Bayesianization of ML models in Forecasting Small Data Sequences with Missing Values

Aleksandra Vatian[1[0000-0002-5483-716X]], Ivan Tomilov[1[0000-0003-1886-2867]], Keram Goguev[1[0009-0000-8092-3099]], Oksana Romakina[1[0000-0001-9468-4404]], Anna Arsenyeva[1[0000-0002-2606-1667]] Dmitry Dobrenko[1[0009-0006-1485-1166]], Natalia Gusarova[1[0000-0002-1361-6037]]

[1]1ITMO University, 49 Kronverksky av., St. Petersburg, Russia
alexvatyan@gmail.com

Abstract. Forecasting small sequences with missing values poses significant challenges in machine learning, particularly when dealing with limited training data and incomplete information. In this paper we present a novel unified methodological standpoint, which comparatively evaluates the effectiveness of various machine learning models in predicting sequential data under constrained conditions. We investigate four model families: ARIMA, Gradient Boosting, Fully Connected Neural Networks, and Transformers, examining their performance through both frequentist and Bayesian implementations. Utilizing five diverse datasets representing different sequencing patterns, we systematically analyze model behavior under varying data availability and missingness scenarios. By randomly reducing dataset sizes and introducing missing values, we explore how model complexity and Bayesian probabilistic approaches impact predictive accuracy. The experimental results demonstrate that Bayesianization consistently improves model performance across different datasets, with an average SMAPE reduction of 1-5%. Notably, neural network models, particularly Fully Connected Neural Networks, showed the most significant improvements through Bayesian techniques. This research provides insights into handling small, sparse sequential data and highlights the potential of Bayesian methods in enhancing predictive modeling under data-constrained conditions.

Keywords: Bayesian Inference, Time Series Forecasting, Small Data, Missing Values, Machine Learning Models

1 Introduction and Motivation

Sequential data prediction is one of the typical tasks in ecology, education, medicine, finance, to name a few areas. Analysts must use disparate data, collecting every available element from all possible sources, some of which can seem unusual. These cases are of interest for the research of predicting small data sequences with missing values. Both constraints may cause some problems in supervised learning models: both small datasets and missing data can result in biased or incomplete models, making it difficult to achieve accurate predictions [1].

Handling missing values and small datasets in sequence predicting tasks has a long and well-developed background [2, 3]. The basic challenge for both of them is model overfitting. A natural way to prevent it is to balance the amount of training data with the number of parameters of the machine learning (ML) model being trained. This can

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_15

be reached by using various methods, among which Autoregressive Integrated Moving Average (ARIMA) models [4] remain one of the most popular time series forecasting techniques. Some more sophisticated models have also been proposed, namely, a tool based on genetic programming and Kalman filter as suggested in [5].

The emergence of transformers [6], which have de facto become ubiquitous for a wide range of IT solutions [7], could not but affect the area of sequence prediction [8]. However, in most known solutions, transformers are primarily used to impute the missing values into a sequence [9]. The only exception detected is the study by [10] that aims to predict clinical outcomes from irregularly sampled multivariate clinical time-series without any imputation required. The effective use of transformers for predicting time sequences is reported in [9, 11, 12], however, [13] has reported the opposite results. In terms of the performance of transformers on short-term forecasting on small samples, the only work detected [14] experimentally compared the efficiency of the ARIMA-, LSTM- and transformer-based algorithms. As stated by the authors themself, this result is not exhaustive and does not neglect further studies of the transformers capabilities for short sequence forecasting.

Our analysis shows that the implementation of all the models described above relies on frequential statistics, that is on point estimates of parameters. A new impetus to enriching the available data may be received from a transition from point estimates of parameters to probabilistic estimates using Bayes theorem, the so-called Bayesianization [15, 16]. This enables defining a prior distribution over the missing values so that they can be inferred with the other unobserved parameters when fitting the models [17]. In the perspective of balancing the number of parameters and the amount of training data, this transition could enrich the dataset as such while allowing the model to extract more information from each sample.

