Master's thesis

2021

Martine Lysebo

Master's thesis

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

NTNU
Norwegian University of Science and Technology

Martine Lysebo

Sampling decisions for energy efficient IoT sensors using conditional copula models

A case study of noise level indicators at the student working space Koopen

June 2021

NTNU
Norwegian University of Science and Technology

Norwegian University of
Science and Technology

Sampling decisions for energy efficient IoT sensors using conditional copula models

A case study of noise level indicators at the student working space Koopen

Martine Lysebo

Industrial Mathematics
Submission date: June 2021
Supervisor: Ingelin Steinsland

Norwegian University of Science and Technology
Department of Mathematical Sciences

Preface

This thesis is submitted as the final requirement of my MSc in Industrial Mathematics at the Department of Mathematical Sciences at the Norwegian University of Science and Technology (NTNU) in Trondheim. The main work was carried out from January to May 2021.

I would like to give a special thanks to my supervisor Ingelin Steinsland. Our weekly meetings this past year has been filled with good discussions, great suggestions and excellent guidance. Not the least has it been a positive environment, with a lot of encouragement.

On a personal note, I would like to thank my family for always believing in me, being there, being positive, and reading through my thesis. I would also like to give a heartfelt thank to all the great people I have had the honor to meet, voluntare with, and learn from these past years. I have had the best time in my life so far, thanks to these new best friend. It has really been a blast, and I cant wait for more adventures in the future!

Abstract

To be able to make IoT sensors more energy efficient, we explore using dependency in data to decrease their sampling rate. In this thesis, we look at noise levels in the working environment Koopen, and a minimum of one sample each 15 minutes is set as basis for the investigation of down sampling. Based on an energy-accuracy tradeoff, the goal is to make a decision on whether the IoT sensors should sample the next quarter, or let the noise levels be estimated based on a model that conditions on the last observation.

In this work two approaches for modelling are taken; 1) to model the process conditioned on the last observed noise level directly, and 2) to model the process conditioned on the discrepancy between a weekly reference noise level and the observed one. For each of these approaches a Student t copula or a Gaussian Normal copula are set up for the time-dependencies.

The conditional copula models are evaluated based on how well they replicate the environments noise characteristics given by the peak noise level, median noise level, background noise level and variability in noise level. These noise level indicators are defined by quantiles within the given time period of 15 minutes, and how they classify is the basis for our decision making.

Generating densities of the simulated noise level indicators for different copulas, shows that the Normal copula fit the use case slightly better than the Student t copula. The choice of conditional model has great impact, and the one based on discrepancies are the best fit for the Koopen data. The modelling analysis do though conclude that the background noise level and variance in noise level can not be replicated using the models suggested in this thesis.

To evaluate the energy-accuracy tradeoff, a utility function is defined for the Koopen use case as basis for decision making. The results from the decision analysis of the conditional normal copula model based on discrepancies, show an opportunity of reduction to 1.1% and 5.8% samples a week, for the peak noise level and median noise level, respectively. This given an expected loss in energy of 0.05 for sampling 15 minutes, and no extreme measures as previous sampled noise level value from the IoT sensors.

Sammendrag

For å gjøre IoT-sensorer mer energieffektive, undersøkes bruk av avhengighetsstrukturer for å redusere samplingsfrekvensen. I denne oppgaven hvor vi ser på støynivåer i arbeidsmiljøet Koopen, er en observasjon hvert 15. minutt satt som et minimum. Basert på en avveining mellom energibruk og nøyaktighet i observasjoner, er målet å ta en avgjørelse på hvorvidt IoT-sensorene skal observere neste kvarter, eller la støynivået estimeres av en modell basert på forrige observasjon.

I dette arbeidet ser vi på to betingede modeller; 1) å modellere basert på det sist observerte støynivået direkte, og 2) å modellere basert på avviket mellom et ukentlig referansestøynivå og den siste observasjonen. For hver av disse tilnærmingene settes det opp en Student t kopula eller en Gaussisk Normal kopula for tidsavhengighetene.

De betingede kopulamodellene blir evaluert ut ifra hvor godt de replikerer omgivelsenes støyegenskaper gitt av støytopper, median støynivå, bakgrunnsstøy og variasjon i støynivå. Disse støynivåindikatorene er definert av kvantiler innen den gitte tidsperioden på 15 minutter, og hvordan de klassifiseres er grunnlaget for vår beslutningstaking.

Å generere tettheter av de estimerte støynivåindikatorene for forskjellige kopulaer, viser at Normal kopulaen passer arbeidsmiljøet noe bedre enn Student t kopulaen. Valget av betinget modell har stor innvirkning, og den som er basert på avvik, passer best for Koopen dataene. Modelleringsanalysen konkluderer i midlertidig med at bakgrunnsstøyen og variansen i støynivå ikke kan replikeres ved hjelp av modellene som er foreslått i denne oppgaven.

For å evaluere avveiningen mellom bruk av energi og nøyaktighet, defineres en nyttefunksjon tilpasset Koopen som grunnlag for beslutningstakingen. Resultatene fra beslutningsanalysen av den betingede normale kopulamodellen basert på avvik, viser en mulighet for reduksjon til henholdsvis 1,1 % og 5,8 % observasjoner i uken, for henholdsvis støytopper og median støynivå. Dette gitt et forventet tap i energi på 0,05 for observasjoner i 15 minutter, og ingen ekstreme støynivåer som tidligere observert verdi av IoT-sensorene.

Contents

Preface . . . . **i**
Abstract . . . . **iii**
Sammendrag . . . . **v**
Contents . . . . **vii**
Figures . . . . **ix**
Tables . . . . **xiii**
1 Introduction **1**
2 Mathematical formulation of noise level indicators **5**
2.1 Empirical quantiles 5
2.2 Sound data 5
2.2.1 Noise level indicators 7
3 The Koopen data **9**
3.1 Case study 9
3.2 Explanatory analysis 10
3.2.1 Raw noise level data 11
3.2.2 Noise level indicators 13
3.2.3 Weekly reference 15
4 Background **17**
4.1 Stochastic processes and time series models 17
4.2 Copulas 17
4.2.1 Empirical copula estimation 18
4.2.2 Copula selection 19
4.2.3 Conditional models based on copulas 20
4.3 Evaluation methods used for decision making in the Koopen use case 21
4.3.1 Brier score 21
4.3.2 Classification error 22
5 Methods **23**
5.1 Dependency models for noise 23
5.1.1 Conditional noise Model 1 23
5.1.2 Conditional noise Model 2 24
5.1.3 Conditional copula models for the Koopen use case 25
5.1.4 Evaluation of conditional copula models for the Koopen use case 25
5.2 Decision model for the Koopen use case 25

5.2.1 Utility function for the Koopen use case 25
5.2.2 Decision network formulation 27
5.3 Statistical software 28
6 Results 29
6.1 Fitting the dependency models 29
6.1.1 Simulation of noise level data 31
6.2 Model evaluation for noise level indicators. 32
6.3 Results for sampling decisions 37
6.3.1 Decision based on different times of week 37
6.3.2 Decision based on different sampled previous value 38
6.3.3 Evaluation on choice of expected loss in energy 41
7 Conclusion 43
Bibliography 45

Figures

2.1 An illustration of the amplitude, wavelength and oscillations of a sound wave. 6

3.1 Students working at Koopen. Taken from [15]. 9
3.2 Physical setup of Libelium devices in Koopen. Five sensors, one router and a central database connected through Wi-Fi [7]. 10
3.3 The raw data x_t observed for the whole time period. 11
3.4 The raw data x_t from Sunday 17.02.2019 to Saturday 23.02.2019. 11
3.5 The kernel density for different data sets d of the raw data x_t. . . . 12
3.6 The autocorrelation function for different data sets of the raw data. A lag of 1800 corresponds to one hour, and a lag of 10 000 in excess of five and a half hour. 13
3.7 A quarter of raw data, x_t from 09:00 to 09:15 Wednesday, Feburary 20th 2019 with its respective noise level indicators. 14
3.8 Overview of the noise level indicators over time for d =all. 14
3.9 Weekly reference noise level in dB. 15

4.1 An illustration of the transformation process for a bivariate distribution function between the random variables X_j and the uniform variables U_j. Here $X_j \sim \mathcal{N}(0,1)$ with 2000 random generated samples. (a) Scatterplot of the margins X_1 and X_2. (b) The cumulative distribution function $F(X)$ used to transform the data. (c) Scatterplot of the uniform variables U_1 and U_2 on [0,1]. 18
4.2 Scatter plot of 2000 random samples from the Gaussian copula for $\rho = \{0.3, 0.9\}$. Taken from [22]. 20
4.3 Scatter plot of 2000 random samples from the Student t copula for $\rho = \{0.3, 0.9\}$ and $df = \{2, 7\}$. Taken from [22]. 21
4.4 An illustration on the modelling process from a copula. (a) The given bivariate copula. (b) The probability distribution function given previous value=0.2. (c) Pointer to the next value after generating a random variable on [0,1]. 21

5.1 An overview of the 16 different conditional copula models $C^d_{M,c}$ used in this thesis. 26

5.2 Influence diagram for the decision used in this thesis on whether to sample with an IoT sensor or model the next 15 minutes. 28

6.1 Scatterplot using the conditional noise Model 1 of the marginals unitary transformations for all data sets. 30
6.2 Scatterplot using the conditional noise Model 2 of the diff marginals unitary transformations for all data sets. 30
6.3 Bivariate Normal copula and Student t copula with parameters $\rho = 0.92$ and $\{\rho = 0.94, df = 2.9\}$, respectively for conditional model 1. 31
6.4 Bivariate Normal copula and Student t copula with parameters $\rho = 0.81$ and $\{\rho = 0.92, df = 2.0\}$, respectively for conditional model 2. 32
6.5 A random modelled quarter using conditional noise Model 1 for both the Student t and Normal copula for d =all. Given previous value $x_0 = 45.9$. 32
6.6 A random modelled quarter using conditional noise Model 2 for both the Student t and Normal copula for d =all. Given previous value $x_0 = 45.9$ and $tow =$ Wednesday 09:00. 33
6.7 Model comparison for different conditional copula models of the data set d =all for all noise level indicators. 34
6.8 Model comparison for different conditional copula models of the data set d =work for all noise level indicators. 34
6.9 Model comparison for different conditional copula models of the data set d =no work for all noise level indicators. 35
6.10 Model comparison for different conditional copula models of the data set d =busy hour for all noise level indicators. 35
6.11 QQ-plots for all indicators of the modelled data compared to the raw data for conditional noise Model 1 and 2, copula, $c =$ Normal and data set, $d =$ all. The red line is plotted as reference to ideal behaviour. 36
6.12 Binary decision on whether to sample(grey) or not(white) throughout a week based on the conditional copula model $C^{all}_{M2,N}$ and L^{all}_{10}. Previous sampled values are taken from the raw data at given time of week 16.02.2019 to 23.02.2019. (a) E=0.005 (b) E=0.05 (c) E=0.15. 38
6.13 Binary decision on whether to sample(grey) or not(white) throughout a week based on the conditional copula model $C^{all}_{M2,N}$ and L^{all}_{50}. Previous sampled values are taken from the raw data at given time of week 16.02.2019 to 23.02.2019. (a) E=0.005 (b) E=0.05 (c) E=0.15. 39
6.14 The loss in accuracy, $A(t)$ for the conditional copula model $C^{all}_{M2,N}$ and L^{all}_{10}, with different sampled previous noise levels. The loss in energy, E are plotted as dotted lines for comparison. 39

6.15 The loss in accuracy, $A(t)$ for the conditional copula model $C_{M2,N}^{all}$ and L_{50}^{all}, with different sampled previous noise levels. The loss in energy, E are plotted as dotted lines for comparison. 40

Tables

2.1 Common noise levels defined by [1]. 7
2.2 Noise level indicators used in this thesis. 7
2.3 Indicator ranges for classification of working environments defined by [7]. 8

3.1 Data sets d; all, work, no work and busy hour. 12

6.1 Conditional bivrariate Normal and Student t copula models used in this thesis with selected best fit parameters ρ and df = degrees of freedom. 31

Chapter 1

Introduction

Everywhere around us we hear sounds. From cars driving by, children playing hide and seek in the park, bird chirps and church bells in the distance. Vibrating objects cause slight changes in air pressure, and travel as waves through the air. When these waves reach your ear, you hear it as sound. Unwanted sound is often referred to as noise, but this depends upon circumstances and the person listening. If you are exposed to loud noise levels over time, it may cause problems to your hearing [1].

Housing, working and relaxing have different standards for acceptable noise levels. The question is whether such standards are taken into account in our everyday life? Do construction companies, employers and authorities consider them when they set up houses, establish working environments and construct cities? Recommended noise limits are exceeded worldwide according to recent evidence[2].

The effects of noise are considerable; interference with communication, disturbance of sleep, stress, annoyance, effects on performance and in the worst case hearing loss. Therefore, it is important to be aware of what amount of noise levels people are exposed to in different situations [3]. Traditional measuring of noise levels requires manual operation and expensive equipment, and have shown to come short in reflection of actual noise characteristics.

Internet of things (IoT) has lately been raised as a hot topic in communication technology [4]. One use smart sensors to monitor different aspects around us. From your pulse, to the number of bicycles that passes a specific spot in Trondheim every day. Fast output is generated by machines that automates and controls the huge amounts of information logged. The human effort is minimized because the devices of IoT interact and communicate directly with each other through the internet. Many sensors do not need to be connected to the electric grid, and are instead driven by battery power. This gives the opportunity of placing them almost anywhere. To exploit the full potential of these sensors, we want to increase their lifetime and effect as much as possible. An ideal sensor would use minimal energy on sampling and communicating, and still log enough data to gain the desired insight. This tradeoff depends on the IoT sensors use case. It is stated that there are huge potentials for down sampling by energy-accuracy tradeoffs [5].

The data set Koopen let us examine noise levels monitored every other second for almost three months by IoT sensors. This is done in a working environment at the Norwegian University of Science and Technology(NTNU) in Trondheim.

To be able to adapt the time and type of work, it is of interest to know the noise level's characteristics at given times. Thus, we want to point out the sounds that are common for the environment, which peaks it has, and what type of background noise you could expect. For noise, it is stated [6] that the 50-, 90- and 10-percentile describes these properties, and that the variability in noise level can be expressed by the subtraction between the 90- and 10-percentile. To set up such quantiles we need to define a time period, within which they are calculated. For this type of work environment, 15 minutes seems to be an appropriate choice of time period, and are used throughout the thesis. In Chapter 3 the Koopen data are explored, and shown to be time dependent, with a non-gaussian distribution function.

This master thesis builds upon my own project thesis motivated by the same need, using the same case study. The data are provided by researchers at NTNU that also consider down sampling for energy efficient IoT sensors [7]. Their approach to the possibility of down sampling have been a direct cut in samples, and by a predictive random forest strategy [5], as well as adaptive sensing based on deep reinforcement learning [8].

