European Journal of Research Development and Sustainability (EJRDS)
Available Online at: https://www.scholarzest.com
Vol. 1 No. 2, October 2020,
ISSN: 2660-5570

IMPROVED ANALYSIS FOR TIME SERIES WITH MISSING DATA

Mukta Agarwal
Assistant Professor, Sabarmati University,
Ahmedabad , Gujarat, India,
mukta09agarwal@gmail.com

Article history:

Received 26th September 2020
Accepted: 13th October 2020
Published: 31st October 2020

Abstract:

Particular range examination (SSA) is an incredible method for time arrangement investigation. In light of the property that the first run through arrangement can be imitated from its essential parts, this commitment builds up an improved SSA (ISSA) for preparing the fragmented time arrangement and the adjusted SSA (SSAM) of Schoellhamer (2001) is its unique case. The methodology is assessed with the manufactured and genuine inadequate time arrangement information of suspended-residue fixation from San Francisco Bay. The outcome from the manufactured time arrangement with missing information shows that the overall mistakes of the important segments recreated by ISSA are a lot more modest than those remade by SSAM. Additionally, when the level of the missing information throughout the entire time arrangement arrives at 60 %, the upgrades of relative blunders are up to 19.64, 41.34, 23.27 and 50.30 % for the initial four head parts, individually. Both the mean supreme mistake and mean root mean squared blunder of the recreated time arrangement by ISSA are likewise more modest than those by SSAM. The particular enhancements are 34.45 and 33.91 % when the missing information represents 60 %. The outcomes from genuine fragmented time arrangement additionally show that the standard deviation (SD) inferred by ISSA is 12.27 mg L−1 , more modest than the 13.48 mg L−1 determined by SSAM.

Keywords: Time series prediction, missing data, ensemble prediction.

1.INTRODUCTION

Particular range investigation (SSA) presented by Broomhead and King (1986) for examining dynamical frameworks is an incredible toolbox for separating short, boisterous and disordered signs (Vautard et al., 1992). SSA first exchanges a period arrangement into a direction network, and completes the foremost part examination to select the predominant segments of the direction lattice. In view of these predominant parts, the time arrangement is remade. In this way the reproduced time arrangement improves the sign to-commotion proportion and uncovers the attributes of the first run through arrangement. SSA has been broadly utilized in geosciences to examine an assortment of time arrangement, for example, the stream and ocean surface temperature (Robertson and Mechoso, 1998; Kondrashov and Ghil, 2006), the seismic tomography (Oropeza and Sacchi, 2011) and the month to month gravity field (Zotova and Shum, 2010). Schoellhamer (2001) built up an altered SSA for time arrangement with missing information (SSAM), which was effectively applied to break down the time arrangement of suspended-dregs fixation (SSC) in San Francisco Bay (Schoellhamer, 2002). This SSAM approach doesn't have to fill missing information. All things being equal, it registers every key segment (PC) with noticed information and a scale factor identified with the quantity of missing information. Shen et al. (2014) built up another primary segment investigation approach for separating regular mode mistakes from the time arrangement with missing information of a territorial station organization. The other sort of SSA approach measure the time arrangement with missing information by filling the information holes recursively or iteratively, for example, the "Caterpillar" SSA technique (Golyandina and Osipov, 2007), the attribution strategy (Rodrigues and Carvalho, 2013) or the iterative technique (Kondrashov and Ghil, 2006). This paper is propelled by Schoellhamer (2001) and Shen et al. (2014) and builds up an improved SSA (ISSA) approach. In our ISSA, the slacked relationship grid is processed similarly as by Schoellhamer (2001) – the PCs are straightforwardly registered with both the eigenvalues and eigenvectors of the slacked connection lattice. Nonetheless, the PCs in Schoellhamer (2001) were determined with the eigenvec peaks and a scale factor to make up for the missing worth. Additionally, we don't have to fill in the missing information recursively and iteratively as in Golyandina and Osipov (2007). The remainder of this paper is coordinated as follows: the improvement of SSA for time arrangement with missing information continues in Sect. 2, manufactured and genuine mathematical models are introduced in Sects. 3 and 4 separately, and afterward ends are given in Sect. 5.

