Robert-Jan Bodewes ANR: 343259

Modelling and Forecasting Higher Education Demand in the Netherlands

MSc. Information Management Thesis

Supervisor: Dr. Arash Saghafi
2nd Reader: Dr. Martin Smits

Tilburg 2022

Abstract

The aim of this research is twofold. Firstly, this paper aims to identify an accurate time series forecasting method for higher education enrollments in the Netherlands. Secondly, this paper also identifies several determinants of higher education demand in the Netherlands. Location, institution ranking, and discipline have a statistically significant effect on enrollments. Moreover, this paper estimates enrollments on the discipline level with a mean absolute percentage error of as low as .13 percent using a double exponential smoothing approach. Furthermore, this thesis leverages publicly available data on the DUO (Dienst Uitvoering Onderwijs) API.

Preface

This thesis is written to meet the graduation requirements for the Information Management Master's program at Tilburg University.

I want to thank my supervisor Dr. Arash Saghafi, who provided guidance and feedback on the project, and with whom I had genuinely enjoyable meetings discussing the issues of this work. Secondly, I'd like to thank my parents, who made it possible to attend university. Moreover, I'd like to thank Thijs van Vugt for the support he offered me while combining this thesis with a full-time internship. Lastly, Jacopo Gutterer deserves some recognition here for taking time out of his busy schedule and helping me get the data I needed to complete this analysis.

- Robert-Jan Bodewes, June 2022

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Company Introduction

Studyportals B.V. is an Eindhoven-based company that provides an online education choice platform, listing more than 200,000 undergraduate, postgraduate, distance learning, preparation programs worldwide and other international education resources. The website also lists scholarships. The company charges clients (usually higher education institutions) who want to promote themselves on the platform. More specifically, the amount charged to a client is proportional to the number of students enrolled in the client's programs. In 2019, the site attracted 44 million users. Six hundred forty thousand students used the Studyportals platform to enroll that same year. Studyportals B.V. was originally founded in 2009 due to a lack of information surrounding English-taught HE programs in Europe. Currently, the Studyportals team consists of over 150 people. Studyportals has offices in Eindhoven, Bucharest, Manchester, Melbourne, and Boston.

This research was conducted in the "Analytics and Consulting" (ACT) business unit. ACT was founded in 2015. In short, this BU sells dashboards to universities. These dashboards show relative demand and supply for all disciplines and programs. For example, these dashboards can identify the demand for engineering courses in the Netherlands. Clients can then determine whether they should launch a program based on demand or competition. Moreover, ACT also provides consulting services to corporate clients. For example, ACT informs ASML on possible sources of engineering talent. Lastly, ACT sells its pageview data to other market insight firms like Intelligence Group.

1.2 Topic Introduction

As of late, predictive analytics techniques have been applied by organizations to "forecast your buying decisions, the likelihood of you leaving a job, your health and date of death" (Siegel, 2013). Awareness surrounding predictive analytics increased following the controversy around Cambridge Analytica, which leveraged Facebook data to influence the outcome of the Brexit referendum in 2016 (Kaminska, 2020). Predictive analytics includes "statistical models and other empirical methods that are aimed at creating empirical predictions" and "assessing the quality of those predictions" (Shmueli & Koppius, 2011). LaRivere et al. (2017) list several use cases for predictive analytics. Firstly, predicting demand is the foremost use case. Accurate demand forecasts are valuable because

storing inventory is costly, and stockouts are disastrous. Secondly, predictive analytics contributes to pricing decisions. More specifically, predictive analytics helps estimate the price sensitivity of a customer. Providing a single price is inefficient because it alienates customers who could have been profitably served through price discrimination. Predictive analytics enables companies to target customers based on their price responsiveness (through targeted discounts) instead of demographics. Thirdly, predictive analytics allows predictive maintenance. Predictive maintenance is significant because it prevents supply chain disruptions. Moreover, the main cost associated with machine downtime is the opportunity cost (forgone profits).

Increasingly, predictive analytics is being adopted in a higher education (HE) context. Researchers at HE institutions regularly forecast enrollments (Brinkman McIntyre, 1997). Enrollment forecasts are "fundamental elements of planning and budgeting at any higher education institution that depends on student enrollments" (Brinkman McIntyre, 1997). Tuition policy, budget forecasts, faculty staffing, and institutional closures depend on enrollment forecasts (Weiler, 1987). Predictive analytics is also being used in other areas. According to one report by McKinsey, Northeastern University leveraged analytics to boost its U.S. News & World Report rankings from 115 in 2006 to 40 in 2017. Northeastern University also uses a predictive model to identify "best-fit" students. The same report also claims University of Maryland University College (UMUC) used analytics to achieve a 20 percent increase in new student enrollment while reducing its marketing expenditure by 20 percent. In 2013, UMUC analysts noticed that significant advertisement spending generated a low amount of leads. Upon further investigation, the analysts spotted a bottleneck. More specifically, the UMUC call center was understaffed. After investing in new call-center capabilities, UMUC experienced a 20% rise in student enrollments while decreasing advertising spending by 20% (Krawitz et al., 2021).

1.3 Problem Description

Although many papers discuss methods for forecasting higher education (HE) enrollments, many do not produce actionable insights for Studyportals clients. Studyportals clients are not interested in what time series model is used. Instead, they are interested in the predictions themselves. Secondly, enrollment predictions for 2023 per educational field do not exist for the Netherlands. The following paragraphs will cover existing papers related to HE enrollment forecasting. The subsequent literature shows that enrollment forecasting has not been attempted for the Netherlands (where several key clients are located).

Lavilles and Arcilla (2012) forecast enrollments for six community colleges in the United States. This study uses regression, autoregression, and a three-component model to

project enrollments. These time-series models had a 20% lower MAPE (mean absolute percentage error) than a naive model used at the time. Brinkman and McIntyre (1997) contrast enrollment forecasting methods. The most important factors affecting enrollment include demographic factors (age structure, racial and ethnic composition), economic factors (disposable income, unemployment rate), and competitor-related factors (tuition fees of competing programs). The study concludes that enrollments can be accurately "forecast entirely based on prior enrollment patterns." This thesis will also adopt one of the curve-fitting techniques mentioned in Brinkman and McIntyre (1997).

1.4 Problem Statement

The problem is that data on enrollments by discipline exists for the Netherlands, but forecasts do not. HE institutions have some data on the current demand for each discipline across different countries. However, it is better to let forward-looking metrics inform decision-making (Rajni & Malaya, 2015). Moreover, many HE forecasting literature projects enrolments for a single institution. Studyportals clients are more interested in how enrolments per discipline will change over time. Hence, this paper will evaluate a time series model's ability to predict enrolments for every discipline at the country level.

The solution will be a report which forecasts enrollments for the Netherlands. In addition, this report will also include the determinants of HE demand. More specifically, institutions should know what drives demand for their courses. Studyportals clients are clustered around several key markets (The Netherlands, Germany, Italy, UK). However, this study only focuses on the Dutch HE market. Moreover, the error needs to be relatively low for the findings to be actionable. Since enrollment figures are forecasted for all educational fields, the results are relevant for all HE institutions in the Netherlands. In addition, Studyportals is interested in forecasting as a possible future line of business.

1.5 Research Question

In order to solve the problems described in the previous section, the following question must be answered:

Which time series model generates the most accurate enrollment forecasts by discipline?

Additionally, to understand the nature of demand within the Dutch HE sector, the following must also be answered:

What are the determinants of demand for higher education in the Netherlands?

Furthermore, these research questions can be divided into the following sub-questions. These include the following:

1. Which educational fields will have the most demand in the future?
2. Which educational fields will have the least demand in the future?
3. What is the best metric for assessing the accuracy of a time-series model?
4. How much enrollment data is required to generate an accurate forecast?
5. Does enrollment data generally have a trend of any kind?

The first two sub-questions reflect the interests of HE decision-makers. The following sub-questions are more technical in nature.

1.6 Methodology and Approach

This section introduces the research method adopted in this paper. In short, this thesis aims to identify a model which can predict enrollments with sufficient accuracy. More specifically, this thesis focuses on a group of forecasting methods labeled "time-series" by Sinuany-Stern (2021). These time-series methods include; moving averages, exponential smoothing, fuzzy time series, and autoregressive methods (ARIMA). Mean absolute percentage error (MAPE) will be used to evaluate each model. MAPE is widely adopted in HE forecasting research for model evaluation. Moreover, the most accurate forecasts will be produced for the relevant HE decision-makers. Furthermore, identifying determinants of demand for HE will be done using a weighted least squares regression.

Secondly, this method mainly uses data the DUO (Dienst Uitvoering Onderwijs) API. DUO is a subsidiary of the Dutch Ministry of Education and was founded in 2010. DUO is an institution that provides student grants, collects tuition fees, and handles the application process, in addition to many other tasks. The API contains enrollment data by educational field for all Dutch HE institutions. Although enrollment data only goes as far back as 2015, the API is updated regularly to include the most recent statistics. This distinguishes the DUO API from other sources of enrollment data like UNESCO (UIS). Additionally, the data is quite granular. More specifically, enrollments are recorded at the program level per institution.

1.7 Thesis Structure

This study is structured as follows. The first chapter introduces the topic and motivation. Secondly, Chapter 2 covers relevant literature and forecasting methods. Furthermore, forecasting and modeling methodologies are included in Chapter 3. Modeling results are highlighted in Chapter 4, and forecasts are shown in Chapter 5. Chapter 6 summarizes all findings and Chapter 7 recommends future research efforts.

2. Literature Review

This chapter covers relevant literature that relates to the topic mentioned above. Section 2.1 describes several forecasting methods. Secondly, section 2.2 summarizes the findings of past literature, which investigates the determinants of higher education demand. Thirdly, section 2.3 includes hypothesis development.

2.1 Relevant Forecasting Methods

Sinuany-Stern (2021) divides forecasting methods into six main groups. These include the following:

- **Time series** includes moving averages, exponential smoothing, fuzzy time series, and ARIMA.
- **Ratio forecasting** methods incorporate Markov analysis and the cohort-survival method.
- **Regression analysis** is the most popular forecasting method.
- **Simulation** includes system dynamics amongst other approaches.
- **Data mining** consists of regression, decision trees, random forests, neural networks, genetic programming, and grey models.
- **Questionnaires** are mostly used in marketing efforts.

The following is an overview of what HE forecasting papers use each of the aforementioned methods. Note that most papers on this topic adopt multiple forecasting models. However, each paper has a "main" method that it espouses.

