Can students' progress data be modeled using Markov chains?

FILIP CARLSSON
Can students' progress data be modeled using Markov chains?

FILIP CARLSSON
Authors

Filip Carlsson
*Industrial Engineering and Management*
*KTH Royal Institute of Technology*

Place for Project

KTH Royal Institute of Technology
Stockholm, Sweden

Examiner

Jörgen Säve-Söderbergh
*Department of Mathematics*
*KTH Royal Institute of Technology*

Supervisors

Jimmy Olsson
*Associate Professor and Head of Division Mathematical Statistics*
*Department of Mathematics*
*KTH Royal Institute of Technology*

Julia Liljegren
*Doctoral Student*
*Department of Industrial Economics and Management*
*KTH Royal Institute of Technology*
Abstract

In this thesis a Markov chain model, which can be used for analysing students’ performance and their academic progress, is developed. Being able to evaluate students progress is useful for any educational system. It gives a better understanding of how students resonates and it can be used as support for important decisions and planning. Such a tool can be helpful for managers of the educational institution to establish a more optimal educational policy, which ensures better position in the educational market.

To show that it is reasonable to use a Markov chain model for this purpose, a test for how well data fits such a model is created and used. The test shows that we cannot reject the hypothesis that the data can be fitted to a Markov chain model.

The data used for the thesis contains information about 22551 students from LTH between the years 1993 - 2016 from 15 different programs, i.e all master of engineering programs offered.
Abstract


Den data som används i examensarbetet innehåller information om 22551 studenter från LTH mellan åren 1993 - 2016 från 15 olika program.
Acknowledgements

I would like to express my gratitude to Jimmy Olsson for providing me with guidance and expertise within the subject of this thesis, especially for the inspiration to check validity of fitting a Markov chain model to the dataset, which came to be an important and interesting part of the thesis. I would also like to thank Jörgen Säve-Söderbergh for his guidance and support throughout the process of writing, and Timo Koski for the guidance in the initial part of the thesis.
## Contents

1 Introduction ............................................. 7  
  1.1 Background .......................................... 7  
  1.2 Problem statement .................................. 7  
  1.3 Purpose ............................................... 8  
  1.4 Literature Review .................................. 8  
  1.5 Data collection ..................................... 9  
  1.6 Data driven decision making ....................... 9  

2 Mathematical Theory ................................. 10  
  2.1 Theoretic background for Markov chains ............ 11  
    2.1.1 Transition probabilities ....................... 11  
    2.1.2 Chapman Kolmogorov ......................... 11  
    2.1.3 Absorbing states .............................. 12  
    2.1.4 Absorbing Markov chains .................. 12  
  2.2 \( \chi^2 \) test .................................. 13  

3 Methodology ........................................... 14  
  3.1 Data collection .................................... 14  
  3.2 The model .......................................... 14  
  3.3 Probability transition matrix ..................... 17  
  3.4 Frequency analysis ................................ 17  
  3.5 Validation .......................................... 18  
  3.6 Validity of fitting a first-order Markov chain model to data .................. 18  

4 Results ................................................. 20  
  4.1 Model 1 ............................................. 20  
  4.2 Model 2, comparing male and female students .... 22  
  4.3 Prediction .......................................... 23  
  4.4 Results of validation ................................ 24  
    4.4.1 First approach of \( \chi^2 \)-test ........... 24  
  4.5 Checking validity of fitting a first-order Markov chain model to the dataset 26  

5 Discussion ............................................. 28  
  5.1 The result ......................................... 28  

6 Conclusion ............................................. 30
1 Introduction

1.1 Background

Evaluation of students progress is an essential part of any educational system. It gives a better understanding of how students resonates, and it can be used as support for important decisions and planning. This can be helpful for managers of the educational institution to establish a more optimal educational policy, which ensures better position in the educational market. Results from prior similar studies indicates that the students progression through their education towards their degree possess pertinent stochastic characteristics, and can therefore be modelled as a Markov chain.

A Markov chain is a sequence of random variable through which a process moves through where the current state can effect the outcome of the next state. It can thus be used for describing systems that follow a chain of linked events, where what happens next depends on the current state of the system [1].

A Markov chain can be a suitable probability model for time series in which the observation at a given time is the category in which an individual falls. The simplest version of the Markov chain is that in which there are a finite number of states or categories and a finite number of equidistant time points at which observations are made, this means that the chain is of first order, and the transition probabilities are the same for each time interval [4].

1.2 Problem statement

The aim of this paper is to develop an absorbing Markov chain model, which can be used for analysing the students performance and their academic progress. The model enables estimation and continuous monitoring of the following indicators:

- The fraction of students that have successfully progressed toward different stages of the study program.
- The expected time a student spends at a particular stage of a study programme and the expected duration of the study.
- The fraction of students who have finished the study successfully with graduation as well the fraction of students who have withdrawn from the study.

The model also enables the prediction of future enrolment of students.

The model will also be used to determine if there are differences between the genders of how students accomplishes their studies. In 2007 the master in engineering program become 10 semesters long instead of 9 semesters. The model will be used to see if this development have generated changes in students’ progression.
Goodness of fit
The thesis will also include a statistical analysis of how well a first-order Markov model fits the data. When creating a statistical model it is of importance to know that the data can be described by the model of choice. I.e that it can be fitted, and how well. This thesis contains a test of how predictions made with the model relate to the actual outcome. The thesis also validates if a first-order Markov chain model can be fitted to the data-set. That is, it validates if students’ progress can be described using Markov chains.

