Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Endre Bjørge Urheim

Bayesian statistical modelling and analysis of a DAS data set

Master's thesis in Applied Physics and Mathematics
Supervisor: Håkon Tjelmeland
June 2023

NTNU
Norwegian University of
Science and Technology

Endre Bjørge Urheim

Bayesian statistical modelling and analysis of a DAS data set

Master's thesis in Applied Physics and Mathematics
Supervisor: Håkon Tjelmeland
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Preface

This project is my Master's thesis in Industrial Mathematics at the Norwegian University of Science and Technology, conducted in the spring of 2023, supervised by Professor Håkon Tjelmeland. The data set is provided by Centre for Geophysical Forecasting at NTNU. The work in this project is a continuation of my work in Urheim (2022), in the course *TMA4500 - Industrial Mathematics, Specialization Project*, performed in the autumn of 2022. The tasks are natural extensions of the work from this course, decided together by Professor Tjelmeland and myself. For reading this thesis, I recommend having basic knowledge about statistics, in addition to some knowledge about time series and how to calculate with probability distributions. However, new theory is explained or referred to when introduced.

I want to thank my supervisor Professor Håkon Tjelmeland, for guidance throughout this thesis. Our weekly meetings have been informative and exciting, and Professor Tjelmeland has shown great interest in my work. I would also like to thank Robin André Rørstadbotnen at Centre for Geophysical Forecasting for always being available and answering my questions regarding the data set. Finally, I am grateful for my fellow students and friends, as we have all been working towards an end goal simultaneously, which has given me motivation throughout the semester.

Endre Bjørge Urheim
Trondheim, June 2023

Abstract

Distributed acoustic sensing (DAS) is a system that uses fiber-optic cables as sensors to retrieve seismic information from the area around the cables. In this project, we analyze a DAS data set obtained by Centre for Geophysical Forecasting (CGF) at NTNU, Norway. The data set contains data collected over ten minutes along the railway tracks between Marienborg station and Støren, south of Trondheim, Norway. Our goal is to model probabilities of events around the railway tracks in a Bayesian framework, which can be used to detect possibly dangerous situations fast. The first part of the analysis is about obtaining structures in the data set for detecting events easier. We analyze differentiated time series and the autocorrelation function at lag one for small time subsets for the differentiated series. In that way, we restrict the amount of data to process without losing information about the signals across all positions and times.

For the modelling part, we assume Markov properties in space to account for dependence. This is used to construct hidden Markov models (HMMs) in spatial direction. We present four models, each representing versions of the general hidden Markov model, where the first three work as building blocks towards the fourth model, which is the most important in this project. For the first two models, we assume independence in time, while the third model introduces dependence between two adjacent layers in time. For the main model, we use the information from the data at previous times to obtain probabilities of events in the current time layer.

We find that the models manage to detect events well, and their results are relatively similar. Determining the hyperparameters in the models is important because the results are sensitive to the values of the hyperparameters. The detected events from the models correspond well to the observations from CGF.

Sammendrag

"Distributed acoustic sensing" (DAS) er et system som bruker fiber-optiske kabler som sensorer for å innhente seismisk informasjon fra området rundt kablene. I dette prosjektet analyserer vi et DAS-datasett fra Centre for Geophysical Forecasting (CGF) ved NTNU i Norge. Datasettet inneholder data samlet inn over ti minutter langs jernbanelinja mellom Marienborg stasjon og Støren, sør for Trondheim i Norge. Målet med prosjektet er å modellere sannsynligheter for hendelser rundt jernbanelinja med en Bayesiansk tilnærming. Dette kan brukes til å oppdage potensielt farlige situasjoner raskt. Den første delen av analysen handler om å identifisere strukturer i datasettet, for å gjøre det enklere å oppdage hendelser. Vi analyserer differensierte tidsrekker, og autokorrelasjonsfunksjonen mellom nabopunkter i mindre tidssett, innenfor tidsrekkene. På den måten begrenser vi mengden data vi behandler, uten å miste informasjon om signalene ved alle posisjoner og tider.

I modellene vi lager, antar vi Markov-egenskaper i rom for å modellere avhengighet. Dette brukes til å konstruere skjulte markovkjeder i romlig retning. Vi presenterer fire modeller, hvorav de tre første er byggesteiner for den fjerde modellen, som er den viktigste i dette prosjektet. For de to første modellene antar vi uavhengighet i tid, mens i den tredje modellen introduserer vi avhengighet mellom to og to lag i tid. I hovedmodellen bruker vi informasjon fra dataene ved tidligere tidspunkt for å beregne sannsynligheter for hendelser i nåtid.

Resultatene tilsier at modellene klarer å oppdage hendelsene godt, og resultatene er relativt like for alle modellene. Hvordan vi velger hyperparametere i modellene er viktig, fordi resultatene er sensitive til deres verdier. Hendelsene modellene oppdager samsvarer godt med observasjoner gjort av CGF.

Table of Contents

1 Introduction **1**

2 The DAS system **3**

2.1 Physical theory 3

2.2 Preprocessing 6

3 The Trondheim data **7**

3.1 The data set 8

3.2 Data transformation 11

3.2.1 Autocorrelation at lag one 13

4 Statistical theory **20**

4.1 Bayesian statistics 20

4.2 Empirical Bayes methods 22

4.3 Hidden Markov models 23

4.4 Forward-backward algorithm 25

4.4.1 The general forward recursion formula 27

4.4.2 The general backward recursion formula 28

4.5 Clique graphs 29

5 Models applied to the Trondheim data **31**

5.1 First-order HMM in space 31

5.2 Second-order HMM in space 33

5.3 First-order HMM in space and dependence in time 35

5.4 Main model 37

5.4.1 Concept of the model 37
5.4.2 Solution for the model 39

6 Algorithms **42**
6.1 First-order HMM in space 43
6.2 Second-order HMM in space 43
6.3 First-order HMM in space and dependence in time 44
6.4 Main model 45

7 Results and discussion **48**
7.1 Parameter estimation 48
7.2 First-order HMM in space 48
7.3 Second-order HMM in space 53
7.4 First-order HMM in space and dependence in time 55
7.5 Main model 57
7.6 Parameter values for the main model 58
7.6.1 Parameters in the likelihood 60
7.6.2 Conditional probabilities in time 63
7.6.3 Further discussion 64

8 Concluding remarks **65**

Bibliography **67**

Appendix **70**
A Standard deviations in time sets 70
B ACF at lag 2 and 3 70

1 Introduction

Distributed acoustic sensing (DAS) is a system that uses fiber optic cables to record seismic activity. It is part of a class of techniques called distributed fiber optic sensing (DFOS) (Zhan 2020), which share the use of fibers in the cables as the sensors instead of external devices. One of the main advantages of the DFOS techniques is the high density of the sensors because they do not need external sensors placed around the system (Kislov and Gravirov 2022). In addition, the collection of data can happen over great distances almost instantaneously, since it is based on the phase of backscattered light in the fibers, which means that the information is transferred with the speed of light. DAS devices can be connected to different kinds of fibers, including dark (unused) fibers, which makes the system relatively inexpensive, because the DAS device is the only new part necessary for data collection (Zhan 2020).

According to Zhan (2020), oil companies were the first to develop the DAS system to explore seismic activity. However, the technology has later been used in several domains such as underwater positioning and earthquake monitoring (Shang et al. 2022). A case study from Taweesintananon et al. (2021) demonstrates how the DAS technology can be applied to optical cables in the Trondheimsfjord, Norway, for constructing seismic images. In parts of the study, they compare the data obtained by the DAS system to data collected by another seismic system. Despite the greater amount of background noise in the DAS data, they conclude that the data qualities from the two systems are approximately the same. Since the DAS system can record over large distances almost instantaneously, it has some advantages in this situation.

Along the railway tracks in the area south of Trondheim, there are dark fibers. Personnel from Centre for Geophysical Forecasting (CGF) attached a DAS device to such fibers at Marienborg station in Trondheim and recorded the seismic activity along the approximately 51 km long distance to Støren, south of Trondheim. The goal with this project is to create statistical models to detect events in the DAS data set from this experiment. We call this data set the Trondheim data. The Trondheim data include preprocessed DAS data, but we understood that doing analysis on the preprocessed data was not appropriate due to random noise and sensitive data. Since the Trondheim data includes points in time and space, we look at differenced data in each time series at every spatial position. There we seem to have a more defined structure, and more potential for finding interesting results in the data set. Then we consider autocorrelations in time between the differenced time points in small time sets, and define Bayesian models including dependence in space and time. Initially, first- and second-order hidden Markov

models in space are fitted, before dependence in time is introduced. These models are the building blocks for developing the main model, where we use dependence in time and space, and use the information at earlier times when calculating probabilities for current events.

There are several ways to approach the analysis of a DAS data set. One possibility involves using all the available data to analyze specific situations that are already known. This approach may include data points at a later time than the occurrence of the event, and they can be used to learn how events look in the data set. However, in this project, we take a different approach by using previous and current observations to detect events as they happen. Therefore, the algorithms must be efficient, to detect events rapidly. A final goal could be to create a system where new data is loaded directly into the algorithm, which calculates the probability of a data point being an event, such as cars or animals crossing or trees falling on the railway tracks.

There have been some previous studies on event detection using DAS data, but with the utilization of machine learning techniques (Shiloh et al. 2019). According to Shiloh et al. (2019), developing efficient algorithms for detecting and classifying events in the areas around fiber optic cables is extremely important. They argue that DAS systems are attractive for this purpose due to their high density of sensors and rapid data update rate over tens of kilometres. As more DAS systems are deployed continuously, detection algorithms with DAS data must keep pace with the evolution. Shiloh et al. (2019) propose a method using generative adversarial networks (GAN) for detecting and classifying events. They conclude with a proof that the concept of using deep learning methodologies is possible for these purposes.

This project is an extension of Urheim (2022) from the course TMA4500 - Industrial Mathematics, Specialization Project. Urheim (2022) analyze the autocorrelation functions at lag one for small time sets, and make different models, including a Bayesian model where they assume the points to be independent of each other. However, this approach is a simplification, since events such as trains moving or cars crossing the railway tracks occur over multiple points in time and space. The exploratory analysis and the Bayesian model in Urheim (2022) are important foundations for this project, as they accomplish a lot of important time-consuming results.

This thesis is structured as follows. In Section 2 we explain the physical theory behind the DAS system and the preprocessing steps necessary to analyze the data. Section 3 presents the Trondheim data and the transformations we apply before doing the analysis. Further, in Section 4 we go through the theory behind

the statistical methods we use in this project, before we present four different models that we apply to the Trondheim data in Section 5. The algorithms that follow the four models are described in Section 6. After this, in Section 7, we present the results from applying the four models to the Trondheim data, focusing on the main model. In addition, we discuss the results and the affection of hyperparameters to the results. Finally, we conclude the project and give examples of how this analysis and the Trondheim data can be investigated further in Section 8.

2 The DAS system

In this section, we present the physical theory of how the DAS system works. Additionally, we explain the necessary preprocessing steps, to analyze the data. CGF applied the theory presented in this section to the data set before we received it.

2.1 Physical theory

In the DAS system, laser beams are sent through the fibers, with some of the light reflecting due to Rayleigh backscattering (Taweesintananon et al. 2021) because of strain in the fibers. Strain in the fibers is a result of seismic activity. The system constantly sends out laser beams at a high rate and measures the phase change between the backscattered light at the previous sweep and the backscattered light at the current sweep, at equally spaced positions along the fiber. In that way, the DAS system can measure changes in the fibers over time, over large distances almost simultaneously (Liu et al. 2017).

Figure 1 shows the three main parts of how a DAS system is typically constructed. The system is described in SEAFOM (2018). The first part is the distributed sensor, which is the sensor in the system. In this case, the distributed sensor is the dark fiber in the cables running along the railway tracks. Attached to the sensor is the second part, which is the interrogation unit (IU). The IU sends out light and records the phase change between the light sent out and the backscattered light, at every position. It also converts this phase change into fiber strain, which is the physical unit in the Trondheim data. We denote the fiber strain ϵ. The preprocessing steps of calculating the fiber strain from the phase change are described in Section 2.2. The last main part of the system consists of three smaller parts, with different tasks. The processor processes and stores the data in

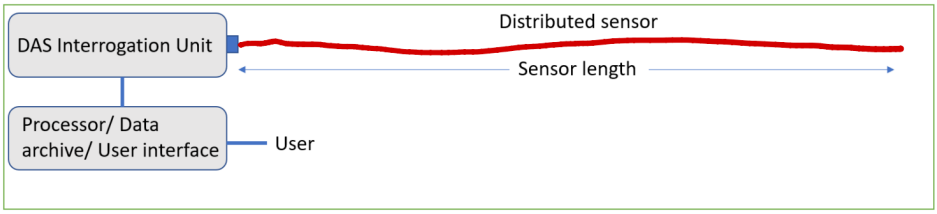


Figure 1: These are the three parts that constitute the DAS system. The illustration is taken from Urheim (2022).

the data archive, and the user interface has an interface where a user can interact with the system.

Haukanes (2021) describes in detail how the IU handles the backscattered light before it is preprocessed. The description includes calculations that use parameters specified in the DAS setup, called signal parameters. The IU sends out light with a propagation delay between spatial points of $\Delta\tau$ seconds, corresponding to a spatial sampling interval of

$$dx = \frac{c}{2 \cdot refractiveIndex \cdot fiberOverLength} \Delta\tau, \tag{1}$$

where $c \simeq 3,0 \cdot 10^8 m/s$ is the speed of light in vacuum, and *refractiveIndex* and *fiberOverLength* contain some information about the properties of the fiber. The light beams are sent out every dt seconds. The backscattered light is used to calculate the phase change between the current phase and the phase of the backscattered light of the previous light beam, dt seconds earlier, for each spatial sample separated by dx meters. The phase change at index i, located at $i \cdot dx$ is denoted $\dot{\Phi}_i$, which has unit $[rad/s]$. Since the amount of data rapidly becomes large, the IU saves data at every fourth spatial sample, and these are referred to as channels. Figure 2 shows the spatial and temporal sampling points, and explains some of the signal parameters mentioned in this section.

Further, a spatial moving average is applied to the phase change over $nAvgTau$ samples, which corresponds to $nAvgTau \cdot dx$ in meters. The resulting averaged phase change centered in $i \cdot dx$ is denoted $\dot{\Phi}_{avg(i)}$. The reason for applying a spatial moving average to $\dot{\Phi}_i$ is to avoid an effect which is called Rayleigh fading, further described in Sklar (1997).

Finally, the IU calculates the differential phase

$$\dot{\phi}_i = \dot{\Phi}_{avg(i+nDiffTau)} - \dot{\Phi}_{avg(i)}, \tag{2}$$

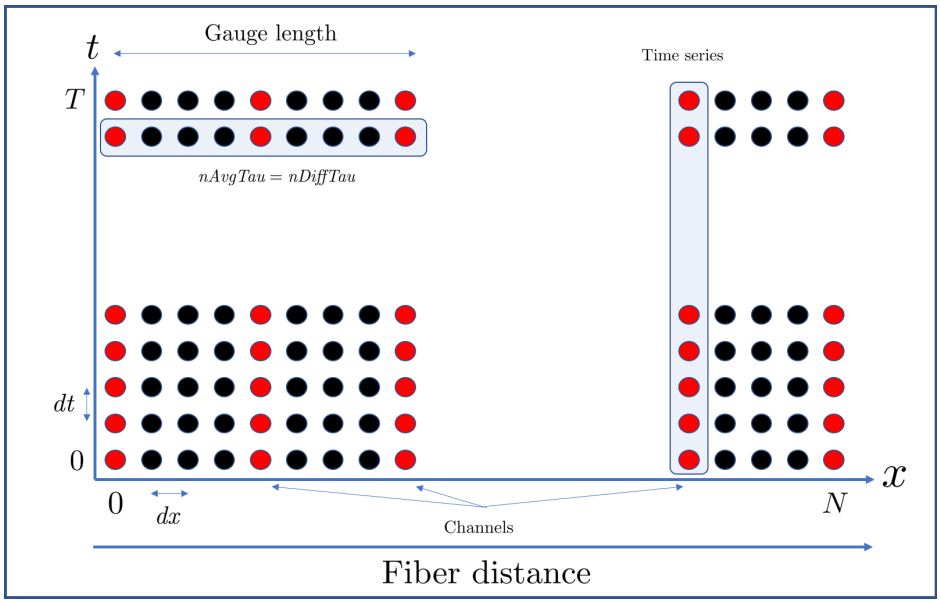


Figure 2: The signal parameters and an explanation of the data collection. The x-axis run along the fiber, and there are N samples in space. The IU collects data at each channel (marked with red dots) which is every fourth dx, and every index in time, separated by dt, with a total of T points. In this system, the gauge length is $nDiffTau \cdot dx$, where $nDiffTau = nAvgTau = 8$. Each series of red dots constitutes the time series at the specific channels. The illustration is inspired by a corresponding illustration in SEAFOM (2018).

between two averaged phase changes separated by *nDiffTau* spatial samples. Here, *nDiffTau* is a specified constant that describes the number of spatial samples we differentiate over. The reason for taking the difference between two points is that the phase in one point alone is unsuited for seismic analysis (Dean et al. 2016). The physical distance between the two spatial samples, separated by *nDiffTau*, is called the gauge length (GL), and this is an important parameter for a DAS experiment (Dean et al. 2016). Too high gauge length leads to poor resolution of the signals, while too low gauge length results in a bad signal-to-noise ratio. The formula for the gauge length used in this experiment is

$$GL = \frac{c \cdot \Delta\tau}{2 \cdot refractiveIndex} \cdot nDiffTau. \tag{3}$$

The differential phases $\dot{\phi}_i$ at each position i are the values stored in the data set before the preprocessing steps described in Section 2.2, are applied.

2.2 Preprocessing

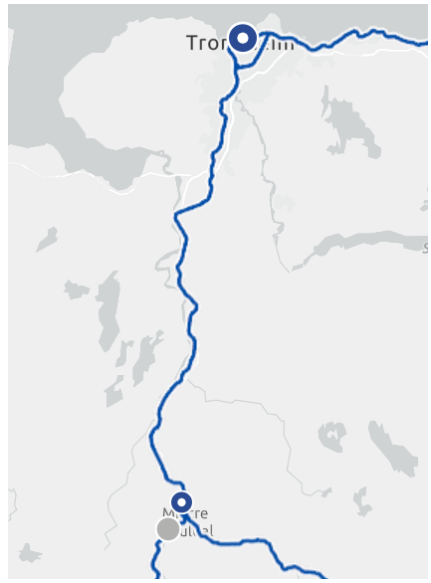
The preprocessing steps convert the differential phase $\dot{\phi}_i$ at spatial position i into fiber strain ϵ_i, which is the physical quantity we want to analyze. The procedure for this conversion is described by Haukanes (2021). In most cases, there are three necessary steps for converting the differential phase into fiber strain.

The first step is to scale the differential phase, to get the unit $[rad/m/s]$. We call this experiment-specific scaling factor *dataScale*. After the scaling, the phase is wrapped in a range between $[-spatialUnwrRange/2,\ spatialUnwrRange/2)$. Therefore, the next step is to unwrap the phases which go more than one time around the unit circle, see (Li et al. 2018; Padilla et al. 2023) for a detailed explanation. Finally, the data is integrated over time to get the unit $[rad/m]$. The resulting quantity is the phase change over one gauge length centered in i, denoted ϕ_i. After this, we want to transform the radians into a physical measure.

Haukanes (2021) and Taweesintananon et al. (2021) describe that the resulting phase change over one gauge length in units $[rad/m]$ has a linear relationship with the strain in the fiber ϵ. The unit for the strain is $[\varepsilon]$, which is the standardized measurement for DAS performance parameters (SEAFOM 2018). The linear relation between the phase change ϕ_i and the fiber strain ϵ_i is

$$\epsilon_i = \frac{\phi_i}{sensitivity}, \tag{4}$$

(a) Map of the area from Google Maps (2023). Marienborg station is pinned by the black and white symbol between Trondheim and Sluppen.



(b) Map of the railway system from Bane NOR (2023). Støren is located at the large blue pin on the bottom of the map.

Figure 3: Map of the area (a) and the railway system (b) between Marienborg station and Støren. The railway tracks run along the road E6 in this area.

where the constant *sensitivity* depends on some properties of the fiber and the wavelength of the laser beams.

3 The Trondheim data

The following section presents the Trondheim data, and we go through the transformations we apply to obtain more structure in the data. The Trondheim data was collected on September 1st, 2021, by CGF using a DAS device on a dark fiber running along railway tracks from Marienborg station to Støren. A map of this geographical area, together with the railway system are displayed in Figure 3.

Constant	Value
dt	$5 \cdot 10^{-4} s$
$\Delta\tau$	$10^{-8} s$
refractiveIndex	1.47
fiberOverLength	1
nAvgTau	8
nDiffTau	8
DataScale	$1.43 \cdot 10^{-6} s^2/m$
spatialUnwrRange	6152.14
sensitivity	9362208.90 $rad/m/\varepsilon$

Table 1: Signal parameters in the experiment where the Trondheim data is collected. Some of them do not have units.

3.1 The data set

Table 1 shows the values for the signal parameters in (1) and (3), in addition to *dt*, *spatialUnwrRange*, *dataScale* and *sensitivity* for this experiment. Thus, the spatial sampling interval in (1) becomes

$$dx = \frac{3 \cdot 10^8}{2 \cdot 1.47 \cdot 1} \cdot 10^{-8} \simeq 1.02m. \tag{5}$$

We can see from (1) and (3), with *fiberOverLength* $= 1$, that $GL = dx \cdot nDiffTau$. Thus, the gauge length in (3) becomes

$$GL = \frac{3 \cdot 10^8 \cdot 10^{-8}}{2 \cdot 1.47} \cdot 8 \simeq 8.17m. \tag{6}$$

The Trondheim data is stored in files that each span ten seconds in the time direction and cover a distance of approximately 51km along the railway tracks in the spatial direction. The DAS device records data every $dt = 0.0005$ second, resulting in 20,000 data points in the time direction per file. As mentioned above, the spatial sampling interval dx is equal to $1.02m$. In the spatial direction, data is collected every fourth spatial sample, which is every $4 \cdot 1.02m = 4.08m$. These spatial sample locations where data is collected are referred to as channels. With a data range over 51km, there are a total of 12500 channels with time series. As a result, each file contains $20,000 \times 12,500$ data points, and we have access to 61 files covering the time interval between 13:30:03 and 13:40:13. Due to the considerable amount of data, we limit the analysis to data points recorded

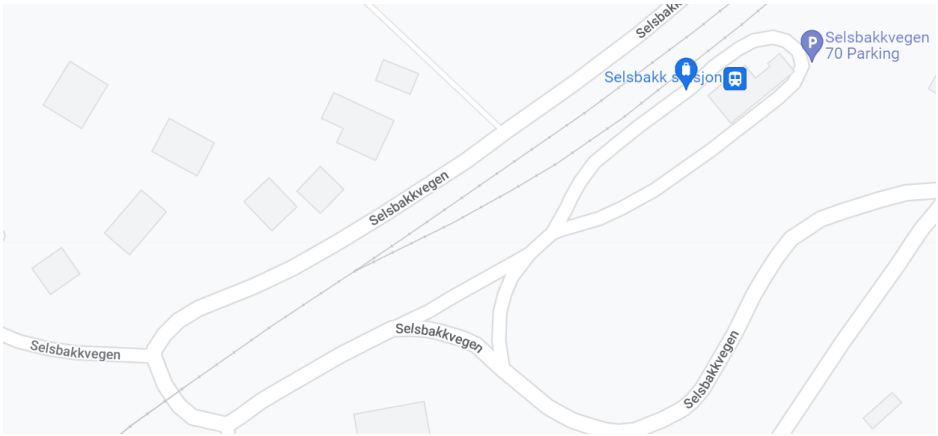


Figure 4: The area around Selsbakk station from Google Maps (2022). The railway tracks are the thin grey lines with evenly spaced marks.

between 13:34:53 and 13:36:53, with one exception in the data exploration in Section 3.2.