2.1.1 Time Series Techniques

Moving Averages (MA)

In its simplest form, the forecast F_{t+1} is a function of the average of the observations over N periods:

$$\sum_{j=t-N+1}^{t} \frac{X_t}{N}$$

Lavilles and Arcilla (2012) adopt MA as one of their forecasting methods. They forecast

enrolments for six community colleges in the United States and find that exponential smoothing methods outperform MA in terms of accuracy. More specifically, exponential smoothing yielded a 20.5% lower mean absolute percentage error (MAPE) than the MA approach.

Exponential Smoothing

The forecast F_{t+1} is a function of the last forecast, F_t, the actual observation, X_t, and a smoothing constant, α:

(1) $F_{t+1} = \alpha X_t + (1 - \alpha)F_t$

This approach is popular in HE forecasting literature. For example, Anggrainingsih et al. (2015) adopt exponential smoothing to predict the number of Sebelas Maret University (UNS) website visitors. The authors achieved an error of 12.95%. The authors compare single exponential smoothing (SES), double exponential smoothing (DES), and triple exponential smoothing (TES) forecasts. DES and TES expand on the equation shown above. DES (a.k.a Holt's model) assumes the data has a linear trend. DES is expressed the following way mathematically:

(2) $F_t = \alpha D_t + (1 - \alpha)(F_{t-1} - T_{t-1})$
(3) $T_t = \gamma(F_t - T_{t-i}) + (1 - \gamma)T_{t-i}$
(4) $F_{t+1} = F_t + T_t$
Where:
T_t is the trend in period t and γ is the smoothing constant for trends.

Anggrainingsih et al. (2015) used mean absolute percentage error (MAPE) to determine the accuracy of each model. Data pattern analysis is required to choose between SES, DES, and TES. For example, SES is appropriate if no seasonality or linear trend is present. When adopting SES constant refinement of the smoothing coefficient is required (Anggrainingsih et al., 2015). If a linear trend is present, then DES is appropriate and the optimum and needs to be solved for. Lastly, TES is the best option if the data exhibits seasonality. More information on how to solve for smoothing coefficients is provided in chapter 3.

Anggrainingsih et al. (2015) use web traffic statistics from the UNS website between January 2008 and June 2014. All in all, the authors consider a MAPE less than 10% a good result. Using a TES model achieves a MAPE of 12,45%.

Fuzzy Time Series

Fuzzy time series models produce an interval prediction rather than a point forecast. Fuzzy time series models adopt fuzzy logic where a particular item may belong and simultaneously not belong to the same set, such that its membership is a value on the interval [0,1]. Fuzzy time series was originally introduced by Song and Chissoum (1993).

Aladag et al. (2010) claim that high order time series models can produce more accurate forecasts than first-order models. However, adopting high-order fuzzy time series models requires complex matrix operations. The authors implement neural networks to eliminate the need for these matrix operations. Also, neural networks can identify nonlinear fuzzy relationships, generating more accurate forecasts. The authors use enrollment data from the University of Alabama from 1971 to 1992. Data from 1989 to 1992 was used as test data. The authors forecast enrollments with second, third, and fourth-order fuzzy time series models. The authors also experiment with the size of the hidden layer of the neural network that fuzzyfies and defuzzyfies the observations. The results suggest that using neural networks to implement high-order fuzzy time series models yields a better result than previous approaches. Moreover, the authors include a table ranking conventional time series forecasting methods. These results are presented in Figure 1. ARIMA is the most accurate conventional enrollment forecasting method by a large margin.

ARIMA (Auto Regressive Integrated Moving Average)

ARIMA belongs to a group of models which generates predictions solely based on past values. Three types of ARIMA models exist. These include; ARIMA (non-seasonal), SARIMA (seasonal ARIMA), and SARIMAX (seasonal ARIMA with exogenous variables). Chen (2008) uses an ARIMA and a linear regression to forecast enrollments for Oklahoma State University (OSU). Unforeseen changes in enrollment patterns significantly impact the accuracy of ARIMA forecasts. Hence, researchers have to select appropriate variables. For example, HE enrollment is affected by high school graduation rates. Specific migration patterns also affect HE enrollments (like net in-migration). The condition of the economy also determines HE enrollments to an extent. More specifically, enrollments are affected by the "rate of increase in college tuition relative to growth in family income, trends in federal and state financial aid, and employment prospects for new graduates" (Chen, 2008).

In order to identify the most appropriate ARIMA (p, d, q) model, Chen (2008) observes the enrollment series, the autocorrelation function (ACF), and the partial autocorrelation function (PACF) of the residual series. p is the pth order of the autoregressive effect. d is the number of differencing required to make the time series stationary. q is the qth order of

the moving average.

Upon choosing the correct p, d, and q, the author estimates the coefficients of the independent variables using Melard's parameter estimation. In short, this estimation method involves determining the coefficients such that the estimated enrollment series is nearly identical to the actual enrollment time series. All in all, the authors find that an ARIMA (1, 1, 0) with 7 independent variables forecasts enrollments with a MAPE of 2.11%.

2.1.2 Ratio Methods

Cohort-survival Method

Cohort-survival is widely used in enrollment forecasting because of "its simplicity and intuitive meaning" (Sinuany-Stern, 2021). The approach calculates retention ratios, which are a function of academic attainment (grade point average). The total enrollments are calculated at the lowest GPA level for several years. This low-performing group of students is "aged" through the following grades. More specifically, if 60% of the graduating high-school students in a given region enroll in HE institutions. Predicting HE enrollments for the following year is given by $F_{t+1} = 0.6Z_t$ where Z_t is the number of high school students about to graduate. Burkett (1985) uses cohort survival to predict freshman enrollment at the University of Mississippi during 1985-1989, based on data of 7-12th grade students in the state of Mississippi and the number of out-of-state freshmen. Murtaugh et al. (1999) use survival analysis to predict student retention at Oregon State University (OSU) between 1991 and 1996. More specifically, the authors analyze the number who leave OSU prior to graduating to identify the main sources of student attrition. All in all, their sample consists of 8,867 students. Over four years, 40% of students withdraw prior to graduating.

Markov Analysis

The advantage of Markov analysis is that it is "relatively straightforward to set up and does not require extensive use of computers to find parameters like steady-state probabilities" (Sinuany-Stern, 2021). A Markov analysis is based on a system with K possible states. In addition, Markov analysis involves calculating transition probabilities where P_{ij} is the probability of moving from state i to state j, based on historical data. Bessent and Bessent (1980) adopt Markov analysis to model the progression of doctoral students at the University of Texas at Austin. Their dataset contains enrollment data from 1969 until 1978. For example, the authors determine that the probability of a student advancing to candidacy, given that that student has enrolled in a doctorate program is 0.27. In addition, the probability of graduation for candidates is 0.18. Hence, the attrition rates for doctorate students are astonishingly high. The authors find that there is no effective way to expedite

the graduation process for doctorate students without sacrificing the quality of the program. Instead, Bessent and Bessent (1980) recommend developing a more selective admissions process. Kwak et al. (1986) studied student retention in the Transportation, Travel, and Tourism (TTT) department at Parks College in St. Louis. The Markov analysis used the grade and enrollment data of 117 students between 1978 and 1984. Parks College divides its academic calendar into three trimesters. The second trimester starts in May and ends in July. Unsurprisingly, the authors find that 29% of the winter trimester's students take the summer trimester off. Additionally, the authors estimate that out of all full-time students 15.6% will drop out before the following year.

2.1.3 Regression Analysis

A regression detects the relationship between a dependent variable (to be forecasted) and a set of independent variables. Regression models are seldom used for HE enrollment forecasting (Sinuany-Stern, 2021). For example, Sinuany-Stern (1984) uses linear regression to predict the number of students in each campus of Cuyahoga Community College in Ohio. The main independent variables in the regression were: the rate of unemployment in the country, the amount of financial aid, the opening of a new campus facility, and the number of high school graduates in the surrounding area.

2.1.4 Simulation Models

Simulation models imitate a system or a process to observe its behavior over time. Moreover, simulation allows researchers to observe a system's behavior under different scenarios. Simulation models consist of multiple components. The behavior of these components is prescribed with a series of equations.

1. System Dynamics (SD) uses differential equations to describe the relationships between each subsystem. It was developed by Forrester (1961). It enables policy analysis. More specifically, SD allows researchers to simulate the effects of a policy change. The most popular SD software packages include Dynamo, iThink, Powersim, and Vensim.
2. Discrete event simulation models a system as a discrete series of events. Each event marks a change in the state of the system.
3. Agent-based simulation (ABS) involves modeling the interactions between autonomous agents using discrete event simulation.

Xiao and Chankong (2017) model the demand for medicine students and the supply of

medical talent in universities using system dynamics. The supply of medical students is a function of birth rate, mortality rate, investment, number of students enrolled in medical fields, graduation rates, attrition rates, and the state of the job market. More specifically, the authors aim to predict the number of medical talents in Jiangsu Province in China using data from the Jiangsu Provincial Statistics Bureau (2010-2014). The authors achieve a percentage error below 0.5%. However, this approach has several limitations. Firstly, the model performs poorly when major policy changes occur. Secondly, estimating the time lag of each variable in the model is complex. For example, some time is required for a graduate to achieve medical doctor status. This time lag exists because some residency period is required for a graduate to achieve doctor status. Thirdly, the authors make numerous assumptions. For example, Xiao and Chankong's (2017) model assumes that doctors only come from medical colleges. However, alternative sources like web-based medical training programs may also exist.

Saltzman and Roeder (2012) also adopt an SD approach to simulate the progression of business students in a college program. San Francisco State University (SFSU) implemented deep budget cuts following the 2008 recession. Hence, SFSU had to be more diligent when spending its scarce resources. More specifically, the authors observed the impact of changes in the business curriculum on students. In addition, the authors wanted to identify bottleneck courses in the SFSU college of business. They compare and contrast the simulation approach with the Markov analysis approach. Firstly, the Markov assumption states that any state is solely dependent on the previous state. To construct a realistic model, the model must contain all variables that could change from one state to another. However, including all variables leads to the "curse of dimensionality." One advantage of simulations is that they allow the system to behave differently over time. By contrast, the state transition probabilities remain fixed in a Markov analysis. For example, application rates and course offerings may change over time. SD enables this dynamism. Moreover, entities are not independent in real life. For example, the authors note how both female and male students may compete for the same fixed number of positions in a class or program. All in all, Saltzman and Roeder successfully modeled the flow of students through the college of business at SFSU. However, some assumptions in their model caused some discrepancy between predicted values and reported values. For example, some students require additional courses to fulfill their graduation requirements.