1.3 Purpose

The purpose of this thesis is to examine if it is possible to fit a Markov chain model to students’ flow data, and thereby use the Markov chain properties to gain information. The purpose of building a model is to demonstrate its usability. The thesis also works as an example of how institutions such as universities can use data to extract valuable information.

1.4 Literature Review

There have been similar studies of how one can model students progression towards higher education degree done before. There is quite a lot data available since the universities keeps track of their students enrolment and performance. It is also an interesting subject which can give useful information for the institutions. Examples can be seen as [6, Crippa, Mazzoleni & Zenga, 2016] which uses a non-homogeneous fuzzy Markov Chain approach. [7, Rahim, et.al, 2016] create a projection model of postgraduate student flow based on the Markov Chain for postgraduate students at the college of Arts and Science in Universiti Utara Malaysia and [10, Alenka, et.al, 2017] have done a Markov Analysis of students performance and academic process in Slovenian higher education environment. The two latter uses a similar approach as the one which will be used in this paper, thus they use the properties of Markov chains to extract information.

Other examples of attempts to apply Markov chains to analyse higher education process are Moody DuCloy (2014) who applied Markov chains to analyse and predict the mathematical achievement gap between African American and white American students, Hlavatý Dömeová (2014) presented a Markov chain model for students progress throughout the particular course. Symeonaki & Kalamatianou (2011) used the theory of Non-Homogenoious Markov Systems with fuzzy states to describe students educational progress in Greek universities[10]. Some of these papers have been used for
inspiration and help for this paper. Many reports uses the same mathematical background, but differ since the used data-set is different and what is estimated differs.

Since the theory of Markov chains is quite simple and this type of application indicates high practical value, it offers good opportunities for implementation in practice.

Markov chains stems from stochastic processes and is a established subject in statistics and there is therefore a lot of literature on the mathematics. The hardest part was to find a suitable significance test for the model, thus most other papers just documented the results and not the validation of the model.

The method used to validate if a first-order Markov chain model could be fitted to the data, was highly inspired by M.H. Eggers’ article; Validity of fitting a first-order Markov chain model to data [13]. Egger described an elegant, but not completely intuitive way of checking if an observed sequence is compatible with the assumptions of a Markov chain model.

1.5 Data collection

The data used for the paper contains information about 22551 students from LTH between the years 1993 - 2016 from 15 different programs, i.e all master of engineering programs offered. The data set contains several rows containing the following information:

- The given student referred to with an anonymous code
- Which study state the student is in
- Which year and which semester the student began his/her studies at the program
- Which year the row refers to
- Number of unit points finished during the given semester

This means that each student have several rows which describes its way throughout the study. In this paper, the most interesting is the study state, which in the data is divided into 15 states, describing enrolled semester, exchange study, study break, inactivity, graduation and withdrawal.

1.6 Data driven decision making

Being able to predict and evaluate students’ progression through the program is an essential part of any educational system. This paper proposes a decision support system based on an absorbing Markov chain, which is used for helping decision makers at the faculty of engineering in controlling students’ flow transition enrolment. Some
important controlling criteria that govern students’ flow performance during semesters have been evaluated. Including estimation of students flow between different study levels, the average life time a student spends at each level, the number of semesters required for graduation, and students graduation probability. Advantages with the system includes finding any bottle necks to be solved during students transition study from one semester to another, and helping to know students needed facilities to planning for future required resources, hence achieving good quality and efficient university education. This information is valuable for managers of the educational institution to improve their processes, as well for the policy makers at government agencies, to supervise the effectiveness of existing educational policies. Having estimated the future minimum enrolment, the school management can be able to adjust the policy when necessary. Such information is also worthwhile for students, education planners, employers and other actors in the labour market to help them make informed decisions on investment in education. [12]

How universities can become better at making data driven decisions:
The first step is to address and improve the quality of the data. This is essential for anyone who wishes to use data for forecasting and to gain a decision base from data. Find the variables on which the outcome depends on. Often this means finding a lot more data and then finding which variables the decision should be based on. The data also must be representative. As in this case where the data represents students from the faculty of engineering. This should be applicable at the faculty of engineering only, and not other faculties such as law or social sciences. Since there are a lot of differences in the study programs, and perhaps even the students enrolled. For even better results one should perhaps look even closer and look at each different engineering program, instead of the whole faculty.

Lower the costs of access to information. One reason that industries do not use data-driven decisions as standard today is that it can be quite expensive to implement a framework if it doesn’t already exists. Many university institutions have gained a lot of data over the years. Mainly for administrative purposes. This means that the data is there, but if there is no sophisticated tool for searching and sorting the data, the making of models will be both time consuming and expensive.

They also need to improve the way in which information is presented. Make information easier to find. Increase the speed at which information is made available. Raise awareness of business intelligence at senior management levels. Foster a collaborative style of decision-making.[11]

2 Mathematical Theory

The computation is based on theory of Absorbing Markov chains, which is a case of Markov chains with finite states.
2.1 Theoretic background for Markov chains

A stochastic process is called a Markov chain if

\[ P(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, ..., X_n = i_n) = P(X_{n+1} = i_{n+1} | X_n = i_n) \] (1)

for all \( n \) and states \( i_0, i_1, ..., i_{n+1} \). This means that the only relevant information is in the current state and the process does not regard knowledge of historic states. The definition of the Markov property, is a core element of what constitutes a Markov process, i.e that the process is memoryless [3].

