Forecasting the Number of AI Publications: A Glimpse at the Rapid Development of AI

Zhang, Haoqi
New York University
hz3223@nyu.edu

1 Introduction

Artificial intelligence (AI) has witnessed an unprecedented surge in research activity in the past decade. The number and frequency of publications are closely tied with the growth and development of a scientific field. The number of AI-related publications has grown rapidly, signaling the field's increasing importance and rapid development. In this study, we try to glimpse at the developments in the field of artificial intelligence through the lens of forecasting its publications every month. Forecasting this growth will hopefully provide valuable insights on the scale of AI research and the importance of publications.

Arxiv is a research archive platform for scientific publications. Nearly all AI related We utilized the Arxiv API to collect data and form a dataset on the number of AI publications for each month.

We then use the autoregressive integrated moving average (ARIMA) method to model and forecast the number of AI publications for the following 50 months. By analyzing historical publication data, we aim to provide insights into the future development of AI. Our findings highlight the high momentum of AI research and the rapid expansion of the field.

2 Data Description

The dataset consists the number of publications under the category cs.AI published every month that are archived in Arxiv from January 1995 March 2025. Nearly all AI related publications are archived in Arxiv under the category cs.AI.

As there is no such dataset available, we created our own dataset using the Arxiv API. In the metadata for each publication, the published date and the category of the publication is recorded. Using the API, We grouped the published date into months and obtained the number of published publications per month under the category cs.AI. The data can be found on my github page[1].

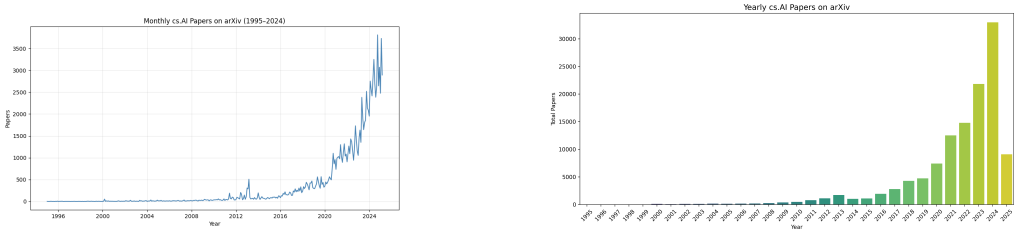
Here, I would like to arXiv for use of its open access interoperability that enabled me to create this dataset.[2]

3 Data Analysis & Plots

We plot out the number of Monthly AI papers from 1995 to 2025 in Figure 3a as we can see in both plots (this study is finished in March 2025, and the number of AI papers has already surpassed many previous years), there is a clear **exponential upward trend.**

The ACF of the Log-Transformed Series can be found here 6, we can see clearly that the series are not stationary, thus differencing is needed.

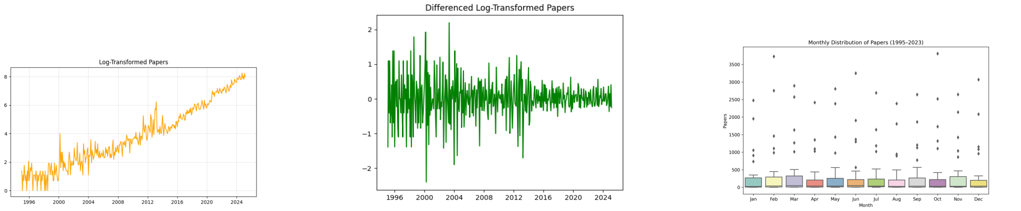
In Figure 2a, we can see that after we conduct **Log-Transformation**, the graph shows linear increasing trend, and in Figure 2b shows the Differenced Log Transformed papers. The Differenced Log-Transformed series appears to be **stationary**, which allows us to fit an ARIMA model.



(a) Monthly AI papers from 1995 to 2025

(b) Bar Plot of AI Papers from 1995 to 2025

Figure 1: Comparison of AI Publications Over Time



(a) Log Transformed Monthly AI papers

(b) Differenced Log Paper

(c) The Monthly Distributions of the papers

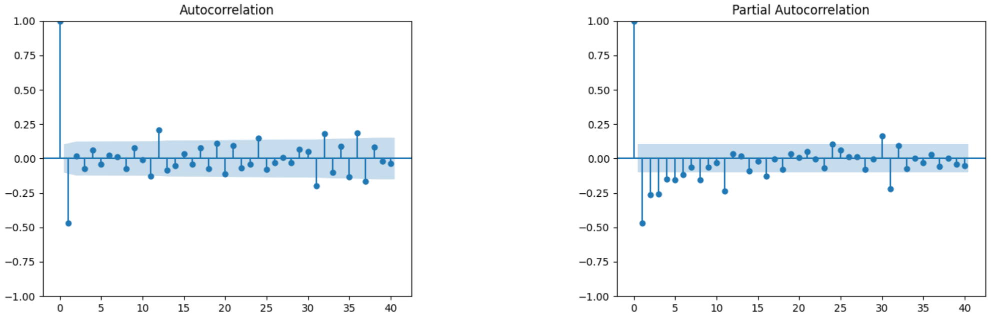
Figure 2: Conducting Log-Transformation and Differencing