The most essential location in the Trondheim data is Selsbakk station, located approximately $4000m$ along the railway tracks. The area is displayed in Figure 4, consisting of the station, and a bridge over a road. Here, personnel from CGF have noted the times of cars crossing under the railway tracks, in addition to experimenting with personnel jumping up and down on the station. With that as help, we can learn what these types of events look like in the data set and understand how we can classify data points as events.

Figure 5 shows the fiber strain between 3000 and 5500 meters along the railway tracks and between 13:34:53 and 13:36:53. The highest and the lowest strain values in absolute value are truncated to display some of the events better. In the heatmap, we can see large vertical lines in red and blue. At around $4000m$, where personnel from CGF have taken notes, these large vertical lines do not relate to anything they have reported. This indicates that the values in the data set differ a lot, and most of these signals are not interesting for our analysis, as they are not related to any events. At approximately 4000 meters, we can also see small diagonal lines and, if examined closely, small horizontal lines for the first 20 seconds. In addition, even more indistinct horizontal lines can be observed at approximately 50 seconds after 13:34:54 at Selsbakk station. These observations match the experiments and notes by the personnel from CGF on people jumping and cars crossing.

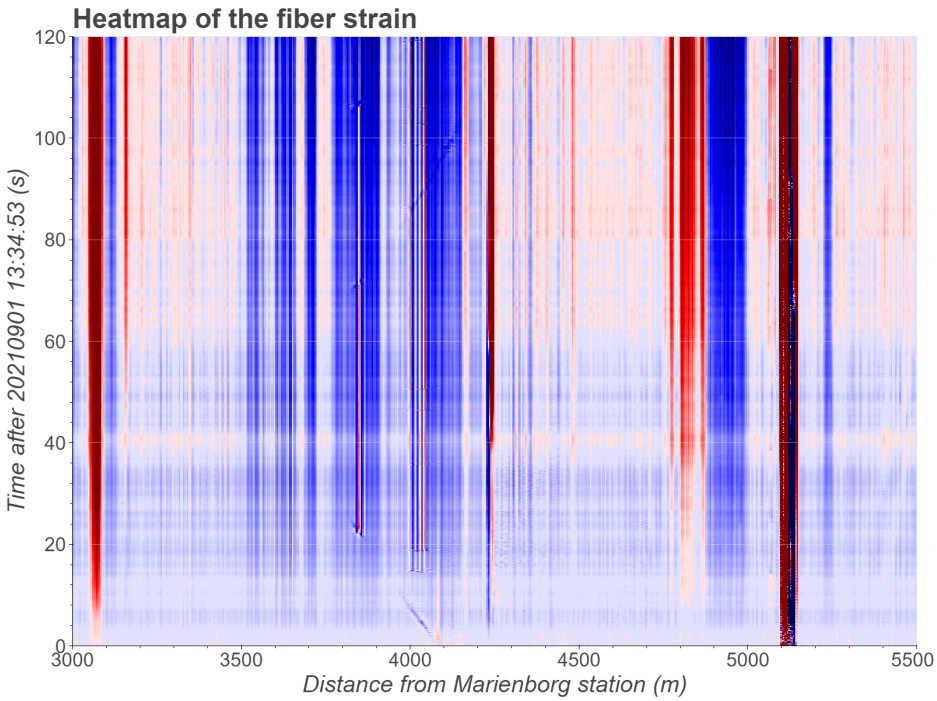


Figure 5: Fiber strain ϵ between 3000 and 5500 meters, and for two minutes after 13:34:53, displayed as a heatmap. The values of ϵ are truncated between $[-90 \cdot 10^{-9}\varepsilon,\ -5 \cdot 10^{-9}\varepsilon]$ and $[5 \cdot 10^{-9}\varepsilon,\ 90 \cdot 10^{-9}\varepsilon]$. We have included every hundredth time point, as they are sufficient for showing the structures in the data set.

As the recording of data has a high sensitivity (Taweesintananon et al. 2021), the Trondheim data contains a considerable amount of noise (Isken et al. 2022), which is evident from Figure 5. Therefore, a standard procedure is to apply filtering to the data set (Ma et al. 2022; Taweesintananon et al. 2021), and remove high and low frequencies, since much of the noise is assumed to exist due to these frequencies. Initially, we attempted to filter out frequencies lower than 5 Hz and higher than 90 Hz by using a Butterworth filter (Butterworth 1930). However, this removed many interesting signals in addition to the noise. A lot of the observed unwanted noise are removed by the data transformation described in Section 3.2 instead, which we consider sufficient.

3.2 Data transformation

Now, we discuss the transformations we apply to the Trondheim data. They follow the same procedure conducted in Urheim (2022). They aim to remove the unwanted signals and noise described in Section 3.1 from the Trondheim data, and to obtain clear structures in the data. In the following, we presume that the reader has some prior knowledge about time series, especially the mean, autocovariance and autocorrelation functions, stationarity, and how we can transform a time series from non-stationary to stationary by applying differences. We refer to Shumway and Stoffer (2017) for the complete background theory. We perform the transformations by applying theory from the field of time series, since the Trondheim data contains time series at each spatial sampling location.

We denote the time series in the Trondheim data $\{\epsilon^t; t = 1, 2, \ldots, T\}$, where ϵ^t is the strain in the fiber, and t is the time index. Figure 6 shows four time series $\{\epsilon^t; t = 1, 2, \ldots, T\}$ from different channels in the Trondheim data. We see that they are not stationary as their means are not constant. Stationarity is important because it indicates that the process behaves with some regularity over time, which we desire when doing analysis. In this study, we want to detect events that affect the fiber strain ϵ in the Trondheim data, and therefore, having regularity in the data can make it easier to sort out the real events from noise.

Transforming a non-stationary time series into a stationary one can be achieved using differences. Thus, we want to apply differences to $\{\epsilon^t; t = 1, 2, \ldots, T\}$ at every channel to see if we get stationary time series. Differences of order k are denoted

$$\Delta\epsilon^t = \Delta^{(1)}\epsilon^t = \epsilon^t - \epsilon^{t-1}, \tag{7}$$

$$\Delta^{(k)}\epsilon^t = \Delta^{(k-1)}\epsilon^t - \Delta^{(k-1)}\epsilon^{t-1}. \tag{8}$$

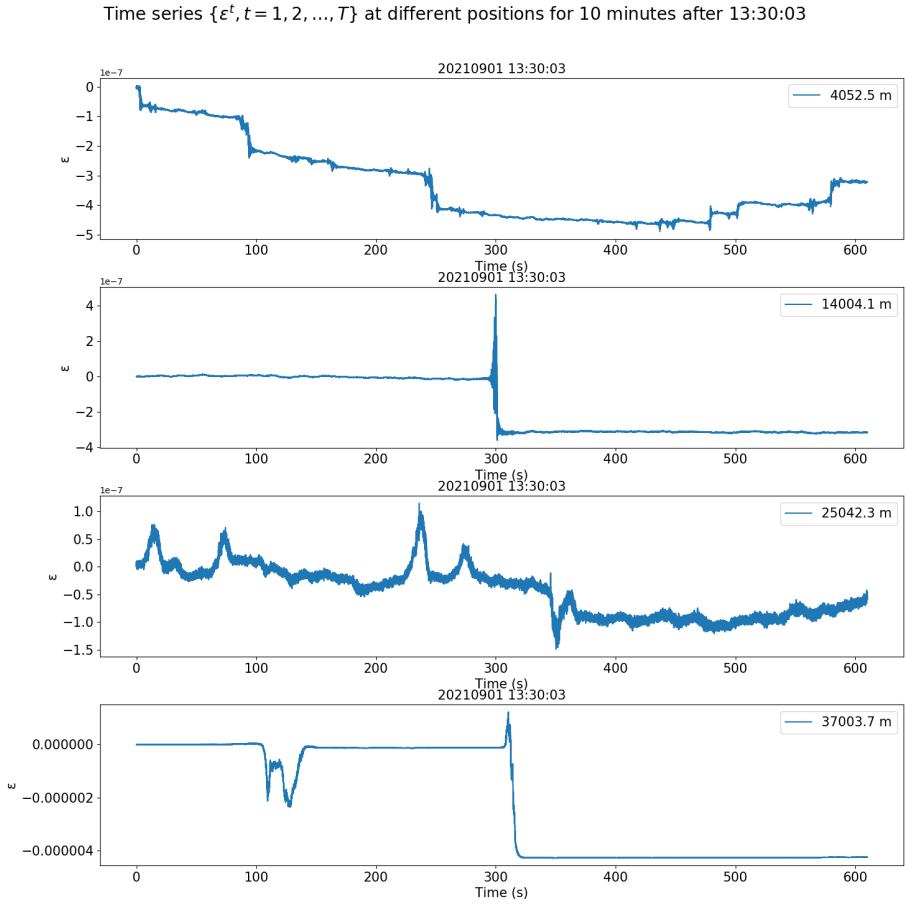


Figure 6: Four time series with fiber strain $\{\epsilon^t, t = 1, 2, \ldots, T\}$ at four different locations for ten minutes after 13:30:03. Note the different vertical scale in the four plots.

Figure 7 shows what happens to the time series in Figure 6, after applying first-order differences from (7). At several positions and times, we observe spikes in the differenced time series, and for the series at $4052.5m$, the events reported by CGF match the spikes. We can see that the time series appear to be stationary everywhere there are no such spikes. By inspecting Figure 7, we observe that the mean seems to be constant and approximately equal to 0 at every point in time, which is expected since we look at a differenced series. In addition, we see that the autocovariance function only seems to depend on two time points t and $t + h$ through the difference h at the times where there are no spikes, because the time series behave similarly during periods of length h. Since we make these observations at different locations, we assume that the first-order differenced time series $\{\Delta\epsilon^t; t = 1, 2, \ldots, T\}$ in the Trondheim data are stationary wherever no events take place. This implies regularity over time. Everywhere there are events, we make no assumptions about the time series.

Figure 8 displays first-order differenced time series for the fiber strain between 3000 and 5500 meters in a heatmap. Some of the structures from Figure 5, like the small horizontal lines at 4000 meters are noticeable. In addition, there seem to be fewer dominant signals overshadowing other signals than we observed in Figure 5. This observation matches the expectation from the first-order differenced time series in Figure 7. In addition we see a large vertical line at 5100 meters. However, several reported signals from CGF at Selsbakk station are still hard to detect.

3.2.1 Autocorrelation at lag one

We have obtained stationary time series where there are no events from first-order differences at every channel. Now, we desire to have some quantity that catches the structures in the data which we can analyze. The differenced series include more time points than we consider necessary, and processing them is hard computationally. By simply excluding many of them, we can remove interesting signals in the data set. Hence, we want a quantity summarizing the information in smaller time sets. Therefore, we divide the series into small time sets. We believe that an event is happening over some time, and therefore we decide that the small time sets contain a quarter of a second of data. Since the time difference between two consecutive samples is $dt = 0.0005s$, each time set contains 500 points. The mean of a time set for the differenced series is not interesting to investigate because when we sum over differenced data points, every point cancels out except for the first and the final. We tried the standard deviations of each time set, observing some structures, but the signals are weak. We refer to Appendix A for a heatmap of the standard deviations of the time sets. Another potential measure that summarizes the information in the time sets is the autocorrelation

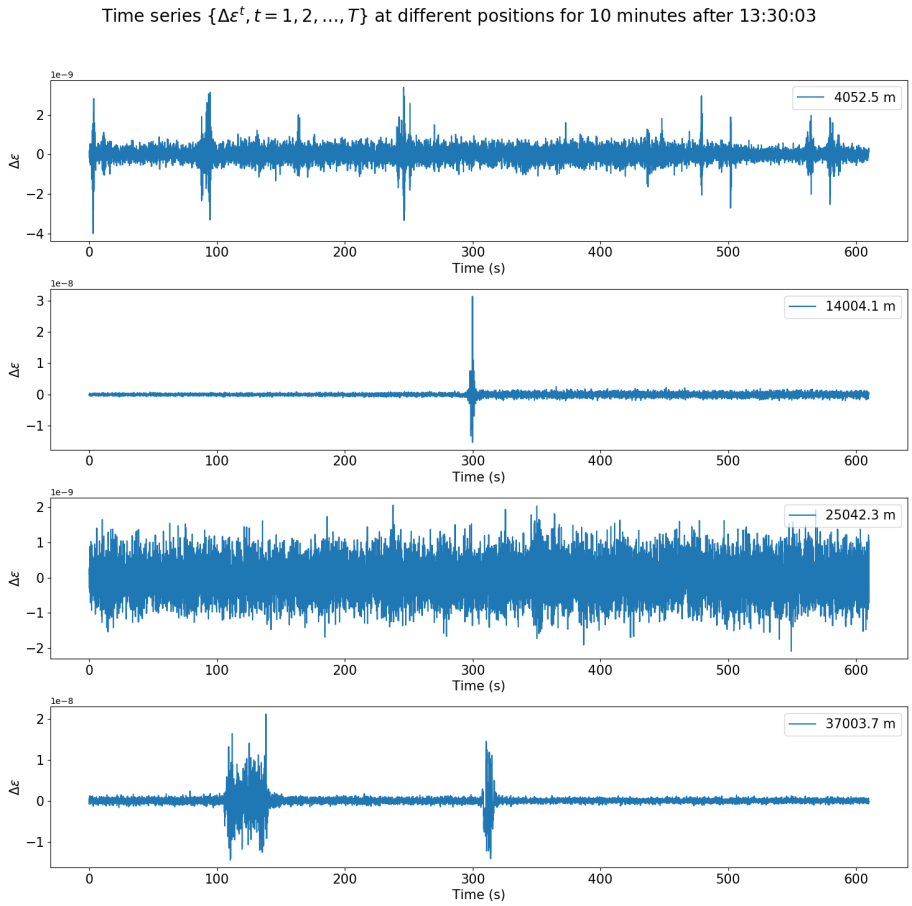


Figure 7: Four time series with first-order differenced fiber strain $\{\Delta\epsilon^t, t = 1, 2, \ldots, T\}$ at four different locations for ten minutes after 13:30:03. Note the different vertical scale in the four plots.

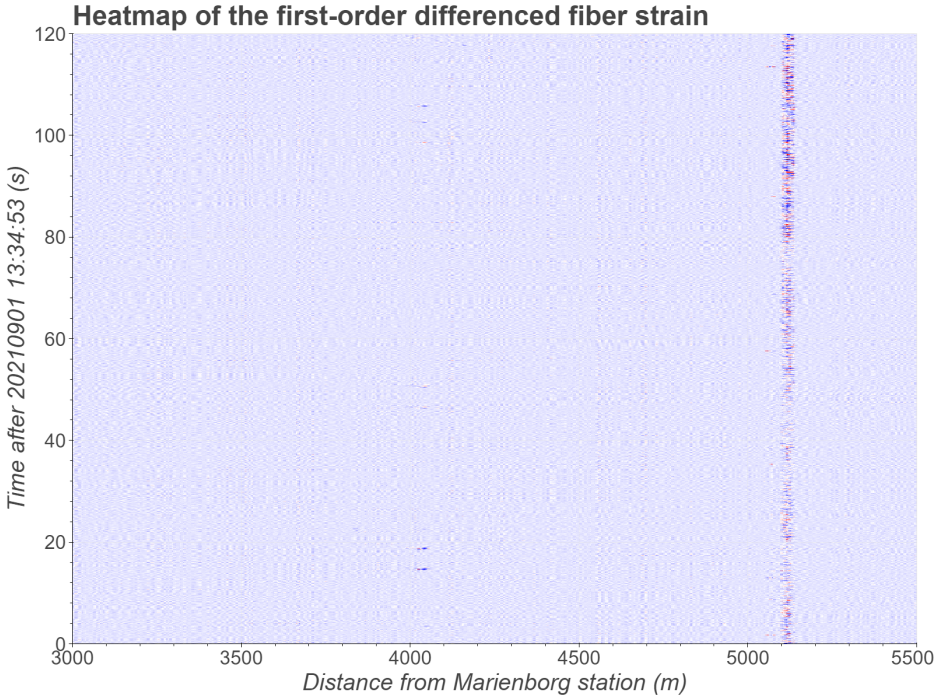


Figure 8: First-order differenced fiber strain $\Delta\epsilon^t$ between 3000 and 5500 meters, and for two minutes after 13:34:53, displayed as a heatmap.

function (ACF) at different lags h, for each time set. The ACF at lag h describes how ϵ^{t+h} can be predicted linearly from ϵ^t, and it is bounded between -1 and 1. High or low values for the ACF mean high positive or negative correlation, respectively, between ϵ^t and ϵ^{t+h}. Since there is one ACF value for each time set, the number of temporal points is reduced to 40 per ten seconds, which is 480 points in time for two minutes. We believe this is sufficient for obtaining a grid of points that is not too coarse and is suitable for event detection.

Figure 9 shows the ACF values at the first 15 lags for four different time sets at the same position as the first plots in Figures 6 and 7, which is at Selsbakk station. Personnel from CGF have jumped up and down at Selsbakk station in the time interval we have used in the second plot in Figure 9. In this second plot, we can see what happens to the structure of the ACF during a jump, and it deviates from the other, showing high correlation between points several lags apart. The three other plots show times when personnel from CGF have not reported anything. Figure 10 shows the autocorrelations at $h = 1, 2, 3$ at Selsbakk station. We only include $h = 1, 2, 3$ as we assume that the time points closest to each other are the most related, since we have no previous knowledge about seasonal patterns in the time series. For $h = 1$, the ACF fluctuates around approximately -0.5 where nothing is assumed to happen, indicating that the points closest to each other have a slight negative correlation. For $h = 2$ and $h = 3$, the ACF fluctuates around 0, suggesting a minimal correlation between time points more than one step away from each other, where nothing is happening.

Figure 10 indicates that the ACF values at different lags give the same information from the Trondheim data, since the spike patterns are similar for all three. ACF values at lag one for the same area as in Figures 5 and 8 are displayed as a heatmap in Figure 11. We refer to Appendix B for heatmaps of ACF at lag two and three for the same area, confirming the suspicion that they give the same information. Therefore, we continue with the lag one values, as Figure 10 shows that the spikes are more distinct and easier to recognize than for the other lag values. We easily notice the events in the heatmap in Figure 11, and they propagate over multiple points in space and time. At Selsbakk station at approximately 4000 meters, we can see the horizontal lines which are jumps performed by the personnel from CGF. In addition, we see the diagonal patterns at Selsbakk station, which are cars driving under the railway tracks. Furthermore, there is one large vertical line at 5100 meters, which might come from a construction site, or another source that can create signals constant in time. We also notice some other sources that create periodic signals in time at around 3150 and 4550 meters. Finally, we notice smaller signals in Figure 11, which might be small events or noise due to small movements in the ground. It is also easy to notice that a lot of the noise in Figure 5 have been significantly reduced with the ACF

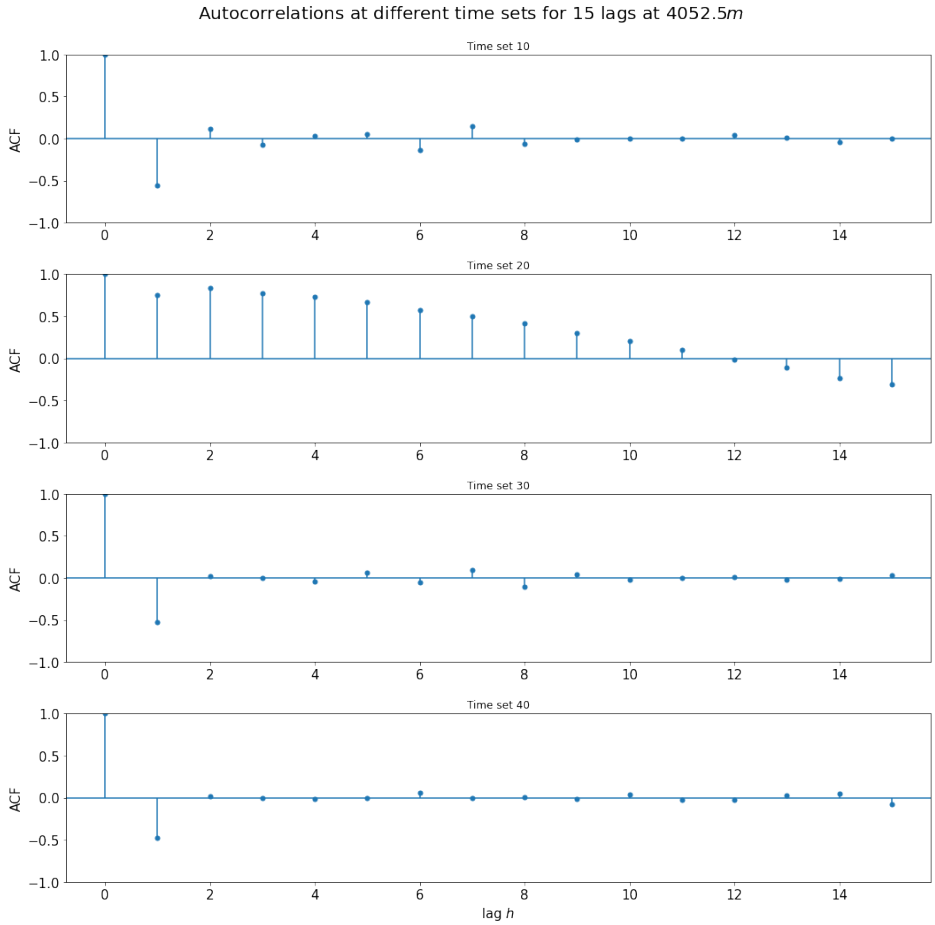


Figure 9: ACF at lag $1, \ldots, 15$ for four different time sets at Selsbakk station.

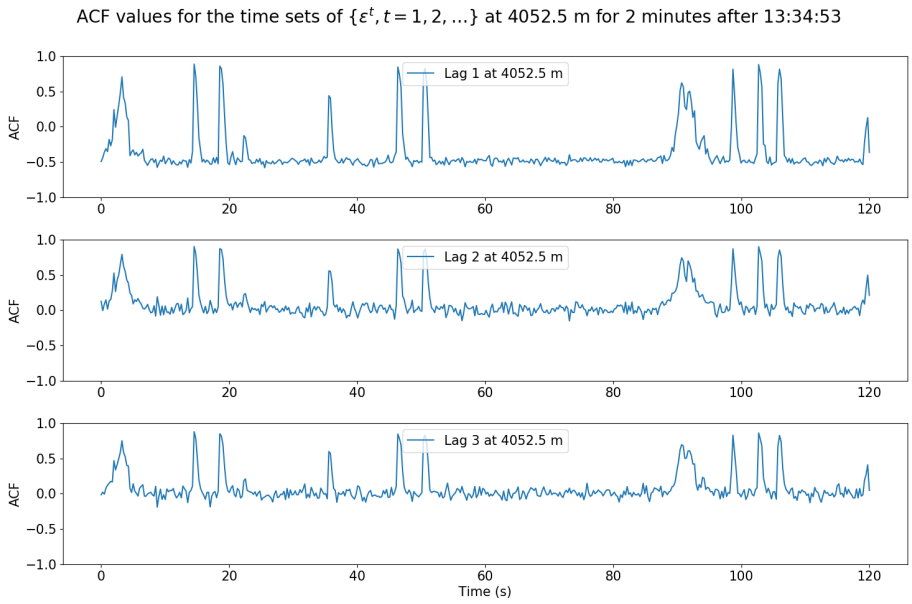


Figure 10: ACF at lag one, two and three for the time sets in two minutes at Selsbakk station.