A revival of interest in the Bayesianization has recently been due to the advent of transformers [19]. Although earlier works [18] confirmed its effectiveness for predicting short sequences, other authors including [20] tended to claim a danger of overbias in this case.

To sum up, the existing landscape of proposed approaches in the field of forecasting small sequences with missing values does not appear to be holistic and to provide a reasonable choice of IT solution in any specific situation. In an effort to fill this gap, we set the goal to compare, from a unified methodological standpoint, the effectiveness of the most popular approaches to forecasting small sequences with missing values in their frequentist and Bayesian implementations.

We selected the models typical for benchmarking the efficiency of sequences prediction, and, following the approach of [21], used model complexity as the basic parameter for characterizing the model. Specifically, we used models from the families for ARIMA, Gradient boosting with trees (GB), fully connected neural networks (FCNN), and transformers. All model implementations were examined both in their classical and in Bayesian forms.

We used the Missing Completely at Random (MCAR) model, where the probability of having the missing data does not depend on covariates; however, we apply the removal procedure either to single samples with extremely short sequences or to the entire sequences when it represents the whole dataset. We have chosen 1-step forecasting as

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_15

the basic task in the time series forecasting. We have chosen forecasting in a regression form, as categorical sequential data can be forecasted likewise by choosing a link function [22].

2 Method and Materials

2.1 Models and their Bayesianization

We aimed to explore the dependence of the prediction quality of on the ML model type and its complexity. The models selected for comparison are described below.

In implementing a GB regression we followed [23] to handle a multidimensional sequence forecasting task. The model's complexity was regulated by the variable depth of trees. The number of trees was fixed at a standard value of 100 trees. To handle missing values with the ARIMA model, the maximum likelihood approach was used as an example of an approach adopted to MCAR type missing values. A FCNN was implemented and applied in a standard way for sequence forecasting as described in [36]. The number of layers was chosen as the complexity parameter for this model. A transformer model (Figure 1) was built in the simplest form [8] as a composition of multi-head attention layers with standard feed forward network and normalization between them to elicit the effect of a small training dataset in its pure form.

To convert models to Bayesian form we used a scheme (Figure 2) with the parameters estimated by fully factorized gaussians [27]. This scheme was chosen as one of the simplest, the fastest and the most stable among Bayesian inference for deep learning (DL) models to increase the reliability of results. This method was applied for ARIMA in the same way as in [28] to obtain its analogue in the form of Bayesian structural model. Bayesianization was only applied to the last fully connected layers of the transformer model (Figure 3) to increase stability and convergence [29].

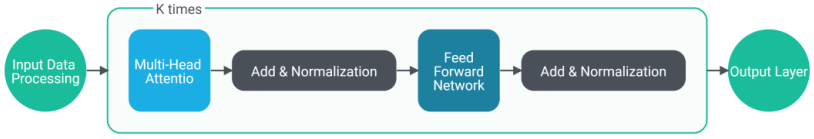

Fig. 1. Transformers scheme in the frequentist variant.

Fig. 2. Bayesianization scheme.

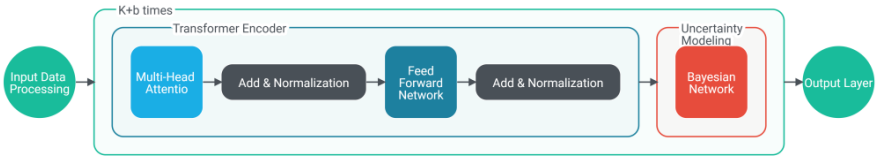

Fig. 3. Transformers scheme: Bayesian variant.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_15

2.2 Data sets used

We used five datasets with different types of sequencing. The datasets were classified into the following groups: (i) a set of extremely short sequences; (ii) a long sequence with regular patterns; (iii) a sequence with irregular patterns.