2.1.1 Transition probabilities

This yields to the introduction of the transition probabilities for time homogeneous Markov chains. Let \( \{X_n; n \in N\} \) be a Markov chain. If the conditional probabilities \( P(X_{n+1} = j | X_n = i) \) is independent of \( n \) for all \( i, j \) in the state space, the chain is said to be time homogeneous. In that case, the transition probabilities is defined as

\[ p_{ij} = P(X_{n+1} = j | X_n = i) \] (2)

All possible transition probabilities form a transition matrix \( P \), defined as

\[
P = \begin{bmatrix}
p_{11} & p_{13} & p_{13} & \cdots & p_{1N} \\
p_{21} & p_{22} & p_{23} & \cdots & p_{2N} \\
p_{31} & p_{32} & p_{33} & \cdots & p_{3N} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
p_{N1} & p_{N2} & p_{N3} & \cdots & p_{NN}
\end{bmatrix}
\] (3)

For each row, \( r \), in the probability transition matrix it holds that

\[ \sum_{i=1}^{N} p_{ri} = 1 \] (4)

[3]

2.1.2 Chapman Kolmogorov

Below the Chapman-Kolmogorov Equations are derived. These relations are used when derivating a transition matrix after a certain, \( n \), amount of jumps. Here, \( p^{(n)}_{ij} \), denotes
the probability to go from state $i$ to state $j$ in $n$ jumps.

\[
p_{ij}^{(m+n)} = \sum_{k \in \Omega} p_{ik}^{(m)} p_{kj}^{(n)}
\]

b) \( P^{(m+n)} = P^{(m)} P^{(n)} \)

c) \( P^{(n)} = P^n \)

d) \( p^{(n)} = p^{(0)} P^{(n)} = p^{(0)} P^n \)

[3]

2.1.3 Absorbing states

Absorbing states can be defined as a state to which the process can enter and then never leave. Therefore the following holds, where \( T_i \) is the number of time-steps required to reach the absorbing state, $i$,

\[
P(T_i < \infty) = 1
\]

Then, of course, \( p_{ii} = 1 \), i.e. the probability of staying in the same state given that you have entered it is 1 [3].

Non-absorbing states are transient in the meaning that if state $i$ communicates with state $j$, and state $j$ communicates with state $k$, then state $i$ communicates with state $k$.

2.1.4 Absorbing Markov chains

The general form of the probability transition matrix, $P$, of an absorbing Markov chain with $r$ absorbing states and $t$ transient states is defined as:

\[
P = \begin{bmatrix}
Q & R \\
0 & I
\end{bmatrix}
\]

(7)

Where

- $Q$ - $t \times t$ matrix expressing transitions between the transient states
- $R$ - $t \times r$ matrix expressing transitions from the transient states to absorbing states
- $0$ - $r \times t$ zero matrix
- $I$ - $r \times r$ identity matrix
Some useful characteristics of the absorbing Markov chain are the expected time until absorption and probability of absorption. To determine these values, the fundamental matrix

\[ N = (I - Q)^{-1} \]  

is needed. The elements \( n_{ij} \) of the fundamental matrix express in average how many times a Markov chain reaches the transient state \( j \) if it starts in the transient state \( i \). Expected time until absorption or more precisely expected number of steps before a Markov chain is absorbed into one of the absorbing states when it started in the transient state \( i \), is denoted as \( \mu_i \). It can be obtained from

\[ \mu = N1 \]  

where \( 1 \) is a column identity vector .

The probability of absorption in absorbing state \( j \) when starting in transient state \( i \), \( f_{ij} \), can be obtained from

\[ f = NR \]

where \( R \) is the sub-matrix from the probability transition matrix (1).

The distribution over states in a given time \( n \) can be written as a stochastic row vector

\[ p^{(n)} = p^{(0)}P^n \]  

where \( p^{(0)} \) represents the initial distribution of students. The element \( p_i^{(n)} \) represents the probability that a Markov chain is in state \( i \) in time \( n \).[10]

### 2.2 \( \chi^2 \) test

The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. The aim is to determine the zero hypothesis:

\[ H_0: \text{There is no significant difference between the observed and the expected value.} \]

Take the probability space \( \{A_1, A_2, ..., A_n\} \) where \( A_i \) is the states in the model. With this notation \( p_t = P(A_i) \) and \( \sum_{i=1}^n p_i = 1 \). Then \( x_t(i) \) is the number of times the stage \( A_i \) have been observed in \( N_t \) observations, where \( \sum_{i=1}^n p_i = N_t \) and \( N_t \) is the total number of students year \( t \).

Then the estimated value is \( \hat{x}_t(i) = N_t \hat{p}_t(i) \), where \( \hat{p}_t = \hat{p}_{t-1}P \).

In this case the test variable, \( Q(t)_{obs} \), becomes:

\[ Q_{obs} = \sum_{i=1}^n Q_{obs}(i) = \sum_{i=1}^n \frac{(\hat{x}_t(i) - x_t(i))^2}{x_t(i)} \]  

13
where
\[ \hat{x}_t = x_0 P^t \]

If \( H_0 \) is true, i.e. the probabilities in \( P \) are satisfying, one can show that \( Q_{obs} \) is an observation of \( Q \) which is approximately \( \chi^2(k-1) \)-distributed. If \( Q_{obs} \) seems to be too big to come from a \( \chi^2(k-1) \)-distribution, there is reason to reject \( H_0 \).

*Test of significance:*

- Reject \( H_0 \) if \( Q_{obs} > \chi^2_\alpha(k-1) \)
- Do not reject \( H_0 \) if \( Q_{obs} \leq \chi^2_\alpha(k-1) \)

Where \( \alpha \) is the level of significance.[9]

### 3 Methodology

#### 3.1 Data collection

The data used for the paper contains information about 22551 students from LTH between the years 1993 - 2016 from 15 different programs, i.e. all master of engineering programs offered. The data set is in form of a txt-file where each row contains data about one semester for a given student. This means that each student have several rows which describes its way throughout the study. Each row contains the following information:

- The given student referred to with a anonymous code
- Which study state the student is in
- Which year and which semester the student began his/her studies at the program
- Which year the row refers to
- Number of unit points finished during the given semester

In this paper, the most interesting is the study state, which in the data is divided into 15 states, describing enrolled semester, exchange study, study break, inactivity, graduation and withdrawal.