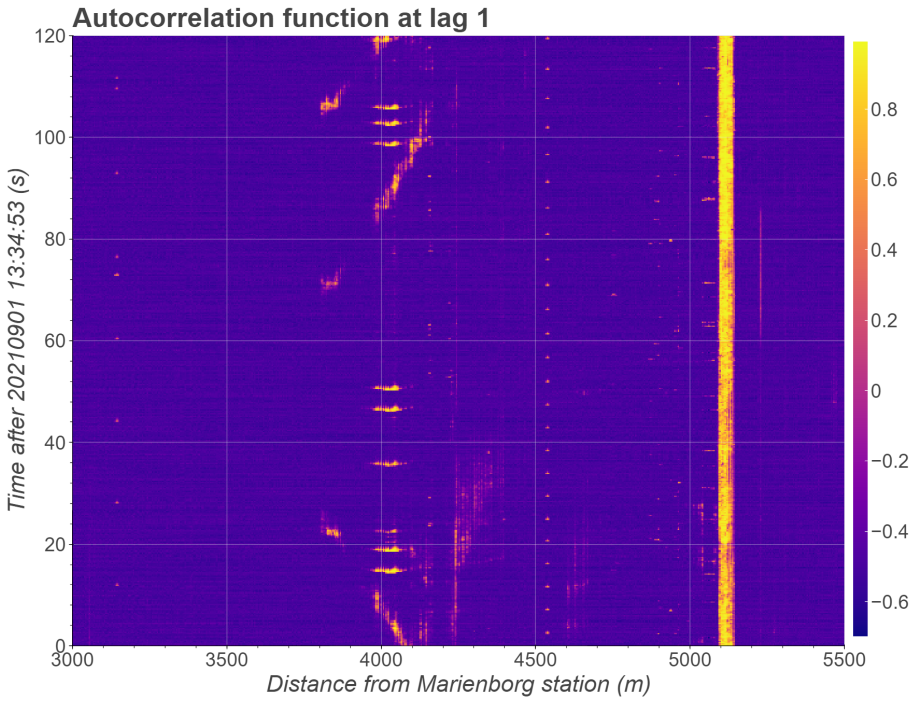


Figure 11: Heatmap of the autocorrelations at lag one between 3000 and 5500 meters, and between 13:34:53 and 13:36:53.

values at lag one, which is one of the things we wanted to achieve with the data transformation.

As we mentioned above, the ACF measures how well ϵ^{t+h} can be predicted based only on the value of ϵ^t through a linear relationship $\epsilon^{t+h} = a\epsilon^t + b$. If ϵ^{t+h} can be predicted perfectly through this linear relationship, then the ACF at lag h is positive or negative one, depending on the sign of a. The ACF values are, by definition, bounded between -1 and 1 due to Cauchy-Schwarz inequality. For the modelling in Section 5, we scale and shift the ACF-values at lag one to be in the interval $[0, 1]$, because we want to use a probability distribution that takes values in $[0, 1]$. These values are denoted y_i for spatial index i. Each y_i has R points in time, one for each time set, where each set contains information from 500 data points in time. We denote the collection of all y_i in spatial direction $\mathbf{y} = \{y_1, \ldots, y_i, \ldots, y_N\}$. Now that we have a quantity that reduces the size of the data set and seems to preserve most of the relevant information in the original

data set, in addition to removing a lot of noise and unwanted signals, it is easier to analyze the information in the data set. Thus, y_i is used for modelling the DAS data set throughout the project.

4 Statistical theory

Here, we present some statistical theory we use when modelling the transformed data from Section 3.2. First, we present the Bayesian approach to statistics, which is the framework for the models we make in this project. Next, we explain the theory of how we estimate the parameters in this project using empirical Bayes estimators. Then, we show the general hidden Markov model structure, which is the base for our models. Further, we describe the forward-backward algorithm, which is a powerful tool for computing specific joint distributions over large amounts of data. Finally, we discuss some theory on clique graphs in general graphs.

4.1 Bayesian statistics

In this section, we describe the Bayesian approach to statistics, which is the foundation for the methods in this project. A detailed introduction to Bayesian statistics can be found in Casella and Berger (2002).

The fundamental difference between frequentist and Bayesian statistics is how a parameter θ, describing the nature of the data, is interpreted. In the frequentist approach, θ is believed to be a fixed quantity, which is unknown. Then, a sample $\mathbf{y} = \{y_1, \ldots, y_N\}$ of size N is drawn from a population described by θ, to get knowledge of the unknown parameter. On the other hand, in the Bayesian approach, the parameter θ is considered a stochastic variable that follows a probability distribution. This distribution is called the prior distribution, and is independent of the data. Therefore it should be decided before the data is observed. We denote the prior distribution $p(\theta)$. The goal is to update the prior distribution with information gained from observing the data. This information is contained in the likelihood of the data, $p(\mathbf{y}|\theta)$. The updated distribution for the parameter is called the posterior distribution, and it is computed using Bayes' theorem (Joyce 2021),

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta) \cdot p(\theta)}{p(\mathbf{y})} = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})}, \tag{9}$$

where $p(\mathbf{y})$ is the marginal distribution of $\mathbf{y}$. This marginal distribution depends on whether θ is continuous or discrete. In the continuous case, θ must be integrated out of the joint distribution $p(\theta, \mathbf{y})$,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\theta) \cdot p(\theta) d\theta, \tag{10}$$

while for the discrete case, the integral is replaced with a sum,

$$p(\mathbf{y}) = \sum_{\theta} p(\mathbf{y}|\theta) \cdot p(\theta). \tag{11}$$

In the case where $\theta \in \{0, 1\}$, Bayes' theorem in (9) becomes

$$p(\theta = 1|\mathbf{y}) = \frac{p(\mathbf{y}|\theta = 1) \cdot p(\theta = 1)}{p(\mathbf{y}|\theta = 1) \cdot p(\theta = 1) + p(\mathbf{y}|\theta = 0) \cdot p(\theta = 0)}, \tag{12}$$

$$p(\theta = 0|\mathbf{y}) = \frac{p(\mathbf{y}|\theta = 0) \cdot p(\theta = 0)}{p(\mathbf{y}|\theta = 1) \cdot p(\theta = 1) + p(\mathbf{y}|\theta = 0) \cdot p(\theta = 0)}. \tag{13}$$

In this setting, we may use the Bernoulli distribution as prior for θ, which is on the form

$$p(\theta) = \gamma^{\theta} \cdot (1 - \gamma)^{1-\theta}, \tag{14}$$

where $\gamma \in (0, 1)$ is an assumed known value. An example of a likelihood is when each element in the sample follows a beta-distribution,

$$p(y_i|\theta = l) = Beta(y_i; \alpha_l, \beta_l) = \frac{1}{B(\alpha_l, \beta_l)} y_i^{\alpha_l - 1}(1 - y_i)^{\beta_l - 1}, \quad 0 \leq y_i \leq 1, \tag{15}$$

where $l \in \{0, 1\}$, $\alpha_l, \beta_l > 0$ and $B(\alpha_l, \beta_l)$ is the beta function, see Chaudhry et al. (1997). The Bernoulli distribution is a conjugate prior for the beta likelihood (Diaconis and Ylvisaker 1979), and we use them for modelling in this project. We give an example, using the Bernoulli prior distribution with $\gamma = 0.1$, and a random beta likelihood. Figure 12 displays the random beta likelihood and the resulting posterior distributions as functions of y_i to see how $p(\theta|y_i)$ changes with y_i, even though in reality it is a function of θ, for the model in (12) and (13). This example has a scalar parameter θ, but we can generalize the concept to situations where θ is a vector.

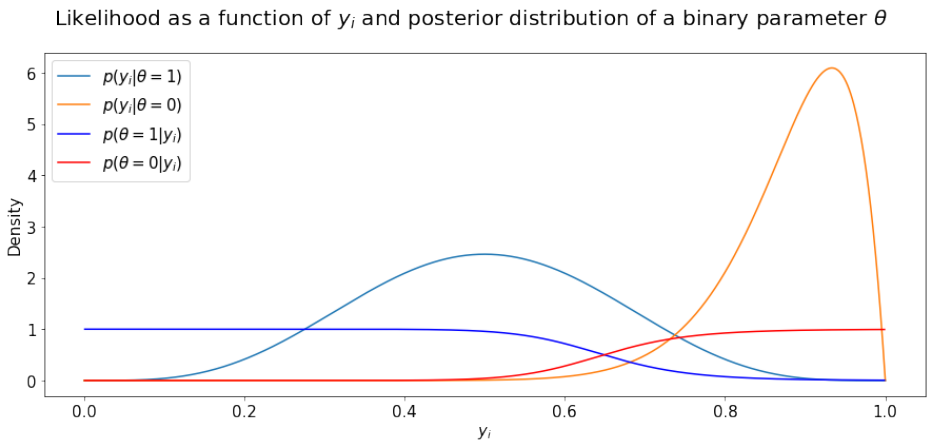


Figure 12: An example of the Bayesian model in (12) and (13). The likelihood consists of two beta distributions, and the prior is Bernoulli distributed with $p(\theta = 1) = \gamma = 0.1$.

4.2 Empirical Bayes methods

Empirical Bayes methods have several things in common with ordinary Bayesian methods but are fundamentally different in one part. We present an example to illustrate the difference. In Section 4.1, we assumed that the prior distribution $p(\theta)$ was Bernoulli distributed from (14). Further, we can create a hierarchical model by assuming that the parameter γ in (14) follows a beta distribution from (15) with hyperparameters α and β, that is

$$\gamma \sim Beta(\gamma; \alpha, \beta). \tag{16}$$

An ordinary Bayesian model determines the values for α and β before observing the data, while the empirical Bayesian approach will not specify α and β beforehand (Casella 1985). Instead, the hyperparameters are estimated from the data. The hyperparameters α and β can be saved in a vector, which we call τ, and τ can include hyperparameters of the likelihood of the data, $p(y_i|\theta)$ and of the prior distribution, $p(\theta)$.

The information about the hyperparameters is contained in the marginal likelihood for the data, denoted $L(\tau|\mathbf{y})$, and it is defined as

$$L(\tau|\mathbf{y}) = p(\mathbf{y}|\tau). \tag{17}$$

We can maximize this function with respect to τ, and this procedure is called hierarchical maximum likelihood (Farrell and Ludwig 2009). In this way, we can obtain the maximum likelihood estimates of the hyperparameters in the model.

4.3 Hidden Markov models

In this section, we present some theory on hidden Markov models. For a more detailed introduction to the theory, we refer to Rabiner (1989). A hidden Markov model (HMM) is a statistical model used to describe a process with different states, associated with probabilities of moving between the states. These probabilities are called transition probabilities, and they are defined through a Markov chain, see Pinsky and Karlin (2011). The states themselves can not be observed directly, but are related to observations, thereby the name hidden Markov model. We let $\mathbf{\Theta} = \{\theta_{1:N}\} = \{\theta_1, \dots, \theta_N\}$ denote the parameters, which are nodes that can be in the different states, and we let $\mathbf{y} = \{y_{1:N}\} = \{y_1, \dots, y_N\}$ denote the observations related to $\mathbf{\Theta}$. For convenience, we let the set of states for θ_i include two states. In a two-state, $(k+1)$-th order HMM, the Markov property tells that the probability of being in the current state, given all the predecessor states are equal to the probability of being in the current state, given the $k+1$ closest predecessor states,

$$p(\theta_i|\theta_1, \dots, \theta_{i-1}) = p(\theta_i|\theta_{i-k-1}, \dots, \theta_{i-1}), \tag{18}$$

where $0 \leq k < i-1$. Note that θ_i can be a vector in a second dimension, $\theta_i = \{\theta_i^1, \dots \theta_i^R\}$. The transition probabilities following the Markov property in (18) are the fundamental parts in the joint prior distribution of $\mathbf{\Theta}$. For a $(k+1)$-th order HMM, it is given by

$$p(\mathbf{\Theta}) = p(\theta_{1:k+1}) \cdot p(\theta_{k+2}|\theta_{1:k+1}) \cdot p(\theta_{k+3}|\theta_{2:k+2}) \cdot \dots \cdot p(\theta_N|\theta_{N-k-1:N-1}), \tag{19}$$

where $p(\theta_{1:k+1})$ is the probability distribution of the initial states $\theta_{1:k+1} = \{\theta_1, \dots, \theta_{k+1}\}$.

The transition probabilities can be expressed as an invariant transition matrix for all $i = 1, \dots, N$, denoted $\mathbf{P}$, with rows and columns specifying the probabilities of transitioning between the states. The size of the transition matrix depends on k and the size of the vectors θ_i. As mentioned, associated with each θ_i is an observation y_i, while the true state of θ_i is unknown. Each observation y_i conditioned on the associated θ_i is independent of the other states and observations,

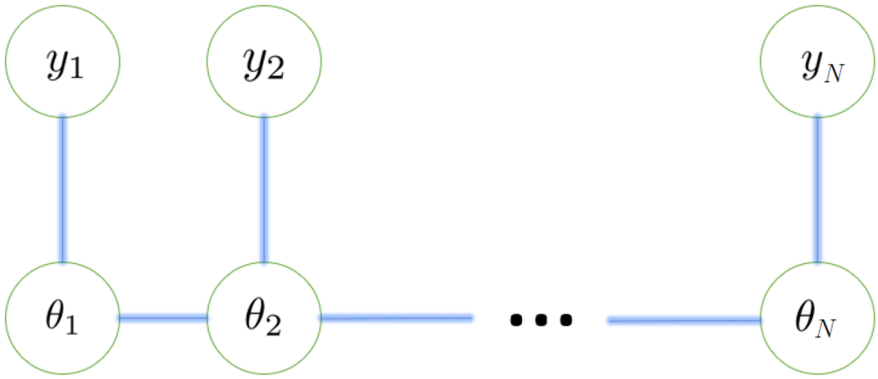


Figure 13: Figure describing the first-order HMM structure. It is inspired by Rios-Munoz et al. (2020).

which means that the likelihood of the data can be factorized as

$$p(\mathbf{y}|\mathbf{\Theta}) = \prod_{i=1}^{N} p(y_i|\theta_i). \tag{20}$$

A first-order model is illustrated in Figure 13, and we can see that the HMM is an undirected graph. The parameters θ_i and their associated observations y_i are the nodes, and the edges between them show the dependence. The indirectness of the graph means that the edges are bidirectional between nodes (Heckmann et al. 2015). We see that each observation is conditionally independent of every other parameter and observation, given its corresponding parameter, as there are no edges between y_i and the other nodes, conditioned on θ_i.

The joint posterior distribution can be calculated using Bayes theorem in (9) (Joyce 2021), and the law of total probability (Walpole et al. 2016),

$$\begin{aligned} p(\mathbf{\Theta}|\mathbf{y}) &= \frac{p(\mathbf{y}|\mathbf{\Theta}) \cdot p(\mathbf{\Theta})}{p(\mathbf{y})}, \\ &= \frac{p(\mathbf{y}|\mathbf{\Theta}) \cdot p(\mathbf{\Theta})}{\sum_{\tilde{\mathbf{\Theta}}} p(\mathbf{y}|\tilde{\mathbf{\Theta}}) \cdot p(\tilde{\mathbf{\Theta}})}. \end{aligned} \tag{21}$$

Replacing $p(\mathbf{\Theta})$ and $p(\mathbf{y}|\mathbf{\Theta})$ with their respective representations in (19) and (20),

we get that

$$p(\boldsymbol{\Theta}|\mathbf{y}) = \frac{p(\theta_{1:k+1}) \cdot \ldots \cdot p(\theta_N|\theta_{N-k-1:N-1}) \cdot \prod_{i=1}^{N} p(y_i|\theta_i)}{\sum_{\tilde{\boldsymbol{\Theta}}} p(\tilde{\theta}_{1:k+1}) \cdot \ldots \cdot p(\tilde{\theta}_N|\tilde{\theta}_{N-k-1:N-1}) \cdot \prod_{i=1}^{N} p(y_i|\tilde{\theta}_i)}. \tag{22}$$

Ideally, we would want to obtain the joint posterior distribution in (22), before marginalizing it over the subset of parameters obtained by excluding θ_i, denoted θ_{-i}, to get $p(\theta_i|\mathbf{y})$. This is the probability of the state of θ_i, given all y_is. However, this requires that we sum over θ_{-i} in the posterior distribution, which is $N-1$ variables. With these variables being binary, we get a sum over 2^{N-1} terms for each time series, which is too computationally intensive when N is large. We must also compute the sum over $\tilde{\boldsymbol{\Theta}}$ in (22), which is a sum over 2^N terms. Therefore, we need a more efficient method for finding the marginal posterior distribution of $\theta_i|\mathbf{y}$. A method for doing this is called the forward-backward algorithm, described in Section 4.4.

4.4 Forward-backward algorithm

We introduce a general form of the forward-backward algorithm for $(k+1)$-th order HMMs, inspired by an algorithm for the first-order HMM from Devijver (1985) and by an algorithm for the second-order HMM from Sung-Hyun et al. (2018). In addition, we discuss how this algorithm can be used to estimate hyperparameters in an HMM, with empirical Bayes estimators, described in Section 4.2. In the end, in Sections 4.4.1 and 4.4.2, we also give a proof of this general algorithm. We base the theory in this section on the HMM described in Section 4.3, and continue with the same notation.

An HMM is shown in Figure 13, where we consider the parameters θ_i and the observations y_i to be vectors. Assume that we want to compute the marginal posterior distributions of $\theta_i|\mathbf{y}$. In principle, we can solve them by calculating the joint posterior distribution in (22) before marginalizing over θ_{-i}. However, as discussed in Section 4.3, we need a more computationally efficient way of solving them.

To efficiently compute the marginal posterior distributions $p(\theta_i|\mathbf{y})$, we first calculate the posterior distributions $p(\theta_{i-k:i}|\mathbf{y})$, using the forward-backward algorithm (Devijver 1985). After that, we can marginalize over $\theta_{i-k:i-1}$ to obtain $p(\theta_i|\mathbf{y})$, which is not hard to do computationally. The idea behind the forward-backward algorithm is to make a decomposition of the joint distribution $p(\theta_{i-k:i}, \mathbf{y})$ and calculate each part of the decomposition recursively before multiplying them together, as this is much less computationally demanding than calculating the joint

posterior in (22) directly and marginalizing over every θ_{-i}. The decomposition is on the form

$$p(\theta_{i-k:i}, \mathbf{y}) = p(\theta_{i-k:i}, y_{1:i}) \cdot p(y_{i+1:N}|\theta_{i-k:i}, y_{1:i}), \tag{23}$$

where $y_{1:i}$ can be removed in the second part due to the aforementioned conditional independence of y_i given θ_i. Thus, we get

$$p(\theta_{i-k:i}, \mathbf{y}) = p(\theta_{i-k:i}, y_{1:i}) \cdot p(y_{i+1:N}|\theta_{i-k:i}). \tag{24}$$

In other words, the decomposition is made of two parts, where the first is a joint distribution of $\theta_{i-k:i}$ and every observation up to i, and the second part is a conditional distribution of the rest of the observations, given $\theta_{i-k:i}$. We call $p(\theta_{i-k:i}, y_{1:i})$ the forward probabilities and $p(y_{i+1:N}|\theta_{i-k:i})$ the backward probabilities. Multiplying them together yields $p(\theta_{i-k:i}, \mathbf{y})$, and the posterior distribution of $\theta_{i-k:i}$ given $\mathbf{y}$ can now be computed from (24) through Bayes' theorem,

$$p(\theta_{i-k:i}|\mathbf{y}) = \frac{p(\theta_{i-k:i}, \mathbf{y})}{\sum_{\tilde{\theta}_{i-k:i}} p(\tilde{\theta}_{i-k:i}, \mathbf{y})}, \tag{25}$$

where $\sum_{\tilde{\theta}_{i-k:i}} p(\tilde{\theta}_{i-k:i}, \mathbf{y})$ requires a sum over relatively few terms. The recursive formulas for computing the forward and the backward probabilities, together with a proof, are given in Sections 4.4.1 and 4.4.2, respectively.

Once we have obtained the posterior distributions in (25), we can sum out the other parameters, such that we obtain a marginal posterior distribution for each θ_i, given all the observations $\mathbf{y}$,

$$p(\theta_i|\mathbf{y}) = \sum_{\theta_{i-k:i-1}} p(\theta_{i-k:i}|\mathbf{y}). \tag{26}$$

Some difficulties come with computing the forward and backward probabilities. We see that the forward probabilities, $p(\theta_{i-k:i}, y_{1:i})$ and the backward probabilities, $p(y_{i+1:N}|\theta_{i-k:i})$ are joint distributions over multiple variables, and can be hard to obtain computationally due to rounding errors. There are several methods to avoid this problem. Rabiner (1989) suggests a method where the forward and backward probabilities are normalized in every recursive step. Hence, the only sizes we deal with are marginal probabilities, bounded in $[0, 1]$. However, this method includes a lot of extra calculations, and is hard to derive analytically for higher-order HMMs. Therefore we avoid the problem by computing everything on the log scale.

As discussed in Section 4.2, we also want to estimate the hyperparameters in a model with an empirical Bayes estimator. For an HMM, the hyperparameters are the parameters in the likelihood of the data, and the transition probabilities. We can achieve this by using information obtained from the forward-backward algorithm. For this purpose, we need the last forward probabilities from the recursive algorithm, described in Sections 4.4.1 and 4.4.2. We can see from (24) that, with N replacing i,

$$p(\theta_{i-k:i}, y_{1:i}) = p(\theta_{N-k:N}, \mathbf{y}). \tag{27}$$

Since the model depends on the hyperparameters, we can marginalize over $\theta_{N-k:N}$ to obtain $p(\mathbf{y})$, which is the marginal likelihood for the data $L(\tau|\mathbf{y})$ from (17), where τ is a vector containing the hyperparameters. $L(\tau|\mathbf{y})$ can be computed, given values for the hyperparameters in τ. In mathematical notation, the marginal likelihood for $\mathbf{y}$ is

$$L(\tau|\mathbf{y}) = p(\mathbf{y}) = \sum_{\theta_{N-k:N}} p(\theta_{N-k:N}, \mathbf{y}). \tag{28}$$

As we stated at the beginning of this section, we have provided a general form of the forward-backward algorithm. In Section 5, we present four different models from the general framework from Section 4.3. Each model has its specific set of hyperparameters τ, specified in Section 5. However, there are some limitations to the first three of the models, and they are presented as a motivation for the main model described in Section 5.4. Further, in Section 6, we provide the algorithms for the models, which are special cases from the general algorithm described in this section.

4.4.1 The general forward recursion formula

The proof in this section holds for calculating the forward probabilities in a general HMM. We denote the forward probabilities $\mathbf{f}_{i-k:i} = p(\theta_{i-k:i}, y_{1:i})$, and they are calculated recursively by the formula

$$\mathbf{f}_{i-k:i} = p(y_i|\theta_i) \cdot \sum_{\theta_{i-k-1}} p(\theta_i|\theta_{i-k-1:i-1}) \cdot \mathbf{f}_{i-k-1:i-1}, \tag{29}$$

starting with

$$\mathbf{f}_{1:k+1} = p(y_{1:k+1}|\theta_{1:k+1}) \cdot p(\theta_{1:k+1}) = p(\theta_{1:k+1}, y_{1:k+1}), \tag{30}$$

using the likelihood of the data and the transition probabilities. We prove this formula by induction. For any index $i \in \{k+2, \ldots, N\}$,

$$\mathbf{f}_{i-k-1:i-1} = p(\theta_{i-k-1:i-1}, y_{1:i-1}).$$

Given $\mathbf{f}_{i-k-1:i-1}$ and using the multiplication rule of probability, conditional independence of y_i given θ_i, and the law of total probability, we have that

$$\begin{aligned}
\mathbf{f}_{i-k:i} &= p(y_i|\theta_i) \cdot \sum_{\theta_{i-k-1}} p(\theta_i|\theta_{i-k-1:i-1}) \cdot \mathbf{f}_{i-k-1:i-1}, \\
&= p(y_i|\theta_i) \cdot \sum_{\theta_{i-k-1}} p(\theta_i|\theta_{i-k-1:i-1}) \cdot p(\theta_{i-k-1:i-1}, y_{1:i-1}), \\
&= p(y_i|\theta_i) \cdot \sum_{\theta_{i-k-1}} p(\theta_{i-k-1:i}, y_{1:i-1}), \\
&= p(y_i|\theta_i) \cdot p(\theta_{i-k:i}, y_{1:i-1}), \\
&= p(\theta_{i-k:i}, y_{1:i}).
\end{aligned}$$

Hence, this recursive formula holds for $i = k+2, \ldots, N$.

