Predicting the long-term influx of new AI master students using VARMA

Niels van 't Leven, 6197019

Supervisors: Prof. dr. G.T. Barkema, Dr. M.A.J.M. Coemans
Second examinor: Dr. A. Gatt

Graduate School of Natural Sciences
Utrecht University
01-07-2022

Abstract

The aim of this paper is to predict the long-term influx of new master students in the Artificial Intelligence program at Utrecht University. A VARMA model was trained on 81.25% of the data, the remaining 18.75% of the data was used as test set. The VARMA model required at least two time series data sets which influence each other. The first was the past influx numbers of students to the Artificial Intelligence program. The second was the number of bachelor diplomas for Artificial Intelligence. The predictions show that the number of new students will rise from 119 in 2022 to 194 in 2031. These results were predicted with a MAPE value of 6.75, which means the model is a good fit. Furthermore the results show a linear trend in the increase of new master students.

Keywords: VARMA, Vector Autoregressive Moving Average, time series prediction, master influx, MAPE

1 Introduction

1.1 Context

In the past two decades studying at a university has become increasingly popular in the Netherlands. 164,638 full-time students were enlisted at Dutch universities in the academic year 2002. This number had more than doubled in nearly two decades to 334,926 full-time students at all Dutch universities combined [1]. Master programs had a relative steeper rise in popularity compared to bachelor programs. 74,781 full-time master students were studying at Dutch universities in 2011, this increased to 122,975 full-time master students at Dutch universities in 2021. Especially the proportion of master students compared to bachelor students has changed a lot. Fig. 1 shows how the number of full-time master students has increased from 2011 and onwards. The total increase was 64.4% from 2011 to 2021, the percentages in the graph are the year-on-year relative increase of full-time master students.

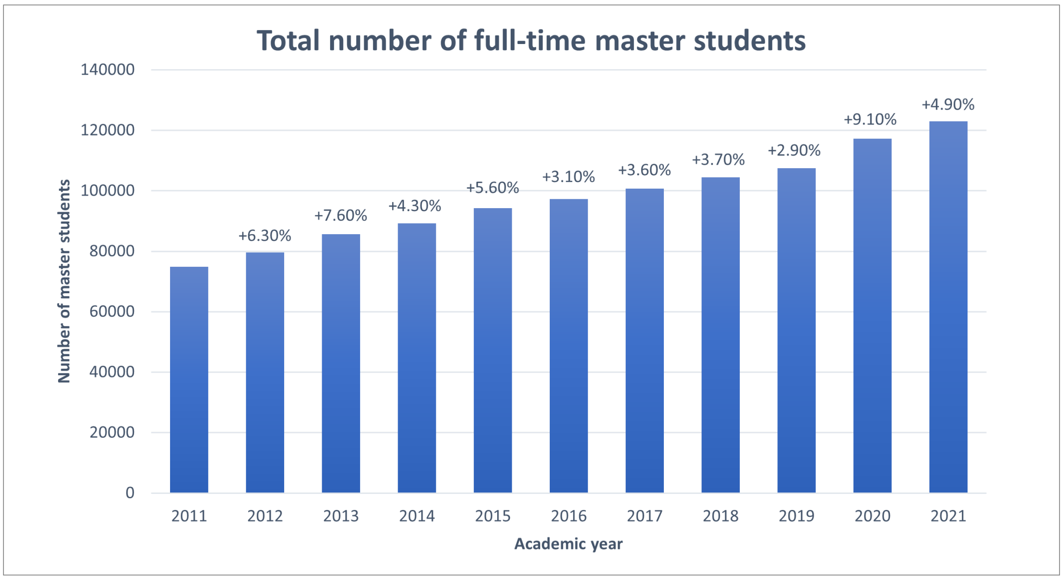

Figure 1: Total number of full-time master students at Dutch universities from 2011 to 2021. On the x-axis are the academic years plotted. On the y-axis are the total number of students plotted that studied at a university master program. The percentages above the bars indicate the amount of growth with respect to the year before. Adapted from [1]

Fig. 2 shows how the number of full-time bachelor students increased from 2011 to 2021. The total increase from 2011 to 2021 was 36%, which is lower compared to the rise in master students. The graph is overall less steep and the total number of full-time bachelor students even decreased a little in 2012. This was due to a policy change that implicated that students who needed more than one year extra to complete their studies would have to pay a fine [2]. This led to many students quickly finishing their studies to avoid a fine. This effect, however, was only temporary because the new legislation was abolished a few months into the academic year. After a relative plateau between 2011 and 2016, the number of full-time bachelor students increased on average by 5.5% anually with a top of 7.3% in

2020.