The strategy for down sampling in this thesis, is based on using time dependence to model the Koopen data. Our working hypothesis is that by letting the IoT sensors sample once each 15 minutes, the next quarter of noise levels are dependent on this previous sample. Making a decision on whether to model this next quarter, or let the IoT sensors sample it is the thesis' goal. Thus, we need to investigate if dependency itself is enough to replicate real noise level characteristics.

Applying copula models as the dependence structure between random variables, has been used as a method in several research areas. Taking it a step further, the copulas can also be used to estimate multiple aspects of real life. The most known examples may be in finance, but wind energy, flood events and environmental sciences are other areas where copula-based estimation are used [9][10][11]. The background for modelling using copulas are presented in Chapter 4.

Our Koopen data can be referred to as time series, and Copulas are in this thesis used as a framework for their dependency. Two different conditional noise models are set up in Chapter 5. The first is based on copula modelling from observations taken straight from the raw data. The second conditional noise model uses the discrepancies from a weekly reference time series model of the Koopen data as basis for the set up copula.

For our decision making, optimizing the utility, lays the foundation for whether we decide to sample the next quarter or not. The utility function weights the energy-accuracy tradeoff for monitoring with IoT sensors. The loss in accuracy by our conditional copula models is set up using Brier score, and a classification success rate for the noise level indicators at given time. The loss inn energy for sampling 15 minutes are considered time independent.

In Chapter 6 the results are presented. A conclusion on this work as well as recommendations for further work can be seen in Chapter 7.

Chapter 2

Mathematical formulation of noise level indicators

In this chapter a mathematical formulation of noise level indicators are introduced. These indicators are based on quantiles, which is defined in Section 2.1. In Section 2.2 sound measuring and the noise level indicators used in this thesis are introduced.

2.1 Empirical quantiles

Quantiles are a property of probability distributions first introduced by Maurice G. Kendall in 1940. The quantile of a distribution is the x_p such that a proportion p of the values are less than or equal to x_p with $0 < p < 1$. Empirical quantiles are often constructed by order statistics by sorting the data in ascending order as a sequence $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ with n data points. Let $F(x)$ be the cumulative probability function. Then the p-th quantile of the probability distribution can be obtained by,

$$Q(p; F(x)) = x_p = \begin{cases} x_{((n+1)p)} & \text{if (n+1)p is an integer} \\ \frac{(x_{([(n+1)p])} + x_{([(n+1)p]+1)})}{2} & \text{otherwise} \end{cases} \tag{2.1}$$

where $[a]$ is the greatest integer not exceeding a [12]. If $(n+1)p$ is not an integer there are several more complex ways to set up the quantile, elaborated by Hyndman and Fan (1996)[13], not covered in this thesis.

2.2 Sound data

Sound is according to the Cambridge dictionary, defined as something that you can hear or that can be heard. Noise is when sound becomes unwanted, unpleasant or loud. The loudness of the sound is represented by it's amplitude, and refers to the magnitude of an oscillation. Frequency is the rate at which a source produces sound waves and is expressed in hertz(Hz). 1 Hz equals one oscillation per second. These terms are illustrated for sound waves in Figure 2.1.

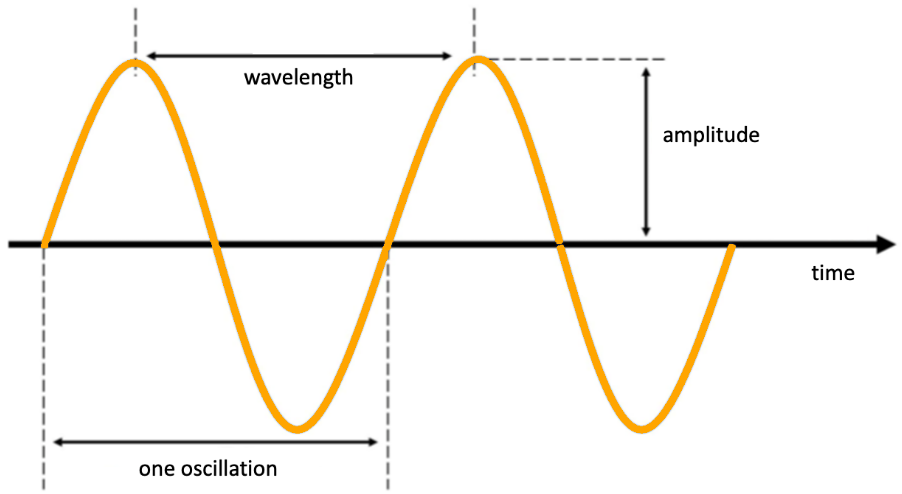

Figure 2.1: An illustration of the amplitude, wavelength and oscillations of a sound wave.

The quantity decibel(dB) is commonly used for measuring sound and is defined by [14] as,

$$S = 10log_{10}(\frac{A}{B}) \text{ dB} \tag{2.2}$$

where B is a reference level and A the measured level, both in intensity, power or pressure. Here the reference level depends on what equipment you use to monitor the noise level. Descibel(dB) is as stated a dimensionless logarithmic unit. To get S as the sound pressure level, we need A and B expressed in air pressure. For sound, the amount of air pressure fluctuation the source creates, is its sound pressure expressed in Pascals (Pa).

Human ears are less sensitive to low and high frequencies of sound, and noise measurement readings can be adjusted to correspond to this peculiarity by using frequency weighting [3]. To adjust the measured sound to a curve, A-weighting is a standard way of electronically filtering noise to represent what the human ear hears. Some commonly known A-weighted sound pressure noise levels are expressed in Table 2.1.

The speed of which a sound level meter measures sounds responds to changes in noise levels, is called time weighting [14]. Modern sound level meters have two options of time-weighting; SLOW and FAST, with respectively time constants of 1 second and 125 milliseconds. The SLOW mode is typically used to determine and observe sounds slowly changing average value, which we use in this thesis.

Everyday Sounds and Noises	Average Sound Pressure Level(dB)
Softest sound that can be heard	0
Soft whisper	30
Normal conversation, air conditioner	60
Washing machine, dishwasher	70
Motorcycle	95
Shouting or barking in the ear	110

Table 2.1: Common noise levels defined by [1].

Sound measurements of the sound pressure levels in an environment can therefore be attained by (2.2) using a sound level meter. The time- and frequency weighting are chosen by the user, and depend on type of equipment.

2.2.1 Noise level indicators

In this thesis the sound level observations can be referred to as,

$$x_{,}x_2, \dots, x_t \dots, x_{n-1}, x_n,$$

with x_t as sound measurement at time point t with a total of n observations.

To evaluate noise characteristics in an environment one uses noise level indicators. These are set up by empirical quantiles calculated within a time period T. The sound pressure level observations for each time period can be expressed as,

$$x_{t:T} = (x_t, x_{t+1}, \dots, x_{t+T}).$$

The choice of time period T is determined suitable for the context. Using a working environment as case study, T=15 minutes seems like a suitable choice for looking at relevant noise characteristics that effect it's users.

The noise level indicators used in this thesis are chosen based on present work on the Koopen data set [7][5] with T =15 min, and are presented in Table 2.2.

$\boldsymbol{L_{10}}$	The peaks of noise defined by the sound pressure level exceeded for 10% of the time period T considered.
$\boldsymbol{L_{50}}$	The median noise level for the time period T considered
$\boldsymbol{L_{90}}$	The background noise level defined by the sound pressure level exceeded for 90% of the time period T considered
$\boldsymbol{L_{10-90}}$	The variability in noise level inside the interval i considered. It is defined by the difference between the peaks of noise L_{10} and the background noise L_{90}.

Table 2.2: Noise level indicators used in this thesis.

Having $x_{t:T}$ as the distribution of sound in the time interval $[t, t+T]$, let $F(x)$ denote the cumulative distribution. The evaluated noise indicators used in this thesis can be expressed as,

$$L_\xi(x_{t:T}) = Q(\frac{1-\xi}{100}, F(x)) \qquad \text{for } \xi \in \{10, 50, 90\}. \tag{2.3}$$

Here $L_\xi(x_{t:T})$ is the noise level indicator in decibel and Q the empirical quantile from (2.1). For ξ=10-90 we have,

$$L_{10-90}(x_{t:T}) = L_{10}(x_{t:T}) - L_{90}(x_{t:T}). \tag{2.4}$$

To evaluate and classify the noise level indicators, ranges customized to the given context are needed. In this project we follow the previous work in [7] and use,

$$Class(L_\xi(x_{t:T})) = k_\xi(x_{t:T}) \in \{\text{good,fair,poor}\} \tag{2.5}$$

to classify the noise level indicators. The ranges for different classes are based on international standards for acceptable levels in working environments set by Table 2.3.

Indicator	**Good**	**Fair**	**Poor**
L_{10}	< 50	[50,60]	> 60
L_{50}	< 45	[45,55]	> 55
L_{90}	< 42	[42,52]	> 52
L_{10-90}	< 3	[3,5]	> 5

Table 2.3: Indicator ranges for classification of working environments defined by [7].

Chapter 3

The Koopen data

In this Chapter the data set used in this project is presented and explored. Section 3.1 introduces the Koopen case study, and the measurements used. In Section 3.2 an explanatory analysis of this data is presented.

3.1 Case study

At the Norwegian university of science and technology(NTNU) in Trondheim, there is a working environment for students named Koopen. It has a variation of stations for working and different types of learning, with a full capacity of about 80 students. The ceiling is high, and the space is connected directly to other parts of the building without walls separating them. The Koopen-area can be seen in Figure 3.1.

Figure 3.1: Students working at Koopen. Taken from [15].

At Koopen the sound pressure levels were monitored every other second from 06.02.2019 to 26.04.2019 with five different sensors. The background for the setup of this data is presented in [7]. Libelium devices [16] were used as IoT sensors connected to power and logging through a router to a central database. These consist of sound level meters measuring A-weighted sound pressure levels with SLOW time weighting as presented in Section 2.2. The physical setup of the IoT sensors can be seen in Figure 3.2. Even though the sensors are connected to power, the use case can be applied to investigate energy efficient down sampling strategies for sensors driven by batteries as well.

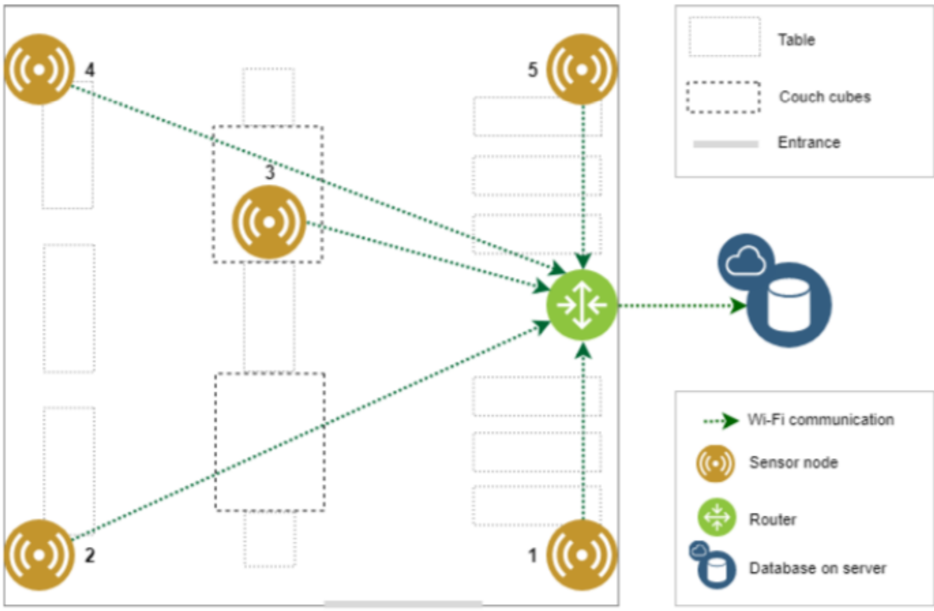

Figure 3.2: Physical setup of Libelium devices in Koopen. Five sensors, one router and a central database connected through Wi-Fi [7].

In this thesis, only observations from one of the sensors, sensor 1 are considered. Our data should therefore ideally consist of about 3.4 million observations. 12.5% are missing, which results in about 3 million logged data points. We denote these raw observations as x_t with the time point t. They are presented in Figure 3.3 for the whole logged time period, and in Figure 3.4 for a week late February 2019.

3.2 Explanatory analysis

In this Section the data is explored. Our hypothesis' are that there are systematic patterns following time of week and time dependencies due to the type of environment we study sound from. Expecting higher and more variable noise levels within work hours, than at night and in weekends.

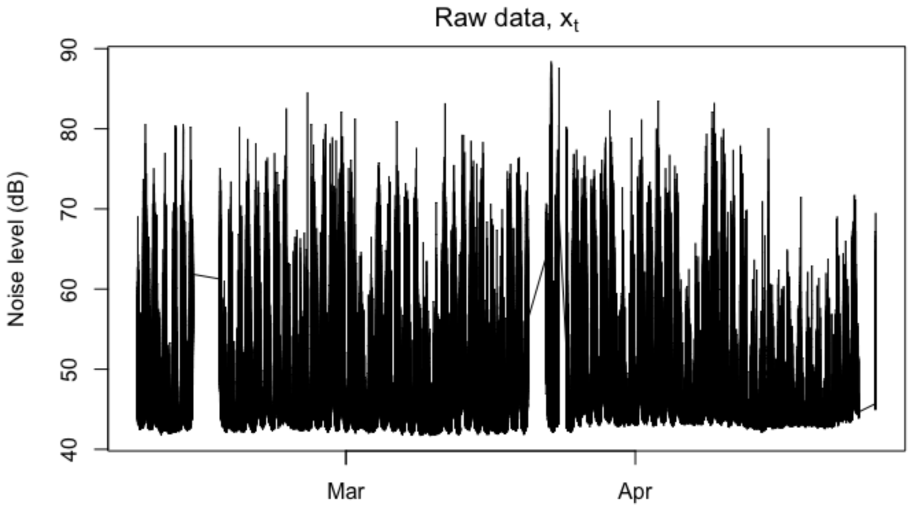

Figure 3.3: The raw data x_t observed for the whole time period.

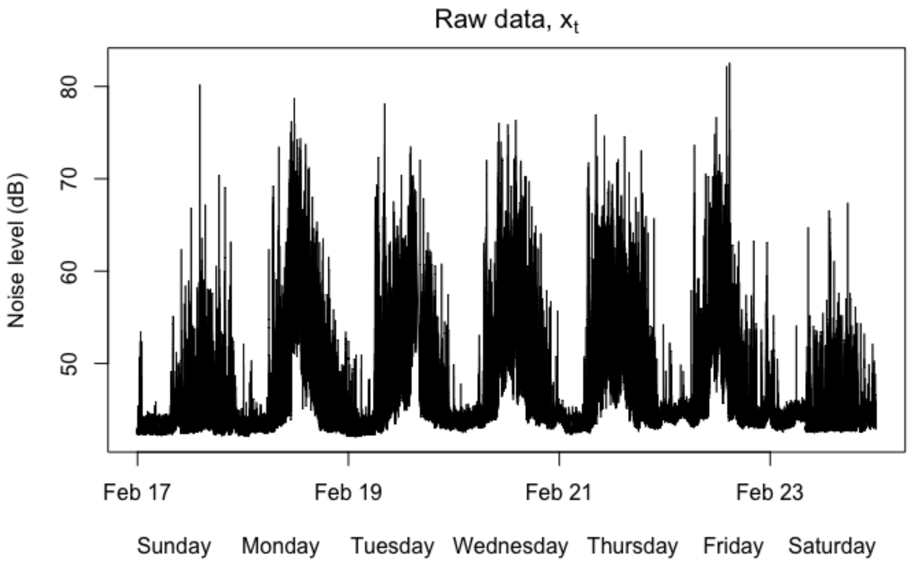

Figure 3.4: The raw data x_t from Sunday 17.02.2019 to Saturday 23.02.2019.