The main parameters of the used datasets are summarized in Table 1.

Table 1. Main parameters of the datasets used.

Dataset name	Original size	Type of sequencing
Students Performance	480 sequences	(i)
Global Mean Sea	30000 records	(ii)
Long-term Weather	52560 records	(ii)
Microsoft	5000 records	(iii)
Stock Market	25161 records	(iii)

2.3 Experimental scenario and metrics

We pursued two objectives: to investigate how different models configurations would address the prediction task with a variable training set size and the volume of missing values; to compare the quality of Bayesian alternatives with classical frequentist models. We generally followed the methodology of the previous benchmarking [35].

For this purpose, the fragments of various lengths (100%, 75%, 50%, 25%) were randomly selected from the training part of each experimental dataset. Data fragmentation for the training set was performed so as to preserve the semantic coherence of the original data set. Namely, sequences of adjacent data were cut out from the data sets consisting of a single time series; individual time series were selected without disturbing them from the data sets representing a set of short time series.

To simulate the missing values, following MCAR model, we randomly replaced varied (0%, 25%, 50%, 75%) proportions of values with the missing ones.

We repeated the above procedures 30 times with each dataset, thereby forming subdatasets for training the models. namely, 480 sub-datasets for each original dataset. All the analyzed models were trained independently and from scratch on each such subdataset. Their efficiency was measured using the SMAPE metric:

$$\mathrm{SMAPE} = \frac{100}{\mathrm{n}} \sum_{i=1}^{n} \frac{|y_i - \bar{y}_i|}{(|y_i| + |\bar{y}_i|)/2},$$

where y_i is the target value of i-th example, $\bar{y}_i$ is the prediction for i-th example.

3 Results and Discussion

As shown in Section 2.3, the full experimental results are too extensive to be included in the paper. They can be provided upon request. To demonstrate the identified patterns, the most representative cases were selected for each of the dataset groups described in Section 2.2. Namely, the following groups of cases were formed:

1) Full dataset without missing values, referred to below as “Full dense”;

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_15

2) 10% of the original dataset, but without missing values, referred to below as "Small dense";

3) Full dataset, but with 75% missing values, referred to below as "Full missing";

4) 10% of the original dataset with 75% missing values, referred to below as "Small missing".

Figure 4 demonstrates the behavior of model's SMAPE for the datasets of group (i).

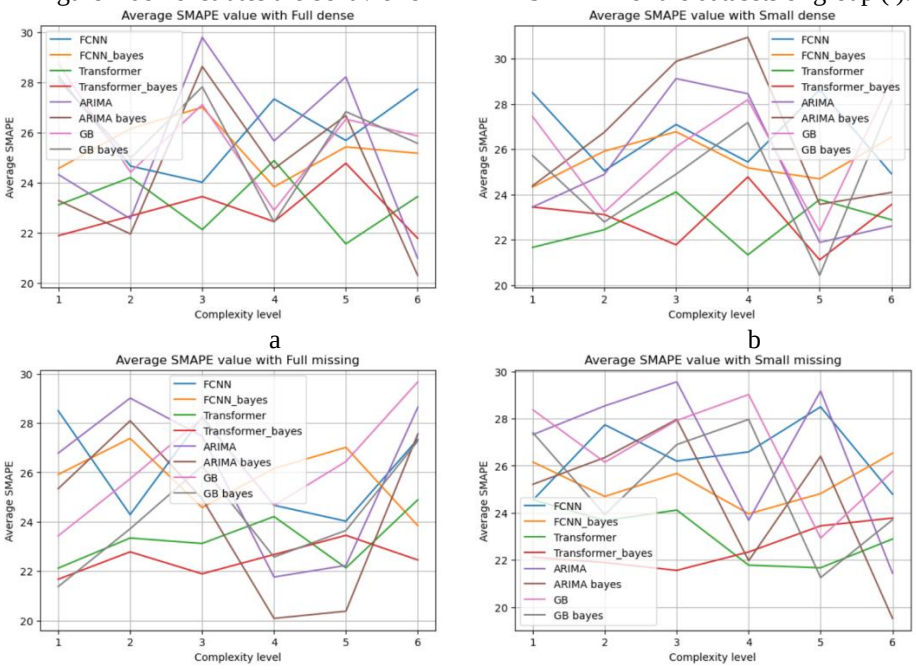

Fig. 4. SMAPE models for datasets of group (i) – a set of extremely short sequences

