Citation Count Prediction Using Different Time Series Analysis Models

1st Priya Porwal
Computer Engineering Department
Amity University, Panvel
Mumbai, India
priya.porwal20@gmail.com

2nd Manoj H. Devare
Computer Engineering Department
Amity University, Panvel
Mumbai, India
mhdevare@amity.mum.edu

Abstract **— The paper helps to predict the future citation value of a fresh dataset of research papers by considering the past values of the citation count of paper using univariate time series analysis models and evaluating their performance through various evaluation metrics. It is important to predict future citation count as it helps to assess researcher's achievements, promotions, fund allocation, etc. This research is in addition to past research where for prediction, different parameters like content of paper, author details, venue impact etc. were considered. The real and original data for the dataset was extracted from the Google Scholar profile of top ranked authors. Three models of time series, Autoregressive Integrated moving average(ARIMA), Simple exponential smoothing (SES), and Holt winter's exponential Smoothing (HWES) are applied to observe the result variations. The models obtained error metric values for the complete dataset. All four-evaluation metrics were calculated. The best results for the predictions for citation count were obtained from the Simple exponential smoothing and Holt winter's exponential Smoothing models, whose values were almost the same for all evaluation metrics because of almost no change in formula. Among all four-error metrics mentioned in the design, MASE gave sensible results, with almost all values being less than 1. The results showed similar graphs for both Simple exponential smoothing and Holt winter's exponential smoothing models for actual and predicted values of citation count as there is negligible difference in formula.**

Keywords **— Citation count prediction, Holt winter's exponential smoothing, Mean absolute scaled error, Simple exponential smoothing, Time series analysis**

I. INTRODUCTION

Nowadays, researchers are working on projects and publishing their work through research papers. To assess the quality of research paper, it is very important topic to consider. These scientific documents are knowledge-bearers. Analyzing its impact for any researcher is very significant to assess the researcher's achievements, promotions, fund allocation, recruitments etc. [15].
The quality of research paper article is influenced by many factors like venue impact, author expertise, paper content [7] and found that the content feature alone is not satisfactory for citation prediction. To evaluate the quality of scientific document, the most important metric used is citation count [16]. Citation is regarded as an indicator of article's strength. Citation count is also used to measure other important metrics like h-index [14], i-10 index.

In previous studies the techniques used to predict citation count were deep neural network techniques using sequence pattern of early year citations [17], back-propagation neural network using paper, journal, author, reference and early citations features [13], deep learning using metadata semantic features [18] and large scale biblio-features [19].
Citation prediction in advance leads to know impact of paper's author and that helps in allocation of research funds, promotions, faculty recruitments.

In this paper, a method is proposed to predict future citations using past citation values of the paper using time series analysis model. It is possible to create a time series out of a collection of data points that are ordered chronologically. There is often a predetermined amount of time between each data point. (Every minute,hour, day, week, and so forth). With the advancement of technology, a growing number of applications across a wide range of areas are producing massive amounts of time series data.

In this method the other features related to paper like author, venue, journal information and paper content are not considered. There is a huge variation in citation pattern of the papers as few papers get noticed easily and referred whereas few remains unnoticed. There is no any single method to predict citation count.

The paper is organized as follows. In section 2, we discuss the related literature work. Section 3 presents the time series analysis models and the evaluation metrices. Section 4 details the dataset, methodology used and results and in the last section 5 represents the conclusion and future work.

II. LITERATURE SURVEY

There are many existing literatures available for citation prediction based on multi feature models. Altmetrics also predict the research impact. The impact was measured in days not in years. Altmetrics measures scholar impact by measuring activity on social media. There is a weak correlation between citations and altmetric which makes doubtful to trust altmetrics. It is based on diverse audiences. Based on the social web, altmetrics is an inappropriate tool to predict future citations [2].

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

This paper explained four key techniques to predict article impact, named as data mining, statistical analysis, machine learning, and network science techniques [1]. Author used machine learning gradient boosting decision tree model for prediction [15]. The author used neural network model to predict high citation papers of ESI and achieved up to mark results [20]. The paper concluded that as the citation window increases the journal impact factor decreased gradually. The author has used paper, journal, author and reference features [16]. In this paper first two years citations were used to trin the model using RNN technique [17]. The author predicted the five-year citation count using BP neural network with five features; publication month, paper length, first two years citation, first-cited age and journal self-citation [13]. The citation prediction was done using deep learning techniques using semantic features extracted from metadata [18].

Author used deep learning techniques to predict long term citations using regression models and earson coefficients to measure the correlation. The data was extracted from network links [19]. The Author predicted highly cited ESI papers. Author established relationship between four features and document scientific impact. Author used neural network [20].