3.2.1 Raw noise level data

From Figure 3.3 it is clear that the noise levels variate over time. However, the noise levels are never below 42 dB, which from Table 2.3 indicates a high background noise level, L_{90}, that will never classify as good. This seems to be the constant minimum noise level, probably coming from ventilation. Since we are studying a working environment, patterns in noise levels are to be expected. Figure 3.3 of the raw noise levels show seasonal variations in late April, which corresponds to the Easter break. In Figure 3.4 of one week of raw noise level data, we see weekly and daily variations as expected, depending on the usage of the work space. Due to these variations four data sets; d={All, Work, No work, Busy hour} are introduced and presented in Table 3.1. The norwegian calender[17] is used to set weekdays and holidays.

Data set, d	**Description**
All	All logged data.
Work	Weekdays between 07-19, excluded holidays
No work	All data in the data set all that is not in the data set work.
Busy hour	All data between 12-13 within the data set work.

Table 3.1: Data sets d; all, work, no work and busy hour.

The kernel density plots of different data sets d, can be found in Figure 3.5. The density functions are all dominated by a peak at low noise levels (about 44 dB), and a long right tail to higher noise levels. This is not as prominent for work and busy hour, where the kernel density has more mass on louder noise levels. This is expected from Figure 3.4 and our hypothesis with high noise levels occurring within work hours. The density functions shows that the distributions are skewed by forming tails, implying non-Gaussian distributions.

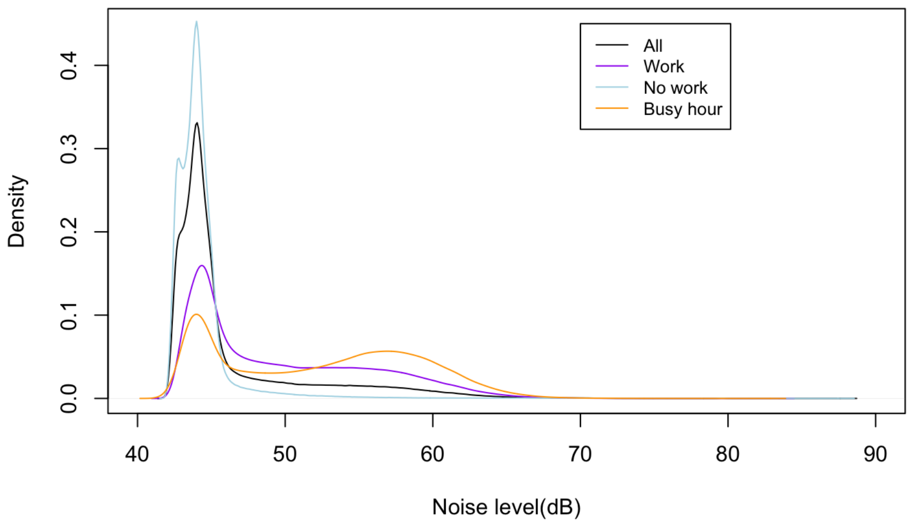

Figure 3.5: The kernel density for different data sets d of the raw data x_t.

To evaluate how well the present value of the time series data are related to its past values, we use the empirical auto correlation function. It is plotted for different data sets d in Figure 3.6.

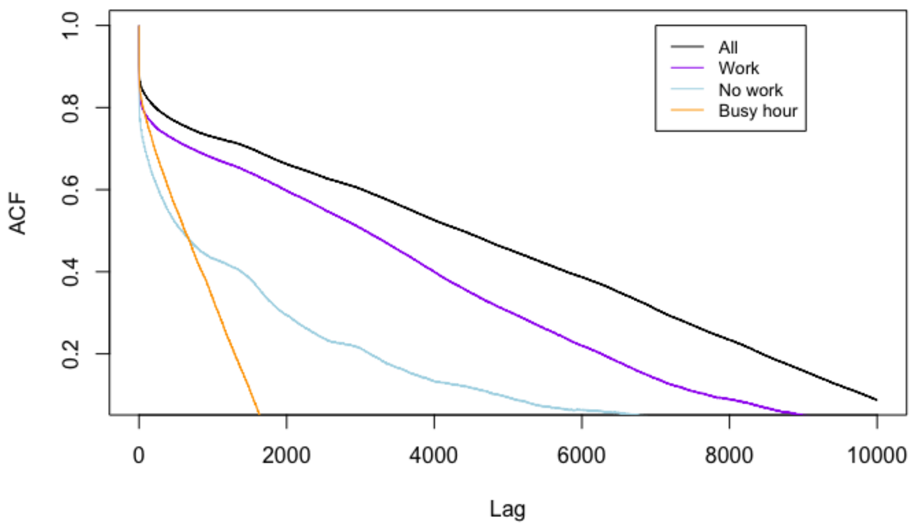

Figure 3.6: The autocorrelation function for different data sets of the raw data. A lag of 1800 corresponds to one hour, and a lag of 10 000 in excess of five and a half hour.

We see from Figure 3.6 that the serial correlation in the data differs some between data sets. The fact that the no work data set is faster than the one for work, seems in line with our hypothesis of more variation within work hours. That the slowest auto correlation function belongs to all, makes sense since it contains both work and no work data.

3.2.2 Noise level indicators

To get a clear picture of the noise characteristics of the environment, we use noise level indicators as presented in Section 2.2.1. L_{10}, L_{50}, L_{90} and L_{10-90} respectively gives us information about the peak noise level, median noise level, background noise level and variability in noise level. The sample period we calculate the quantiles for each indicator within, is set as stated to T=15 minutes. For our measured data every other second this correspond to n_T=450 samples within each time period. An example of this quarter of raw data with its respective noise level indicators, are plotted in Figure 3.7.

An overview of the noise level indicators for all intervals over time are presented in Figure 3.8. To classify whether the indicator refers to a poor, fair or good level the ranges in Table 2.3 are used.

In Figure 3.8 we can see that the peak noise level(L_{10}^{all}) and the median noise level(L_{50}^{all}) are classified as good for almost all time outside peak hours Monday to Friday. The background noise level(L_{90}^{all}) is never classified as good, and this indicates that the general noise, from by example ventilation, always exceeds the

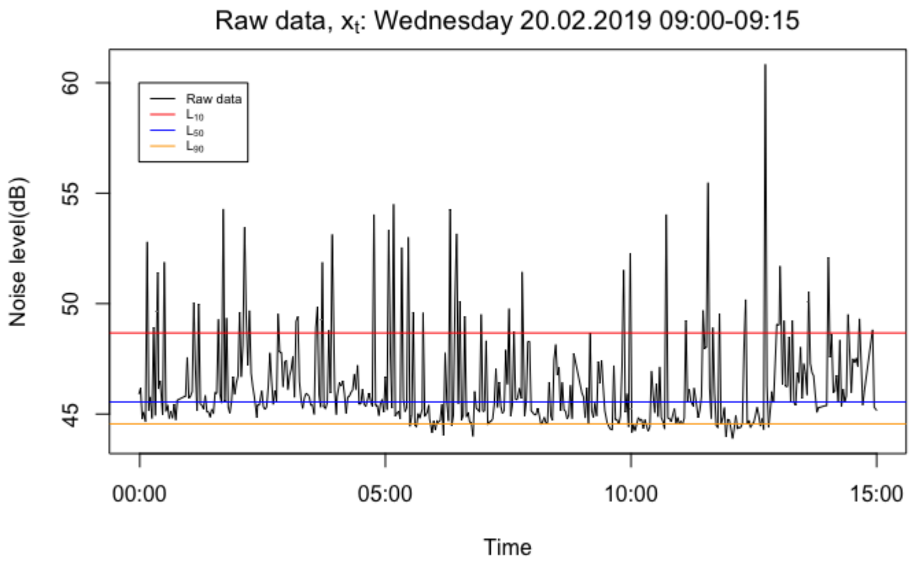

Figure 3.7: A quarter of raw data, x_t from 09:00 to 09:15 Wednesday, Feburary 20th 2019 with its respective noise level indicators.

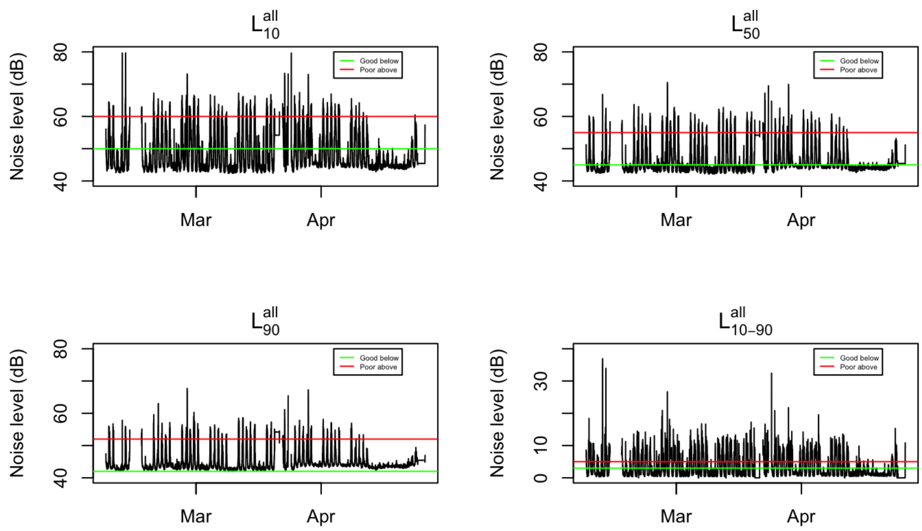

Figure 3.8: Overview of the noise level indicators over time for d =all.

desired sound level for such an environment. The variance in noise level(L_{10-90}^{all}) seems to classify as poor most often.

3.2.3 Weekly reference

To investigate how the noise levels varies throughout a week, the variable tow is introduced referring to the time of week. Within a week there are 672 intervals of 15 minutes to calculate noise level indicators within, and therefore $tow \in 1, 2, \ldots, 672$. As weekly reference we use the average of $L_{50}^{all}(x_{t:T})$ at all tow calculated over all weeks j defined as,

$$\mu_{50}^{tow} = \frac{1}{12} \sum_{j=1}^{12} L_{50}^{j}(x_{t:T}) = (\mu_{50}^{1}, \mu_{50}^{2} \ldots, \mu_{50}^{672}). \tag{3.1}$$

Here tow is dependent on the time t for which the random variable $x_{t:T}$ occur and the week number $j \in \{1, 2, \ldots 12\}$. The weekly references are presented in Figure 3.9.

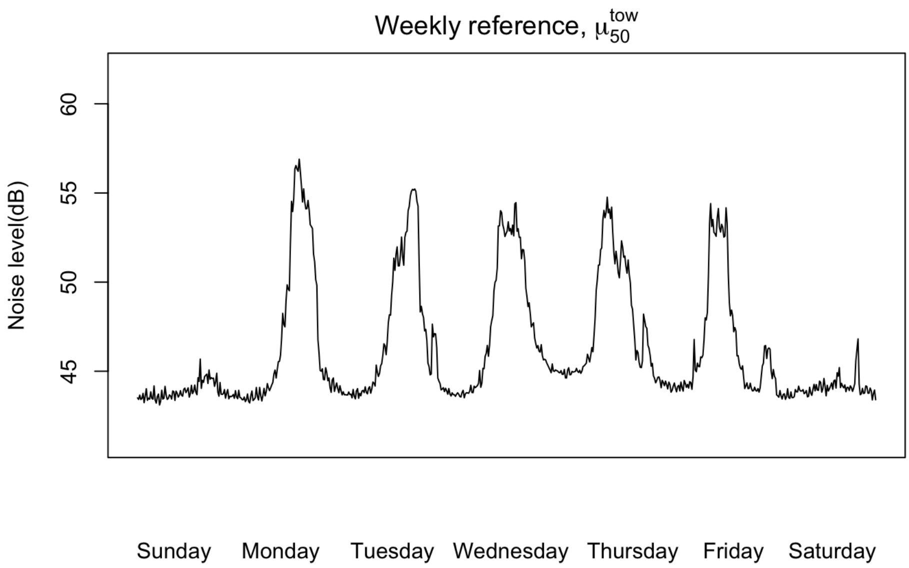

Figure 3.9: Weekly reference noise level in dB.

Figure 3.9 confirms our hypothesis of highest noise levels within working hours from Monday to Friday.

Chapter 4

Background

In this chapter the background for the modelling and methods used are presented. Stochastic processes and time series models are introduced in Section 4.1. Copulas as a framework for dependency is presented in Section 4.2, and the evaluation methods used for decision making in this thesis, are introduced in Section 4.3.

4.1 Stochastic processes and time series models

The noise levels measured over time, can be seen as a stochastic process being sampled as time series. This section is based on theory from Rausand and Høyland (2004) [18]. A collection of discrete random variables X_t indexed by time, forms a stochastic process and can be denoted by,

$$\{X_t\}_{t \in N}$$

where N is the total time space. A common assumption is that future states of this process only depends on the current state. This Markov property is defined as,

$$X_{t+1}|(X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = X_{t+1}|(X_t = x_t). \tag{4.1}$$

The process is said to be stationary when for any $t \in N$ with n set of indexes, $X_{t_1}, \ldots X_{t_n}$ all have the same probability distribution.

4.2 Copulas

The evaluation of the data presented in Chapter 3 with kernel densities from Figure 3.5, shows that our data are non-Gaussian, hence we need another framework for dependency than the multivariate Gaussian distribution gives. To model the dependency between random variables in our data, we use copulas as a method for describing the dependencies between cumulative distribution functions based on their marginal distributions. By transforming the marginal distributions into uniform distributions when using copulas, it allows us to describe the marginal distributions, and their joint dependencies(the copulas) separately.

This section is based on theory from Jaworski, Durante, Härdle and Rychlik (2009) [19]. A d-dimensional copula is defined as a d-variate distribution function on $\mathbb{I}^d$ whose univariate marginals are uniformly distributed on $\mathbb{I}$. Let $\mathbf{U} = (U_1, U_2, \dots U_d)$ be a random variable associated to this d-copula such that $U_j \sim \mathscr{U}(\mathbb{I})$ for every $j \in \{1, 2, \dots d\}$ and $\mathbf{U} \sim C$.