Figure 4 shows that in the cases of Full dense, Small dense and Full missing (Figure 4a, 4b, 4c), Bayesianization improves the efficiency of all models - FCNN, GB and transformers - regardless of their complexity. On the other hand, for FCNN and GB, a stable decrease by approximately 5% is observed in SMAPE, and for transformer models, the average SMAPE drop makes about 1% and is less stable. The exception is the case of Small missing (Figure 4d), where for all types of models, the decrease in SMAPE is nearly the same (approximately 4%).

A comparison of Figures 4a–4d shows that for ARIMA, Bayesianization yields a SMAPE drop of approximately 3% on average. However, overall they underperform the other models, with an average SMAPE of around 26–30%, while most other models have SMAPE in the range of 21–27%.

Figure 5 demonstrates the behavior of model's SMAPE for datasets of group (ii). On small versions of the dataset (Figure 5b - Small dense, Figure 5d - Small missing), regardless of the data sparsity, SMAPE remains approximately the same for all models. For FCNN and GB, it grows by approximately 0.5-1.5% with Bayesianization, but no

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_15

stable effect is observed on the quality of the transformers. At the same time, GB increases quality with increasing complexity (SMAPE drops by 2%), and FCNN lose about the same amount. In the case of complete data (Figure 5a - Full dense, Figure 5c - Full missing), the quality varies somewhat more, but the same ratio of values is preserved as on small data.

In the case of Small missing data (Figure 5d), a trend towards a decrease in the quality of FCNN with increasing complexity and a small increase in the quality of transformer models is again evident, but this time only for the Bayesian version. ARIMA and its Bayesian version still do not exhibit a stable dependence of quality on complexity.

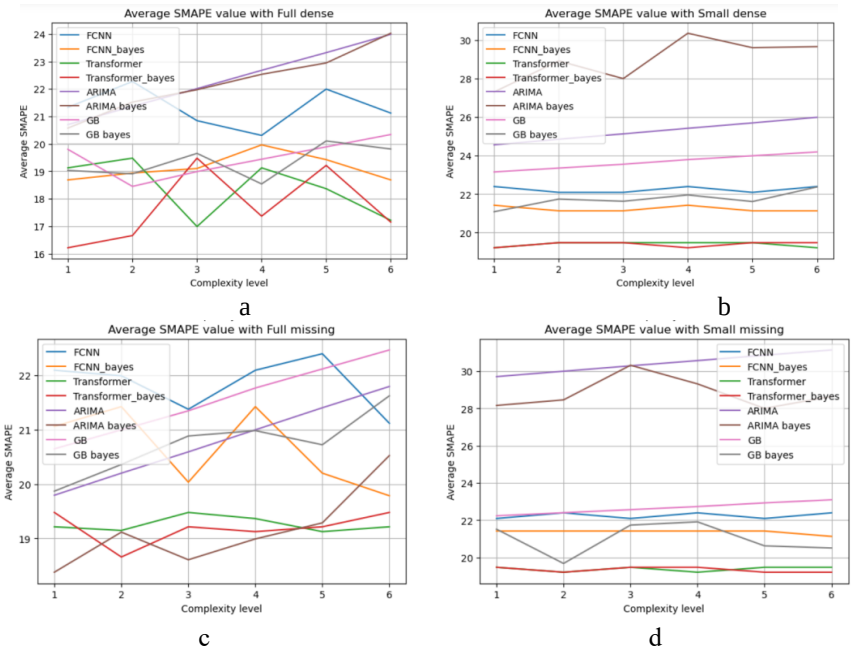

Fig. 5. SMAPE models for datasets of group (ii) – large sequence with a large proportion of missing values