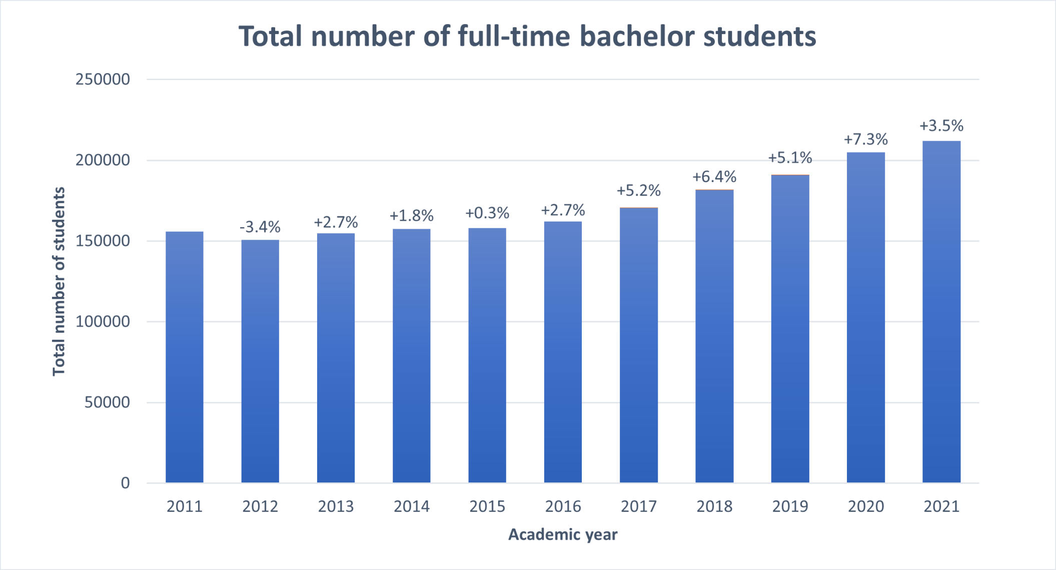

Figure 2: Total number of full-time bachelor students at Dutch universities from 2011 to 2021. On the x-axis are the academic years plotted. On the y-axis are the total number of students plotted that studied at a university bachelor program. The percentages above the bars indicate the amount of growth with respect to the year before. Adapted from [1]

The yearly increasing influx of new master students has a large influence on the increase of the total number of master students. The stacked bar chart in Fig. 3 shows the influx of new master students from 2011 tot 2021. One of the first notable percentile annual increases happened in 2012, which might be explained by the threat that students had to pay a fine if they studied too long [2]. This could have caused many more bachelor students to graduate the year before and therefore starting a master program in 2012 [1]. The next relatively high increase happened in 2015 which is largely explained by the increase in students who were not enlisted at a Dutch university before, the yellow bar in the chart in Fig. 3 represents this group. This group predominantly consists of international students. The highest relative increase of new master students happened in 2020. This was mainly due to a large rise in students who did their bachelor at the same university, represented by the blue bar in the chart shown in Fig. 3. The academic year of 2020 was the second academic year impacted by the COVID-19 virus. Students who could not complete their bachelors in time due to COVID-19 related reasons could that year already start their master. This way more people were able to apply for a master and start it already without finishing their bachelor. Another reason could be that students who were planning to have a gap year, decided to start their masters because their travel or work plans were canceled due to COVID-19 [3]. While there exists past data about the rise in student influx, there are no published works available as of yet about predicted influx numbers. This study will aim to fill this gap in time series forecasting.

