DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2018

Examining Handovers in a Telecommunications Network Using Markov Chains and Dissimilarity Matrices

PONTUS RESARE

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Examining Handovers in a Telecommunications Network Using Markov Chains and Dissimilarity Matrices

PONTUS RESARE

Degree Projects in Mathematical Statistics (30 ECTS credits)
Degree Programme in Applied and Computational Mathematics (120 credits)
KTH Royal Institute of Technology year 2018
Supervisor at Ericsson AB: Johan Larsson
Supervisor at KTH: Camilla Landén
Examiner at KTH: Camilla Landén

TRITA-SCI-GRU 2018:215
MAT-E 2018:35

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Abstract

A telecommunications network is divided into cells, which have radio properties to lessen interference. Users move between these cells with their equipment. If the equipment is actively used, it goes through a process called handover when it moves between cells, this creates sequences of visited cells. This thesis investigates these handovers and the corresponding sequences of visited cells.

In this thesis there are two objectives related to the handovers between cells. The first is to determine if different types of sequences have different proportions of unwanted behaviour, the second is to develop a method to detect changes in the patterns of the handovers, between different time periods.

For both objectives it is examined if the sequences of visited cells can be modelled as r-order Markov chains. For the first objective, it is examined if there are different proportions of unwanted behaviour for the r most recently visited cells, using a Markov chain approach. The sequences are also examined as a whole with a clustering method using dissimilarity matrices. For the second objective, it is first examined if it is possible to model the sequences of visited cells from different time periods as Markov chains and then perform a homogeneity test between them. After that it is examined if dissimilarity metrics could be used to detect changes between time periods, this is done using dissimilarity matrices.

In the end it can be concluded that different types of sequences have different proportions of unwanted behaviour. Furthermore, it can be concluded that the approach of modelling the sequences as Markov chains in order to detect changes in handover behaviour between time periods, does not work. Finally, it is concluded that dissimilarity metrics could be used to detect changes between time periods, and additionally, some suitable dissimilarity metrics are presented.

Sammanfattning

Ett telekommunikationsnätverk är uppdelat i celler, dessa har radioegenskaper som ska minska interferensen. Användare rör sig mellan cellerna med sin utrustning. Om utrustningen används aktivt, så kommer den gå igenom en process kallad handover när den rör sig mellan celler, och sekvenser av besökta celler skapas. Detta examensarbete undersöker dessa handovers och de motsvarande cellsekvenserna.

I detta examensarbete finns det två mål relaterade till handover mellan celler. Det första målet är att bestämma om olika typer av cellsekvenser har olika proportioner av oönskat beteende, det andra målet är att skapa en metod som kan upptäcka skillnader i handovermönster mellan olika tidsperioder.

För båda målen så undersöks det om cellsekvenserna kan modelleras som Markovkedjor av ordning r. För att uppnå det första målet, så undersöks det med hjälp av en Markovkedjemetod, om sekvenser med samma r första celler har samma proportion av oönskat beteende. Cellsekvenserna undersöks också i sin helhet genom att de klassificeras med hjälp av en olikhetsmatris. För att uppnå det andra målet, så undersöks det först om det är möjligt att modellera cellsekvenserna från olika tidsperioder som Markovkedjor för att sedan göra ett homogenitettest dem emellan Efter detta så undersöks det om olikhetsmått kan användas för att upptäcka skillnader mellan tidsperioder, detta görs med hjälp av olikhetsmatriser.

I detta examensarbete så kan man konstatera att olika typer av sekvenser har olika proportioner av oönskat beteende. Dessutom så kan det konstateras att det inte fungerar att detektera skillnader i handovermönster genom att modellera cellsekvenserna som Markovkedjor och sedan göra homogenitetstest. Slutligen så kan det även konstateras att det fungerar att använda olikhetsmått för att upptäcka skillnader i handovermönster, dessutom så finns det förslag på några lämpliga olikhetsmått.

Acknowledgements

This thesis has been done in collaboration with the company Ericsson AB. Ericsson provided the initial ideas for this thesis, it also provided the data needed for this project. Furthermore, Ericsson provided equipment, and office space for the duration of the project. I would like to thank my supervisor at Ericsson, Johan Larsson, who explained mobile communications for me, and came with feedback during the process of doing this thesis. I would also like to thank all the other employees at Ericsson who helped in any way. Additionally, I would like to thank my supervisor at KTH, Camilla Landén, who helped with feedback and proofreading. Finally, I would like to thank my mother, Anneli Resare, who among other things helped me with proofreading this thesis.

Pontus Resare, Sollentuna, Sweden

May 31, 2018

Contents

Chapter 1

Introduction

This introduction will start with a brief overview of telecommunications, combined with a definition of what a cell is in telecommunications. Furthermore, it will state the objectives of this thesis, as well as briefly describe the methods employed to solve them. Finally, it will give an overview of the content of the thesis.

1.1 Telecommunications and Cells

Ericsson, the company where this thesis was done is a large telecommunication company. In telecommunication the primary objective is sending data to and from an end user in the form of SMS, phone calls, web-surfing, etc. In telecommunication the end user's device is called *User Equipment* (UE) [Cox, 2014, p. 2], most of the time this is a phone.

The first mobile telecommunication systems were introduced in the early 1980s [Cox, 2014, p. 2], and since then there have been improvements to the technology. These improvements are usually divided into different generations 1G, 2G, 3G, 4G. A generation can be implemented using different standards, one example is 2G where the system GSM (*Global System for Mobile Communications*) is used in Europe, and cdmaOne is used in the United States [Cox, 2014, p. 6]. This thesis is done on data collected from a network which is using the 4G system known as LTE-Advanced.

One key property in mobile telecommunication systems is the division of the area to be covered by the network into smaller areas called *cells*. In each small area the communication between the UE and the radio base station has some radio properties that lessens the interference with UEs that are in other cells in the network. The most common way to model this division is as hexagons [Miao et al., 2016, p. 95], see Figure 1.1.1.

Another key property is a *session*. When a UE is not actively used in an LTE-Advanced network, it is in a mode called RRC_IDLE [Cox, 2014, p. 42]. When this UE becomes active a *session* starts. If the UE moves between different cells during a session, it will go through a process called *handover*, if the UE has one or several handovers during a session, sequences of visited cells will be created. The sequences of visited cells will alternatively be referred to as UE cell movement, or UE movement between cells. This session can end in different ways, for example it can end when the UE becomes inactive, this ending is called *release*. There are several different release outcomes and they will be divided into desired behaviour and undesired behaviour, which will be referred to as *normal* and *abnormal*, respectively.

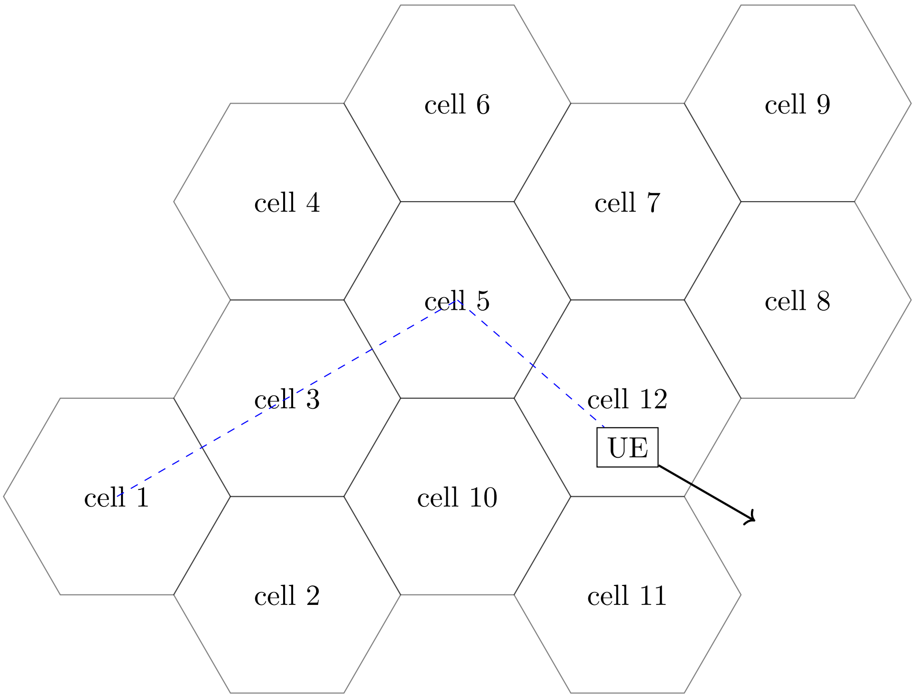

Figure 1.1.1: This figure shows how the area is divided into the smaller areas called cells. In the figure a UE is travelling though the network and creating a sequence of visited cells. In this case: *cell 1* → *cell 3* → *cell 5* → *cell 12.*

1.2 The Objectives

The thesis was done for a team at Ericsson which was interested in two subjects related to cell movement in the network. These subjects will be stated as objectives below.

Objective A: Look into if certain UE movements between cells coincide with unwanted behaviour. The unwanted behaviour in this case is if the UE is released from its session in an abnormal way. This will be referred to as *abnormal release*. In short:

- *Determine if different types of sequences have different proportion of abnormal releases.*

Objective B: Find a way to determine if there have been changes in the patterns of UE movements between cells for different time periods. This could for example help the team at Ericsson where this thesis was done, with discovering if certain software updates coincide with changed patterns of the UEs movements between cells. In short:

- *Develop a method to detect changes in patterns of UE movements between cells.*

1.3 The Tasks

In order to reach the two objectives, a number of tasks were created. A short motivation and summary of the tasks will be given below.

It was decided that the sequences of visited cells for each session would be modelled as Markov chains, either 1-order Markov chains or Markov chains of a higher order. The motivation for using Markov chains as a model is that the probabilities of handovers between cells, only depending on the most recently visited cells seems like a reasonable assumption. It is probable that what happened several handovers earlier does not matter for what happens next.

1.3.1 Tasks for Objective A

First, in order to find a solution to *Objective A*, that is, if certain movements between cells coincide with unwanted behaviour, it had to be classified what constitutes unwanted behaviour. This specification will be done in Subsection 2.4.3.

A question that emerges after defining *Objective A* is what is meant by different types of sequences. Two approaches were available. The first approach is to look at the most recently visited cells before a release is recorded. The second approach is to consider the sequences as a whole and find a way to categorise them into different types.

For the first approach, the idea is to determine if the probability of unwanted behaviour is dependent on the most recently visited cells. And if so, how far back in the history of visited cells this dependency goes. This can be seen as using a Markov assumption of order r. Consequently, task A.1 is to first define a hypothesis, then find a suitable statistical test and finally use and analyse this test.

The second approach, task A.2 is something similar to one of the methods that will be used to reach *Objective B*. First, choose a way to determine which sequences that are alike. Then divide the sequences into different classes depending on which sequences that are alike, using a clustering algorithm and finally check if any class is more prone to unwanted behaviour. This will for reasons that will be discussed in Chapter 5 *(Discussion)*, be more of an extra task.

Below is a short summary of the aforementioned tasks:

Task A.1: Using a Markov assumption, the probability of abnormal behaviour is dependent on the r latest visited cells.

1. Find a suitable statistical test.
2. Use this statistical test.

Task A.2: Divide sequences into classes using a dissimilarity metric.

1. Find a suitable sequence dissimilarity metric.
2. Find a suitable clustering algorithm.
3. Check if the different sequence classes have different proportions of unwanted behaviour.

1.3.2 Tasks for Objective B

Before describing the tasks used to solve *Objective B*, some basic properties of network data has to be stated.

In this thesis some local knowledge at Ericsson about general traffic patterns in telecommunication networks is going to be used. It is known from a large number of observations in many networks that *key performance indicators* (KPI) follows a 24-hour pattern. That is, if the time of day is the same, the KPI is roughly the same. This is due to the user's behaviour being similar. Most users commute to work in the morning, and work until they commute back home in the afternoon, with a lunch in the middle. Additionally, there is usually a difference in KPIs between weekdays and the weekend, this is due to most users working on weekdays, and not working on the weekend. There is also a difference between days of the week, e.g. a difference between Tuesdays and Thursdays. It is probable that UE cell movement also fulfils these properties. Consequently, in this thesis it will be assumed UE cell movement follows the aforementioned pattern.

Objective B can be rephrased as: Given two datasets representing two different time periods, find a way to determine if there are any meaningful differences in the behaviour of the UEs cell movement between these two time periods. Two different ideas for determining meaningful difference were evaluated in this thesis.

The first idea, which will be called task B.1, is to model the sequences of visited cells from the two time periods as being outcomes of two Markov chains which have equal transition probabilities. Then construct a test for this hypothesis that the two Markov chains have equal transition probabilities against the alternative hypothesis that they do not have equal transition probabilities. The test could then be used to determine changes in patters of UE movement between cells, if the hypothesis that the time periods have the same transition probabilities is rejected, there have been a change in the pattern of UE movement.

The aforementioned idea has to be checked, this check consists of two subtasks. The first (task B.1.1), is to validate that the homogeneity test works when many transition probabilities are low. Specifically, it should give reasonable test result, when performing the test on two simulated groups of visited cell sequences which are outcomes of Markov chains with the same transition probabilities. The second (task B.1.2), is to perform the test on real network data, and see if the test results seem reasonable, that is the test result should reflect the knowledge about the general traffic pattern.

The second idea, which will be called task B.2, is to instead use some form of dissimilarity metric. The dissimilarity metric assigns a value that tells how different the UE cell movements is between different time periods, which could be used to detect changes in patterns of UE movement between cells. This dissimilarity metric should also match the knowledge in Ericsson on which time periods that are alike. It should assign high dissimilarity between time periods where the time of day differs, and low dissimilarity if the time of the day is the same. This dissimilarity metric needs to be found.

It will be found in the following way: First some candidate dissimilarity metrics are created, this is task B.2.1. After that, the different dissimilarity metrics must be evaluated, this is done using network data. A large number of time periods are compared, creating a dissimilarity matrix, this is done for each dissimilarity metric. The different time periods are then divided into classes using the dissimilarity matrix. This will be referred to as task B.2.2. Finally, the classification of time periods for each dissimilarity metric will be compared with the knowledge at Ericsson about general traffic patterns, this is task B.2.3.

Below is a short summary of the aforementioned tasks:

Task B.1: Construct a homogeneity test using a Markov chain assumption.

> **Task B.1.1:** Do a validation of the statistical test using simulated data.
>
> **Task B.1.2:** Do the test on real network data and see if it performs well.

Task B.2: Find a suitable dissimilarity metric.

> **Task B.2.1:** Find suitable candidates for the dissimilarity metric.
>
> **Task B.2.2:** Use the different dissimilarity metrics to divide different time-periods into different classes of UE cell movement.
>
> **Task B.2.3:** Check if the classification matches the knowledge at Ericsson on which time-periods that have similar movements between cells.

1.3.3 Preliminary Task

As a preliminary task it was decided to check how well the modelling of the sequences of visited cells as Markov chains works, and if any order fits with the data.

Task C.1 Find the order of the Markov chain describing sequences of visited cells.

1. Find a suitable statistical test.
2. Use this statistical test.

1.4 Outline of Report

Below, there is a short outline of the report:

After this introductory chapter, an overview of the data extraction is given in Chapter 2 *(Background)*. This is combined with information about cells, handovers, release statuses, and the sequences of visited cells.

The next chapter, Chapter 3 *(Theory)* consists of two parts. Both parts contain theory that is needed to solve the problems stated in *Objective A* and *Objective B*. The first part contains information about Markov chains and the statistical test that will be used. These tests will be employed to solve task A.1, B.1, and C.1. The second part consists of several smaller parts. First a short description on how to cluster objects using a dissimilarity matrix. Secondly, it contains an overview of the different dissimilarity metrics (task B.2) that are going to be used to classify UE cell movement. Furthermore, it describes the dissimilarity metric used when classifying single sequences, which is going to be used in task A.2. Finally, it contains an overview of the PAM-algorithm, which is the algorithm used for classification.

In Chapter 4 *(Methodology)* the different tests and clusterings done are described. First the datasets used in the report are described, then the tasks related to *Objective A* are described, and finally all tasks related to *Objective B* are described. Most of the tables containing the results are in the Appendix. The penultimate chapter, Chapter 5 *(Discussion)* contains a discussion of the results and the conclusions. The final chapter, Chapter 6 *(Future Research)* contains ideas about possible future research subjects.

Chapter 2

Background

This chapter will expand on the concepts introduced earlier in Section 1.1, such as cells, handovers and release statuses. It will start with a basic description of an LTE-network, and continue with the handover procedure, and possible outcomes of the handover. Then it will continue with describing possible release status after the handover. Furthermore, it will describe some of the challenges of the data extraction, and caveats in the way the data is extracted.

2.1 Basic Network Description

As said previously in Chapter 1 *(Introduction)* this thesis is done using data from an LTE-Advanced system. A full description of an LTE-Advanced system is out of the scope of this report. However, in order to go forward, a rudimentary description of relevant parts of the system is still needed. Fortunately, a lot of complexities can be left out without losing relevant information. The simplified plan of the network can be seen in Figure 2.1.1. As can be seen in the figure, the relevant network is quite simple. There is the UE *(User Equipment)*, the eNBs *(evolved Node B)*[1], where the latter are responsible for radio communications between the UE and the rest of the network, and finally there is the rest of the network, containing infrastructure which is out of the scope of this report. Also marked in the figure are the relevant interfaces. The eNBs communicate with each other using an interface called X2, and communicate with the rest of the network using an interface called S1. The eNB that currently is handling the communication between the UE and the rest of the network is called the *serving eNB*.

In the introduction, the division of the area to be covered into cells was mentioned. These cells are serviced by eBNs, usually an eNB services groups of cells, a typical example is when an eNB controls three cells through three antennas that each control an angle of 120° [Cox, 2014, p. 3], which gives the hexagonal structure mentioned earlier. See Figure 2.2.1.

[1] The B stands for base station

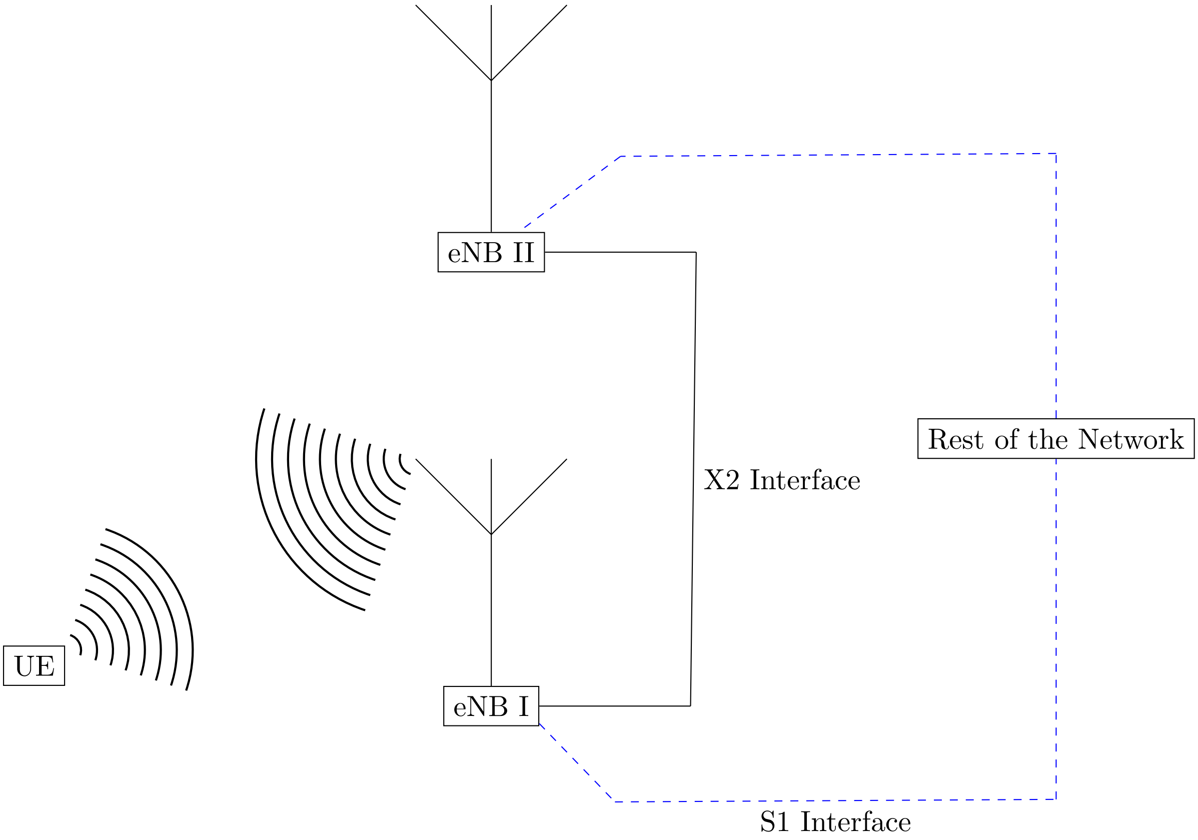

Figure 2.1.1: A very basic description of the network. Currently the UE is communicating with eNB I. That is, eNB I is the serving eNB.