We also examined the behavior of SMAPE models for datasets of group (iii). In the cases of missing values, regardless of the size, the same dependencies are traced as in Fig. 5, but the influence of Bayesianization on both FCNN and transformers becomes stronger. Additionally, it should be noted that Bayesianization has a weak influence on the quality of both ARIMA and boosting in this case - no more than 0.5%. In the case of Small missing data, the transformer approximately maintains the error level, varying it within 0.7%. The quality of a FCNN with Bayesianization increases by a significant 6% in both Full missing and Small dense cases. At the same time, for both regular and Bayesianized FCNN, on Full missing, an increase in complexity leads to a drop in error within 3.5%, and on Small dense, on the contrary, the error increases by 2.1%. GB and ARIMA along with their Bayesian versions lose quality by 2-3% with increasing complexity.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_15

4 Conclusion and Future Works

In this paper, we compared forecasting methods for small sequences with missing values using frequentist and Bayesian approaches. We focused on 1-step forecasting with models like ARIMA, Gradient boosting, neural networks, and transformers. We used three dataset types: extremely short sequences, a large sequence with many missing values, and a small sequence with many missing values, creating 480 sub-datasets each. Models were trained independently on each sub-dataset, using SMAPE to measure prediction quality. Results showed Bayesianization reduced SMAPE by 2% for short sequences, 3-5% for long sequences, and 1-3% for small sequences. Bayesianization is promising for improving forecasting quality in challenging conditions.

Acknowledgments. This work was supported by the Ministry of Science and Higher Education of the Russian Federation, FSER-2025-0013

References

1. Salman N.A., Hasson S.T. A Prediction Approach for Small Healthcare Dataset. 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech), Split/Bol, Croatia, 2023, pp. 1-5, doi: 10.23919/SpliTech58164.2023.10193552.
2. Hyndman, R.J., Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on 02.02.2025.
3. Little RJ, Rubin DB. Statistical Analysis with Missing Data. Wiley; 2019.
4. Box, G.E.P. and Jenkins, G.M. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
5. Kone L.A., Leonteva A.O., Diallo M.T., et al. (2023). Short Time Series Forecasting Method Based on Genetic Programming and Kalman Filter. In: Collet, P., Gardashova, L., El Zant, S., Abdulkarimova, U. (eds) Complex Computational Ecosystems. CCE 2023. Lecture Notes in Computer Science, vol 13927. Springer, Cham. https://doi.org/10.1007/978-3-031-44355-8_6
6. Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need. In Advances in Neural Information Processing Systems 30, 2017, pages 5998–6008. Curran Associates, Inc.
7. Amatriain X., Sankar A., Bing J., et al. Transformer models: an introduction and catalog. arXiv:2302.07730v4 [cs.CL] 31 Mar 2024
8. Wen Q., et al. Transformers in Time Series: A Survey. arXiv:2202.07125v5 [cs.LG] 11 May 2023
9. Nejad A. S. et al. SERT: A transformer-based model for multivariate temporal sensor data with missing values for environmental monitoring. Computers & Geosciences, Volume 188, June 2024, 105601
10. Tipirneni S., Reddy C.K. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. ACM Trans. Knowl. Discov. Data, 16 (6) (2022), pp. 1-17
11. Chen S.-A., Li C.-L., Yoder N.C., et al. TSMixer: An All-MLP Architecture for Time Series Forecasting. Transactions on Machine Learning Research (09/2023). arXiv:2303.06053v5 [cs.LG] 11 Sep 2023.
12. Peng S., Xiong Y., Zhu Y., et al. Mamba or Transformer for Time Series Forecasting? Mixture of Universals (MoU) Is All You Need. arXiv:2408.15997v1 [cs.LG] 28 Aug 2024.
13. Kafritsas N. Will Transformers Revolutionize Time-Series Forecasting? Towards Data Science. Jul 31, 2024. https://towardsdatascience.com/will-transformers-revolutionize-time-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_15