4.4.2 The general backward recursion formula

The proof in this section holds for calculating the backward probabilities in a general HMM. In a similar way as for the forward probabilities, the backward probabilities are calculated recursively. We denote the backward probabilities $\mathbf{b}_{i-k:i} = p(y_{i+1:N}|\theta_{i-k:i})$, and they are calculated by the formula

$$\mathbf{b}_{i-k:i} = \sum_{\theta_{i+1}} p(y_{i+1}|\theta_{i+1}) \cdot p(\theta_{i+1}|\theta_{i-k:i}) \cdot \mathbf{b}_{i-k+1:i+1}, \tag{31}$$

starting with

$$\mathbf{b}_{N-k:N} = 1, \tag{32}$$

using the likelihood of the data and the transition probabilities. We also prove this formula by induction. For any index $i \in \{N, \ldots, k+2\}$,

$$\mathbf{b}_{i-k:i} = p(y_{i+1:N}|\theta_{i-k:i}). \tag{33}$$

Figure 14: Figure describing the graph for an HMM with cliques marked in colored rectangles between pairs of adjacent nodes. Each node is a vector containing two elements, meaning each clique contains four elements.

Given $\mathbf{b}_{i-k:i}$, and using the multiplication rule of probability, conditional independence of y_i given θ_i, and the law of total probability, we have that

$$\begin{aligned}
\mathbf{b}_{i-k-1:i-1} &= \sum_{\theta_i} p(y_i|\theta_i) \cdot p(\theta_i|\theta_{i-k-1:i-1}) \cdot \mathbf{b}_{i-k:i}, \\
&= \sum_{\theta_i} p(y_i|\theta_i) \cdot p(\theta_i|\theta_{i-k-1:i-1}) \cdot p(y_{i+1:N}|\theta_{i-k:i}), \\
&= \sum_{\theta_i} p(\theta_i, y_{i:N}|\theta_{i-k-1:i-1}), \\
&= p(y_{i:N}|\theta_{i-k-1:i-1}).
\end{aligned}$$

Hence, this recursive formula holds for $i = N, \ldots, k+2$.

4.5 Clique graphs

In Section 4.3, we discussed the system in Figure 13, which is an undirected graph, where the θ_i's and y_i's are the nodes, and the lines represent the edges between them. We continue with the notation introduced in Sections 4.3 and 4.4. Now we consider the system in Figure 14, which provides an example of a graph for the first-order HMM from Figure 13, however without the observations $\mathbf{y}$. A clique c is a subset of the nodes, with an edge between every pair of nodes (Wang and Guo 2008). Figure 14 displays three cliques for a first-order HMM as rectangles in different colors. In Figure 14, we consider each node θ_i as a vector containing two elements. Consequently, we have defined the cliques to be each 2×2 square between consecutive nodes and elements within the nodes, which in this example is $\{\theta_{i-1:i}\}$. As we can see, two adjacent cliques share a common node. The common nodes are called the separators, defined as the intersection between two cliques.

The cliques and the separators in Figure 14 can be represented as a clique graph, which is a convenient way of representing the graph to do inference (Barber

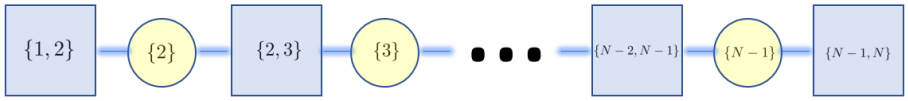


Figure 15: Figure describing a clique graph, where the squares represent the cliques, and the circles represent the separators. The edges between them show the dependence. The figure is inspired by Barber (2012).

2012). We simply let $\{i-1, i\}$ denote a clique, and we let $\{i\}$ denote a separator, for index i. A clique graph can be displayed graphically, where the cliques are represented as squares, and the separators are represented as circles. A clique graph representation of an HMM with cliques defined between two consecutive nodes is displayed in Figure 15.

Doing inference in a clique graph requires the joint distribution over the graph. For the 2×2 cliques in our example, the joint distribution of interest is $p(\mathbf{\Theta}) = p(\theta_{1:N})$. Barber (2012) prove that this distribution can be expressed as

$$\begin{aligned} p(\mathbf{\Theta}) &= \frac{p(\theta_{1:2}) \cdot \ldots \cdot p(\theta_{N-1:N})}{p(\theta_2) \cdot \ldots \cdot p(\theta_{N-1})}, \\ &= \frac{\prod_{i=2}^{N} p(\theta_{i-1:i})}{\prod_{i=2}^{N-1} p(\theta_i)}, \end{aligned} \tag{34}$$

where the nominator consists of the probability distributions of the cliques, while the denominator consists of the distributions of the separators. We prove (34) using conditional probability. To do this, we consider the joint prior distribution of $\mathbf{\Theta}$ defined through the first-order Markov chain in (19) with $k = 0$. Using (19) with the distribution for the first clique defined as $p(\theta_{1:2})$, this has the form

$$p(\mathbf{\Theta}) = p(\theta_{1:2}) \cdot \prod_{i=2}^{N} p(\theta_{i+1}|\theta_i). \tag{35}$$

For the first two consecutive cliques and the separator between them, we have that

$$p(\theta_{1:2}) \cdot p(\theta_3|\theta_2) = p(\theta_{1:2}) \cdot \frac{p(\theta_{2:3})}{p(\theta_2)}, \tag{36}$$

and further

$$p(\theta_{1:2}) \cdot p(\theta_3|\theta_2) \cdot p(\theta_4|\theta_3) = p(\theta_{1:2}) \cdot \frac{p(\theta_{2:3}) \cdot p(\theta_{3:4})}{p(\theta_2) \cdot p(\theta_3)}. \tag{37}$$

Following (35), we see that the pattern in (37) continues for $i > 3$, due to conditional probability, which means that (35) can be written on the form in (34).

Thus, when we define the cliques, we simultaneously define the joint prior distribution for $\boldsymbol{\Theta}$. The representation for the prior distribution in (34) is useful when we want to model dependence between elements in cliques.

5 Models applied to the Trondheim data

In this section, we introduce four models that we apply to the Trondheim data, and they are all special cases of the general HMM described in Section 4.3. For each model, we define a Markov chain in the spatial direction, and in the last two, we include temporal dependence. We let the superscript r denote the temporal index, such that $\theta^{1:R} = \{\theta^r\}_{r=1}^R$. Thus, the binary parameter θ_i^r describes the presence or absence of an event in spatial index i and time index r, and the parameters are correlated in a $(k+1)$-th order Markov chain in space. Since θ_i^r is binary, we have a two-state Markov chain. This Markov chain defines the prior distribution for $\boldsymbol{\Theta}^r = \{\theta_1^r, \ldots, \theta_N^r\}$. In addition, the observations y_i^r depend on their respective parameters and are conditionally independent of the other observations and parameters, as illustrated in Figure 13. We use the prior distribution in (19), specified by the order $k+1$ of the Markov chain, and the likelihood function in (20), to calculate the posterior distribution of θ_i given the observed data $\mathbf{y}$. Notation introduced in Section 4 still applies to this section.

The first three models, presented in Sections 5.1, 5.2 and 5.3 respectively, are simpler than the fourth model in Section 5.4, and they are stepping stones towards the fourth model. However, comparing how the models perform on the Trondheim data is interesting. In the following, we explain the Markov chains and the resulting prior distributions, the likelihood of the data, and how the hyperparameters are estimated, while Section 6 presents the algorithms following the model specifications.

5.1 First-order HMM in space

In the first model, we consider a first-order HMM in space, which means that $k = 0$ in (19), and the parameter θ_i consists of one point in time, denoted θ_i^r. We assume that the parameters are independent in time direction. The current state of θ_i^r depends only on the closest neighbors, θ_{i-1}^r and θ_{i+1}^r. Therefore, the

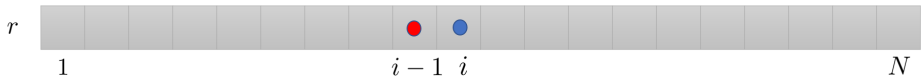


Figure 16: A system with N points in spatial direction. The conditional probabilities in (38) are described through the blue and the red dots, where the blue dot represents θ_i^r, which is conditioned on the red dot that represents θ_{i-1}^r.

prior distribution for $\mathbf{\Theta}^r$ becomes

$$p(\mathbf{\Theta}^r) = p(\theta_1^r) \cdot p(\theta_2^r|\theta_1^r) \cdot \ldots \cdot p(\theta_N^r|\theta_{N-1}^r), \tag{38}$$

where $p(\theta_i^r|\theta_{i-1}^r)$ represents transition probabilities between states in a Markov chain. These are the probabilities of transitioning from one state to another. Figure 16 shows the dependence in the Markov chain, where the blue dot is conditioned on the red dot. The transition probabilities $p(\theta_i^r|\theta_{i-1}^r)$ are invariant in space and can be represented by a 2×2 transition matrix, which is on the form

$$\mathbf{P}_1 = \begin{bmatrix} \eta_0 & 1-\eta_0 \\ 1-\zeta_0 & \zeta_0 \end{bmatrix},$$

where $\eta_0 = p(\theta_i^r = 0|\theta_{i-1}^r = 0)$ and $\zeta_0 = p(\theta_i^r = 1|\theta_{i-1}^r = 1)$. The transition probabilities in the first row represent the probabilities of transitioning from $\theta_{i-1}^r = 0$, while the transition probabilities in the second row represent the probabilities of transitioning from $\theta_{i-1}^r = 1$. To ensure that the Markov chain is stationary, we define the prior probability for the first parameter in space, $p(\theta_1^r)$, as the limit of the transition probabilities (Pinsky and Karlin 2011). It is known from Pinsky and Karlin (2011) that the limiting transition matrix of a two-state Markov chain on this form is

$$\lim_{m\to\infty} \mathbf{P}_1^m = \begin{bmatrix} \frac{1-\zeta_0}{1-\eta_0+1-\zeta_0} & \frac{1-\eta_0}{1-\eta_0+1-\zeta_0} \\ \frac{1-\zeta_0}{1-\eta_0+1-\zeta_0} & \frac{1-\eta_0}{1-\eta_0+1-\zeta_0} \end{bmatrix}.$$

Hence, we get that

$$p(\theta_1^r = 1) = \frac{1-\eta_0}{1-\eta_0+1-\zeta_0},$$
$$p(\theta_1^r = 0) = \frac{1-\zeta_0}{1-\eta_0+1-\zeta_0}.$$

In addition to the joint prior distribution for the parameters, we need the likelihood of the data, which is on the form

$$p(\mathbf{y}^r|\mathbf{\Theta}^r) = p(y_1^r, \ldots, y_N^r|\theta_1^r, \ldots, \theta_N^r) = \prod_{i=1}^{N} p(y_i^r|\theta_i^r), \tag{39}$$

due to the assumption of conditional independence. We assume that the likelihood of the data points, $p(y_i^r|\theta_i^r)$ follows a beta distribution from (15), depending on whether $\theta_i^r = 0$ or $\theta_i^r = 1$,

$$p(y_i^r|\theta_i^r = 0) = Beta(y_i^r; \alpha_0, \beta_0), \tag{40}$$
$$p(y_i^r|\theta_i^r = 1) = Beta(y_i^r; \alpha_1, \beta_1). \tag{41}$$

The reason for this assumption is that the beta distribution takes values between 0 and 1, which y_i^r also does after scaling the ACF at lag one. In addition, the beta distribution is flexible, and its parameters can be estimated to fit the data well. Finally, we can calculate the posterior distribution $p(\theta_i^r|\mathbf{y}^r)$, using the forward-backward algorithm from Section 4.4. The model-specific algorithm is described in Section 6.1. This is performed independently for each time layer $\theta_i^r, 1 \leq r \leq R$.

For the empirical Bayes parameter estimation for this model, the hyperparameters consist of the parameters in the likelihood function, $(\alpha_{0:1}, \beta_{0:1})$, as well as η_0 and ζ_0 in the transition matrix $\mathbf{P}_1$. Therefore, we denote the set of hyperparameters for this model

$$\tau_1 = (\alpha_{0:1}, \beta_{0:1}, \eta_0, \zeta_0). \tag{42}$$

We follow the procedure in Section 4.4 to compute the marginal likelihood function in (17). The model-specific algorithm is described in Section 6.1. The marginal likelihood is then maximized with respect to τ_1 using numerical methods to obtain the maximum likelihood estimates for the hyperparameters in τ_1.

5.2 Second-order HMM in space

A second-order HMM in space has $k = 1$ in (19), and the parameters consist of only one point in time. The parameters θ_i^r depends on the two neighboring parameters on each side, θ_{i-2}^r, θ_{i-1}^r, θ_{i+1}^r, θ_{i+2}^r, and we assume independence in time. Hence, the joint prior distribution is

$$p(\mathbf{\Theta}^r) = p(\theta_{1:2}^r) \cdot p(\theta_3^r|\theta_{1:2}^r) \cdot \ldots \cdot p(\theta_N^r|\theta_{N-2:N-1}^r), \tag{43}$$

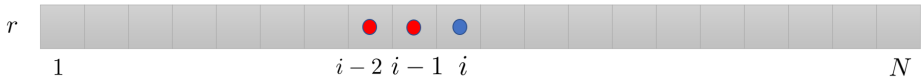


Figure 17: A system with N points in spatial direction. The conditional probabilities in (43) are described through the blue and the red dots, where the blue dot represents θ_i^r, which is conditioned on the red dots that represent $\theta_{i-2:i-1}^r$.

where $p(\theta_i^r|\theta_{i-2:i-1}^r)$ are transition probabilities between θ_{i-1}^r and θ_i^r, where the transition depends on θ_{i-2}^r. We assume that these transition probabilities are invariant in space. Figure 17 illustrates the dependence in the Markov chain, where the blue dot is conditioned on the two red dots. From (43), we get that the transition probabilities between $\theta_{i-2:i-1}$ and $\theta_{i-1:i}$ can be represented by a 4×4 transition matrix

$$\mathbf{P}_2 = \begin{bmatrix} \eta_1 & 0 & 1-\eta_1 & 0 \\ \eta_2 & 0 & 1-\eta_2 & 0 \\ 0 & 1-\zeta_1 & 0 & \zeta_1 \\ 0 & 1-\zeta_2 & 0 & \zeta_2 \end{bmatrix},$$

where

$$\begin{aligned} \eta_1 &= p(\theta_i^r = 0|\theta_{i-2}^r = 0, \theta_{i-1}^r = 0), \\ \eta_2 &= p(\theta_i^r = 0|\theta_{i-2}^r = 1, \theta_{i-1}^r = 0), \\ \zeta_1 &= p(\theta_i^r = 1|\theta_{i-2}^r = 0, \theta_{i-1}^r = 1), \\ \zeta_2 &= p(\theta_i^r = 1|\theta_{i-2}^r = 1, \theta_{i-1}^r = 1). \end{aligned}$$

We want the Markov chain in (43) to be stationary, and therefore we compute $p(\theta_{1:2}^r)$ as the limit $\lim_{m\to\infty} \mathbf{P}_2^m$. We compute this limit by applying matrix multiplication to $\mathbf{P}_2$ with itself until we have something that seems to have converged.

As for the first-order HMM, the likelihood of the data has the same form as in (39) since we are assuming conditional independence of the observations given the parameters. We still assume that the likelihood follows the beta distributions in (40) and (41).

We follow the forward-backward algorithm described in Section 6.2 to compute the posterior distributions $p(\theta_i^r|\mathbf{y}^r)$ for each time layer independently. Compared to the first-order HMM, we expect that this model gives smoother structures in space as we model the correlation between more points. However, correlation in

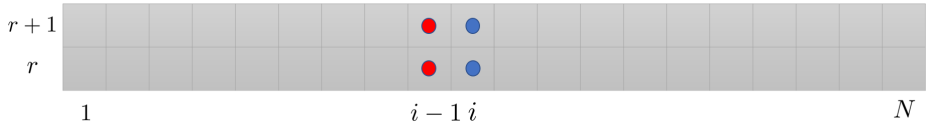


Figure 18: A system with N points in spatial direction and two layers in time direction. The Markov chain is described through the blue and the red dots, where the blue dots represent $\theta_i^{r:r+1}$, conditioned on the red dots that represent $\theta_{i-1}^{r:r+1}$.

time has not been considered yet, which is the goal of the models in Sections 5.3 and 5.4.

The hyperparameters for this second-order HMM are the parameters in the likelihood function, $(\alpha_{0:1}, \beta_{0:1})$, as well as the elements in the transition matrix $\mathbf{P}_2$, namely $\eta_{1:2}$ and $\zeta_{1:2}$. We denote the set of hyperparameters

$$\tau_2 = (\alpha_{0:1}, \beta_{0:1}, \eta_{1:2}, \zeta_{1:2}). \tag{44}$$

We follow the procedure from Section 6.2 for computing the marginal likelihood function in (17). Then, the marginal likelihood function is maximized numerically with respect to τ_2, to obtain the maximum likelihood estimates for the hyperparameters in τ_2.

5.3 First-order HMM in space and dependence in time

To account for dependence in the time direction, we can define the prior through the general Markov chain in (19) with $k = 0$ and with θ_i being a vector with two elements in time, denoted $\theta_i^{r:r+1}$. Thus, we get a first-order Markov chain in space, over two layers in time simultaneously. Figure 18 illustrates how this Markov chain looks on an $N \times 2$ grid, where the blue dots are conditioned on the red dots. We assume pairs of time layers are independent of the other time layers. Hence, we can express the joint prior distribution as

$$p(\mathbf{\Theta}^{r:r+1}) = p(\theta_1^{r:r+1}) \cdot p(\theta_2^{r:r+1} | \theta_1^{r:r+1}) \cdot \ldots \cdot p(\theta_N^{r:r+1} | \theta_{N-1}^{r:r+1}), \tag{45}$$

where $p(\theta_i^{r:r+1} | \theta_{i-1}^{r:r+1})$ represents transition probabilities from the two red dots to the two blue dots in Figure 18. As for the first two models, the transition probabilities are assumed invariant in space. Since there are four binary variables,

this leads to 2^4 transition probabilities. Therefore, they can be stored in a 4×4 transition matrix $\mathbf{P}_3$, where the rows must sum to one,

$$\mathbf{P}_3 = \begin{bmatrix} \eta_1 & \zeta_1 & \gamma_1 & 1-\eta_1-\zeta_1-\gamma_1 \\ \eta_2 & \zeta_2 & \gamma_2 & 1-\eta_2-\zeta_2-\gamma_2 \\ \eta_3 & \zeta_3 & \gamma_3 & 1-\eta_3-\zeta_3-\gamma_3 \\ \eta_4 & \zeta_4 & \gamma_4 & 1-\eta_4-\zeta_4-\gamma_4 \end{bmatrix}.$$

$\mathbf{P}_3$ is organized such that all the elements in one row are conditioned on the same values for $\theta_{i-1}^{r:r+1}$, i.e.

$$\begin{aligned} \eta_j &= p(\theta_i^r = 0, \theta_i^{r+1} = 0 | \theta_{i-1}^{r:r+1}), \\ \zeta_j &= p(\theta_i^r = 1, \theta_i^{r+1} = 0 | \theta_{i-1}^{r:r+1}), \\ \gamma_j &= p(\theta_i^r = 0, \theta_i^{r+1} = 1 | \theta_{i-1}^{r:r+1}), \\ 1-\eta_j-\zeta_j-\gamma_j &= p(\theta_i^r = 1, \theta_i^{r+1} = 1 | \theta_{i-1}^{r:r+1}), \end{aligned}$$

for $j = 1, \ldots, 4$. To keep the Markov chain stationary, we define the four initial probabilities $p(\theta_1^{r:r+1})$ from $\lim_{m\to\infty} \mathbf{P}_3^m$, and they are obtained similarly as in the second model in Section 5.2. Similar to the two previous models, we assume the likelihood of the data follows the form in (39) and the beta distributions in (40) and (41).

By following the forward-backward algorithm in Section 6.3, we obtain the joint posterior distribution $p(\theta_i^{r+1} | \mathbf{y}^{r:r+1})$ for each pair of adjacent time layers independently. This is the distribution of a single point conditioned on all the data within two layers in time. We proceed with this for every two layers in time, obtaining posterior distributions of each point, given two time layers of observations. Note that we can include more layers in time in the posterior distributions by making models with more time points included in θ_i. We have not tried this, due to the increased computational complexity. We expect that the model described in this section shows smoother structures in time compared to the previous two models.

The hyperparameters for this model are the parameters in the likelihood function, $(\alpha_{0:1}, \beta_{0:1})$, and the elements in the transition matrix $\mathbf{P}_3$, namely $(\eta_{1:4}, \zeta_{1:4}, \gamma_{1:4})$. Thus, we denote the set of hyperparameters

$$\tau_3 = (\alpha_{0:1}, \beta_{0:1}, \eta_{1:4}, \zeta_{1:4}, \gamma_{1:4}). \tag{46}$$

We compute the marginal likelihood function in (17), using the algorithm described in Section 6.3. Thus, we obtain the maximum likelihood estimates for

the hyperparameters in τ_3 by maximizing the marginal likelihood numerically with respect to τ_3.

The limitation of the first three models presented is that they do not include the information from earlier time points when calculating the probabilities at later time points. For this model, the dependence between points in time is considered only between pairs of adjacent layers. To include the information from earlier times, we need a different approach, motivating the main model in this project.

5.4 Main model

For the main model in this project, we want to incorporate information from previous points in time when computing probabilities of events in the current time layer. Ideally, we want to calculate $p(\theta_i^R|\mathbf{y}^{1:R})$, which is the posterior distribution of an event at spatial point i in the last time layer R, given all the previous and current observations. In principle, this can be obtained by computing the joint posterior distribution $p(\mathbf{\Theta}^{1:R}|\mathbf{y}^{1:R})$ using Bayes' theorem, before marginalizing over the set of every parameter except θ_i^R, denoted $\theta_{-i}^{1:R}$. However, this can not be performed computationally, as we discussed in Section 4.3. Therefore, we need a more clever way of solving $p(\theta_i^R|\mathbf{y}^{1:R})$. First, we present a way to do this conceptually in Section 5.4.1, before explaining how the conceptual solution is approximated in Section 5.4.2.

5.4.1 Concept of the model

In the previous three models, we have defined the prior distribution of $\mathbf{\Theta}$ through a Markov chain in space, where each θ_i in $\mathbf{\Theta} = \{\theta_1, \ldots, \theta_N\}$ is either a scalar or a vector with two elements adjacent in time. In addition, we have used the assumption of conditional independence of $y_i|\theta_i$ in (20) to compute the posterior probabilities $p(\theta_i|\mathbf{y})$, using the forward-backward algorithm. This is performed for each layer, or pair of layers, in time independently, without carrying on information at earlier time points.

