: Visualization of uncertainty by superimposing two standard deviations from the mean lines

Figure 4.4 presents the segmentation results based on the MCMC estimation. Each color represents the three states. Red for State 1, green for State 2 and blue for State 3. Overall segmentation results are very close to the ground truth except for a few segments. In the bottom part of Figure 4.4, the probability of being a corresponding state can be visualized. We can observe that the probability is relatively low in the case of erroneous segments. Also, at the boundaries where the states change, the probability is lower due to the increase of uncertainty.

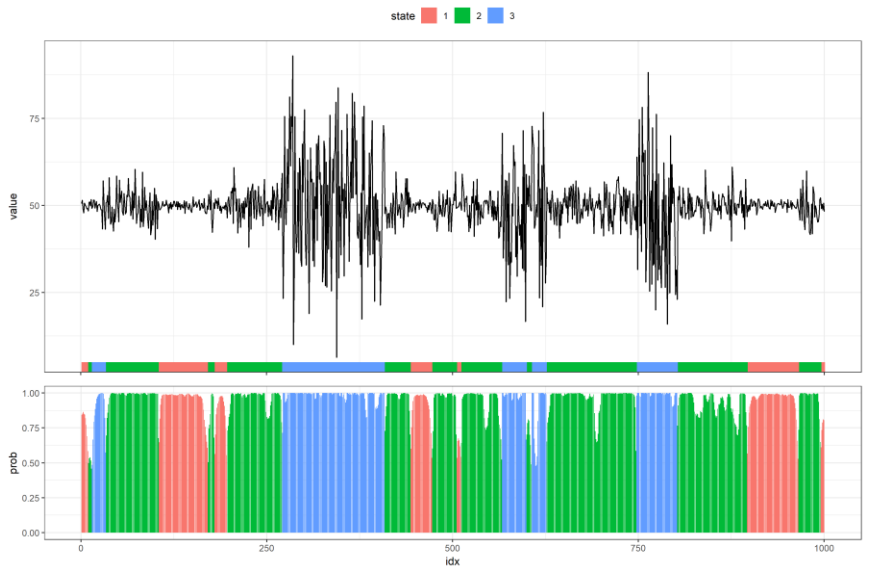

Figure 4.4: Segmentation result from MCMC estimation with probabilities of a point belongs to the corresponding state

4.3.2 Real Data

As another example, we applied the HMM segmentation method using MCMC to the real-world pavement condition score data a highway in Texas, using the 140-mile-long State Highway 46. Figure 4.5 presents the results. Each point in the graphic represents the ride score per half-mile sections, as measured in 2016.

One of the issues we confronted when using the MLE approach to HMM segmentation is that we cannot set a minimum segment length. Thus, the algorithm results in a number of short segments, of little utility for practical use. Those short segments are produced when there is a sudden jump or drop in a pavement condition score at a location. Although it is not certain whether such sudden changes occur due to measurement error, a feature to control the short segment issue might be advantageous in practice.

To prevent this problem, we incorporated a sticky parameter that forces a reduction in the state-to-state transitions. The sticky parameter, which is a fixed value, is added to a hyperparameter of the Dirichlet prior that corresponds to the self transition probability. Because the non-uniform hyperparameter of the state i determines the distribution of the Dirichlet, it results in a skewed distribution that the probability of transition from state i to the same state i is much higher than that of state i to a different state. For example, with a three-state HMM, a set of the Dirichlet parameters for a state $\alpha = (1, 1, 1)$ would result in fairly uniform transition probabilities while the non-uniform Dirichlet parameter $\alpha = (10, 1, 1)$ may result in a very skewed distribution, for which the probability of transition from state 1 to state 1 is much higher than that of transition from state 1 to state 2 or state 1 to state 3.

The top part of Figure 4.5 depicts the segmentation results without using the sticky parameter. Several short length segments are caused by a sudden jump or drop in ride score. As a result of incorporating the sticky parameter, we can observe fewer too-short segments, which can be a more practical solution, as illustrated in the bottom half of Figure 4.5.

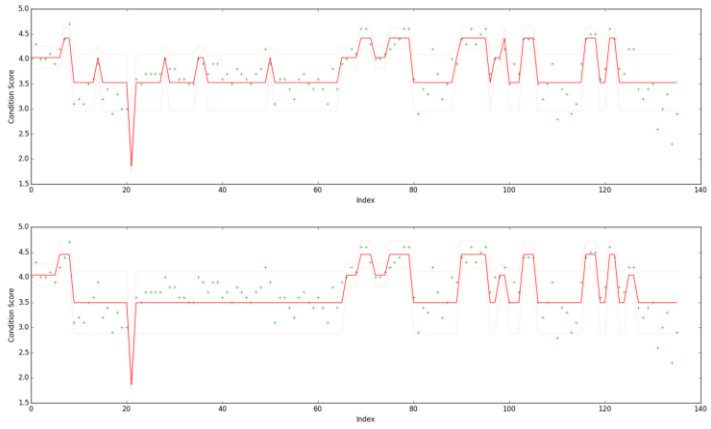

Figure 4.5: Comparison of HMM segmentation results from (a) without a sticky parameter and (b) with a sticky parameter on the real data (ride score of SH0046 K)

4.4 Discussion

This chapter presented a Bayesian approach using the Gibbs sampler to estimate HMM parameters, providing analysis examples using the simulated and real data. When comparing use of the EM algorithm to the Gibbs sampling for HMM segmentation, both approaches result in reasonably good segmentation that can be useful in practice.

Therefore, in circumstances where only a point estimate is needed or using maximal likelihoods to obtain the best model is sufficient, then the MLE approach provides a more straightforward and quicker solution than the Bayesian approach. However, to estimate parameters with probability with the goal of obtaining the model with global optima, the MCMC method would be beneficial.

When used with the simulated and real data, in a Bayesian context, the MCMC method tended to require more computation time to converge to the global optima. That is because the Bayesian approach involves difficulties such as a longer running time of the Markov chains for convergence, and sufficiently large samples drawn for accurate estimation. In contrast, the MLE approach required less computation time. However, the multiple estimations of HMMs, which require as much computation time as the MCMC method, were needed to obtain the best model in terms of the likelihood, as the global optima are not guaranteed in the MLE approach.

Despite the extended computation time required, the MCMC method has some characteristics that benefit its use in practice. First of all, thanks to the Bayesian property, the MCMC method can capture the uncertainty of a segmentation result—which allows use of this uncertainty to evaluate the resulting segments. Secondly, with a proper number of iterations in the MCMC procedure, the global optimum can be achieved so that the most probable segmentation is provided. Lastly, because the MCMC approach is very flexible, we can control the minimum length of a segment by modifying the prior of the transition probability. Also, by using a specific prior for the emission probability, we can obtain the more favorable means of each state. For example, the condition score in TxDOT is categorized in five different levels: very good, good, fair, poor, and very poor. We can assign priors that correspond to the range of each level to produce five states for which mean values are close to the predefined range.

The MCMC method does have some limitations. As mentioned earlier, the approach needs many iterations to converge so that bias caused by the starting values and Monte Carlo error can be negligible. Thus, the computation time for estimating a large network, like Texas highway networks, would be significant. In addition, determining the hyperparameter of the prior distribution is difficult. Since the results significantly depend on the value, they should be selected subjectively to obtain reasonably practical segmentation.

There are several opportunities to further enhance the Bayesian approach for future work. A sensitive study should be done to explore the effect of prior selection for both transition and emission probability. By doing so, the minimum length of a segment and the mean of each state can be controlled in a more sophisticated way. Furthermore, a method to estimate the HMM using multivariate observations was not implemented in this study. Incorporating multiple measurements of ride quality and pavement condition simultaneously, a method using multivariate observations would provide more practical and accurate segmentation results.

Chapter 5. HMM Segmentation for Identifying M&R Project History

In the previous chapters, we suggested HMMs as a new method for the highway segmentation. This chapter presents the result of a practical application using the HMM models in maintenance operation.

The effectiveness of a maintenance activity can be estimated in various ways. The simplest way is to compare the pavement performance before and after the maintenance activity. For instance, the difference in the IRI value before and after treatment can be a good indicator of how effective the treatment is, as shown in Figure 5.1. Having access to the work history, which indicates when and what treatment action was done on a pavement section, is essential for this estimation. However, an estimation may be required even when the work history is not recorded properly. In such a case, context clues can fill the information gap. For example, a significant drop in IRI or rutting measurements at a certain time can imply that some treatment actions would have been done.

