Imputation of missing aerosol data over Indian region

P. Kumar[1], P. Dave[2] and M. Bhushan[1]

[1]Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai, 400076, India
[2]Climate Studies, Indian Institute of Technology Bombay, Mumbai, 400076, India
Keywords: AOD; Missing value; Temporal correlation
Presenting author email: pratyushkumar@iitb.ac.in

The amount of aerosols present in atmosphere is typically quantified as Aerosol optical depth (AOD), where higher AOD implies higher amount of aerosols in the atmospheric column. In literature, AOD retrieved using remote sensing techniques have been widely used. However, the presence of clouds in the atmosphere along with non-uniform satellite overpass time over a given location poses a challenge in retrieving the AOD values at a frequent and constant rate (*Lou and Obradovic, 2011*). In recent literature techniques have been proposed to estimate the missing AOD values (*Lou and Obradovic, 2011; Radosavljevic et al., 2010*). In particular, imputation technique proposed by *Lou and Obradovic* (*2011*) was found to accurately estimate AOD even when the data had large fraction of missing values. However these techniques have not been validated over the Indian region where complex orography and aerosol characteristics such as predominance of anthropogenic aerosol over Indo-gangetic plain (IGP) (*Ramchandran and Kedia, 2013*) makes the AOD estimation problem a challenging one.

The current work focuses on application of imputation technique proposed by *Lou and Obradovic* (*2011*) to estimate missing AOD values over the Indian region. In particular, we exploit temporal correlation amongst AOD values to build stochastic linear time series models to estimate the missing values. Towards this end, the temporal correlation amongst AOD data is captured by use of a latent variable which is assumed to evolve as a first order Gaussian Markov process. The observed AOD value is assumed to be a noisy measurement of this latent variable, where the measurement noise is assumed to be Gaussian as well.

MODIS retrieved AOD L3 data at $1^o x 1^o$ resolution was used for the analysis for period of March-September (Pre-monsoon: March-May; Post-monsoon: June-September; total 214 days) for year 2000. Longitude and latitude of spatial domain was from 66.5^o-100.5^o and 6.5^o-40^o respectively. Fraction of missing values in the AOD data was largest over North-western India (>0.8) while over the IGP (R1) and Peninsular India (R3) fraction of missing values were small (<0.4) as shown in Fig.-1(a).

To characterize the performance of the stochastic linear time series modelling approach, only pixels with at-least 15% available values were considered. The available AOD values at each of these pixels were divided into training and testing sets. In particular, randomly chosen 10% of the available values were assigned to the testing set while the remaining 90% were assigned to the training set. At each pixel a separate time series model was obtained using the training set and used to impute the missing values corresponding to testing set. Mean squared error (MSE) was obtained from the differences of the imputed and the corresponding known testing values. This entire procedure was repeated ten times at each pixel and averaged MSE (ARMSE) across these 10 realizations are depicted in Fig.-1(b). It can be seen that the ARMSE values are lower in Western

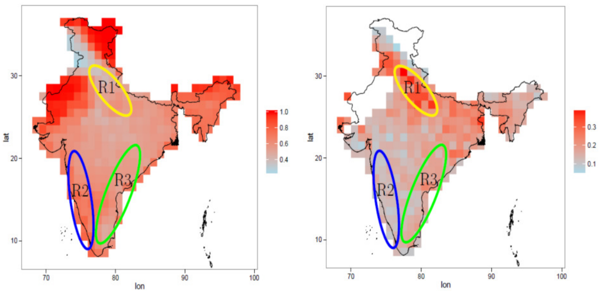

Figure 1: *(a) Fraction of missing values and (b) ARMSE value. R1, R2 and R3 represent regions with differing model performance.*

coastal (R2) and R3 regions while they are higher in R1 region, though the fraction of missing values is lower in R1. This suggests that the model performs relatively poorly in R1 as compared to Peninsular region. This could be due to the variability of aerosol characteristics over R1 where more anthropogenic aerosols are emitted during pre-monsoon period leading to significant changes in aerosol characteristic in this period. The improved model performance over R2 and R3 could be due to abundance of dust and sea-salt aerosols that show less seasonal variability during the period under consideration. The results thus indicate that the models performed better when the aerosol variability is low despite the missing value fraction being large.

The current work, to the best of our knowledge, is a first attempt at imputing observational AOD values over Indian subcontinent and further work is required to improve the imputation performance over regions with large aerosol variability.

Lou. Q, Obradovic. Z, "*Modelling multivariate spatio-temporal remote sensing data with large gaps*", Proc. 22nd Int. joint Conf. on Artificial Intelligence, Barcelona, Spain, 2011.

Radosavljevic, V., Vucetic, S., Obradovic, Z., "*A data mining technique for aerosol retrieval across multiple accuracy measures*", IEEE Geoscience and Remote Sensing Letters, vol. 7, no. 2, 2010, pp. 411-415.

Ramchandran, S., Kedia, S., "*Aerosol, clouds and rainfall: inter-annual and regional variations over India*", Climate Dynamics, vol. 40, 2013, pp 1591-1610.