From figure 2c we can see the monthly distribution of the papers, the papers published are roughly distributed evenly across the diferent months, suggesting there is no seasonality, and hence, we do not have to worry about it.

4 ARIMA Model Selection and Parameter Estimation

4.1 ARIMA Model Selection

The ACF and PACF as show in Figure, suggests that the Log-Differenced Series is stationary, and is ready to use ARIMA model to conduct forecasting.



(a) Monthly AI papers from 1995 to 2025

(b) Bar Plot of AI Papers from 1995 to 2025

Figure 3: ACF and PACF of the Differenced Log Series

From the plots, we observe that the autocorrelation is significant at lag 1, while the partial autocorrelations are significant at lags 1 and 2. Based on this, one possible choice is the MA(1) model, or equivalently, ARIMA(0,1,1), as the PACF gradually dies down and the ACF appears to cut off at lag 1.

To Validate my assumption, I conducted a grid search. I tested every p and q from 0,1,2,3, and d from 0,1,2. The Best Model I obtained is indeed ARIMA(0,1,1).

Hence, we select the **ARIMA(0,1,1)** model. **Its AIC$_c$ is also the lowest among the choices with 493.2158209673091.** More AICC Values can be found here 4. The AIC$_c$ is calculated by:

$$\text{AICc}(p, q) = -2\log\left[\text{likelihood}(p, q)\right] + 2(p + q + 1) \cdot \frac{n}{n - p - q - 2}$$

4.2 Parameter Estimation

We estimate the parameters from the following results:

Coefficient	Coef	Std Err	z	P>|z|	[0.025]	[0.975]
`x1`	0.0199	0.004	5.419	0.000	0.013	0.027
`ma.L1`	-0.8574	0.024	-35.178	0.000	-0.905	-0.810
`sigma2`	0.2148	0.010	21.091	0.000	0.195	0.235

Table 1: Estimation Results of the Model

Given the estimation result, the fitted model is:

$$\Delta x_t = 0.0199 + (-0.8574)\Delta x_{t-1} + \epsilon_t$$

We obtain the complete form of the fitted model,

$$x_t - x_{t-1} = 0.0199 + (-0.8574)(x_{t-1} - x_{t-2}) + \epsilon_t$$

$$x_t = 1.0199x_{t-1} - 0.8574x_{t-2} + 0.0199 + \epsilon_t$$

5 Results Report & Diagnostic Checking

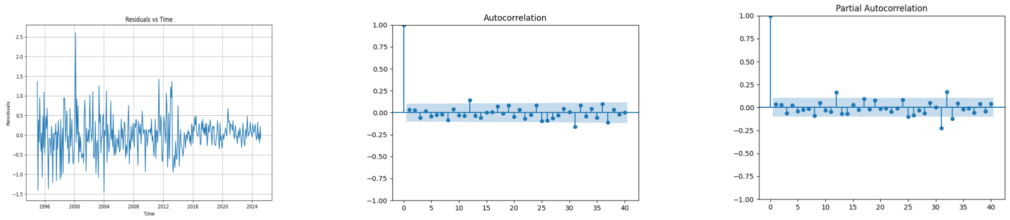
5.1 Results Report

For the selected model, we also show some statistics, plot the residuals and the ACF and PACF of the residuals. The Ljung-Box values can be found here 3, in supplementary Graphs and Tables. See the following

Test	Value
Ljung-Box (L1) (Q)	1.53
Jarque-Bera (JB)	167.28
Prob(Q)	0.22
Prob(JB)	0.00
Heteroskedasticity (H)	0.12
Skew	0.45
Prob(H) (two-sided)	0.00
Kurtosis	6.21

Table 2: Statistical Tests of the Model

The referred figures 4 are the residuals and the ACF and PACF of the residuals.