For the main model, we still use the assumption of conditional independence for the likelihood in (20), which allows us to perform the necessary calculations. However, the joint prior distribution is defined a bit differently. Conceptually, we assume Markov structure in time, and define conditional probabilities between two layers, $p(\mathbf{\Theta}^{r+1}|\mathbf{\Theta}^r)$. These are assumed to be invariant in time, so that we can carry on the information from previous time steps. In addition, we can define a joint prior distribution for the first layer in time, $p(\mathbf{\Theta}^1)$, to calculate the joint

prior distribution for the first two layers in time, $p(\mathbf{\Theta}^{1:2})$. Applying Bayes' rule, we get that

$$p(\mathbf{\Theta}^{1:2}|\mathbf{y}^1) = \frac{p(\mathbf{\Theta}^{1:2}) \cdot p(\mathbf{y}^1|\mathbf{\Theta}^{1:2})}{p(\mathbf{y}^1)} = \frac{p(\mathbf{\Theta}^{1:2}) \cdot p(\mathbf{y}^1|\mathbf{\Theta}^1)}{p(\mathbf{y}^1)}. \tag{47}$$

In principle, we can marginalize this expression over $\mathbf{\Theta}^1$ to obtain $p(\mathbf{\Theta}^2|\mathbf{y}^1)$. For the next time step, we can use the conditional probabilities we have defined in time, $p(\mathbf{\Theta}^{r+1}|\mathbf{\Theta}^r)$, and the posterior distribution $p(\mathbf{\Theta}^2|\mathbf{y}^1)$. Thus, we get

$$p(\mathbf{\Theta}^{2:3}|\mathbf{y}^1) = p(\mathbf{\Theta}^2|\mathbf{y}^1) \cdot p(\mathbf{\Theta}^3|\mathbf{\Theta}^2), \tag{48}$$

because $\mathbf{y}^1$ is conditionally independent of $\mathbf{\Theta}^3$ given $\mathbf{\Theta}^2$. Now, we can use the resulting distribution in (48) as a prior distribution for the next pair of layers in time. Thus, the next Bayesian update becomes

$$p(\mathbf{\Theta}^{2:3}|\mathbf{y}^{1:2}) = \frac{p(\mathbf{\Theta}^{2:3}|\mathbf{y}^1) \cdot p(\mathbf{y}^2|\mathbf{\Theta}^2)}{p(\mathbf{y}^2|\mathbf{y}^1)}, \tag{49}$$

before we marginalize over $\mathbf{\Theta}^2$ to get the posterior distribution $p(\mathbf{\Theta}^3|\mathbf{y}^{1:2})$. This procedure continues for every time layer.

Note that we are not using the joint likelihood $p(\mathbf{y}^{1:2}|\mathbf{\Theta}^{1:2})$ in (47) because then the update in time would require that we condition on each $\mathbf{y}^r$ twice, which we, of course, should not do. However, since we want to use the information in the current time step to model the posterior distribution, we can also use the joint likelihood $p(\mathbf{y}^{r:r+1}|\mathbf{\Theta}^{r:r+1})$ in every time step, to obtain

$$p(\mathbf{\Theta}^{r:r+1}|\mathbf{y}^{1:r+1}) = \frac{p(\mathbf{\Theta}^{r:r+1}|\mathbf{y}^{1:r-1}) \cdot p(\mathbf{y}^{r:r+1}|\mathbf{\Theta}^{r:r+1})}{p(\mathbf{y}^{r:r+1}|\mathbf{y}^{1:r-1})}, \tag{50}$$

and we can marginalize to get $p(\theta_i^{r+1}|\mathbf{y}^{1:r+1})$. This distribution can be used to present the results in every time layer. Thus, for the final layer, we can obtain $p(\theta_i^R|\mathbf{y}^{1:R})$, which is the probability of an event, given all the observations.

While this is the concept behind the model, we have a computational problem when marginalizing the distributions presented in this section. If we define $p(\mathbf{\Theta}^{r+1}|\mathbf{\Theta}^r)$ and $p(\mathbf{\Theta}^1)$ such that $p(\mathbf{\Theta}^{1:2})$ has a Markov structure, (47) will also have a Markov structure. However, when we marginalize to obtain $p(\mathbf{\Theta}^2|\mathbf{y}^1)$, we break the Markov structure. To compute the posterior probabilities in a computationally feasible way, we want to use the forward-backward algorithm, which

requires a Markov structure. Therefore, we replace $p(\boldsymbol{\Theta}^2|\mathbf{y}^1)$ with an approximation, denoted $\tilde{p}(\boldsymbol{\Theta}^2|\mathbf{y}^1)$, that has a Markov structure. Further, we compute

$$\tilde{p}(\boldsymbol{\Theta}^{2:3}|\mathbf{y}^1) = \tilde{p}(\boldsymbol{\Theta}^2|\mathbf{y}^1) \cdot p(\boldsymbol{\Theta}^3|\boldsymbol{\Theta}^2), \tag{51}$$

which is an equivalent step to (48), using $\tilde{p}(\boldsymbol{\Theta}^2|\mathbf{y}^1)$. The distribution in (51) can be used in the forward-backward algorithm because it has a Markov structure. After using $\tilde{p}(\boldsymbol{\Theta}^{2:3}|\mathbf{y}^1)$ in the Bayesian update in (49), when we marginalize over $\boldsymbol{\Theta}^2$ in $\tilde{p}(\boldsymbol{\Theta}^{2:3}|\mathbf{y}^{1:2})$, the Markov structure is broken again. Thus, we have to replace $\tilde{p}(\boldsymbol{\Theta}^3|\mathbf{y}^{1:2})$ with a new approximation, denoted $\tilde{\tilde{p}}(\boldsymbol{\Theta}^3|\mathbf{y}^{1:2})$, which has a Markov structure. Repeating this procedure in time, we make new approximations in every time step. To simplify the notation, after r approximations we denote the resulting approximated distribution that has a Markov structure, $p^*(\boldsymbol{\Theta}^r|\mathbf{y}^{1:r-1})$. Section 5.4.2 explains the details regarding these approximations.

5.4.2 Solution for the model

In this section, we first define the conditional probabilities in time, $p(\boldsymbol{\Theta}^{r+1}|\boldsymbol{\Theta}^r)$. Then, we discuss how we obtain the joint prior distribution for the first two layers in time, $p(\boldsymbol{\Theta}^{1:2})$. Further, we explain how $\tilde{p}(\boldsymbol{\Theta}^2|\mathbf{y}^1)$, which is an approximation to the posterior distribution $p(\boldsymbol{\Theta}^2|\mathbf{y}^1)$, is defined. Then, we discuss how the conditional probabilities in time, $p(\boldsymbol{\Theta}^{r+1}|\boldsymbol{\Theta}^r)$, are used together with $\tilde{p}(\boldsymbol{\Theta}^2|\mathbf{y}^1)$, to proceed in time. At last, we present the hyperparameters in this model.

We define conditional probabilities in time with a first-order Markov structure and by the result in (34), this can be expressed as

$$p(\boldsymbol{\Theta}^{r+1}|\boldsymbol{\Theta}^r) = \frac{p(\theta^{r+1}_{1:2}|\boldsymbol{\Theta}^r) \cdot \ldots \cdot p(\theta^{r+1}_{N-1:N}|\boldsymbol{\Theta}^r)}{p(\theta^{r+1}_{2}|\boldsymbol{\Theta}^r) \cdot \ldots \cdot p(\theta^{r+1}_{N-1}|\boldsymbol{\Theta}^r)}. \tag{52}$$

Additionally, we set the restriction

$$p(\theta^{r+1}_{i-1:i}|\boldsymbol{\Theta}^r) = p(\theta^{r+1}_{i-1:i}|\theta^r_{i-1:i}), \tag{53}$$

which gives the conditional probabilities in time displayed in Figure 19, where the blue dots are conditioned on the red dots. The restriction in (53) leads to 16 different conditional probabilities, and we assume they are invariant in space and

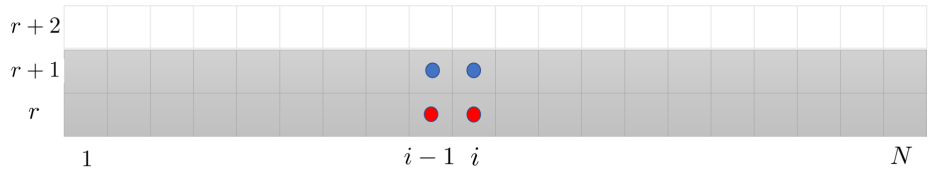


Figure 19: A system with N points in spatial direction and three layers in time direction. The conditional probabilities in time are described through the blue and the red dots, where the blue dots represent $\theta^{r+1}_{i-1:i}$, which are conditioned on the red dots that represent $\theta^{r}_{i-1:i}$.

time. Therefore, $p(\theta^{r+1}_{i-1:i}|\theta^{r}_{i-1:i})$ can be expressed in a 4×4 transition matrix,

$$\mathbf{P}_4 = \begin{bmatrix} \eta_1 & \zeta_1 & \gamma_1 & 1-\eta_1-\zeta_1-\gamma_1 \\ \eta_2 & \zeta_2 & \gamma_2 & 1-\eta_2-\zeta_2-\gamma_2 \\ \eta_3 & \zeta_3 & \gamma_3 & 1-\eta_3-\zeta_3-\gamma_3 \\ \eta_4 & \zeta_4 & \gamma_4 & 1-\eta_4-\zeta_4-\gamma_4 \end{bmatrix}.$$

$\mathbf{P}_4$ is organized such that every element in row j are conditioned on the same values for θ^{r}_{i-1} and θ^{r}_{i},

$$\begin{aligned} \eta_j &= p(\theta^{r+1}_{i-1} = 0, \theta^{r+1}_{i} = 0|\theta^{r}_{i-1:i}), \\ \zeta_j &= p(\theta^{r+1}_{i-1} = 1, \theta^{r+1}_{i} = 0|\theta^{r}_{i-1:i}), \\ \gamma_j &= p(\theta^{r+1}_{i-1} = 0, \theta^{r+1}_{i} = 1|\theta^{r}_{i-1:i}), \\ 1-\eta_j-\zeta_j-\gamma_j &= p(\theta^{r+1}_{i-1} = 1, \theta^{r+1}_{i} = 1|\theta^{r}_{i-1:i}), \end{aligned}$$

for $j = 1, \ldots, 4$.

To obtain $p(\boldsymbol{\Theta}^{1:2})$, we need $p(\boldsymbol{\Theta}^{1})$. We imagine analyzing data many time layers later than the initial layer, indicating that the initial layer in time is not important for the results. For simplicity, $p(\theta^{1}_{i-1:i})$ are computed from the limiting probabilities of the conditional probabilities in time, $\lim_{m\to\infty} \mathbf{P}_4^m$, and they are used to define $p(\boldsymbol{\Theta}^{1})$ through a first-order Markov chain.

Now, we can obtain the joint prior distribution for 2×2 cliques in the first two time layers because of the Markov structure for the conditional probabilities in time,

$$p(\theta^{1:2}_{i-1:i}) = p(\theta^{2}_{i-1:i}|\theta^{1}_{i-1:i}) \cdot p(\theta^{1}_{i-1:i}). \tag{54}$$

This joint prior distribution allows us to calculate transition probabilities in space, $p(\theta^{1:2}_{i}|\theta^{1:2}_{i-1})$, which has Markov properties in space. They are needed to perform

the forward-backward algorithm on the Markov chain in the spatial direction. Hence, the Markov structure for the first two time layers is the same as for the model in Section 5.3. Using (54), the joint prior distribution of $\mathbf{\Theta}^1$ and $\mathbf{\Theta}^2$ can be expressed as (34),

$$p(\mathbf{\Theta}^{1:2}) = \frac{\prod_{i=2}^{N} p(\theta_{i-1:i}^{1:2})}{\prod_{i=2}^{N-1} p(\theta_i^{1:2})}. \tag{55}$$

Again, we assume the likelihood of the data has the form in (20) and follow the beta distributions in (40) and (41). Using this and the prior distribution in (55), we follow the model-specific forward-backward algorithm described in Section 6.4 to obtain the posterior distribution for each clique in the first two time layers, $p(\theta_{i-1:i}^{1:2}|\mathbf{y}^1)$. Further, we marginalize over $\theta_{i-1:i}^1$ to obtain the posterior distribution $p(\theta_{i-1:i}^2|\mathbf{y}^1)$ for the second layer. This distribution can in principle be obtained by marginalizing $p(\mathbf{\Theta}^2|\mathbf{y}^1)$. However, as $p(\mathbf{\Theta}^2|\mathbf{y}^1)$ does not have a Markov structure, it can not be computed by $p(\theta_{i-1:i}^2|\mathbf{y}^1)$. Therefore, we replace it with an approximation $\tilde{p}(\mathbf{\Theta}^2|\mathbf{y}^1)$, which has a Markov structure defined such that $\tilde{p}(\theta_{i-1:i}^2|\mathbf{y}^1) = p(\theta_{i-1:i}^2|\mathbf{y}^1)$. We believe this is a good approximation, because the local structures in $p(\mathbf{\Theta}^2|\mathbf{y}^1)$ are preserved in $\tilde{p}(\mathbf{\Theta}^2|\mathbf{y}^1)$. Then, we multiply the conditional probabilities in time, $p(\theta_{i-1:i}^{r+1}|\theta_{i-1:i}^r)$ with $\tilde{p}(\theta_{i-1:i}^2|\mathbf{y}^1)$, to compute the joint prior distribution

$$\tilde{p}(\theta_{i-1:i}^{2:3}|\mathbf{y}^1) = p(\theta_{i-1:i}^3|\theta_{i-1:i}^2) \cdot \tilde{p}(\theta_{i-1:i}^2|\mathbf{y}^1), \tag{56}$$

for the cliques in the second and third time layers. For the next iteration, we use the forward-backward algorithm on (56). Thus, we obtain $\tilde{p}(\theta_{i-1:i}^{2:3}|\mathbf{y}^{1:2})$. Then, we marginalize over $\theta_{i-1:i}^2$ to obtain $\tilde{p}(\theta_{i-1:i}^3|\mathbf{y}^{1:2})$. As in the first time step, $\tilde{p}(\theta_{i-1:i}^3|\mathbf{y}^{1:2})$ can not be used to compute $\tilde{p}(\mathbf{\Theta}^3|\mathbf{y}^{1:2})$, because the Markov structure is broken again. Therefore, we replace $\tilde{p}(\mathbf{\Theta}^3|\mathbf{y}^{1:2})$ with an approximation $\tilde{\tilde{p}}(\mathbf{\Theta}^3|\mathbf{y}^{1:2})$, with a Markov structure defined such that $\tilde{\tilde{p}}(\theta_{i-1:i}^3|\mathbf{y}^{1:2}) = \tilde{p}(\theta_{i-1:i}^3|\mathbf{y}^{1:2})$. This procedure is repeated iteratively in time, displayed in Figures 19 and 20.

To summarize, we define approximations to the distributions in Section 5.4.1 that has Markov structures in space, for each pair of adjacent layers in time. The approximations are constructed such that the local dependencies in the joint distributions are preserved. The Markov structures are crucial to perform the forward-backward algorithm.

After r iterations in time, we replace a lot of tildes with p^* to simplify the notation. Thus, we obtain $p^*(\theta_{i-1:i}^{r+1}|\mathbf{y}^{1:r})$, which is an approximation to $p(\theta_{i-1:i}^{r+1}|\mathbf{y}^{1:r})$ for the reasons described above. Additionally, as we explained in Section 5.4.1,

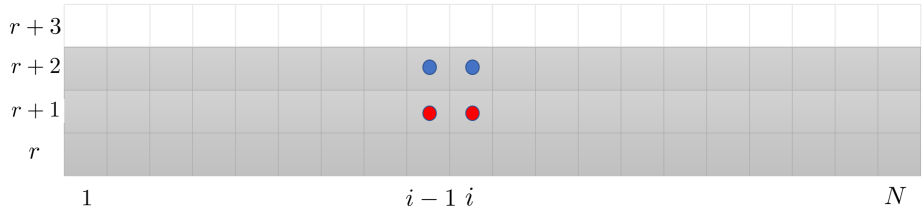


Figure 20: A system with N points in spatial direction and four layers in time direction. The conditional distributions in time are described through the blue and the red dots, where the blue dots represent $\theta_{i-1:i}^{r+2}$, which are conditioned on the red dots that represent $\theta_{i-1:i}^{r+1}$.

we compute $p^*(\theta_i^{r+1}|\mathbf{y}^{1:r+1})$ for each layer and present them as the results in Section 7.5.

This model is expected to be the most accurate, due to the dependence in space from the Markov chain, and in time obtained by defining conditional probabilities in time and using the previous information to model the current situation.

The hyperparameters for this model include the elements in the transition matrix $\mathbf{P}_4$, which are $(\eta_{1:4}, \zeta_{1:4}, \gamma_{1:4})$, and the parameters in the likelihood function, $(\alpha_{0:1}, \beta_{0:1})$. Therefore, we denote the set of hyperparameters for this model

$$\tau_4 = (\alpha_{0:1}, \beta_{0:1}, \eta_{1:4}, \zeta_{1:4}, \gamma_{1:4}). \tag{57}$$

The marginal likelihood in (17) is computed by the algorithm described in Section 6.4. To obtain the maximum likelihood estimates for the hyperparameters in τ_4, we maximize the marginal likelihood numerically with respect to τ_4.

6 Algorithms

Now, we present the specific forward-backward algorithms for each of the four models described in Section 5. In addition, we provide the algorithms for the empirical Bayes parameter estimation specific to each model. Throughout this section, we adopt the same notation as in Section 5.

6.1 First-order HMM in space

For the HMM with $k = 0$ in (19), and θ_i representing a single parameter in time, we want to find the posterior probabilities $p(\theta_i^r|\mathbf{y}^r)$ for each time layer r. For this purpose, we have defined the prior distribution of $\mathbf{\Theta}^r$ through the first-order Markov chain in (38), and we use the forward-backward algorithm described in Section 4.4. The forward probabilities are $\mathbf{f}_i^r = p(\theta_i^r, y_{1:i}^r)$ and the backward probabilities are $\mathbf{b}_i^r = p(y_{i+1:N}^r|\theta_i^r)$. To compute these probabilities, we refer to the general formulas in Sections 4.4.1 and 4.4.2, using the forward and backward probabilities defined above.

We multiply the forward probabilities $p(\theta_i^r, y_{1:i}^r)$ and the backward probabilities $p(y_{i+1:N}^r|\theta_i^r)$, to obtain the joint distributions $p(\theta_i^r, \mathbf{y}^r)$ as in (24). The posterior distributions are then computed by

$$p(\theta_i^r|\mathbf{y}^r) = \frac{p(\theta_i^r, \mathbf{y}^r)}{\sum_{\tilde{\theta}_i^r} p(\tilde{\theta}_i^r, \mathbf{y}^r)}, \tag{58}$$

where we sum over two terms only in the denominator. We perform this procedure for every layer $r = 1, \dots R$ in time.

To estimate the hyperparameters in τ_1 in (42), we use (28) with the last forward probability in space to calculate the marginal likelihood function, which can be expressed as

$$L(\tau_1|\mathbf{y}^r) = p(\mathbf{y}^r) = \sum_{\theta_N^r} p(\theta_N^r, \mathbf{y}^r). \tag{59}$$

By calculating (59) for each layer r from 1 to R, and multiplying them together, we obtain

$$L(\tau_1|\mathbf{y}^{1:R}) = \prod_{r=1}^{R} L(\tau_1|\mathbf{y}^r), \tag{60}$$

which is the marginal likelihood including all layers in time. Then, the marginal likelihood is maximized numerically with respect to τ_1 to obtain the maximum likelihood estimates for the hyperparameters in τ_1.

6.2 Second-order HMM in space

Now, we consider the second-order HMM in space. In this case, the prior distribution for time layer r, $p(\mathbf{\Theta}^r)$ is defined through the second-order Markov chain

in (43). We use the forward-backward algorithm as before, and for $k = 1$ in (19) and θ_i consisting of a single parameter in time, the forward probabilities are $\mathbf{f}^r_{i-1:i} = p(\theta^r_{i-1:i}, y^r_{1:i})$ and the backward probabilities are $\mathbf{b}^r_{i-1:i} = p(y^r_{i+1:N}|\theta^r_{i-1:i})$. Again, for computing these probabilities recursively, we refer to the formulas in Sections 4.4.1 and 4.4.2.

By multiplying the forward probabilities $p(\theta^r_{i-1:i}, y^r_{1:i})$ and the backward probabilities $p(y^r_{i+1:N}|\theta^r_{i-1:i})$, we obtain the joint distributions $p(\theta^r_{i-1:i}, \mathbf{y}^r)$. Further, we calculate the joint posterior distribution $p(\theta^r_{i-1:i}|\mathbf{y}^r)$, using Bayes theorem.

To compute the marginal posterior distributions for each θ^r_i, we marginalize over one parameter θ_{i-1},

$$p(\theta^r_i|\mathbf{y}^r) = \sum_{\theta^r_{i-1}} p(\theta^r_{i-1:i}|\mathbf{y}^r). \tag{61}$$

Note that for the initial points in spatial direction for each layer in time, we must sum over θ^r_2 in $p(\theta^r_{1:2}|\mathbf{y}^r)$, to obtain the marginal posterior distribution $p(\theta^r_1|\mathbf{y}^r)$.

To estimate the hyperparameters in τ_2 from (44) for this model, we utilize the last forward probabilities in space to calculate the marginal likelihood from (28). The marginal likelihood becomes

$$L(\tau_2|\mathbf{y}^r) = p(\mathbf{y}^r) = \sum_{\theta^r_{N-1:N}} p(\theta^r_{N-1:N}, \mathbf{y}^r). \tag{62}$$

As for the first model, we calculate the marginal likelihood for each layer in time before multiplying them together,

$$L(\tau_2|\mathbf{y}^{1:R}) = \prod_{r=1}^{R} L(\tau_2|\mathbf{y}^r). \tag{63}$$

Finally, (63) is maximized numerically with respect to τ_2 to obtain the maximum likelihood estimates for the hyperparameters in τ_2.

6.3 First-order HMM in space and dependence in time

In the model with a first-order Markov chain in space, $k = 0$ in (19), for vectors containing two points in time, $\theta_i = \theta^{r:r+1}_i$, our goal is to compute the posterior distributions $p(\theta^{r+1}_i|\mathbf{y}^{r:r+1})$. The forward probabilities are $\mathbf{f}^{r:r+1}_i = p(\theta^{r:r+1}_i, y^{r:r+1}_{1:i})$ and the backward probabilities are $\mathbf{b}^{r:r+1}_i = p(y^{r:r+1}_{i+1:N}|\theta^{r:r+1}_i)$.

Thus, the model-specific decomposition from (24) is

$$p(\theta_i^{r:r+1}, \mathbf{y}^{r:r+1}) = p(\theta_i^{r:r+1}, y_{1:i}^{r:r+1}) \cdot p(y_{i+1:N}^{r:r+1} | \theta_i^{r:r+1}). \tag{64}$$

To compute the forward and backward probabilities, we refer to the general recursive formulas in Sections 4.4.1 and 4.4.2, using the forward and backward probabilities specified above.

After obtaining the forward and backward probabilities, we multiply them together to get $p(\theta_i^{r:r+1}, \mathbf{y}^{r:r+1})$ in (64). Further, we apply Bayes' theorem to compute the posterior distribution $p(\theta_i^{r:r+1} | \mathbf{y}^{r:r+1})$. Then, we marginalize over θ_i^r to get the marginal posterior distribution, conditioned on two layers of observations in time

$$p(\theta_i^{r+1} | \mathbf{y}^{r:r+1}) = \sum_{\theta_i^r} p(\theta_i^{r:r+1} | \mathbf{y}^{r:r+1}). \tag{65}$$

We repeat this process for every pair of adjacent layers in time.