Letting F be a d-dimensional distribution function with univariate margins $F_1, F_2, \dots, F_d$ Sklar's theorem states that there exists a copula C such that for all $(x_1, x_2, \dots x_d) \in \mathbb{R}^{-d}$,

$$F(x_1, x_2, \dots, x_d) = C(F_1(x_1), F_2(x_2), \dots F_d(x_d)). \tag{4.2}$$

The copula C can then be obtained by the formula,

$$C(u_1, u_2, \dots u_d) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), \dots F_d^{-1}(u_d)) \tag{4.3}$$

where F_j^{-1} is the pseudo-inverse of F_j. Hence having uniform marginals on $\mathbb{I}$ and preserving the components dependence, copulas are a way of transforming the random variables $(X_1, X_2, \dots, X_d)$ into other random variables $(U_1, U_2, \dots, U_d) = (F_1(X_1), F_2(X_2), \dots F_d(X_d))$. An illustration of the transformation process for a bivariate distribution function is illustrated in Figure 4.1 for 2000 random generated samples $X_j \sim \mathscr{N}(0, 1)$.

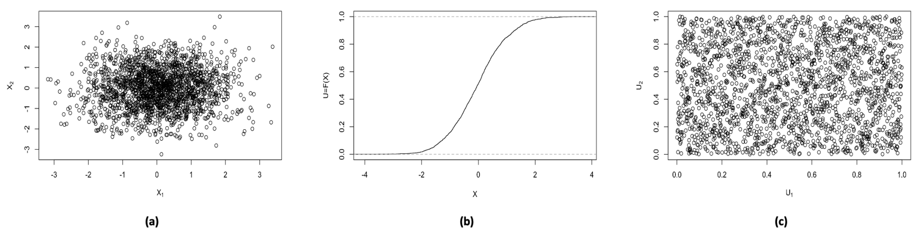

Figure 4.1: An illustration of the transformation process for a bivariate distribution function between the random variables X_j and the uniform variables U_j. Here $X_j \sim \mathscr{N}(0, 1)$ with 2000 random generated samples. (a) Scatterplot of the margins X_1 and X_2. (b) The cumulative distribution function $F(X)$ used to transform the data. (c) Scatterplot of the uniform variables U_1 and U_2 on [0,1].

4.2.1 Empirical copula estimation

The Koopen data does not have any known marginal distribution, hence we are in need of an nonparametric approach to our copula estimation based on observations. Assume multivariate data observations $(X_1^t, X_2^t, \dots, X_d^t)$ from a random vector $(X_1, X_2, \dots, X_d)$ with $t = 1, 2, \dots T$ as the time point. From $(U_1^t, U_2^t, \dots, U_d^t) = (F_1(X_1^t), F_2(X_2^t), \dots F_d(X_d^t))$ we can set up an estimator,

$$\hat{C}(u_1, u_2, \ldots, u_d) = \hat{F}(\hat{F}_1^{-1}(u_1), \hat{F}_2^{-1}(u_2), \ldots \hat{F}_d^{-1}(u_d)) \tag{4.4}$$

Here $\hat{F}$ is a nonparametric estimator of the d-dimensional distribution function with,

$$\hat{F}_j^{-1}(s) = \{t | \hat{F}_i(t) \geq s\} = \hat{U}_j(s), \tag{4.5}$$

as a nonparametric estimator of its pseudo-inverse. These marginal distributions $\hat{F}_i$ are usually unknown, but by using the empirical distribution functions they can be found as,

$$\hat{F}_j^T(x) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}(\hat{X}_j^t \leq x_j) \text{ for } x \in \mathbb{R}, \tag{4.6}$$

with **1** known as the indicator function.

4.2.2 Copula selection

In this thesis we focus on bivariate copulas with $d = 2$. Then the uniform random variables are U_1, U_2, and F a 2-dimensional distribution function with marginals $F_1(x_1)$ and $F_2(x_2)$. This subsection is based on Kurowicka and Joe (2010) [20]. To select an appropriate bivariate copula, the parameters need to be estimated for all evaluated copulas based on the marginal distributions.

Using a nonparametric approach with observed multivariate data X_1^t and X_2^t we use the following maximum likelihood estimation for estimating the copula parameter(s) $\hat{\theta}_C$,

$$\hat{\theta}_C = ArgMax_{\theta_C} \sum_{t=1}^{T} ln\ C(\hat{U}_1^t, \hat{U}_2^t | \theta_C) \tag{4.7}$$

with U_1^t and U_2^t as the pseudo inverses from (4.5) of the empirical distribution functions from (4.6). For all evaluated copula families the Akaike Information Criteria (AIC)[21] are computed by,

$$AIC := -2 \sum_{t=1}^{T} \ln[C(\hat{U}_1^t, \hat{U}_2^t | \theta_C)] + 2k \tag{4.8}$$

with $k = 1$ for one parameter copulas and $k = 2$ for two parameters. The copula family with minimum AIC value is selected.

Two common copulas; the Gaussian Normal copulas and Student t copulas are selected for analysis of the data in this thesis. They are both elliptical with a Spearman's rho parameter, $\theta_C = \rho$. ρ is a nonparametric measure of statistical dependence between two variables. It is defined as,

$$\rho = \frac{cov(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}} \tag{4.9}$$

with $cov(x_1, x_2)$ as the covariance between the variables and σ_{x_1} and σ_{x_2} as their standard deviations. The Student t copula also has degrees of freedom, df as parameter.

In this thesis we will have F as a 2-dimensional unknown distribution with empirical margins. The Gaussian copula $C_N(u_1, u_2|\rho)$ and the Student t copula $C_{St}(u_1, u_2|\rho, df)$ is set up by (4.4). Examples of 2000 random variables generated by the Gaussian and student t copulas for different parameter values, are presented in Figure 4.2 and 4.3, respectively. It is clear that the closer you get to (0,0) and (1,1) the stronger correlated copulas. The student t copulas produce pseudo observations that appear in a star liked shape. As the df increases(closer to Gaussian copula) this tail dependence gets smaller.

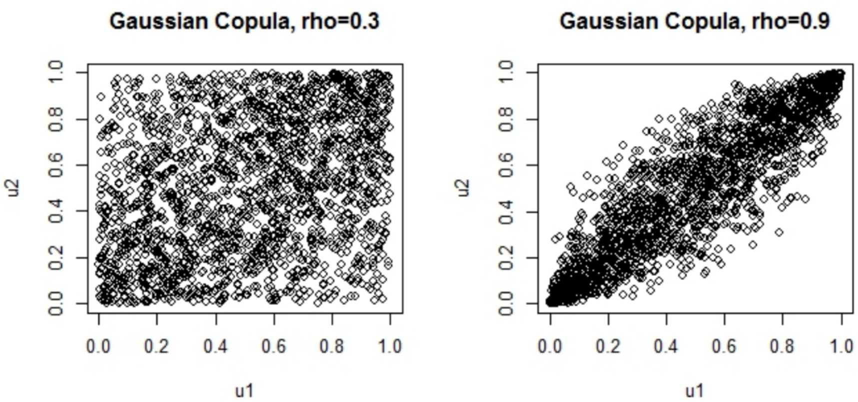

Figure 4.2: Scatter plot of 2000 random samples from the Gaussian copula for $\rho = \{0.3, 0.9\}$. Taken from [22].

4.2.3 Conditional models based on copulas

To model $\hat{U}_j^t \sim \hat{C}(u_j)$ we assume that the Markov property from (4.1) holds. $\hat{U}_j^{t-1}$ is used as input to extract a probability distribution for that previous uniform value associated with the copula $\hat{C}(u_j)$. The conditional distribution function of this probability distribution is then calculated, and by a random generated value on $[0, 1]$ we point to the next uniform value $\hat{U}_j^t$. This process is illustrated in Figure 4.4 and generated for the number of wanted modelled points.

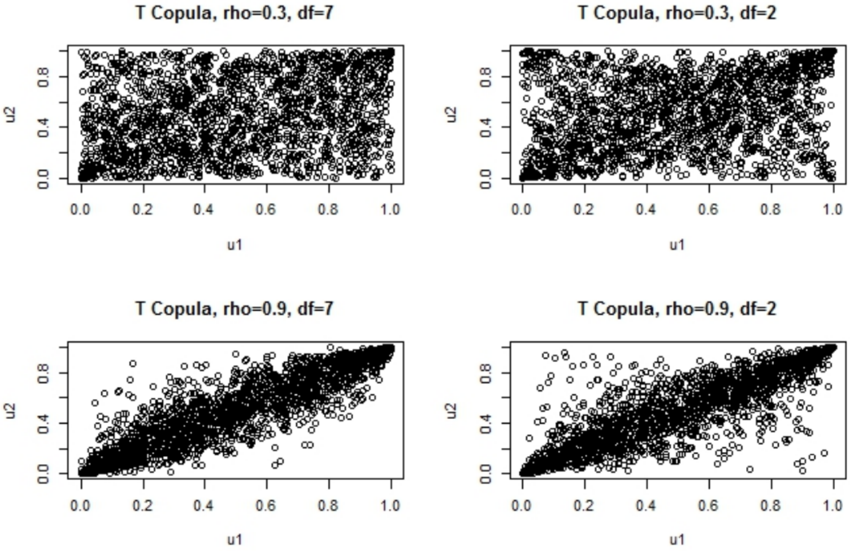

Figure 4.3: Scatter plot of 2000 random samples from the Student t copula for $\rho = \{0.3, 0.9\}$ and $df = \{2, 7\}$. Taken from [22].

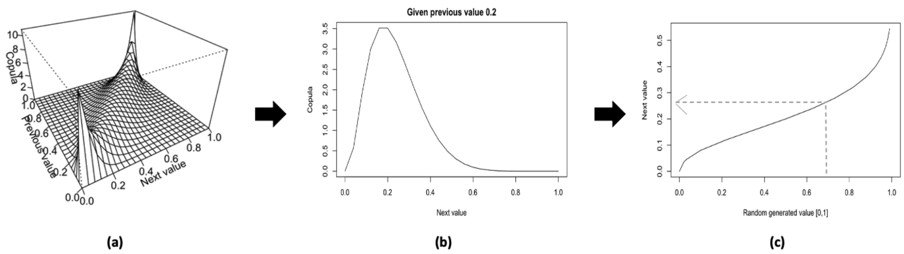

Figure 4.4: An illustration on the modelling process from a copula. (a) The given bivariate copula. (b) The probability distribution function given previous value=0.2. (c) Pointer to the next value after generating a random variable on [0,1].

4.3 Evaluation methods used for decision making in the Koopen use case

In this Section the evaluation methods used for decision making in the Koopen use case are presented.

4.3.1 Brier score

In this thesis optimizing utility will be the ground for decision making. For the accuracy of the probabilistic predictions a cost function is needed. Expected Brier score is such a function, and measures the mean square difference between the expected probability assigned to the possible outcomes at time t and the actual outcomes $o(t)$. It is used for binary predictions and can be denoted as,

$$BS = (p(t) - o(t))^2, \quad (4.10)$$

with $p(t)$ as the expected probability of a prediction. $o(t) = 1$ if the outcome equals the prediction and $o(t) = 0$ if not [12]. The most accurate predictions will be found at low Brier scores, and the higher the BS the less accurate prediction. It classifies as a good cost function because of its properties as strictly defined(here between [0,1]), non-negative($BS \geq 0$) and having no fixed cost($BS(p(t) = 0) = 0$).

4.3.2 Classification error

In this thesis the conditions for classification of noise level indicators with T=15 min are the ranges in Table 2.3.

We define the classification success rate(CSR) as the proportion of cases for which the predicted class $\hat{k}_\xi(X_{t:T})$ equals the true class $k_\xi(x_{t:T})$.

$$CSR_{L_\xi(X_{t:T})} = \frac{\sum_{s=1}^{S} \mathbf{1}(\hat{k}_{\xi,s}(X_{t:T}) = k_\xi(x_{t:T}))}{S}, \quad (4.11)$$

with $\mathbf{1}$ as the indicator function and S as the number of classifications done for the same estimated indicator $L_\xi(X_{t:T})$ to evaluate the classification success rate.

Chapter 5

Methods

In this Chapter we set up conditional copula models in Section 5.1 and a decision framework in Section 5.2 for the Koopen case study presented in Chapter 3. Finally, the use of statistical software is presented in Section 5.3.

5.1 Dependency models for noise

In this Section we introduce how to model noise levels the next 15 minutes, instead of letting the IoT sensors sample them.

The working hypothesis is that we can use the dependency in the last sampled value to model the next quarter. From Figure 3.5 the kernel densities do not form a common known probability distribution, and this calls for further work to be able to model the data set. Copulas are in Section 4.2 presented as a method for representing dependent data with general marginal distributions. This seems like an appropriate choice for modelling of the Koopen data. The data are divided in different data sets d according to Table 3.1.

5.1.1 Conditional noise Model 1

In this thesis we set up two empirical marginals based on observations of noise levels. Each marginal in a marginal pair with the same d has a two second displacement from each other, but are practically equal.

$$X_t^d : \text{ Noise level at time t.}$$

The marginals are transformed to unitary values $U^d(t)$ on $[0,1]$, by the process described in Figure 4.1, where the empirical cumulative distribution $F^d(X^d(t))$ is used.

$$U^d = F_{M1}^d(X^d).$$

To set up the a copula by (4.4) the unitary values, U^d are used. Its parameters θ_C are found by maximum likelihood estimation by (4.7). Assuming the Markov property as presented in (4.1) holds for our data, $U_t^d|U_{t-1}^d$. The copula C_{M1}^d can then be used for conditional modelling of $U_{t:T}^d|u_0^d$ described by the process in Section 4.2.3. Having set the time period T to 15 minutes this translates to 450 conditional modelled U_t^d by,

$$U_t^d|U_{t-1}^d \sim C_{M1}^d(U^d|\theta_C), \tag{5.1}$$

which is straight forward to get Monte Carlo samples from. The translation back to noise level values are done by,

$$X_{t:T}^d = {F_{M1}^d}^{-1}(U^d),$$

to finally achieve the noise level indicators $L_\xi^d(X_{t:T}^d)$ calculated within the time period T, given a previous value X_0^d.

5.1.2 Conditional noise Model 2

Model 1 is based on modelling from observations taken straight from the raw data. The fact that the Koopen data are time dependent, are elaborated in Chapter 3.1. From Figure 3.9 it is clear that the noise level varies with time of week. Therefore, we introduce a model including the weekly reference from (3.1). This model is based on statistical models for the discrepancy between the weekly reference and the observed noise level X_t^d at time t,

$$\text{diff}X_t = X_t - \mu_{L50}^{tow}.$$

$\text{diff}X_t$ are transformed to unitary values $\text{diff}U^d(t)$ by the empirical cumulative distribution for the discrepancy terms $F_{M2}^d(\text{diff}X^d(t))$ by the process illustrated in Figure 4.1,

$$\text{diff}U^d = F_{M2}^d(\text{diff}X_t^d)$$

The copula are set up by (4.4) using these unitary values $\text{diff}U^d(t)$. Its parameters θ_C are found by maximum likelihood estimation by (4.7). Assuming the Markov property as presented in (4.1) holds for our data, $\text{diff}U_t^d|\text{diff}U_{t-1}^d$. The copula, C_{M2}^d based on discrepancies can then be used for conditional modelling of $\text{diff}U_{t:T}^d|\text{diff}U_0^d$, by the process in Section 4.2.3. Having set the time period T to 15 minutes this translates to 450 conditional modelled $\text{diff}U_t^d$ by,

$$\text{diff}U_{t+1}^d|\text{diff}U_t^d \sim C_{M2}^d(\text{diff}U^d|\theta_C) \tag{5.2}$$

which is straight forward to get Monte Carlo samples from. The translation back to noise level values are done by,

$$\text{diff}X_{t:T}^d = {F_{M2}^d}^{-1}(\text{diff}U^d)$$

$$X_{t:T}^d = \text{diff}X_{t:T}^d + \mu_{L50}^{tow}$$

to finally achieve the noise level indicators $L^d_\xi(X^d_{t:T})$ calculated within the time period T, given a previous value X^d_0 and time of week tow.