#### 3.2 The model

The duration of an engineering degree in Sweden is generally five years, or ten semesters. In the model, there is one stage for each semester. There is also stages for exchange
semesters, inactivity and study break. These are the transient stages. There are also two absorbing stages, graduation and withdrawal.

For the model, the following stages is defined:

- 1 - 10 the student is enrolled into the $i$th semester of the study programme
- 11 the student has announced study break the current semester
- 12 the student is currently inactive
- 13 the student is enrolled into an exchange semester
- 14 the student has graduated and successfully finished the study programme
- 15 the student has withdrawn from the study programme

In developing the model, some assumptions are considered.

- A student who is currently enrolled in semester $i$ can in the following year not enrol in a semester $j$ where $j < i$, i.e. $p_{ij} = 0$ if $j < i$.
- Irrespective of the stage of the study, after the end of each year, some students become inactive or announce study break.
- A student who is inactive or have study break for more than 10 semesters is classified as having withdrawn from the study programme.
- A student who has withdrawn will never finish the study program. If the student enrolls in a new program, he/she will be classified as a new student.
A student who has successfully graduated will leave the program and never enrol in any semesters again.

Since the length of the degree was changed year 2007, from 9 semesters to 10 semesters, the dataset is splitted into two parts, one for modelling the first part of the data and one for the latter. The models are the same, with the difference that there is only stage 1-9 for the semesters in the first model.

With the states in the initial model, the states \{14, 15\} are absorbing, while the states \{1,2,...,13\} are transient. That is the upper left $13 \times 13$ sub matrix in the probability transition matrix represents $Q$ and the upper right $13 \times 2$ sub matrix represents $R$.

The fundamental matrix $N$ represents the expected time that a student spends at a particular stage. The fundamental matrix is calculated as (8), using the sub matrix, $Q$, from (12). The fundamental matrix is obtained in the following form:

$$
\begin{bmatrix}
  n_{11} & n_{12} & n_{13} & \ldots & n_{113} \\
  n_{21} & n_{22} & n_{23} & \ldots & n_{213} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  n_{131} & n_{132} & n_{133} & \ldots & n_{1313}
\end{bmatrix}
$$

(13)

The elements $n_{ij}$ in $N$ represents the expected time a student spends for the $j$-th study stage when he/she started in stage $i$.

The expected duration of the study is estimated by the expected time until absorption according to equation (9). This results in the column vector, $\mu$. The value in each $\mu_i$ can be seen as the expected duration until graduation or withdrawal for a randomly selected student at stage $i$ in the study programme.

The probability of graduation och withdrawal for students in each stage can be obtained as the probability of absorption, equation (10), using the fundamental matrix (13) and the sub matrix, $R$, from (12). This gives the matrix, $f$. The values in the first column, $f_{i14}$, represents the fraction of students, currently in the $i$th stage, who will successfully graduate. The values in the second column, $f_{i15}$, represents the fraction of students, currently in the $i$th stage, who will withdraw from the programme and never finish. It is convenient that the sum of each row is equal to one[10].
3.3 Probability transition matrix

As defined in Section 4.3 the states is set to be $i = 1, 2, ..., 15$. Let the time of observation be $t = 0, 1, ..., T$, where the time of observation means the students’ number of semesters in the system. Let $p_{ij}$ be the probability of state $j$ at time $t$, given state $i$ at time $t - 1$. Let $n_{ij}(t)$ denote the number of individuals in state $i$ at time $t - 1$ and $j$ at time $t$.

3.4 Frequency analysis

A Markov chain with probability space, $\{1, 2, ..., N\}$, starts at time 1 in state $x_1$ and jumps at time 2, 3, ..., $n$ to the states $x_2, x_3, ..., x_n$. The aim is now to estimate the transition matrix $P = (p_{ij})_{N \times N}$.

The intuitive way to estimate the transition probabilities $p_{ij}$ is the relative frequency of jumps from state $i$ to state $j$ is the relative frequency of jumps from $i$ to $j$, i.e

$$p_{ij}^* = \frac{n_{ij}}{\sum_{k=1}^{N} n_{ik}}$$

where $n_{ij}$ is the number of observed jumps from $i$ to $j$.

To be more stringent, it can be seen that the maximum-likelihood method gives the same estimate of the transition matrix. The Likelihood function is the probabilities to get the observations that one actually have obtained, thus the Markov property gives

$$L(x_1, x_2, ..., x_n; P) = p_{x_1x_2}p_{x_2x_3}...p_{x_{n-1}x_n} = \prod_{i=1}^{N} \prod_{j=1}^{N} p_{ij}^{n_{ij}}$$

(15)

where $n_{ij}$ is the number of jumps from $i$ to $j$. This function should be maximised under the condition that the summation of each row in the transition matrix equals 1, i.e

$$p_{i1} + p_{i2} + ... + p_{iN} = 1, \quad i = 1, 2, ..., N$$

(16)

The logarithm of $L$ under these conditions is maximised with the Lagrange multiplicative method. Thus

$$\Lambda = \ln(L) - \lambda_1(p_{11} + p_{12} + ... + p_{1N}) - \lambda_2(p_{21} + p_{22} + ... + p_{2N}) - ... - \lambda_N(p_{N1} + p_{N2} + ... + p_{NN}) =$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} n_{ij} \ln(p_{ij}) - \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i p_{ij} = \sum_{i=1}^{N} \sum_{j=1}^{N} (n_{ij} \ln(p_{ij}) - \lambda_i p_{ij})$$