The paper [7] used statistical techniques to preprocess data, select features includes paper content, author expertise, venue impact and apply regression models to predict the impact of top articles. Results showed that author and venue features make the paper more attractive. The best accuracy was achieved after a period of 10 years. To assess citation count of a particular article after a given time period, the system accepted a few specific features [8]. In the learning process, regression models are used, and performance is evaluated using the coefficient of determination (R^2).

Machine learning is used to design effective algorithms and predict future impact. Data mining is used to find out structural patterns in scholarly data. Network analysis is helpful to identify key nodes and their relationships, and to explore structural features [1]. There is an early contribution to citation-based metrics for measuring article impact, including PageRank and the HITS algorithm [3].

The author predicts the future impact of published articles. The features were derived from full text using time series analysis, rhetorical sentence analysis [4]. They predict term frequencies for the first time. They experimented and proved that features like n-grams didn't have much impact on results. Their experiments showed that just text features alone gave more accurate prediction compared to just metadata features.

Datasets are personalized. There are few sources of data collection like SCOPUS, Web of Science, PubMed, CiteULike, Mendeley, DBLP, and MAG. The main problems with the available datasets are incompletion and loss of data, which result in poor evaluation [1]. The one prepared for this experiment is a fresh dataset by web scraping from Google scholar profiles of the top 1000 scientists by their H-index value in computer science and electronics.

In this paper time series analysis model was used to predict citation count. A time series is a sequence in which a metric is recorded at regular intervals of time. In this research specifically univariate time series forecasting was used which take the past values to predict its future values.
The different univariate time series models used are Autoregressive Integrated moving average (ARIMA), simple exponential smoothing (SES), Holt winter's exponential smoothing (HWES).

III. TIME SERIES FEATURES

A. Models

1) *Autoregressive Integrated moving average (ARIMA):* ARIMA is a forecasting algorithm based on the concept of using past values information to foresee future values [9]. ARIMA models can be used to model any "nonseasonal" time series that has patterns and isn't just random noise.

2) *Simple Exponential Smoothing (SES):* It is a time-series forecasting method for univariate data that lacks seasonality or a trend. It follows that a prediction is a weighted sum of prior observations, where the most recent observation is given a higher weight [10]. It requires a single parameter called alpha, also called the smoothing factor or smoothing coefficient.

3) *Holt Winter's Exponential Smoothing (HWES):* It is triple exponential smoothing. It can be used when there is no seasonality. It is made up of level (alpha) and trend (beta) components.

B. Evaluation Metrics

1) *RMSE:* RMSE stands for root mean square error. RMSE is used to estimate the square root of average squared differences between observed and predicted values [11].

$$RMSE = \sqrt{\frac{1}{N}\sum_{d=1}^{N}(c^{d}(t) - r^{d}(t))^{2}} \quad (1)$$

2) *MAE:* MAE stands for mean absolute error. It is the absolute difference between the predicted and actual value.

$$MAE = \frac{1}{N}\sum_{d=1}^{N} |c^{d}(t) - r^{d}(t)| \quad (2)$$

3) *MAPE:* MAPE stands for mean absolute percentage error, which measures the average deviation between true and predicted values for N papers. It is measured in percentages. $c^{d}(t)$ represents the predicted number of citation counts of paper for the time period t. Whereas $\hat{r}^{d}(t)$ is the actual citation count value [6].

$$MAPE = \frac{1}{N}\sum_{d=1}^{N} \left(\frac{\hat{c}^{d}(t) - r^{d}(t)}{\hat{r}^{d}(t)}\right) \quad (3)$$

4) *MASE:* MASE stands for mean absolute scaled error, which calculates the mean absolute scaled error (MASE) between the forecast and the eventual outcomes.

$$MASE = \frac{1}{N}\sum_{d=1}^{N} q_{d} \quad (4)$$

$$q_{d} = \frac{e_{d}}{\frac{1}{n-1}\sum_{i=2}^{n} |y_{i} - y_{i-1}|} \quad (5)$$

Where d = set of forecasting sample periods, y_i = actual observations time series and e_d = forecast error for a given period.

It is observed in his research that among all four metrics mentioned above; MASE gives sensible results [5]. If the MASE value is greater than 1, it means the model is performing worse [12]. The results obtained for MASE in this research is less than 1 in the case of SES and HWES Models.