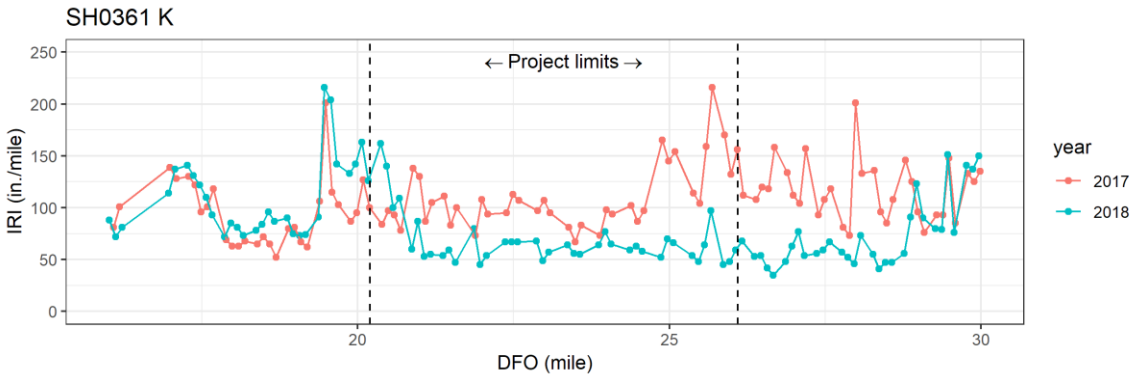

Figure 5.1: IRI measurements in 2017 and 2018 with project limits from work history on SH0361 K

Identifying M&R project boundaries is a critical component of pavement management for maintaining pavements effectively. When M&R project histories are not readily available, segmentation can be useful to determine the project boundaries, given that some pavement performance measurements are collected annually. A segmentation model can detect the differences of measurements over time and give some insights about resulting segments. The HMM segmentation method can be used for detecting a difference between before and after a maintenance action to identify M&R projects. In this chapter, we provide a practical framework for work history detection with real pavement data.

5.1 Framework

Figure 5.2 shows a workflow for HMM segmentation to identify M&R project boundaries using multivariate data. First, data processing is needed to run the HMM segmentation. Then, a model is selected, including the variable and number of states that will be used in the model. The selected model will be evaluated after learning and decoding the HMM and the model selection procedure

might be repeated if necessary. Post processing will be considered to merge states and segments, thus reducing redundant states and ensuring the minimum length of segments.

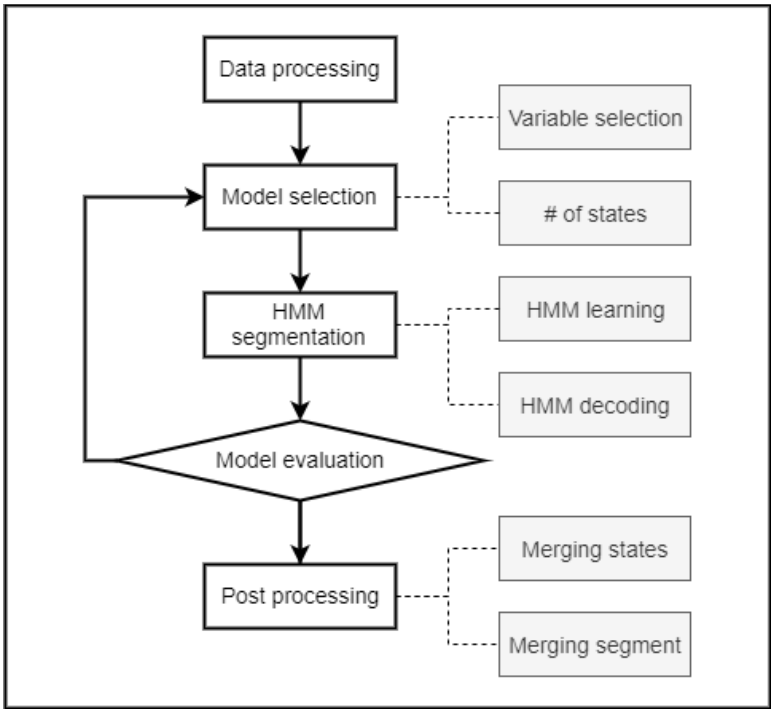

Figure 5.2: Flow chart for HMM segmentation framework

5.1.1 Data Processing

Whitening and Coloring Multivariate Data

Data whitening is one of the typical transforming processes to make multivariate data applicable to statistical analysis. When dealing with multidimensional observations to be used in HMM, it is necessary to transform observation data into a form that can be handled in the model.

The whitening transform is comprised of two main steps: decorrelation and scaling. Firstly, in the decorrelation step, the data is rotated such that it falls along the principal axes. This rotation removes the correlation between the components and results in a diagonal covariance matrix. Secondly, the scaling step squeezes or stretches the decorrelated data to make the unit variances. The resulting covariance matrix is identity.

In the scope of pavement segmentation, the whitening plays a critical role because some variables used in the segmentation model are highly correlated; e.g., pavement distresses such as rutting and cracking are highly correlated at the same location. Therefore, there should be a remedy for the correlation problem that might cause numerical instability in the course of model estimation and the whitening transform is one of the remedies available. In addition to that, the scaling step is beneficial for using multiple measures at the same time with the same importance in segmentation due to the identical variances for all variables in the data.

In the remainder of this section, we present how to transform to whitened data and also present how to reverse the transform by coloring the data.

Let X be a multivariate Gaussian random vector with mean μ and covariance matrix Σ. The covariance matrix is calculated thusly:

$$\Sigma = E[XX^T]$$

The resulting covariance matrix can be decomposed as follows:

$$\Sigma = \Phi \Lambda \Phi^{-1}$$

where Λ is a diagonal matrix with the eigenvalues of Σ and Φ is the eigenvector. A random vector with a decorrelated multivariate Gaussian distribution can be obtained by the multiplication of the transposed eigenvector by the original vector X.

$$Y = \Phi^T X$$

Y has a diagonal covariance matrix with the eigenvalues. Thus, to scale to a standard multivariate Gaussian distribution, the following step must be taken.

$$W = \Lambda^{-\frac{1}{2}} Y$$

Now, W has an identity matrix as its covariance matrix.

The coloring process is the reverse of the whitening process. When one wants to obtain the original multivariate Gaussian distribution from the white noise data, the coloring process must be completed.

Let S be a random vector from a whitened distribution. We can reverse the scale of S back to that of the original distribution by multiplying the square root of the diagonal matrix of the eigenvalues.

$$Y = \Lambda^{\frac{1}{2}} S$$

And Y is still the decorrelated data, so by rotating the data, we can obtain the correlated data again.

$$X = \Phi Y$$

To illustrate what the whitening transform actually does to data, a thousand generated data points were sampled from the bivariate Gaussian distribution with $\mu = 0$ and covariance matrix

$$\Sigma = \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}$$

Figure 5.3 shows the plot of the generated data from X to Y to W. Figure 5.3(a) shows a scatter plot of the original samples. The shape of the linear relationship definitely presents strong

correlations between variables. By the eigenvectors of X covariance, the decorrelated data Y can be found by rotating the original data to align it with the principal axes of the data as shown in Figure 5.3(b). The covariance matrix of Y proves that there is no correlation between components. Furthermore, this decorrelated data Y can become the whitened data W by scaling each variance to unit one. Now the normalized data demonstrates a circular shape in the scatter plot, as can be seen in Figure 5.3(c), and the covariance matrix becomes an identity matrix, which was the desired outcome of the procedure.