(17)

By differentiating with respect to $p_{ij}$ it is obtained that

$$\frac{\partial \Lambda}{\partial p_{ij}} = \frac{n_{ij}}{p_{ij}} - \lambda_i = 0$$

(18)
which gives

\[ \frac{p_{i1}}{n_{i1}} = \frac{p_{i2}}{n_{i2}} = \ldots = \frac{p_{iN}}{n_{iN}} = \frac{1}{\lambda_i} \] (19)

thus

\[ p_{ij} = \frac{n_{ij}}{\lambda_i} \] (20)

Together with the condition (18) it holds that

\[ 1 = \sum_{j=1}^{N} p_{ij} = \sum_{j=1}^{N} \frac{n_{ij}}{\lambda_i} = \frac{n_i}{\lambda_i} \] (21)

where \( n_i \) is the number of times the chain have been in state \( i \) until the time \( n - 1 \). Thus it is obtained that

\[ \frac{1}{\lambda_i} = \frac{1}{n_i}, \text{ thus } p_{ij}^* = \frac{n_{ij}}{n_i} = \frac{n_{ij}}{\sum_{k=1}^{N} n_{ik}} \] (22)

This holds under the assumption that the transition probabilities are not related to each other, except that the rows sums to 1 [5].

### 3.5 Validation

The first step in the validation was to split the data set into two groups, containing the same number of students, call them validation group and prediction group. To accomplish this, a MATLAB function which divided the data set in two random groups from the students reference number.

Then a probability transition matrix were created from the prediction data. The transition matrix was then multiplied with a vector describing the situations students from the prediction data were in at a particular semester. The process was repeated the same number of times as it takes as the comparison covered. The result was compared to the states of the students in the validation data after the same number of semesters.

### 3.6 Validity of fitting a first-order Markov chain model to data

An alternative way to check the validity of the model is to look at the validity of fitting a first-order Markov chain model to the data-set. This part describes a way of testing whether there is a Markov chain of first order that is a likely model to fit the given sequence of data.

A natural question to ask is

*is an observed sequence \( X(1), \ldots, X(N) \) compatible with the assumption of an MCM of order 1 for a suitable choice of the constants \( p_{ij} \)?*
An alternative way, suggested by M.H. Egger in validity of fitting a first-order Markov chain model to data [13], to answer this question is to convert the problem, into \( k \) cases of, a problem which is not concerned with Markov chains. Namely into something like

*a set of \( r \) coloured balls has \( r_j \) balls of colour \( j \) (\( 1 < j < k \)), where \( r = r_1 + \ldots + r_k \); the balls are placed in a row by selecting one ball at a time (without replacement) from the set; test whether the resulting sequence of colours is compatible with the null hypothesis that each ball was selected at random from those balls that had not already been placed in the row.*

The idea is that if for each state \( X_i \) one peruse the data sequence \( X(1), \ldots, X(N) \) and write out the subsequence of the states that are immediately preceded by an occurrence of \( X_i \). The sequence contains \( R_i \) entries, of which \( n_{ij} \) are followed by the state \( X_j, 1 \leq j \leq k \). If a MCM of order 1 underlies the original sequence then the subsequence should be a random permutation.

This lands in an simple and elegant way to find out whether a sequence comes from a MCM of order 1 or not by checking randomness in the subsequence. This can be done by checking whether the relative frequency of each state in these subsequences changes significantly along the sequence. A suitable test would be to divide the subsequence into \( K \) parts and then the count the number of each state appearing in each part of the subsequence. Here then the alternative hypothesis is that the proportion changes between the parts of the subsequence.

<table>
<thead>
<tr>
<th>State</th>
<th>I</th>
<th>II</th>
<th>...</th>
<th>K</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_1 )</td>
<td>( f_{S_1I} )</td>
<td>( f_{S_1II} )</td>
<td>\ldots</td>
<td>( f_{S_1K} )</td>
<td>( T_{S_1} )</td>
</tr>
<tr>
<td>( S_2 )</td>
<td>( f_{S_2I} )</td>
<td>( f_{S_2II} )</td>
<td>\ldots</td>
<td>( f_{S_2K} )</td>
<td>( T_{S_2} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>\ldots</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( S_{15} )</td>
<td>( f_{S_{15I}} )</td>
<td>( f_{S_{15II}} )</td>
<td>\ldots</td>
<td>( f_{S_{15K}} )</td>
<td>( T_{S_{15}} )</td>
</tr>
<tr>
<td>Total</td>
<td>( T_I )</td>
<td>( T_{II} )</td>
<td>\ldots</td>
<td>( T_K )</td>
<td>( T )</td>
</tr>
</tbody>
</table>

Figure 1: Observed frequencies for sequence

If the sequence is random, the expected frequency for the the first state in the first part is given by \( T_I \times T_{S_1}/T \) and the observed frequency is given by \( f_{S_1I} \). The value of

\[
\sum (f - e)^2/e \quad (23)
\]

where \( f \) is the observed frequency and \( e \) is the expected frequency, and the sum is taken over the whole table.
The confidence with which the hypothesis can be rejected can then be approximated by comparing the sum, $\chi^2$, with a $\chi^2$ variate with $(N - 1)(K - 1)$ degrees of freedom. Where $N$ is the number of states and $K$ is the number of parts of which the subsequence is divided into.[13]

To be able to use this method on the students process, the sequence is obtained by looking at all students as a sequence of one student's progress at the time. This means that it is assumed that when an absorbing state is reached, the probability of going to the initial state for the next student equals 1.