2.2 Handovers

2.2.1 Handover Introduction

As mentioned previously, movements between cells correspond to the real world event when a UE moves across the border between cells in Figure 2.2.1. That is, when the UE moves from its old cell over the border to the new cell, it will become harder and harder to communicate with the old cell. Consequently the network will then try to do a handover to the new cell [Cox, 2014, p. 237-254]. After the handover the UE is communicating with the new cell instead. The handover could be between cells that have the same serving eNB or between cells that have different serving eNBs. If the handover is between cells that have different eNBs, the communication between the UE and the rest of the network will be routed though the new eNB after the handover.

Example of Handover

An example of a handover can be seen in Figure 2.2.1. In the figure a car containing a UE (phone) is currently travelling from the lower left corner to the upper right corner along a highway. When the UE moves between cell 5 and cell 7, the signal strength from cell 5 becomes weaker and the signal from cell 7 becomes stronger. After some threshold is reached the UE will contact eNB II and send a report about the different signal strength levels. When eNB II decides that it would be favourable for the UE to be in contact with cell 7 instead of cell 5, it will do the *handover* to eNB III, and eNB III will become the serving eNB.

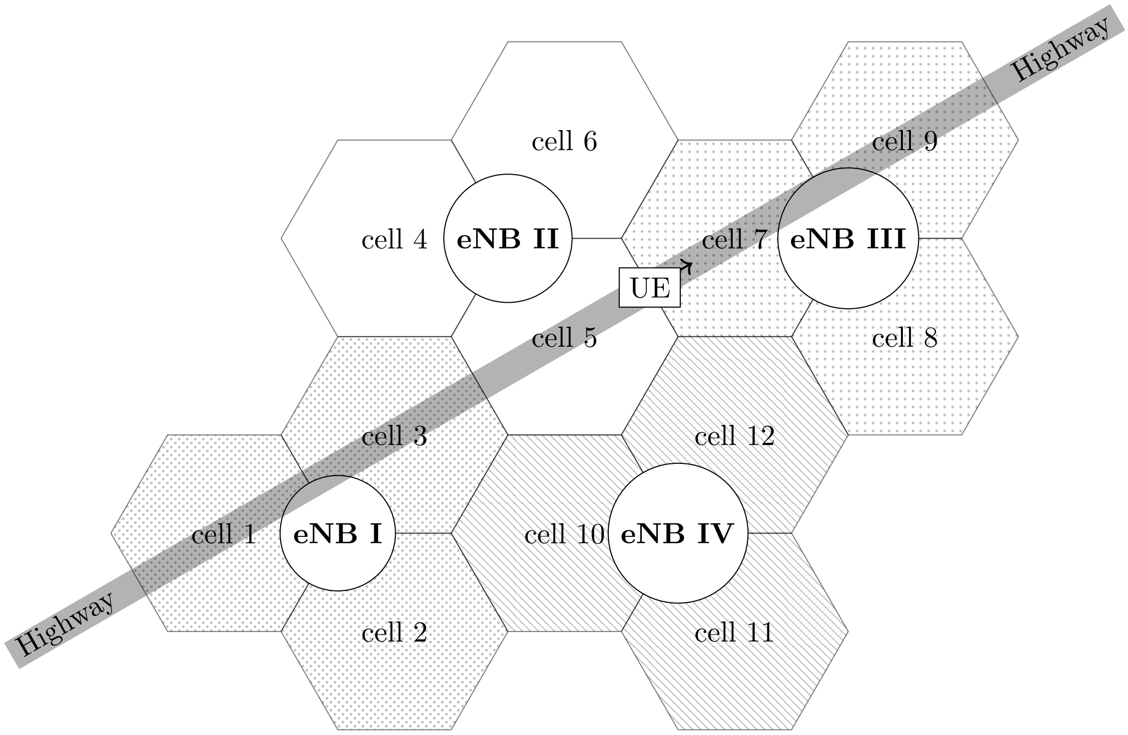

Figure 2.2.1: Sketch on how the area to be covered is divided into cells and corresponding eNBs. With a grey bar representing a highway.

2.2.2 The X2-interface Handover Procedure

There are a few different handover scenarios, the handover could for example either go between cells with the same eNB or between cells with different eNBs. There are a few different few scenarios when there is a handover between different eNBs, the most common is the *X2-based handover* procedure [Cox, 2014, p. 250]. These are handovers over the X2-interface which is the interface directly between different eNBs, see Figure 2.1.1. The data collected in this report will mainly come from handovers using this interface.[2]

As stated earlier, a handover is started when the current eNB (II in Figure 2.2.1) receives a measurement report from the UE that indicates that a handover would be favourable. Then it will start the procedure by sending a *Handover Request* over the X2-interface to the eNB (III in Figure 2.2.1) that services the more favourable cell. The new eNB will answer this request with a *Handover Request Acknowledge* over the X2-interface. After these initial messages, other messages are sent and more procedures are executed. More information can be found in e.g. [Cox, 2014, p. 250-253].

[2] The method used to extract data from handovers over the X2-interface, also extract some data from handovers between cells with the same serving eNB, see Subsection 2.4.1.

2.2.3 Handover Outcomes

Worth mentioning is that the handover is not always successful. There are several possible outcomes of the handover procedure, simplified these could be:

Possible Handover Outcomes:

- Successful handover
- Handover cancelled
- Handover failed

One aspect of this report is finding out if there are any correlations between the movement of UEs between cells and problematic events in the network, which is problematic events in the new eNB after a handover. Therefore, it is worth noting that the handover being cancelled is not a problematic event.

2.3 Release Status

Assuming there was a successful handover, several new things could happen in the new eNB. A simplified version is that one of the following three things happen:

- UE disconnects normally
  - The UE disconnects from the eNB in a normal fashion, the user could for example become inactive.
- UE disconnects abnormally
  - The UE disconnects from the eNB in an abnormal fashion, here an example could be that the eNB looses signal with the UE.
- New handover
  - The UE could have a new handover. Either back to the eNB that the UE was in earlier or to some new eNB.

The two most common endings of UE communication with an eNB is either a successful new handover or user inactivity. These outcomes in the new eNB, presented above, will be referred to as the *Release Status*.

2.4 Data Extraction

Ericsson has several ways of logging the data in their customer's networks. One way that the data is logged is as CTR-data, which is basically just events with their attributes. Each event that happens in the network is logged with attributes in the following form.

```
{
        eventName = xxxxx
        eventId = xxxxx
        time = xxxxx
        RAC − UE − REF = xxxxx
        .
        .
        .
        other attributes
        .
        .
        .
}
```

2.4.1 Extracting History of Visited Cells

One key to finding the patterns of UE movement between cells in this report is that the CTR-logs of the event *Handover Request* contain information on which cells a UE has previously visited. Both the *Handover Request* event and the *Handover Request Acknowledge* event are saved in the CTR logs, both in the old eNB (I in Figure 2.2.1) and the new eNB (II in Figure 2.2.1).

Additionally, handovers that occur between cells with the same serving eNB are also recorded as CTR-data. Fortunately the internal handover events in the serving eNB are recorded in the same way as handovers over the X2-interface. Consequently these handovers are also extracted it the same way.

These two events the *Handover Request* and the *Handover Request Acknowledge* contain among other things, a 3GPP message. This message contains a lot of information, relevant to this thesis is that the message corresponding to the event *Handover Request Acknowledge*, contains a list of names of which cells the UE has visited and the amount of time spent in each cell, in seconds. Simplified the data is on the format.

```
{
        {Cell identifier (cell it tries to do a handover from),
         time connected to that cell in seconds},

        {Cell identifier (cell it visited before the one above),
         time connected to that cell in seconds},

        {Cell identifier (cell it visited before the one above),
         time connected to that cell in seconds},

        and so on...
}
```

Note that the cell the UE tries to make a handover from is included in the visited cell list, that is, the cell that is sending the *Handover Request*. In the case of the UE moving as in Figure 2.2.1 with the UE staying in 2 seconds in each cell, the list would look as:

```
{
        {name: 5,
         time: 2},

        {name: 3,
         time: 2},

        {name: 1,
         time: 2},

}
```

A slightly simplified and commented copy of one such message 3GPP message can be seen in Chapter D *(Example of a Handover Request Message)* in the Appendix. The list of visited cells and the corresponding time spent in them is easily extracted using regular expressions. The CTR-log corresponding to the *Handover Request Acknowledge* also contain a cell identifier for the new cell, which will be added to the list of visited cells.

2.4.2 Matching History of Visited Cells with Behaviour

When a UE connects to an eNB, the eNB will assign a temporary identification number to the UE called RAC-UE-REF. The UE has this RAC-UE-REF as long as it is actively connected to the eNB. If the UE becomes inactive or if there is a successful handover this RAC-UE-REF will no longer be in use. The RAC-UE-REF is useful because it can link together the list of visited cells with possible handover outcomes as well as possible outcomes in the new eNB.

There is however one caveat. When the *Handover Request* (which contains the list of visited cells) is received in the new eNB, no new RAC-UE-REF is assigned in the CTR-logs. This RAC-UE-REF is assigned later, when the new eNB answers with a *Handover Request Acknowledge.*

Fortunately the *Handover Request Acknowledge* is sent out very soon after *Handover Request* is received. This together with that each pair of *Handover Request Acknowledge* and corresponding *Handover Request* has an id called UE-X2AP-ID that is unique, for a short time, makes it possible to assign a RAC-UE-REF to the *Handover Request* retroactively. Using this, the history of visited cells from the *Handover Request* can be matched with the RAC-UE-REF from the *Handover Request Acknowledge* and the behaviour in the new eNB, that is *Handover Outcomes* and *Release Status Outcome.* A summary can be seen below.

1. *Handover Request*
    - Sent from old eNB to new eNB
    - Contains information about the UE's previously visited cells
    - Lacks RAC-UE-REF
    - Has UE-X2AP-ID
2. *Handover Request Acknowledge*
    - Sent from new eNB to old eNB

- Has RAC-UE-REF
- Has UE-X2AP-ID

3. *Handover Outcome*
    - Gives information on the the result of the Handover
    - Has RAC-UE-REF
4. *Release Status Outcome*
    - Gives information about what happened in the new eNB
    - Has RAC-UE-REF

2.4.3 Labelling the Data

As mentioned earlier in the introductory chapter, one aspect that will be looked into in this report is *Objective A*, that is, does the proportion of unwanted behaviour depend on the history of visited cells. In order to determine the former, it has to be specified what constitutes unwanted behaviour. The unwanted behaviour has two sources, the first unwanted behaviour is if the handover to the cell, where the eNB is recording, fails. It may be a bit confusing to say that one of the behaviours of a UE in a new eNB is that the UE never connects to the new eNB. In this case it makes sense however, a failed handover to the new eNB is unwanted behaviour, and it could depend which cells the UE has previously visited. The second unwanted behaviour is if there is a successful handover to the new eNB, and that the UE disconnects abnormally. The event when there is a successful handover to the new eNB, and then a successful handover from the new eNB will be labelled as normal behaviour. In this report the unwanted behaviour will be refereed to as abnormal and the wanted behaviour as normal. A summary can be seen in Table 2.4.1.

Outcome of handover from old eNB	Release status in new eNB	Resulting label
failed	–	abnormal
cancelled	–	normal
succeeded	handover from new eNB	normal
succeeded	UE disconnects normally	normal
succeeded	UE disconnects abnormally	abnormal

Table 2.4.1: Overview on the labelling of sequences. All sequences are labelled depending on handover outcomes and release status in new eNB. In the Appendix in Table A.1.5 a more thoroughly description of the labelling can be seen.

2.4.4 Example of Data

If one only considers the cell the UE visits and not how long the UE is in each cell, the data becomes a list on the format.

$$(\textit{Older Cells} \rightarrow \textit{Cell sending Handover Requenst} \rightarrow \textit{Cell where data is collected}), (\textit{label})) \tag{2.4.1}$$

For example,

$$((2786 \to 3043 \to 2786), (normal)) \tag{2.4.2}$$

or

$$(9587 \to 3378 \to 2786 \to 3378 \to 2786), (abnormal)) \tag{2.4.3}$$

Here the numbers represents different cell identification numbers. In general some sequences are more common than others, this can for example be seen in Figure 2.4.1.

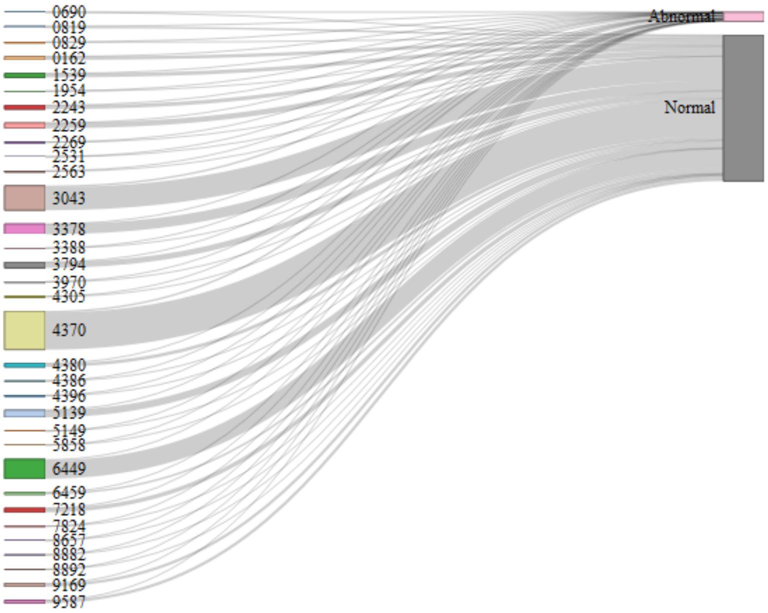

Figure 2.4.1: A visualization on how some UEs cell sequences are more common than other. The thickness of the path indicates how many UE handovers normal and abnormal there are between different cells and cell 2786.

2.5 Challenges with the Data Extraction

2.5.1 Limitations with the Data Recording

There are some properties of the data collection that needs to be pointed out. It is important to note that all data in the network is not collected. This is because the recording of the CTR-data only happens in some eNBs. Data is not collected if the UE do not pass though a cell serviced by a eNB where data is collected. This is illustrated in Figure 2.5.1

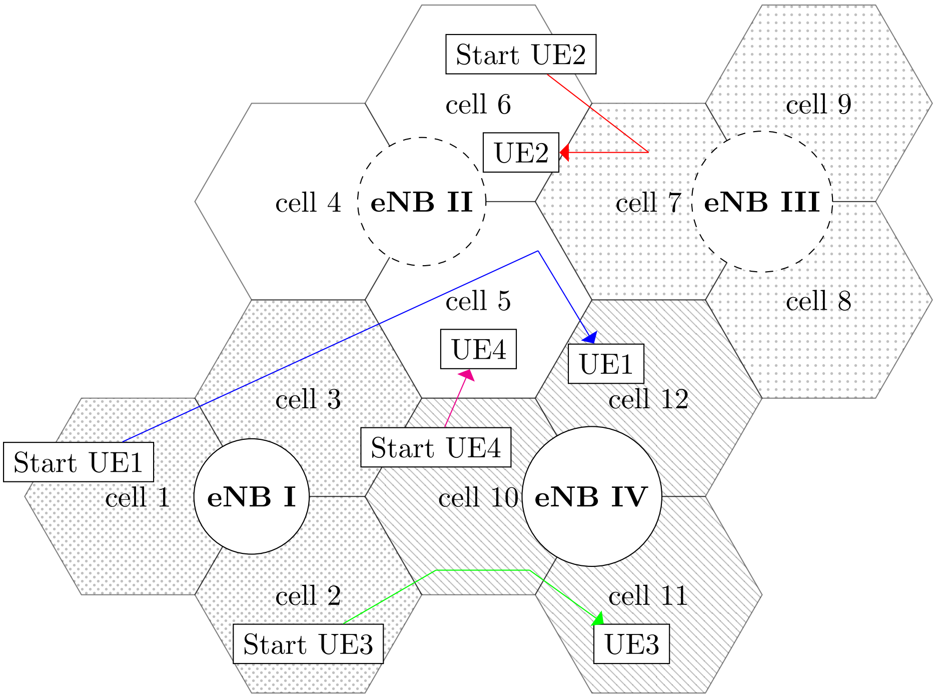

Figure 2.5.1: Cell plan showing four sequences of handovers. Data is being recorded in the eNBs with dashed outline. That is, eNB II, and eNB III.

In Figure 2.5.1 the actual visited cell sequences are:

UE 1: cell 1, cell 3, cell 5, cell 12.

UE 2: cell 6, cell 7, cell 6.

UE 3: cell 2, cell 10, cell 11.

UE 4: cell 10, cell 5.

And the recorded visited cell sequences are:

UE 1: cell 1, cell 3, cell 5.

UE 2: cell 6, cell 7.

UE 2: cell 6, cell 7, cell 6.

UE 4: cell 10, cell 5.

That is, some visited cell sequences will be recorded twice (UE 2), some will not be recorded at all (UE 3), and some will only have a part of the sequence recorded (UE 1). The first problem with visited cell sequences being recorder twice is easily fixed. There is a temporary identification number *Masked IMEISV* that is the same in the CTR-data in all eNBs where the data is recorded. This together with the fact that the time spent in each cell is also recorded, makes it possible to determine if two recorded sequences describe the same sequence, and therefore one can merge the recordings into one.

The last problem with data of UE movement between cells not being recorded, if they do not pass through the eNBs where recordings happen, is not easily fixed and could have consequence for the analysis in this report. It will at least lead to a bias in the data towards longer sequences, because those will have a higher probability of being recorded.

2.5.2 Limitations with the Visited Cells List

There are a few thing to note about the message accompanying the *Handover Request*.

First, the number of visited cells listed in the message accompanying the *Handover Request* is always between 1 and 16. After comparing sequences that had the same *Masked IMEISV* and if was concluded that the reason for this is that the list of visited cells is constrained to the 16 most recently visited. When the UE has visited more than 16 cells, the oldest cell is dropped from the message. This was not a large limitation however, as sequences can be linked together into to longer sequences with the same technique used to determine if sequences are unique, i.e. time spent in cells and the *Masked IMEISV*.