For estimation of the hyperparameters in τ_3 from (46), we calculate the marginal likelihood for the data in (28), using the last forward probability in space, for each pair of adjacent layers $\mathbf{y}^{r:r+1}$. For this model, we express the marginal likelihood as

$$L(\tau_3 | \mathbf{y}^{r:r+1}) = p(\mathbf{y}^{r:r+1}) = \sum_{\theta_N^{r:r+1}} p(\theta_N^{r:r+1}, \mathbf{y}^{r:r+1}). \tag{66}$$

To obtain the marginal likelihood for all the layers in time, we multiply every second marginal likelihood $L(\tau_3 | \mathbf{y}^{r:r+1})$ for $r = 1, 3, \ldots R$, since each likelihood contains two layers in time and the same layers should not be used twice. Thus we obtain

$$L(\tau_3 | \mathbf{y}^{1:R}) = L(\tau_3 | \mathbf{y}^{1:2}) \cdot L(\tau_3 | \mathbf{y}^{3:4}) \cdot \ldots \cdot L(\tau_3 | \mathbf{y}^{R-1:R}). \tag{67}$$

$L(\tau_3 | \mathbf{y}^{1:R})$ is then maximized numerically with respect to τ_3 to get the maximum likelihood estimates for the hyperparameters in τ_3.

6.4 Main model

The main model in this project is described in Section 5.4. We define conditional probabilities for cliques in time to include the information from previous time points. Then, we use the conditional probabilities in time to compute transition

probabilities in space, which we need in the forward-backward algorithm. Using the notation from Section 5.4, for time layer r,

$$p^*(\theta_i^{r:r+1}|\theta_{i-1}^{r:r+1}, \mathbf{y}^{1:r-1}) = \frac{p(\theta_{i-1:i}^{r+1}|\theta_{i-1:i}^{r}) \cdot p^*(\theta_{i-1:i}^{r}|\mathbf{y}^{1:r-1})}{p^*(\theta_{i-1}^{r:r+1}|\mathbf{y}^{1:r-1})}. \tag{68}$$

The Markov structure in space for two layers in time is the same as the one in Section 6.3. However, as we want to use the conditional probabilities in time to incorporate the previous information further in time, shown in (56), we need the forward-backward algorithm to return a joint distribution over two parameters in space. If we use the algorithm from Section 6.3, and include information from previous time steps, we obtain $p^*(\theta_i^{r+1}|\mathbf{y}^{1:r})$, which can not be used in (56). Therefore, we consider the algorithm for a second-order HMM with $k = 1$ in (19) and $\theta_i = \theta_i^{r:r+1}$ for this model. Additionally, as discussed in Section 5.4, we need to compute two posterior distributions for the main model. The first posterior distribution is used as prior in the next time layer, while the second posterior distribution is used to present the results. Therefore, we require two separate algorithms for solving each case, respectively.

For the first case, the forward probabilities are $\mathbf{f}_{i-1:i}^{r:r+1} = p^*(\theta_{i-1:i}^{r:r+1}, y_{1:i}^r|\mathbf{y}^{1:r-1})$ and the backward probabilities are $\mathbf{b}_{i-1:i}^{r:r+1} = p^*(y_{i+1:N}^r|\theta_{i-1:i}^{r:r+1}, |\mathbf{y}^{1:r-1})$. The resulting decomposition in (24) for the first case is

$$p^*(\theta_{i-1:i}^{r:r+1}, \mathbf{y}^r|\mathbf{y}^{1:r-1}) = p^*(\theta_{i-1:i}^{r:r+1}, y_{1:i}^r|\mathbf{y}^{1:r-1}) \cdot p^*(y_{i+1:N}^r|\theta_{i-1:i}^{r:r+1}, |\mathbf{y}^{1:r-1}). \tag{69}$$

The forward and backward probabilities are computed with the general formulas described in Sections 4.4.1 and 4.4.2 with $\mathbf{f}_{i-1:i}^{r:r+1}$ and $\mathbf{b}_{i-1:i}^{r:r+1}$ specified above, using the likelihood for one layer in time, $p(y_i^r|\theta_i^r)$, instead of two layers, $p(y_i^{r:r+1}|\theta_i^{r:r+1})$.

Once we have obtained (69), we apply Bayes' rule and marginalize over $\theta_{i-1:i}^r$ to get $p^*(\theta_{i-1:i}^{r+1}|\mathbf{y}^{1:r})$. We use this posterior distribution together with the conditional probabilities in time, $p(\theta_{i-1:i}^{r+2}|\theta_{i-1:i}^{r+1})$ to obtain the prior distribution for each clique in the next two layers,

$$p^*(\theta_{i-1:i}^{r+1:r+2}|\mathbf{y}^{1:r}) = p^*(\theta_{i-1:i}^{r+1}|\mathbf{y}^{1:r}) \cdot p(\theta_{i-1:i}^{r+2}|\theta_{i-1:i}^{r+1}). \tag{70}$$

This process is repeated iteratively in time, using (68) to calculate new transition probabilities in space from (70) for the next time layers, which we utilize in the forward and backward recursions.

For the second case, we define the decomposition of the joint distribution of a clique, $\theta_{i-1:i}^{r:r+1}$ and every observation in two time layers, $\mathbf{y}^{r:r+1}$ as

$$\begin{aligned} & p^*(\theta_{i-1:i}^{r:r+1}, \mathbf{y}^{r:r+1}|\mathbf{y}^{1:r-1}), \\ =& p^*(\theta_{i-1:i}^{r:r+1}, y_{1:i}^{r:r+1}|\mathbf{y}^{1:r-1}) \cdot p^*(y_{i+1:N}^{r:r+1}|\theta_{i-1:i}^{r:r+1}, \mathbf{y}^{1:r-1}). \end{aligned} \tag{71}$$

For this case, the forward proabilities are $\mathbf{f}_{i-1:i}^{r:r+1} = p^*(\theta_{i-1:i}^{r:r+1}, y_{1:i}^{r:r+1}|\mathbf{y}^{1:r-1})$, and the backward probabilities are $\mathbf{b}_{i-1:i}^{r:r+1} = p^*(y_{i+1:N}^{r:r+1}|\theta_{i-1:i}^{r:r+1}, \mathbf{y}^{1:r-1})$.

Using the general formula to compute the forward and backward probabilities, we obtain the joint distribution $p^*(\theta_{i-1:i}^{r:r+1}, \mathbf{y}^{r:r+1}|\mathbf{y}^{1:r-1})$ from (71). Then, we apply Bayes' rule and marginalize over $\theta_{i-1:i}^{r}$ and θ_{i-1}^{r+1} to get the marginal posterior distribution $p^*(\theta_i^{r+1}|\mathbf{y}^{1:r+1})$, which we use as result in Section 7.5.

To estimate the hyperparameters in τ_4 from (57), we calculate the marginal likelihood in (28), with the last forward probability in space, using the algorithm from the first case in this section. The last forward probability in space is obtained for every layer in time while we follow the iterations in time from Section 5.4.2. Thus, we obtain $p^*(\theta_{N-1:N}^{r:r+1}, \mathbf{y}^r|\mathbf{y}^{1:r-1})$ for each time layer r. Then, for each layer except the last two, we obtain the marginal likelihood

$$L(\tau_4|\mathbf{y}^r) = p^*(\mathbf{y}^r|\mathbf{y}^{1:r-1}) = \sum_{\theta_{N-1:N}^{r:r+1}} p^*(\theta_{N-1:N}^{r:r+1}, \mathbf{y}^r|\mathbf{y}^{1:r-1}), \tag{72}$$

which contains a sum of 16 variables. For the last two layers in time, we use the last forward probabilities in space from the algorithm in the second case in this section to obtain

$$L(\tau_4|\mathbf{y}^{R-1:R}) = p^*(\mathbf{y}^{R-1:R}|\mathbf{y}^{1:R-2}) = \sum_{\theta_{N-1:N}^{R-1:R}} p^*(\theta_{N-1:N}^{R-1:R}, \mathbf{y}^{R-1:R}|\mathbf{y}^{1:R-2}). \tag{73}$$

Multiplying (72) for every layer in time and (73), we obtain

$$\begin{aligned} L(\tau_4|\mathbf{y}^{1:R}) &= L(\tau_4|\mathbf{y}^1) \cdot L(\tau_4|\mathbf{y}^2) \cdot \ldots \cdot L(\tau_4|\mathbf{y}^{R-1:R}), \\ &= p^*(\mathbf{y}^1) \cdot p^*(\mathbf{y}^2|\mathbf{y}^1) \cdot \ldots \cdot p^*(\mathbf{y}^{R-1:R}|\mathbf{y}^{1:R-2}). \end{aligned} \tag{74}$$

Once we have obtained the marginal likelihood, $L(\tau_4|\mathbf{y}^{1:R})$, we maximize it numerically with respect to τ_4 to obtain the maximum likelihood estimates for the hyperparameters in τ_4.

7 Results and discussion

In this section, we present the results from the four models, which are the posterior probabilities for an event based on the observations and the model specifications. We begin by explaining how the hyperparameters in the models are estimated before we go through the results for each model. For the main model, we also investigate how different choices of values for the hyperparameters affect the results. Throughout this section, we adopt the notation introduced in Sections 5 and 6.

7.1 Parameter estimation

To compare the performance of the different models, we estimate the hyperparameters in τ at the same locations within the same time frame for all the models. This area is between 16000 and 17000 meters along the railway tracks and between 13:34:43 and 13:36:53 in time, and it is a different area than where we present the results. We choose this area because it contains several events, but most of it consists of non-events, such that it is a representative sample of the entire data set. The area is displayed in Figure 21.

As previously stated, the hyperparameters in an HMM are estimated by maximizing the marginal likelihood function $L(\tau|\mathbf{y}^{1:R})$ with respect to τ, where τ consists of the elements in the transition matrix $\mathbf{P}$ and the parameters in the likelihood of the data, namely $\alpha_{0:1}$ and $\beta_{0:1}$. The probabilities in the transition matrix $\mathbf{P}$ are model specific, but we denote them η_j, ζ_j, and γ_j, depending on the model. Thus we get that

$$\tau = (\alpha_{0:1}, \beta_{0:1}, \eta_j, \zeta_j, \gamma_j). \tag{75}$$

Once the hyperparameters are estimated, we use them in the models to obtain the results presented in this section. To calculate the maximum likelihood estimates numerically, we use the library `scipy.optimize` with a trust-region algorithm for constrained optimization (Virtanen et al. 2020) on the marginal likelihood function, with specified bounds and constraints for the hyperparameters.

7.2 First-order HMM in space

Using the procedure to obtain the marginal likelihood $L(\tau_1|\mathbf{y}^{1:R})$ described in Section 6.1, and maximizing this with respect to τ_1, we obtain the maximum like-

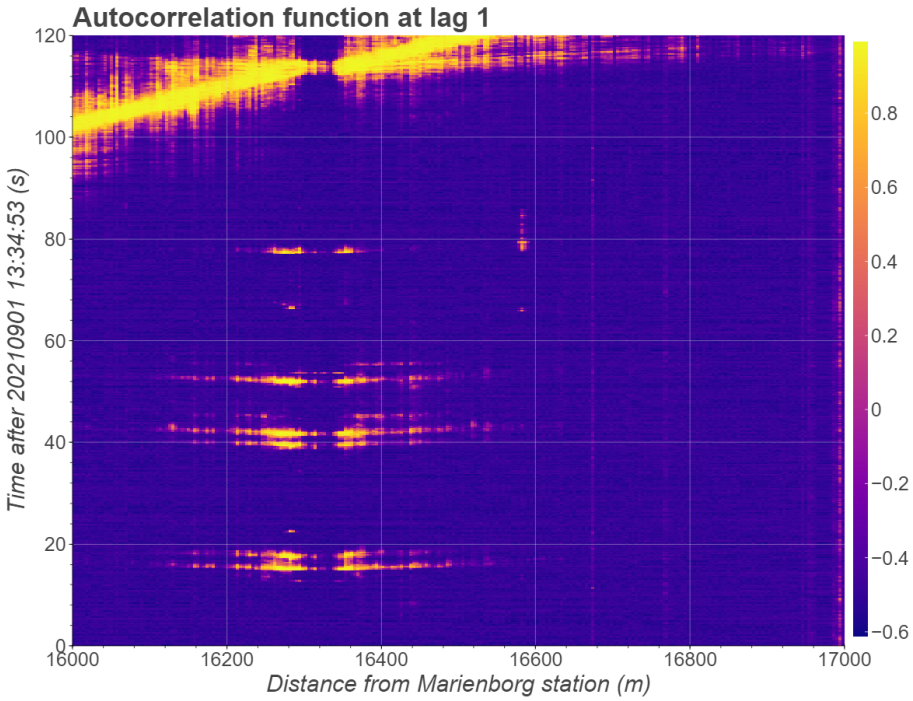


Figure 21: Heatmap of the autocorrelation at lag one in the area where we estimate the parameters.

lihood estimates for the hyperparameters in this model. These hyperparameters are the transition probabilities in $\mathbf{P}_1$, which is the transition matrix described in Section 5.1, and the parameters in the likelihood function $p(y_1^r|\theta_i^r = l) = Beta(y_i; \alpha_l, \beta_l)$, which are α_l and β_l for $l \in \{0, 1\}$. The transition probabilities in $\mathbf{P}_1$ are bounded in (0,1), while $\alpha_l, \beta_l > 0$. The estimated transition probabilities for the area in Figure 21 are

$$\mathbf{P}_1 = \begin{bmatrix} 0.992 & 0.008 \\ 0.079 & 0.921 \end{bmatrix},$$

which means that

$$\begin{aligned} p(\theta_i^r = 0|\theta_{i-1}^r = 0) &= 0.992, \\ p(\theta_i^r = 1|\theta_{i-1}^r = 1) &= 0.921. \end{aligned} \tag{76}$$

Thus, the probability of staying in the same state when moving between spatial nodes is significantly higher than the probability of transitioning to the other state. Additionally, transitioning to state 0, given the current node is in state 1, has a higher probability than the opposite transition. This suggests that non-events are more likely to occur. We obtain the initial probabilities in space from the limit of the transition matrix, $\lim_{m\to\infty} \mathbf{P}_1^m$, and they become

$$\begin{aligned} p(\theta_1^r = 0) &= 0.904, \\ p(\theta_1^r = 1) &= 0.096, \end{aligned} \tag{77}$$

which reflects the probability of an event in the estimation area. The estimated parameters of the likelihood function, $\alpha_{0:1}$ and $\beta_{0:1}$ are

$$\begin{aligned} \alpha_0 &= 143.382, \\ \beta_0 &= 417.061, \\ \alpha_1 &= 1.611, \\ \beta_1 &= 1.139, \end{aligned} \tag{78}$$

where α_0, β_0 represent the likelihood where $\theta_i^r = 0$, while α_1, β_1 represent the likelihood where $\theta_i^r = 1$. The likelihood as a function of y_i^r is displayed in Figure 22. We can see that $p(y_i^r|\theta_i^r = 0)$ takes most values between 0.2 and 0.3, implying the model is confident that the observed values fall within this range when there are no events. On the other hand, $p(y_i^r|\theta_i^r = 1)$ suggests that when there are events, y_i^r can take a wide range of values, with a higher likelihood for larger values.

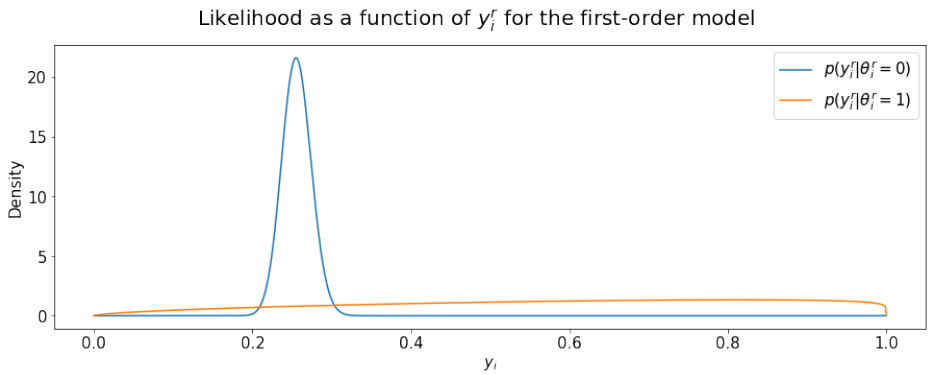


Figure 22: Beta likelihood displayed as a function of y_i^r, with the hyperparameters $\alpha_{0:1} = (143.382, 1.611)$ and $\beta_{0:1} = (417.061, 1.139)$ estimated with an empirical Bayes estimator for a first-order HMM in space.

Following the procedure described in Sections 5.1 and 6.1, we obtain the posterior probabilities for an event $p(\theta_i^r = 1|\mathbf{y}^r)$ for this model, for every point in every layer in time. They are displayed as a heatmap in Figure 23 at the area around Selsbakk station. The posterior probabilities capture the structures from Figure 11, which are the results we expected. We also notice that some of the signals in Figure 11 which resemble noise, have high probabilities of being events in this model. The structures we observe are discussed in Section 3.2.1.

When studying the heatmaps, it is important to remember that probabilities at earlier times are not dependent on probabilities at later times in any of the models. Even though this model does not account for dependence in time, this thesis aims to find the probabilities of events in the current time layer. Looking at the posterior probabilities in the last time layer is an easy way to interpret the model's output for the newest data. Hence, we display the posterior probabilities for the final time layer between 3000 and 5500m in Figure 24. As we expected from the likelihood in Figure 22 and the estimated transition probabilities in (76), the model assigns most probabilities close to 0 and 1, which means that the model is confident in determining whether there is an event in each point. The structures we see in Figure 24 are the large vertical line at 5100 meters, one of the periodic dots at around 4550 meters, and one of the diagonal lines that is a moving car at 4000 meters, discussed in Section 3.2.1.

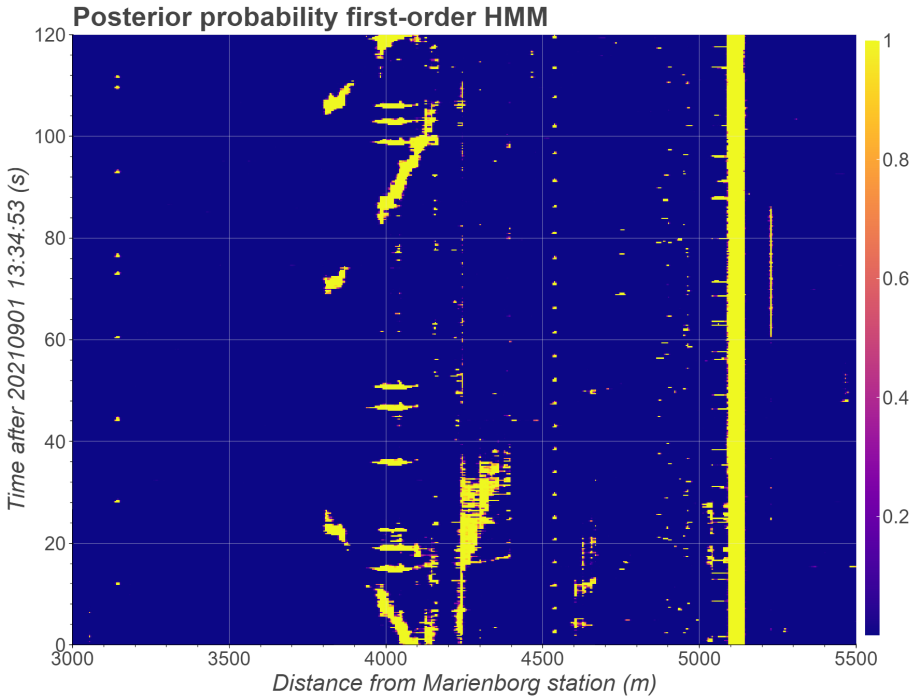


Figure 23: Heatmap of the posterior probabilities $p(\theta_i^r = 1|\mathbf{y}^r)$ for the first-order HMM in space.

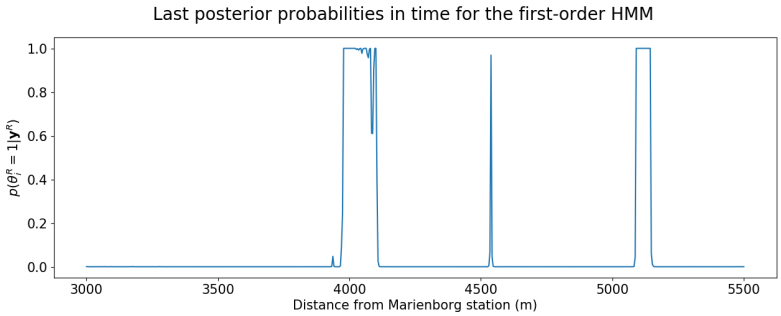


Figure 24: Posterior probabilities for the last time layer R, $p(\theta_i^R = 1|\mathbf{y}^R)$ for the first-order HMM in space.

7.3 Second-order HMM in space

For the second-order HMM, we follow the procedure described in Section 6.2 to obtain the marginal likelihood $L(\tau_2|\mathbf{y}^{1:R})$, before we maximize it with respect to τ_2. The hyperparameters have the same bounds as we described in Section 7.2. We get that the estimated values for the transition probabilities are

$$\mathbf{P}_2 = \begin{bmatrix} 0.988 & 0 & 0.012 & 0 \\ 0.759 & 0 & 0.241 & 0 \\ 0 & 0.596 & 0 & 0.404 \\ 0 & 0.050 & 0 & 0.950 \end{bmatrix},$$

where each row sums to one. This means that

$$\begin{aligned} p(\theta_i^r = 0|\theta_{i-2}^r = 0, \theta_{i-1}^r = 0) &= 0.988, \\ p(\theta_i^r = 0|\theta_{i-2}^r = 1, \theta_{i-1}^r = 0) &= 0.759, \\ p(\theta_i^r = 1|\theta_{i-2}^r = 0, \theta_{i-1}^r = 1) &= 0.404, \\ p(\theta_i^r = 1|\theta_{i-2}^r = 1, \theta_{i-1}^r = 1) &= 0.950. \end{aligned} \tag{79}$$

The estimated transition probabilities in (79) show that when two consecutive parameters in space have the same value, there is a high probability that the following parameter also has the same value. However, when two consecutive parameters have different values, the probability of the next parameter being 0 is higher than the probability of it being 1. The fact that $p(\theta_i = 1|\theta_{i-2} = 0, \theta_{i-1} = 1)$ is smaller than $p(\theta_i = 0|\theta_{i-2} = 0, \theta_{i-1} = 1)$, means that the model allows for a reasonable probability of events in individual points. This is not desirable for our purpose since we believe signals from an event propagate over multiple points in space and time, as we saw in Figure 11.