2. IMPROVED SINGULAR SPECTRUM ANALYSIS FOR TIME SERIES WITH MISSING DATA

For a fixed time arrangement xi (1 ≤ I ≤ N), we can build a L × (N − L + 1) direction grid with a window size L. Its Toeplitz slacked connection lattice C is planned by where both xi and xi+j must be noticed as opposed to missed, and Nj is the quantity of the results of xi and xi+j inside the example file I ≤ N − j . At that point we figure the eigenvalues and eigenvectors from the slacked relationship network C. The PCs are likewise determined with noticed information: where Li is the quantity of noticed information inside the example file from I to I + L − 1. The recreation system of time arrangement from PCs is equivalent to SSA. The scale factor L/Li is utilized to make up for the missing worth. To determine the statement of registering PCs for the time arrangement with missing information, Eq. (3) is reformulated aswhere 1 ≤ I ≤ N − L + 1, and Si and Si are the file sets of testing information and missing information individually inside the whole number span [i, I + L − 1], for example Si ∩Si = 0 and Si ∪Si = [i, I + L − 1]. On the off chance that PCs are accessible, we can duplicate the missing qualities. Subsequently, the missing qualities in Eq. (7) can be subbed with PCs asSince λk, the difference of the kth RC, is arranged in plunging request, the initial a few RCs contain a large portion of the signs of the time arrangement, while the excess RCs contain essentially the commotions of time arrangement. Subsequently the first run through arrangement is remade with the initial a few RCs. The SSAM approach created by Schoellhamer (2001) figures the components c(j) of the slacked relationship framework by Since Gi is a symmetric and rank-lacking network with the quantity of rank insufficiency rising to the quantity of missing information inside the stretch [xi , xi+L−1], the PCs ak,i (k = 1, 2, . . . , L) are fathomed with Eq. (10) in view of the accompanying model (Shen et al. 2014): min : ξ T I 3 −1 ξi, (13) where 3 is askew network of eigenvalues λk, which is the covariance lattice of PCs. The arrangement of Eq. (10) is as per the following: Supposing that v1,k = v2,k = ··· = vL,k = 1/√ L at the missing information focuses, the arrangement of Eq. (15) will be decreased to Eq. (6). Accordingly, the SSAM approach is an uncommon instance of our ISSA approach. The initial a few PCs contain most fluctuation; the component xi+j−1 can be roughly recreated with the initial a few PCs in Eq. (8). The primary contrast of our ISSA come closer from the SSAM approach of Schoellhamer (2001) is in ascertaining the PCs. We produce the PCs from noticed information with Eq. (14) as indicated by the force range (eigenvalues) and eigenvectors of the PCs, while Schoellhamer (2001) computes the PCs from noticed information with Eq. (6) simply as indicated by the eigenvectors and utilizations the scale factor L/Li to remunerate the missing worth. We have called attention to that this scale factor can be gotten from Eq. (15), which is the disentangled variant of our ISSA approach, by assuming the missing information focuses with the equivalent eigenvector components. Along these lines the presentation of our ISSA approach is superior to SSAM of Schoellhamer (2001). The main drawback of our technique is that it will cost more computational exertion.