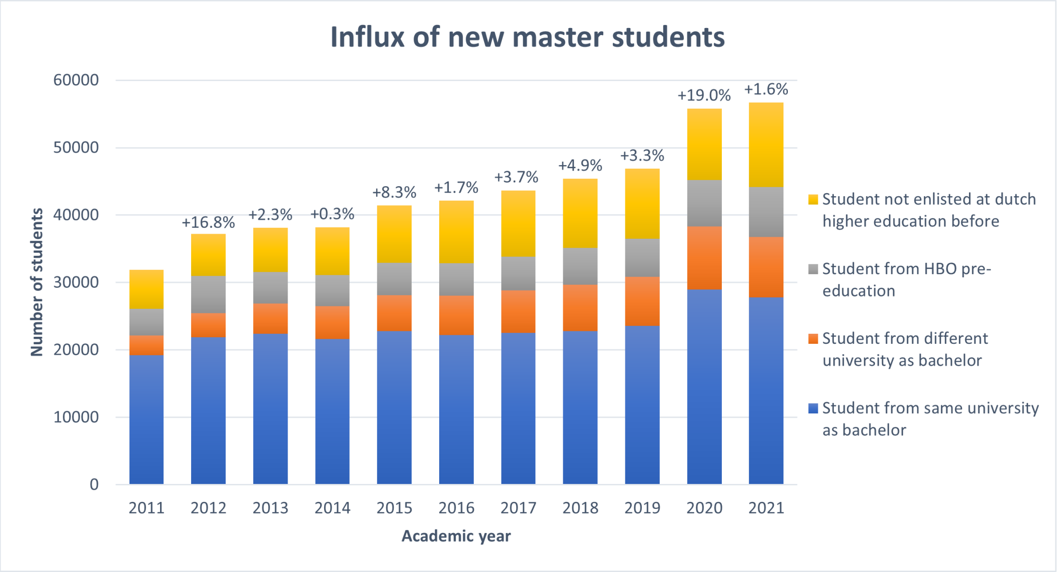

Figure 3: Influx of new master students at Dutch universities from 2011 to 2021. On the x-axis are the academic years displayed and on the y-axis are the number of students. The percentages above the bars indicate the amount of growth with respect to the year before. Adapted from [1]

1.2 Literature overview

The Vector Autoregressive Moving Average (VARMA) model used in this study is a combination of an Autoregressive Integrated Moving Average (ARIMA) model, and a Vector Autoregression (VAR) model. The ARIMA model is a model used to forecast uni-variate time series data. The VAR model is used to forecast multivariate time series data. A VARMA model is the multivariate alternative of the ARIMA model and can take multiple time series into account [4]. The mathematical characteristics of the VARMA model had already been obtained in 1957, but it took till the 80's and 90's for software to be at hand to implement it [5]. Several important studies have been done which use VARMA models, such as a study on the effects of parameter estimation on the forecasts of VARMA models [6].

A fairly recent example of a study using a VARMA model compared the ARIMA-X and VARMA-X model in their ability to predict the price of rice in six provinces on Java [7]. A VARMA-X model is a VARMA model with the addition of an exogenous variable in the model. An exogenous variable is a variable that is measured outside of the model but is imposed on the model.In this study the reseachers used the price of milled dry grain as an exogenous variable. They calculated several accuracy measures after employing both models and the VARMA-X model proved best in predicting the price of rice overall. With the use of a VARMA-X model, the price of rice in one of the provinces was influenced by the rice prices in the other provinces a month earlier, this mechanism is not possible in the ARIMA-X model. The price of rice in two of the four locations however, was a bit more accurately predicted by the ARIMA-X model [7].

The example above is a good indication that VARMA models are still being used. Other studies tried to build upon the already existing VARMA model to improve it or adapt it to certain circumstances. One study for example tried to improve the model by using k-means to cluster the errors of the VARMA model [8]. In addition to this, they used a Bayesian network to learn the association between the data and the trend attached to the VARMA error. The probabilities of the corresponding VARMA errors belonging to each trend could then be used to compensate the values estimated by the VARMA model. Instead of focusing on making the predictions more accurate, a study in 2018 sought to improve parameter estimation for real-time prediction [9]. The parameters of the VARMA model are often estimated in a batch way, which is not fast enough for real-time prediction. They came up with two algorithms, VARMA - Online Gradient Descent (VARMA-OGD) and VARMA- Online Newton Step (VARMA-ONS). The VARMA-OGD algorithm makes use of online gradient descent and only works with a general convex loss function. The VARMA-ONS algorithm is based on the online newton step algorithm and is only effective with an exp-concave loss function.

This paper uses the regular version of the VARMA model mentioned in the various studies before. The objective of this paper is to get an accurate prediction of the influx of new Artificial Intelligence (AI) master students in the ten years after the academic year 2021. This resulted in the following research question:

"To what extent can the influx of new master students for Artificial Intelligence be accurately predicted for ten years to come?"

2 Data

Several files were used in this study which were taken from a larger 1 Cijfer HO dataset. The 1 Cijfer HO dataset is a file from DUO, which is a government agency from the Dutch ministry of education. DUO's task among others, is to provide financing to students, recognize their diplomas and organize exams. DUO processes data from BRON-HO, which stands for Basis Register Onderwijs - Hoger Onderwijs, into the 1 Cijfer HO file [10].

2.1 Data Exploration

The 1 Cijfer HO data consisted of five large CSV files which contained for example information about students who were enlisted in a master or students who graduated. These data sets went back to the 1980s and proved to be useful for predicting the influx of new master students. The first exploration of the data sets showed that the columns did not have column names. The column names were stated in separate word documents which accompanied the data sets. The word documents referenced the column names to the respective column number and explained what the column names meant and what the possible values were for that field.