(a) Log Transformed Monthly AI papers

(b) Differenced Log Paper

(c) The Monthly Distributions of the papers

Figure 4: Residuals, Reidual ACF and PACF

5.2 Diagnostic Checking

- **Ljung-Box (Q):** The p-value of 0.22 suggests no significant autocorrelation in the residuals, meaning the model's errors are independent. This might mean model may be overestimating the uncertainty in the predictions.
- **Jarque-Bera (JB):** The p-value of 0.00 indicates that the residuals are not normally distributed, which could affect the model's reliability.
- **Heteroskedasticity (H):** The p-value of 0.00 shows that the residuals have varying variance over time, violating model assumptions. This can lead to inefficient estimates of the model parameters.
- **Kurtosis:** The kurtosis value of 6.21 indicates heavy tails.

6 Final Forecast

The referred figure 5 is the final forecast of the number of AI publications of every month for the next 50 months, also their corresponding 95% confidence intervals.

The forecasts seem reasonable, as they align with the trend of rapid growing number of publications. The confidence intervals are indeed to some degree excessively high. The width is relatively smaller at first, then for longer forecasts horizons often result in wider intervals because the uncertainty grows as the forecast moves further into the future.

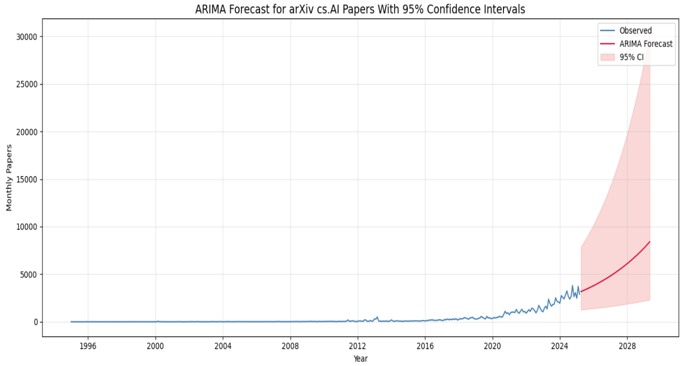


Figure 5: Forecast For the Next 50 Months

Interval Width Analysis

- **Model uncertainty:** The wide confidence intervals could be a result of the model's high variability in the forecast, indicating uncertainty in the parameters or residuals.
- **Heteroskedasticity:** As shown earlier, the residuals exhibit non-constant variance, which can lead to less reliable forecasts with larger confidence intervals.

7 Supplementaty Graphs and Tables

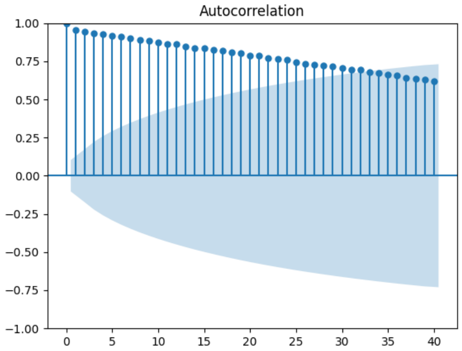


Figure 6: ACF of Log Transformed Series

Lag	Ljung-Box p-value
1	0.494408
2	0.676794
3	0.568155
4	0.702264
5	0.729728
6	0.807265
7	0.867110
8	0.662249
9	0.688532
10	0.742575
11	0.778162
12	0.255970
13	0.293325
14	0.289589
15	0.356003
16	0.424735
17	0.362314
18	0.425908
19	0.320878
20	0.338535

Table 3: Ljung-Box p-values for different lags

ARIMA Order (p, d, q)	AICC Value
(0, 0, 0)	1586.4492
(0, 0, 1)	1260.7836
(0, 0, 2)	1055.4421
(0, 1, 0)	649.7758
(0, 1, 1)	**493.2158**
(0, 1, 2)	494.9168
(1, 0, 0)	652.5943
(1, 0, 1)	502.6820
(1, 0, 2)	504.2278
(1, 1, 0)	561.2356
(1, 1, 1)	494.9023
(1, 1, 2)	495.8349
(2, 0, 0)	568.7014
(2, 0, 1)	504.2130
(2, 0, 2)	505.2897
(2, 1, 0)	538.4566
(2, 1, 1)	496.8730
(2, 1, 2)	497.2633

Table 4: AICC Values for Different ARIMA Orders

References

[1] Haoqi Zhang. Cs.ai papers on arxiv from jan 1995 to mar 2025. `https://github.com/kevin-zhang-1/cs.AI-Papers-on-Arxiv-from-Jan-1995-to-Mar-2025`, 2025.

[2] arXiv. arxiv api documentation, 2025.