Furthermore, the time spent in each cell is recorded in the message accompanying the *Handover Request* in whole seconds and the rounding is a bit peculiar. All values are rounded downwards to the nearest whole second. The exception is if less than 1 second has been spent in a cell, then the time spent is rounded upwards to 1, i.e. if the time spent in each cell is recorded as 1, the UE could have been there for 0 - 2 seconds. Additionally, a lot of the time spent in each cell are recorded as 4095, this seems to be some form of error value.

Chapter 3

Theory

3.1 Mathematical Description of the Visited Cell Sequences

3.1.1 Unlabelled Visited Cell Sequences

In order to model the visited cell sequences, the state variable x_t is first introduced, this corresponds to which cell the UE is in after the t-th handover. Consequently, x_0 is the first cell the UE visits and x_1 is the second cell and so on. The sequences of visited cells will end when the UE is recorded for the last time, this time will be denoted T and sometimes called the recording time. The recording time can take any values equal or larger than 1 ($T = 1, 2, \dots$). Using the aforementioned each sequence can be written:

$$x_0, x_1, \dots, x_{T-1}, x_T \, . \tag{3.1.1}$$

Each such sequence could be seen as outcomes of a stochastic process $X_0, X_1, \dots, X_{T-1}, X_T$.

3.1.2 Labelled Visited Cell Sequences

The visited cell sequences can be labelled. This labelling will be modelled as an extra state after X_T and will be denoted y_T. Using the aforementioned each sequence can be written:

$$x_0, x_1, \dots, x_{T-1}, x_T, y_T \, . \tag{3.1.2}$$

Each such sequence could be seen as outcomes of a stochastic process, $X_0, X_1, \dots, X_{T-1}, X_T, Y_T$.

3.1.3 Using Time Spent in Cells

There is data available of the time the UE spends in each cell. Therefore, it would for example be possible to model the visited cell sequences as continuous-time Markov process. Due to some problems it was decided that this would not be done, more details can be found in Subsection 6.1.1. Additionally, it would be possible to incorporate the time spent in each cell when using Markov chains, a description on how can be found in Subsection 6.1.2. In the end it was decided to not use the time spent in each cell for the analysis.

3.2 Markov Chains

This thesis will model the sequences of visited cells introduced earlier as Markov chains. Therefore, it would be of help to introduce some basic concepts, definitions and properties of Markov chains. As well as a definition of higher order Markov chains.

3.2.1 Introduction on Markov Chains

A Markov chain is a stochastic process $\{X_t\}$ in discrete time with a finite state space $X_t \in \{1, \ldots, m\}$, that has the Markov property, that is

$$P(X_t|X_{t-1}, X_{t-2}, X_{t-3}, \ldots, X_0) = P(X_t|X_{t-1}) \ . \tag{3.2.1}$$

If the probability of transition for all pairs of X_t and X_{t-1} are the same for all t, that is, if

$$P(X_t = j|X_{t-1} = i) = P(X_{t-1} = j|X_{t-2} = i) = \ldots = P(X_1 = j|X_0 = i) \quad \forall\ i,\ j \ , \tag{3.2.2}$$

then the Markov chain is said to be time homogeneous. In this report the Markov chain is assumed to be time homogeneous.

A key concept when dealing with Markov chains is the probability matrix $\mathbf{P}$, also called transition matrix. If one first defines

$$p_{ij} = P(X_t = j|X_{t-1} = i) \quad i = 1, 2 \ldots, m \quad j = 1, 2 \ldots, m \ . \tag{3.2.3}$$

Then the probability matrix $\mathbf{P}$ is defined as

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1m} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & p_{m3} & \cdots & p_{mm} \end{bmatrix} \ . \tag{3.2.4}$$

The initial probabilities will be written

$$\lambda_i = P(X_0 = i) \ . \tag{3.2.5}$$

With notation as above the probability of some sequence $(x_0, x_1, \ldots, x_T)$ is,

$$P(x_0, x_1, \ldots, x_T) = \lambda_{x_0} p_{x_0 x_1} p_{x_1 x_2} \cdots p_{x_{T-1} x_T} \ . \tag{3.2.6}$$

If f_{ij} is the number of transitions from state i to state j in this sequence, then

$$P(x_0, x_1, \ldots, x_T) = \lambda_{x_0} \prod_{i,j} p_{ij}^{f_{ij}} \ . \tag{3.2.7}$$

3.2.2 Estimation of Probabilities from Data

The maximum likelihood estimates of the transition probabilities, p_{ij} can be found in the following way.

First assume there are n sequences, each of length T. Furthermore, assume the sequences are outcomes of the same Markov chain described by the same transition matrix $\mathbf{P}$, and that there are m different possible states. Additionally, assume that the initial probabilities λ_i are known. Then the likelihood $\mathcal{L}$ of some n sequences is

$$\mathcal{L} = \prod_{i,j} p_{ij}^{f_{ij}} \ , \tag{3.2.8}$$

where f_{ij} is the total number of transitions from state i to state j in all n sequences, f_{ij} will be referred to as frequency counts. The maximum likelihood estimates $\hat{p}_{ij}$ are those p_{ij} that maximize $\mathcal{L}$ under the constraints

$$\sum_{j=1}^{m} p_{ij} = 1 \quad \text{and} \quad p_{ij} \geqslant 0 \ . \tag{3.2.9}$$

This maximization problem can be solved using for example *Lagrange Multipliers* (it will not be done here), and the following maximum likelihood estimates can be obtained

$$\hat{p}_{ij} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}} = \frac{f_{ij}}{f_{i*}} \ . \tag{3.2.10}$$

3.2.3 Homogeneity Test on Markov Chains

One thing of interest in this report is testing if two Markov chains have the same transition matrix $\mathbf{P}$. There are several different test statistics available, one example is given in [Anderson and Goodman, 1957], the same test with clear description on how to reduce the degrees of freedom when dealing with frequency counts that are zero is described in [Bickenbach and Bode, 2003].

First, assume that there are two groups of sequences $S_1^{(1)}$, $S_2^{(1)}$, ..., $S_{n_1}^{(1)}$ and $S_1^{(2)}$, $S_2^{(2)}$, ..., $S_{n_2}^{(2)}$ available. The groups of sequences has n_1 and n_2 sequences respectively. Secondly assume there are m possible states, which can be given integer identifications $1, 2, \ldots, m$ without loss of generality.

Furthermore assume that both groups of sequences are 1-order Markov chains, where the transition probabilities matrices are assumed to be $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ respectively. The maximum likelihood transition probability estimate, for transitions from state i to state j for group $h \in 1, 2$ is mentioned in Subsection 3.2.2:

$$\hat{p}_{ij}^{(h)} = f_{ij}^{(h)} / f_{i*}^{(h)} \ . \tag{3.2.11}$$

Where $f_{ij}^{(h)}$ is the number of transitions from state i to state j in group h, and $f_{i*}^{(h)}$ is the number of transitions from state i. If the assumption is that the transition probabilities are the same for both groups, the maximum likelihood transition probability estimate is instead

$$\hat{p}_{ij} = f_{ij} / f_{i*} \ . \tag{3.2.12}$$

Where f_{ij} is the number of transitions from state i to j in both groups of sequences together and f_{i*} is the number transitions from state i in both groups together. Using the definition of $\hat{p}_{ij}^{(h)}$ and $\hat{p}_{ij}$

$$Q = \sum_{h=1}^{2} \sum_{\hat{p}_{ij} \neq 0} f_{i*}^{(h)} \frac{\left(\hat{p}_{ij}^{(h)} - \hat{p}_{ij}\right)^2}{\hat{p}_{ij}} \tag{3.2.13}$$

can be used as a test statistic. This is the same test statistic as in [Anderson and Goodman, 1957, p. 101] with a slight modification, in the previously mentioned source, it is assumed that $\hat{p}_{ij} > 0$, this assumption does not hold in this thesis. When the terms in the sum where $\hat{p}_{ij} = 0$ are removed, the degrees of freedom have to be reduced, this is for example done in [Bickenbach and Bode, 2003, p. 369]. The test statistic Q is asymptotically distributed as a chi-square random variable with $\sum_{i=1}^{m}(r_i - 1)(c_i - 1)$ degrees of freedom. Which will be written as:

$$Q \sim \text{asy}\,\chi^2\left(\sum_{i=1}^{m}(r_i-1)(c_i-1)\right) . \tag{3.2.14}$$

In the equation above r_i is the number of positive $\hat{p}_{ij}$ for each state i. Secondly $c_i = 2$ when there are at least one positive element in $\hat{p}_{i1}^{(1)}, \ldots, \hat{p}_{im}^{(1)}$, and also at least one positive element in $\hat{p}_{i1}^{(2)}, \ldots, \hat{p}_{im}^{(2)}$. If only one of $\hat{p}_{i1}^{(1)}, \ldots, \hat{p}_{im}^{(1)}$ and $\hat{p}_{i1}^{(2)}, \ldots, \hat{p}_{im}^{(2)}$ has at least one positive element, then $c_i = 1$.

This test is equivalent to m chi-square test of homogeneity on contingency tables with dimensions $r_i \times c_i$. Finally, write

$$k = \sum_{i=1}^{m}(r_i-1)(c_i-1) . \tag{3.2.15}$$

If our assumptions that both groups of sequences are Markov chains is true, then

$$p = P(X > Q), \quad \text{where} \quad X \in \chi^2(k) . \tag{3.2.16}$$

is the probability that, the test statistic Q or a larger value would be obtained if $\mathbf{P}^{(1)} = \mathbf{P}^{(2)}$ was true.

3.2.4 Markov Chains of Higher Order

A possible generalization of a Markov chain is to define Markov chains of order r (also called r-order Markov chains) as a stochastic time series with the property

$$P(X_t|X_{t-1}, X_{t-2}, X_{t-3}, \ldots, X_0) = P(X_t|X_{t-1}, X_{t-2}, X_{t-3}, \ldots, X_{t-r}) . \tag{3.2.17}$$

In essence, a normal Markov chain with a longer memory. The following short-hand notation will be used from this section and onwards

$$p_{ij|k} = P(X_t = k|X_{t-1} = j, X_{t-2} = i) \tag{3.2.18}$$

for a second-order Markov chain. Furthermore

$$p_{ijk|l} = P(X_t = l|X_{t-1} = k, X_{t-2} = j, X_{t-3} = i) \tag{3.2.19}$$

for a third-order, and so on. The reason it is written $p_{ijk|l}$ in this thesis and not p_{ijkl} is to make the distinction clear between the l state which the processes is transitioning to and ijk which are the states that the processes have been in.

In this report, one question is if a group of sequences $S_1, S_2, \ldots S_n$ each representing a sequence of visited cells with different lengths T_i, that is $(X_0, \ldots, X_{T_i})$, can be modelled as outcomes from a Markov chain of order r.

In [Anderson and Goodman, 1957] and a way to construct a test for this is stated. The basic idea of the test is to first create a null hypothesis that the Markov chain is of order $r-1$. Then test this against the alternative hypothesis that the Markov chain is of order r.

Below, the case of testing the null hypothesis of a Markov chain of order 1 against an alternative hypothesis of a Markov chain of order 2 will be stated first. Afterwards the general test will be stated.

Test of Order 1 versus Order 2

First define $f_{ijk}(t)$ to denote the number of times in which $X_t = k$, $X_{t-1} = j$, and $X_{t-2} = i$ in all sequences (S_1, S_2, …, S_n) where $i, j, k = 1, 2, \ldots, m$ and $t = 2, 3, \ldots, T$. Then define

$$f_{ijk} = \sum_{t=2}^{T} f_{ijk}(t) \ , \tag{3.2.20}$$

and

$$f_{ij*} = \sum_{k=1}^{m} f_{ijk} \ . \tag{3.2.21}$$

Under the assumption that the alternative hypothesis is true, that is, the sequences are Markov chains of order 2. Then the maximum likelihood estimate of $p_{ij|k}$ is

$$\hat{p}_{ij|k} = f_{ijk}/f_{ij*} \ . \tag{3.2.22}$$

Furthermore, under the null hypothesis of a Markov chain of order 1, the maximum likelihood estimate of $p_{j|k}$ is

$$\hat{p}_{j|k} = \sum_{i=1}^{m} f_{ijk} \Bigg/ \sum_{i=1}^{m}\sum_{k=1}^{m} f_{ijk} = \frac{f_{*jk}}{f_{*j*}} \ . \tag{3.2.23}$$

The test statistic is

$$Q = \sum_{i=1}^{m} \sum_{\hat{p}_{j|k} \neq 0} f_{ij*} \frac{\left(\hat{p}_{ij|k} - \hat{p}_{j|k}\right)^2}{\hat{p}_{j|k}} \ . \tag{3.2.24}$$

Which is stated in [Anderson and Goodman, 1957, p. 101], where it is stated that Q is asymptotically χ^2 distributed with, $m(m-1)^2$ degrees of freedom. Terms can be excluded from the test, if some of the estimated probabilities are zero, that is, if $\hat{p}_{j|k} = 0$, or if certain transitions do not happen, that is, if $f_{ij*} = 0$. Then the degrees of freedom have to be reduced, this is for example done in [Bickenbach and Bode, 2003]. With the reduced degrees of freedom Q instead becomes

$$Q \sim \text{asy}\, \chi^2 \left(\sum_{j=1}^{N} (r_j - 1)(c_j - 1) \right) = \chi^2(k) \ . \tag{3.2.25}$$

Here c_j is the number of positive $\hat{p}_{j|k}$ for each j, and r_j is the number of positive f_{ij*} for each j.

Then, assuming that the collection of sequences $S_1, S_2, \ldots S_n$ are Markov chains with the same probabilities $p_{ij|k}$,

$$p = P(X > Q), \quad \text{where} \quad X \in \chi^2\,(k) \tag{3.2.26}$$

is the probability that a test statistic Q, or larger, would be obtained if $p_{ij|k} = p_{j|k}$.

Test of Order $r-1$ versus Order r

In the general case the approach is similar. Firstly, use $f_{ij\cdots kl}(t)$ to denote the number of times in which $X_{t-r}=i$, $X_{t-r+1}=j$, $\ldots$, $X_{t-1}=k$, and $X_t=l$, in all sequences $(S_1, S_2, \ldots, S_n)$. Furthermore $\{i, j, \cdots, k, l\}=1,2,\ldots,m$ and $t=r,r+1,\ldots,T$. In $f_{ij\cdots kl}(t)$ the centred dots $(\cdots)$ stands for the extra needed indexes.

First define

$$f_{ij\cdots kl}=\sum_{t=r}^{T} f_{ij\cdots kl}(t)\ , \tag{3.2.27}$$

secondly, define

$$f_{ij\cdots k*}=\sum_{l=1}^{m} f_{ij\cdots kl}\,, \tag{3.2.28}$$

and

$$\hat{p}_{j\cdots k|l}=\frac{\sum_{i=1}^{m} f_{ij\cdots kl}}{\sum_{i=1}^{m} f_{ij\cdots k*}}\ , \tag{3.2.29}$$

furthermore, define

$$\hat{p}_{ij\cdots k|l}=\frac{f_{ij\cdots kl}}{f_{ij\cdots k*}}\,. \tag{3.2.30}$$

By computing the test statistic

$$Q_{j\cdots k}=\sum_{i=1}^{m}\sum_{l=1}^{m} f_{ij\cdots k*}\frac{\left(\hat{p}_{ij\cdots k|l}-\hat{p}_{j\cdots k|l}\right)^2}{\hat{p}_{j\cdots k|l}} \tag{3.2.31}$$

a test of the hypothesis can be done. Here, with the notation mentioned earlier

$$Q_{j\cdots k}\sim \text{asy}\,\chi^2\left((m-1)^2\right)\ , \tag{3.2.32}$$

according to [Anderson and Goodman, 1957]. That is $Q_{j\cdots k}$ is asymptotically χ^2 distributed with $(m-1)^2$ degrees of freedom. The sum of all these $Q_{j\cdots k}$ is

$$Q=\sum_{j=1}^{m}\cdots\sum_{k=1}^{m} Q_{j\cdots k}\sim \text{asy}\,\chi^2\left(m^{r-1}(m-1)^2\right)\ , \tag{3.2.33}$$

that is, the sum of earlier test statistics is also asymptotically χ^2 distributed with $m^{r-1}(m-1)^2$ degrees of freedom.

With Reduced Number of Degrees of Freedom

In the case of some $\hat{p}_{j\cdots k|l}$ or $f_{ij\cdots k*}$ being zero, the corresponding terms are removed because they are undefined and the number of degrees of freedom are reduced,

$$Q_{j\cdots k}\sim \text{asy}\,\chi^2\left((r_{j\cdots k}-1)(c_{j\cdots k}-1)\right)\ . \tag{3.2.34}$$

Where $c_{j\cdots k}$ is the number of $\hat{p}_{j\cdots k|l}$ where $\hat{p}_{j\cdots k|l}>0$, and $r_{j\cdots k}$ is the number of $f_{ij\cdots k*}$ where $f_{ij\cdots k*}>0$. Then

$$Q=\sum_{j=1}^{m}\cdots\sum_{k=1}^{m}\sim \text{asy}\,\chi^2\left(\sum_{j=1}^{m}\cdots\sum_{k=1}^{m}(r_{j\cdots k}-1)(c_{j\cdots k}-1)\right)\ , \tag{3.2.35}$$

because all $Q_{j\cdots k}$ are independent.

3.2.5 Test of Homogeneity between Markov Chains of Order r

The test given in Subsection 3.2.3 can be extended to test for homogeneity between Markov chains of order r instead. First assume there are two groups of sequences, identified by $h \in \{1, 2\}$. They each have the transition count $f^{(h)}_{ij\cdots kl}$ and the transition count $f_{ij\cdots kl}$ together. Now, let

$$\hat{p}^{(h)}_{ij\cdots k|l} = \frac{f^{(h)}_{ij\cdots kl}}{f^{(h)}_{ij\cdots k*}} \tag{3.2.36}$$

and

$$\hat{p}_{ij\cdots k|l} = \frac{f_{ij\cdots kl}}{f_{ij\cdots k*}} \,. \tag{3.2.37}$$

Under the assumption that $\hat{p}_{ij\cdots k|l} > 0$, then

$$Q_{i\cdots k} = \sum_{h=1}^{2} \sum_{l=1}^{m} f^{(h)}_{ij\cdots k*} \frac{\left(\hat{p}^{(h)}_{ij\cdots k|l} - \hat{p}_{ij\cdots k|l}\right)^2}{\hat{p}_{ij\cdots k|l}} \,, \tag{3.2.38}$$

forms a test statistic for the test of the null hypothesis that two groups of sequences are r-order Markov chains with the same transition probabilities against the alternative hypothesis that they are r-order Markov chains with different transitions probabilities.

The test statistic is

$$Q_{i\cdots k} \sim \text{asy}\, \chi^2\left(m^r(m-1)\right) \,. \tag{3.2.39}$$

If $\hat{p}_{ij\cdots k|l} = 0$, or $f^{(h)}_{ij\cdots k*} = 0$ the terms in the sum are removed and the degrees of freedom can be reduced using the method described in earlier sections.

3.2.6 Test of Homogeneity for Labelled Sequences

Below follows a slightly different test than the ones introduced above. Here the data is as described in Subsection 3.1.2. Assume that there are a total of n sequences on the form

$$X_0, X_1 \ldots, X_T, Y_T \tag{3.2.40}$$

where $Y_T \in \{"normal", "abnormal"\}$. That is Y_T describes the different outcomes in Subsection 2.4.3. Furthermore $X_t \in 1, 2, \cdots, m$ represent the different states.