5.1.3 Conditional copula models for the Koopen use case

An overview of the in total 16 conditional copula models that are used in this thesis are presented in Figure 5.1. There are two conditional noise models, M described in Subsection 5.1.1 and 5.1.1 for $M1$ (based on direct observations) and $M2$ (based on the discrepancy from the observations weekly reference), respectively. These are adjusted based on four different data sets $d \in$ {all, work, no work, busy hour}, defined in Table 3.1. For each of these eight models we then use two different selected copulas c; a Gaussian normal copula(N) and a Student t copula(St) as described in 4.2.2. Each of the estimated conditional copula models have copula parameters θ_C, with $\theta_C = \rho$ and $\theta_C = \{\rho, df\}$ for the Normal and Student t copulas, respectively.

5.1.4 Evaluation of conditional copula models for the Koopen use case

The different conditional copula models $C^d_{M,c}$ are evaluated based on how well they replicate the noise characteristics of the peak noise level, median noise level, background noise level and variability in noise level by the noise level indicators L^d_{10}, L^d_{50}, L^d_{90} and L^d_{10-90}, respectively. The evaluation is done by visually comparing these to the distributions of $L^d_\xi(X^d_{t:T})$ that the different conditional copula models $C^d_{M,c}$ estimates. In addition we look at some QQ-plots for the different conditional noise models and noise level indicators.

5.2 Decision model for the Koopen use case

The decision to be made by the sensor is binary; sample (every other second) for the next 15 minutes, or not. The decision is to be made, and optimized, based on a noise level observation at present, and our chosen conditional copula model. Independent on the decision, we will therefore always sample one's each 15 minutes. Based on the results in Section 6.2 we use conditional noise Model 2 and a normal copula based on the whole data set with d =all as the basis for all decision making described in this thesis. In addition to the present sampled noise level observation x_0, this model also uses knowledge about time of week, tow.

5.2.1 Utility function for the Koopen use case

To conclude which action optimizes the energy-accuracy tradeoff between energy consumption and modelling accuracy, we need to set up a utility function. In this

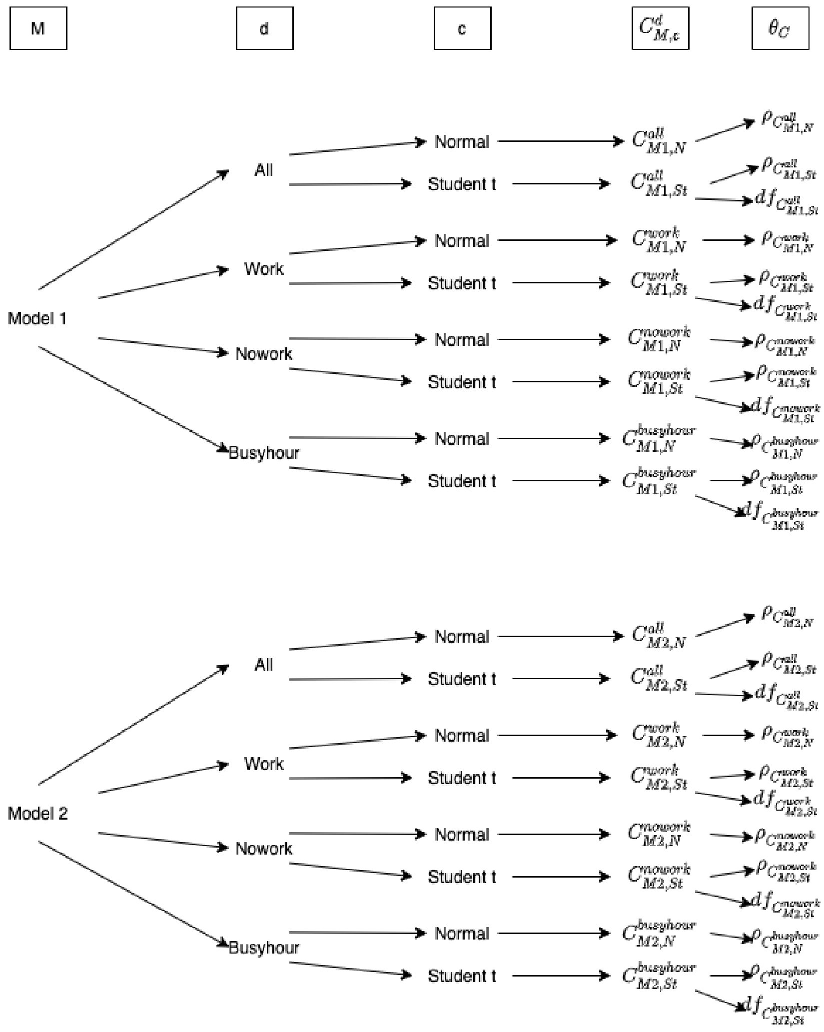

Figure 5.1: An overview of the 16 different conditional copula models $C^{d}_{M,c}$ used in this thesis.

thesis we address the decision problem as an optimization problem were we maximize the utility based on choice of action $a = \{\text{sample, no sample}\}$. The utility function for given a can then be denoted as,

$$U_a(t) = -COST(t) = -(E + A(t)), \tag{5.3}$$

for the Koopen use case. Here $COST$ denotes the total cost, E the cost of energy for sampling 15 minutes and A the cost in accuracy loss.

The expected loss in energy when sampling is given by the use case. To be able to evaluate and discuss how and why different costs in energy affects the decision problem used in this thesis we let E take four different values. The expected loss as cost in energy is a binary operator that takes the value $E = 0$ for a=no sample and $E \in \{0.005, 0.05, 0.15\}$ for a=sample.

$A(t)$ is also dependent on use case and the time t that sets the previous sampled value x_0 and the time of week, tow. In this work we have chosen Brier score from (4.10) as basis for our accuracy cost function. The expected loss in accuracy when we choose not to sample can therefore be denoted as,

$$A(t) = (p(t) - o(t))^2, \tag{5.4}$$

and $A(t) = 0$ for a=sample. The expected value $p(t)$ is estimated using the classification success rate, CSR defined by (4.11). The expected value $L_\xi^d(X_{t:T})$ of the noise level indicators $L_\xi^d(x_{t:T})$ are estimated using Monte Carlo simulation for $s = 1, 2, \ldots, S$ simulations as described in Subsection 5.1.2. For the Koopen use case we have three classes $k \in$ {good, fair, poor} for which these noise level indicators can classify within. Examination of the simulated data shows us that each estimated $L_\xi^d(X_{t:T})$ at time point t only differs between two of the classes. Therefore, we may use the binary cost function Brier score. The classification in CSR is done deterministically, meaning we always use the estimated class $\hat{k}_\xi(X_{t:T})$ with highest probability, as true class $k_\xi(x_{t:T})$. Then we for a = no sample have,

$$A(t) = (CSR_{L_\xi(X_{t:T})} - o(t))^2 = \left(\frac{\sum_{s=1}^{S} \mathbf{1}(\hat{k}_{\xi,s}(X_{t:T}) = k_\xi(x_{t:T}))}{S} - 1\right)^2. \tag{5.5}$$

The observation $o(t)$ is always set to 1, according to the deterministic classification assuming we observe the expected observation. This leads to $A(t) \in [0, 0.25]$, independent on choice of conditional copula model.

In total the utility is,

$$U_a(t) = \begin{cases} -E & \text{if } a = \text{sample} \\ -A(t) & \text{if } a = \text{no sample}, \end{cases} \tag{5.6}$$

which makes us able to choose an action based on which maximizes the utility. The decision is a =sample if $E < A(t)$ and a =no sample otherwise.

5.2.2 Decision network formulation

Another framework for presenting the decision problem is based on common artificial intelligence methodologies as presented by Russel and Stuart(2016) [23]. An influence diagram is used to clarify the different parts of the decision network and how they interact with each other to form a decision. For the Koopen use case the influence diagram is presented in Figure 5.2. The known features previous value and time of week, as well as the random variables energy cost and modelling accuracy are ovals and represent the chance nodes. The decision makers choice(to sample or not) is drawn as a rectangle and represent the decision node. Finally, the diamond is the utility function and represent a utility node.

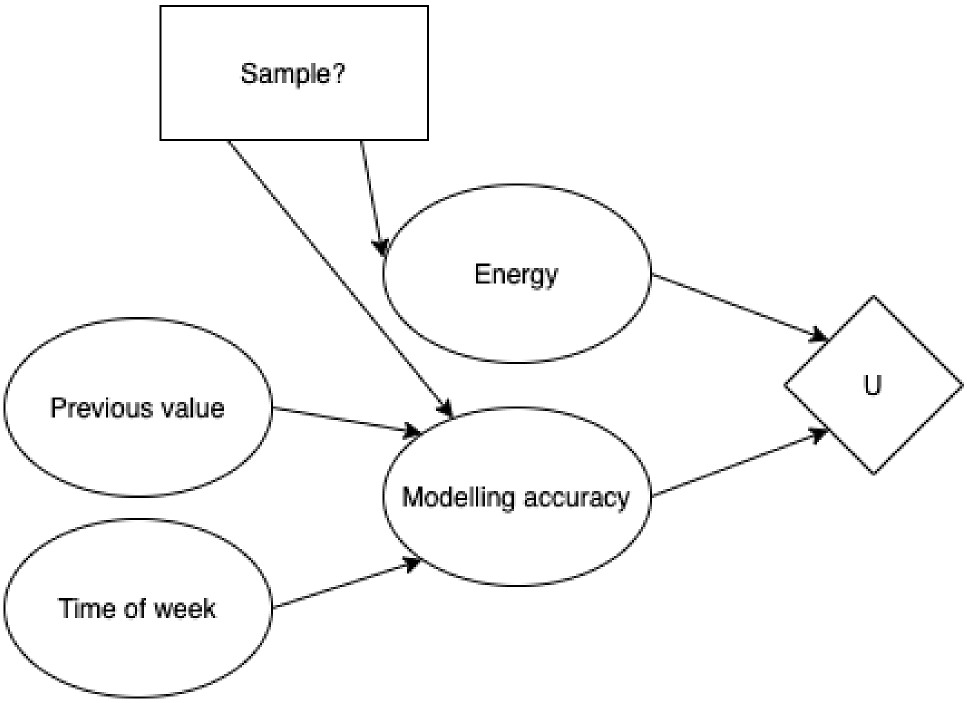

Figure 5.2: Influence diagram for the decision used in this thesis on whether to sample with an IoT sensor or model the next 15 minutes.

5.3 Statistical software

R is used for the analysis in this thesis. The data were loaded from [24] and transformed into data frames by the work of my project thesis [25]. The library 'copula' [26] and 'VineCopula' [27] were used to select and set up the Gaussian and Student t copulas. The code for all analysis done in this thesis can be found at Github [28].

Chapter 6

Results

In this Chapter the results are presented and discussed. The results regarding choice of copulas are presented in Section 6.1. The results from all conditional copula models are presented and compared in Section 6.2. In Section 6.3 the results for sampling decisions are presented and evaluated for different choices of expected loss in energy.

6.1 Fitting the dependency models

The margins for different data sets d are by the process described by Figure 4.1 in Section 4.2 transformed to unitary values, U on $[0,1]$. The scatterplots of these unitary values can be seen in Figure 6.1 and 6.2 for conditional noise Model 1 and 2, respectively.

Looking at u_1 and u_2 in Figure 6.1 and 6.2, we observe a clear tendency for values on $u_1 = U2$ and $\text{diff}u_1 = \text{diff}u_2$, as well as a representation through the whole unitary space. u_1 and u_2 as well as $\text{diff}u_1$ and $\text{diff}u_2$ seems symmetric for all d except busy hour. To investigate the impact choice of copula has to our modelling results, we have chosen to use different elliptical copulas; Gaussian Normal and Student t in this thesis.

In Section 4.2.2 methods for selecting an appropriate copula based on AIC from (4.8) are presented. For all models we get the bivariate Student t copula as best fit compared to the Normal copula.

Visually by comparing the tendencies in Figure 6.1 and 6.2 to Figure 4.2 and 4.3 we would expect quite high Spearman's rho parameters, ρ from (4.9) and degrees of freedom, $df < 7$ for the Student t copulas.

To fit Student t and Normal copulas to the different conditional copula models, the maximum likelihood estimates are found as described in Section 4.2.2 by (4.7). All parameters for the conditional bivariate copula models set up in Subsection 4.2.3 are presented in Table 6.1.

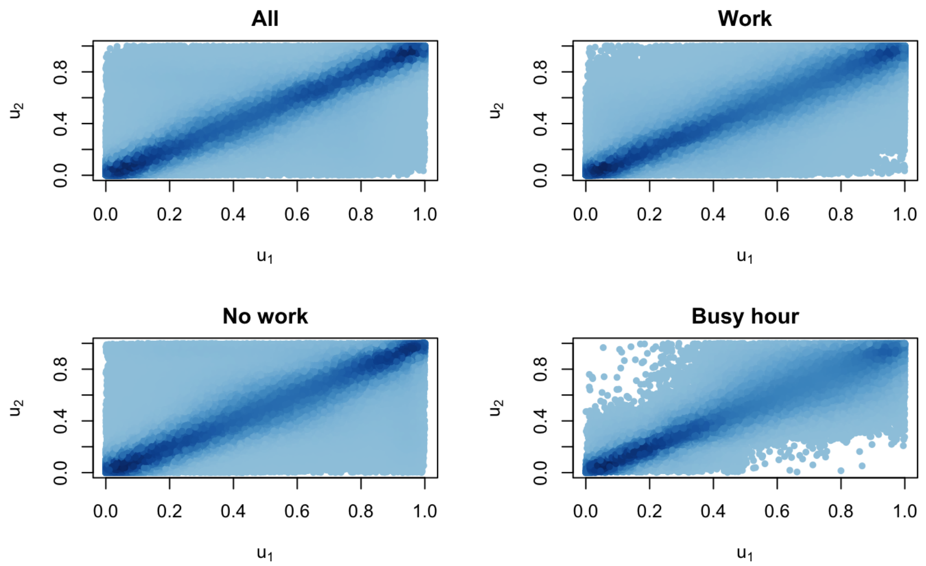

Figure 6.1: Scatterplot using the conditional noise Model 1 of the marginals unitary transformations for all data sets.