2.1.5 Data Mining (DM)

In short, DM combines computer science, statistics, and analytics. More specifically, DM uses several algorithms for forecasting. These include linear modeling, regression analysis, decision trees, random forests, density estimation, neural networks, KNN (k-nearest

neighbor), variance analysis, text mining, principal components, hierarchical clustering, and density-based methods. In addition, DM automatically tests various methods and selects the most performant one (Sinuany-Stern, 2021).

Below are some descriptions of numerous DM sub-methods:

1. Random forests are a type of ensemble machine learning algorithm. This algorithm simultaneously constructs numerous decision trees while training and returning the class that is the mode (classification) or the mean prediction (regression).
2. Neural networks use multiple layers of nodes to generate predictions from raw data. Aladag et al. (2010) use a neural network to help predict enrollments for the University of Alabama.
3. Genetic Programming (GP) involves searching for the most performant program in the space of all programs. Amber et al. (2015) use genetic programming to forecast electricity consumption by HE buildings in the HE sector. Generally, GP consists of three key components. Firstly, the terminal pool contains all input variables required to generate predictions. Secondly, the function pool contains all functions used to make the mathematical formulas to solve the GP problem. Thirdly, the fitness function identifies how performant each mathematical function in the function pool is. Amber et al. (2015) found that their GP model slightly outperformed a comparable multiple regression.
4. Grey models combine "the theoretical structure of the forecasting model with data to complete the model" (Sinuany-Stern, 2021). For example, this may include combining a linear regression and a neural network.
5. Decision trees are machine learning algorithms that can solve classification and regression problems. Every decision tree has a root node, where inputs pass through. This root node is divided into additional sets of nodes. Nodes that do not split into other nodes are called leaf nodes. Lastly, Gini impurity is used to assess whether observations have been classified correctly.

2.1.6 Questionnaires and Surveys

Questionnaires and surveys are often used to complement one of the aforementioned techniques in HE forecasting literature. Questionnaires are a set of questions given to a group of individuals who are the subject of a statistical study. Surveys are not used for forecasting but to understand students' behavior or assess different teaching methods' effectiveness.

For example, Sinuany-Stern (1976) used questionnaires to understand the 20% growth

in enrollments at Ohio community college. The regression model was adapted based on survey responses. Bannerjee and Igbaria (1993) use surveys to determine how organizations inform their computer capacity planning decisions. More specifically, 1400 directors of computer centers at US and Canadian institutions were involved. The response rate was 13.1 percent.

2.2 Determinants of Higher Education Demand

This section will cover the relevant factors that affect higher education enrollments. More specifically, the following articles study the economics of higher education.

Mixon et al. (1994) studied the determinants of out-of-state migration in the U.S. The authors sampled 220 (4-year) institutions for 1990. Colleges were selected at random from all states. In addition, both private and public colleges were adequately represented. In terms of explaining the variation in out-of-state enrollments, four variables are significant at the 99% level. These include tuition fees, entrance difficulty, whether the institution is private or public, and whether the institution is "historically black." Mainly, Mixon et al. (1994) find that the income effect of tuition for out-of-state students is more significant than the price effect. Since out-of-state applicants are not sensitive to price changes, university administrators can raise fees to increase revenues. Moreover, the exclusivity of an institution heavily affects demand. Neill (2009) studies the effect of tuition fee changes on the demand for universities in Canada. The author takes an instrumental variable approach to address endogeneity concerns. In Canada, tuition fees are endogenous because the provincial government determines the cost of education. In addition, private higher education institutions are relatively rare in Canada. Hence, the tuition fee policy affects most Canadian universities. As a result, the author uses the political party in power in a given province for a given year as an instrument. In addition, five different parties have been in control during the relevant period. Moreover, the political party is unlikely to affect HE demand through channels other than the tuition fee, making it a valid instrument. All in all, a $1000 increase in tuition corresponds to a 2-3 percentage point decline in enrollments. Prospective students with parents who do not have a tertiary degree are affected the most by tuition fee changes.

In addition to tuition fees, Neill (2009) also identifies other likely determinants of demand. For example, the unemployment rate for high school graduates represents forgone income if students choose not to pursue higher education. The author finds that a higher unemployment rate for high school graduates contributes to more enrollments. This is expected since the unemployment rate for high school graduates represents the opportunity cost of not pursuing higher education. If jobs are scarce for high school graduates, there is a larger

incentive to enroll in universities.

Fredriksson (1997) studies the impact of returns to higher education in Sweden on enrollments. This form of analysis is possible in Sweden due to the significant wage compression that occurred during the 1960s and 1970s. More specifically, the difference between high school graduate salaries and university graduate salaries changed substantially over time compared to other countries. This salary difference is referred to as the "wage premium" (return to higher education). In the late 1960s, the wage premium was at 44 percent. At its lowest, wage premium was below 20 percent by 1981. In Sweden, this wage premium is a function of the number of employment opportunities, the tax system, and subsidies offered to students. Firstly, unemployment benefits were found to have no effect on HE demand. Secondly, the primary determinant of higher education demand is the level of the wage premium.

Mueller and Rockerbie (2004) model HE demand using tuition fees, income per capita, and other variables related to benefits accruing to university graduates. The country determines much of how the HE supply and demand curves interact. In Canada, most universities are public and face tuition fee constraints. Hence, institution administrators cannot set tuition fees such that the market for higher education clears. As a result, there is constant excess demand for university slots in Canada. Furthermore, the authors assume that Canadian tuition fees are exogenous. The authors can estimate an aggregate demand function using a least-squares model by making this assumption.

Additionally, Mueller and Rockerbie (2004) use application data instead of enrollment data. Application data is more representative of HE demand if not all students who apply for a given program are admitted. This is because enrollments are affected by supply changes. For example, if the amount of tertiary education spending increases or alumni donations increase, enrollments are affected. However, there are drawbacks associated with measuring demand through application counts. For example, students may apply to as many institutions as they want. This may make it seem like demand is high when students are just less selective in terms of applying to institutions. In addition, the opportunity cost of attending university for a year is equal to the yearly wage of an unskilled worker. Hence, the authors include the annual wage for a service sector employee in their model.

Mueller and Rockerbie (2004) find that ranking has a statistically significant effect on the number of applications. More specifically, a one-place increase in Maclean's ranking corresponds to 1.3% more applications for the institution. Also, the income elasticity of median income is statistically insignificant, making higher education an inferior good for high school students. More importantly, the weekly unskilled wage and the real interest

rate has no statistically significant effect on applications. Hence, the opportunity costs of attending university have no impact on the decision to attend.

Christofides et al. (2001) investigate whether higher education subsidies disproportionately benefit higher income families. They find that parents' income level has a statistically significant effect on postsecondary attendance rates. Moreover, this study includes provincial dummies to capture the effect of regional differences in demand. These dummies are statistically significant. Surprisingly, tuition fees are not a factor. However, the lack of variance in tuition fees during the given time period may explain why tuition fees are insignificant. Furthermore, between 1975 and 1993, the regressive impact on tuition fee subsidies decreased.

Huijsman et al. (1986) identify several determinants of demand for higher education in the Netherlands. The author explains a number of unique features that the Dutch education system has. For example, all universities are public. Generally, the type of highschool one graduated from determines whether admission is granted or not. Furthermore, institutions do not issue financial aid. However, the government does provide interest-free loans and grants to HE students. Moreover, the authors analyze enrollments between 1950 and 1982. Finally, the authors find that per capita income has a significant effect on enrollment.

2.3 Hypothesis Development

This section outlines some hypotheses to be tested in later chapters. More specifically, these attempt to identify drivers of higher education demand.

Firstly, numerous papers comment on the relationship between location and enrollment. For example, Christofides et al. (2001) find that the degree of urbanization and enrollment are related. Hence, the first hypothesis investigates the following:

Hypothesis 1 (H1): Location has a statistically significant impact on enrollments.

Secondly, many theories comment on the link between institution ranking and demand. Mueller Rockerbie (2005) finds that in Canada a superior ranking corresponds to more demand.

Hypothesis 2 (H2): A superior institution ranking contributes towards more enrollments.

3. Methodology

This chapter covers the various methodologies adopted by this research. Firstly, section 3.1 explains the data and models used for modeling the demand for higher education in the Dutch HE sector. Section 3.2 covers enrollment forecasting methodologies.

3.1 Modelling HE Demand and Data

Firstly, the HE demand analysis combines numerous data sources to explain the variation in registrations at the institution-discipline-year level. Hence, enrollments were collected for every discipline at a Dutch HE institution between 2017 and 2021. This enrollment data is publicly available on the DUO (Dienst Uitvoering Onderwijs) API. Secondly, rankings per institution come from the Studyportals AWS data warehouse. The data engineering team scrapes popular rankings each year and stores this information in an AWS Redshift environment. This paper uses QS (Quacquarelli Symonds) institution rankings. QS rankings are based on academic reputation, employer reputation, faculty/student ratio, citations per faculty, international faculty ratio, and international student ratio. Numerous media outlets and institutions also produce rankings. Times Higher Education World University Rankings and ARWU rankings (also known as the Shanghai ranking) adopt similar criteria to evaluate institutions. National unemployment statistics were collected from Statista. Moreover, average income data was collected from CEIC.

Secondly, the main variable being analyzed is first-year students' enrollments for a given discipline per institution for a particular year. Note that a prospective first-year university student may not have multiple enrollments at a time. Students who have enrolled in university for the first time by 1 October for a given year are part of the dataset for that year. Comparable studies use either enrollment or application data to reflect demand.

This paper aims to construct an aggregate demand curve for university places in the Netherlands. In the Netherlands, almost all students are admitted, given that they have successfully completed VO (Voortgezet Onderwijs). VO is a form of secondary education that grants high school students access to universities. Hence, this paper assumes that the supply curve for university places is perfectly elastic. In other words, capacity increases or decreases depending on the number of applicants. Assuming perfectly elastic demand is quite common in the existing literature. For example, Huijsman et al. (1986), Mueller & Rockerbie (2005), and Neill (2009) make the same assumption. Similar papers do this such

that a least-squares approach can be used to estimate the demand curve. In the Netherlands, this assumption is warranted because tuition fees are exogenous. The Dutch government determines the tuition fees and adjusts them for inflation. Hence, all institutions in the sample charge identical fees. In addition, the supply curve is not likely to shift due to changes in alumni donation amounts and shocks to endowment funding.