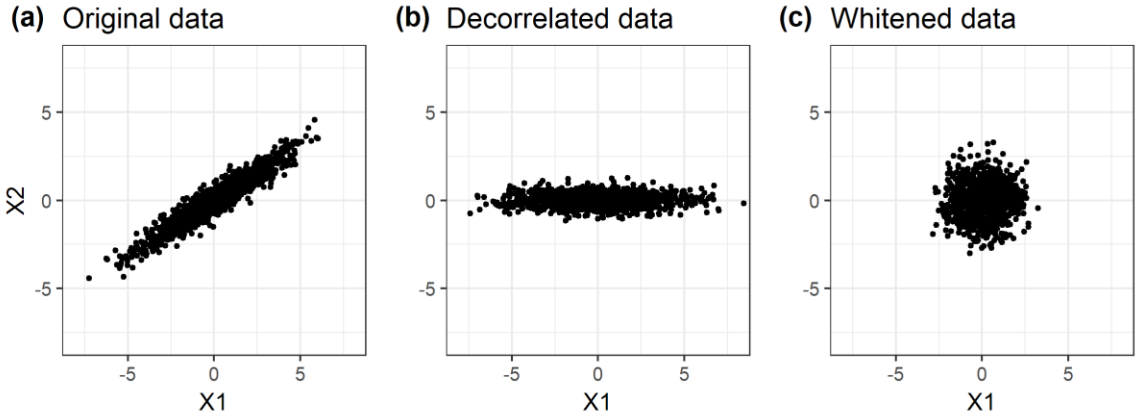

Figure 5.3: Scatter plots of bivariate Gaussian distribution. (a) Original, (b) Decorrelated, and (c) Whitened data

The covariance matrices of the sample data changed as follows:

$$(a)\begin{bmatrix}5.50 & 3.28\\ 3.28 & 2.14\end{bmatrix} \rightarrow (b)\begin{bmatrix}7.50 & 0.00\\ 0.00 & 0.14\end{bmatrix} \rightarrow (c)\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

The aforementioned procedures are based on the assumption that the data is from multivariate Gaussian distribution. Also, whitened data needs to be centered about zero. To achieve a zeroed center, one can easily subtract the mean of each variable from the data points. Since real-world data is not always normally distributed, another transformation may be needed so that the data form a normal distribution prior to apply the whitening transformation.

5.1.2 Model Selection

Selecting the Number of States

Typically for comparing between models for evaluating which model is better than others, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used.

The likelihood is always improving as the number of states increases. However, if a large number of states is used for a model, overfitting is expected to occur. That is, there would be too many short-length segments, although the likelihood of a model is better. AIC and BIC penalize log-likelihood by adding a penalty term to prevent this issue. Thus, a lower AIC or BIC value indicates a better model. Two criteria are as follows:

Akaike information criterion (AIC):

$$-2\log L + 2p$$

Bayesian information criterion (BIC):

$$-2\log L + p\log T$$

where, $p = m^2 + km - 1$ m: number of states k: number of parameters.

5.1.3 HMM Segmentation

HMM learning and decoding can be done by using the MLE approach or the Bayesian method introduced in previous chapters. Here, we used the MLE approach to implement the segmentation method for identifying an M&R project.

5.1.4 Post-processing

Merging States

For the purpose of identifying M&R projects, we can merge all other states except the state representing the M&R projects. On the other hand, if one wants to obtain both segmentation and project identification, we need to merge segments whose properties are similar to reduce redundant states.

Merging Segments

Once we have merged states, there will be short-length segments that are shorter than the minimum segment length required as an M&R project. For instance, the TxDOT experts consider one mile as the minimum length for a single project.

By merging a short segment with an adjacent segment, the minimum length can be secured. We suggest merging those short segments, joining the neighboring segments in accordance with certain criteria, such as similar mean and variance level.

5.2 Example

5.2.1 Data Processing

As an example, we applied the segmentation framework to real-life pavement data to identify M&R project sections. Since 2016, TxDOT has outsourced data collection to obtain pavement condition data using automatic and accurate methods at highway speed. A vendor measures roughness in IRI and distresses on the pavement by using technologies involving lasers and images, and reports the summarized values for 0.1-mile intervals such that each data point represents a section that is one-tenth of a mile. A data set most recently collected from 2017 to 2018 in the Austin area, one of 25 districts in TxDOT, was used for this study. Since rut depth measurement was considered as a variable in the segmentation model, the data was filtered to include only asphalt concrete pavements. The TxDOT-maintained highway system in the Austin

District consist primarily of asphalt concrete highways, so this network size is suitable for the study.

To process the data, highways were grouped by roadbed name, which entails a highway name and a label that indicates left or right lanes, and main or frontage roads for divided highways (e.g., IH0035 L refers to the left lanes of the Interstate Highway 35). Undivided highways were labeled with the letter "K". The resulting group of highway roadbeds was divided into subgroups when there is a gap between data points greater than two miles within the same group of highway. Because the HMM segmentation assumes the dependency of neighboring sections, a hidden state of the next section depends on that of the current section by Markov chain. It is not reasonable that highway sections farther apart than two miles affect the performance of each other's pavements. Other inventory information such as county and pavement type was not used for grouping, to prevent interference with the segmentation process from factors other than pavement condition.

Within a grouped highway, we matched the identical 0.1-mile sections over two years based on location information—Distance From Origin (DFO)—such that the difference between two years at a specific DFO can be calculated. Sections with a missing measurement or ones that do not have pairs at a location over two years were discarded from the data set. Also removed were highway groups for which a center-line length was less than five miles, to secure an adequate number of data points per highway group. As a result, overall, the process yielded 2484.3 miles of 162 highway groups.

5.2.2 Variable Exploration

Two measures, i.e, roughness in IRI and rut depth, were chosen to detect the M&R project boundaries. The values of IRI and rut depth are continuous, so the multivariate Gaussian distribution can be used as the emission probability of the HMM. Also, the fact that IRI and rut depth are not usually correlated is useful to accommodate the multivariate Gaussian distribution in the model. Figure 5.4(a) shows the scatter plot of the two measures in the processed data set. A slight positive correlation is observed between IRI and rut depth. The paired correlation was calculated as 0.19.

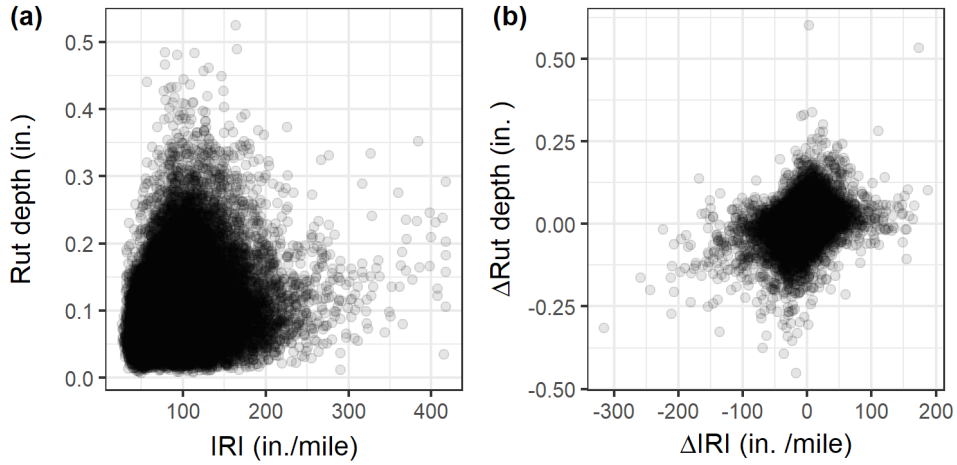

Figure 5.4: Scatter plot of (a) Rut depth versus IRI; (b) ΔRut depth versus ΔIRI

In order to identify an M&R project, detecting a meaningful change of values per each variable over a two-year period is important. The change of values, Δ is defined as

$$\Delta IRI = IRI_{t+1} - IRI_t; \;\; \Delta Rut = Rut_{t+1} - Rut_t$$

where t is time in year.

Therefore, we can expect that sections with negative values less than a specific value in delta might have M&R project because both IRI and rut measurements decrease as a pavement section is improved. Whereas, sections with no M&R project might experience increases in the measurements. Also, we should consider errors involved in the data because of operation errors by human and equipment, and location differences over the years. Although ΔIRI and ΔRut are independent, they might have relatively higher correlations since they would be both negative in the project sections. However, Figure 5.4(b) shows that there is no apparent high correlation between them. The correlation value was 0.31, which is higher than that between IRI and rut measurements.

Figure 5.5 shows the histogram of ΔIRI and ΔRut. The distributions are almost symmetric with a single peak around zero. It was expected that the difference between two years of performances was not significant. In addition, we cannot expect left-skewed distributions because there might have been only a few M&R projects for two years. In Table 5.1, a mean and standard deviation of each variable is presented. The mean of ΔIRI is slightly negative and that of ΔRut is positive; however, they are not statistically different from zero—that is, the overall changes in two variables are not so meaningful.