One thing of interest could be to answer the question: if, for some r,

$$P(Y_T|X_T, X_{T-1}, \ldots, X_0) = P(Y_T|X_T, X_{T-1} \ldots, X_{T-r+1}) \,. \tag{3.2.41}$$

That is, is the probability of *"normal"* or *"abnormal"* dependent on the history of visited states, and if it is dependent, how far back does the dependency go?

Let us first start with some notation. As before we will write

$$P(Y_T = y|X_T = l, X_{T-1} = k, \ldots, X_{T-r+1} = j, X_{T-r} = i) = p_{ij\cdots kl|y} \,. \tag{3.2.42}$$

And $f_{ij\cdots l,y}$ are the number of sequences where

$$X_{T-r} = i, , X_{T-r+1} = j, \ldots, X_{T-1} = k, X_T = l, Y_T = y \ . \tag{3.2.43}$$

Note that there are only two different y:s, $f_{ij\cdots l,normal}$ and $f_{ij\cdots l,abnormal}$. Additionally, $f_{ij\cdots l,*}$ is the sum of the two aforementioned, i.e.

$$f_{ij\cdots l,*} = f_{ij\cdots l,normal} + f_{ij\cdots l,abnormal} \ . \tag{3.2.44}$$

Assuming that Equation 3.2.41 holds then the probability

$$p_{ij\cdots kl|y} = p_{j\cdots kl|y} \ . \tag{3.2.45}$$

The maximum likelihood estimate of $\hat{p}_{j\cdots kl|y}$ is

$$\hat{p}_{j\cdots kl|y} = \frac{\sum_{i=1}^{m} f_{ij\cdots l,y}}{\sum_{i=1}^{m} f_{ij\cdots l,*}} \ . \tag{3.2.46}$$

Note that the standard form for a chi-square test is:

$$Q_{j\cdots l} = \sum_{i=1}^{m} \frac{\left(O_i^{normal} - E_i^{normal}\right)^2}{E_i^{normal}} + \sum_{i=1}^{m} \frac{\left(O_i^{abnormal} - E_i^{abnormal}\right)^2}{E_i^{abnormal}} \ . \tag{3.2.47}$$

With E_i^{normal} as the expected amount of normal release for visited cell i, O_i^{normal} as the observed amount of normal release for visited cell i, and correspondingly for $E_i^{abnormal}$ and $O_i^{abnormal}$. This can be rewritten, using that

$$O_i^{normal} = f_{ij\cdots l,normal} \quad \text{and} \quad E_i^{normal} = f_{ij\cdots l,*} \cdot \hat{p}_{j\cdots l|normal} \ , \tag{3.2.48}$$

and that

$$O_i^{abnormal} = f_{ij\cdots l,abnormal} \quad \text{and} \quad E_i^{abnormal} = f_{ij\cdots l,*} \cdot \hat{p}_{j\cdots l|abnormal} \ . \tag{3.2.49}$$

The test statistic for each $Q_{j\cdots l}$ then becomes

$$Q_{j\cdots l} = \sum_{i=1}^{m} \frac{\left(f_{i\cdots l,normal} - f_{ij\cdots l,*} \cdot \hat{p}_{j\cdots l|normal}\right)^2}{f_{ij\cdots l,*} \cdot \hat{p}_{j\cdots l|normal}} + \sum_{i=1}^{m} \frac{\left(f_{i\cdots l,abnormal} - f_{ij\cdots l,*} \cdot \hat{p}_{j\cdots l|abnormal}\right)^2}{f_{ij\cdots l,*} \cdot \hat{p}_{j\cdots l|abnormal}} \ . \tag{3.2.50}$$

This can be rewritten, with $Y = \{normal,abnormal\}$ as

$$Q_{j\cdots l} = \sum_{i=1}^{m} \sum_{y \in Y} \frac{\left(f_{i\cdots ly} - f_{ij\cdots l,*} \cdot \hat{p}_{*j\cdots l|y}\right)^2}{f_{ij\cdots l,*} \cdot \hat{p}_{*j\cdots l|y}} = \sum_{i=1}^{m} \sum_{y \in Y} f_{ij\cdots l,*} \frac{\left(\frac{f_{i\cdots ly}}{f_{ij\cdots l,*}} - \hat{p}_{*j\cdots l|y}\right)^2}{\hat{p}_{*j\cdots l|y}} \ . \tag{3.2.51}$$

By setting

$$\hat{p}_{ij\cdots l|y} = \frac{f_{i\cdots ly}}{f_{ij\cdots l,*}} \ , \tag{3.2.52}$$

the equation Equation 3.2.51 can be written as

$$Q_{j\cdots l} = \sum_{i=1}^{m} \sum_{y \in Y} f_{ij\cdots l,*} \frac{\left(\hat{p}_{ij\cdots l|y} - \hat{p}_{*j\cdots l|y}\right)^2}{\hat{p}_{*j\cdots l|y}} \ . \tag{3.2.53}$$

This can be compared with Equation 3.2.31, after a comparison it can be noted that Equation 3.2.53 is just a simplification of the test given in Equation 3.2.31. Consequently, it is asymptotically chi-square distributed. The degrees of freedom are calculated just like before.

3.2.7 Asymptotic Properties

A quick explanation why the asymptotic properties hold is given below.

Consider the test statistic $Q_{j\cdots k}$ in Equation 3.2.31 for some $j, \ldots, k$. First of all note that $\hat{p}_{j\cdots k|l} > 0$ is equivalent to $f_{*j\cdots k|l} > 0$. Consequently, by assuming the $r_{j\cdots k}$ first $f_{ij\cdots k|*}$ are positive, and likewise for the $c_{j\cdots k}$ first $f_{*j\cdots k|l}$. Then all non-zero frequency counts $f_{ij\cdots k|l}$ used to compute $Q_{j\cdots k}$ can be written in the following table:

	l = 1	l = 2	$\cdots$	$l = c$
i = 1	$f_{1j\cdots k1}$	$f_{1j\cdots k2}$	$\cdots$	$f_{1j\cdots kc}$
i = 2	$f_{2j\cdots k1}$	$f_{2j\cdots k2}$	$\cdots$	$f_{2j\cdots kc}$
i = 3	$f_{3j\cdots k1}$	$f_{3j\cdots k2}$	$\cdots$	$f_{3j\cdots kc}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$i = r$	$f_{dj\cdots k1}$	$f_{dj\cdots k2}$	$\cdots$	$f_{dj\cdots kc}$

where $r = r_{j\cdots k}$ and $c = c_{j\cdots k}$. If indexes $j, \ldots, k$ are dropped and the row and column sums are added, the table instead becomes

	l = 1	l = 2	$\cdots$	$l = c$	
i = 1	f_{11}	f_{12}	$\cdots$	f_{1c}	f_{1*}
i = 2	f_{21}	f_{22}	$\cdots$	f_{2c}	f_{2*}
i = 3	f_{31}	f_{32}	$\cdots$	f_{3c}	f_{3*}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$i = r$	f_{d1}	f_{d2}	$\cdots$	f_{dc}	f_{d*}
	f_{*1}	f_{*2}	$\cdots$	f_{*c}	f_{**}

The null hypothesis that $p_{ij\cdots k|l} = p_{j\cdots k|l}$ can be written as $p_{i|l} = p_l$ when the indexes $j, \ldots, k$ are dropped. Under the null hypothesis each row are outcomes from f_{i*} trials of a multinomial distribution. That is there are f_{i*} trials, and the probability for a transition to state l is p_l. The standard chi-square test statistic for homogeneity is, with $\hat{p}_l = f_{*l}/f_{**}$

$$Q = \sum_{i=1}^{r} \sum_{l=1}^{c} \frac{\left(f_{il} - f_{i*} \cdot \hat{p}_l\right)^2}{f_{i*} \cdot \hat{p}_l} = \sum_{i=1}^{r} \sum_{l=1}^{c} f_{i*} \frac{\left(\frac{f_{il}}{f_{i*}} - \hat{p}_l\right)^2}{\hat{p}_l} . \tag{3.2.54}$$

And it is know that $Q \sim \text{asy}\chi^2((r-1)(c-1))$, see for example [Conover, 1999]. After inserting the indexes $j \ldots k$ again, this becomes

$$Q_{j\cdots k} = \sum_{i=1}^{r} \sum_{l=1}^{c} f_{ij\cdots k*} \frac{\left(\frac{f_{ij\cdots kl}}{f_{ij\cdots k*}} - \cdot\hat{p}_{j\cdots k|l}\right)^2}{\hat{p}_{j\cdots k|l}} = \sum_{i=1}^{r} \sum_{l=1}^{c} f_{ij\cdots k*} \frac{\left(\hat{p}_{ij\cdots k|l} - \cdot\hat{p}_{j\cdots k|l}\right)^2}{\hat{p}_{j\cdots k|l}} . \tag{3.2.55}$$

Which is the same as Equation 3.2.31 with reduced degrees of freedom, and consequently Equation 3.2.34 holds.

3.3 Clustering Using a Dissimilarity Matrix

3.3.1 Description of the Method

As mentioned in the Chapter 1 *(Introduction)* it was decided to use a form of clustering to solve both task A.2, and B.2. The goal is similar for both tasks. There are objects

$$A_1, A_2, A_3, \ldots, A_m \tag{3.3.1}$$

that should be divided into different classes, according to some similarity, that is, similar objects should be placed in the same class. In task A.2 the objects are sequences and in task A.2 the objects are groups of sequences. Consequently, the approach is the same for both tasks. First some suitable dissimilarity $d(A_i, A_j)$ metric is chosen. Then the different dissimilarities can be computed and put into an $m \times m$ dissimilarity matrix $\mathbf{D}$, where the elements are

$$d_{ij} = d(A_i, A_j) \ . \tag{3.3.2}$$

Then a clustering algorithm could be used to classify each of the objects in Equation 3.3.1. The PAM algorithm [Kaufman and Rousseeuw, 1987] was chosen because it is a clustering algorithm that can take an $m \times m$ dissimilarity matrix as input.

The rest of this section contain the following: first, the different dissimilarity metrics used to compare UE cell movements (groups of sequences) between time periods, after that, the dissimilarity metric used to compare single sequences with each other, and finally the PAM-algorithm.

3.3.2 Dissimilarity Metrics for Comparing Groups of Sequences

For these metrics assume there are two groups of visited cells sequences, one contains n_1 sequences, i.e. the sequences are $S_1^{(1)}$, $S_2^{(1)}$, …, $S_{n_1}^{(1)}$ and the other one contains n_2 sequences, i.e. the sequences are $S_1^{(2)}$, $S_2^{(2)}$, …, $S_{n_2}^{(2)}$. Denote the former set of sequences A_1 and the latter A_2.

When comparing groups of visited cell sequences, there is a problem. There is no existing way to determine the dissimilarity (or similarity) between different groups of visited cell sequences, at least not to the best of our knowledge. However, one possible way to compare the two sets of visited cell sequences could be to estimate the transition matrices $\hat{\mathbf{P}}^{(1)}$ and $\hat{\mathbf{P}}^{(2)}$ corresponding to both sets of sequences, and then compare the transition matrices, using a suitable dissimilarity metric. These will be introduced first.

Dissimilarity Metrics on Transition Matrices

There has been earlier work done on comparing transition matrices, one example is credit risk modelling, where the changes in a company's credit rating can be modelled as a Markov chain. Consequently, there has also been an interest to compare different credit rating transition matrices with each other. Different metrics for comparing transition matrices are investigated in for example [Trück, 2004].

One type of metrics mentioned in [Trück, 2004] are *element-by-element* metrics, of which the simplest ones are the L_1-norm and the L_2-norm, which are

$$d_1(A_1, A_2) = \sum_{i=1}^{m} \sum_{j=1}^{m} \left| \hat{p}_{ij}^{(1)} - \hat{p}_{ij}^{(2)} \right| \tag{3.3.3}$$

and

$$d_2(A_1, A_2) = \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\hat{p}_{ij}^{(1)} - \hat{p}_{ij}^{(2)} \right)^2 . \tag{3.3.4}$$

Where the estimation of $\hat{p}_{ij}^{(1)}$ and $\hat{p}_{ij}^{(2)}$ has been defined in Subsection 3.2.2. There are several other types of metrics, even more *element-by-element* based, some *eigenvalue* based, and some *eigenvector* based. However, there is a lack of information on how the transition matrices that are estimated from handovers correspond to UE cell movement. It is unsure if the eigenvalues or the eigenvectors of the transition matrix correspond to any meaningful UE cell movement behaviour. Consequently, the only metrics using the estimated transition probabilities are Equation 3.3.3 and Equation 3.3.4.

Dissimilarity Metrics on Frequency Count Matrices

It seemed reasonable to compare the dissimilarity metrics stated above with other metrics. The dissimilarity metrics found in the literature only used the transition probability matrix. An alternative metric could instead of being restricted by the transition probabilities, compare the number of transitions from cell i to cell j in the different time periods, with some general normalization (in Equation 3.3.3 and Equation 3.3.4 the normalization is the number of transitions from cell i). Because no alternative metrics were available in the literature, some alternative metrics had to be constructed. In the end four alternative dissimilarity metrics were constructed.

Before they are stated, note that the total number of transitions for time period t will be written as

$$f_{**}^{(t)} = \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij}^{(t)} , \tag{3.3.5}$$

These are the four other dissimilarity metrics: First

$$d_3(A_1, A_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left| f_{ij}^{(1)} - f_{ij}^{(2)} \right|}{f_{**}^{(1)} + f_{**}^{(2)}} , \tag{3.3.6}$$

that is, the sum of the absolute value of the difference in number of transitions from state i to state j, divided by the total number of transitions in both sets of sequences. Second

$$d_4(A_1, A_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left(f_{ij}^{(1)} - f_{ij}^{(2)} \right)^2}{\left(f_{**}^{(1)} + f_{**}^{(2)} \right)^2} , \tag{3.3.7}$$

that is, the sum of the squared value of the difference in number of transitions from state i to state j, divided by square of the total number of transitions in both sets of sequences. Additionally,

$$d_5(A_1, A_2) = \sum_{i=1}^{m} \sum_{j=1}^{m} \left| \frac{f_{ij}^{(1)}}{f_{**}^{(1)}} - \frac{f_{ij}^{(2)}}{f_{**}^{(2)}} \right| , \tag{3.3.8}$$

that is the sum of the absolute value of difference of the transitions that are done from state i to state j as a percentage of all the transitions. Finally

$$d_6(A_1, A_2) = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{f_{ij}^{(1)}}{f_{**}^{(1)}} - \frac{f_{ij}^{(2)}}{f_{**}^{(2)}}\right)^2 , \tag{3.3.9}$$

that is the sum of the squared value of difference of the transitions that are done from state i to state j as a percentage of all the transitions. Additionally, as mentioned earlier, by using the definition

$$f_{*j}^{(t)} = \sum_{i=1}^{m} f_{ij}^{(t)} , \tag{3.3.10}$$

where $f_{ij}^{(t)}$ is the number of transitions from cell i to cell j in the group of sequences A_t, Equation 3.3.3 can be written as

$$d_1(A_1, A_2) = \sum_{i=1}^{m}\sum_{j=1}^{m}\left|\frac{f_{ij}^{(1)}}{f_{*j}^{(1)}} - \frac{f_{ij}^{(2)}}{f_{*j}^{(2)}}\right| . \tag{3.3.11}$$

This can be done in the same way for Equation 3.3.4. That is Equation 3.3.3 and Equation 3.3.4 is the difference in the number of transitions from cell i to cell j in the different time periods, with a normalization.

3.3.3 Dissimilarity Metric for Comparing Single Sequences

An example of classifying sequences using a dissimilarity metric and a dissimilarity matrix can be found in [Chandola, 2009, p. 61]. There it is used to classify sequences in order to find sequences that are anomalies. Several different dissimilarity metrics are mentioned there, the one that is going to be used in this report is the normalized length of the *longest common subsequence*.

Here the definition of a subsequence is the following. S_u is a subsequence of S_i if S_u can be obtained from S_i by only removing elements in the sequence. The longest common subsequence $LCS(S_i, S_j)$ is the longest possible sequence that is a subsequence of both S_i and S_j. The normalized length of the *longest common subsequence* is the following

$$d(S_i, S_j) = 1 - \frac{|LCS(S_i, S_j)|}{\sqrt{|S_i||S_j|}} . \tag{3.3.12}$$

Where $|Z|$ is the length of the sequence Z.

Example of longest common subsequence

The sequences

$$S_1 = \text{ A, } \textbf{B}\text{, C, B, B, } \textbf{A}, \textbf{B}, \textbf{D} \tag{3.3.13}$$

and

$$S_2 = \textbf{ B}, \textbf{A}, \textbf{B}\text{, A, }\textbf{D}\text{, C, A} \tag{3.3.14}$$

has the *longest common subsequence*, **B, A, B, D**.

3.3.4 The PAM Algorithm

The PAM (*Partition around medoids*) algorithm [Kaufman and Rousseeuw, 1987] is a clustering algorithm that divides some objects $(A_1, A_2, \dots, A_m)$ into a user specified number of clusters. It takes an $m \times m$ dissimilarity matrix $\mathbf{D}$ as input. Here $\mathbf{D}$ describes the dissimilarity of the objects $(A_1, A_2, \dots, A_m)$ in the way that if $d(x, y)$ is some form of dissimilarity metric, then the entries in $\mathbf{D}$ is

$$d_{ij} = d(A_i, A_j)\,. \tag{3.3.15}$$

In the case of the objects being points in $\mathbb{R}^n$ this dissimilarity could for example be Manhattan distance or Euclidean distance.

The key concept in this algorithm is the concept of representative objects (also called medoids). First assume k representative objects $(B_1, B_2, \dots, B_k)$ are already chosen. Each representative object represents a class. All objects belong to the same representative object (or class) as the representative object it is most similar to. That is an object $A_i \in (A_1, A_2, \dots, A_m)$ is assigned to the representative object B_s, if

$$d(A_i, B_s) \leqslant d(A_i, B_t) \quad \text{for all} \quad t \neq s\,. \tag{3.3.16}$$

In the case of ties, the PAM-algorithm assigns the object to the representative object it encounters first. In order to state the actual minimization problem that the PAM-algorithm solves, the following variable z_{it} is defined as:

$$z_{is} = \begin{cases} 1 & \text{if } A_i \text{ belongs to the representative object } B_s \\ 0 & \text{otherwise} \end{cases} \tag{3.3.17}$$

Then given a desired number of k representative objects and a matrix $\mathbf{D}$ describing the dissimilarities of some objects $(A_1, A_2, \dots, A_m)$, the PAM-algorithm solves the following minimization problem:

Find the set of representative objects $(B_1^*, B_2^*, \dots, B_k^*)$ out of all possible representative objects $(B_1, B_2, \dots, B_k)$ that minimizes the function

$$\sum_{i=1}^{m} \sum_{s=1}^{k} d(A_i, B_s^*) z_{is}\,. \tag{3.3.18}$$

Then for the classification each representative object $(B_1^*, B_2^*, \dots, B_k^*)$ represents a class and each object in $(A_1, A_2, \dots, A_m)$ belongs to the same class as the representative object it is most similar to.

Chapter 4

Methodology

4.1 The Datasets

The experiments done in this report were done on different datasets. In order to get an easier overview of the properties of these datasets they will be presented below. Focus is on how many eNBs the data was gathered from, the number of visited cell sequences the datasets contained, and the total number of visited cells in all those visited cell sequences combined.