Because real tuition fees are the same for all institutions in the sample, tuition is not included as an independent variable in this analysis. Moreover, a recent meta-analysis by Havranek et al. (2017) finds that the effect of tuition on enrollment is negligible. Additionally, student grants are issued by the government and not the institutions themselves (Huijsman et al., 1986). Hence, the likelihood of receiving aid is constant regardless of the institution one has registered for.

The enrollment data used in this study exhibit heteroskedasticity. Namely, the variance in the residuals increases as the fitted values increase. Heteroskedasticity is a common problem faced by researchers. There are numerous econometric responses to this problem. Responses include transforming the dependent variable or specifying the dependent variable differently. This thesis adopts a weighted least squares (WLS) model to address the issue. WLS models assign weights to each observation depending on the variance of the corresponding fitted value. Hence, these weights are calculated using the fitted values and residuals of a linear model. Then, these weights are incorporated into estimating a robust weighted model. An overview of the fully specified model is presented in Figure 1.

Figure 1. Explaining Variation in Enrollments

$$Registrations_{i,t} = \beta_0 + \beta_1 Top50_{i,t} + \beta_2 Top100_{i,t} + \beta_3 Top200_{i,t} + \beta_4 Top300_{i,t} + \beta_5 Top400_{i,t} + \beta_6 Unranked_{i,t} + disciplinedummies + locationdummies + \epsilon$$

Registrations are captured at the discipline level for every institution per year. Hence, each of the 2095 observations corresponds to a discipline-institution-year. Not all institutions in the sample received a QS ranking for every year. Hence, the rankings were one-hot encoded. In other words, an institution's ranking is captured by numerous dummy variables. Additionally, a dummy variable is included for every province in the Netherlands. Hence, these location dummies will account for differences in the mean number of registrations owing to location. In addition, dummy variables for each discipline were included to show how demand varies per educational field. Due to the large number of variables included in the model, time-fixed effects were excluded to avoid multicollinearity. Table 1 includes a description of every independent variable in the WLS model.

Table 1. Variable Descriptions

Variable Name	Description
Registrations_top50	This dummy variable takes value one if an institutions QS ranking equal to 50 or lower for a given year. The variable takes value zero otherwise.
Registrations_top100	This dummy variable takes value one if an institutions QS ranking equal to 100 or lower for a given year. The variable takes value zero otherwise.
Registrations_top200	This dummy variable takes value one if an institutions QS ranking equal to 200 or lower for a given year. The variable takes value zero otherwise.
Registrations_top300	This dummy variable takes value one if an institutions QS ranking equal to 300 or lower for a given year. The variable takes value zero otherwise.
Registrations_top400	This dummy variable takes value one if an institutions QS ranking equal to 400 or lower for a given year. The variable takes value zero otherwise.
registrations$unranked	This dummy variables takes value one if an institution did not receive a QS ranking for a given year. The variable takes value zero otherwise.
Location dummies	These variables indicate the province of a certain observation.
Discipline dummies	These variables indicate the discipline of a certain observation.
Unemployment	The unemployment rate in the Netherlands for a given year.
Interest_rate	The short term interest rate in the Netherlands for a given year.

Table 2. Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Unemployment_rate	2,095	4.616	1.132	3.400	6.900
Interest_rate	2,095	0.033	0.005	0.026	0.043
Top50	2,095	0.007	0.081	0	1
Top100	2,095	0.080	0.272	0	1
Top300	2,095	0.164	0.370	0	1
Top400	2,095	0.091	0.287	0	1
Unranked	2,095	0.120	0.325	0	1
Friesland	2,095	0.058	0.233	0	1
Gelderland	2,095	0.116	0.321	0	1
Groningen	2,095	0.112	0.315	0	1
Limburg	2,095	0.067	0.250	0	1
Noord_Brabant	2,095	0.130	0.336	0	1
Noord_Holland	2,095	0.090	0.286	0	1
Primary Education Teaching	2,095	0.005	0.069	0	1
Economics	2,095	0.108	0.310	0	1
Behavior and Society	2,095	0.128	0.334	0	1
Healthcare	2,095	0.124	0.330	0	1
Agriculture and Environment	2,095	0.013	0.115	0	1
Nature	2,095	0.087	0.282	0	1
Law	2,095	0.113	0.316	0	1
Language and Culture	2,095	0.128	0.335	0	1
Engineering	2,095	0.113	0.316	0	1
Multidisciplinary	2,095	0.115	0.319	0	1

3.2 Forecasting HE Enrollments

In terms of forecasting, this thesis mainly focuses on forecasting enrollments through exponential smoothing and double exponential smoothing. Triple exponential smoothing is irrelevant for enrollment forecasting because enrollments are recorded on a yearly basis. Hence, incorporating seasonality into the forecasting equation does not improve its ability to make predictions (Lavilles and Arcilla, 2012).

The exponential smoothing algorithm is expressed in the following way:

(1) $F_{t+1} = \alpha X_t + (1 - \alpha)F_t$

Exponential smoothing is generally applied to forecast time series data that do not have a trend. There are two decisions that have to be made when applying simple exponential smoothing. Firstly, one must choose the value of the forecast F at time $t = 1$. Secondly, one must choose a value for α (the smoothing constant). In this study, F_1 is equal to the first observation. Moreover, mean squared error was calculated for $\alpha = 0.1, \alpha = 0.2, ..., \alpha = 0.9$. Lastly, the alpha with the lowest corresponding MSE was chosen.

Furthermore, the double exponential smoothing algorithm is expressed as follows:

(2) $F_t = \alpha D_t + (1 - \alpha)(F_{t-1} - T_{t-1})$
(3) $T_t = \gamma(F_t - T_{t-i}) + (1 - \gamma)T_{t-i}$
(4) $F_{t+1} = F_t + T_t$
Where:
T_t is the trend in period t and γ is the smoothing constant for trends.

DES is applied to data which has a trend of some sort. DES uses two smoothing constants, α and γ. Two decisions have to be made when forecasting with DES. Firstly, one must determine the initial values of F_t and T_t. Secondly, the optimal α and γ must be determined. F_1 shall be equal to the first observation. Secondly, T_1 is equal to the second observation minus the first. In addition, the optimum combination of α and γ were found by setting each variable to a value between 1 and 0. α and γ were chosen such that MSE was minimized.

4. Results

4.1 Determinants of Higher Education Demand

The model output described in section 3.1 is displayed in Table 3. Model 1 is the most robust WLS regression. It omits the unemployment and interest rate variables to minimize multicollinearity. Hence, model 1 is the main focus of this analysis. Model 2 was included to demonstrate the effects of unemployment and interest rates' effects on enrollments. Both models are constructed using the WLS approach outlined in the previous chapter.

Table 3. WLS Regression Results

	Dependent variable:	
	Registrations	
	Model 1	Model 2
Unemployment_rate		−27.693**
		(11.408)
Interest_rate		−7,815.259***
		(2,301.467)
Top50	472.950	445.739
	(628.520)	(369.804)
Top100	170.281	229.627***
	(128.032)	(73.950)
Top300	56.929	−43.200
	(66.241)	(37.389)
Top400	−221.951***	−297.046***
	(50.515)	(31.170)
Unranked	−28.237	58.902**
	(39.609)	(28.427)
Friesland	−736.997***	−687.820***

	(58.818)	(32.760)
Gelderland	−128.381**	−74.763**
	(51.453)	(29.178)
Groningen	−29.573	−33.244
	(76.506)	(44.123)
Limburg	−13.291	34.378
	(87.516)	(47.162)
Noord_Brabant	14.810	51.239**
	(16.947)	(19.987)
Noord_Holland	−103.861	−57.858
	(85.409)	(37.629)
Primary Education Teaching	−110.959	−127.253
	(147.824)	(82.228)
Economics	1,009.576***	986.380***
	(111.134)	(65.002)
Behavior and Society	881.403***	846.328***
	(95.043)	(55.246)
Healthcare	663.962***	657.665***
	(79.615)	(45.593)
Agriculture and Environment	2,844.819***	2,768.197***
	(681.215)	(397.061)
Nature	716.633***	695.946***
	(113.281)	(66.140)
Law	603.659***	562.726***
	(68.669)	(38.465)
Language and Culture	507.994***	484.904***
	(55.671)	(29.585)

Engineering	729.590***	661.315***
	(105.059)	(58.647)
Multidisciplinary	201.444***	181.945***
	(43.157)	(21.588)
Constant	225.040***	619.285***
	(48.703)	(80.299)
Observations	2,095	2,095
R^2	0.255	0.517
Adjusted R^2	0.247	0.512
Residual Std. Error	2.128 (df = 2073)	1.259 (df = 2071)
F Statistic	33.757*** (df = 21; 2073)	96.541*** (df = 23; 2071)
Note:		*p<0.1; **p<0.05; ***p<0.01

Model 1 shows that twelve independent variables have a statistically significant effect on the dependent variable. All in all, the model explains 24.7 percent of the cross-sectional variance in first-year university enrollments.

Firstly, the Top400 dummy variable is negatively correlated with enrollments in model 1. More specifically, institutions with a QS ranking between 300 and 400 accumulate 222 fewer enrollments per discipline per year. The remaining ranking dummy variables are statistically insignificant and have positive coefficients. These results indicate that enrollments are affected when a Dutch institution's QS ranking drops below 300. Surprisingly, the coefficients of the Top50, Top100, and Top300 dummy variables are statistically insignificant. A possible explanation could be that prospective students value subject-level rankings over institution-level rankings. Subject-level and institution-level rankings can differ enormously for some institutions in the sample. For example, Tilburg University as an institution is currently ranked 356th. However, its Economics program is ranked 41st globally. Additionally, Erasmus University Rotterdam is ranked 43rd in the Economics subject area but 179th globally.

Secondly, the "Friesland" location dummy variable is statistically significant at the one percent level. More specifically, institutions in Friesland experience 737 fewer enrollments per discipline than other institutions in the sample due to location effects. Moreover, the Gelderland dummy variable is also statistically significant at the one percent level. Institutions in Gelderland accumulate 128 fewer enrollments per discipline year compared

Table 4. Breusch-Pagan Test Results

Studentized Breusch-Pagan Test	
BP	0.00016985
df	21
p-value	1

to others. This could be explained by the fact that the two provinces are less densely populated or less urbanized. Christofides et al. (2001) find that there is a positive correlation between the degree of urbanization and enrollments.

Thirdly, almost all discipline dummy variables are statistically significant at the one percent level. These variables are designed to reflect differences in demand across various educational fields. The larger the coefficient in model 1, the more popular that discipline is. Namely, agriculture and environment, economics, and behavior and society are the three most popular disciplines. Based on the coefficient size, the three least popular disciplines are multidisciplinary (programs that combine multiple fields), language and culture, and law.