IV. Methodology

A. Dataset

The data is extracted from Google Scholar profiles of top scientists using web crawling with the Beautiful Soup python library. The year-wise citation count value from the year the paper was published till 2019 is scrapped from Google Scholar and saved in a csv file. Fig. 2 shows the sample of dataset. The dataset consists of approximately 13824 rows, including the papers from 1980 till 2018. Prediction after a long period achieves high accuracy [7]. Fig. 3 represents block diagram of proposed method.

Most research focuses on particular disciplines and sub-disciplines. In this research, we focus on papers with mixed disciplines because this research is irrespective of disciplines. The Corpus is prepared from the papers of scientists having the top h-index for computer science and electronics. At different time points, the value can be represented as x(t1), x(t2),.....x(tn), where x represents the citation value and t1, t2,...tn represents the year the paper was published. The proposed method employed only past values from time series to anticipate future values, which is known as univariate time series forecasting, where only a single parameter is considered for evaluation.

B. Results

A few graphs for actual and predicted citation counts are shown in Fig. 4. A graph is generated for each row of the dataset for all models. The X-axis in this graph denotes the years, and the Y-axis the number of citations. Consequently, the graph's (x,y) values at each point display (year, citation count in that year). For instance, (2008, 12) indicates that the paper had 12 citations in 2008.There are different color lines which represent the outcomes from different models. Yellow represents ARIMA, Blue represents actual citation count, green is HEWS, and red is SES. The green and red lines are overlapping because of the same values. In Fig. 4, the first row shows the results of ARIMA, HWES, and SES with actual citations where the lines of HWES and SES overlap due to the same values but more synchronized with the actual citation count and the second row shows the same graphs with ARIMA, HWES, and SES with actual citations.

This paper is a part of an ongoing research project titled "Research Impact Prediction through Text Analysis". In which we analysed the citation count through various text parameters. In Fig. 1, the x axis represents authors, and the y axis represents the average error rate of ARIMA, SES and HWES for best evaluation metric MASE. From Fig. 1 it is clear that SES and HWES give values less than 1 as compared to ARIMA, which gives high values.

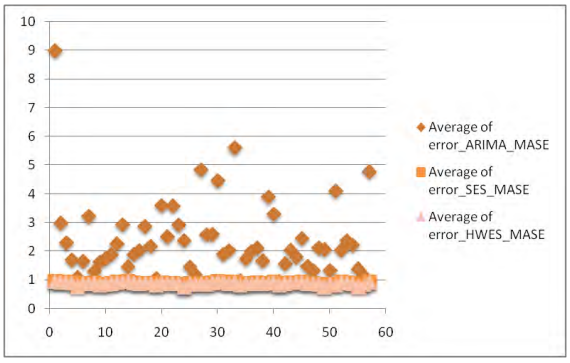


Figure 1. ARIMA, SES and HWES MASE Result Summary (x-axis – authors, y-axis – Average of error)