4.1.1 Dataset i

Data set i contains data collected from 12 different eNBs. It consists of 7 different 1-hour periods. A summary can be seen in Table 4.1.1.

Day	Date	Time	No. of sequences	No. of visited cells
Thursday	2017-10-26	12:00 to 13:00	11077	40771
Thursday	2017-11-02	12:00 to 13:00	9618	34773
Thursday	2017-10-26	14:00 to 15:00	3484	12274
Thursday	2017-11-02	14:00 to 15:00	4001	12578
Sunday	2017-10-22	13:00 to 14:00	5496	19915
Sunday	2017-11-12	13:00 to 14:00	5145	19564
Tuesday	2017-11-07	10:00 to 11:00	21594	81950

Table 4.1.1: Information on dataset i.

No. of visited cells	No. of cell sequences	The percentage of sequences with equal of lower No. of visited cells
2	37157	61.50 %
3	10268	78.50 %
4	4395	85.77 %
5	2491	89.90 %
6	1423	92.25 %
7 and longer	4681	100.00 %

Table 4.1.2: Overview of distribution of visited cell lengths, in all of Dataset *i* together.

4.1.2 Dataset *ii*

The second dataset is data collected from five different eNBs between the dates 2018-01-03 and 2018-01-28. This data is divided into 1-hour periods, e.g. 2018-01-03, 09:00:00 to 2018-01-03, 09:59:59. Because of of the low amount of sequences during the night, only the 1-hour time periods between 08-22 were used. Furthermore, only time periods with full data for the whole hour in all eNBs were used. Dataset *ii* contains 201 different 1-hour time periods after this filtering.

4.1.3 Dataset *iii*

The third dataset have data that is collected from a single eNB, between the dates 2018-01-02 and 2018-01-29, which is servicing two cells. This data was not divided into 1-hour periods. Special is that the sequences are not necessary unique. In total it contained 87 607 sequences, and a total of 399 657 visited cells in all sequences combined. An overview on sequence lengths can be seen below.

No. Visited Cells	No. of Cell Sequences	The percentage of sequences with equal of lower No. of visited cells
2	38442	43.88 %
3	13849	59.69 %
4	8235	69.09 %
5	5699	75.59 %
6	4160	80.34 %
7 and longer	17222	100.00 %

Table 4.1.3: Overview of distribution of sequence lengths of visited cells in dataset *iii*.

4.2 Preliminary Task

4.2.1 Finding the Order of the Markov Chain (Task C.1)

The preliminary task, i.e. task C.1, was to find a possible candidate of the order r, when modelling the sequences of visited cells as a Markov chain of order r. The data is on the format described in Subsection 3.1.1 and the test done is described in Subsection 3.2.3. For $r = 1$, it is a test of the null hypothesis that the probability of handover to cell j is independent of which cell the UE is in (cell i), i.e. 0-order, against the alternative hypothesis that the visited cell sequences are 1-order Markov chains. In short, a 0-order versus 1-order test. Secondly, for $r = 2$ it is a test of the null hypothesis that the sequences of visited cells are 1-order Markov chains, against the alternative hypothesis that they are 2-order Markov chains. In short, a 1-order versus 2-order test. And so on for $r = 3, 4, 5$, it is a 2-order versus 3-order test, a 3-order versus 4-order test, and a 4-order versus 5-order test.

The tests were done on the whole of dataset i together, for $r = 1, 2, 3, 4, 5$. The results can be seen in Table C.1.1. For this dataset some properties of the testing will be reported. Of interest when doing χ^2 test is the expected frequency counts in the tables under the null hypothesis, if they are low, the test will be misleading. There is a common rule of thumb first stated by [Cochran, 1954] that no more than 20 % of the expected count should be smaller than 5, although this is considered too harsh by some [Conover, 1999, p .202]. In [Conover, 1999] for example, it is written that the chi-square approximation is satisfactory if all expected frequency counts are larger than 0.5, and that most a larger than 1. Therefore an overview on the percentage of comparisons that is done when expected frequency counts where smaller than 5 ($f_{ij...k*} \cdot \hat{p}_{j...k|l} < 5$), can be found in Table C.1.2 in the Appendix at page 71.

Additionally, the same procedure as above was also done on each of the groups of sequences (time periods) in dataset i separately. The results can be seen in Table C.1.3 to Table C.1.9 in the Appendix, at pages 71 to 72. An overview on expected frequency counts were done as well, see Table C.1.10 in the Appendix at page 72.

4.3 Objective A: Proportion of Abnormal Release

In the introduction *Objective A* was stated, which was to find the answer to the question:

- *Determine if different types of sequences have different proportion of abnormal releases.*

Specifically, to find out if there is any connection between the UEs mobility between cells and unwanted behaviours. This would mean that different sequences of visited cells have a different percentage of abnormal behaviour. There are two ideas how to check this. The first idea is to use a form of Markov chain approach, that is task A.1. The second idea, task A.2, is to cluster the different sequences as mentioned in Section 3.3.

Here the sequences are on the form described in Subsection 3.1.2, as mentioned earlier in Section 2.4, (page 17) the sequences are labelled *normal* or *abnormal* depending on release status and handover result.

4.3.1 Using the Markov Chain Assumption (Task A.1)

In this section the connection between the UEs visited cells history and abnormal releases is examined backwards in time, the test can be found in Subsection 3.2.6.

For each $r = 1, 2, 3, 4$ the following test is performed:

The null hypothesis that the probability of an abnormal state is homogeneous for all sequences with the same r latest visited cells is tested against the alternative hypothesis that the probability of an abnormal state is dependent on the $r + 1$ latest visited cells. If the p value is smaller than 0.05 the null hypothesis will be rejected.

That is, the approach is straightforward, first to see if it is possible that the probability is independent of the most recently visited cell, secondly to see if it is possible the probability is independent of the two most recently visited cells, and so on. In this task a new cell state is introduced: *NO-CELL*. If the visited cell sequence is shorter than the length of visited cells used in the hypothesis, the padding *NO-CELL* is added.

The test was done on dataset *iii*, the results can be seen in Table A.1.1, and the corresponding information can be seen in Table A.1.2. The total number of sequences used was 87 607. Dataset *iii* was used mainly to get a high frequency counts for the longer sequences. A short example on one of the sub-tables can be seen in Table 4.3.1.

$X_{t-2} = i, X_{t-1} = j, X_t = k$	$f_{ijk,norm.}$ No. norm.	$f_{ijk,abnorm.}$ No. abnorm.	$f_{ijk,*} \cdot \hat{p}_{jk\|norm.}$ No. norm. exp.	$f_{ijk,*} \cdot \hat{p}_{jk\|abnorm.}$ No. norm. exp.
NO-CELL, 3043, 2786	7560	611	7578.35	592.65
2786, 3043, 2786	352	3	329.25	25.75
3053, 3043, 2786	139	14	141.90	11.10
4864, 3043, 2786	3	2	4.64	0.36
2796, 3043, 2786	2	0	1.85	0.15

Table 4.3.1: An example of a Q_{jk}. Here for $Q_{3043,2786}$

Filtering the Data

It will be discussed in greater detail why filtering is needed see Chapter 5 *(Discussion)*. There was a problem with the expected frequency counts $f_{ij\cdots k,*} \cdot \hat{p}_{j\cdots k|y}$ being low. Remember the rule of thumb from [Cochran, 1954] that no more than 20% of the expected frequency count being lower than five. Therefore, the tests were done on filtered data as well. The filtering is quite simple, all combinations of sequences containing the five most recently visited cells, with counts smaller than 10, were excluded, i.e. where $f_{ijklm} < 10$.

The test was done on dataset *iii* as well, the results can be seen in Table A.1.3, corresponding information can be seen in Table A.1.4. The total number of sequences used after the filtering was 62 554.

4.3.2 Using Clustering of Sequences (Task A.2)

The second task A.2 was also tested. First all sequences from the time period 2017-10-26, 12-13 in dataset i was extracted. The dissimilarity matrix of all sequences was computed as described in Subsection 3.3.1 using the dissimilarity metric described in Subsection 3.3.3. This created a $11\,077 \times 11\,077$ matrix. All sequences were then divided into $k = 2, 3, \ldots, 10$ different classes using the PAM-algorithm. A standard homogeneity test was then performed on the number of sequences that are labelled normal and abnormal for each k. Table A.2.1 contains the number of clusters, the test statistic, the degrees of freedom, and the probability of acquiring such an extreme test statistic if the null hypothesis of all classes having the same probability of abnormal behaviour is true.

4.4 Objective B: Changes in UEs Cell Movement

In the introduction Objective B was also stated, which was to find the answer to the question:

- *Develop a method to detect changes in patterns of UE movements between cells.*

The methods to solve this objective was briefly stated in Subsection 1.3.2 *(Tasks for Objective B)*. The method will be described more thoroughly in this section, with references to relevant parts in the earlier theory chapter.

4.4.1 Evaluating the Homogeneity Test (Task B.1)

The first task was B.1. Which was *"Construct a homogeneity test using Markov chains assumption"*. This is the same as testing the null hypothesis that the two groups of sequences from the different time periods are outcomes of the two 1-order Markov chains with equal transition probabilities, against the alternative hypothesis that they are outcomes of 1-order Markov chains with different transition probabilities. The test for this was described in Subsection 3.2.3 *(Homogeneity Test on Markov Chains)*.

Recall, this task consisted of two subtasks:

Task B.1.1: Do a validation of the statistical test using simulated data.

Task B.1.2: Do the test on real network data and see if it performs well.

The methodology for task B.1.1 is explained first, with task B.1.2 after.

Validation of the Homogeneity Test

First of all, the decision to add task B.1.1 should be explained more thoroughly. It was noted that a lot of the estimated transition probabilities from the groups of sequences from the different time periods were very small. It is possible that the asymptotic properties described in Subsection 3.2.3 would not be fulfilled. Mostly because some of the more unlikely transitions occur very seldom, which leads to low frequency counts in the tests. These low frequency counts are a problem because they could lead to poor asymptotic properties for the test statistic. Therefore, it seemed prudent to make some form of validation of the test, and see if it works for some testing data that has the same properties as the real data.

The following validation was performed. First some validation data was generated. This was done using actual network data to generate a transition matrix $\mathbf{P}$, and initial probabilities p_0, using maximum likelihood estimates.

Then P and p_0 were used to create two sets of sequences, denoted A_1 and A_2. Then these two sets of sequences were used as input for the test described in Subsection 3.2.3. This was done using a test strength of 5%. If the test is correct it should reject the hypothesis around 5% of the time (the hypothesis is true by design).

The validation was done for different numbers of total transitions, that is the total number of transitions in both of the sets of sequences A_1 and A_2, was n where

$$n \in (100, 200, 500, 1000, 2000, \ldots, 14\,000, 15\,000)\,. \tag{4.4.1}$$

This was done a 1000 times for each n. The estimation of the true rejection rate is the same as estimation of α in a $Bin(\alpha, 1000)$ stochastic variable. Because there is a large number of tests, a normal approximation can be used and an estimation of the 95% confidence interval of the true rejection rate is

$$\hat{\alpha} \pm 1.96\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{1000}}\,, \tag{4.4.2}$$

with $\hat{\alpha} = s/1000$ where s is the total number of successes of the 1000 trials. By comparing the actual rejection rate with the 5% specified, a quick check on the validity of the test would be obtained. For example, if the test always succeeds or always fail there is a definite problem with the test.

The validation of the homogeneity test described above was done, the result can be seen in Table B.1.1. In Figure B.1.1 the distribution of the test statistic can be seen. The validation was done using the time period 2017-10-26, 12:00 to 13:00 in dataset i.

Homogeneity Test on Real Network Data

In order for the homogeneity test to be useful, it should match the assumption that UE cell movement is more similar during the same time of day. And furthermore, time periods during weekdays should be more similar to each other in contrast to the weekend.

Therefore, the homogeneity test should not reject that time periods, which are on the same time of day and same day of the week, have the same transition probabilities. Moreover, it should reject that some time periods from different days of the week and different times of day have the same transition probabilities. The test is not going to be useful if these conditions are not fulfilled.

Consequently, this test was done on dataset i. It contained a pair two Thursdays 12-13, which should have the same UE cell movement, and a pair of two Thursdays 14-15, which also should have the same UE cell movement, and a pair of two Sundays 13-14, which also should have the same cell movement. In all these cases the null hypothesis should not be rejected. Other combinations of time periods should be rejected, at least most of the time.

Three tables were produced to summarise the results of these tests, they can be found on page 61. Table B.1.2 contain the test statistic Q from Equation 3.2.13 for all possible combinations of the test. Moreover, Table B.1.3 contains the degrees of freedom k from Equation 3.2.15 for all possible combinations of the test. Finally Table B.1.4 contains the probability that such a large test statistic Q would be obtained if the hypnosis that the two time periods have Markov chains with the same transition probabilities was true. If the p value is smaller than 0.05 the null hypothesis will be rejected.

4.4.2 Evaluating Possible Dissimilarity Metrics (Task B.2)

The Dissimilarity Matrices

The analysis on the dissimilarity metrics mentioned in Subsection 3.3.2, was done using the idea mentioned in Subsection 1.3.2 *(Tasks for Objective B)*. This evaluation needs a lot of 1-hour periods for the clustering, additionally it needs as many eNBs as possible. Since all eNBs do not have data from all possible time periods, a compromise between getting as many 1-hour time periods and eNBs as possible had to be done, this compromise is dataset *ii*, which is the dataset used for this analysis.

The dissimilarity metrics between each of the 201 different 1-hour timer periods, were computed, creating the dissimilarity matrix $\mathbf{D}$, where each element i, j is the dissimilarity metric between the i-th and j-th hour periods. Note that due to symmetry only the lower triangle of the matrix is needed. Because the matrix is large (201×201) only a part of two of the matrices are shown in the report, the two matrices can be seen in the Appendix, in Table B.2.1, (page 62), and Table B.2.2, (page 63) respectively.

Clustering Using Dissimilarity Matrices

Then next step is clustering using the PAM algorithm (Subsection 3.3.4) with the 6 different dissimilarity matrices mentioned in the previous subsection. It was decided that the PAM algorithm should divide the different time periods into four classes. The first part of the clustering result is shown in Table B.2.3. The rest is excluded because it would be too large.

Summarising the Clusterings

Secondly all the clusterings were summarised to get an overview of the data. This was done by first dividing the clusters into their corresponding one-hour-period (06 - 07, 07 - 08, ...), and then arranging them after the number of handovers that each set of sequences had. Then display this in tables, with one-hour-periods corresponding classes and corresponding number of handovers. The tables summarising each clustering can be seen in Table B.2.4 to Table B.2.9.

Chapter 5

Discussion

This chapter contains a discussion of the results from Chapter 4 *(Methodology)*. It starts with a discussion on the preliminary task (the order of the Markov chain). Afterwards there is a discussion on *Objective A* and *Objective B*. Finally, there is a summary of the conclusions.

5.1 On the Preliminary Task, The Order of the Markov Chain

This task (C.1) was checking the plausibility of the handovers between cells being modelled as a Markov chain of order r. The test is described in Subsection 4.2.1 and the results can be found in the Appendix in Section C.1.

When all the data in dataset i was combined and tested together, the null hypotheses of a Markov chain of order 0, 1, 2, 3, and 4 could be rejected in favour of the alternative hypotheses of a Markov chain of order 1, 2, 3, 4, and 5 respectively.

Furthermore, when the different one-hour periods in dataset i was tested separately in Section C.1, the results were the same. The null hypotheses of a Markov chain of order 0, 1, 2, 3, and 4 could be rejected in favour of the alternative hypotheses of a Markov chain of order 1, 2, 3, 4, and 5 respectively.

Observations

First of all, in the data it can be seen that although the test statistic is very extreme if the null hypotheses are true, it is less extreme for higher r. For larger degrees of freedom k a $\chi^2(k)$, is approximately normally distributed with mean k and standard deviation $\sqrt{2k}$. When testing on all time periods together, the test statistic is approximately: $k + 15621\sqrt{2k}$ for 0 versus 1, $k + 1998\sqrt{2k}$ for 1 versus 2, $k + 282\sqrt{2k}$ for 2 versus 3, $k + 121\sqrt{2k}$ for 3 versus 4, and $k + 85\sqrt{2k}$ for 4 versus 5. That is there are large improvements when going from a 0-order Markov chain to a 1-order, and from a 1-order Markov chain to a 2-order Markov chain, after that the improvements are much smaller.

Secondly in Table 4.1.2 it can be noted that most sequences of visited cells are quite short. For example, 78.50 % of all sequences contain 2 or 3 visited cells. There is little use to model the sequences as a 3-order Markov chain when the sequences are shorter than that.

Furthermore, in Table C.1.2 and Table C.1.10 it can be seen that a lot of the comparisons in the tests were done, when the expected frequency counts $f_{ij\ldots k*} \cdot \hat{p}_{j\ldots k|l}$ were lower than 5. The percentage of expected frequency counts that are less than five, is much larger than the 20 % recommended by the rule of thumb. Consequently, all result should be viewed with some scepticism.

A possibility to using χ^2 tests is to compute the probability of exactly. This can for example be done using an algorithm by [Mehta and Patel, 1983], which is implemented several different statistical programs, see [Kroonenberg and Verbeek, 2018, p. 3]. One such algorithm is *fisher.test* in R.[1] The aforementioned could not be done due to the computations being too complex. For example, the order 0 versus 1 test using all dataset i together, would have to be made on a 265×258 table.

The key take-away from this is that the probability of handover for UEs between cells is dependent on the history of earlier visited cells. However, the tests for a Markov chain of a larger order is unsure, due to the problem with many entries having a small expected frequency count.

5.2 On Objective A

Recall, *Objective A* was:

Determine if different types of sequences have different proportion of abnormal releases.

In this section task A.1 will be discussed first, and A.2 will be discussed afterwards.

5.2.1 On Task A.1, Dependency of the r Most Recently Visited Cells

Unfiltered Data

Recall, for each r:

The null hypothesis that the probability of an abnormal state is homogeneous for all sequences with the same r *latest visited cells is tested against the alternative hypothesis that the probability of an abnormal state is dependent on the* $r+1$ *latest visited cells.*

The conclusions from Table A.1.1 is that the null hypothesis above, i.e. that the probability being homogeneous for all sequences with the same r latest visited cells, can be rejected for all r. That is, $r = 1$, $r = 2$, $r = 3$, and $r = 4$, can all be rejected, at a level of 5 %.

There is a caveat however. In Table A.1.2 it is seen that most of the combinations of sequences $X_{T-r}, \ldots, X_{T-1}, X_T, Y_T$ are uncommon. Consequently, many of the expected frequency counts $f_{ij\cdots l,*} \cdot \hat{p}_{j\cdots l|normal}$ and $f_{ij\cdots l,*} \cdot \hat{p}_{j\cdots l|abnormal}$ are less than 5. Furthermore, it is not shown in the table due to space restriction, but most of those expected frequency counts that less than 5, are less than 1. One example: When testing the hypothesis for homogeneity with $r = 3$, 86.54 % of the labelled sequences has the expected frequency counts being less than 5, and 64.42 % of the labelled sequences has the expected frequency counts being less than 1.