4.2 Robustness Tests

4.2.1 Heteroskedasticity

Figure 2 demonstrates why a weighted least squares approach was chosen over a linear model. Figure 2 plots the fitted values of a linear model on the x-axis and the corresponding residuals on the y-axis. Heteroskedasticity is present because the variance of the residuals changes according to the value of the fitted values.

A Breusch-Pagan test was used to determine whether or not heteroskedasticity was present in model 1 of Table 3. The results of this heteroskedasticity test are reported in Table 4. Note that the p-value is higher than 0.05. Hence, heteroskedasticity is not present in model 1.

4.2.2 Multicollinearity

This section describes numerous tests which detect the extent of multicollinearity within model 1 of Table 3. More specifically, Table 5 outlines the results of six tests. The Farrar Chi-Square test and Theil's method detect multicollinearity. This is only an issue if the extent of multicollinearity is high. Hence, further analysis is required. Figure 3 shows the

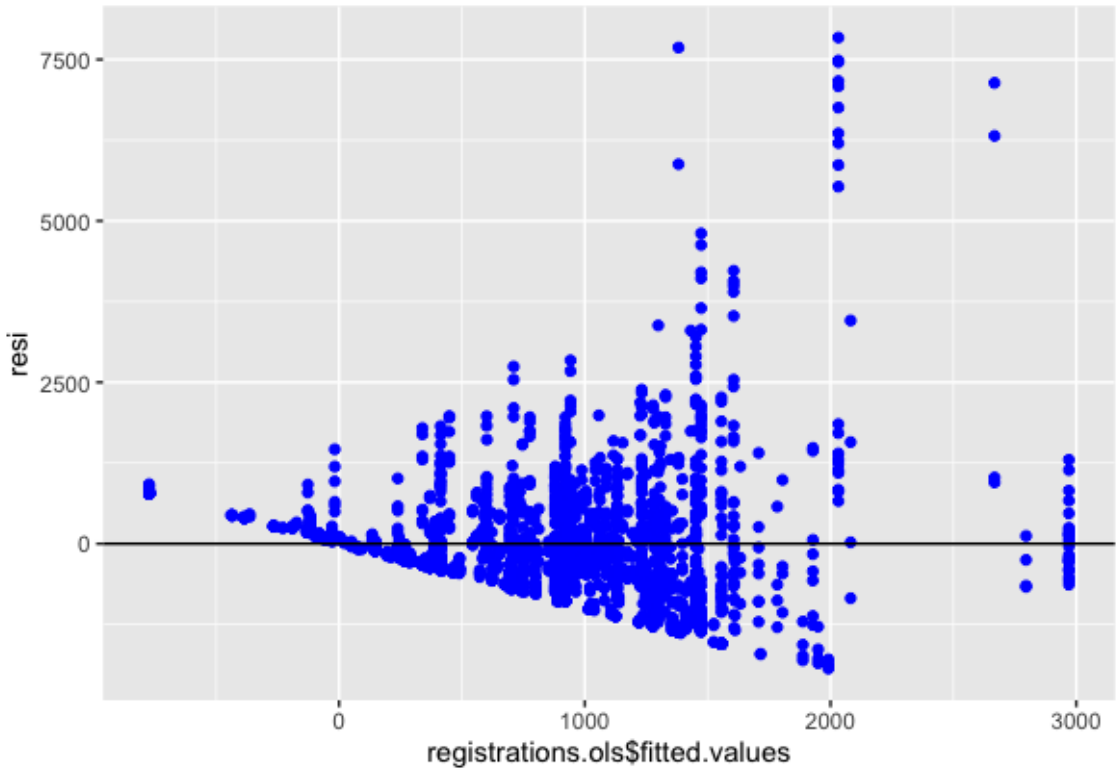

Figure 2. Residuals vs. Fitted

corresponding VIF (variance inflation factor) for every independent variable in the analysis. The VIF for any given variable is equal to the ratio of the overall model variance and a model that solely includes that variable. Figure 3 shows that the VIF values are quite low across the board. Hence, the extent of multicollinearity in model 1 of Table 3 is minimal. Appendix B includes a correlation matrix. Note that all variables that exhibited collinearity were dropped from the study.

Table 5. Multicollinearity Tests

	MC Results	Detection
Determinant |X'X|	0.0263	0
Farrar Chi-Square	7592.5154	1
Red Indicator	0.0977	0
Sum of Lambda Inverse	38.0645	0
Theil's Method	3.0413	1
Condition Number	10.0018	0

1 –>COLLINEARITY is detected by the test
0 –>COLLINEARITY is not detected by the test

All Individual Multicollinearity Diagnostics Result

	VIF	TOL	Wi	Fi	Leamer	CVIF	Klein	IND1	IND2
registrations$top50	1.0232	0.9774	2.4032	2.5309	0.9886	0.9667	0	0.0094	0.0591
registrations$top100	1.1079	0.9026	11.1898	11.7844	0.9501	1.0468	0	0.0087	0.2540
registrations$top300	1.8623	0.5370	89.4178	94.1694	0.7328	1.7595	1	0.0052	1.2076
registrations$top400	2.0151	0.4963	105.2670	110.8608	0.7045	1.9039	1	0.0048	1.3138
registrations$unranked	1.1220	0.8913	12.6475	13.3196	0.9441	1.0601	0	0.0086	0.2835
registrations$Friesland	1.1271	0.8872	13.1824	13.8829	0.9419	1.0649	0	0.0086	0.2941
registrations$Gelderland	1.6547	0.6043	67.8899	71.4975	0.7774	1.5634	1	0.0058	1.0319
registrations$Groningen	1.1367	0.8797	14.1754	14.9287	0.9379	1.0740	0	0.0085	0.3136
registrations$Limburg	1.4013	0.7136	41.6108	43.8220	0.8448	1.3239	1	0.0069	0.7468
registrations$Noord_Brabant	2.0840	0.4799	112.4057	118.3789	0.6927	1.9690	1	0.0046	1.3566
registrations$Noord_Holland	1.5142	0.6604	53.3205	56.1539	0.8127	1.4306	1	0.0064	0.8856
registrations$leraar_basisonderwijs	1.1172	0.8951	12.1563	12.8022	0.9461	1.0556	0	0.0086	0.2737
registrations$economie	2.3693	0.4221	141.9974	149.5430	0.6497	2.2386	1	0.0041	1.5073
registrations$gedrag_en_maatschappij	2.5888	0.3863	164.7562	173.5112	0.6215	2.4459	1	0.0037	1.6006
registrations$gezondheidszorg	2.5331	0.3948	158.9841	167.4324	0.6283	2.3933	1	0.0038	1.5785
registrations$landbouw_en_natuurlijke_omgeving	1.3563	0.7373	36.9530	38.9167	0.8586	1.2815	1	0.0071	0.6852
registrations$natuur	2.1359	0.4682	117.7930	124.0524	0.6842	2.0180	1	0.0045	1.3870
registrations$recht	2.4324	0.4111	148.5356	156.4287	0.6412	2.2981	1	0.0040	1.5358
registrations$taal_en_cultuur	2.5920	0.3858	165.0865	173.8591	0.6211	2.4489	1	0.0037	1.6019
registrations$techniek	2.4543	0.4075	150.8073	158.8210	0.6383	2.3188	1	0.0039	1.5454
registrations$sectoroverstijgend	2.4369	0.4104	149.0059	156.9239	0.6406	2.3024	1	0.0040	1.5378

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

Figure 3. Detecting Multicollinearity

5. Forecasts

This Chapter contains the results of the forecasting approaches outlined in Chapter 3. In addition, some comparisons are drawn between the output of this study and comparable papers in the past.

5.1 Enrollment Forecast Analysis

Firstly, the optimal α and γ values per discipline are outlined in Table 5. 0.9 was the α value that yielded the lowest error for all disciplines. Hence, recent data is more valuable when predicting the future number of enrollments. These α calculations are in line with the findings of Lavilles and Arcilla (2012).

Interestingly, the optimal γ value highly depends on the broad educational field. As shown in Table 5, the gamma values range from 0.3 to 0.9. These results are also in line with the findings of Lavilles and Arcilla (2012). However, Anggrainingsih et al. (2015) use a significantly lower alpha to predict the amount of web traffic an institution will receive in the future.

Table 6 displays point forecasts and MAPE values for every discipline (for both ES algorithms). MAPE was calculated by comparing actual data with forecasted values for each model. Mathematically, MAPE is expressed the following way:

$MAPE = \frac{1}{n} \sum_{t-1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$

Where A_t is the actual value and F_t is the forecasted value.

For all disciplines, DES yields a more accurate forecast. In addition, the findings of Table 6 show that Dutch university administrators could adopt DES to forecast registrations for the following academic year. Additionally, DES significantly outperforms ES because most time series used in this study have a trend. Appendix A shows that enrollments were strictly increasing for most disciplines in the sample.

Table 6. Optimal Alphas and Gammas per Discipline

Discipline	Optimal Alpha	Optimal Gamma
Economics	0.9	0.6
Behavior and Society	0.9	0.9
Healthcare	0.9	0.9
Law	0.9	0.3
Language and Culture	0.9	0.4
Engineering	0.9	0.5
Multidisciplinary	0.9	0.9
Nature	0.9	0.8
Tertiary Education Teaching	0.9	0.3
Agriculture and Environment	0.9	0.9

Table 7. Percentage Error and Forecasts per Discipline

Discipline	ES		DES	
	Point Forecast 22/23	MAPE	Point Forecast 22/23	MAPE
Economics	49838.38	9.7%	53206.61	0.43%
Behavior and Society	56208.85	8.9%	59752.53	0.47%
Healthcare	34743.77	9.5%	36059.79	0.59%
Law	35065.88	8.1%	37476.94	0.39%
Language and Culture	30522.84	5.6%	32134.98	0.36%
Engineering	49492.42	6.7%	52358.11	0.64%
Multidisciplinary	12666.24	15.7%	13765.54	0.44%
Nature	34021.54	10.5%	36323.42	0.53%
Tertiary Education Teaching	1383.71	9.5%	1269.57	0.13%
Agriculture and Environment	13142.68	5.4%	13462.93	0.23%

6. Summary

To conclude, this thesis attempts to answer two main questions. Firstly, this study identifies an appropriate model for enrollment forecasting in the Netherlands. Secondly, several determinants of demand for Dutch HE institutions were also identified. This chapter summarizes all findings and insights throughout this paper.