The first and most important data set of all the 1cijferHO data sets was the Inschrijvingen_aggr_VSNU_2020.csv data set. This was by far the largest file and contained the enlistments of students in programs of all the universities in the Netherlands excluding the theological universities, Open university and the University of Humanistic Studies. The number of enlisted students is measured every year on the first of October, which is the reference date, and goes back from 2020 to 1982. This data set contained 16,505 missing values in the column 'Iscedf2013rubriek', a column that indicates the UNESCO educational sector to which the program of the student belongs. Fortunately, this information is not useful for the analysis, so this variable can be ignored entirely. The second dataset was the Gediplomeerdencohort_VSNU_2020.csv data set. This data set included all the records of students that graduated from all the universities except the aforementioned universities going back from 2020 to 2002. This data set had missing values in the columns listed below in Table 1.

The function of the column 'Isjcedf2013rubriek' is the same in this data set as in the first one (namely the UNESCO educational sector) and is also not useful for the analysis. The 'Maand type ho' column contains the number of months a student is enlisted in higher education till the month of their graduation. With only five missing values this column is largely complete and none of the missing values did a bachelor or master in AI so these records can be discarded anyway. The 'Maand equivalent' column shows how many months the student enlisted in the program at that particular university till their graduation. The 'Masterin' column indicates if the student enlisted for a master program in the year they graduated for their bachelor or in the year after. The 'Masterintwee' column indicates if the student enlisted for a master program in the year they graduated for their bachelor or in the two years after. The column 'Masterintot' indicates if a student has enlisted for a master program since they graduated from their bachelor. The column 'Masterex3' indicates if the student got their master diploma in the year they graduated or in the three years after that. The column 'Masterex5' indicates if the student got their master diploma in the year they graduated or in the five years after that. The column 'Masterextot' indicates if the student got their master diploma in the year they graduated or in any of the years after that. The column 'JaarMasterex' contains the year in which the student graduated from their master programs. If they graduated in multiple master programs it will contain the year for the first one. The column 'Instroomcategorie' indicates, for students with their master diploma, if they came from within the same university from which they received their master diploma or from another university.

Missing Values List	
Variable Name	Nr. of missing values
Iscedf2013rubriek	7687
Maand type ho	5
Maand equivalent	1
Masterin	671134
Masterintwee	671134
Masterintot	671134
Masterex3	671134
Masterex5	671134
Masterextot	671134
JaarMasterex3	849796
Instroomcategorie	600045

Table 1: Columns with missing values in the Gediplomeerdencohort_VSNU_2020.csv data set.

2.2 Data preparation

The required data from the first dataset Inschrijvingen_aggr_VSNU_2020 is the influx of new students every year at the Artificial Intelligence master program at Utrecht University. To get this data, the data set was filtered on the following variables: Opleiding actueel equivalent, Indicatie eerstejaars actuele opl.-instelling, Inschrijvingsvorm and Actuele instelling. Opleiding actueel equivalent, is the current CROHO code of the program the student is enlisted in. CROHO stands for Centraal Register Opleidingen Hoger Onderwijs and is a registry for all the programs of higher education in The Netherlands. CROHO codes are the five digit numbers in that registry which each indicate a program. For example, the national CROHO code for Artificial Intelligence is 66981. This specific code applies to all the Artificial Intelligence master programs in The Netherlands. Indicatie eerstejaars actuele opl.-instelling, indicates whether or not the student is newly enrolled into the current program at that university and thus whether they started their first year at that program or not. This variable can be three different numbers one, three, or six, with a one or three indicating that it is a newly enrolled student at that program and a six indicating that it is a second or higher year student. Inschrijvingsvorm indicates whether the record is a student or not and has three possible values; S, A, and E. An S stands for a record being a student, A for a record being an Auditor, and E for a record which is a partial student and thus can not go to seminars and lectures but can attend exams. Actuele instelling indicates what university the student is attending. The possible values are the BRIN-numbers of Dutch universities, 21PD is the BRIN-number for Utrecht University for example. The data from Inschrijvingen_aggr_VSNU_2020 was selected with the parameters shown in Table 2 below.

Selection Parameters	
Variable Name	Value
Opleiding actueel equivalent	66981
Indicatie eerstejaars actuele opl. -instelling	1 OR 3
Inschrijvingsvorm	S
Actuele instelling	21PD

Table 2: Selection parameters to filter for the influx of new AI master students. The left column contains the variable names and the right column contains the values that were chosen for those variables.

The selected data was then grouped on the Inschrijvingsjaar variable and counted how many records there were in every year. After this the numbers for the influx of new students every year at the Artificial Intelligence master program at the Utrecht University were ready to be analyzed.

The required data from the second dataset Gediplomeerdencohort_VSNU_2020 is the amount of newly graduated students each year for the Artificial Intelligence bachelor at Utrecht University. To get this data, the dataset was filtered on the following variables: Opleiding actueel equivalent, Maand examenresultaat and Actuele instelling. Maand examenresultaat explains what month of the year the student graduated. The selection parameters are shown in Table 3 below.