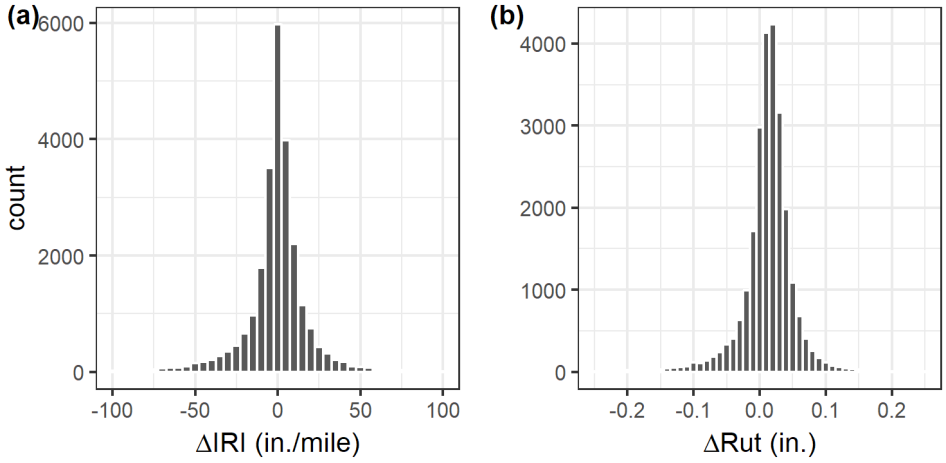

Figure 5.5: Histogram of (a) ΔIRI; (b) ΔRut

Table 5.1: Mean and standard deviation values of ΔIRI and ΔRut

ΔIRI Mean	ΔIRI SD	ΔRut Mean	ΔRut SD
−0.875	22.256	0.013	0.042

Another set of variables was considered for use in the model: the 2017 initial IRI and rut measurements, extracted from a two-year dataset. When a measurement of the previous year is higher, the difference of measurements before and after an M&R project might become greater. Also, by using those measurements as a variable, the model can detect segments incorporating the conditions of pavement sections in terms of IRI and rut depth. That means that we can obtain both project boundaries and highway segmentation simultaneously. Therefore, it is worth adding the initial values in the HMM.

Figure 5.6(a) and (b) display the distribution of IRI and rut depth, respectively. They are both right-skewed distributions, so log transformations were applied to them to make the distributions closer to normal distributions. As can be seen in Figure 5.6(c) and (d), the distributions do become closer to a normal distribution after the transformation. Q-Q plots per each distribution from (a) to (d) were drawn as shown in Figure 5.6(e) to (h) to check the normality, additionally. The log-transformed variables have straighter lines in the Q-Q plots (f) and (h); that is, the normality is visually evident.

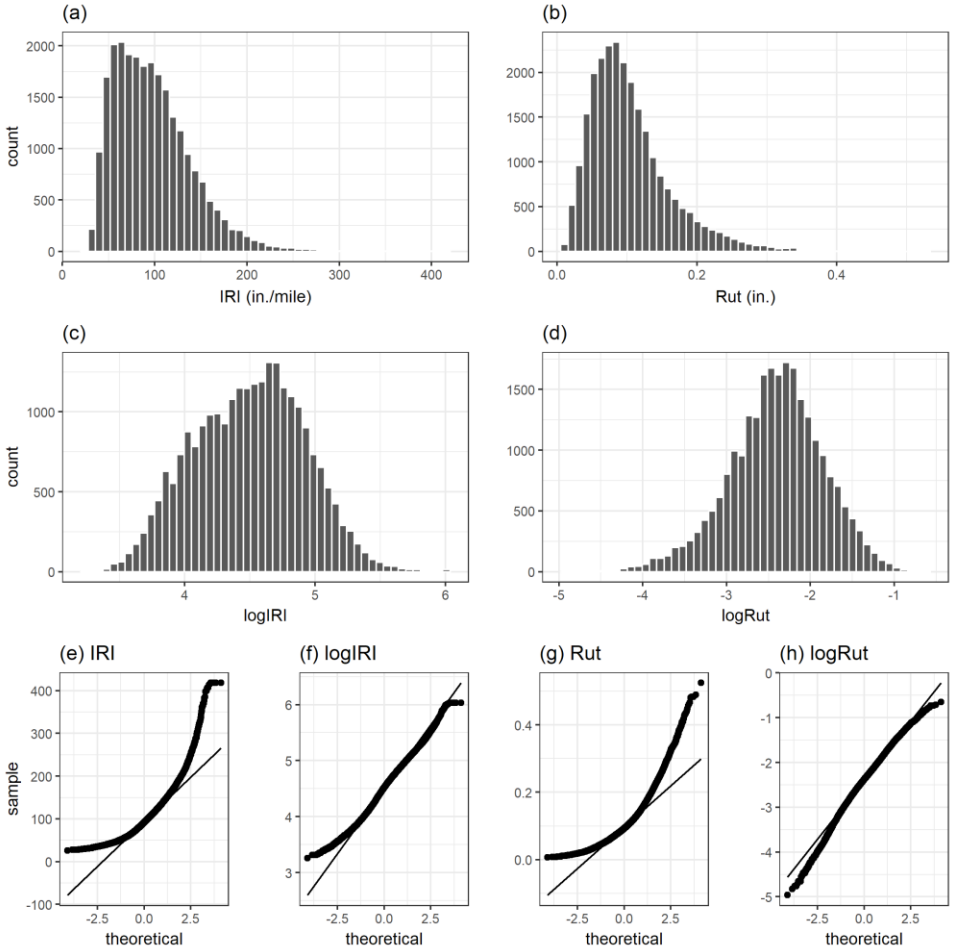

Figure 5.6: Distributions of the initial IRI and rut depth. (a) IRI and (b) rut depth; (c) logIRI and (d) logRut; (e), (f), (g), and (h) Q-Q plots of before and after log transformation for IRI and rut, respectively

The normality is not a critical problem for the HMM itself because the normality assumptions are involved only when a hidden state emits the observations by determining the emission probability to a multivariate Gaussian. The reason for the transformation is the whitening process, in which the normality assumption is made. For multivariate analysis, as explained in the previous section, whitening is necessary to eliminate correlations between variables and standardize variables to make their scale the same. Thus, this linear transformation was also applied to the data set. Once the model produced results, they were unwhitened by a coloring process to be presented.

Lastly, one more set of variables was considered. When IRI and rut are measured, a vendor collects the measurement on the left and right wheel paths separately. Typically, the right wheel path is prone to develop more damages over time, so that the measurements on the left and right wheel paths are expected to be different. Although the discrepancy is not significant, and the average value of both paths are used in practice, two new variables were created, ΔIRI_{diff} and ΔRut_{diff}, to accommodate this inevitable difference:

$$IRI_{diff} = |IRI_{left\ wheel\ path} - IRI_{right\ wheel\ path}|$$

$$\Delta IRI_{diff} = IRI_{diff,t+1} - IRI_{diff,t}$$

$$Rut_{diff} = |Rut_{left\ wheel\ path} - Rut_{right\ wheel\ path}|$$

$$\Delta Rut_{diff} = Rut_{diff,t+1} - Rut_{diff,t}$$

If a section has a significant discrepancy in values between two wheel paths, the absolute difference between them would be greater than the one after an M&R treatment. Hence, we expect a negative $\Delta value_{diff}$ for sections experience a treatment.

In the next section, a variable selection process is outlined to determine which set of variables is more effective for the model.

5.2.3 Model Selection with Partial Data

A model selection process aims at two major aspects. One is to select variables to be entered into HMM; another is to select a proper number of hidden states for the HMM. For the model selection purpose, the ten longest highway groups in the full dataset and another two groups that show a visually obvious M&R project were selected as a partial dataset, which has 598.3 miles in total center-line length. For estimating the HMM parameters closed to the global optima, two hundred iterations of the EM procedure were performed per each variation of the model with different sets of variables and number of states. With the relatively smaller dataset with 12 highway groups, it was possible to run a significant number of models to evaluate and determine the optimal setup for the problem.

Variable Selection

The variable selection was explored with the three sets of variable combinations, including two, four, and six variables. Firstly, two variables, such as ΔIRI and ΔRut, assumed to be most important factors for detecting a project, were included in the model. An HMM with five hidden

states produced the segmentation result presented in Figure 5.7. The figure is a frontage road of the Interstate Highway 35 (IH0035 X). The IRI and rut values are displayed as lines, and each color and line type represents the measurements in corresponding year. On the x-axis, a color-coded rectangular bar depicts the segmentation result. The bar is separated by colors located on a specific segment range; each bar color represents a hidden state. We can observe the State 3 is the most probable project segment located from approximately milepost 22 to 26 in DFO. That segment is placed where significant decreases in both variables are observed. It also can be seen in Table 5.2 that State 3 has the most negative means for both variables. Other states also have physical meaning. For instance, State 1 has negative means in ΔIRI, but has a slightly positive mean in ΔRut. Therefore, State 1 has the potential to be recognized as a project. However, it should be noted that the result was obtained from only a portion of the full data that was used for this task.