Firstly, the results outlined in Chapter 4 indicate that location, institution ranking, and discipline determine the number of enrollments an educational field will receive at a university. These findings are in line with the literature covered in Chapter 2. Since the acceptance rate for institutions in the Netherlands is relatively high, prospective students are attracted to institutions with a superior QS ranking. Only a few select HE programs in the Netherlands face excess demand. Some examples of these are given in the following chapter. All in all, the main robust model explained 24.7% of the variation in enrollments. Including more variables such as unemployment or the interest rate would have resulted in biased estimates. However, the most significant contribution is the inclusion of QS ranking. Comparable studies like Mueller Rockerbie (2005) do not incorporate the rankings of the most popular media outlets.

Secondly, the findings of the forecasting efforts are also promising. As outlined in Chapter 2, univariate forecasting methods are typically outperformed by more sophisticated approaches like neural networks or fuzzy time series models. However, using seven years of enrollment data and relatively simple double exponential smoothing algorithms yielded accurate forecasts. More specifically, the least accurate DES forecast yielded a MAPE of 0.64%. These findings are relevant because Studyportals clients may find it cumbersome and costly to implement the more advanced methods mentioned in the literature review. In addition, MAPE is the most widely adopted metric for measuring the accuracy of a model in HE forecasting literature. MAPE is preferred because positive and negative errors cannot cancel each other out.

7. Recommendations for Future Research

Many papers which construct demand curves for higher education incorporate ranking data. However, it is not clear which rankings disproportionately affect demand. In addition, do students value the individual program ranking over the institution's ranking? These questions remain unexplored in HE demand literature. Additionally, the answers to these questions shall likely differ depending on the country being studied.

Secondly, most papers that model demand assume that supply is perfectly elastic. It may be wise to revisit this assumption. Some programs in the Netherlands face excess demand. For example, several programs have a "numerus fixus." In short, this means that the number of seats is limited. These programs are often faced with excess demand. For example, TU Delft's aerospace engineering program has 440 seats while receiving 1154 applications in 2016, according to bachelors.nl. In addition, several biomedical sciences programs would have to double or triple in size to meet demand. More specifically, researchers could investigate which variables cause significant shifts in the supply curve.

8. References

Aladag, C. H., Yolcu, U., Egrioglu, E. (2010). A high order fuzzy time series forecasting model based on adaptive expectation and artificial neural networks. Mathematics and Computers in Simulation, 81(4), 875-882.

Amber, K. P., Aslam, M. W., Hussain, S. K. (2015). Electricity consumption forecasting models for administration buildings of the UK higher education sector. Energy and Buildings, 90, 127-136.

Anggrainingsih, R., Aprianto, G. R., Sihwi, S. W. (2015, October). Time series forecasting using exponential smoothing to predict the number of website visitor of Sebelas Maret University. In 2015 2nd international conference on information technology, computer, and electrical engineering (ICITACEE) (pp. 14-19). IEEE.

Banerjee, S., Igbaria, M. (1993). An empirical study of computer capacity planning in US universities. Information management, 24(4), 171-182.

Bessent, E. W., Bessent, A. M. (1980). Student flow in a university department: Results of a Markov analysis. Interfaces, 10(2), 52-59.

Brinkman, P. T., McIntyre, C. (1997). Methods and Techniques of Enrollment Forecasting. New directions for institutional research, 93, 67-80.

Burkett, H. E. (1986). THE USE OF A COHORT-SURVIVAL MODEL TO FORECAST THE UNIVERSITY OF MISSISSIPPI FRESHMAN ENROLLMENT, 1985 TO 1989 (PREDICT).

Chen, C. K. (2008). An Integrated Enrollment Forecast Model. IR Applications, Volume 15, January 18, 2008. Association for Institutional Research (NJ1).

Christofides, L. N., Cirello, J., Hoy, M. (2001). Family Income and Postsecondary Education in Canada. Canadian Journal of Higher Education, 31(1), 177-2087.

Forrester, J. W. (1961). Industrial Dynamics. MIT Press: Cambridge, MA.

Fredriksson, P. (1997). Economic incentives and the demand for higher education. Scandinavian Journal of Economics, 99(1), 129-142.

Huijsman, R., Kloek, T., Kodde, D. A., Ritzen, J. M. M. (1986). An empirical analysis of college enrollment in the Netherlands. De Economist, 134(2), 181-190.

Hwang, J. R., Chen, S. M., Lee, AC. H. (1998). Handling forecasting problems using fuzzy time series. Fuzzy sets and systems, 100(1-3), 217-228.

Ismail, Z., Efendi, R. (2011). Enrollment forecasting based on modified weight fuzzy time series. Journal of Artificial Intelligence, 4(1), 110-118.

Kaminska , I. (2020, October 8). Become an FT subscriber to read: ICO's final report into Cambridge Analytica invites regulatory questions. Financial Times. Retrieved February 16, 2022, from https://www.ft.com/content/43962679-b1f9-4818-b569-b028a58c8cd2

Krawitz, M., Law, J., Litman, S. (2021, November 11). How higher-education institutions can transform themselves using advanced analytics. McKinsey Company. Retrieved March 15, 2022, from https://www.mckinsey.com/industries/education/our-insights/how-higher-education-institutions-can-transform-themselves-using-advanced-analytics

Kwak, N. K., Brown, R., Schiederjans, M. J. (1986). A Markov analysis of estimating student enrollment transition in a trimester institution. Socio-Economic Planning Sciences, 20(5), 311-318.

LaRivere, J., McAfee, P., Rao , J., Narayanan, V. K., Sun, W. (2017, April 24). Where predictive analytics is having the biggest impact. Harvard Business Review. Retrieved February 16, 2022, from https://hbr.org/2016/05/where-predictive-analytics-is-having-the-biggest-impact

Lavilles, R. Q., Arcilla, M. J. B. (2012). Enrollment forecasting for school management system. International Journal of Modeling and Optimization, 2(5), 563.

Mixon Jr, F. G., Hsing, Y. (1994). The determinants of out-of-state enrollments in higher education: A tobit analysis. Economics of Education Review, 13(4), 329-335.

Mueller, R. E., Rockerbie, D. (2005). Determining demand for university education in Ontario by type of student. Economics of Education Review, 24(4), 469-483.

Murtaugh, P. A., Burns, L. D., Schuster, J. (1999). Predicting the retention of university students. Research in higher education, 40(3), 355-371.

Neill, C. (2009). Tuition fees and the demand for university places. Economics of Education Review, 28(5), 561-570.

Rajni, J., Malaya, D. B. (2015). Predictive analytics in a higher education context. IT Professional, 17(4), 24-33.

Romero, C., Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

Saltzman, R. M., Roeder, T. M. (2012). Simulating student flow through a college of business for policy and structural change analysis. Journal of the Operational Research Society, 63(4), 511-523.

Shmueli, G., Koppius, O. R. (2011). Predictive analytics in information systems research. MIS quarterly, 553-572.

Siegel, E. (2013). Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley Sons.

Sinuany-Stern, Z. (1984). A financial planning model for a multi-campus college. Socio-Economic Planning Sciences, 18(2), 135-142.

Sinuany-Stern, Z. (2021). Forecasting methods in higher education: An overview. Handbook of Operations Research and Management Science in Higher Education, 131-157.

Sinuany-Stern, Z. (1976). Enrollment forecasting models for a multi-campus community college. ORSA/TIMS Meeting.

Song, Q., Chissom, B. S. (1993). Fuzzy time series and its models. Fuzzy sets and systems, 54(3), 269-277.

Tsevi, L. (2018). Survival Strategies of International Undergraduate Students at a Public Research Midwestern University in the United States: A Case Study. Journal of International Students, 8(2), 1034-1058.

Universitaire Bachelors met een numerus fixus. Universitaire bachelors - alle 400+ bach-

eloropleidingen van de Nederlandse universiteiten. (n.d.). Retrieved June 9, 2022, from https://bachelors.nl/numerus-fixus/

Weiler, W. C. (1980). A model for short-term institutional enrollment forecasting. The journal of higher education, 51(3), 314-327.Weiler, W. C. (1980). A model for short-term institutional enrollment forecasting. The journal of higher education, 51(3), 314-327.

Xiao, B., Chankong, V. (2017). A system dynamics model for predicting supply and demand of medical education talents in China. Eurasia Journal of Mathematics, Science and Technology Education, 13(8), 5033-5047.