[1] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/fisher.test.html

Filtered Data

As a consequence of the problems described above, the test was done on filtered data as well. The filtering is quite simple. All combinations of sequences containing the five most recently visited cells, with frequency counts being smaller than 10, were excluded, i.e. where $f_{ijklm} < 10$. No consideration was shown to the proportion of Abnormal or Normal in the frequency counts. Additionally remember than i, j, k, l are filled out with the cell *"NO-CELL"* if the visited sequence is too short. In total around 30 % of the total data was excluded. For the filtered data the following test is now done for each r:

The null hypothesis that the probability of an abnormal state is homogeneous for all sequences, with the same r latest visited cells, where there are at least 10 instances of the sequence, is tested against the alternative hypothesis that the probability of an abnormal state is dependent on the $r+1$ latest visited cells.

The conclusions from Table A.1.3 is that the hypothesis above of the probability being homogeneous for all sequences with the same r latest visited cells, can be rejected for $r = 1$, $r = 2$, and $r = 3$ can all be rejected, at a level of 5 %. However, the hypotheses that $r = 4$ cannot be rejected for filtered sequences.

As a result of the filtering, it can be seen in Table A.1.4 that more of the expected frequency counts $f_{ij\cdots k,*} \cdot \hat{p}_{j\cdots k|normal}$ and $f_{ij\cdots k,*} \cdot \hat{p}_{j\cdots k|abnormal}$ are now larger than 5. Additionally, it can be noted that the number of sequences that can be used for the analysis is smaller in the filtered dataset, especially for higher r.

Observations

Some observations can be made. First, it can be noted that the most common source for Abnormal release status in dataset *iii* is *Handover-Desirable-For-Radio-Reasons*, at 7.31 % of all sequences. The second largest source for Abnormal releases is *Failed-Trrcconnectionreconfiguration-Expired* at 2.08 %. It could for example be possible that only *Handover-Desirable-For-Radio-Reasons* is dependent on the history of visited cells and that the other release status are independent.

Furthermore, it can also be noted, that just like in dataset i, most of the sequences in dataset *iii* are quite short, 60 % contain 2 or 3 visited cells.

Due to the large amount of counts being lower than 5 in the unfiltered data, the rejection of the hypothesis of homogeneity for the probability given the last r visited cells, should be viewed with some scepticism. So the main conclusion is that the probability of abnormal drops for more common sequences is dependent on the history of the last 3 visited cells. The fact that it cannot be rejected that the 4th most recently visited does not matter for common sequences, combined with that 60 % percent of sequences are 2 or 3 visited cells long, makes it reasonable that the probability of abnormal release can be modelled as only dependent on the 3 most recently visited cells.

5.2.2 On Task A.2, Different Proportions Using Clustering of Sequences

The first thing that can be concluded from the results in Table A.2.1 is that the null hypotheses of homogeneity can be rejected for $k = 7, 8, 9, 10$ with a confidence level of 5%.

Observations

As mentioned in the introduction this became more of an extra task. The reason that this approach was not used more was because it was slow and memory intensive, no specific benchmark on time complexity was done, but it is significantly slower than using the Markov assumption. It can be noted that this method uses an $m \times m$ distance matrix, where m is the number of sequences. So for this analysis the matrix contained $11077 \cdot 11077 \approx 1.227 \cdot 10^8$ elements, which used a disk space of 468.6 MB. If this analysis would be done on a dataset on the same size as dataset *iii*, the matrix would contain $87607 \cdot 87607 \approx 7.675 \cdot 10^9$ elements, which is close to 65 times larger than the matrix used. It is also worth mentioning that this is not a sparse matrix, so there is not sparsity to use.

5.3 On Objective B

5.3.1 On Task B.1, Usability of the Homogeneity Test

Validation

The validation in Table B.1.1 shows that the method should work for real data. The estimation of the real rejection rate of the test is around 10 %. That is, the hypothesis that the two groups of sequences have the same transition probabilities is rejected as implausible 10% of the time. In Figure B.1.1 it is seen that the test statistic is normally distributed, this is because the chi-square distribution becomes approximately normally distributed when the degrees of freedom is large. The main conclusion is that the tests seem good enough to be used on the real datasets.

Real Data

From the test on real data, it can be concluded that the hypothesis that the different time periods have the same transition probabilities can be rejected, for tests between all 1-hour time periods in dataset i. Furthermore, note that the test statistic is very large for all tests, much larger than the degrees of freedom.

The conclusion from this, is that a test of homogeneity cannot be used to determine if there have been changes to the UEs movements between cells after software updates. The problem is that the test of homogeneity is always going to reject the null hypothesis that the transition probabilities are the same.

Observations

Note that the test described in Subsection 3.2.5 can be used to test homogeneity for higher order Markov chains. That is, test the null hypothesis that two groups of sequences are r-order Markov chains with the same transition probabilities against the alternative hypothesis that the groups of sequences describe r-order Markov chains with different transitions probabilities. A homogeneity test could be done assuming the data describes Markov chains of order 2, 3, 4. Recall from the preliminary task that is better to model the visited cell sequences as Markov chains of a higher order. However, this was not done in this thesis, the main reason was that it would probably lead to the tests being done when the expected frequency counts are low. The different time periods in dataset i contain around 200 different cells. If the data was modelled as a 2-order Markov chains, there would be $200 \cdot 200 = 40\,000$ possible states.

5.3.2 On Task B.2, A Suitable Dissimilarity Metric

For task B.2, the central part is to discuss task B.2.3, i.e. how well the classifications for the different dissimilarity metrics matches with the knowledge about with time periods that have similar movement between cells. These are Table B.2.4 through B.2.7, at page 65 through 68.

The first two alternatives is when the metrics using the sum of the absolute or squared difference in conditional probability. That is Equation 3.3.3 and Equation 3.3.4, with the classification of the time periods in Table B.2.4 and Table B.2.5 respectively. Both those approaches did not lead to a classification that matches the previous knowledge at Ericsson. This is easily seen because the time periods which are on the same time of day are not in the same classes at all.

The third alternative is when the metrics were the sum of the absolute difference in the number of handovers, divided by the total number of handovers in both sets of sequences. That is Equation 3.3.6, with the classification of the time periods in Table B.2.6. This one seems to work better. However, there is a problem, by comparing the classes with the total number of visited cells in the time period, it can be seen that this classification is equivalent to classifying using total number of visited cells. This can be seen even clearer by ordering Table B.2.3 by the number of visited cells, unfortunately this would take up a lot of space in this report and consequently it is excluded.

The fourth alternative is when the metrics was the sum of the squared difference in the number of handovers, divided by the total number of handovers in both sets of sequences. That is Equation 3.3.7, with the classification of the time periods in Table B.2.7. Here the problem is the same as above, it is basically just a classification on the number of handovers in the time period.

The fifth and sixth alternatives are when the metric were the sum of absolute or squared difference of the percent of total handovers that are handovers from cell i to cell j. That is Equation 3.3.8 and Equation 3.3.9, with the classification of the time periods in Table B.2.8 and Table B.2.9 respectively. These two metrics seemed to fit the best with the ideas at Ericsson. They divided the day into three main parts: Late Morning & Afternoon, Lunch, and Early morning & Evening. This division into classes is not the same as classifying by the number of handovers in the period.

The main conclusion here is that Equation 3.3.8 or Equation 3.3.9 can be used to compare time periods.

Observations

As mentioned earlier the null hypothesis that the sequences of visited cells are 1-order Markov chains was rejected. Therefore, it could be argued that this should be reflected in how the clustering of time periods was done. This was not done, the main reason for this is because the 1-hour time periods contain around 200 different cells. If the data was modelled as a second-order Markov Chains, there would be $200 \cdot 200 = 40\,000$ possible states. Which it turn would lead to trying to compare $40\,000 \times 40\,000$ matrices.

5.4 Summary of Conclusions

A summary on the conclusions is done in order to get a clear view of them.

5.4.1 Summary on Conclusions for the Preliminary Task

Recall that the preliminary task was:

Task C.1 (Preliminary) **:** *Find the order of the Markov chain.*

In the end no good candidate order for the Markov chain was found. There is a problem with the accuracy of the tests, this will likely not be improved much by more data. It is important to note however that the sequences of visited cells can still be modelled as a Markov chain of order r, with the caveat that this model does not fit that well with the actual data.

5.4.2 Summary on Conclusions for Objective A

Recall that *Objective A* was

Determine if different types of sequences have different proportion of abnormal releases.

The two tasks used to solve this objective was:

Task A.1: Using a Markov assumption, the probability of abnormal behaviour is dependent on the r latest visited cells.

Task A.2: Divide sequences into classes using a dissimilarity measure.

For task A.1 it can be concluded the 3 most recently visited cells matter for the probability of an abnormal release for more common sequences of visited cells. Furthermore, for task A.2 it can be concluded that the method works. However, the method is slow and memory intensive, consequently the other approach is preferred to this one. In general, it can be concluded certain sequences of visited cells have a larger probability of abnormal behaviour. This is highly interesting when trying to find the cause of abnormal behaviour.

5.4.3 Summary on Conclusions for Objective B

Recall that *Objective B* was:

Develop a method to detect changes in patterns of UE movements between cells.

The two tasks used to solve this objective was:

Task B.1: Construct a homogeneity test using Markov Chains assumption.

Task B.2: Find a suitable dissimilarity metric.

Firsts, for task B.1 it can be concluded that the approach does not work. It is not possible to use a homogeneity test to detect changes in UE movement between cells. As a consequence of task B.1 not working, it can be concluded that a dissimilarity metric has to be used to detect changes. In task B.2 it can be concluded that Equation 3.3.8 and Equation 3.3.9 can be used as dissimilarity metrics. These metrics could be used to detect changes in UE movement, after for example software updates, more on how this could be done is in Chapter 6 *(Future Research).*

Chapter 6

Future Research

In this chapter some ideas about future research are described.

6.1 Modelling Cell Sequences

There could be future research done on how to model the resulting visited cell sequences from the UEs handovers between cells. Some ideas are stated below.

6.1.1 As a Continuous-time Markov Process

One example, could be to model the sequence as a continuous-time Markov process instead of the discrete Markov Chain used in this thesis. This can be quite easily done due to the message sent with the *Handover Request* containing the time spent in each cell. There are three caveats to consider however. The time recorded is in rounded seconds, and a lot of those are 1, 2 or 3. Also note that the rounding is different depending on if the time spent in each cell is larger or smaller than 1 (see Section 2.5). Another problem is that it would be harder to incorporate previously visited cells in the transitioning probabilities, that is the same problem as in Subsection 6.1.2. Therefore it is not entirely sure how good it is to model the sequences as a continuous-time Markov process, it would probably be better to used the idea mentioned Subsection 6.1.2.

6.1.2 As a Markov Chain, Using Time Spent in Each Cell

Another possibility would be to use the time spent in each cell when modelling visited cell sequences as a Markov chain. The idea is to model the sequences as spending 1 second in each cell. After each second the UE does a cell transition, most of the time this would be to the same cell as the one the UE is currently in. One of the problems with this idea is that it would be harder to incorporate previously visited cells in the transition probabilities.

6.2 More General UE Cell Movement Behaviour

Furthermore, there are possibilities for future research on the proportion of abnormal releases for visited cell sequences. For example, this thesis differentiates between cell identification, that is, there is a difference between

$$((\mathit{NO\text{-}CELL} \rightarrow 2786 \rightarrow 3043 \rightarrow 2786), (\mathit{normal})) \tag{6.2.1}$$

and

$$((\mathit{NO\text{-}CELL} \rightarrow 3043 \rightarrow 2786 \rightarrow 3043), (\mathit{normal}))\ . \tag{6.2.2}$$

Note that both these sequences have the same pattern. They start in one cell, have a handover to a new cell and the have a handover back to the old cell. Consequently, they could both be considered to be sequences on the form

$$((\mathit{NO\text{-}CELL} \rightarrow A \rightarrow B \rightarrow A), (\mathit{normal}))\ . \tag{6.2.3}$$

Sequences could be classified in this way, and then the Markov assumption method could be used to for example find if certain types of movement coincide with different proportion of abnormal releases. One typical example would be determining if ping-pong behaviour (handovers back and forth between two cells) coincides with a higher number of abnormal releases.[1]

Additionally, the size of the cell could be included in some way, this information is given in the *Handover Request* message, see Appendix D.

6.3 Using the Dissimilarity Metric to Find Changes After Software Updates

Another area of future research is to use the results from task B.2, and construct a method to determine changes in UE cell movement.

One example could be to construct a set of reference time periods for each 1-hour time period. The set of reference time periods could for example be all such 1-hour time periods for a month. A new time period could be compared to each of its reference time periods using a dissimilarity metric, with a mean dissimilarity being computed out of all of those comparisons. This mean dissimilarity could be computed for each new time period creating a time series of dissimilarities. If there for example is a large jump in all time series (each 1-hour time period of the day has its own reference time period) after a software updates, it can be concluded that the software update lead to changed UE cell movement.

[1] Some preliminary work was done on this. In general sequences with ping-pong behaviour ($\ldots \rightarrow B \rightarrow A \rightarrow B \rightarrow A$), had a lower proportion of abnormal releases. That is, ping-pong behaviour does not coincide with a higher abnormal behaviour.

Bibliography

Theodore W Anderson and Leo A Goodman. Statistical inference about markov chains. *The Annals of Mathematical Statistics*, pages 89–110, 1957.

Frank Bickenbach and Eckhardt Bode. Evaluating the markov property in studies of economic convergence. *International Regional Science Review*, 26(3):363–392, 2003.

Varun Chandola. *Anomaly detection for symbolic sequences and time series data.* University of Minnesota, 2009.

William G Cochran. Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4): 417–451, 1954.

W. J Conover. *Practical nonparametric statistics.* John Wiley, New York ; Chichester, 3. ed.. edition, 1999. ISBN 0-471-16068-7.

Christopher Cox. *An introduction to LTE, LTE-advanced, SAE, VoLTE and 4G mobile communications.* Wiley, 2014.

Leonard Kaufman and Peter Rousseeuw. *Clustering by means of medoids.* North-Holland, 1987.

P. M. Kroonenberg and Albert Verbeek. The tale of cochran's rule: My contingency table has so many expected values smaller than 5, what am i to do? *The American Statistician*, 0(0):1–9, 2018. doi: 10.1080/00031305.2017.1286260. URL `https://doi.org/10.1080/00031305.2017.1286260`.

Cyrus Mehta and Nitin Patel. A network algorithm for performing fisher's exact test in r × c contingency tables. 78:427–434, 06 1983.

Guowang Miao, Jens Zander, Ki Won Sung, and Slimane Ben Slimane. *Principles of cellular systems.* Cambridge University Press, 2016. doi: 10.1017/CBO9781316534298.006.

S Trück. Measures for comparing transition matrices from a value-at-risk perspective. *University of Karlsruhe*, 2004.

Appendices

Appendix A

Results for Objective A, Proportion Abnormal

A.1 Markov Assumption

Unfiltered data

Hypothesis	Test statistic	*df*	p
$P(Y_T \vert X_T,\ldots,X_0) = P(Y_T \vert X_T)$	5322.22	108	0.00
$P(Y_T \vert X_T,\ldots,X_0) = P(Y_T \vert X_T, X_{T-1})$	2704.53	1511	0.00
$P(Y_T \vert X_T,\ldots,X_0) = P(Y_T \vert X_T,\ldots,X_{T-2})$	5618.61	4666	0.00
$P(Y_T \vert X_T,\ldots,X_0) = P(Y_T \vert X_T,\ldots,X_{T-3})$	5270.09	4702	0.00

Table A.1.1: Test of homogeneity that the behaviour of a visited cell sequences ends in an abnormal state, for different conditional probabilities. That is the experiment described in Section 4.3. Total number of sequences are 87607.

Type of Sequence	Total entries*	Expected count ≤ 5, in %**	Used sequences***
X_{T-1}, X_T, Y_T	220	34.55 %	87607
$X_{T-2}, X_{T-1}, X_T, Y_T$	3192	73.25 %	87490
$X_{T-3}, X_{T-2}, \ldots, X_T, Y_T$	10516	86.54 %	46287
$X_{T-4}, X_{T-3}, \ldots, X_T, Y_T$	11446	93.27 %	23095

Table A.1.2: Corresponding information table to Table A.1.1.
* The number of different versions of $ij \cdots kl, y$, where there are at least two different i's.
** The percentage of those aforementioned entries that have a expected frequency count where $f_{ij\cdots k,*} \cdot \hat{p}_{j\cdots k|y} < 5$.
*** The number of sequences where the condition stated in * is fulfilled.

Filtered data

Hypothesis	Test statistic	df	p
$P(Y_T\|X_T,\ldots,X_0) = P(Y_T\|X_T)$	4564.93	57	0.00
$P(Y_T\|X_T,\ldots,X_0) = P(Y_T\|X_T,X_{T-1})$	855.09	210	0.00
$P(Y_T\|X_T,\ldots,X_0) = P(Y_T\|X_T,\ldots,X_{T-2})$	252.21	188	0.00
$P(Y_T\|X_T,\ldots,X_0) = P(Y_T\|X_T,\ldots,X_{T-3})$	238.20	213	0.11

Table A.1.3: Test of homogeneity that the behaviour of a visited cell sequences ends in an abnormal state, for different conditional probabilities. That is the experiment described in Section 4.3. This time with filtered data, that is only done on sequences with more than 10 occurrences. Total number of sequences after filtering are 62554.

Type of Sequence	Total entries*	Expected count ≤ 5, in %**	Used sequences***
X_{T-1}, X_T, Y_T	118	12.71 %	62554
$X_{T-2}, X_{T-1}, X_T, Y_T$	518	27.22 %	62364
$X_{T-3}, X_{T-2}, \ldots, X_T, Y_T$	564	39.01 %	21972
$X_{T-4}, X_{T-3}, \ldots, X_T, Y_T$	642	46.88 %	10572

Table A.1.4: Corresponding information table to Table A.1.3.
* The number of different versions of $ij\cdots kl, y$, where there are at least two different i's.
** The percentage of those aforementioned entries that have a expected frequency count where $f_{ij\cdots k,*} \cdot \hat{p}_{j\cdots k|y} < 5$.
*** The number of sequences where the condition stated in * is fulfilled.

Occurrences of release statuses

Missing release cause	Session outcome	Label	Percentage
YES	SUCCESSFUL-HANDOVER (8)	Normal	57.75 %
YES	USER-INACTIVITY (10)	Normal	30.71 %
YES	HANDOVER-DESIRABLE-FOR-RADIO-REASONS (26)	Abnormal	7.31 %
NO	FAILED-TRRCCONNECTIONRECONFIGURATION-EXPIRED (1)	Abnormal	2.08 %
YES	RELEASE-DUE-TO-EUTRAN-GENERATED-REASON (9)	Abnormal	0.82 %
NO	HANDOVER-CANCELLED (4)	Normal	0.35 %
YES	Missing-Acknowledge	Abnormal	0.27 %
YES	RADIO-CONNECTION-WITH-UE-LOST (11)	Abnormal	0.16 %
YES	NORMAL-RELEASE (1)	Normal	0.16 %
NO	SUCCESSFUL (0)	Normal	0.11 %
YES	CS-FALLBACK-TRIGGERED (5)	Normal	0.10 %
YES	TX2RELOC-OVERALL-EXPIRY (30)	Abnormal	0.09 %
YES	DETACH (3)	Normal	0.04 %
YES	HANDOVER-FAILURE-IN-TARGET-EPC-ENB-OR-TARGET-SYSTEM (17)	Abnormal	0.03 %
NO	FAILED-TIME-OUT-OF-PATH-SWITCH-REQUEST (3)	Abnormal	0.01 %
NO	Na	Normal	0.01 %
YES	TS1RELOC-OVERALL-EXPIRY (19)	Abnormal	0.00 %

Table A.1.5: This is an overview of the actual labelling, if the release cause is missing in the data, the result of the handover is examined. Most of the time, missing release means that the handover was cancelled or that the handover failed.