3. PERFORMANCE OF ISSA WITH SYNTHETIC TIME SERIES

A similar manufactured time arrangement as in Schoellhamer (2001) are utilized to dissect the exhibition of ISSA contrasted with SSAM. The manufactured SSC time arrangement is communicated as where R(t) is a period arrangement of Gaussian background noise zero mean and unit standard deviation; cs(t) is the intermittent sign communicated as cs(t) = 100 − 25cosωs t + 25(1 − cos2ωs t)sinωsnt + 25(1 + 0.25(1 − cos2ωs t)sinωsnt)sinωat. (17) The occasional sign sways about the mean worth 100 mg L−1 incorporating the signs with occasional recurrence ωs = 2π/365 day−1 , spring/neap rakish recurrence ωsn = 2π/14 day−1 and shift in weather conditions precise recurrence ωa = 2π/12.5/24 day−1 . The 1 year of engineered SSC time arrangement c(t), beginning at 1 October with 15 min time step, is introduced at the lower part of Fig. 1, the relating occasional sign cs(t) is appeared at the highest point of Fig. 1. In spite of the fact that the determination of window length is a significant issue for SSA (Hassani et al., 2012; Hassani and Mehmoudvand, 2013), this paper picks a similar window length (L = 120) as that in Schoellhamer (2001) to look at the exhibition of the proposed strategy with that of Schoellhamer. Utilizing the manufactured time arrangement we figure the slacked relationship lattice and the fluctuations of every mode. The initial four modes contain the occasional segments, which represent 72.3 % of the absolute fluctuation; specifically, the primary mode contains 50.2 % of the complete difference. To assess the exactnesses of recreated PCs from the time arrangement with various rates of missing information, following the methodology of Shen et al. (2014), we register the overall blunders of the initial four modes inferred by ISSA and SSAM with the accompanying articulation: from information missing time arrangement, and a0 means the PCs reproduced from the time arrangement without missing information. We plan the examination of missing information by arbitrarily erasing the information from the engineered time arrangement. The level of erased information is from 10 to 60 % with an expansion of 10 % each time. At that point, we reproduce the initial four PCs from the information erased engineered time arrangement utilizing both SSAM and ISSA, and rehash the tests multiple times. The general mistakes of the initial four PCs are introduced in Fig. 2, from which we unmistakably observe that the exactnesses of reproduced PCs by our ISSA are clearly higher than those by SSAM, particularly for the second and fourth PCs. On account of 60 % missing information, the exactness enhancements are up to 19.64, 41.34, 23.27 and 50.30 % for the initial four PCs, individually. We recreate the time arrangement c(t) ˆ utilizing the initial four PC modes and afterward assess the nature of remade arrangement by analyzing the mistake 1c(t) ˆ = ˆc(t) − cs(t). For the cases whose missing information are between 10 to 50 % throughout the entire time arrangement, the reproduced part of the time arrangement demonstrated rate (IMP) of ISSA regarding SSAM is likewise recorded in Table 1. As the measure of missing information expands, the IMPs of both MARE and MRMSE increment too. Additionally, when the engineered time arrangement with the missing information is same as the genuine SSC time arrangement of Fig. 4, the IMPs of MARE and MRMSE are 8.87 and 15.19 %, separately.

4. PERFORMANCE OF ISSA WITH TIME SERIES

The mid-profundity SSC time arrangement at San Mateo Bridge is introduced in Fig. 4, which contains around 61 % missing information. This time arrangement was accounted for by Buchanan and Schoellhamer (1999) and Buchanan and Ruhl (2000), and broke down by Schoellhamer (2001) utilizing SSAM. We break down this time arrangement utilizing our ISSA with the window size of 30 h (L = 120) contrasting and SSAM. The initial 10 modes speak to predominant intermittent segments as appeared in Schoellhamer (2001) which contain 89.1 % of the absolute fluctuation. Along these lines, we remake the time arrangement with the initial 10 modes when the missing information in a window size is under 50 %. The lingering time arrangement, for example the distinctions of noticed short reproduced information, are introduced in Fig. 5. The greatest, least and mean outright residuals just as the SD are introduced in Table 2. Unmistakably both greatest and least residuals are altogether decreased by utilizing ISSA approach. The SD of our ISSA is decreased by 8.6 %. The squared relationship coefficients between the perceptions and the remade information from ISSA and SSAM are 0.9178 and 0.9046, individually, which mirror that the reproduced time arrangement with our ISSA can undoubtedly, to enormous degree, indicate the ongoing arrangement. nificantly more modest than those by SSAM. As the portion of missing information builds, the improvement of the overall mistake gets more prominent. At the point when the level of missing information arrives at 60 %, the upgrades of the initial four head segments are up to 19.64, 41.34, 23.27 and 50.30 %, separately. Besides, when the missing information represent 60 %, the MARE and MRMSE inferred by ISSA are 3.52 and 4.60 mg L−1 , and by SSAM are 5.37 and 6.96 mg L−1 . The comparing upgrades of ISSA concerning SSAM are 34.45 and 33.91 %. At the point when the missing information of manufactured time arrangement is equivalent to the genuine SSC time arrangement, the enhancements of MARE and MRMSE are 8.87 and 15.19 %, individually. The SD got from the genuine SSC time arrangement at San Mateo Bridge by ISSA and SSAM are 12.27 and 13.48 mg L−1 , and the squared connection coefficients between the perceptions and the remade information from ISSA and SSAM are 0.9178 and 0.9046, separately. Thusly, ISSA can surely, by and large, recover the useful signs from the first inadequate time arrangement.