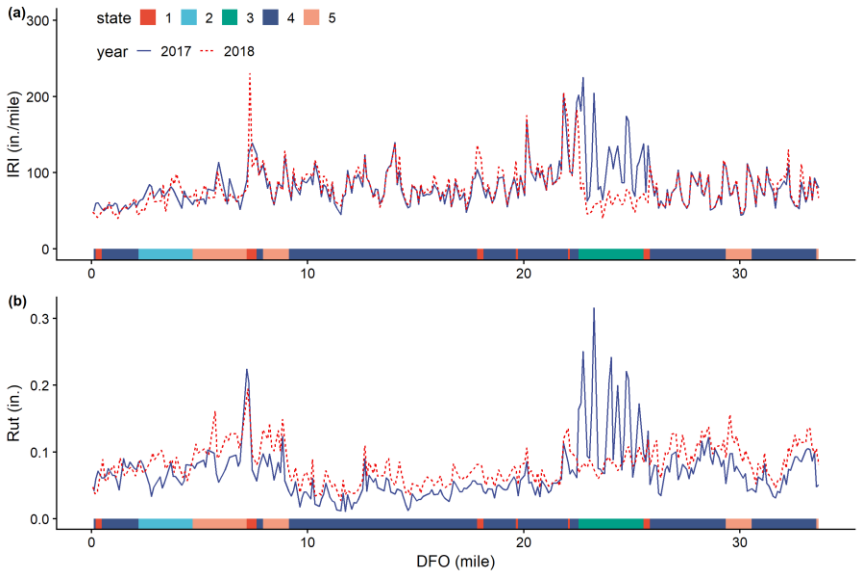

Figure 5.7: Segmentation result with 2-variable 5-state HMM on the partial data: (a) IRI; (b) rut depth (IH0035 X)

Table 5.2: Estimated state parameters with 2-variable 5-state HMM on the partial data

State	mean_dIRI	sd_dIRI	mean_dRut	sd_dRut
1	-7.145	45.772	0.016	0.032
2	2.361	17.133	0.030	0.054
3	-31.782	33.034	-0.081	0.064
4	0.638	6.460	0.016	0.015
5	2.283	9.971	0.037	0.018

Secondly, four variables, such as ΔIRI, ΔRut, initial IRI and rut depth, were included in the model to take into account the effect of initial measurement before an M&R construction. Also, we expected the model would offer a segmentation based on pavement condition as well as project boundaries. Figure 5.8 demonstrates the results of the four-variable model. Similar to the result of

the two-variable model, project segments were correctly recognized. This time, State 5 represents the project segment, unlike the previous model. Each state now has additional meaning in terms of the initial conditions of pavement, as is observable in Table 5.3.

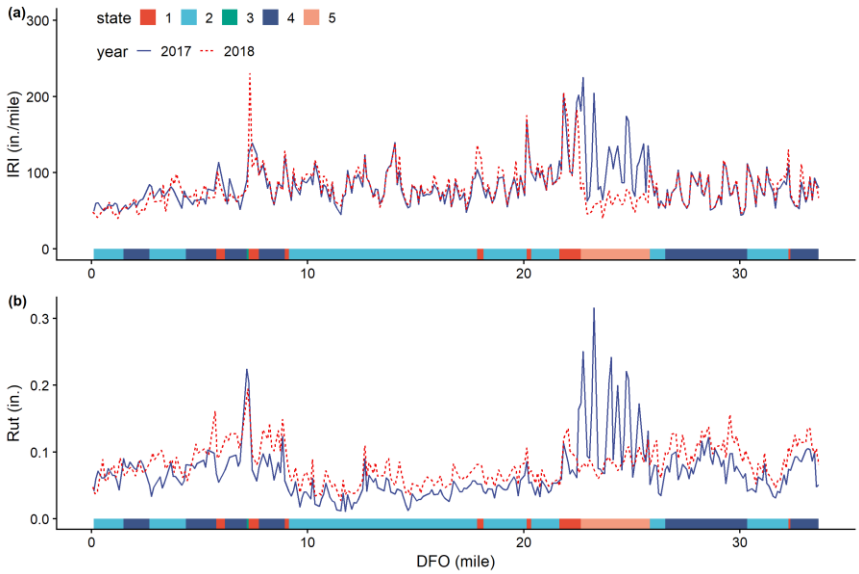

Figure 5.8: Segmentation result with 4-variable 5-state HMM on the partial data: (a) IRI; (b) rut depth (IH0035 X)

Table 5.3: Estimated state parameters with 4-variable 5-state HMM on the partial data

State	mean_dIRI	sd_dIRI	mean_dRut	sd_dRut	mean_IRI	sd_IRI	mean_Rut	sd_Rut
1	-8.693	44.457	0.016	0.037	113.020	73.277	0.071	0.047
2	2.195	9.171	0.025	0.018	60.989	38.542	0.045	0.030
3	1.884	14.200	0.027	0.045	103.412	64.300	0.176	0.114
4	0.709	7.694	0.023	0.020	62.163	39.026	0.104	0.066
5	-21.792	32.131	-0.058	0.075	87.631	56.010	0.147	0.098

Lastly, the result of the 6-variable model is presented in Figure 5.9 and Table 5.4.

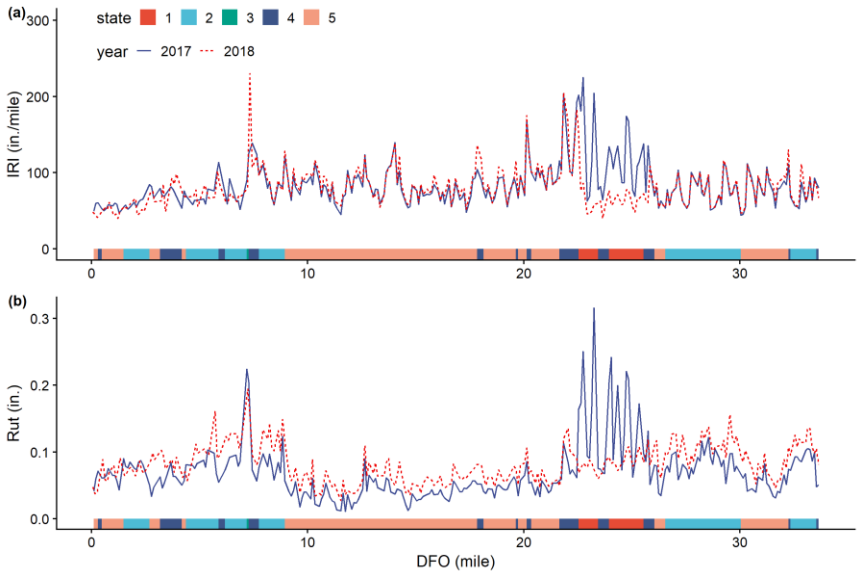

Figure 5.9: Segmentation result with 6-variable 5-state HMM on the partial data: (a) IRI; (b) rut depth (IH0035 X)

Table 5.4: Estimated state parameters with 6-variable 5-state HMM on the partial data

State	mean_dIRI	sd_dIRI	mean_dRut	sd_dRut	mean_IRI	sd_IRI	mean_Rut	sd_Rut	mean_dIRId	sd_dIRId	mean_dRutd	sd_dRutd
1	-15.148	35.036	-0.025	0.091	94.177	60.477	0.162	0.107	-3.337	22.127	-0.001	0.083
2	0.500	8.070	0.023	0.021	62.195	38.952	0.105	0.066	-0.358	7.537	0.011	0.034
3	1.646	11.331	0.022	0.033	103.109	63.989	0.171	0.111	0.770	11.975	0.003	0.039
4	-6.391	40.969	0.018	0.034	107.245	69.673	0.067	0.044	-3.075	23.582	0.002	0.036
5	2.068	8.227	0.025	0.017	59.530	37.621	0.046	0.031	-0.317	6.787	0.005	0.022

Since so far we evaluated the model results with one highway example, a look at overall segmentation quality is needed. Therefore, 2-D scatter plots with ΔIRI and ΔRut, which are the most important factors for the purpose of this study, were drawn for three models with the different number of variables, as shown in Figure 5.10. The result shows that the 2-variable and 4-variable models detect project segment well. However, the 6-variable model could not identify the states correctly.