series-forecasting-1ac0eb61ecf3/Perez-Lebel A., et al. Benchmarking missing-values approaches for predictive models on health databases. Gigascience. 2022 Apr 15;11:giac013. doi: 10.1093/gigascience/giac013
14. Ammann G. Using LSTMs And Transformers To Forecast Short-term Residential Energy Consumption. Tilburg University June 2022. https://arno.uvt.nl/show.cgi?fid=160767
15. Vatian A.S., Gusarova N.F., Dobrenko D.A., Pankova K.S., Tomilov I.V. Using topological data analysis for building Bayesian neural networks. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 6, pp. 1187–1197. doi: 10.17586/2226-1494-2023-23-6-1187-1197
16. Shi H., Wang Y., Han L., et al. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. arXiv:2412.05723v1 [stat.ML] 7 Dec 2024
17. Tan M.T., Tian G.-L., Ng K.W. Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation. Chapman & Hall, 2019. 346 p. ISBN 9780367385309
18. Hryniewicz O., Kaczmarek K. (2015). Forecasting Short Time Series with the Bayesian Autoregression and the Soft Computing Prior Information.. Advances in Intelligent Systems and Computing, vol 315. Springer, Cham. https://doi.org/10.1007/978-3-319-10765-3_10
19. Muller S., et al. Transformers Can Do Bayesian Inference. ICLR 2022. https://arxiv.org/pdf/2112.10510
20. Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2019). Bayesian Versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review. *Structural Equation Modeling: A Multidisciplinary Journal*, 2019, *27*(1), 131–161. https://doi.org/10.1080/10705511.2019.1577140
21. Nakkiran P., Kaplun G., Bansal Y., et al. Deep Double Descent: Where Bigger Models and More Data Hurt. 4 Dec 2019; arXiv:1912.02292 [cs.LG]
22. Fokianos K., Kedem B. Regression Theory for Categorical Time Series. Statistical Science 2003, Vol. 18, No. 3, 357–376
23. Zhang Z., Jung C, GBDT-MO: Gradient Boosted Decision Trees for Multiple Outputs. arXiv:1909.04373v2 [cs.CV] 28 Dec 2019
24. Khosravi P,, Vergari A., Choi Y. Handling Missing Data in Decision Trees: A Probabilistic Approach. arXiv:2006.16341v1 [cs.LG] 29 Jun 2020
25. Velicer W.F., Colby S.M. A Comparison of Missing-Data Procedures for Arima Time-Series Analysis. August 2005Educational and Psychological Measurement 65(4):596-615. DOI:10.1177/0013164404272502
26. Borovykh A., Oosterlee C.W., Bohte S.M. Generalization in fully-connected neural networks for time series forecasting. arXiv:1902.05312v2 [stat.ML] 27 Jul 2019
27. Kochurov M., Garipov T., Podoprikhin D., et al. Bayesian Incremental Learning for Deep Neural Networks. arXiv:1802.07329v3 [stat.ML] 27 Mar 2018.
28. Zhang F., Li Y., Li X. Comparison of ARIMA and Bayesian Structural Time Series Models for Predicting the Trend of Syphilis Epidemic in Jiangsu Province. Infection and Drug Resistance. December 2024. 17:5745-5754. DOI:10.2147/IDR.S462998
29. Okolie C.J., Adeleke A.K., Smit J.L., et al. Performance analysis of Bayesian optimised gradient-boosted decision trees for digital elevation model (DEM) error correction: interim results. June 2024ISPRS Annals of the Photogrammetry Remote Sensing and Spatial Information Sciences X-2-2024:179-183. DOI:10.5194/isprs-annals-X-2-2024-179-2024

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_15