A. DES Forecasting Vizualizations

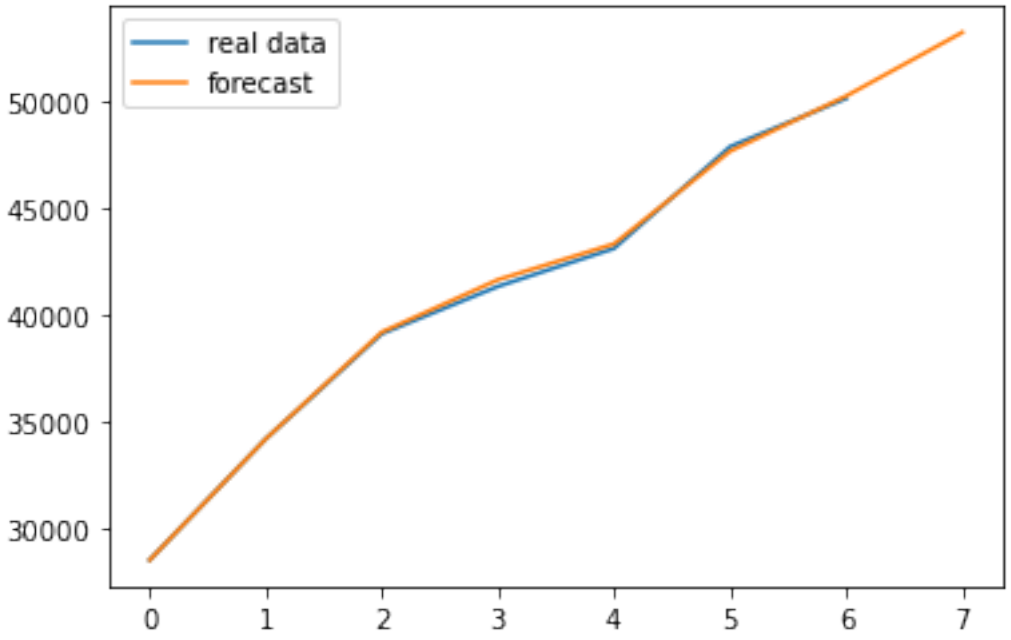

Figure 4. Economics DES Forecast

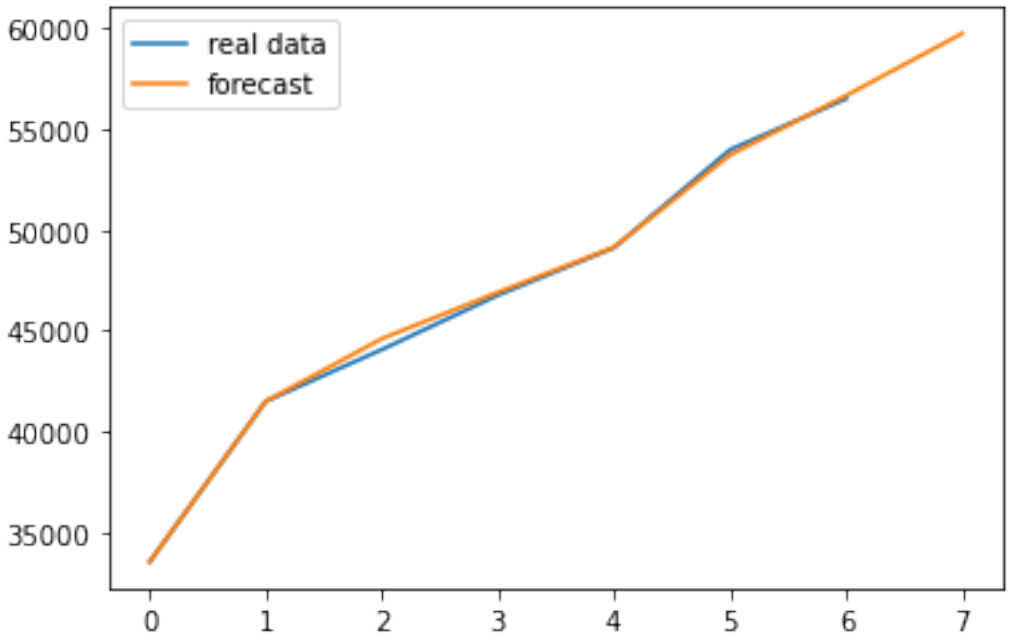

Figure 5. Behavior and Society DES Forecast

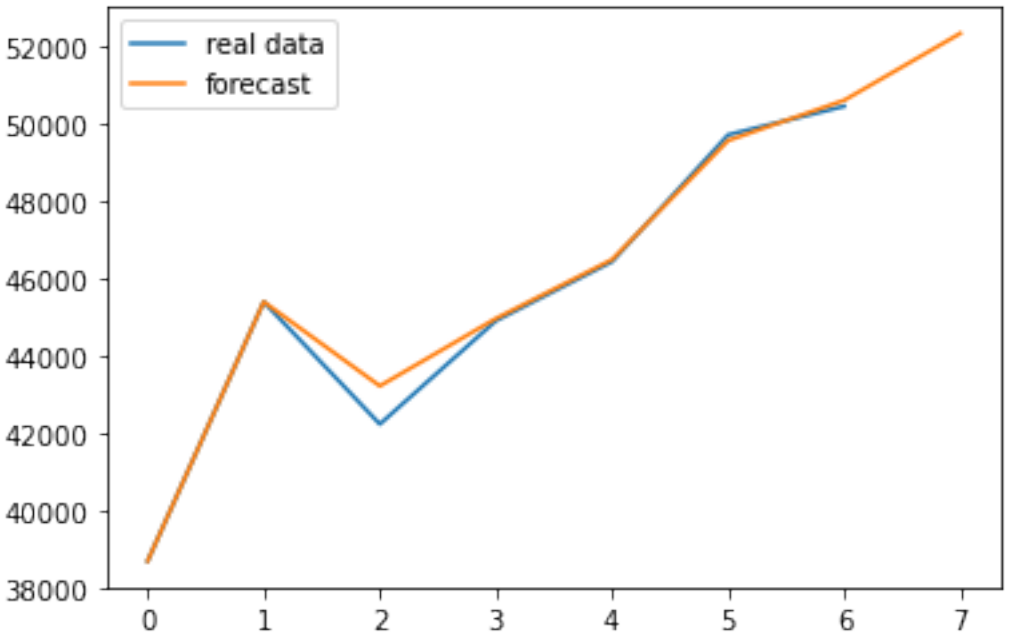

Figure 6. Engineering DES Forecast

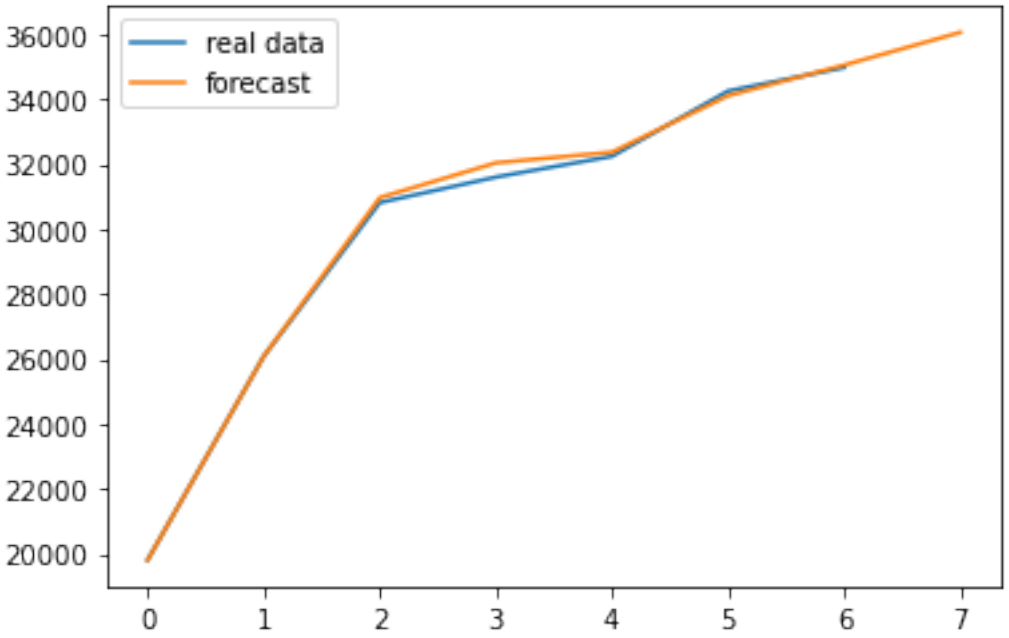

Figure 7. Healthcare DES Forecast

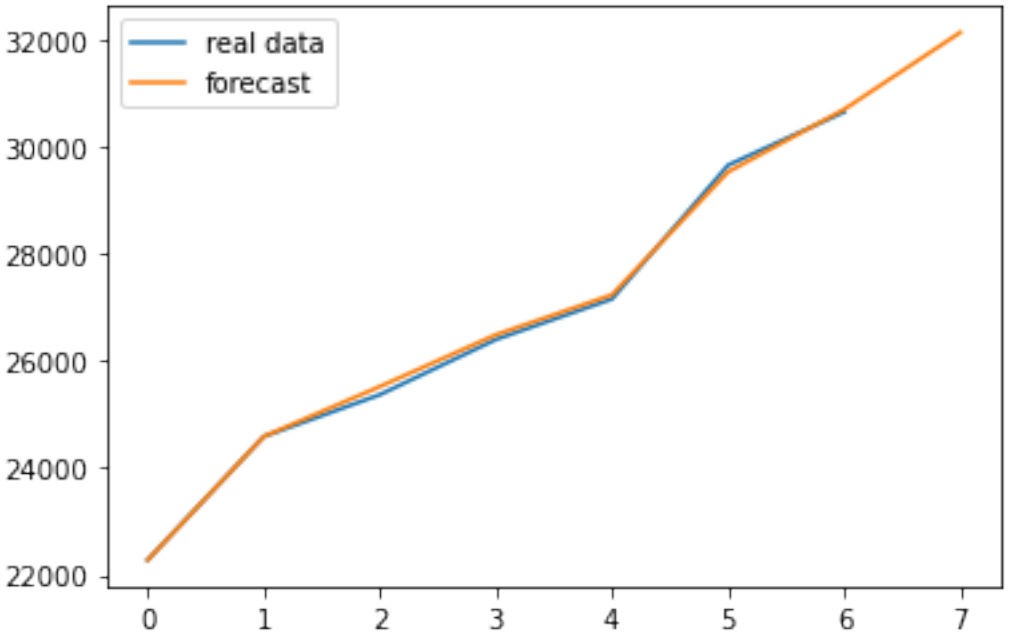

Figure 8. Language and Culture DES Forecast