5. CONCLUSION

We have built up the ISSA approach in this paper for preparing the inadequate time arrangement by utilizing the rule that a period arrangement can be imitated utilizing its primary parts. We demonstrate that the SSAM created by Schoellhamer (2001) is a unique instance of our ISSA. The exhibitions of ISSA and SSAM are shown with a manufactured time arrangement, and the outcomes show that the general mistakes of the initial four head segments by ISSA are sig-nificantly more modest than those by SSAM. As the division of missing information expands, the improvement of the general mistake gets more noteworthy. At the point when the level of missing information arrives at 60 %, the upgrades of the initial four head segments are up to 19.64, 41.34, 23.27 and 50.30 %, separately. Besides, when the missing information represent 60 %, the MARE and MRMSE determined by ISSA are 3.52 and 4.60 mg L−1 , and by SSAM are 5.37 and 6.96 mg L−1 . The relating upgrades of ISSA as for SSAM are 34.45 and 33.91 %. At the point when the missing information of engineered time arrangement is equivalent to the genuine SSC time arrangement, the enhancements of MARE and MRMSE are 8.87 and 15.19 %, separately. The SD got from the genuine SSC time arrangement at San Mateo Bridge by ISSA and SSAM are 12.27 and 13.48 mg L−1 , and the squared relationship coefficients between the perceptions and the recreated information from ISSA and SSAM are 0.9178 and 0.9046, separately. In this manner, ISSA can for sure, generally, recover the educational signs from the first fragmented time arrangement.

REFERENCES

1. Brockwell, P.J., Davis, R.A.: Introduction to Time Series and Forecasting. Springer, New York (1996)
2. Hill, H.S., Mjelde, J.W.: Challenges and opportunities provided by seasonal climate forecasts: a literature review. Journal of Agricultural and Applied Economics 34, 603–632 (2002)
3. Abu-Mostafa, Y., Atiya, A., Magdon-Ismail, M., White, H. (eds.): Special Issue on Neural Networks in Financial Engineering. IEEE Transactions on Neural Networks 12, 653–656 (2001)
4. Parlos, A., Oufi, E., Muthusami, J., Patton, A., Atiya, A.: Development of an intelligent long-term electric load forecasting system. In: Proc. Intelligent Systems Applications to Power Systems Conference, ISAP 1996, pp. 288–292 (1996)
5. Atiya, A., El-Shoura, S., Shaheen, S., El-Sherif, M.: A comparison between neural network forecasting techniques - case study: river flow forecasting. IEEE Transactions Neural Networks 10, 402–409 (1999)
6. Clements, M.P., Hendry, D.F.: Macro-economic forecasting and modelling. The Economic Journal 105, 1001–1013 (1995)
7. Chu, C.-W., Zhang, G.P.: A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of Production Economics, 217–231 (2003)
8. Box, G., Jenkins, G.: Time Series Analysis, Forecasting and Control. Holden-Day Inc., San Francisco (1976)

9. Andrawis, R., Atiya, A.F.: A new Bayesian formulation for Holt's exponential smoothing. Journal of Forecasting 28, 218–234 (2009)
10. Ahmed, N.K., Atiya, A.F., El Gayar, N., El-Shishiny, H.: An empirical comparison of machine learning models for time series forecasting. Econometric Reviews 29(5- 6), 594–621 (2010)
11. Lendasse, A., Francois, D., Wertz, V., Verleysen, M.: Vector quantization: a weighted version for time-series forecasting. Future Generation Computer Systems 21, 1056–1067 (2005)
12. Allison, P.D.: Missing Data. Sage Publications, Inc. (2001)
13. Andrawis, R., Atiya, A.F., El-Shishiny, H.: Forecast combination model using computational intelligence/linear models for the NN5 time series forecasting competition. International Journal of Forecasting 27, 672–688 (2011)
14. Rubin, D.B.: An overview of multiple imputation, in Survey Research Section. American Statistical Association (1988)
15. Honaker, J., King, G.: What to do about missing values in time series cross-section data. American Journal of Political Science 54, 561–581 (2010)
16. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B 39, 1–38 (1977)
17. Denk, M., Weber, M.: Avoid filling Swiss cheese with whipped cream: imputation techniques and evaluation procedures for cross-country time series. IMF Working Paper WP/11/151 (2011)