A.2 Clustering of Sequences

Number of clusters	Test statistic	*df*	*p*
2	2.69	1	0.10
3	2.46	2	0.29
4	2.50	3	0.47
5	8.01	4	0.09
6	8.03	5	0.15
7	16.55	6	0.01
8	20.54	7	0.00
9	22.49	8	0.00
10	30.50	9	0.00

Table A.2.1: Standard homogeneity test that the sequences in all clusters have equal probabilities of abnormal behaviour.

Appendix B

Results for Objective B, Changes in UE Cell Movement

B.1 Homogeneity Test

B.1.1 Validation of the Homogeneity Test

n	Estimate of real rejection rate
15 000	0.091 ± 0.018
14 000	0.105 ± 0.019
13 000	0.087 ± 0.017
12 000	0.100 ± 0.019
11 000	0.120 ± 0.020
10 000	0.095 ± 0.018
9000	0.111 ± 0.019
8000	0.143 ± 0.022
7000	0.166 ± 0.023
6000	0.155 ± 0.022
5000	0.180 ± 0.024
4000	0.212 ± 0.025
3000	0.225 ± 0.026
2000	0.263 ± 0.027
1000	0.298 ± 0.028
500	0.209 ± 0.025
200	0.121 ± 0.020
100	0.045 ± 0.013

Table B.1.1: Here, n is the number total number of visited cells. The test was done a 1000 times for each n.

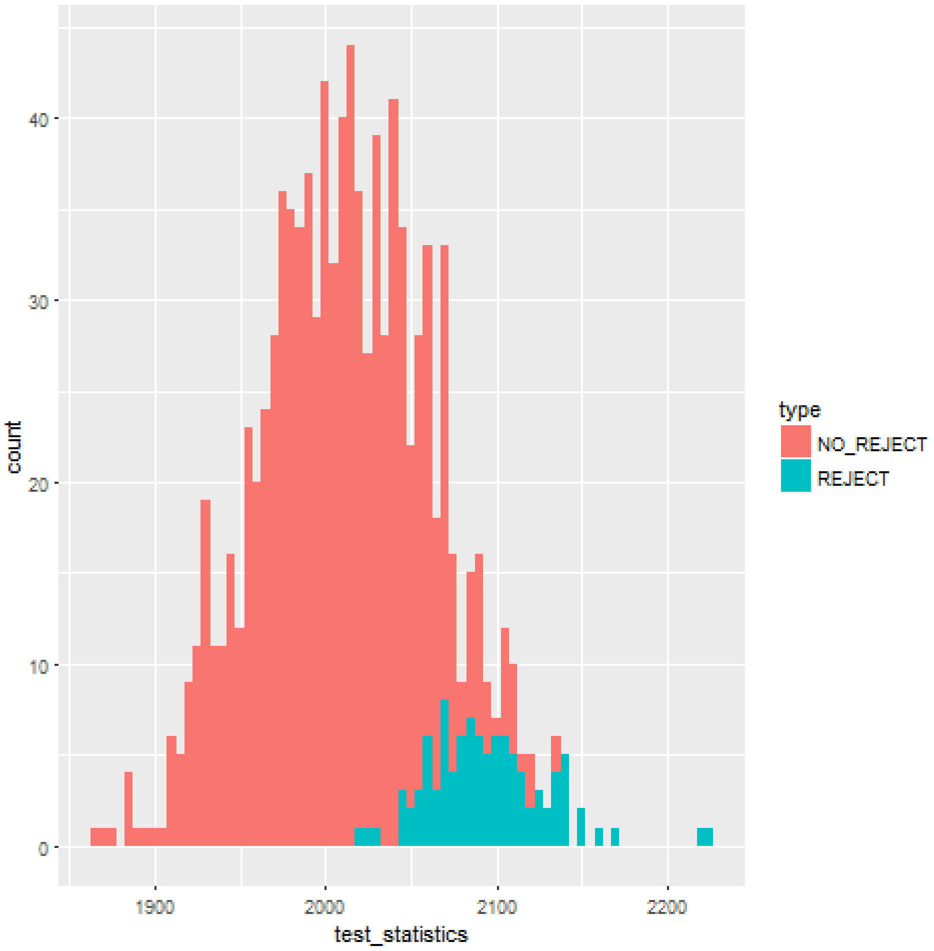

Figure B.1.1: Distribution of test statistic for the homogeneity test described in Subsection 4.4.1. With both the test statistic when the hypothesis is rejected and the test statistic when hypothesis is not rejected.

B.1.2 Homogeneous Test on Real Network Data

1-hour time period	Thu. 26 Oct. 12 - 13	Thu. 02 Nov. 12 - 13	Thu. 26 Oct. 14 - 15	Thu. 02 Nov. 14 - 15	Sun. 22 Oct. 13 - 14	Sun. 12 Nov. 13-14
Thu. 26 Oct, 12 - 13	-	-	-	-	-	-
Thu. 02 Nov, 12 - 13	8974.9	-	-	-	-	-
Thu. 26 Oct, 14 - 15	5095.1	7279.6	-	-	-	-
Thu. 02 Nov, 14 - 15	7291.3	5053.6	5152.2	-	-	-
Sun. 22 Oct, 13 - 14	8518.3	8171.2	6390.9	5576.9	-	-
Sun. 12 Nov, 13 - 14	10018.8	8664.0	6819.5	5810.1	7024.9	-
Tue. 07 Nov, 10 - 11	10256.8	10452.6	9774.6	8581.9	11068.0	12244.9

Table B.1.2: Table of Q the *test statistic*, for comparison between different 1-hour time period in dataset i. The formula for calculating the test statistic can be seen in Equation 3.2.13.

1-hour time period	Thu. 26 Oct. 12 - 13	Thu. 02 Nov. 12 - 13	Thu. 26 Oct. 14 - 15	Thu. 02 Nov. 14 - 15	Sun. 22 Oct. 13 - 14	Sun. 12 Nov. 13 - 14
Thu. 26 Oct, 12 - 13	-	-	-	-	-	-
Thu. 02 Nov, 12 - 13	2523	-	-	-	-	-
Thu. 26 Oct, 14 - 15	2048	1945	-	-	-	-
Thu. 02 Nov, 14 - 15	2076	1936	1306	-	-	-
Sun. 22 Oct, 13 - 14	2252	2114	1506	1560	-	-
Sun. 12 Nov, 13 - 14	2221	2085	1475	1525	1700	-
Tue. 07 Nov, 10 - 11	3102	3077	2629	2601	2791	2803

Table B.1.3: Table of k, the degrees of freedom for the test when comparing different 1-hour time period in dataset i. The formula for calculating the test statistic can be seen in Equation 3.2.15

1-hour time period	Thu. 26 Oct. 12-13	Thu. 02 Nov. 12-13	Thu. 26 Oct. 14-15	Thu. 02 Nov. 14-15	Sun. 22 Oct. 13-14	Sun. 12 Nov. 13-14
Thu. 26 Oct, 12 - 13	-	-	-	-	-	-
Thu. 02 Nov, 12 - 13	0.00	-	-	-	-	-
Thu. 26 Oct, 14 - 15	0.00	0.00	-	-	-	-
Thu. 02 Nov, 14 - 15	0.00	0.00	0.00	-	-	-
Sun. 22 Oct, 13 - 14	0.00	0.00	0.00	0.00	-	-
Sun. 12 Nov, 13 - 14	0.00	0.00	0.00	0.00	0.00	-
Tue. 07 Nov, 10 - 11	0.00	0.00	0.00	0.00	0.00	0.00

Table B.1.4: Table of p, the probability that the test statistic Q in Table B.1.2 would be obtained if both 1-hour time periods would have the same transition probability matrix. That is, the result of Equation 3.2.16

B.2 Evaluation of Dissimilarity Measures

B.2.1 The Resulting Dissimilarity Metrics

	03 Jan. 08 - 09	03 Jan. 09 - 10	03 Jan. 10 - 11	03 Jan. 11 - 12
03 Jan, 08 - 09	0.00	-	-	-
03 Jan, 09 - 10	101.76	0.00	-	-
03 Jan, 10 - 11	105.59	116.85	0.00	-
03 Jan, 11 - 12	116.83	123.92	127.63	0.00
03 Jan, 12 - 13	103.38	117.78	116.92	125.24
03 Jan, 13 - 14	95.71	106.55	105.98	121.58
03 Jan, 14 - 15	99.40	103.57	103.14	119.09
03 Jan, 15 - 16	105.83	115.92	118.95	124.72
03 Jan, 16 - 17	106.64	118.55	118.00	119.15
03 Jan, 17 - 18	91.60	110.12	112.26	118.38
03 Jan, 18 - 19	83.27	113.90	112.30	125.61
03 Jan, 19 - 20	93.68	118.98	119.63	126.33
03 Jan, 20 - 21	84.93	98.52	105.69	115.43
03 Jan, 21 - 22	88.03	101.53	109.85	115.24
04 Jan, 08 - 09	82.41	106.38	119.31	124.71
04 Jan, 09 - 10	80.92	98.47	106.47	115.49
04 Jan, 10 - 11	97.89	109.02	105.85	122.39
04 Jan, 11 - 12	92.82	109.35	102.10	117.32
04 Jan, 12 - 13	100.67	121.79	112.83	120.24
04 Jan, 13 - 14	121.20	125.48	125.48	137.42

Table B.2.1: Cropped version of the dissimilarity matrix, using Equation 3.3.3 as a dissimilarity metric.

	03 Jan. 08 - 09	03 Jan. 09 - 10	03 Jan. 10 - 11	03 Jan. 11 - 12
03 Jan, 08 - 09	0.00	-	-	-
03 Jan, 09 - 10	0.32	0.00	-	-
03 Jan, 10 - 11	0.52	0.36	0.00	-
03 Jan, 11 - 12	0.52	0.35	0.22	0.00
03 Jan, 12 - 13	0.49	0.32	0.28	0.23
03 Jan, 13 - 14	0.47	0.31	0.30	0.27
03 Jan, 14 - 15	0.47	0.31	0.25	0.21
03 Jan, 15 - 16	0.50	0.34	0.26	0.23
03 Jan, 16 - 17	0.43	0.26	0.30	0.29
03 Jan, 17 - 18	0.35	0.23	0.35	0.34
03 Jan, 18 - 19	0.39	0.39	0.50	0.50
03 Jan, 19 - 20	0.47	0.43	0.55	0.54
03 Jan, 20 - 21	0.48	0.46	0.61	0.61
03 Jan, 21 - 22	0.50	0.55	0.68	0.68
04 Jan, 08 - 09	0.27	0.36	0.55	0.53
04 Jan, 09 - 10	0.25	0.24	0.41	0.39
04 Jan, 10 - 11	0.44	0.28	0.27	0.28
04 Jan, 11 - 12	0.48	0.31	0.22	0.19
04 Jan, 12 - 13	0.50	0.36	0.28	0.23
04 Jan, 13 - 14	0.48	0.34	0.30	0.26

Table B.2.2: Cropped version of the dissimilarity matrix, using Equation 3.3.8 as a dissimilarity metric.

B.2.2 The Classification Using PAM

Time period	Dissimilarity measure						Number of visited cells
	3.3.3	3.3.4	3.3.6	3.3.7	3.3.8	3.3.9	
03 Jan, 08 - 09	1	1	1	1	1	1	1827
03 Jan, 09 - 10	1	2	1	1	1	1	2849
03 Jan, 10 - 11	1	2	2	2	1	1	4996
03 Jan, 11 - 12	2	2	2	2	1	1	4847
03 Jan, 12 - 13	1	2	2	2	2	2	4541
03 Jan, 13 - 14	1	2	2	2	2	2	4254
03 Jan, 14 - 15	1	2	2	2	2	2	4387
03 Jan, 15 - 16	1	1	2	2	1	1	4673
03 Jan, 16 - 17	1	1	2	1	1	1	3903
03 Jan, 17 - 18	1	3	1	1	1	1	2950
03 Jan, 18 - 19	3	3	3	3	3	3	2045
03 Jan, 19 - 20	3	4	3	3	3	3	2041
03 Jan, 20 - 21	3	4	3	3	3	3	1866
03 Jan, 21 - 22	3	4	3	3	3	3	1133
04 Jan, 08 - 09	3	4	1	3	1	1	1792
04 Jan, 09 - 10	1	1	1	1	1	1	2473
04 Jan, 10 - 11	1	1	2	1	1	1	3986
04 Jan, 11 - 12	1	2	2	2	1	1	4561
04 Jan, 12 - 13	1	1	2	2	2	2	5081
04 Jan, 13 - 14	4	2	2	2	2	2	4446
04 Jan, 14 - 15	3	4	2	2	2	2	4153
04 Jan, 15 - 16	1	2	2	2	1	1	4849
04 Jan, 16 - 17	1	1	2	2	1	1	4401
04 Jan, 17 - 18	3	2	2	1	1	1	3550
04 Jan, 18 - 19	3	3	3	3	3	3	2313
04 Jan, 19 - 20	3	3	3	3	3	3	1666
04 Jan, 20 - 21	3	3	4	3	3	3	1082
04 Jan, 21 - 22	3	4	4	3	3	3	867
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table B.2.3: Shows the classification of each time period according to the six different dissimilarity metrics. It is cropped because of length reasons.

B.2.3 The Summary of the Classifications

The figures are on the next page in landscape mode.

Time period	Classes	Number of visited cells
06 - 07	3 3 3 1 1 1 3 3 3	530 614 672 683 699 853 882 998 1156
07 - 08	1 1 1 3 1 1 3 3 1	1054 1074 1081 1153 1157 1194 1301 1371 1834
08 - 09	1 3 1 1 1 1 1 1 1 1 3	1731 1792 1827 2013 2087 2221 2250 2400 2424 2532 2889
09 - 10	1 1 1 1 1 1 1 1 1 1 1	2120 2473 2539 2828 2849 2850 2961 3110 3218 3338 3459
10 - 11	3 1 1 1 1 1 1 1 3 1 1 1	3208 3359 3593 3924 3976 3986 4009 4185 4206 4381 4485 4996
11 - 12	1 1 1 1 1 1 1 1 1 1 1 1 2	3883 4006 4114 4419 4456 4495 4502 4543 4561 4605 4698 4709 4847
12 - 13	1 3 1 1 1 1 1 1 3 1 3 1 1	3904 3982 4328 4388 4393 4523 4541 4634 4650 4761 5029 5081 5374
13 - 14	3 3 3 1 1 1 4 1 1 1 3 1 1	3663 4083 4161 4254 4287 4300 4446 4469 4557 4628 4706 4864 5720
14 - 15	1 1 1 1 1 1 3 3 1 1 1 1 1	3440 3665 3679 3764 3971 4024 4139 4153 4276 4387 4420 4811 5015
15 - 16	1 3 1 1 1 1 1 1 3 1 1 1	3322 3399 3684 3693 3893 4143 4262 4353 4518 4673 4849 4874
16 - 17	3 1 3 1 1 1 1 1 1 1 1 1 1	3099 3379 3408 3506 3637 3675 3903 3907 3943 4036 4159 4401 4498
17 - 18	1 1 3 1 1 3 1 3 3 1 3 1 1	2747 2950 2990 3087 3119 3135 3250 3550 3562 3576 3959 4151 4359
18 - 19	3 3 1 3 3 3 1 3 3 3 1 1 3 3	1789 1796 1838 1893 1963 2045 2116 2313 2357 2432 2522 2579 2813 2819
19 - 20	3 3 3 3 3 1 3 3 3 3 3 3 1 3 3	1332 1367 1376 1386 1474 1501 1610 1666 1743 1763 1806 1892 1972 2031 2041
20 - 21	3 3 3 3 3 3 1 3 3 3 3 3 3 3 3	951 980 1051 1055 1082 1144 1212 1355 1358 1414 1599 1607 1698 1866 1942
21 - 22	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	697 710 786 789 811 828 850 867 1132 1133 1198 1215 1229 1548 1803

Table B.2.4: Using Equation 3.3.3 as dissimilarity metric, i.e. the sum of the absolute value of the difference in the probability of handover to cell j given that the UE was in cell i. Each number (1,2,3,4) in the "Classes" column corresponds to a classification of the data according to the PAM-algorithm. Each classification has their corresponding number of visited cells in the column "Total Number of visited Cells".

Time period	Classes	Number of visited cells
06 - 07	4 4 4 1 1 3 1 4 3	530 614 672 683 699 853 882 998 1156
07 - 08	4 1 1 4 1 1 4 1 1	1054 1074 1081 1153 1157 1194 1301 1371 1834
08 - 09	1 4 1 1 1 4 4 1 1 1 4	1731 1792 1827 2013 2087 2221 2250 2400 2424 2532 2889
09 - 10	3 1 1 1 2 2 1 2 1 1 1	2120 2473 2539 2828 2849 2850 2961 3110 3218 3338 3459
10 - 11	4 3 3 3 1 1 2 2 3 1 1 2	3208 3359 3593 3924 3976 3986 4009 4185 4206 4381 4485 4996
11 - 12	1 1 4 1 4 2 1 1 2 2 1 1 2	3883 4006 4114 4419 4456 4495 4502 4543 4561 4605 4698 4709 4847
12 - 13	2 1 3 1 2 2 2 1 2 1 4 1 1	3904 3982 4328 4388 4393 4523 4541 4634 4650 4761 5029 5081 5374
13 - 14	4 3 3 2 3 2 2 2 2 1 3 1 1	3663 4083 4161 4254 4287 4300 4446 4469 4557 4628 4706 4864 5720
14 - 15	1 1 2 3 1 1 4 4 1 2 2 1 2	3440 3665 3679 3764 3971 4024 4139 4153 4276 4387 4420 4811 5015
15 - 16	3 4 1 1 2 1 2 1 3 1 2 2	3322 3399 3684 3693 3893 4143 4262 4353 4518 4673 4849 4874
16 - 17	4 1 3 3 3 3 1 1 1 1 1 1 2	3099 3379 3408 3506 3637 3675 3903 3907 3943 4036 4159 4401 4498
17 - 18	1 3 3 3 1 4 1 2 3 1 3 3 3	2747 2950 2990 3087 3119 3135 3250 3550 3562 3576 3959 4151 4359
18 - 19	1 4 1 3 3 3 1 3 4 3 1 1 3 1	1789 1796 1838 1893 1963 2045 2116 2313 2357 2432 2522 2579 2813 2819
19 - 20	4 4 3 3 1 3 4 3 4 4 4 3 1 4 4	1332 1367 1376 1386 1474 1501 1610 1666 1743 1763 1806 1892 1972 2031 2041
20 - 21	3 3 3 4 3 4 1 4 3 1 3 1 4 4 3	951 980 1051 1055 1082 1144 1212 1355 1358 1414 1599 1607 1698 1866 1942
21 - 22	4 4 3 4 4 4 4 4 4 4 4 4 3 4 3	697 710 786 789 811 828 850 867 1132 1133 1198 1215 1229 1548 1803

Table B.2.5: Using Equation 3.3.4 as dissimilarity metric, i.e. the sum of the squared value of the difference in the probability of handover to cell j given that the UE was in cell i. Each number (1,2,3,4) in the "Classes" column corresponds to a classification of the data according to the PAM-algorithm. Each classification has their corresponding number of visited cells in the column "Number of visited Cells".