4 Results

4.1 Model 1

The probability transition matrix is obtained by looking at each student's way through the study in the data. For each student taking a certain step $i \rightarrow j$ a counter at position $ij$ in the matrix adds one. When the whole data set is added to the model, each row is weighted to sum up to one.

The probability transition matrix for the initial model is then calculated to be:

$$P_{st} = \begin{pmatrix}
    14 & 3545 & 31 & 19 & 3545 & 6 & 3545 & 0 & 175 \\
    0 & 175 & 467 & 1 & 0 & 0 & 0 & 0 & 97 \\
    0 & 0 & 467 & 1 & 0 & 0 & 0 & 0 & 97 \\
    0 & 0 & 0 & 175 & 31 & 1 & 0 & 0 & 97 \\
    0 & 0 & 0 & 0 & 175 & 31 & 1 & 0 & 97 \\
    0 & 0 & 0 & 0 & 0 & 175 & 31 & 1 & 97 \\
    0 & 0 & 0 & 0 & 0 & 0 & 175 & 31 & 97 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 175 & 97 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 175 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}$$

(24)
It is also interesting to see how long a randomly selected student is expected to stay at a stage $i$ until entering the stage $j$. This is obtained directly from the fundamental matrix, $N$.

$$\mu_{\cdot j} = \begin{pmatrix}
11.92 & 9.98 \\
11.7 & 9.59 \\
11.3 & 9.34 \\
10.4 & 8.6 \\
9.51 & 7.86 \\
8.37 & 6.93 \\
6.87 & 4.97 \\
5.6 & 3.91 \\
4.62 & 3.8 \\
7.21 & 5.78 \\
8.49 & 3.6 \\
9.17 & 6.28
\end{pmatrix}$$

$$\mu_{\cdot i} = \begin{pmatrix}
11.92 & 9.98 \\
11.7 & 9.59 \\
11.3 & 9.34 \\
10.4 & 8.6 \\
9.51 & 7.86 \\
8.37 & 6.93 \\
6.87 & 4.97 \\
5.6 & 3.91 \\
4.62 & 3.8 \\
7.21 & 5.78 \\
8.49 & 3.6 \\
9.17 & 6.28
\end{pmatrix}$$
The graduation - withdrawal probabilities are calculated as the absorption probabilities from a given transient state. For this purpose the fundamental matrix, (13), as well as sub-matrices $R$ gathered from (12) were used. This gives the result:

$$f_{-07} = \begin{pmatrix} 0.5599 & 0.4381 \\ 0.6002 & 0.3999 \\ 0.6406 & 0.3596 \\ 0.6612 & 0.3382 \\ 0.6955 & 0.3037 \\ 0.7154 & 0.2862 \\ 0.7481 & 0.2496 \\ 0.7618 & 0.2377 \\ 0.772 & 0.2266 \\ 0.7456 & 0.2529 \\ 0.4226 & 0.5759 \\ 0.5218 & 0.479 \end{pmatrix} \quad f_{07-} = \begin{pmatrix} 0.6173 & 0.3865 \\ 0.6529 & 0.3458 \\ 0.7064 & 0.2934 \\ 0.7303 & 0.2713 \\ 0.7606 & 0.2395 \\ 0.7697 & 0.2288 \\ 0.8028 & 0.1965 \\ 0.816 & 0.1851 \\ 0.8293 & 0.1699 \\ 0.7932 & 0.2098 \\ 0.8013 & 0.1979 \\ 0.3743 & 0.6236 \\ 0.485 & 0.5165 \end{pmatrix}$$  

(27)

### 4.2 Model 2, comparing male and female students

Comparing female and male students performance and study process can be interesting for many reasons. Both for planning and forecasting if the gender trends changes. If there is a significant difference in the performance gender-related average, it can be in the schools interest to promote itself towards the better performing sex. It can also be of interest to pinpoint what it is that makes one sex perform better than the other, so that the education can be better for everyone.
number of new students, and the rest of the elements will be zero. Then Chapman for upcoming classes. In this case the first element of the transition matrix can as mentioned earlier also be used for prediction of how a new generation of students will go through their education. If the prediction is accurate, this information is valuable for the institution when it comes to for example planing for upcoming classes. In this case the first element of the $p^{(0)}$ vector will contain the number of new students, and the rest of the elements will be zero. Then Chapman Kolmogorov gives that the number of students in each state after $n$ years should be

$$p^{(n)} = p^{(0)}P^n. \quad (30)$$

To refine the data before creating the transition matrix $P$ will probably give a more accurate prediction. It can for example be a good idea to only use the data for a
particular sex or program. The validation described later can be used to see which criteria that gives better predictions.

<table>
<thead>
<tr>
<th>State</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>...</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>900</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>811</td>
<td>13</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
<td>703</td>
<td>33</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>640</td>
<td>45</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>580</td>
<td>56</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>19</td>
<td>33</td>
<td>11</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>28</td>
<td>428</td>
<td>68</td>
<td>16</td>
<td>11</td>
<td>8</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>29</td>
<td>400</td>
<td>72</td>
<td>21</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>11</td>
<td>47</td>
<td>378</td>
<td>81</td>
<td>26</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>11</td>
<td>15</td>
<td>25</td>
<td>55</td>
<td>309</td>
<td>252</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>57</td>
<td>38</td>
<td>30</td>
<td>14</td>
<td>7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>32</td>
<td>30</td>
<td>24</td>
<td>21</td>
<td>18</td>
<td>20</td>
<td>19</td>
<td>25</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>24</td>
<td>51</td>
<td>38</td>
<td>42</td>
<td>34</td>
<td>55</td>
<td>43</td>
<td>35</td>
<td>25</td>
<td>17</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>23</td>
<td>34</td>
<td>60</td>
<td>167</td>
<td>269</td>
<td>548</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>39</td>
<td>86</td>
<td>117</td>
<td>146</td>
<td>164</td>
<td>181</td>
<td>197</td>
<td>212</td>
<td>225</td>
<td>239</td>
<td>344</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Example of prediction of the progress for 900 students starting in semester 1.