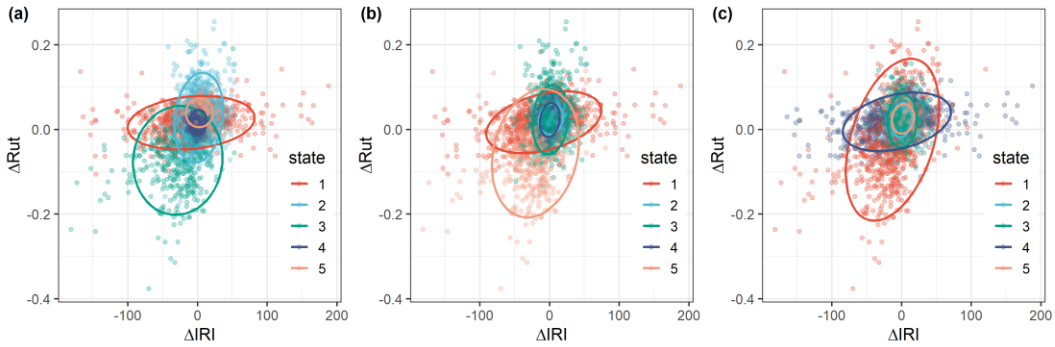

Figure 5.10: 2-D scatter plots of ΔRut vs. ΔIRI for 5-state models on the partial data: (a) 2 variables; (b) 4 variables; (c) 6 variables

Figure 5.10 (b) demonstrates that State 2 and State 4 overlapped, which means they are very similar in terms of the rut and IRI differences. The reason why there are two separate states can be observed in a 1-D plot for showing mean and variance of each state with respect to log rut depth (see Figure 5.11). The states of a 2-variable model cannot detect differences in log rut depth; however, the 4-variable model definitely differentiates states. Meanwhile, the 6-variable model failed to differentiate between states in terms of ΔIRI_{diff}, as shown in Figure 5.12.

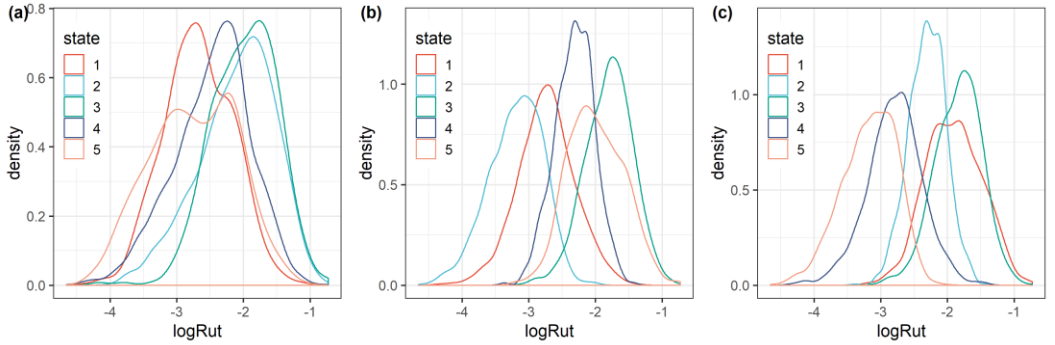

Figure 5.11: Density plots of logRut for 5-state models on the partial data: (a) 2 variables; (b) 4 variables; (c) 6 variables

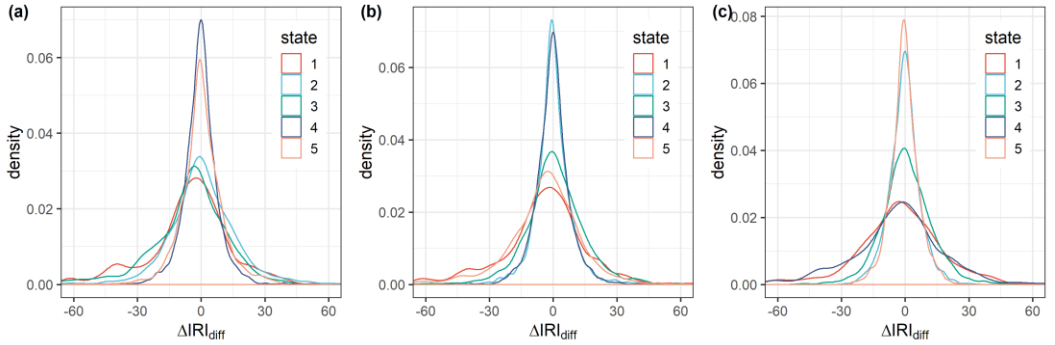

Figure 5.12: Density plots of ΔIRI_{diff} for 5-state models on the partial data: (a) 2 variables; (b) 4 variables; (c) 6 variables

Number of States Selection

In order to select the number of states of HMM segmentation, the Akaike's Information Criteria (AIC) and the Bayesian Information Criteria (BIC) were calculated for the number of states ranging from 3 to 15, as shown in Figure 5.13. The result of AIC indicates that nine is the optimal number of states. A BIC result is obtained in six states.

Because there is no universal method to determine the number of states for HMM, we should consider not only the results from calculation of AIC and BIC but also from a practical point of view. In TxDOT, pavement condition is usually categorized into five classes (very good, good, fair, poor, and very poor). Thus, using five states can be an appropriate choice for both practical and analytical perspectives.

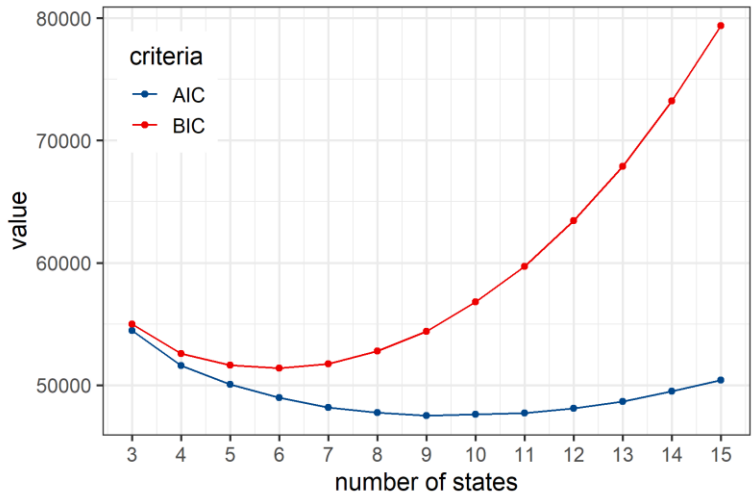

Figure 5.13: AIC and BIC versus number of states

5.3 Discussion

This chapter presented an application of the segmentation method to identify M&R project boundaries using multi-attribute data. The method yielded very promising results with significant benefits—not only identifying the M&R project limits but also producing segmentation based on current pavement conditions by incorporating the initial value of condition measurements.

There are several limitations of the framework. Because the MLE method was used to estimate HMM parameters, it does not guarantee the global optima when it comes to obtaining the maximum likelihood estimators. Therefore, multiple runs of the MLE process should be conducted to reach the near-optimal solution. Also, when the label switching problem occurs, we cannot obtain a state corresponding to M&R project segments with a consistent label. Furthermore, to ensure the minimum length of a segment is greater than a practical length, post-processing is required to merge some segments using somewhat subjective judgement.

For future study, we suggest estimating HMM using the Bayesian approach presented in Chapter 4 instead of using the MLE approach to obtain the global optima. In addition, the Bayesian approach can control the minimum length of segment such that it can minimize the involvement

of post-processing. Furthermore, since we used only a limited size of pavement data in the Austin area to demonstrate the framework, it is necessary to apply the method to more extended data.

Chapter 6. Summary and conclusion

Optimal planning of pavement maintenance and rehabilitation (M&R) activities is essential for highway and transportation agencies to manage a sustainable transportation infrastructure system. Pavement maintenance consists of routine and preventive activities such as filling cracks, patching, chip seal, thin overlays, microsurfacing, etc. Pavement rehabilitation includes actions such as thick overlays and partial to complete reconstruction that increase the structural capacity of pavement. Due to the size of road networks, M&R is one of the major investments in a transportation system. Accordingly, planning M&R and resource allocation are a major issue that challenges administrators and decision makers because they need to determine which pavement road section has to be treated and when, and how that treatment should be conducted. In addition, the decision making process must take into account budget limitations, meet specific goals for maintaining pavement performance, and allocating budgets to maximize cost effectiveness. Finally, other non-engineering external factors also affect the decision process such as political-based decision, extreme weather events, last minute policy changes, etc.