Time period	Classes	Number of visited cells
06 - 07	4 4 4 4 4 4 4 4 4	530 614 672 683 699 853 882 998 1156
07 - 08	4 4 4 4 4 4 4 1 1	1054 1074 1081 1153 1157 1194 1301 1371 1834
08 - 09	1 1 1 1 1 1 1 1 1 1 1	1731 1792 1827 2013 2087 2221 2250 2400 2424 2532 2889
09 - 10	1 1 1 1 1 1 1 1 1 1 1	2120 2473 2539 2828 2849 2850 2961 3110 3218 3338 3459
10 - 11	1 2 2 2 2 2 2 2 2 2 2 2	3208 3359 3593 3924 3976 3986 4009 4185 4206 4381 4485 4996
11 - 12	2 2 2 2 2 2 2 2 2 2 2 2 2	3883 4006 4114 4419 4456 4495 4502 4543 4561 4605 4698 4709 4847
12 - 13	2 2 2 2 2 2 2 2 2 2 2 2 2	3904 3982 4328 4388 4393 4523 4541 4634 4650 4761 5029 5081 5374
13 - 14	2 2 2 2 2 2 2 2 2 2 2 2 2	3663 4083 4161 4254 4287 4300 4446 4469 4557 4628 4706 4864 5720
14 - 15	2 2 2 2 2 2 2 2 2 2 2 2 2	3440 3665 3679 3764 3971 4024 4139 4153 4276 4387 4420 4811 5015
15 - 16	2 2 2 2 2 2 2 2 2 2 2 2	3322 3399 3684 3693 3893 4143 4262 4353 4518 4673 4849 4874
16 - 17	1 2 2 2 2 2 2 2 2 2 2 2 2	3099 3379 3408 3506 3637 3675 3903 3907 3943 4036 4159 4401 4498
17 - 18	1 1 1 1 1 2 2 2 2 2 2 2 2	2747 2950 2990 3087 3119 3135 3250 3550 3562 3576 3959 4151 4359
18 - 19	3 3 1 3 3 3 3 3 3 3 3 3 3 3	1789 1796 1838 1893 1963 2045 2116 2313 2357 2432 2522 2579 2813 2819
19 - 20	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	1332 1367 1376 1386 1474 1501 1610 1666 1743 1763 1806 1892 1972 2031 2041
20 - 21	4 4 4 4 4 4 4 4 3 3 3 3 3 3 3	951 980 1051 1055 1082 1144 1212 1355 1358 1414 1599 1607 1698 1866 1942
21 - 22	4 4 4 4 4 4 4 4 4 3 3 4 4 3 3	697 710 786 789 811 828 850 867 1132 1133 1198 1215 1229 1548 1803

Table B.2.6: Using Equation 3.3.6 as dissimilarity metric, i.e. the sum of the absolute value of the difference in number of handover from cell i to cell j, divided by the total number of handovers in both sets of sequences. Each number (1,2,3,4) in the "Classes" column corresponds to a classification of the data according to the PAM-algorithm. Each classification has their corresponding number of visited cells in the column "Number of visited Cells".

Time period	Classes	Number of visited cells
06 - 07	3 3 3 3 3 3 3 4 4	530 614 672 683 699 853 882 998 1156
07 - 08	1 3 3 3 1 3 1 1 4	1054 1074 1081 1153 1157 1194 1301 1371 1834
08 - 09	1 1 1 1 1 1 1 1 1 1 4	1731 1792 1827 2013 2087 2221 2250 2400 2424 2532 2889
09 - 10	1 1 1 1 1 1 1 1 1 4 1	2120 2473 2539 2828 2849 2850 2961 3110 3218 3338 3459
10 - 11	1 1 1 1 1 1 1 1 1 1 1 1	3208 3359 3593 3924 3976 3986 4009 4185 4206 4381 4485 4996
11 - 12	2 1 2 1 2 1 1 2 1 2 1 1 1	3883 4006 4114 4419 4456 4495 4502 4543 4561 4605 4698 4709 4847
12 - 13	2 2 2 2 2 2 2 2 2 2 2 2 2	3904 3982 4328 4388 4393 4523 4541 4634 4650 4761 5029 5081 5374
13 - 14	2 2 2 2 2 2 2 2 2 2 2 2 2	3663 4083 4161 4254 4287 4300 4446 4469 4557 4628 4706 4864 5720
14 - 15	2 2 2 1 2 2 2 2 2 2 2 1 1	3440 3665 3679 3764 3971 4024 4139 4153 4276 4387 4420 4811 5015
15 - 16	1 1 1 2 1 2 1 1 1 1 1 1	3322 3399 3684 3693 3893 4143 4262 4353 4518 4673 4849 4874
16 - 17	1 1 1 1 1 1 1 1 1 1 1 1 1	3099 3379 3408 3506 3637 3675 3903 3907 3943 4036 4159 4401 4498
17 - 18	1 1 1 1 1 1 1 1 1 1 1 3 1	2747 2950 2990 3087 3119 3135 3250 3550 3562 3576 3959 4151 4359
18 - 19	3 3 3 3 3 3 3 3 3 3 3 3 3 3	1789 1796 1838 1893 1963 2045 2116 2313 2357 2432 2522 2579 2813 2819
19 - 20	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	1332 1367 1376 1386 1474 1501 1610 1666 1743 1763 1806 1892 1972 2031 2041
20 - 21	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	951 980 1051 1055 1082 1144 1212 1355 1358 1414 1599 1607 1698 1866 1942
21 - 22	3 4 3 3 3 3 3 3 3 3 3 3 4 3 3	697 710 786 789 811 828 850 867 1132 1133 1198 1215 1229 1548 1803

Table B.2.7: Using Equation 3.3.7 as dissimilarity metric, i.e. is the sum of squared value of the difference in handover from cell i to cell j, divided by square of the total number of handovers in both sets of sequences. Each number (1,2,3,4) in the ”Classes” column corresponds to a classification of the data according to the PAM-algorithm. Each classification has their corresponding number of visited cells in the column ”Number of visited Cells”.

Time period	Classes	Number of visited cells
06 - 07	4 4 4 4 4 4 4 3 3 4	530 614 672 683 699 853 882 998 1156
07 - 08	4 4 4 4 4 4 4 4 4	1054 1074 1081 1153 1157 1194 1301 1371 1834
08 - 09	1 1 1 1 1 1 1 1 1 1 1	1731 1792 1827 2013 2087 2221 2250 2400 2424 2532 2889
09 - 10	1 1 1 1 1 1 1 1 1 1 1	2120 2473 2539 2828 2849 2850 2961 3110 3218 3338 3459
10 - 11	1 1 2 1 1 1 1 1 1 1 1 1	3208 3359 3593 3924 3976 3986 4009 4185 4206 4381 4485 4996
11 - 12	2 1 2 1 2 1 1 2 1 2 1 1 1	3883 4006 4114 4419 4456 4495 4502 4543 4561 4605 4698 4709 4847
12 - 13	2 2 2 2 2 2 2 2 2 2 2 2 2	3904 3982 4328 4388 4393 4523 4541 4634 4650 4761 5029 5081 5374
13 - 14	2 2 2 2 2 2 2 2 2 2 2 2 2	3663 4083 4161 4254 4287 4300 4446 4469 4557 4628 4706 4864 5720
14 - 15	2 2 2 1 2 2 2 2 2 2 2 1 1	3440 3665 3679 3764 3971 4024 4139 4153 4276 4387 4420 4811 5015
15 - 16	1 1 1 2 1 1 1 1 1 1 1 1	3322 3399 3684 3693 3893 4143 4262 4353 4518 4673 4849 4874
16 - 17	1 1 1 1 1 1 1 1 1 1 1 1 1	3099 3379 3408 3506 3637 3675 3903 3907 3943 4036 4159 4401 4498
17 - 18	1 1 1 1 1 1 1 1 1 1 1 1 1	2747 2950 2990 3087 3119 3135 3250 3550 3562 3576 3959 4151 4359
18 - 19	3 3 4 3 4 3 4 3 3 3 3 3 4 3	1789 1796 1838 1893 1963 2045 2116 2313 2357 2432 2522 2579 2813 2819
19 - 20	3 3 4 3 3 3 3 3 3 4 3 3 3 3 3	1332 1367 1376 1386 1474 1501 1610 1666 1743 1763 1806 1892 1972 2031 2041
20 - 21	3 3 3 3 3 4 4 3 3 3 3 3 3 3 3	951 980 1051 1055 1082 1144 1212 1355 1358 1414 1599 1607 1698 1866 1942
21 - 22	3 4 3 4 3 3 4 3 3 3 3 3 3 3 3	697 710 786 789 811 828 850 867 1132 1133 1198 1215 1229 1548 1803

Table B.2.8: Using Equation 3.3.8 as dissimilarity metric, i.e. the sum of the absolute value of the difference in percent of a handover from cell i to cell j. Each number (1,2,3,4) in the "Classes" column corresponds to a classification of the data according to the PAM-algorithm. Each classification has their corresponding number of visited cells in the column "Number of visited Cells".

Time period	Classes	Number of visited cells
06 - 07	3 3 3 3 3 3 3 4 4	530 614 672 683 699 853 882 998 1156
07 - 08	1 3 3 3 1 3 1 1 4	1054 1074 1081 1153 1157 1194 1301 1371 1834
08 - 09	1 1 1 1 1 1 1 1 1 1 4	1731 1792 1827 2013 2087 2221 2250 2400 2424 2532 2889
09 - 10	1 1 1 1 1 1 1 1 1 4 1	2120 2473 2539 2828 2849 2850 2961 3110 3218 3338 3459
10 - 11	1 1 1 1 1 1 1 1 1 1 1 1	3208 3359 3593 3924 3976 3986 4009 4185 4206 4381 4485 4996
11 - 12	2 1 2 1 2 1 1 2 1 2 1 1 1	3883 4006 4114 4419 4456 4495 4502 4543 4561 4605 4698 4709 4847
12 - 13	2 2 2 2 2 2 2 2 2 2 2 2 2	3904 3982 4328 4388 4393 4523 4541 4634 4650 4761 5029 5081 5374
13 - 14	2 2 2 2 2 2 2 2 2 2 2 2 2	3663 4083 4161 4254 4287 4300 4446 4469 4557 4628 4706 4864 5720
14 - 15	2 2 2 1 2 2 2 2 2 2 2 1 1	3440 3665 3679 3764 3971 4024 4139 4153 4276 4387 4420 4811 5015
15 - 16	1 1 1 2 1 2 1 1 1 1 1 1	3322 3399 3684 3693 3893 4143 4262 4353 4518 4673 4849 4874
16 - 17	1 1 1 1 1 1 1 1 1 1 1 1 1	3099 3379 3408 3506 3637 3675 3903 3907 3943 4036 4159 4401 4498
17 - 18	1 1 1 1 1 1 1 1 1 1 1 3 1	2747 2950 2990 3087 3119 3135 3250 3550 3562 3576 3959 4151 4359
18 - 19	3 3 3 3 3 3 3 3 3 3 3 3 3 3	1789 1796 1838 1893 1963 2045 2116 2313 2357 2432 2522 2579 2813 2819
19 - 20	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	1332 1367 1376 1386 1474 1501 1610 1666 1743 1763 1806 1892 1972 2031 2041
20 - 21	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	951 980 1051 1055 1082 1144 1212 1355 1358 1414 1599 1607 1698 1866 1942
21 - 22	3 4 3 3 3 3 3 3 3 3 3 3 4 3 3	697 710 786 789 811 828 850 867 1132 1133 1198 1215 1229 1548 1803

Table B.2.9: Using Equation 3.3.9 as dissimilarity metric, i.e. the sum of the squared value of the difference in percent of a handover from cell i to cell j. Each number (1,2,3,4) in the "Classes" column corresponds to a classification of the data according to the PAM-algorithm. Each classification has their corresponding number of visited cells in the column "Number of visited Cells".

Appendix C

Results for the Preliminary Task

C.1 The Order of the Markov Chain

C.1.1 All Time Periods in Dataset i Together

Test	Test statistic	*df*	*p*
0 vs. 1	5822285.89	67848	0.00
1 vs. 2	835213.40	72823	0.00
2 vs. 3	107258.44	33852	0.00
3 vs. 4	41361.87	18165	0.00
4 vs. 5	21625.79	9701	0.00

Table C.1.1: Test for order of Markov chain, using all time periods in dataset i together.

Test	Expected frequency count smaller than 5
0 vs. 1	91.28 %
1 vs. 2	96.29 %
2 vs. 3	96.24 %
3 vs. 4	96.66 %
4 vs. 5	97.08 %

Table C.1.2: Percent of terms in Equation 3.2.31 that have expected frequency count smaller than 5, for all $Q_{j\cdots k}$ together. That is when $f_{ij\cdots k*} \cdot \hat{p}_{j\cdots k|l} < 5$. This is in all of Dataset i together.

C.1.2 All Time Periods in Dataset i Separately

Test	Test statistic	*df*	*p*
0 vs. 1	1164032.11	54492	0.00
1 vs. 2	102209.30	21679	0.00
2 vs. 3	11426.13	6081	0.00
3 vs. 4	4533.11	2478	0.00
4 vs. 5	2014.57	1047	0.00

Table C.1.3: Test for order of Markov chain, using data from 2017-10-26, 12:00 to 13:00.

Test	Test statistic	*df*	*p*
0 vs. 1	933608.84	51948	0.00
1 vs. 2	80856.78	17847	0.00
2 vs. 3	8355.27	4566	0.00
3 vs. 4	3770.78	1979	0.00
4 vs. 5	1269.56	704	0.00

Table C.1.4: Test for order of Markov chain, using data from 2017-11-02, 12:00 to 13:00.

Test	Test statistic	*df*	*p*
0 vs. 1	326388.53	32900	0.00
1 vs. 2	20798.24	4848	0.00
2 vs. 3	1569.71	961	0.00
3 vs. 4	754.35	348	0.00
4 vs. 5	189.90	97	0.00

Table C.1.5: Test for order of Markov chain, using data from 2017-10-26, 14:00 to 15:00.

Test	Test statistic	*df*	*p*
0 vs. 1	325245.34	33464	0.00
1 vs. 2	17655.59	5137	0.00
2 vs. 3	1861.24	781	0.00
3 vs. 4	512.79	276	0.00
4 vs. 5	252.01	136	0.00

Table C.1.6: Test for order of Markov chain, using data from 2017-11-02, 14:00 to 15:00.

Test	Test statistic	*df*	*p*
0 vs. 1	640480.44	43878	0.00
1 vs. 2	38603.32	8496	0.00
2 vs. 3	3791.14	1615	0.00
3 vs. 4	839.03	548	0.00
4 vs. 5	440.56	217	0.00

Table C.1.7: Test for order of Markov chain, using data from 2017-10-22, 13:00 to 14:00.

Test	Test statistic	*df*	*p*
0 vs. 1	521764.71	38994	0.00
1 vs. 2	75827.44	7558	0.00
2 vs. 3	4018.79	1195	0.00
3 vs. 4	2382.55	408	0.00
4 vs. 5	232.33	143	0.00

Table C.1.8: Test for order of Markov chain, using data from 2017-11-12, 13:00 to 14:00.

Test	Test statistic	*df*	*p*
0 vs. 1	2361273.22	61000	0.00
1 vs. 2	222739.43	38160	0.00
2 vs. 3	34133.50	14821	0.00
3 vs. 4	13341.85	7501	0.00
4 vs. 5	6201.37	3731	0.00

Table C.1.9: Test for order of Markov chain, using data from 2017-11-07, 10:00 to 11:00.

	Expected frequency count smaller than 5 in percent, for Table:						
Test	C.1.3	C.1.4	C.1.5	C.1.6	C.1.7	C.1.8	C.1.9
0 vs. 1	97.77 %	97.90 %	99.13 %	99.24 %	98.91 %	98.94 %	95.34 %
1 vs. 2	97.89 %	97.60 %	98.03 %	98.27 %	98.26 %	98.28 %	97.02 %
2 vs. 3	97.29 %	97.18 %	97.59 %	97.44 %	97.47 %	97.84 %	96.68 %
3 vs. 4	97.13 %	96.98 %	97.37 %	97.37 %	97.65 %	96.58 %	97.01 %
4 vs. 5	97.19 %	97.07 %	95.31 %	97.94 %	96.20 %	96.54 %	97.31 %

Table C.1.10: Percent of terms in Equation 3.2.31 that have expected frequency count smaller than 5, for all $Q_{j\cdots k}$ together. That is when $f_{ij\cdots k*} \cdot \hat{p}_{j\cdots k|l} < 5$. This is for each time period in dataset i separately.

Appendix D

Example of a Handover Request Message

This is an overview of the Handover Request Message that is sent to the new eNB.

```
X2AP {
  pdu value X2AP-PDU ::= initiatingMessage : {
    procedureCode 0,
    criticality reject,
    value HandoverRequest : {
      protocolIEs {
        {
          id 10,
          criticality reject,
          value UE-X2AP-ID : 3364
          #Above: This ID is matched with UE-X2AP-ID
          # in HANDOVER REQUEST ACKNOWLEDGE
        },
        {
          id 5,
          criticality ignore,
          value Cause : radioNetwork : handover-desirable-for-radio-reasons
        },
        {
          id 11,
          criticality reject,
          value ECGI : {
            pLMN-Identity '05f510'H,
            eUTRANcellIdentifier '1000000110100011111100001011'B
            #Above: Cell identification number in binary
          }
        },
        {
          id 23,
          criticality reject,
          value GUMMEI : {
            gU-Group-ID {
              pLMN-Identity '05f510'H,
              mME-Group-ID 'c545'H
            },
            mME-Code 'a8'H
          }
        },
        {
          id 14,
          criticality reject,
          value UE-ContextInformation : {
          #Text skipped
          #
          #
          #Text skipped
          }
        },
        {
          id 15,
          criticality ignore,
          # Below is the the UE visited cell history.
          value UE-HistoryInformation : {
            LastVisitedCell-Item e_UTRAN_Cell : {
              global-Cell-ID {
                pLMN-Identity '05f510'H,
                eUTRANcellIdentifier '10000001110001111010000001101'B
                # Above: Cell identification number in binary
              },
              cellType {
                cell-Size verysmall
              },
              time-UE-StayedInCell 3 # Time in cell in seconds
            },
            LastVisitedCell-Item e_UTRAN_Cell : {
              global-Cell-ID {
                pLMN-Identity '05f510'H,
                eUTRANcellIdentifier '10000001101110110000000001101'B
                # Above: Cell identification number in binary
              },
              cellType {
                cell-Size verysmall
              },
              time-UE-StayedInCell 2 # Time in cell in seconds
            },
            LastVisitedCell-Item e_UTRAN_Cell : {
              global-Cell-ID {
                pLMN-Identity '05f510'H,
                eUTRANcellIdentifier '10000001110001111010000001101'B
```

```
                # Above: Cell identification number in binary
              },
              cellType {
                cell-Size verysmall
              },
              time-UE-StayedInCell 8 # Time in cell in seconds
            },
            LastVisitedCell-Item e_UTRAN_Cell : {
              global-Cell-ID {
                pLMN-Identity '05f510'H,
                eUTRANcellIdentifier '1000000110111011000000001101'B
                # Above: Cell identification number in binary
              },
              cellType {
                cell-Size verysmall
              },
              time-UE-StayedInCell 1 # Time in cell in seconds
            }
          }
        },
        {
          id 36,
          criticality ignore,
          value SRVCCOperationPossible : possible
        },
        {
          id 98,
          criticality ignore,
          value Masked-IMEISV : '1000011010010101011...' # Too long
          # Above: This Masked-IMEISV is used to
          # find non-unique visited cell sequences
        }
      }
    }
  }
  RRC {
        #Text skipped
        #
        #
        #Text skipped
  }
}
```

TRITA -SCI-GRU 2018:215

www.kth.se