### 4.4 Results of validation

To get a fair set of testdata the test is applied to the students starting their studies between year 1993 and 2005. Thus many students who begin their studies after 2007, will not have made it to graduation until 2016.

#### 4.4.1 First approach of \( \chi^2 \)-test

Let \( x_0 \) be the vector containing all students who are in the system the autumn semester of 2000.
One can see that the predicted values follow the same pattern in size, but that some values are more off than others. This results in the following values of $Q_{obs}$:

<table>
<thead>
<tr>
<th>$i \setminus t$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>6.0</td>
<td>0.5</td>
<td>2.0</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>22.0</td>
<td>4.5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>141.0</td>
<td>3.9</td>
<td>19.0</td>
<td>8.0</td>
</tr>
<tr>
<td>4</td>
<td>9.4</td>
<td>93.0</td>
<td>13.0</td>
<td>111.0</td>
</tr>
<tr>
<td>5</td>
<td>69.0</td>
<td>18.0</td>
<td>100.0</td>
<td>19.0</td>
</tr>
<tr>
<td>6</td>
<td>5.2</td>
<td>44.0</td>
<td>38.0</td>
<td>39.0</td>
</tr>
<tr>
<td>7</td>
<td>188.0</td>
<td>2.2</td>
<td>55.0</td>
<td>27.0</td>
</tr>
<tr>
<td>8</td>
<td>12.0</td>
<td>11.0</td>
<td>6.7</td>
<td>29.0</td>
</tr>
<tr>
<td>9</td>
<td>0.86</td>
<td>6.1</td>
<td>2.7</td>
<td>0.01</td>
</tr>
<tr>
<td>10 (Break)</td>
<td>1.4</td>
<td>0.045</td>
<td>0.05</td>
<td>0</td>
</tr>
<tr>
<td>11 (Inactive)</td>
<td>0.48</td>
<td>0.016</td>
<td>0.37</td>
<td>1.5</td>
</tr>
<tr>
<td>12 (Exchange)</td>
<td>3.7</td>
<td>3.5</td>
<td>2.9</td>
<td>4.7</td>
</tr>
<tr>
<td>13 (Graduation)</td>
<td>1.8</td>
<td>0.99</td>
<td>0.056</td>
<td>0.52</td>
</tr>
<tr>
<td>14 (Withdrawal)</td>
<td>33.0</td>
<td>20.0</td>
<td>42.0</td>
<td>78.0</td>
</tr>
<tr>
<td>$Q_{obs}$</td>
<td>444.89</td>
<td>231.38</td>
<td>288.23</td>
<td>315.75</td>
</tr>
</tbody>
</table>

Figure 4: $Q_{obs}(i)$ and $Q_{obs}$ for the same $t$ as above.

Since $Q_{obs}$ is quite large, the hypothesis $H_0$ can be rejected. One reason for the errors in
the model is that the data is not homogeneous enough. If the model is broken down to only looking at boys at the engineering physics program, F, and put $t_0$ to the autumn semester of 2002. The following result is obtained:

<table>
<thead>
<tr>
<th>$i\backslash t$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.56</td>
<td>0.33</td>
<td>4.0</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>4.5</td>
<td>0.24</td>
<td>25.0</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>0.28</td>
<td>36.0</td>
<td>2.1</td>
<td>5.3</td>
</tr>
<tr>
<td>5</td>
<td>4.0</td>
<td>2.0</td>
<td>81.0</td>
<td>3.9</td>
</tr>
<tr>
<td>6</td>
<td>1.4</td>
<td>7.0</td>
<td>5.9</td>
<td>11.0</td>
</tr>
<tr>
<td>7</td>
<td>3.2</td>
<td>3.4</td>
<td>2.8</td>
<td>5.0</td>
</tr>
<tr>
<td>8</td>
<td>0.7</td>
<td>0.82</td>
<td>1.0</td>
<td>0.76</td>
</tr>
<tr>
<td>9</td>
<td>0.067</td>
<td>0.059</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>10 (Break)</td>
<td>0.57</td>
<td>1.0</td>
<td>0.25</td>
<td>0.14</td>
</tr>
<tr>
<td>11 (Inactive)</td>
<td>0.018</td>
<td>0.96</td>
<td>0.38</td>
<td>0.91</td>
</tr>
<tr>
<td>12 (Exchange)</td>
<td>0.9</td>
<td>2.8</td>
<td>7.2</td>
<td>12.0</td>
</tr>
<tr>
<td>13 (Graduation)</td>
<td>0.17</td>
<td>0.66</td>
<td>0.73</td>
<td>0.16</td>
</tr>
<tr>
<td>14 (Withdrawal)</td>
<td>0.67</td>
<td>0.82</td>
<td>5.3</td>
<td>13.0</td>
</tr>
</tbody>
</table>

| $Q_{obs}$     | 18.0214| 56.1109| 136.8571| 54.6483 |

Figure 5: $Q_{obs}(i)$ and $Q_{obs}$ for the same $t$s as above.