In the Texas Department of Transportation (TxDOT), pavement management systems have been operated since the early 1990s to support the pavement-related decision making processes by storing, retrieving, analyzing and reporting information. Currently, the information is managed in half-mile data collection unit sections in a new system known as Pavement Analyst (PA). This information can be analyzed to objectively support the decision making process such as condition estimation and maintenance needs estimation at the administrative state level. However, at the district level project selection, the half-mile section data are restrictive because typically projects are of any length, combining multiple half-mile sections. Therefore, instead of using the half-mile data collection section, aggregating several of these units into a "management section" consisting of homogeneous sections is necessary. For that reason, obtaining the limits (beginning and end points) of homogeneous sections becomes a key problem in pavement preservation and in pavement and maintenance management. Appropriate segmentation is required for optimal determination of the beginning and end points of the management sections and for a more cost effective M&R plan. Failure to do this will result in suboptimal resource allocation and therefore, waste of limited M&R funding.

In this study, we reviewed and evaluated previous research studies focused on the segmentation of highway pavements. Also, we explored off-the-shelf tools available for detecting change points and compared the results of implementation with the CDA method. Then, we suggested a segmentation method using HMMs as a prospective method for highway management operations. With simulated and real data, the method was tested to demonstrate its benefits as compared to the CDA method. For overcoming some issues that arose in the maximum likelihood estimation of HMMs, a Bayesian approach to estimate HMM parameters was proposed. Throughout data analysis, we demonstrated the application of HMM segmentation to identify M&R project limits with IRI and rut depth differences over time.

The majority of the segmentation methods delineate segments by identifying one change-point at once and repeating the algorithm to detect more changes using the divided segments from the previous run. Even though additional adjustments are suggested as a remedy, this type of approach does not result in optimal multiple change-points. Although finding optimal solutions increases

computational time, the power of modern computer systems and efficient optimization algorithms make it possible to obtain the global optima. Therefore, the developed methods should have the ability to identify the multiple homogeneous segments at once without losing optimality.

Most studies have focused on delineating segments based on different mean levels of segments. Few studies have attempted to develop a method that takes into account changes in variance and autocorrelation. Pavement performance data might be autocorrelated by its very nature. That is, each observation is not independent but correlates to the adjacent one, so that the performance measure of current section has something to do with that of the next section. Thus, it would be of interest to develop a method that can take into account such correlation. In addition, few studies have explored multivariate data for pavement segmentation. Thus, it would be also interesting to develop a method that, for example, can conduct segmentation based on rut and skid data simultaneously.

Throughout the case study of the CDA and two off-the-shelf packages in R, we evaluated and compared the qualitative performance of each method. Although the overall performances of the three methods presented in this study seemed reasonable, the two change-point algorithms produced more reasonable results than the CDA. One benefit of a change-point analysis is that it controls the variability. Also, the change-point algorithms have features to prevent overfitting. Although the PELT and BCP could be implemented conveniently using the packages, establishing a tweaking process—by adjusting the penalty and parameters to obtain desirable segmentation results—is a challenging and subjective task. For both algorithms, the results are highly sensitive to those adjustable user inputs, so visual inspection after multiple runs is the only practical way to optimize the input values.

The PELT algorithm can detect multiple changes in mean and variance but cannot employ multivariate data. As opposed to the PELT, the BCP does not offer variance change detection, but it can handle multivariate data analysis. Due to the nature of the Bayesian approach, the BCP algorithm gives not the location of change-points but the posterior probability of change-points at locations. This property is beneficial in terms of diagnosing the uncertainty of segmentations. However, a post-process is necessary to obtain change-point locations using a threshold value with respect to the posterior probability. This process introduces additional subjectivity to the segmentation results. Also, a potential problem of the BCP is that it would not produce identical results every time it runs because the segmentation results are obtained by the MCMC method. In order to achieve more consistent and rigorous results over multiple runs of the algorithm, proper MCMC settings are required, including an initialization, a burning, and the number of iterations.

As a result of the literature review and comparison between the existing methods, this study suggested the following desirable properties of a segmentation method based on these findings:

- Detect multiple change-points simultaneously;
- Provide optimal or near-optimal solution;
- Detect changes in mean, variance, or autocorrelation;
- Adjust sensitivity in terms of a change in parameters;
- Control the minimum length of a segment;
- Provide a measure of uncertainty; and

- Handle multivariate data.

To integrate these elements, we introduced use of HMM for the task of highway segmentation. HMM consists of two layers. The hidden layer is a sequence of unobserved states that explain another layer, the observation sequence. For instance, the hidden layer can be pavement performance of a road section, and the observation sequence can be a pavement distress such as rut and roughness. Highway segmentation using HMM can be achieved by estimating HMM parameters, such as transition probability and emission probability, through the EM algorithm. Then, the Viterbi algorithm can be used to obtain the most probable state sequence, which can identify segments.

A comparison between the CDA and the HMM approaches to both simulated and real data sets found that the HMM method is more advantageous, as the HMM approach is rigorous to a local variance, can detect variance changes as well as mean changes, and enables multivariate analysis. It cannot be concluded, however, that CDA is always worse than HMM because CDA might present more advantages when used with other data sets. There is no control for adjusting the behavior of the CDA method in finding segments. Hence, HMM can generally produce better results regardless of data.

Some limitations of using the HMM method were identified when applying HMM to real pavement data. One such limitation is the problem of remaining stuck in a local maximum during the EM estimation. To overcome this limitation and also to prevent short-length segments, which are not practical, we can use an alternative approach to estimate HMM parameters using a Bayesian framework. We presented a Bayesian approach using the Gibbs sampler to estimate HMM parameters, providing analysis examples using the simulated and real data. When comparing use of the EM algorithm to the Gibbs sampling for HMM segmentation, both approaches result in reasonably good segmentation that can be useful in practice.

Therefore, in circumstances where only a point estimate is needed or using maximal likelihoods to obtain the best model is sufficient, then the MLE approach provides a more straightforward and quicker solution than the Bayesian approach. However, to estimate parameters with probability with the goal of obtaining the model with global optima, the MCMC method would be more beneficial. When used with the simulated and real data, in a Bayesian context, the MCMC method tended to require more computation time to converge to the global optima. That is because the Bayesian approach involves difficulties such as a longer running time of the Markov chains for convergence, and sufficiently large samples drawn for accurate estimation. In contrast, the MLE approach required less computation time. However, the multiple estimations of HMMs, which require as much computation time as the MCMC method, were needed to obtain the best model in terms of the likelihood, as the global optima are not guaranteed in the MLE approach. Despite the extended computation time required, the MCMC method has some characteristics that favors its use in practice. First of all, thanks to the Bayesian property, the MCMC method can capture the uncertainty of the segmentation results—which allows use of this uncertainty to evaluate the resulting segments. Secondly, with a proper number of iterations in the MCMC procedure, the global optimum can be achieved so that the most probable segmentation is provided. Lastly, because the MCMC approach is very flexible, we can control the minimum length of a segment by modifying the prior of the transition probability. Also, by using a specific prior for the emission

probability, we can obtain the more favorable means of each state. For example, the condition score in TxDOT is categorized in five different levels: very good, good, fair, poor, and very poor. We can assign priors that correspond to the range of each level to produce five states for which mean values are close to the predefined range.

The MCMC method does also have some limitations. As mentioned earlier, the approach needs many iterations to converge so that bias caused by the starting values and Monte Carlo error can be negligible. Thus, the computation time for estimating a large network, such as Texas highway network, would be significant. In addition, determining the hyperparameter of the prior distribution is difficult. Since the results significantly depend on the value, they should be selected subjectively to obtain reasonably practical segmentation. There are several opportunities to enhance the Bayesian approach for future work. A sensitive study should be performed to explore the effect of prior selection for both transition and emission probabilities. By doing so, the minimum length of a segment and the mean of each state can be controlled in a more sophisticated way. Furthermore, a method to estimate the HMM using multivariate observations was not implemented in this study. Incorporating multiple measurements of ride quality and pavement condition simultaneously, a method using multivariate observations would provide more practical and accurate segmentation results.