I21

	A	B	C	D
1	Name	Title	Year	Citation
2	Nathan Eagle	Reality mining: sensi	[2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015,	[12, 35, 90, 130, 177, 231, 208, 269, 321, 297, 264, 304, 252, 200, 9
3	Nathan Eagle	Inferring friendship r	[2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018,	[11, 12, 120, 184, 241, 267, 262, 232, 188, 170, 165, 87]
4	Nathan Eagle	Eigenbehaviors: Ider	[2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018,	[14, 24, 50, 51, 75, 89, 72, 57, 72, 67, 72, 19]
5	Nathan Eagle	Social serendipity: M	[2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015,	[10, 31, 41, 50, 58, 67, 67, 68, 68, 52, 36, 36, 15, 14, 15]
6	Nathan Eagle	Network diversity an	[2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]	[7, 31, 43, 55, 63, 93, 81, 81, 81, 50]
7	Nathan Eagle	Quantifying the impa	[2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]	[2, 42, 66, 88, 82, 86, 88, 65]
8	Nathan Eagle	Smartphones: An em	[2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]	[7, 15, 30, 35, 48, 47, 46, 49, 47, 43, 19]
9	Nathan Eagle	txteagle: Mobile crov	[2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]	[3, 12, 23, 19, 28, 30, 28, 14, 19, 8, 9]
10	Nathan Eagle	Combined short rang	[2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013,	[2, 5, 4, 2, 9, 8, 7, 10, 15, 16, 16, 23, 18, 20, 17, 14, 4]
11	Nathan Eagle	Persistence and peri	[2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018,	[3, 2, 12, 10, 23, 23, 18, 14, 15, 20, 7, 5]
12	Nathan Eagle	Machine perception	[2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015,	[4, 21, 19, 12, 26, 15, 7, 13, 3, 5, 5, 2, 1, 5]
13	Nathan Eagle	Discovering spatiote	[2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]	[4, 8, 16, 14, 20, 17, 13, 13, 11, 7, 3]
14	Nathan Eagle	SMS uprising: Mobile	[2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]	[11, 11, 9, 12, 15, 16, 13, 13, 10, 5]
15	Nathan Eagle	Mobile divides: genc	[2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]	[2, 8, 9, 19, 13, 11, 14, 14, 18, 7]
16	Nathan Eagle	CRAWDAD data set n	[2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017,	[3, 2, 9, 9, 15, 6, 8, 11, 15, 7, 12, 4, 5]
17	Nathan Eagle	The impact of biases	[2013, 2014, 2015, 2016, 2017, 2018, 2019]	[5, 14, 21, 23, 14, 16, 11]
18	Nathan Eagle	Mobility profiler: A f	[2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]	[3, 8, 7, 4, 17, 18, 20, 9, 7, 11]
19	Nathan Eagle	Social serendipity: pi	[2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013,	[1, 5, 6, 3, 2, 5, 4, 3, 14, 3, 5, 9, 8, 12, 5, 6]
20	Nathan Eagle	Divided we call: disp	[2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]	[2, 2, 7, 7, 20, 11, 19, 9, 9]
21	James Philbin	Facenet: A unified en	[2015, 2016, 2017, 2018, 2019]	[54, 350, 728, 1305, 1084]
22	James Philbin	Object retrieval with	[2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017,	[8, 54, 98, 166, 201, 233, 296, 292, 318, 342, 269, 273, 144]
23	James Philbin	Lost in quantization:	[2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018,	[5, 45, 89, 107, 116, 152, 160, 156, 164, 155, 133, 72]
24	James Philbin	Total recall: Automat	[2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017,	[3, 23, 40, 58, 75, 72, 90, 97, 104, 74, 77, 68, 42]
25	James Philbin	Learning fine-grainec	[2015, 2016, 2017, 2018, 2019]	[35, 107, 159, 189, 116]

Figure 2. Screenshot of Dataset

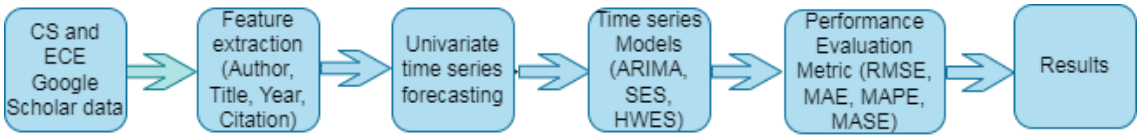
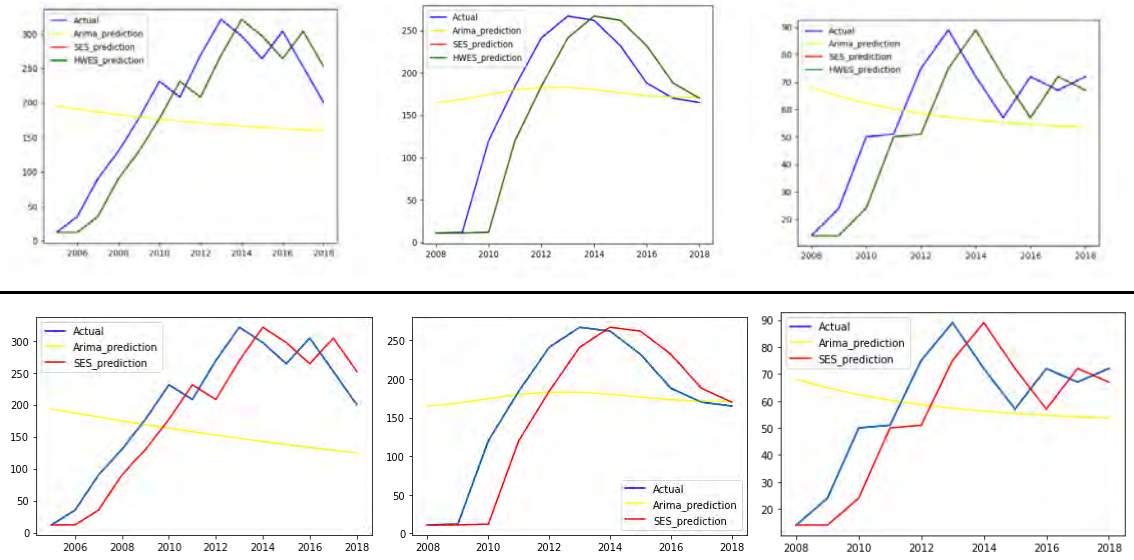