The initial probabilities are $\lim_{m\to\infty} \mathbf{P}_2^m$, which gives $p(\theta_{1:2}^r)$ for each layer in time r. The resulting values are

$$\begin{aligned} p(\theta_1^r = 0, \theta_2^r = 0) &= 0.861, \\ p(\theta_1^r = 1, \theta_2^r = 0) &= 0.015, \\ p(\theta_1^r = 0, \theta_2^r = 1) &= 0.014, \\ p(\theta_1^r = 1, \theta_2^r = 1) &= 0.110. \end{aligned} \tag{80}$$

The estimated parameters for the beta likelihood function for this model are approximately equal to the ones for the first-order model in Section 7.2, however

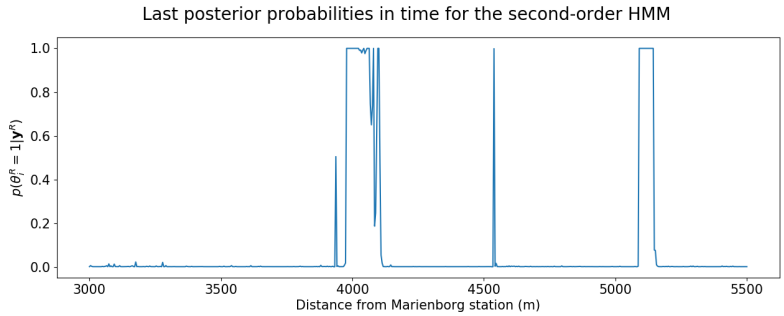


Figure 25: Posterior probabilities for the last time layer R, $p(\theta_i^R = 1|\mathbf{y}^R)$ for the second-order HMM in space.

with slightly higher values for α_0 and β_0,

$$\begin{aligned} \alpha_0 &= 147.433, \\ \beta_0 &= 428.806, \\ \alpha_1 &= 1.638, \\ \beta_1 &= 1.127. \end{aligned} \tag{81}$$

Therefore, we refer to Figure 22 for an illustration, and the discussion regarding the likelihood function is similar to the one we presented in Section 7.2. The slightly higher values for α_0 and β_0 might suggest that this model is more confident of the values for y_i^r where there are no events.

Using the estimated hyperparameter values, we follow the procedure described in Sections 5.2 and 6.2 to obtain the posterior probability of an event, $p(\theta_i^r = 1|\mathbf{y}^r)$, for the second-order HMM. The results are similar to those we obtained from the first-order model. Therefore we omit to present a heatmap.

As for the first-order model, we expect the majority of the probabilities to be close to 0 or 1 due to the parameter values in the likelihood function and the high probabilities of staying in the same state. In Figure 25, we show $p(\theta_i^R|\mathbf{y}^R)$ for the last time layer between 3000 and 5500m. The plot confirms our expectation that most posterior probabilities are close to 0 or 1. Compared to the first model, they look similar, but some spikes at 3900 and 4100m have increased. In addition, the second-order model seems to capture some small structures, particularly noticeable between 3100 and 3300m.

7.4 First-order HMM in space and dependence in time

Again, to estimate the parameters for this model, we calculate the marginal likelihood $L(\tau_3|\mathbf{y}^{1:R})$ by following the procedure from Section 6.3. Then, we maximize this function numerically with respect to τ_3, to get the maximum likelihood estimates for the hyperparameters $\alpha_{0:1}$, $\beta_{0:1}$, $\eta_{1:4}$, $\zeta_{1:4}$ and $\gamma_{1:4}$. For the optimization process, $\alpha_{0:1}, \beta_{0:1} > 0$ and $(\eta_{1:4}, \zeta_{1:4}, \gamma_{1:4}) \in (0, 1)$. Additionally, η_j, ζ_j and γ_j must satisfy the constraint $\eta_j + \zeta_j + \gamma_j < 1$, for $j = 1, \ldots, 4$. For the transition matrix $\mathbf{P}_3$ described in Section 5.3, the estimated parameters are

$$\mathbf{P}_3 = \begin{bmatrix} 0.985 & 0.002 & 0.002 & 0.011 \\ 0.185 & 0.646 & 0.001 & 0.168 \\ 0.182 & 0.001 & 0.685 & 0.132 \\ 0.091 & 0.011 & 0.012 & 0.886 \end{bmatrix},$$

which means that

$$\begin{aligned} p(\theta_i^r = 0, \theta_i^{r+1} = 0 | \theta_{i-1}^r, \theta_{i-1}^{r+1}) &= (0.985, 0.185, 0.182, 0.091), \\ p(\theta_i^r = 1, \theta_i^{r+1} = 0 | \theta_{i-1}^r, \theta_{i-1}^{r+1}) &= (0.002, 0.646, 0.001, 0.011), \\ p(\theta_i^r = 0, \theta_i^{r+1} = 1 | \theta_{i-1}^r, \theta_{i-1}^{r+1}) &= (0.002, 0.001, 0.685, 0.012), \\ p(\theta_i^r = 1, \theta_i^{r+1} = 1 | \theta_{i-1}^r, \theta_{i-1}^{r+1}) &= (0.011, 0.168, 0.132, 0.886), \end{aligned} \tag{82}$$

for each layer r in time. The highest probabilities are that $\theta_i^{r:r+1}$ is $(0, 0)$ or $(1, 1)$ given that $\theta_{i-1}^{r:r+1}$ is $(0, 0)$ or $(1, 1)$, respectively. This means that the states for the next parameters in space are most likely the same as the previous. Considering $p(\theta_i^r, \theta_i^{r+1} | \theta_{i-1}^r = 1, \theta_{i-1}^{r+1} = 0)$ and $p(\theta_i^r, \theta_i^{r+1} | \theta_{i-1}^r = 0, \theta_{i-1}^{r+1} = 1)$, the states for the next two points in space are most likely going to be the same as the previous, and with a slightly higher probability of being $(0, 0)$ than $(1, 1)$.

We have that $\lim_{m \to \infty} \mathbf{P}_3^m$ gives the initial probabilities $p(\theta_1^{r:r+1})$. The resulting values are

$$\begin{aligned} p(\theta_1^r = 0, \theta_1^{r+1} = 0) &= 0.872, \\ p(\theta_1^r = 1, \theta_1^{r+1} = 0) &= 0.009, \\ p(\theta_1^r = 0, \theta_1^{r+1} = 1) &= 0.010, \\ p(\theta_1^r = 1, \theta_1^{r+1} = 1) &= 0.109, \end{aligned} \tag{83}$$

for each layers $r : r + 1$ in time. The parameters of the beta likelihood function

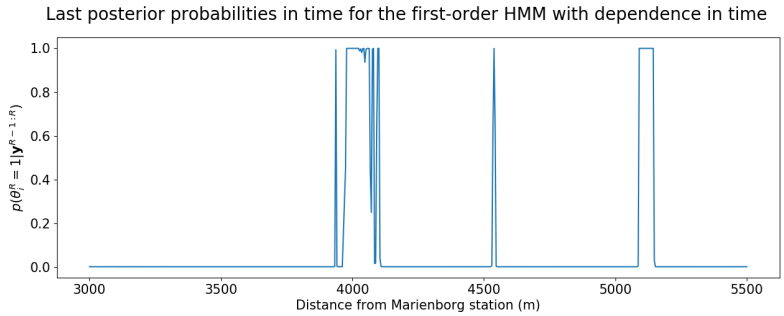


Figure 26: Posterior probabilities for the last time layer R, $p(\theta_i^R = 1|\mathbf{y}^{R-1:R})$ for the first-order HMM in space with two layers in time.

$p(y_i^r|\theta_i^r)$ are estimated to be

$$\begin{aligned} \alpha_0 &= 147.100, \\ \beta_0 &= 427.955, \\ \alpha_1 &= 1.638, \\ \beta_1 &= 1.129, \end{aligned} \tag{84}$$

and they are approximately the same as the estimated values for the second-order HMM from Section 7.3. Therefore, the discussion is the same as the one we presented in Sections 7.2 and 7.3.

We follow the procedure for calculating the posterior probability of θ_i^{r+1} given two adjacent layers of observations in time $\mathbf{y}^{r:r+1}$, described in Sections 5.3 and 6.3, using the estimated hyperparameters for this model. Since it is hard to distinguish between heatmaps for the first three models, we also omit to include a heatmap for this model.

Figure 26 illustrates the posterior probabilities for the last layer in time, namely $p(\theta_i^R|\mathbf{y}^{R-1:R})$. We observe that the two spikes at approximately 3900 and 4100m are larger than for the previous models. This observation suggests that there might be events in the previous time step at these positions, which this model manage to perceive due to the temporal dependence between the two layers. The small structures we noticed in Figure 25 between 3100 and 3300m are not observed in Figure 26.

7.5 Main model

By using the procedure described at the end of Section 6.4, we get the formula for the marginal likelihood function $L(\tau_4|\mathbf{y}^{1:R})$. This is maximized numerically with respect to τ_4, to obtain the maximum likelihood estimates for $\alpha_{0:1}$, $\beta_{0:1}$, $\eta_{1:4}$, $\zeta_{1:4}$ and $\gamma_{1:4}$. As for the maximization process for the third model, we have the bounds $\alpha_{0:1}, \beta_{0:1} > 0$ and $(\eta_{1:4}, \zeta_{1:4}, \gamma_{1:4}) \in (0,1)$, in addition to the constraints that $\eta_j + \zeta_j + \gamma_j < 1$, for $j = 1, \ldots, 4$. The estimated parameters in $\mathbf{P}_4$ from Section 5.4 are

$$\mathbf{P}_4 = \begin{bmatrix} 0.995 & 0.001 & 0.002 & 0.002 \\ 0.032 & 0.914 & 0.001 & 0.053 \\ 0.127 & 0.013 & 0.847 & 0.013 \\ 0.025 & 0.012 & 0.011 & 0.952 \end{bmatrix},$$

which means that

$$\begin{aligned} p(\theta_{i-1}^{r+1} = 0, \theta_i^{r+1} = 0 | \theta_{i-1}^r, \theta_i^r) &= (0.995, 0.032, 0.127, 0.025), \\ p(\theta_{i-1}^{r+1} = 1, \theta_i^{r+1} = 0 | \theta_{i-1}^r, \theta_i^r) &= (0.001, 0.914, 0.013, 0.012), \\ p(\theta_{i-1}^{r+1} = 0, \theta_i^{r+1} = 1 | \theta_{i-1}^r, \theta_i^r) &= (0.002, 0.001, 0.847, 0.011), \\ p(\theta_{i-1}^{r+1} = 1, \theta_i^{r+1} = 1 | \theta_{i-1}^r, \theta_i^r) &= (0.002, 0.053, 0.013, 0.952), \end{aligned} \tag{85}$$

for each pair of layers $r : r+1$ in time. The estimated conditional probabilities in time in $\mathbf{P}_4$ indicate a high probability of staying in the same state when transitioning in time. The conditional probabilities in time are used to calculate the transition probabilities in space with (68), and the transition probabilities in space depend on the time layer, as we use previous information for calculating the prior in each layer, described in Section 5.4. We notice that the estimated probabilities of staying in the same states in time are higher than the probabilities of staying in the same states in space, from Section 7.4. Therefore, we expect longer chains of similar posterior probabilities in time. Furthermore, the initial probabilities in time, $p(\theta_{i-1:i}^1)$, are computed by $\lim_{m\to\infty} \mathbf{P}_4^m$. The resulting values are

$$\begin{aligned} p(\theta_{i-1}^1 = 0, \theta_i^1 = 0) &= 0.903, \\ p(\theta_{i-1}^1 = 1, \theta_i^1 = 0) &= 0.019, \\ p(\theta_{i-1}^1 = 0, \theta_i^1 = 1) &= 0.014, \\ p(\theta_{i-1}^1 = 1, \theta_i^1 = 1) &= 0.064. \end{aligned} \tag{86}$$

We see that the probability of two adjacent points in space being $(0,0)$ is the

highest for the first time layer. The maximum likelihood estimates of the parameters of the beta likelihood $p(y_i^r|\theta_i^r)$ are

$$\begin{aligned}\alpha_0 &= 149.435,\\ \beta_0 &= 434.973,\\ \alpha_1 &= 1.637,\\ \beta_1 &= 1.136.\end{aligned} \tag{87}$$

The values are similar to those obtained for the first three models. Therefore, the discussion is similar to the one we conducted in Sections 7.2 and 7.3, and we omit to include another similar plot.

After we have obtained the estimates for the hyperparameters, we follow the procedure described in Sections 5.4 and 6.4 to obtain the posterior probabilities $p^*(\theta_i^{r+1} = 1|\mathbf{y}^{1:r+1})$ for each spatial point i, and time layer r. They are displayed as a heatmap in Figure 27. Note that the information at earlier times is incorporated in this distribution. It is hard to distinguish between the heatmap in Figure 27 and the one for the first model in Figure 23, but if they are examined closely, the heatmap in Figure 27 shows slightly more temporal dependencies. The differences between Figures 23 and 27 are easier to notice when compared on larger screens. The similarity between Figures 23 and 27 suggests that the parameter values in the likelihood have the most effect on the posterior probabilities. This is explored further in Section 7.6.

We want to focus on what happens in the current time layer, conditioned on the previous ones. Therefore we display the posterior probabilities for the last time layer, $p^*(\theta_i^R|\mathbf{y}^{1:R})$ in Figure 28. We can see that the posterior probabilities around 4000 meters oscillate more than for the third model, which suggests that the model conserves more of the information from previous observations when computing the posterior probabilities in the last layer. The other structures from the third model in Section 7.4 are present in Figure 28.

7.6 Parameter values for the main model

So far, we have presented results using hyperparameters of the HMMs estimated with an empirical Bayes estimator. However, the estimation procedure for the hyperparameters is based only on the values for y_i^r, and not on other information concerning events. The resulting likelihood from the estimation seems to be certain about the values of y_i^r when $\theta_i^r = 0$. Potentially, this overshadows the other hyperparameters, namely the conditional probabilities that decide the depend-

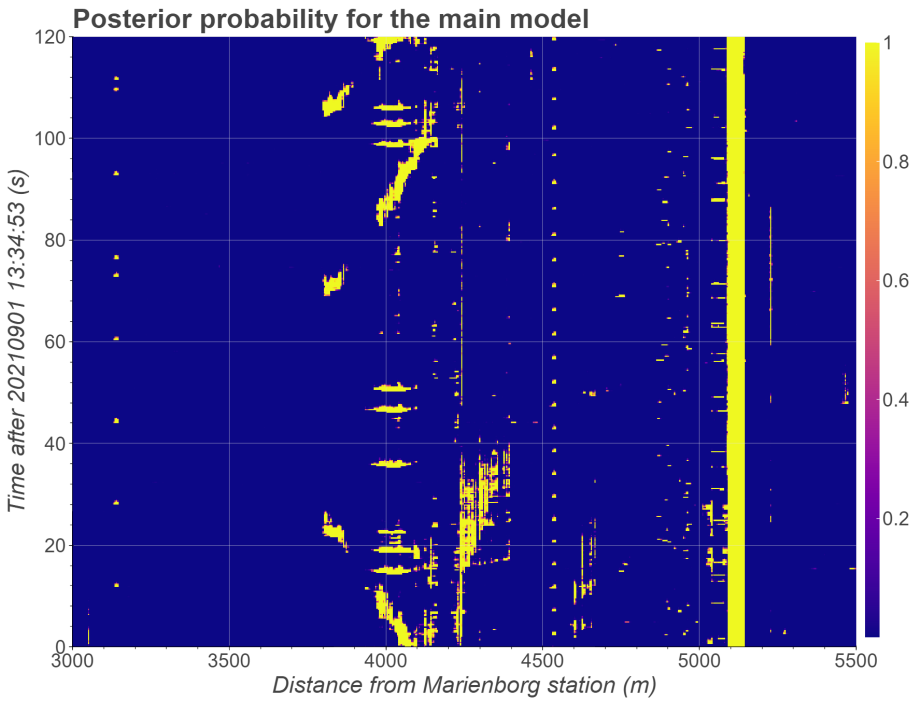


Figure 27: Heatmap of the posterior probabilities $p^*(\theta_i^{r+1} = 1|\mathbf{y}^{1:r+1})$ for the main model. Note that points in time layer r are not dependent on $\mathbf{y}^{r+2:R}$.

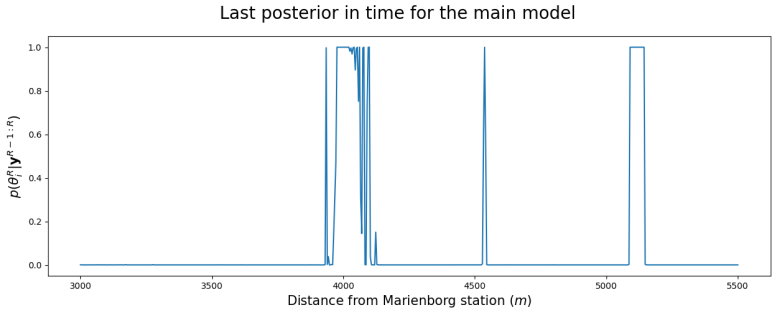


Figure 28: Posterior probabilities for the last time layer, $p^*(\theta_i^R = 1|\mathbf{y}^{1:R})$ for the main model.

ence structures in time and space. Therefore, in this section, we experiment with different values of the hyperparameters for the main model and discuss whether estimating them gives the best description of the reality, based on our knowledge of events along the railway tracks. First, we discuss the hyperparameters in the likelihood function in Section 7.6.1, before we discuss the hyperparameters that are the conditional distributions in time for the cliques in Section 7.6.2. In the end, we present some further discussion in Section 7.6.3

For comparing different values of the hyperparameters, we look at an area with both weak and strong signals for the ACF at lag one, displayed in Figure 29. The region between 29000 and $31000m$ contains weak signals that resemble cars crossing the railway tracks or people walking next to it, and the area between 31000 and $34000m$ exhibits signals that resemble noise. Between 34000 and $40000m$ we see stronger signals that might be trains in motion or vehicles crossing the tracks. In addition, we see a straight vertical line at $36800m$ representing signals that are constant in time at the same position. While the signals between 31000 and $34000m$ do not look like particular events, we do not know whether they are events. The only information we have about them is the values in the data set. Therefore we must choose if such signals should be classified as events or not, when fitting the model.

7.6.1 Parameters in the likelihood

Since the estimated parameters α_0 and β_0 of the likelihood $p(y_i^r|\theta_i^r = 0)$ in (87) are large, the model is certain that the weak signals between 31000 and 34000 meters are events, as we can see in Figure 30a. However, by selecting a beta likelihood with $\alpha_0 = 5$ and $\beta_0 = 12$, the model allows more values for y_i^r when $\theta_i^r = 0$, but the likelihood when $\theta_i^r = 0$ still takes many values between 0.2 and 0.3. This likelihood as a function of y_i^r is shown in Figure 31. Using the obtained estimates for the conditional probabilities in time, the probabilities of most of the points between 31000 and $34000m$ being events become small, as the dependence structure in space and time becomes more significant. We observe this in Figure 30b.

On the other hand, examining the region between 29000 and $31000m$, we observe structures in Figure 30a that are not present in Figure 30b. These structures resemble cars crossing the railway tracks or people walking next to them. Thus, if we decide to fit the model such that the probabilities of events for signals that seem to be noise are low, we also risk classifying other weak signals that might be events as non-events.

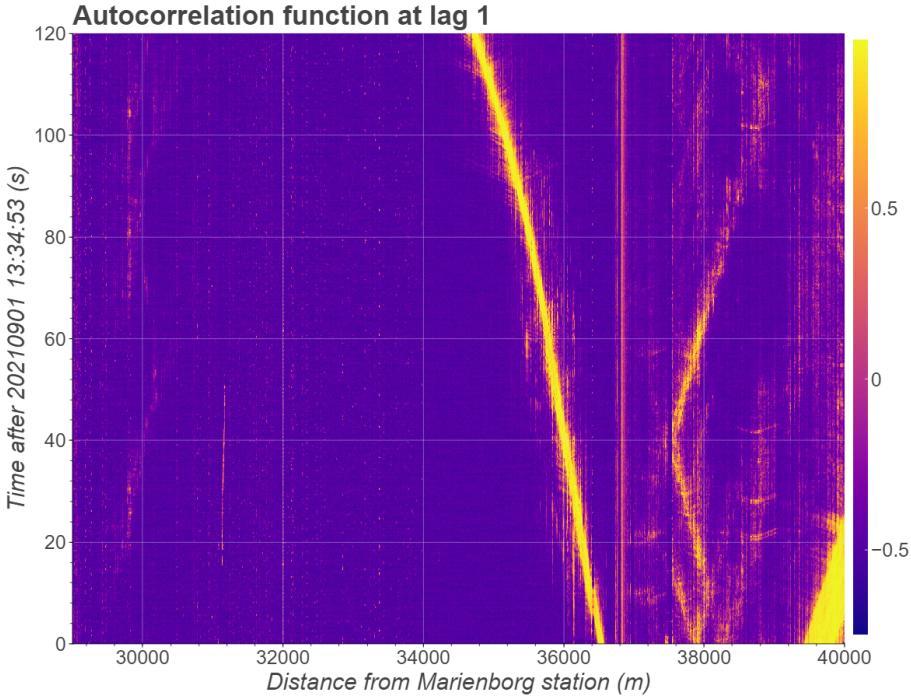
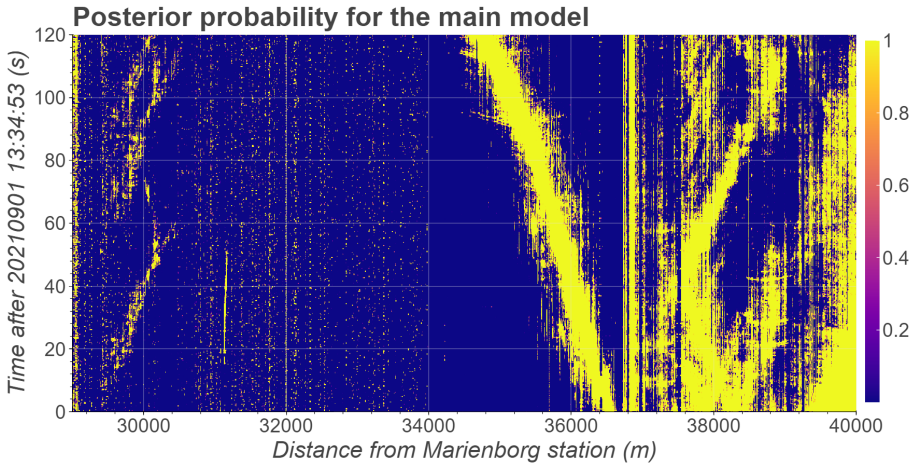
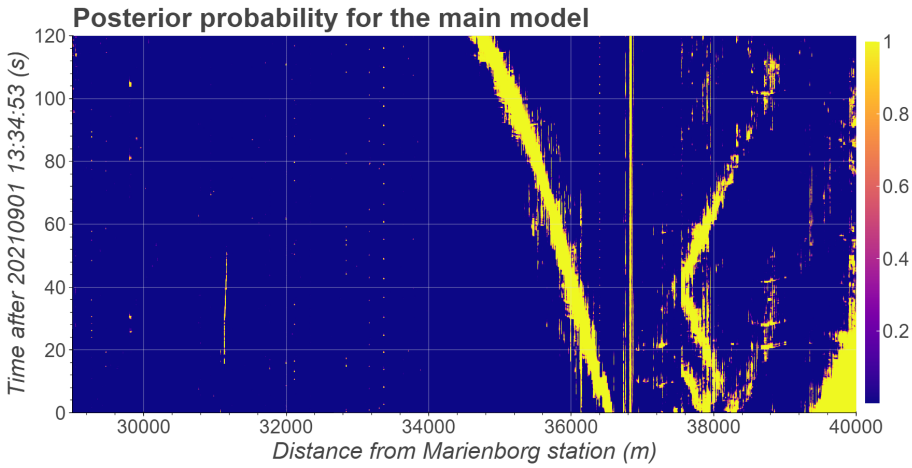


Figure 29: Heatmap of the autocorrelation at lag one between 29000 and 40000 meters.



(a) Estimated parameter values.



(b) Decided parameter values.

Figure 30: Heatmaps of the posterior probabilities $p^*(\theta_i^{r+1} = 1|\mathbf{y}^{1:r+1})$ for the main model with estimated parameter values (a), and decided parameter values (b).