Selection Parameters	
Variable Name	Value
Opleiding actueel equivalent	56981
Maand examenresultaat	<10
Actuele instelling	21PD

Table 3: Selection parameters to filter for the number of bachelor diplomas of AI students. The left column contains the variable names and the right column contains the values that were chosen for those variables.

2.3 Ethical considerations

The records in both 1 cijfer HO data sets have a number as the primary key but no names of individuals are present for example. The information in the 1 Cijfer HO data set files can tell a lot about the career in higher education someone experienced. This however, would require someone to know a lot more personal information to link to identify an individual in the 1 Cijfer HO data set. According to the Algemene Verordening Gegevensbescherming, personal data is data that is directly about an individual or can be traced back to an individual. The sex of the records is recorded in the data set and this is considered to be direct personal data [11]. Nationality, which is also in the data set, is even considered to be special personal data. The presence of only these variables already would require the individuals in possession of the data files to be careful and not share the data in the files with others. This was also explicitly expressed by the supervisors. For this paper the data will only be used to create a model calculating the influx of new master students in the coming years. Afterwards, the data will be removed from every device.

3 Methods

3.1 Method Selection

The purpose of this paper is to predict the influx of new master students in the coming years. Because the data has a time component, a model would have to make use of a time series forecasting method. Selecting the best time series forecasting method is therefore a vital component in answering the research question of this paper. Because the available data was limited in producing good predictors there is a strict focus on past influx numbers. Therefore auto-regressive time series forecasting methods seemed a good fit. In choosing the best auto-regressive time series forecasting method two properties of the data were very important; whether there is a trend or not and whether there is a seasonal component or not. Because the data will predict changes yearly and not during the seasons all the forecasting methods with a seasonal component can be excluded. The bar chart in Fig. 3 shows the influx of new master students in the Artificial Intelligence program and there seems to be a trend in the data. The indication that there is a trend in the time series is something that should be taken into account when deciding on a model. The graph in Fig. 3 shows that people who did their bachelors at a certain university tend to do their masters at the same university. The same counts for the bachelor and master programs in Artificial Intelligence at Utrecht University. That means it would make sense that there would be a connection between the number of students registered in October for the master program in Artificial Intelligence and the number of students who got their bachelors diploma for Artificial Intelligence in the academic year before that. The graph in Fig. 4 shows how the trend of the number of bachelor diplomas for Artificial Intelligence at Utrecht University, closely follows the trend of the influx of new master students for Artificial Intelligence. The VARMA model used in this paper took into account the effect that the number of AI bachelor diplomas has on the number of new AI master students the following year. The AI bachelor diploma time series data consisted of students who got their diploma somewhere in the ten months before the start of the AI master program. This was done because these students seemed likely to join a masters immediately after their studies. There was however a number for the master AI till 2020 but for the bachelor AI only till 2019. Because the time series data has to match on timestamps entirely the 2020 number for the bachelor AI was generated using a simple autoregression method. The VARMA model used in this paper has two important parts in its name. The VAR in VARMA stands for vector autoregression which means the model can deal with multiple time series at a time and understand how they influence each other. The MA in VARMA stands for Moving Average, it is a linear function of the average residual errors at prior time steps to predict the next time step in a time series.