Figure 3. Block diagram of proposed method



(a) (b) (c)

Figure 4. Actual and Predicted citation count (x-axis – year, y-axis – citation count of paper)

V. Conclusions

This research proposes a method for citation prediction using solely time series analysis models ARIMA, SES and HWES. Four assessment measures were employed to assess model performance MAPE, MASE, MAE and RMSE. For all the rows, the output for the simple exponential smoothing and Holt's winter method is the same as there is a small difference in the formula. Among all the three methods, Holt's winter method performed well for all the data as the mean root squared error, mean absolute error, and mean absolute percentage error were less compared to the ARIMA method. Among all evaluation metrics, MASE gives error values of less than 1. Holt's winter method follows the exact trend as the actual data for almost all the cases.

For ARIMA, we need more years of data so that it can fit well with the data. In some graphs ARIMA model negative entries. To apply the ARIMA model, all time series data must first be stationary. As a result, the new entry is made using the difference between the prior and current values. That covers the initial distinction. The Adgtest is used to determine the stationary of time series data (Augmented Dicky-fuller test).

The advantage to this study is the use of newly extracted dataset for analysis, so that it eliminates the issue of incompleteness and data loss. This study's flaw is that it did not compare any state-of-the-art methodologies. As this is completely a time series analysis-based study and previous studies are more on Machine and deep learning with many different features.

In a future study, a multi-model approach might be employed, using elements like author information, Journal impact and article content.

References

[1] Bai, X., Liu, H., Zhang, F., Ning, Z., Kong, X., Lee, I., & Xia, F., "*An overview on evaluating and predicting scholarly article impact*", Information, 8(3), p.73 (2017).

[2] Barnes, Cameron., "*The use of altmetrics as a tool for measuring research impact.*" Australian Academic & Research Libraries, 46.2, pp.121-134 (2015).

[3] Kwok, Roberta., "*Research impact: Altmetrics make their mark.*" Nature, 500.7463, pp.491-493 (2013).

[4] McKeown, K., Daume III, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P & Gravano, L., "*Predicting the impact of scientific concepts using full text features*". Journal of the Association for Information Science and Technology, 67(11), pp. 2684-2696 (2016).

[5] Prestwich, Steven, et al., "*Mean-based error measures for intermittent demand forecasting.*" International Journal of Production Research 52.22, PP. 6782-6791 (2014).

[6] Xiao, S., Yan, J., Li, C., Jin, B., Wang, X., Yang, X. & Zha, H. "*On Modelling and Predicting Individual Paper Citation Count over Time*", In IJCAI, pp. 2676-2682 (2016).

[7] Yan, Lui, et al. "*Citation count prediction: learning to estimate future citations for literature.*" Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, pp.1247-1252 (2011).

[8] Yan, Rui, et al., "*To better stand on the shoulder of giants.*" Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries. ACM, pp.51-60 (2012).

[9] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/ , last seen 3 Dec 2019

[10] https://machinelearningmastery.com/exponential-smoothing-for-time-series-forecasting-in-python/ , last seen 14 April 2021

[11] https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d, last accessed 14 April 2021

[12] https://towardsdatascience.com/how-good-are-your-forecasts-a37cb9cf5c1d (Last accessed 22 April 2021)

[13] Ruan, Xuanmin, et al. "Predicting the citation counts of individual papers via a BP neural network." Journal of Informetrics 14.3 (2020): 101039.

[14] Hirsch, Jorge E. "An index to quantify an individual's scientific research output." Proceedings of the National academy of sciences 102.46 (2005): 16569-16572.

[15] Bai, Xiaomei, Fuli Zhang, and Ivan Lee. "Predicting the citations of scholarly paper." Journal of Informetrics 13.1 (2019): 407-418.

[16] Didegah, Fereshteh, and Mike Thelwall. "Which factors help authors produce the highest impact research? Collaboration, journal and document properties." Journal of informetrics 7.4 (2013b): 861-873.

[17] Abrishami, Ali, and Sadegh Aliakbary. "Predicting citation counts based on deep neural network learning techniques." Journal of Informetrics 13.2 (2019): 485-499.

[18] Ma, Anqi, et al. "A deep-learning based citation count prediction model with paper metadata semantic features." Scientometrics 126.8 (2021): 6803-6823.

[19] Li, Mengiun, et al. "A deep learning methodology for citation count prediction with large-scale biblio-features." 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2019.

[20] Wang, F., Fan, Y., Zeng, A. et al. "Can we predict ESI highly cited publications?". Scientometrics 118, 109–125 (2019).