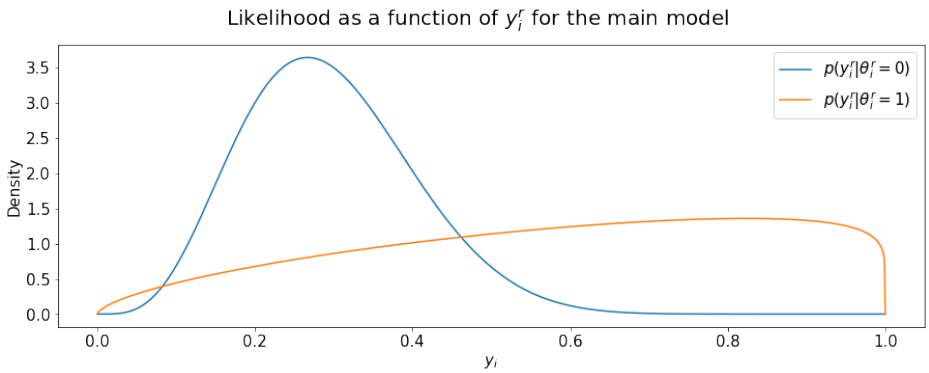


Figure 31: Likelihood displayed as a function of y_i^r, with the hyperparameters $\alpha_0 = 5$ and $\beta_0 = 12$ decided for the main model, and $\alpha_1 = 1.637$ and $\beta_1 = 1.136$ estimated in Section 7.5.

Another difference between Figures 30a and 30b is the effect on the posterior probabilities for the certain events between 34000 and 40000 meters. When we decide that the noisy data between 31000 and 34000 meters have a low probability of being events, we see that many of the posterior probabilities that lie further away from certain events become small. The reason for this is that the values for y_i^r have a larger variation further away from certain events, and the dependence structures in time and space have a bigger impact on the posterior probabilities when we choose small values for α_0 and β_0. An advantage with this is that it is easier to sort out certain events with the most distinct structures in Figure 30b.

Based on the discussion above, determining appropriate parameters for the likelihood is a difficult task. We do not know what approach to take when deciding the parameters, and having more knowledge about the nature of the signals in Figure 29 would help us decide what parameter values to use when fitting a model.

7.6.2 Conditional probabilities in time

We expected that the structures for the posterior probabilities in the main model would be relatively smooth in time and space. This does not seem true for the estimated parameter values in Figure 27 and 30a. Yet, we have not discussed the impact of the conditional probabilities in time, which are $\eta_{1:4}, \zeta_{1:4}$, and $\gamma_{1:4}$. By fitting a model using very high probabilities for adjacent points in time being in the same state, and using $\alpha_0 = 5$, $\beta_0 = 12$ to avoid overshadowing from the

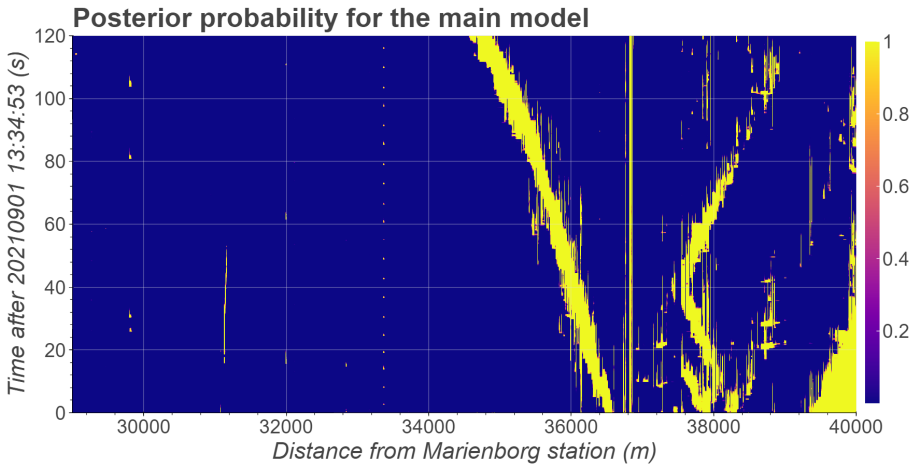


Figure 32: Heatmap of the posterior probabilities $p^*(\theta_i^{r+1} = 1|\mathbf{y}^{1:r+1})$ for the main model, using conditional probabilities in time that are high for adjacent points being in the same state.

likelihood, we obtain the result in Figure 32. As we can see, the dependence structure in time makes long chains of points classified as events, supporting the importance that the spatial and temporal structures do not override the information from the data, contained in the likelihood. The results indicate that we must be careful when deciding the hyperparameters so that neither the dependence structure nor the likelihood of the data dominates the other.

7.6.3 Further discussion

It is important to remember that the information in the heatmaps can be misleading, as the events in time layer r are not dependent on the future time points $r + j$. We present the heatmaps because they are intuitive, and events are easily detected. However, we need to emphasize that the approach we take in this project, is to use data to find the posterior probabilities for events in the present time, conditioned on current and previous observations. Therefore, we believe plots of the last layer, as presented in e.g. Figure 28 are the easiest way to display the posterior probabilities, without getting distracted by previous events.

Even though the different values for the hyperparameters have a large effect on the results, every fitted model detects the major events, which are probably the

most significant in terms of seismic activity. However, we can not be certain, as the weak signals between 29000 and 31000 meters in Figure 21 resemble cars crossing or people walking next to the tracks.

Expanding the models to higher-order Markov chains could potentially remove more short sequences of points classified as events, since using a first-order Markov chain in space only limit the single points. However, the second model in this project uses a second-order Markov assumption, but the results are not significantly different than for the first-order model. In addition, there are computational limitations when applying higher-order HMMs to a large data set such as the Trondheim data, and therefore, this is a difficult task.

8 Concluding remarks

This thesis aims to model probabilities of events by fitting Bayesian models to a DAS data set. The focus is on the main model, which incorporates information at earlier times when computing probabilities of events in the current time points. We observe that all four models detect similar events. Trying different values for the hyperparameters in the main model shows that the results are sensitive to the hyperparameters, which we must choose carefully. However, this is a challenging task because the information about events is mostly unavailable everywhere the data is collected. Having more knowledge concerning the nature behind the signals in the data set will make it easier to choose hyperparameters that fit the reality better. The results in this thesis can be a valuable resource for computing probabilities of events as new data is collected along the railway tracks, to detect possibly dangerous situations rapidly.

An interesting extension of this project can include developing the Bayesian models from detecting events, to also be able to classify them. As we mentioned in Section 1, there have been some previous studies on event detection and classification in DAS data using machine learning techniques. This can be explored in the framework of Bayesian modelling. As discussed above, in this project we take the approach to use the information at previous times to compute probabilities of events in the current time. However, for the extension discussed here, it would make more sense to use all the available data, including data future to the specific events, to gain information about them.

For an alternative analysis of the Trondheim data, a possibility is to try different measures for modelling. In this thesis, we use the autocorrelation function at lag one for small time sets in differentiated data, but there might be other measures

that capture different information from the data set. Zhong et al. (2022) investigate characteristics in background noise in DAS data and propose an algorithm for attenuating the background noise. The results in Zhong et al. (2022) can be interesting to investigate on the Trondheim data, to better understand the properties of the background noise.

Bibliography

Bane NOR, . (2023). *Railway tracks in the Trondheim area, 7028 Trondheim, Norway.* [Accessed 10 May 2023. Available at https://togkart.banenor.no/].

Barber, D. (2012). *Bayesian Reasoning and Machine Learning.* Cambridge University Press, pp. 103–126.

Butterworth, S. (1930). 'On the theory of filter amplifiers'. *Experimental Wireless & the Wireless Engineer* 7, pp. 536–541.

Casella, G. (1985). 'An introduction to empirical Bayes data analysis'. *American Statistician* 39, pp. 83–87.

Casella, G. and R.L. Berger (2002). *Statistical Inference.* 2nd ed. Duxbury.

Chaudhry, M.A. et al. (1997). 'Extension of Euler's beta function'. *Journal of Computational and Applied Mathematics* 78, pp. 19–32.

Dean, T., T. Cuny and A.H. Hartog (2016). 'The effect of gauge length on axially incident P-waves measured using fibre optic distributed vibration sensing: Gauge length effect on incident P-waves'. *Geophysical Prospecting* 65, pp. 184–193.

Devijver, P. (1985). 'Baum forward - backward algorithm revisited'. *Pattern Recognition Letters* 3, pp. 369–373.

Diaconis, P. and D. Ylvisaker (1979). 'Conjugate priors for exponential families'. *The Annals of Statistics* 7, pp. 269–281.

Farrell, S. and C. Ludwig (2009). 'Bayesian and maximum likelihood estimation of hierarchical response time models'. *Psychonomic Bulletin & Review* 15, pp. 1209–17.

Google Maps, . (2022). *Selsbakk station, 7028 Trondheim, Norway.* [Accessed 30 November 2022. Available at https://www.google.com/maps].

— (2023). *Trondheim area, 7028 Trondheim, Norway.* [Accessed 24 February 2023. Available at https://www.google.com/maps].

Haukanes, A. (2021). 'OptoDAS HDF5 file format description'. *Unpublished report.*

Heckmann, T., W. Schwanghart and J. D. Phillips (2015). 'Graph theory-Recent developments of its application in geomorphology'. *Geomorphology* 243, pp. 130–146.

Isken, M. P. et al. (2022). 'De-noising distributed acoustic sensing data using an adaptive frequency-wavenumber filter'. *Geophysical Journal International* 231, pp. 944–949.

Joyce, J. (2021). 'Bayes' theorem'. *The Stanford Encyclopedia of Philosophy.*

Kislov, K. V. and V. V. Gravirov (2022). 'Distributed acoustic sensing: a new tool or a new paradigm'. *Seismic Instruments* 58, pp. 485–508.

Li, D. et al. (2018). 'Phase unwrapping methods applied to DAS data'. *CREWES Research Report* 30.

Liu, X. et al. (2017). 'Distributed acoustic sensing with Michelson interferometer demodulation'. *Photonic Sensors* 7, pp. 193–198.

Ma, L. et al. (2022). 'Signal activity detection for fiber optic distributed acoustic sensing with adaptive-calculated threshold'. *Sensors* 22.

Padilla, M. et al. (2023). 'Optimized algorithms for temporal phase unwrapping without degrading the signal-to-noise ratio'. *From Optica Open*. URL: https://preprints.opticaopen.org/articles/preprint/Optimized_algorithms_for_temporal_phase_unwrapping_without_degrading_the_signal-to-noise_ratio/21964997.

Pinsky, M. A. and S. Karlin (2011). *An Introduction to Stochastic Modeling*. 4th ed. Boston: Academic Press.

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

Rios-Munoz, G. R. et al. (2020). 'Hidden Markov models for activity detection in atrial fibrillation electrograms'. *2020 Computing In cardiology*. Computing in Cardiology Conference.

SEAFOM, Measurement Specifications Working Group (2018). *SEAFOM measuring sensor performance-02*. Manual available at: https://seafom.com/mdocs-posts/seafom_msp_02-august-2018-pdf/.

Shang, Y. et al. (2022). 'Research progress in distributed acoustic sensing techniques'. *Sensors* 22.

Shiloh, L., A. Eyal and R. Giryes (2019). 'Efficient processing of distributed acoustic sensing data using a deep learning approach'. *Journal of Lightwave Technology* 37, pp. 4755–4762.

Shumway, R. H. and D. S. Stoffer (2017). *Time Series Analysis and Its Applications*. 4th ed. Springer.

Sklar, B. (1997). 'Rayleigh fading channels in mobile digital communication systems, part 1, characterization'. *IEEE Communications Magazine* 35, pp. 90–100.

Sung-Hyun, Y. et al. (2018). 'Log-Viterbi algorithm applied on second-order hidden Markov model for human activity recognition'. *International Journal of Distributed Sensor Networks* 14.

Taweesintananon, K. et al. (2021). 'Distributed acoustic sensing for near-surface imaging from submarine telecommunication cable: A case study in the Trondheimsfjord, Norway'. *Geophysics* 86, pp. 303–320.

Urheim, E. (2022). 'Statistical analysis of a distributed acoustic sensing data set'. *Unpublished report, Department of Mathematical Sciences, Norwegian School of Science and Technology, Trondheim, Norway.*

Virtanen, P. et al. (2020). 'SciPy 1.0: Fundamental algorithms for scientific computing in Python'. *Nature Methods* 17, pp. 261–272.

Walpole, R. E. et al. (2016). *Probability and Statistics for Engineers & Scientists*. 9th ed. Pearson.

Wang, X. and J. Guo (2008). 'Junction trees of general graphs'. *Frontiers of Mathematics in China* 3, pp. 399–413.
Zhan, Z. (2020). 'Distributed acoustic sensing turns fiber-optic cables into sensitive seismic antennas'. *Seismological Research Letters* 91, pp. 1–15.
Zhong, T. et al. (2022). 'Statistical characteristics for the background noise in distributed acoustic sensing: analysis and application to suppression'. *Journal of Seismic Exploration* 31, pp. 131–151.

Appendix

A Standard deviations in time sets

Figure 33 shows a heatmap of the standard deviations for each time set at the area between 3000 and 5500 meters. We see some weak horizontal lines at Selsbakk station around 4000 meters that match the observations from CGF. In addition, we see a vertical line at 5100 meters. However, some of the observations from CGF at Selsbakk station are hard to see for the standard deviations of the time sets.

B ACF at lag 2 and 3

The ACF values at lag two for the area around Selsbakk station, between 3000 and 5500 meters south of Marienborg are displayed as a heatmap in Figure 34, and the ACF values at lag three for the same area are displayed as a heatmap in Figure 35. We see that the signals in these plots are similar to the lag one values in Figure 11, however most of the signals take values around zero, as discussed in Section 3.2.1.

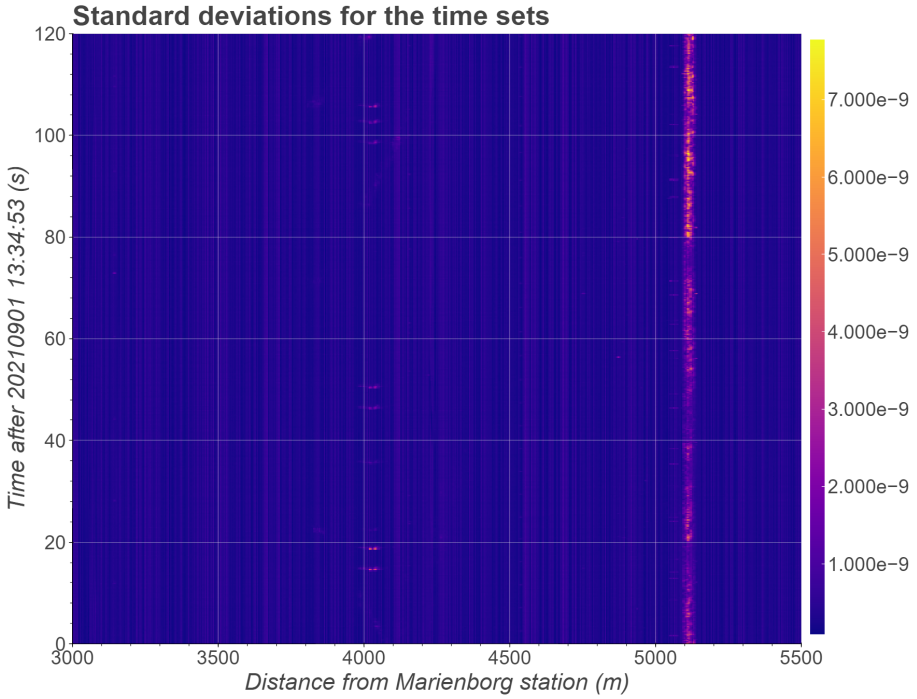


Figure 33: Heatmap of the standard deviations between 3000 and 5500 meters, and between 13:34:53 and 13:36:53.

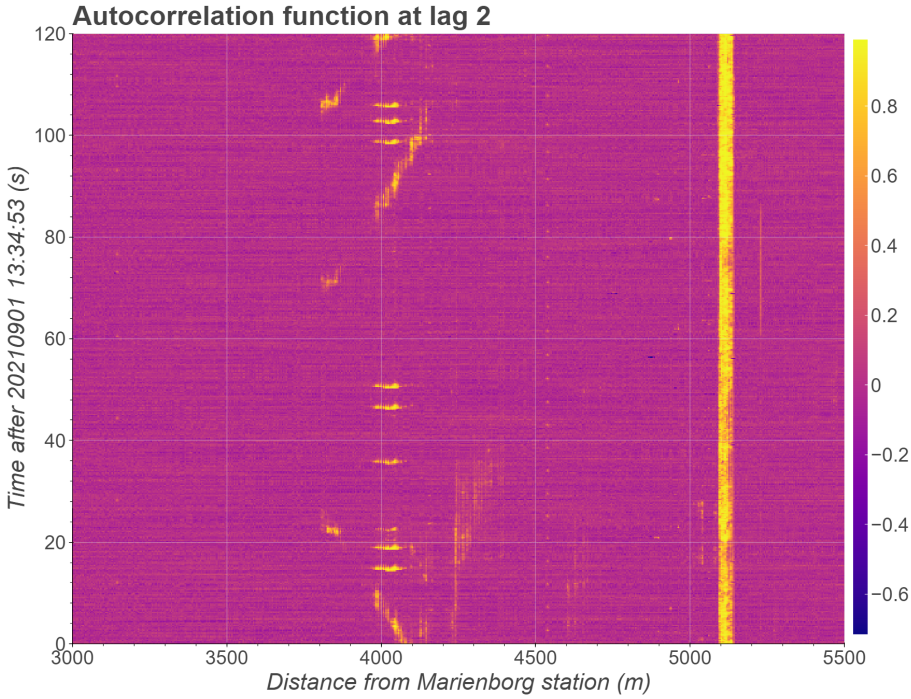


Figure 34: Heatmap of the autocorrelations at lag 2 between 3000 and 5500 meters, and between 13:34:53 and 13:36:53.

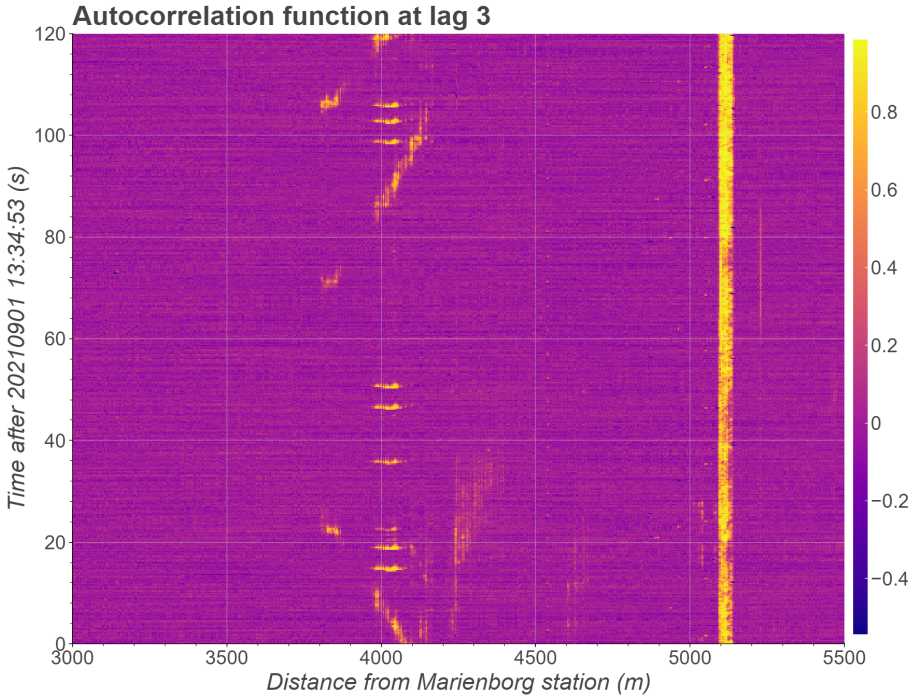
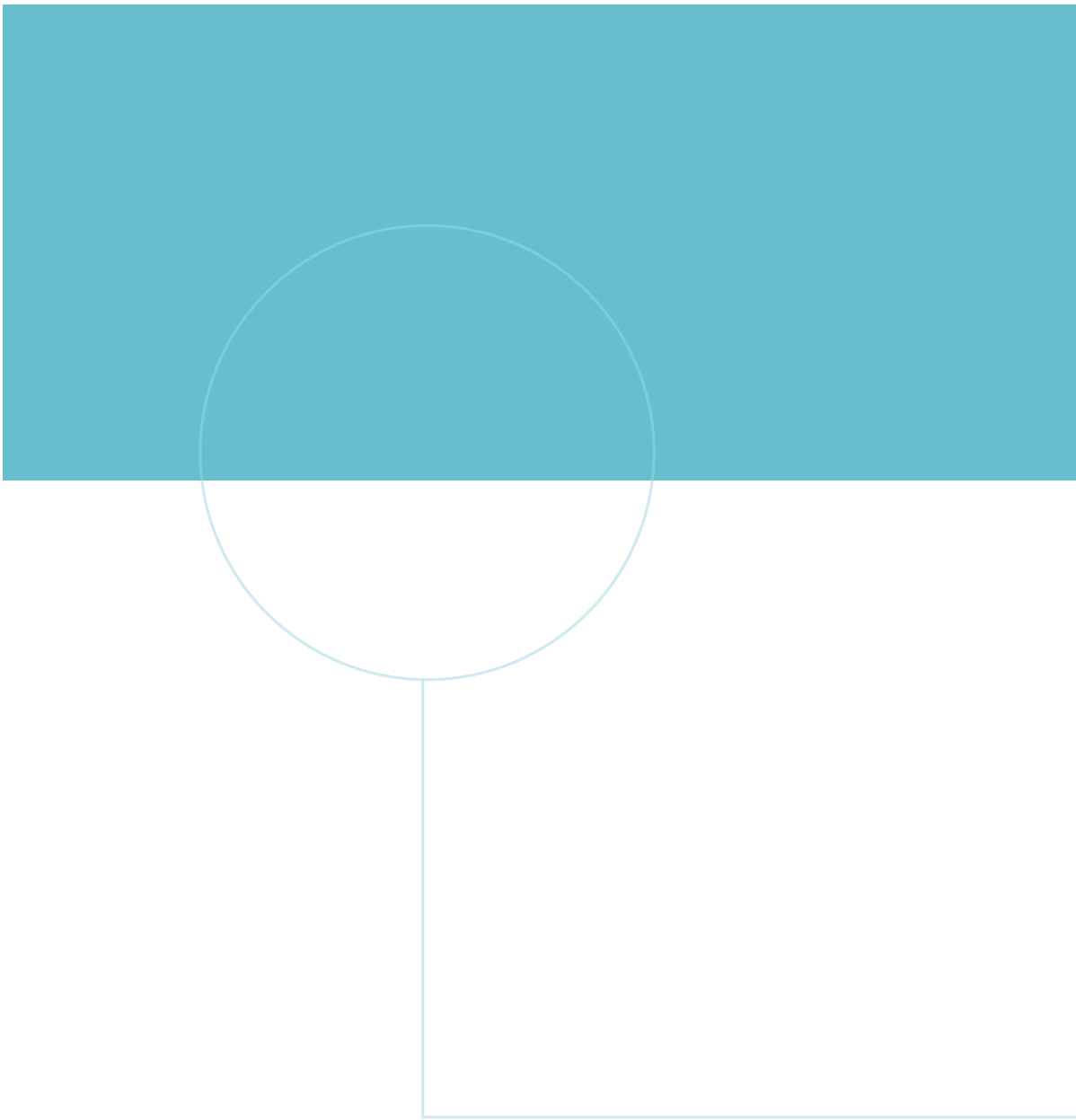


Figure 35: Heatmap of the autocorrelations at lag 3 between 3000 and 5500 meters, and between 13:34:53 and 13:36:53.



NTNU

Norwegian University of
Science and Technology