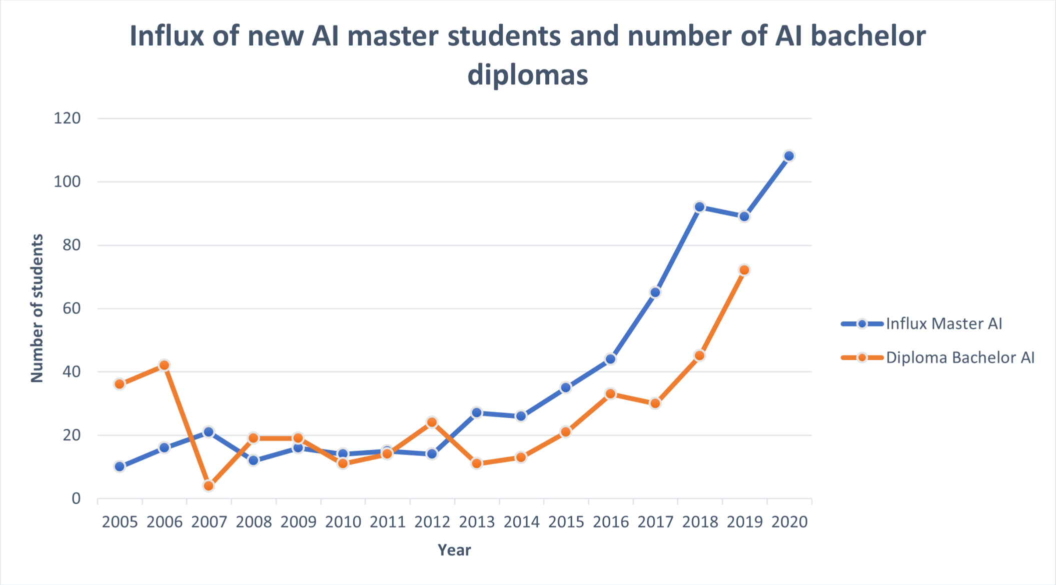

Figure 4: Influx of new AI master students and the number of AI bachelor diplomas. The academic year is plotted on the x-axis and the the number of students is plotted on the y-axis. The blue graph follows the number of new AI master students per academic year. The orange graph follow the number of students who got their bachelor diploma in the 10 months before the start of that academic year.

3.2 Method settings & Accuracy metrics

To train the model, the data was split into a training and test set. The 13 first time-points were used to train the model and the leftover three of the 16 time points were used as the test set. The VARMA order parameter defines the order of the model for the vector autoregression and moving average components. The order ”(1,1)” got the highest accuracy and was used for this model. The first one indicates that the model uses only the first lagged value in both time series for the vector autoregression component. The second one indicates that the moving average component uses only the first lagged error term in both time series. The trend parameter is an important parameter in the VARMA model, it controls the deterministic trend polynomial. The parameter needs a string as input, a ’c’ as input means a constant trend with time, and a ’t’ means a linear trend with time. A ’ct’ would mean both and ’n’ would mean no trend with time. The model was trained with a ’t’ for the trend parameter because this got the highest accuracy on the test set. The model was trained on the portion of 81.25% as training data and the left over 18.75% as the test data. Empirical studies have shown that training data should be around 70-80% of the data [12], [13]. Because the influx data of new master students is time series data, the upper limit for the 70-80% was exceeded by a bit. This had to do with the fact that there were only 16 data points and a bit more data for training a model returned results with higher accuracy, which is preferred in this case.

The most important accuracy measure calculated for the VARMA model is the Mean Absolute Percentage Error (MAPE). MAPE values are used as accuracy measures in various studies which involved VARMA models and other models based on autoregression [14], [15], [7], [16]. The MAPE value is a measure of how large the average percentage difference is between the forecasted value and the actual value. It will allow for calculating the uncertainty boundaries of the predicitions in the future. A smaller MAPE value is considered more accurate because the forecast is closer to the actual value. A MAPE value lower than 10 is considered very good and below 20 means still means the model is a good fit [16]. The MAPE value M is calculated as shown in equation: 1. In formulas 1, 2 and 3 n is the total number of observations, $\mathrm{F_i}$ is the forecasted variable, and $\mathrm{O_i}$ is the observed variable.

$$M = \frac{1}{n}\sum_{i=1}^{n}\frac{|O_i - F_i|}{O_i} \tag{1}$$

The Mean Absolute Error A is calculated with the equation shown in 2, the Mean Absolute Error is the average absolute amount between the observed and predicted number.

$$A = \frac{1}{n}\sum_{i=1}^{n}|O_i - F_i| \tag{2}$$

The Root Mean Squared Error R (RMSE) is calculated with the equation shown in 3. The formula takes the average of the squared residuals after which the root is taken. A Root Mean Squared Error closer to zero is always better.

$$R = \sqrt{\sum_{i=1}^{n}\frac{(O_i - F_i)^2}{n}} \tag{3}$$

The Normalized Root Mean Squared Error N is calculated with the equation shown in 4. The variable R is the result of equation 3the variable Ô is the mean of the observed values.. The formula divides the RMSE by the average of the observed values. A normalized RMSE closer to zero always better.

$$N = \frac{R}{\hat{O}} \tag{4}$$

The model could have included exogenous regressors like [7] used in their VARMA-X model to get an even lower MAPE value. None of the tried exogenous variables however had an improvement on the accuracy of the model so the standard VARMA model was used.

4 Results

The VARMA model with the previously explained parameters predicted the influx of new students at the Utrecht University AI master program as is shown in Table 4. These numbers were rounded to whole numbers as a student can not partially enroll in the program. In the last column the prediction uncertainty is shown based on the MAPE value of the model. The MAPE value is an accuracy measure of the model, and because the MAPE value is a percentage the uncertainty becomes larger with every year if the predicted number increases. The MAPE value of the VARMA model, rounded to two decimals, was 6.75%. The uncertainty borders in Table 4 are measured with the unrounded MAPE value but the actual border values are rounded.