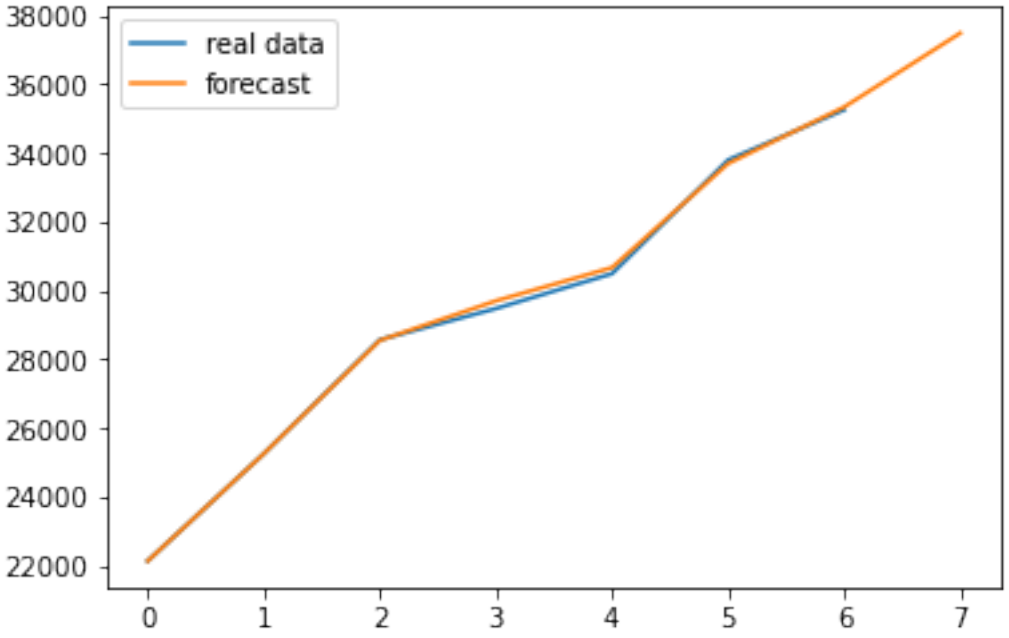

Figure 9. Law DES Forecast

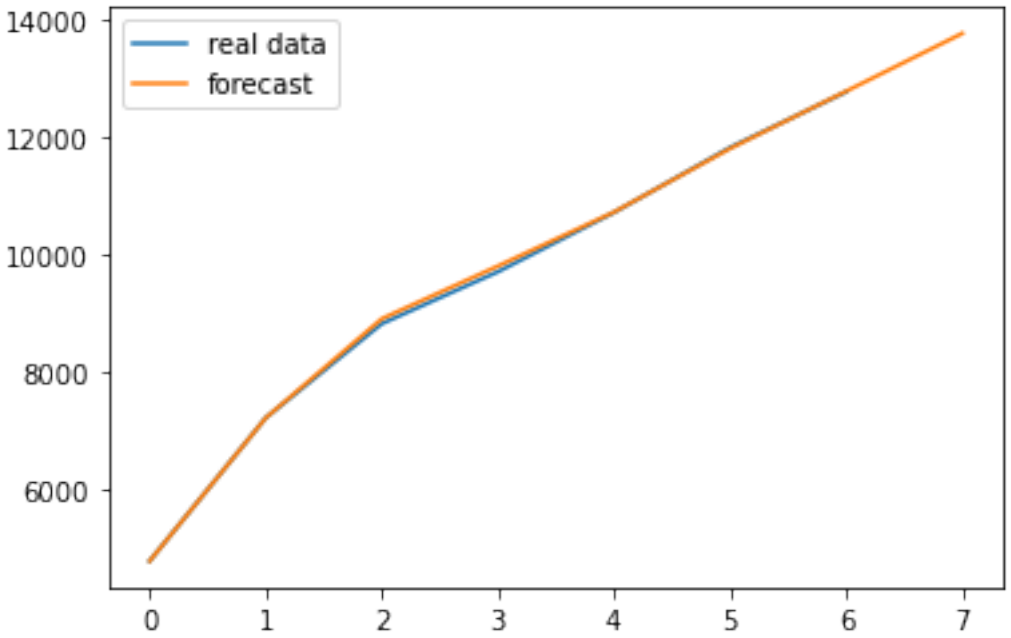

Figure 10. Multidisciplinary DES Forecast

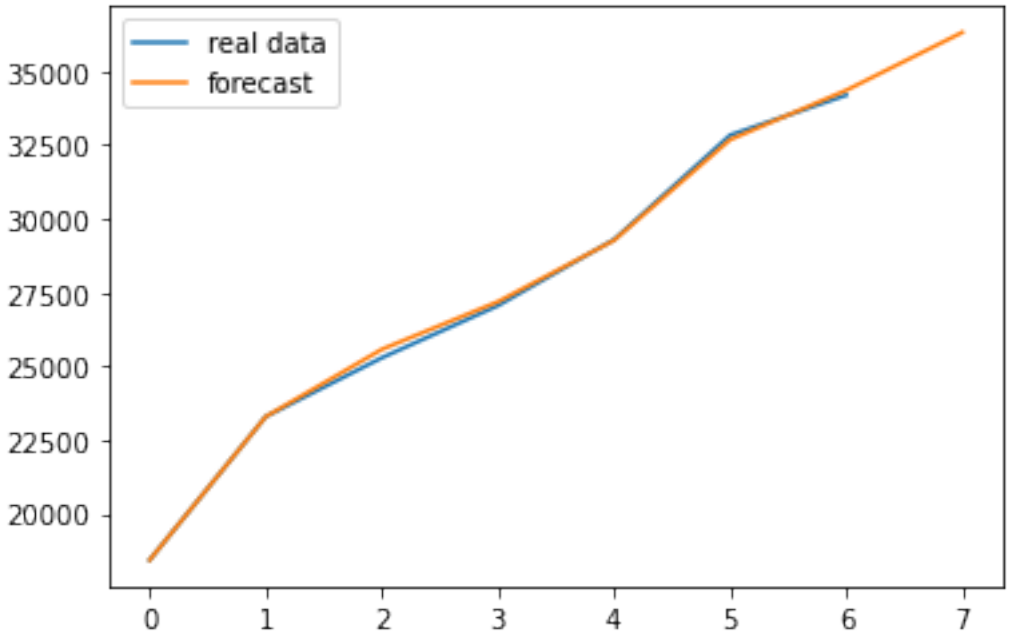

Figure 11. Nature DES Forecast

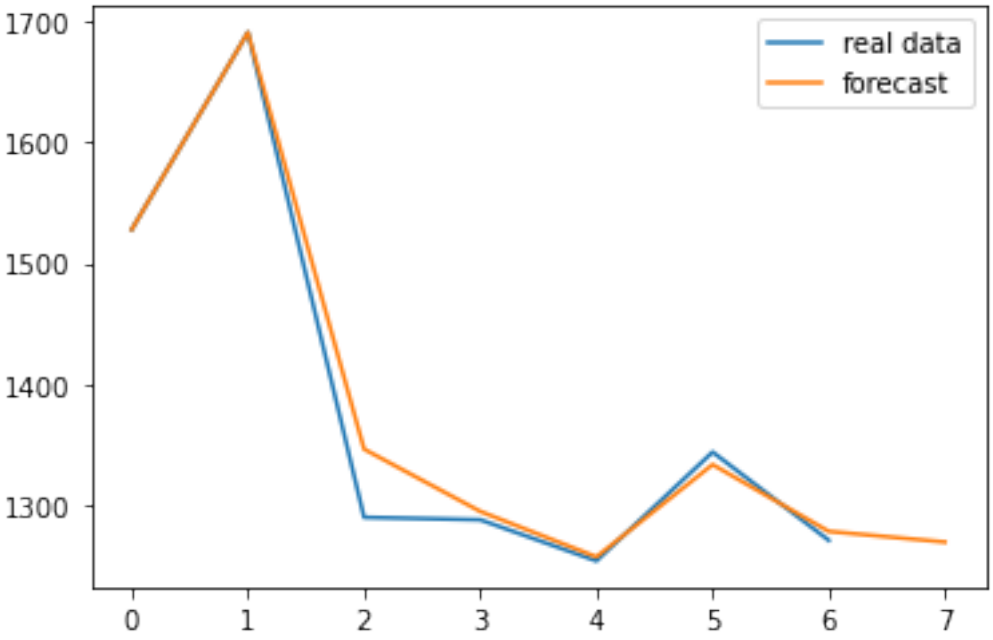

Figure 12. Tertiary Teaching Programs DES Forecast

B. Correlation Matrix

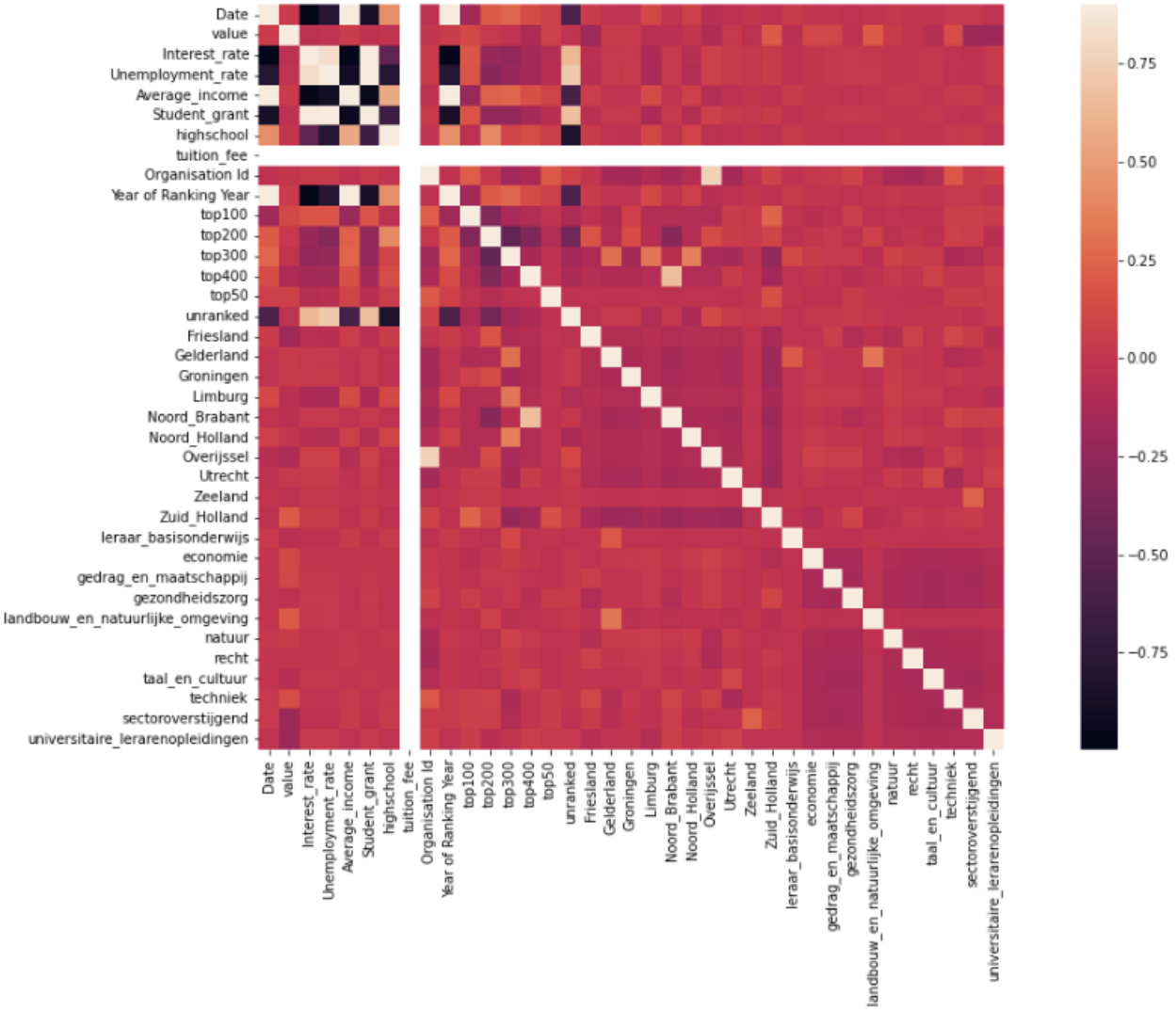

Figure 13. Correlation Matrix