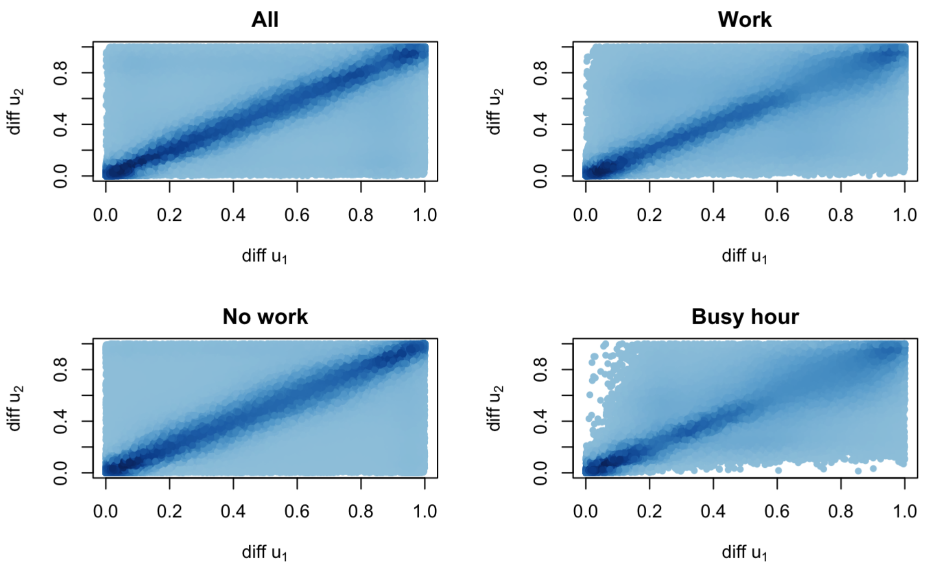

Figure 6.2: Scatterplot using the conditional noise Model 2 of the diff marginals unitary transformations for all data sets.

Examples of a Normal and Student t copula based on the Koopen data for conditional noise model 1 and 2 is presented in Figure 6.3 and 6.4, respectively.

	All	**Work**	**No work**	**Busy hour**
Model 1, Normal : $X_{t+1}\|x_t \sim C^d_{M1,N}$	$\rho = 0.92$	$\rho = 0.89$	$\rho = 0.88$	$\rho = 0.87$
Model 1, Student t : $X_{t+1}\|x_t \sim C^d_{M1,St}$	$\rho = 0.94$ $df = 2.9$	$\rho = 0.90$ $df = 3.8$	$\rho = 0.91$ $df = 2.5$	$\rho = 0.89$ $df = 3.8$
Model 2, Normal : $\text{diff}X_{t+1}\|\text{diff}x_t \sim C^d_{M2,N}$	$\rho = 0.81$	$\rho = 0.78$	$\rho = 0.87$	$\rho = 0.81$
Model 2, Student t : $\text{diff}X_{t+1}\|\text{diff}x_t \sim C^d_{M2,St}$	$\rho = 0.92$ $df = 2.0$	$\rho = 0.82$ $df = 2.0$	$\rho = 0.93$ $df = 2.0$	$\rho = 0.83$ $df = 2.3$

Table 6.1: Conditional bivrariate Normal and Student t copula models used in this thesis with selected best fit parameters ρ and df = degrees of freedom.

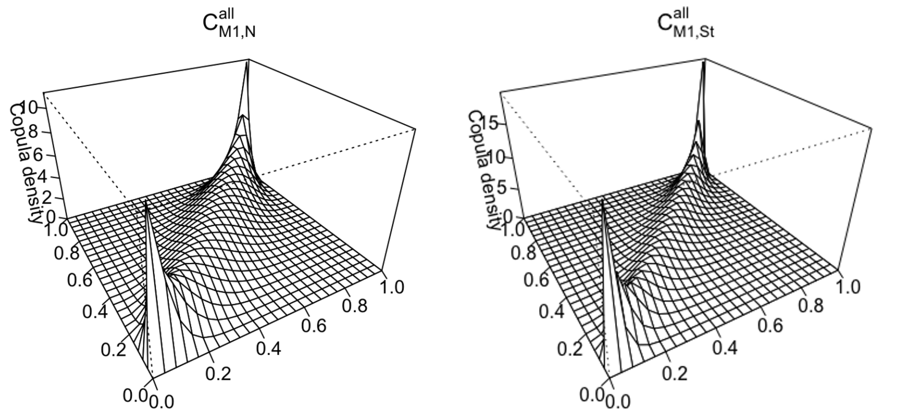

Figure 6.3: Bivariate Normal copula and Student t copula with parameters $\rho = 0.92$ and $\{\rho = 0.94, df = 2.9\}$, respectively for conditional model 1.

6.1.1 Simulation of noise level data

Examples of modelling 15 minutes by the use of both a Normal and Student t copula are presented in Figure 6.5 and 6.6 for conditional noise model 1 and 2, respectively. All are given the previous sampled value $x_0 = 45.9$ from the raw data 20.02.2019 at tow =Wednesday 09:00. To get from a previous value to the next value the process illustrated by Figure 4.4 is used. From the modelled quarters there are no immediate difference between the choices of copula. For the different choice of conditional noise model we can see that Model 2 tends to have some lower noise level values, as well as the largest amount of high noise level values.

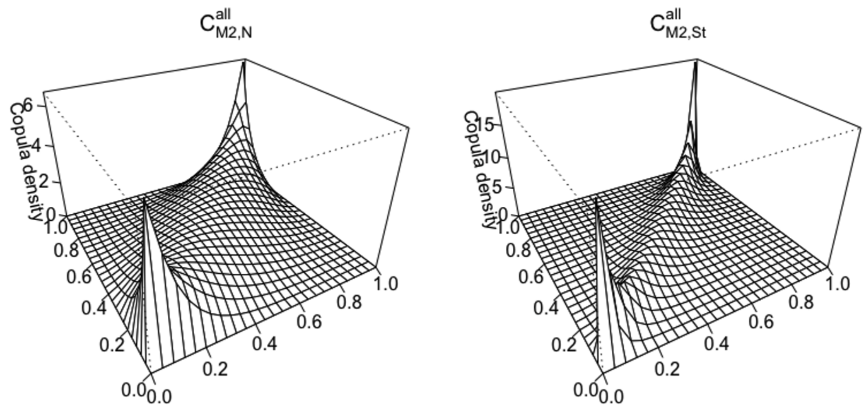

Figure 6.4: Bivariate Normal copula and Student t copula with parameters $\rho = 0.81$ and $\{\rho = 0.92, df = 2.0\}$, respectively for conditional model 2.

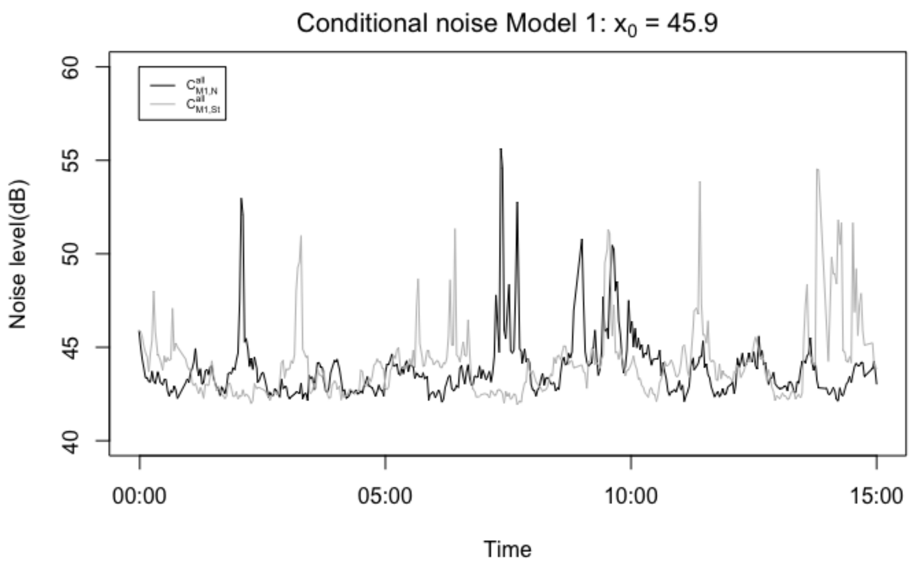

Figure 6.5: A random modelled quarter using conditional noise Model 1 for both the Student t and Normal copula for d =all. Given previous value $x_0 = 45.9$.

6.2 Model evaluation for noise level indicators.

In this Section the results from our dependency modelling using copulas will be presented. The 16 different conditional copula models used, with their respective parameters are shown in Table 6.1. The comparison in this Section is done based on type of copula c, data set d and choice of conditional model M for the estimated noise level indicators L^d_ξ as presented in Section 5.1.4.

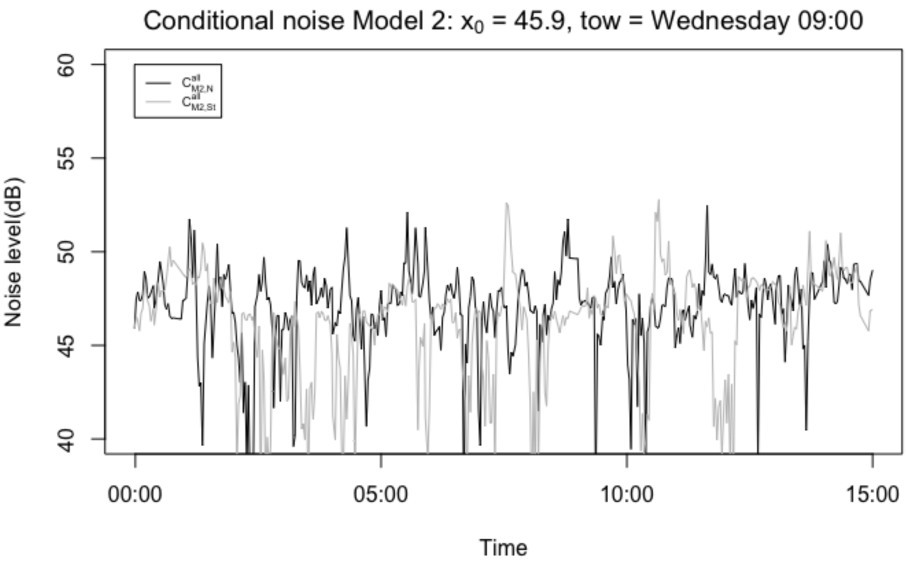

Figure 6.6: A random modelled quarter using conditional noise Model 2 for both the Student t and Normal copula for d =all. Given previous value $x_0 = 45.9$ and $tow =$ Wednesday 09:00.

The simulation of modelled noise level data for one quarter are presented in Subsection 6.1.1. This quarter modelling process is then repeated 1000 times for each conditional copula model $C^d_{M,c}$ with random input values x_0 and tow to get a representation of the total model performance. The densities of the noise level indicators, L^d_ξ for all 16 models from Table 6.1 are compared to the ones for the raw data of the data sets all, work, no work and busy hour in Figure 6.7, 6.8, 6.9 and 6.10, respectively. Here we look at each noise level indicator separately.

The figures show that the choice of copula family seems to have the least impact on the modelling results as their densities always follow each other. For especially d =all in Figure 6.7 the normal copula does seem to capture the noise level indicators closer to their original values based on raw data. Since the noise level indicators are what we want our conditional copula model to replicate, we will in further analysis choose the normal copula for the decision making.

From Figure 6.10, it is clear that none of the models are able to replicate the behaviour of $d =$ busy hour for the raw data. Results for conditional noise Model 1 is really bad, and Model 2 just a little bit better. The choices of copulas may, by looking at the scatterplots in Figure 6.1 and 6.2, be the wrong ones. Using elliptical copulas we are assuming symmetric unitary values U, which is not the case for $d =$ busy hour.

For $d = \{$all, work, no work$\}$ there are no d that stands out as a better fit for the models of the Koopen use case. For that reason the data set $d =$ all is chosen,

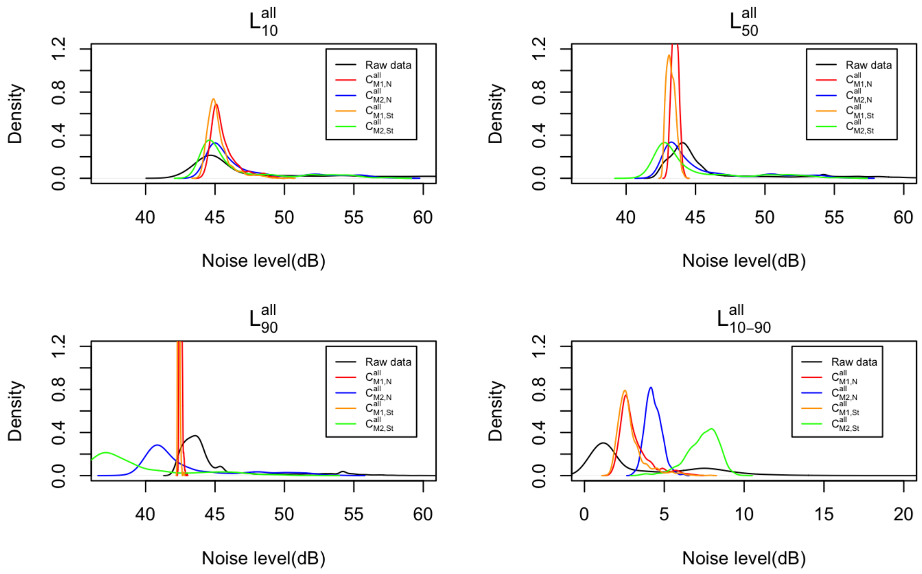

Figure 6.7: Model comparison for different conditional copula models of the data set d =all for all noise level indicators.

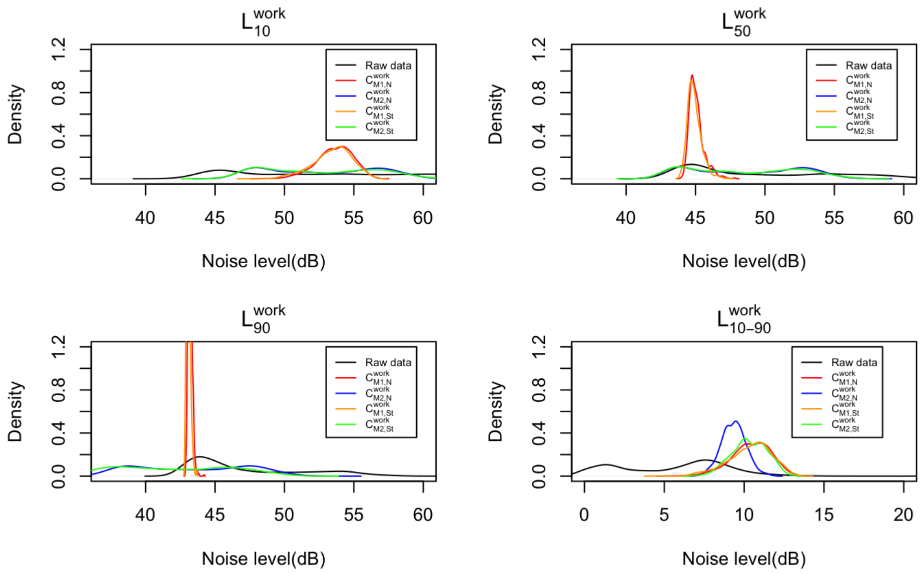

Figure 6.8: Model comparison for different conditional copula models of the data set d =work for all noise level indicators.

since it is convenient that it models all time points, and the IoT sensors will be in no need of switching between models for time points referring to d = work and d = no work.