Predicted influx		
Academic year	Number of students	Uncertainty boundaries
2022	119	111-127
2023	127	119-136
2024	135	126-144
2025	143	134-153
2026	152	141-162
2027	160	149-171
2028	168	157-180
2029	177	165-189
2030	185	173-198
2031	194	181-207

Table 4: Predicted influx of new AI master students at Utrecht University. The first column contains the academic year, the middle column contains the predicted influx of new master students for the master program Artificial Intelligence, and the third column contains the uncertainty boundaries of the prediction.

The graph below in Fig. 5 shows the predicted influx numbers for AI. The MAPE value is taken into account as the uncertainty borders at every time-step.

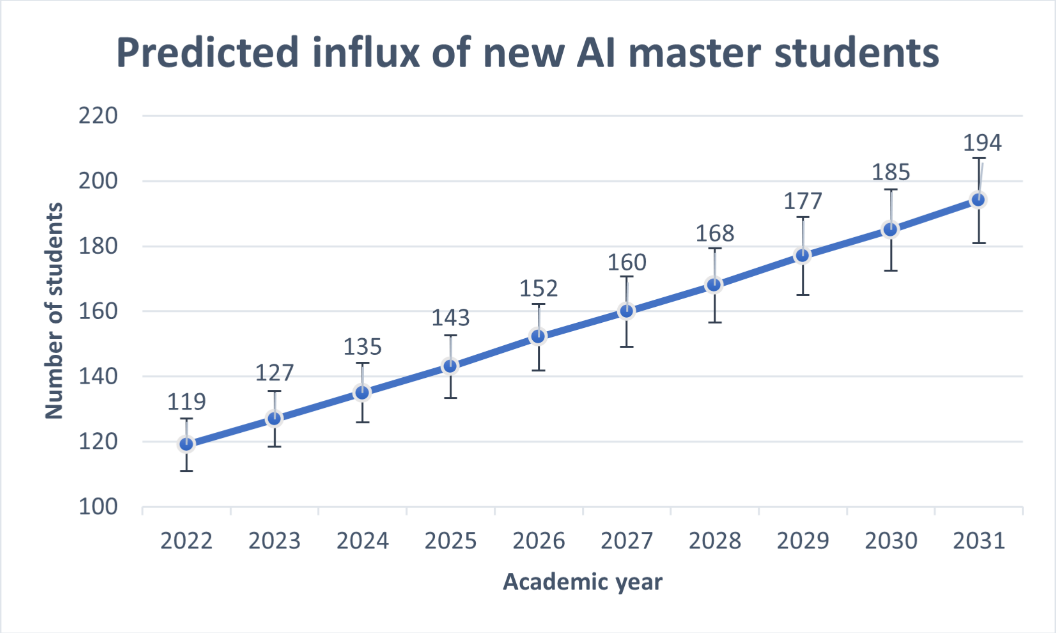

Figure 5: Influx of new master students at the Artificial Intelligence program of Utrecht University. On the X-axis the academic years are plotted and on the y-axis the number of new AI master students are plotted. Uncertainty boundaries at every time step show the uncertainty.

The values of several accuracy metrics are shown in Fig. 5, the values are all rounded to two decimals.

Accuracy metrics of VARMA model	
Accuracy Metric	Metric value
Mean Absolute Error	6.43
Mean Absolute Percentage Error	6.75
Root Mean Squared Error	7.56
Normalized RMSE	0.08

Table 5: Several evaluation metrics of the VARMA model. The first column shows what the name of the accuracy metric and the second column shows the value of the accuracy metric.

5 Discussion

According to the graph shown in Fig. 5, an almost constant yearly increase in new master students is expected for the Artificial Intelligence program at Utrecht University. The predicted influx of new master students of the ten academic years after 2021 are shown in Fig. 4. A good thing to note is that the larger the predicted number, the larger the uncertainty boundaries are for the prediction. The precise predictions for the influx of new AI master students can be found in Table 4. The accuracy metrics in table 5 are all relatively low. The MAPE value is well below 10% which means the model which predicted the future influx numbers is very good [16]. This indicates that the model is a good fit and can predict the test set fairly accurately.

Studies about the long-term influx of new students for any educational program were non-existent until now. This paper could therefore be seen as a start in this domain. The methods used to forecast the long term influx of new master students are, however, employed by other researchers. These studies, however, focus on different domains where time series forecasting is useful such as the prediction of rice prices. This paper can therefore be seen as an addition to the body of research about time series prediction. This study showed that the VARMA model can be an accurate method to deal with time series forecasting problems such as the influx of new master students.

As was explained in the data section the sensitive files were removed after the required data was extracted from them. On its own the model poses no risk to disclosure of personal information or other sensitive data. This model can not be used without the data that was available to train and test it, this would make it only useful and interesting to individuals with access to the data.