We presented an application of the segmentation method to identify M&R project boundaries using multi-attribute data. The method yielded very promising results with significant benefits—not only identifying the M&R project limits but also producing segmentation based on current pavement conditions by incorporating the initial value of condition measurements. There are several limitations of the framework. Because the MLE method is used to estimate HMM parameters, it does not guarantee the global optima when it comes to obtaining the maximum likelihood. Therefore, multiple runs of the MLE process should be conducted to reach the near-optimal solution. Also, when the label switching problem occurs, we cannot obtain a state corresponding to M&R project segments with a consistent label. Furthermore, to ensure the minimum length of a segment is greater than a practical length, post-processing is required to merge some segments using somewhat subjective judgement.

REFERENCES

A. Viterbi. (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm." *IEEE Transactions on Information Theory*, 13(2), 260–269.

AASHTO. (1993). AASHTO guide for design of pavement structures, 1993. AASHTO.

Barry, D., and Hartigan, J. A. (1993). "A Bayesian Analysis for Change Point Problems." *Journal of the American Statistical Association*, 88(421), 309–319.

Baum, L. E., and Eagon, J. A. (1967). "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology." *Bulletin of the American Mathematical Society*, 73(3), 360–363.

Baum, L. E., and Sell, G. R. (1968). "Growth transformations for functions on manifolds." *Pacific J. Math.*, 27(2), 211–227.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *The Annals of Mathematical Statistics*, 41(1), 164–171.

Bennett, C. R. (2004). "Sectioning of road data for pavement management."

Boroujerdian, A. M., Saffarzadeh, M., Yousefi, H., and Ghassemian, H. (2014). "A model to identify high crash road segments with the dynamic segmentation method." *Accident; analysis and prevention*, 73, 274–287.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1983). "Classification and regression trees."

Cafiso, S., and Di Graziano, A. (2012). "Definition of homogenous sections in road pavement measurements." *Procedia - Social and Behavioral Sciences*, SIIV-5th international congress - sustainability of road infrastructures 2012, 53, 1069–1079.

Casella, G., and George, E. I. (1992). "Explaining the Gibbs Sampler." *The American Statistician*, 46(3), 167–174.

Chib, S. (1996). "Calculating posterior distributions and modal estimates in Markov mixture models." *Journal of Econometrics*, 75(1), 79–97.

Cuhadar, A., Shalaby, K. H., and Tasdoken, S. (2002). "Automatic segmentation of pavement condition data using wavelet transform." *IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings (Cat. No.02CH37373)*, 2, 1009–1014 vol.2.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

Divinsky, M., Nesichi, S., and Livneh, M. (1997). "Development of a road roughness profile delineation procedure." *Journal of Testing and Evaluation*, (DR Petersen, ed.), 25(4), 445–450.

Erdman, C., and Emerson, J. W. (2007). "bcp: An R package for performing a Bayesian analysis of change point problems." *Journal of Statistical Software*, 23(3), 1–13.

Frühwirth-Schnatter, S. (1994). "DATA AUGMENTATION AND DYNAMIC LINEAR MODELS." *Journal of Time Series Analysis*, 15(2), 183–202.

Gao, J., and Johnson, M. (2008). "A comparison of Bayesian estimators for unsupervised hidden Markov model POS taggers." *Proceedings of the conference on empirical methods in natural language processing*, EMNLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA, 344–352.

Geman, S., and Geman, D. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721–741.

Haas, R., Hudson, W. R., and Zaniewski, J. P. (1994). *Modern pavement management*. Krieger, Malabar, Fla.

Hong, F., Perrone, E., Mikhail, M., and Eltahan, A. (2017). "Planning pavement maintenance and rehabilitation projects in the new pavement management system in Texas."

Jeffreys, H. (1998). *Theory of probability*. Oxford University Press, Oxford Oxfordshire : New York.

Kaplan, K., Kurtul, C., and Akin, H. L. (2010). "Fast lane tracking for autonomous urban driving using hidden Markov models and multiresolution Hough transform." *Industrial Robot: An International Journal*.

Kennedy, J., Shalaby, A., and Cauwenberghe, R. V. (2000). "Dynamic segmentation of pavement surface condition data."

Killick, R., and Eckley, I. A. (2014). "Changepoint: An R package for changepoint analysis." *Journal of Statistical Software*, 58(3), 1–19.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). "Optimal Detection of Changepoints With a Linear Computational Cost." *Journal of the American Statistical Association*, 107(500), 1590–1598.

Killick, R., Haynes, K., and Eckley, I. A. (2016). *changepoint: An R package for changepoint analysis*.

Kobayashi, K. (2010). "A Poisson Hidden Markov Model for Deterioration Prediction of Road Asset System."

Kobayashi, K., Kaito, K., and Lethanh, N. (2012). "A statistical deterioration forecasting method using hidden Markov model for infrastructure management." *Transportation Research Part B: Methodological*, 46(4), 544–561.

Lethanh, N., and Adey, B. T. (2012). "A Hidden Markov Model for Modeling Pavement Deterioration under Incomplete Monitoring Data."

Lethanh, N., and Adey, B. T. (2013). "Use of exponential hidden Markov models for modelling pavement deterioration." *International Journal of Pavement Engineering*, 14(7), 645–654.

Li, J., He, Q., Zhou, H., Guan, Y., and Dai, W. (2016). "Modeling Driver Behavior near Intersections in Hidden Markov Model." *International Journal of Environmental Research and Public Health*, 13(12).

Miller, N., Thomas, M. A., Eichel, J. A., and Mishra, A. (2015). "A Hidden Markov Model for Vehicle Detection and Counting." *2015 12th Conference on Computer and Robot Vision*, 269–276.

Misra, R., and Das, A. (2003). "Identification of homogeneous sections from road data." *International Journal of Pavement Engineering*, 4(4), 229–233.

Murphy, K. (2005). "Hidden Markov Model (HMM) Toolbox for Matlab." https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html.

Ping, W. V., Yang, Z., Gan, L., and Dietrich, B. (1999). "Development of procedure for automated segmentation of pavement rut data." *Transportation Research Record*, 1655(1), 65–73.

Qi, Y., and Ishak, S. (2014). "A Hidden Markov Model for short term prediction of traffic conditions on freeways." *Transportation Research Part C: Emerging Technologies*, Special Issue on Short-term Traffic Flow Forecasting, 43, 95–111.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE*, 77(2), 257–286.

Rao, V., and Teh, Y. W. (2013). "Fast MCMC sampling for Markov jump processes and extensions." *arXiv:1208.4818 [stat]*.

Scott, S. L. (2002). "Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century." *Journal of the American Statistical Association*, 97(457), 337–351.

Scullion, T., and Smith, R. E. (1997). "TxDOT's pavement management information system : Current status and future directions."

Song, W., Xiong, G., and Chen, H. (2016). "Intention-Aware Autonomous Driving Decision-Making in an Uncontrolled Intersection." *Mathematical Problems in Engineering*, Research Article, https://www.hindawi.com/journals/mpe/2016/1025349/.

Stampley, B. E., Miller, B. S., Smith, R. E., and Scullion, T. (1995). "Pavement management information system concepts, equations, and analysis models."

Thomas, F. (2001). "A Bayesian approach to retrospective detection of change-points in road surface measurements." Doctoral thesis, comprehensive summary, Stockholm University, Stockholm.

Thomas, F. (2003). "Statistical approach to road segmentation."

Thomas, F. (2005). "Automated road segmentation using a Bayesian algorithm."

Van Ravenzwaaij, D., Cassey, P., and Brown, S. D. (2018). "A simple introduction to Markov Chain Monte Carlo sampling." *Psychonomic Bulletin & Review*, 25(1), 143–154.

Wang, X., Peng, L., Chi, T., Li, M., Yao, X., and Shao, J. (2015). "A Hidden Markov Model for Urban-Scale Traffic Estimation Using Floating Car Data." *PLOS ONE*, 10(12), e0145348.

Yang, C., Tsai, Y., and Wang, Z. (2009). "Algorithm for spatial clustering of pavement segments." *Computer-Aided Civil and Infrastructure Engineering*, 24(2), 93–108.

Zou, X., and Levinson, D. (2006). “Modeling Intersection Driving Behaviors: A Hidden Markov Model Approach (I).”

Zucchini, W., MacDonald, I. L., Langrock, R., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov Models for Time Series : An Introduction Using R, Second Edition*. Chapman and Hall/CRC.