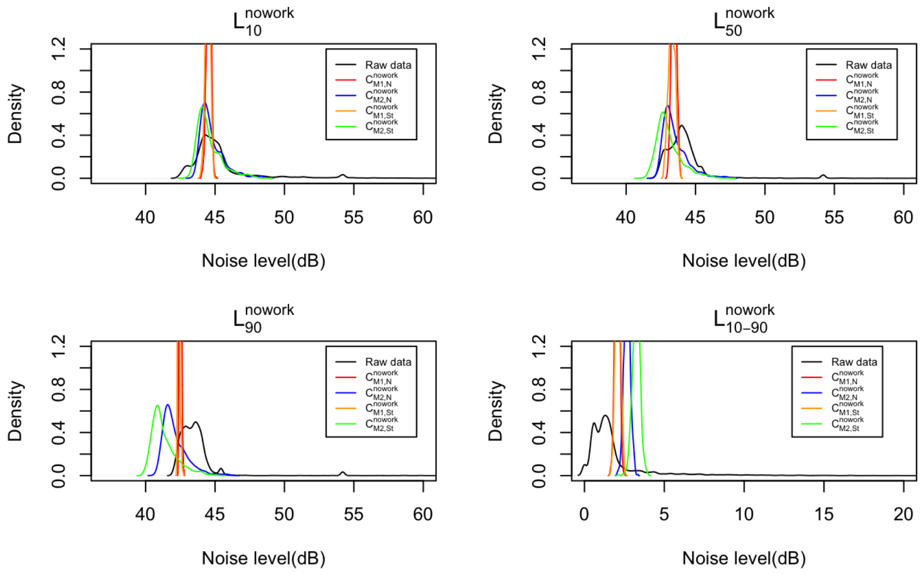

Figure 6.9: Model comparison for different conditional copula models of the data set d =no work for all noise level indicators.

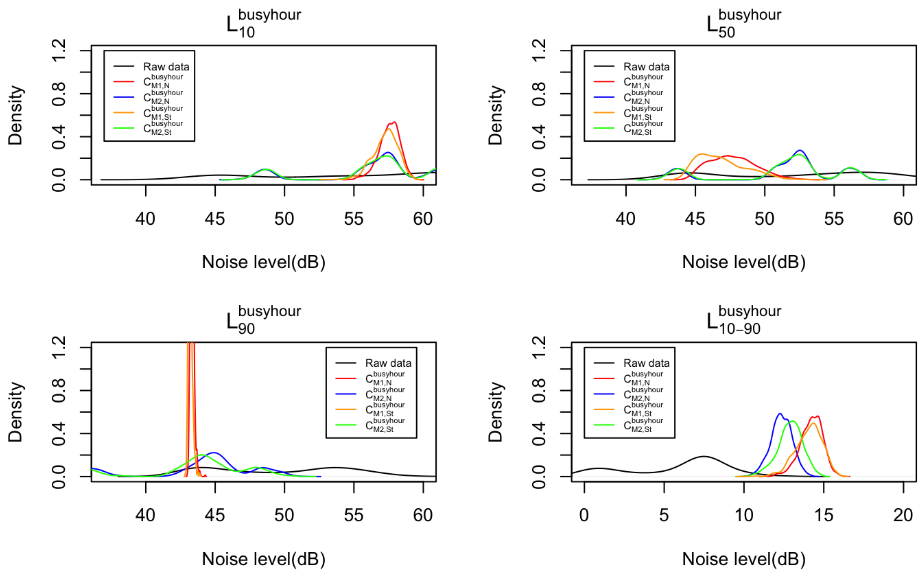

Figure 6.10: Model comparison for different conditional copula models of the data set d =busy hour for all noise level indicators.

Having the 16 conditional copula models narrowed down to only two; $C^{all}_{M1,N}$ and $C^{all}_{M2,N}$ for conditional noise Model 1 and 2, respectively, their quantile-quantile plots for different noise level indicators L^{all}_{ξ} are presented in Figure 6.11. The reference to ideal performance is plotted as a red line.

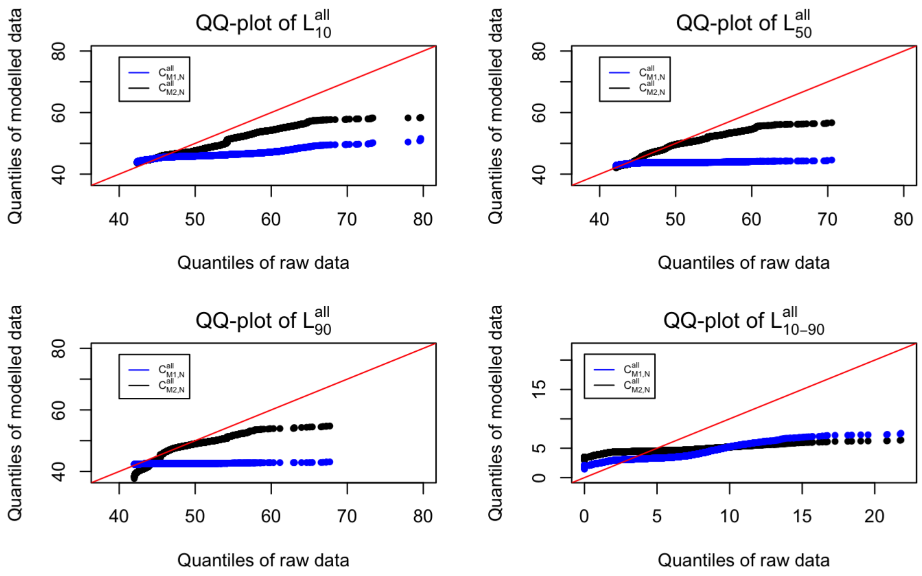

Figure 6.11: QQ-plots for all indicators of the modelled data compared to the raw data for conditional noise Model 1 and 2, copula, c = Normal and data set, d = all. The red line is plotted as reference to ideal behaviour.

Figure 6.7, 6.8, 6.9 and 6.10 all show that conditional noise Model 1 tends to be narrow for all indicators and do not represent a wide enough selection of noise levels for the indicators compared to the raw data. From Figure 6.11 it is also clear that conditional noise Model 2 is the best fit for the Koopen use case. Hence, it will be the preferred choice.

It is clear from both the density plots in Figure 6.7, 6.8, 6.9, 6.10 and the QQ-plot in Figure 6.11 that our conditional copula models do not perform well for L^{all}_{10-90}. If the variance in noise level are important for the monitoring of the Koopen use case, conditional copula models of the data can as a result be rejected as approach for down sampling strategy.

Figure 6.6 of a modelled quarter for conditional noise Model 2 reveal some tendencies to low noise level values that are not monitored by the IoT sensors in the environment we are looking at. This makes especially the background noise level indicator L^{all}_{90} hard to replicate. The density plots in Figure 6.7, 6.8, 6.9 and 6.10 confirms this, and reveal that too low modelled values are furthermore a problem for conditional noise Model 1 as well. As a result, one should not use the conditional copula models presented in this thesis to model the background noise level of the Koopen use case.

According to Figure 6.7 and 6.11, the preferred choice of conditional copula model $C^{all}_{M2,N}$, fit both L^{all}_{10} and L^{all}_{50} quite well. This is especially the case for noise level values below 60 dB, that there are most of in the raw data. If the peaks in noise level and median noise level are of interest in the Koopen use case, one can reduce the sampling rate for the IoT sensors by using $C^{all}_{M2,N}$ for modelling parts of the data. Which parts are further evaluated in Section 6.3.

6.3 Results for sampling decisions

In this Section the result from the decision making are presented. The copula used is $C^{all}_{M2,N}$ as concluded with in Section 6.2. In Section 5.2 the decision model is described. The goal is to find out whether or not to sample the next 15 minutes, given a sampled previous noise level value x_0 and a time of week tow. The cost in energy E is also given as input. To evaluate the different factors in our decision model, the results for different time of week, previous sampled value and choice of expected loss in energy are in Section 6.3.1, 6.3.2, 6.3.3, respectively presented.

Based on the results in Section 6.2 our models do not represent L_{90} and L_{10-90} well. Further, knowing from the utility function (5.6) that the decision making is partly based on the classification success rate, L_{90} may need some adjustments because it never classifies as good, confirmed by Figure 3.8. In Section 5.2.1 we assumed that the largest partition of classes for S simulations of the noise level indicators would classify correctly. To be sure this holds for our decision model, neither L^{all}_{90} nor L^{all}_{10-90} are good choices. Hence, L^{all}_{10} and L^{all}_{50} are the ones being used for decision making for the Koopen use case.

6.3.1 Decision based on different times of week

In this Subsection the previous values are given by the sampled noise levels at each tow from a week of raw data 16.02.2019 to 23.02.2019. The decision for each time of week interval and different E are presented in Figure 6.12 and 6.13 for L^{all}_{10} and L^{all}_{50}, respectively. Grey spaces visualize a =sample and white ones a =no sample. As one would expect, the number of recommended samples by the IoT sensors increases when the expected loss in energy E decreases, as we can see in Figure 6.12 and 6.13.

For the peak noise level L^{all}_{10} in Figure 6.12 the times of week that the decision model tells us to sample, are always within work hours. The Koopen data is based on noise levels in a working environment, and therefore it makes sense that these are most uncertain. When there are people present, the noise levels will change rapidly according to type of work, and how load the people are. It is reasonable to sample some of these times within work hours, in between the set rate of at least one sample each 15 minutes.

For the median noise level L^{all}_{50} in Figure 6.13 there are more times of week that the decision model tells us to sample than for L^{all}_{10}. Here we are interested in the median observed noise, and as we can see the conditional copula model become unsure at by example Friday nights. This indicates that the environment may be used different at this time outside working hours for different weeks, which do make sense. Then it would be a good idea to sample the actual noise levels at those times, as our decision model tells us.

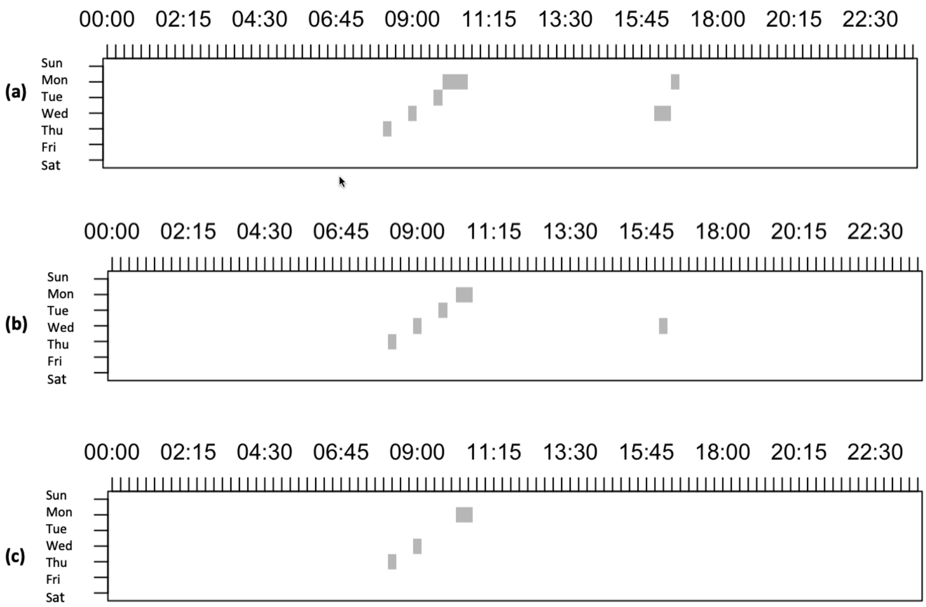

Figure 6.12: Binary decision on whether to sample(grey) or not(white) throughout a week based on the conditional copula model $C_{M2,N}^{all}$ and L_{10}^{all}. Previous sampled values are taken from the raw data at given time of week 16.02.2019 to 23.02.2019. (a) E=0.005 (b) E=0.05 (c) E=0.15.

6.3.2 Decision based on different sampled previous value

In this section we choose to look at three selected times of week, tow for each noise level indicator to get a spectre on how previous sampled value can impact the decision. The previous values are chosen to be $x_0 \in \{40, 45, 60, 80, 90\}$, where both $x_0 = 40$ and $x_0 = 90$ are values out of scope from the given Koopen raw data x_t. The three other $x_0 = 45$, $x_0 = 60$ and $x_0 = 80$ are respectively common, common within work hours and rare.

First we set $tow = \{$Tuesday 08:00, Tuesday 10:00, Wednesday 09:00$\}$ for L_{10}^{all}, and calculate the cost in accuracy $A(t)$ for different previous values. These, as well as all E are presented in Figure 6.14.

For tow =Tuesday 10:00 we see that the decision differs a little for given previous value x_0 =42 dB. $E = 0.005$ lay in between different $A(t)$ and therefore the decision is not independent on previous value for this E. For $tow =$ Wednesday 09:00 in Figure 6.14, A(t) differs more for given previous values. It is almost always quite high, but also here we see that one given previous value x_0 =90 dB could change the decision, here if $E = 0.15$. Figure 6.12 shows that the decision model quite rarely decides on sampling for L_{10}^{all}. If this was only due to the choice of previous value, one should be quite aware of it. Although L_{10}^{all} always categorize with A(t) as in Figure 6.14 where $tow =$ Tuesday 08:00 when the decision is clear

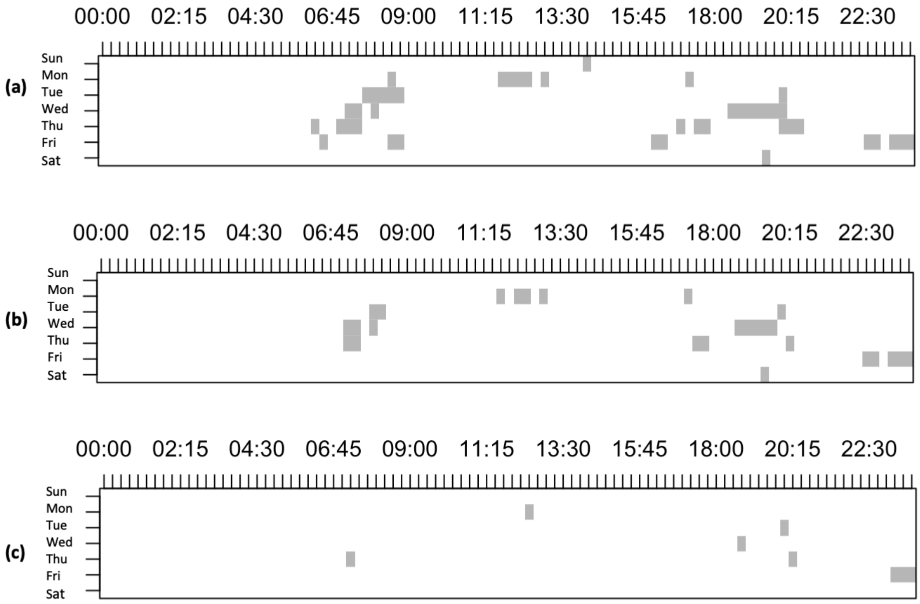

Figure 6.13: Binary decision on whether to sample(grey) or not(white) throughout a week based on the conditional copula model $C^{all}_{M2,N}$ and L^{all}_{50}. Previous sampled values are taken from the raw data at given time of week 16.02.2019 to 23.02.2019. (a) E=0.005 (b) E=0.05 (c) E=0.15.