The VARMA model used in this paper seemed to perform well given the circumstances. The model, however, has its limitations. The accuracy of the model is first of all dependent on the size of the train and test portions of the data. The size of the train and test set were determined to be optimal for the model's accuracy in combination with the other settings. The model however, could give wildly different forecasts with different train-test set ratios. Second of all the model does not account for policy changes and new legislation such as the reintroduction of the study grant for Dutch students [17]. The study grant is a monthly gift from the Dutch government to Dutch students that they do not have to pay back as long as they finish their studies. As mentioned before the study grant was introduced in 1986 and lasted through 2015 after which the study grant was abolished. The study grant is coming back in the academic year 2022 and might make studying more popular among Dutch citizens [17].

Another limitation of this study is that this particular VARMA model seemed to favor an ever increasing linear trend for the influx of new master students. There are, however, limits to the number of new students a master program can accommodate. This is due to a finite number of lecture halls, lecturers and other facilities being available. If this maximum number will not be reached it might be that a plateau will be reached were the number of new master students every year will not differ much from the year before.

Future studies should look into if and when master programs like Artificial Intelligence tend to reach a plateau. This could be achieved by studying when other similar master programs reached a plateau and how long it took those master programs to get there. This information could be useful to make an even more accurate model to predict the influx of new master students for ten years ahead.

6 Conclusion

This study aimed to predict the long-term influx of new master students for the Artificial Intelligence program at Utrecht University. The predictions showed that the influx of new master students will increase linearly the coming ten years. The VARMA time series model used in this paper is specifically designed for this paper and is new in the domain of student influx predictions. It demonstrates that VARMA is an appropriate method to deal with predicting time series data. Therefore it adds to the already existing body of research that concerns forecasting time series data. The model is, however, limited in its flexibility and can not account for external factors influencing the influx of new master students. Future studies could look into when and if a maximum number or plateau of new students will be reached in the coming years.

References

[1] VSNU, "Aantal ingeschreven studenten."

[2] Helen, "Langstudeerboete is erdoor; studentenorganisaties naar rechter."

[3] HOP, "Meer inschrijvingen, minder uitval: universiteiten flink gegroeid door coronacrisis."

[4] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International journal of forecasting*, vol. 22, no. 3, pp. 443–473, 2006.

[5] M. H. Quenouille, "analysis of multiple time-series," 1968.

[6] J. G. de Gooijer and A. Klein, "On the cumulated multi-step-ahead predictions of vector autoregressive moving average processes," *International Journal of Forecasting*, vol. 7, no. 4, pp. 501–513, 1992.

[7] N. Andayani, I. M. Sumertajaya, B. N. Ruchjana, and M. N. Aidi, "Comparison arima-x and varma-x model to space time data: A case study of rice price in six provinces on java island," *International Journal of Applied Mathematics and Statistics*, vol. 55, no. 3, 2016.

[8] H. Guo, X. Liu, and Z. Sun, "Multivariate time series prediction using a hybridization of varma models and bayesian networks," *Journal of Applied Statistics*, vol. 43, no. 16, pp. 2897–2909, 2016.

[9] H. Yang, Z. Pan, Q. Tao, and J. Qiu, "Online learning for vector autoregressive moving-average time series prediction," *Neurocomputing*, vol. 315, pp. 9–17, 2018.

[10] Onderwijsinspectie, "Databronnen en definities."

[11] Autoriteit-Persoongsgevens, "Wat zijn persoonsgegevens?."

[12] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation," 2018.

[13] B. Vrigazova, "The proportion for splitting data into training and test set for the bootstrap in classification problems," *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, vol. 12, no. 1, pp. 228–242, 2021.

[14] S. Prajapati, A. Swaraj, R. Lalwani, A. Narwal, K. Verma, G. Singh, and A. Kumar, "Comparison of traditional and hybrid time series models for forecasting covid-19 cases," *arXiv preprint arXiv:2105.03266*, 2021.

[15] A. Swaraj, K. Verma, A. Kaur, G. Singh, A. Kumar, and L. M. de Sales, "Implementation of stacking based arima model for prediction of covid-19 cases in india," *Journal of Biomedical Informatics*, vol. 121, p. 103887, 2021.

[16] C. P. Da Veiga, C. R. P. Da Veiga, A. Catapan, U. Tortato, and W. V. Da Silva, "Demand forecasting in food retail: A comparison between the holt-winters and arima models," *WSEAS transactions on business and economics*, vol. 11, no. 1, pp. 608–614, 2014.

[17] R. Meijer, "Basisbeurs keert terug, maar wel lager dan voorheen."