This is better than the model above, but should still be rejected according to the $\chi^2$-test. This indicates that there is to much variation between in the "subsets" in the dataset to use the whole set for a prediction.

### 4.5 Checking validity of fitting a first-order Markov chain model to the dataset

The sequence is created by taking data for students starting between 1995 and 2005. Thus many students who begin their studies after 2007, will not have made it to graduation until 2016. The data for each student is lined up after another. In that way a sequence of students process is created. This results in a sequence of states where an absorbing state is followed by the first state. From this sequence a subsequence with immediate following states is created as proposed in the methodology section.

The frequency analysis was made by dividing the sub-sequence into 10 parts and checking whether the frequency of each state changes with time.

<table>
<thead>
<tr>
<th>State</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
<th>$S_5$</th>
<th>$S_6$</th>
<th>$S_7$</th>
<th>$S_8$</th>
<th>$S_9$</th>
<th>$S_{10}$</th>
<th>$S_{11}$</th>
<th>$S_{12}$</th>
<th>$S_{13}$</th>
<th>$S_{14}$</th>
<th>$S_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>13.43</td>
<td>38.37</td>
<td>34.88</td>
<td>18.74</td>
<td>16.18</td>
<td>66.38</td>
<td>45.72</td>
<td>54.41</td>
<td>24.92</td>
<td>1.40</td>
<td>33.30</td>
<td>96.91</td>
<td>29.08</td>
<td>6.29</td>
<td>3.33</td>
</tr>
</tbody>
</table>
This is to be compared with $\chi^2_{126} = 101.074$ to, with a 95% confidence, reject the hypothesis that the frequency changes with time. This indicates that we cannot reject the hypothesis that the data can be fitted to a Markov chain model.
5 Discussion

5.1 The result

The findings of this study indicate that the probability that a student successfully finishes his/her studies and graduates increases for each completed semester. This result is quite expected, since the students who do not drop out is probably more motivated than those who does. Students also tend to become more serious and ambitious during their progression in the study. By looking at the transition matrix, one can also see which semesters that seem to be harder to accomplish than other, i.e it is a lower probability of progression from semester X to Y than from T to Z.

It is realised that students who have announced study break are more likely to finish their studies than students who are inactive. An explanation to this can be that students in the first group have planned their study break for reasons such as internships or travel, while students who are inactive are more careless. By announcing study break you are guaranteed a spot at the programme when you which to revert to the study program, this is not the case for students which are inactive.

To be nominated for an exchange program, performance over the first years is required. Students who apply for exchange programs are often motivated and have shown good study results. It is therefore not surprising that a large part of the students who take a semester abroad will eventually graduate. In some cases is can be hard to find courses at the exchange university that fit ones profile, many students therefore chooses to study language and courses outside their actual education program. It is therefore common that they have more than one year left until graduation even if they take their exchange at the forth year of education. Ambitious students can sometimes be offered jobs before their graduation, this together with the above, can be one reason that about 20 percent of the students who makes it as far as the exchange program chooses to not finish their study and graduate.

In $\mu_{07}$ one can see that students who are enrolled in the 9th semester are expected to stay for 4.62 more semesters in the system. The report ”Ingenjörerna” from SCB[2], shows that only 51% of students who began their studies in the autumn of 2002 had graduated after 8 years, which equals 12 semesters.

When comparing students who began their studies before and after the Bologna process, autumn 2007, one notice that the probability of graduating is slightly higher for students in their first semester for the group starting after 2007. There is a possibility that this is due to the reform. It can also be a result of externalises. The financial crisis in 2008-2009 was followed by a recession, resulting in a less stable labour market, which can explain that people had a higher tendency to stay in school.

Female students tend, on average to perform slightly better than male students. This
result is quite expected, since almost all reports on the subject gives the same result. It is also shown in the data in the report "Ingenjörerna" from SCB[2].
6 Conclusion

A Markov chain have been developed to examine the progress of students at LTH. The model enables estimation of predicted duration of the study, probability of progression between different stages, expected time spent at a particular stage. It can also predict the progress of the students in the university at a particular time.

The performed analysis provide useful information for decision makers and managers of the educational institution to improve their processes. It can also be useful for for managers for policy makers at government agencies for supervising the effectiveness of existing educational policies. Information about the number of expected graduates in the near future can be useful for education planners, employers and other actors in the labour market. If the model is divided into different study programs it can provide information of how hard a program is, which can be interesting for to be students.

One weakness in the model can be that the Markov chain is time homogeneous. That means that temporal stability in the estimated transition probability matrix is assumed. The model presumes that there is no external effects that can influence students behaviour. The assumption is reasonable on a not to long time horizon, because student behaviour is not likely to change too much over short time.

Even though a quite large dataset of many students were used in this research, a limitation is that the results is bounded to only one institution. To be able to make comparisons between for example sex or between students who began their studies at different times, it would be interesting to use data from several institutions.[10]

As stated in paragraph 7.2 can the hypothesis that there is no significant difference between the observed value in the model and the expected value be rejected. One reason for this can be that the group of students in the data is not homogeneous enough. It was seen that the observed values were closer to the expected value when program and sex also were considered. To get even better values one should perhaps consider for example age of the students, grades and taken credits.

The prediction model gives, even if it is not accurate enough to pass validation, values not completely removed from reality. It can therefore still give a hint to what one can expect.

The validity of fitting a first-order Markov chain model to the data does though indicate that we cannot reject the hypothesis that a Markov chain model can be fitted to describe the students’ progression. This is probably the most important finding of this thesis. Thus it opens up for further research and the possibility to apply this kind of modelling on data from other schools or similar processes.
References