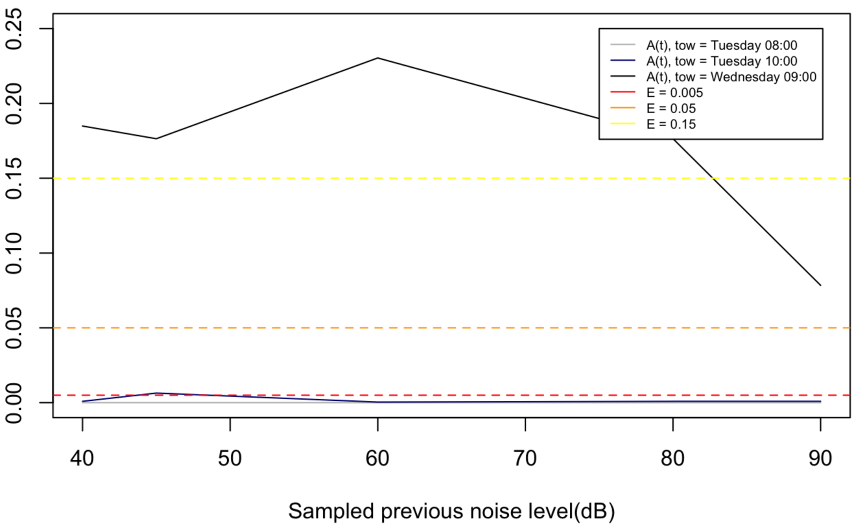

Figure 6.14: The loss in accuracy, $A(t)$ for the conditional copula model $C^{all}_{M2,N}$ and L^{all}_{10}, with different sampled previous noise levels. The loss in energy, E are plotted as dotted lines for comparison.

$E < 0.005$. If Figure 6.12 concludes on no sample for some previous value, the decision is as well independent on previous sampled value in our decision model.

For L_{50}^{all}, we use $tow = \{$Tuesday 08:00, Wednesday 09:00, Friday 23:30$\}$ and calculate the cost in accuracy $A(t)$ for different previous values. These, as well as all E are presented in Figure 6.15.

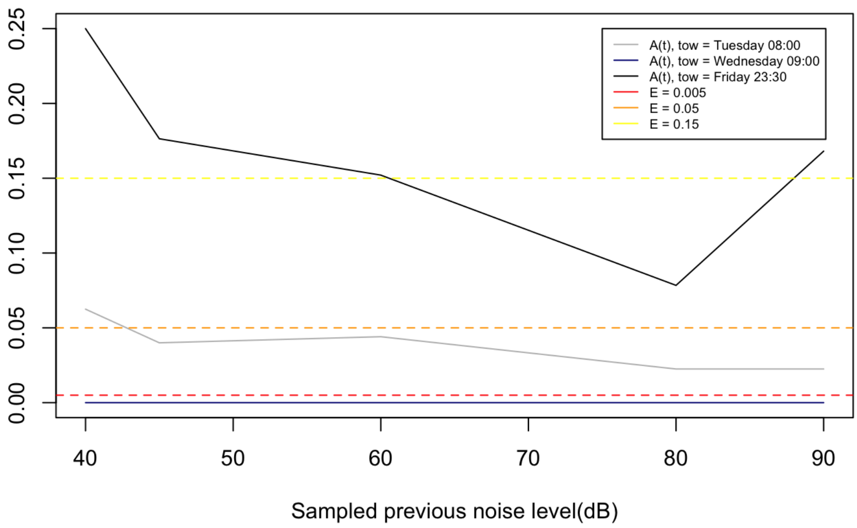

Figure 6.15: The loss in accuracy, $A(t)$ for the conditional copula model $C_{M2,N}^{all}$ and L_{50}^{all}, with different sampled previous noise levels. The loss in energy, E are plotted as dotted lines for comparison.

Here the example of being sure on decision independent on previous sampled noise level is plotted for $tow =$ Wednesday 09:00, as can also be seen by Figure 6.13 is a=no sample for all E. For the other choices of tow in Figure 6.15 the decision is not that clear. For $tow =$ Tuesday 08:00 and $tow =$ Friday 23:30 the decision is dependent on previous noise level for $E = 0.05$ and $E = 0.15$, respectively.

For most times of week, we see that the decision model chooses to not sample as seen in Figure 6.12 and 6.13. It is also shown by examples in Figure 6.14 and 6.15 that this is independent on previous sampled value. Based on this fact the conditional copula model is clearly quite forgetful, an will not catch extreme behaviour. A way to incorporate this complication for the decision making in this thesis, is to always sample for extreme previous sampled values independent on E and $A(t)$. A proper choice for extreme values for the Koopen use case, is noise level values below 42 dB and above 75 dB, which we see from the raw data in Figure 3.3 is outside scope and rare, respectively.

6.3.3 Evaluation on choice of expected loss in energy

The utility function in Section 5.2.1 is based on a weighting of whether the expected loss in energy or loss in accuracy is largest. Therefore the choice of cost in energy E is really important for the decision making. The loss in accuracy is defined in such a way that it exists between 0 and 0.25. The choice of E has from Section 6.3.1 and 6.3.2 shown to impact the decision. Deciding on a E that fits the data and use case is important. The conditional copula model $C_{M2,N}^{all}$ for the Koopen data has shown to be able to reduce the need of sampled quarters considerably by Figure 6.12 and 6.13. From these Figures $E = 0.05$ seems like an appropriate choice of E to contribute to increase the IoT sensors lifetime by down sampling, as well as ensuring model accuracy. $E = 0.05$ translates to a need of over 77% of the noise level indicators classified equally. The number of samples needed throughout a week with the decision model defined in this thesis for $C_{M2,N}^{all}$ with $E = 0.05$, can be reduced to 1.1% and 5.8% wanting L_{10}^{all} and L_{50}^{all}, respectively. This shows a huge possibility for down sampling by the use of conditional copula models as presented in this thesis when you are interested in the peak noise level, and median noise level.

Chapter 7

Conclusion

In this thesis we consider sampling decisions for energy efficient IoT sensors using dependency models. Down sampling makes sensors more efficient and reduces the need of both power and communication, which further contributes to increase their lifetime. The goal is to balance use of energy and accuracy when monitoring noise in the working environment Koopen.

To investigate the environment's noise characteristics, we set up noise level indicators. This is done by using quantile theory in practice. Quantiles are in need of more than just one measure to be set. This challenges the prospect of being able to down sample a lot, when we simultaneous want to explore the noise level indicators within short time periods. If one were able to model these points, instead of sampling them by the IoT sensors, the monitoring would be more energy efficient.

The IoT sensors are set up to sample once each 15 minutes. Dependent on an energy-accuracy tradeoff, the decision model in this thesis chooses between letting the IoT sensors sample within the quarter, and letting a conditional copula model based on dependency in given data, model the quarter of noise levels.

To compare the suggested models with respect to copula families, Akaike Information criteria is used. Their parameters are found by maximum likelihood estimation. In this thesis we use Student t and a Gaussian Normal copula based on our marginals of time dependent noise level data from Koopen. The Normal copula fit the use case the best. Anyway, the modelling results show us that the choice of copula itself, does not have profound impact on how well the conditional copula models perform.

Further we examine if partitioning the Koopen use case data into data sets of all data, only work hours, only outside work hours or a chosen busy hour, makes it easier to be replicated by the dependency models. This shows no large effect, except that the conditional copula models for the busy hour do not fit the data.

The two different types of conditional modelling introduced as conditional noise Model 1 and 2, turns out to have the greatest impact on model performance in this thesis. Model 2 based on discrepancies, instead of observations taken straight from the raw data, seems to be able to replicate the noise level characteristics quite well. This applies especially on higher noise levels, where Model 1 do not even reach up.

The results of conditional copula modelling for the variance in noise level, and the background noise level for the Koopen data, are still shown to be of poor quality and uncertain for some parts of the data, respectively. That being so, these are not recommended modelled by conditional copula models as presented in this thesis.

Knowing noise can be disturbing, annoying and in the worst case dangerous, we are interested in how the different noise level indicators classifies given known standards.

Because it has shown to be the best fit for the Koopen data, we use a conditional Normal copula model based on all the discrepancies from a weekly reference in the data for our decision model. The decision analysis are based on a tradeoff between energy and accuracy for the noise monitoring. To optimize this tradeoff, a binary utility function is set up. The loss in energy is constant for the case of sampling, and the loss in accuracy is based on classification by an adapted Brier score when modelling.

The decision model shows that the sampling rate can be reduced to once every 15 minutes for most part of a week, dependent on which noise level indicator you are interested in. Being interested in the peak noise level, and median noise level, respectively, the IoT sensor samples from one week, can be reduced to 1.1% and 5.8%, respectively for an energy loss set to 0.05. This down sampling will contribute to increase the lifetime of the IoT sensors.

The behaviour of our models for extreme previous values are shown to be insufficient. That being so, independent of the energy-accuracy tradeoff in the utility function for extreme previous values with noise levels below 42dB and above 75dB for the Koopen data, it is recommended to always sample, to catch the noise characteristics for such abnormal behaviour.

This project clearly shows the possibilities for down sampling by the use of conditional copula models. In further analyses it would be of interest to fit models that can capture the background noise level and variance in noise level. Adapting the conditional copula models to be intelligent by learning from new samples could be of interest to make them better, and more robust. The use case could also benefit from investigation of a function for the actual loss in energy.

Bibliography

[1] CDC. (2019). 'What noises cause hearing loss?' [Online]. Available: https://www.cdc.gov/nceh/hearing_loss/what_noises_cause_hearing_loss.html. (accessed: 10.05.2021).

[2] J. L. Mayes, 'Urban noise levels are high enough to damage auditory sensorineural health.,' *Cities and Health (2019)*, 2019.

[3] WHO, 'Noise.,' *Environmental health criteria; 12, 1980*, 1980.

[4] S. Rose. (2019). 'Top 14 iot trends to expect in 2020!' [Online]. Available: https://towardsdatascience.com/top-14-iot-trends-to-expect-in-2020-fa81a56e8653. (accessed: 10.05.2021).

[5] F. A. Kraemer, F. Alawad and I. M. V. Bosch, 'Energy-accuracy tradeoff for efficient noise monitoring and prediction in working environments,' *IoT 2019: Proceedings of the 9th International Conference on the Internet of Things, October 2019*, 2019.

[6] L. Brocolini, C. Lavandier, M. Quoy and C. Ribeiro, *Measurements of acoustic environments for urban soundscapes: Choice of homogeneous periods, optimization of durations, and selection of indicators*. The Journal of the Acoustical Society of America, 2013.

[7] I. M. V. Bosch, 'Efficient noise measurment with energy constrained iot nodes: A case study on working environment quality,' *Master thesis, Norwegian University of Science and Technology*, 2019.

[8] A. Murad, F. A. Kraemer, K. Bach and G. Taylor, 'Information-driven adaptive sensing based on deep reinforcement learning,' *10th International Conference on the Internet of Things (IoT20), October 2020*, 2020.

[9] D. Schindler and C. Jung, 'Copula-based estimation of directional wind energy yield: A case study from germany,' *Energy Conversion and Management, May 2018*, 2018.

[10] F. Li and Q. Zheng, 'Probabilistic modelling of flood events using the entropy copula,' *Advances in Water Resources, November 2019*, 2016.

[11] M. I. Bhatti and H. Q. Do, 'Recent development in copula and its applicationsto the energy, forestry and environmental sciences,' *International journal of hydrogen energy (2019)*, 2019.

[12] A. Stuart and K. Ord, *Kendall's Advanced Theory of Statistics: 1: Distribution theory*. Edward Arnold, 1994.

[13] R. J. Hyndman and Y. Fan, 'Sample quantiles in statistical packages,' *The American Statistician, November 1996, Vol. 50, No. 4*, 1996.

[14] E. H. Berger, *The Noise Manual*. Aiha, 2003.

[15] K. Furberg. (2016). 'Dette kan bli en modell for ntnus nye campusløsning,' [Online]. Available: `https://www.universitetsavisa.no/campus/dette-kan-bli-en-modell-for-ntnus-nye-campuslosning/163139`. (accessed: 22.05.2021).

[16] Libelium. (2020). 'Iot products plug&sense!' [Online]. Available: `https://www.libelium.com/iot-products/plug-sense/`. (accessed: 10.05.2021).

[17] NummeropplysningenAS. (2019). 'Kalender 2019,' [Online]. Available: `https://www.norskkalender.no/default.asp?year=2019`. (accessed: 10.05.2021).

[18] M. Rausand and A. Høyland, *System Reliability Theory Models, Statistical Methods, and Applications*. Wiley, 2004.

[19] P. Jaworski, F. Durante, W. Härdle and T. Rychlik, *Copula Theory and Its Applications*. Springer, 2009.

[20] D. Kurowicka and H. Joe, *Dependence Modeling*. World Scientific, 2010.

[21] H. Akaike, 'Information theory and an extension of the maximum likelihood principle.,' *Proceedings of the Second International Symposium on Information Theory Budapest*, 1973.

[22] J. Krouthén, 'Extreme joint dependencies with copulas. a new approach for the structure of c-vines.,' *Uppsala universitet, project report 2015*, 2015.

[23] S. Russel and P. Norvig, *Artificial intelligence: a modern approach*. Pearson, 2016.

[24] F. Alawad. (2019). 'Koopen,' [Online]. Available: `https://github.com/Faiga91/Koopen`. (accessed: 10.05.2021).

[25] M. Lysebo. (2020). 'Set up data,' [Online]. Available: `https://github.com/martilys/Project-thesis/blob/main/set_up_data`. (accessed: 10.05.2021).

[26] M. Maechler, 'Package 'copula' r,' *CRAN*, 2020.

[27] T. Nagler, 'Package 'vinecopula' r,' *CRAN*, 2021.

[28] M. Lysebo. (2021). 'Rmasterml,' [Online]. Available: `https://github.com/martilys/masterthesis/blob/main/RmasterML.Rmd`. (accessed: 02.06.2021).