Data Mining Support for Improvement of MODIS Aerosol Retrievals

Bo Han[1], Zoran Obradovic[1], Zhanqing Li[2], Slobodan Vucetic[1*]
[1]Center for Information Science and Technology, Temple University, Philadelphia, PA, USA
[2]Department of Meteorology, University of Maryland, College Park, MD, USA
[*]corresponding author <vucetic@ist.temple.edu>

***Abstract*—This paper describes data mining approach for improving the accuracy of aerosol retrieval algorithms. The approach was applied on 1,722 collocated MODIS and AERONET observations over the western part of the continental U.S. Neural networks were trained to predict AERONET Aerosol Optical Thickness (AOT) using attributes derived from observations made by MODIS instrument onboard TERRA satellite. The results showed that neural networks provide more accurate retrievals than the operational MODIS algorithm. Study of differences between neural networks and the MODIS algorithm revealed useful information that can help domain scientists improve quality of the MODIS algorithm.**

Keywords – aerosols, retrieval, MODIS, AERONET, data mining, neural networks, decision trees

I. INTRODUCTION

Aerosols are small particles emanating from natural and man-made sources that both reflect and absorb incoming solar radiation. One of the biggest challenges of current climate research is to characterize and quantify the effect of aerosols on the Earth's radiation budget [1]. There are two major instrument types that collect aerosol data: (1) satellite instruments, such as AVHRR-2, GOME, TOMS, PONDER, MISR, and MODIS [2]; and (2) ground-based instruments, represented by AERONET [3], a global network of about 180 operational sun/sky radiometers. Satellite instruments provide global coverage with high spatial resolution, relatively low temporal resolution and moderately accurate retrievals. AERONET has limited spatial coverage, relatively large temporal resolution and highly accurate retrievals. As a result, AERONET is often used to validate satellite-based retrievals.

Since 2000, MODIS instrument aboard the TERRA satellite represents a major source of high-quality aerosol information. Operational MODIS aerosol retrieval algorithm is an inverse operator of a high-dimensional non-linear function derived from forward-simulation models according to the domain knowledge of aerosol physical properties. It derives the aerosol optical thickness (AOT) by matching the observed spectral reflectance at the top of the atmosphere to the simulated values in lookup tables. Validation studies show that AOT retrieval by MODIS is more accurate over ocean than over land [4]. Main sources of retrieval errors are separation of surface and atmospheric components of the observed radiances, inaccuracies in the forward model, and inversion errors. Some sources of retrieval uncertainties, such as presence of bright surfaces or cloud-contaminated scenes, are inherent to the system and could not be corrected, while others, such as imperfections in the retrieval algorithm, are correctable. A major challenge for aerosol scientists is to understand the major sources of correctable retrieval errors and use this knowledge in improvement of retrieval algorithms. The goal of this study is to explore if data mining can help in that regard. We developed data mining tools to test if MODIS aerosol retrieval over land can be improved and, if the answer is affirmative, to identify conditions over which the improvements are possible.

Our approach has two main components: 1) use neural networks to learn relationships between MODIS observations and AOT; 2) use decision trees to detect conditions when the neural network is more accurate than MODIS retrievals. Neural network trained in the first step can be considered as a completely data-driven retrieval algorithm, which is in contrast to the MODIS operational algorithm that is model-driven. The drawback of neural network retrieval is that it can be accurate only over the conditions similar to those represented by training data. As such, neural networks are not a completely viable alternative to model-driven retrieval algorithms. However, if neural networks can achieve higher retrieval accuracy over the selected set of conditions, this provides a clear signal that accuracy of model-driven algorithm can be further improved. Decision trees developed in the second step should help in identification of such conditions.

II. DATA SETS AND METHODOLOGY

A. Data Sets

AERONET radiometers measure AOT in 10 spectral bands between 340nm and 1640nm in regular intervals during a day. We obtained Level 1.5 cloud-screened AERONET data for 15 sites at the West of the Continental U.S. (see Figure 1 and Table 1) during the three-year period between 2002 and 2004. MODIS instrument has a single camera observing radiances over 36 spectral bands between 410nm and 14μm at three different spatial resolutions (250m, 500m, 1km) [5]. MODIS has a swath width of 2330km and it achieves the global coverage every 1 – 2 days.

We collected MODIS radiance, cloud cover and aerosol data that are spatially and temporally collocated with the AERONET observations. Upon merging, a total of 1,722

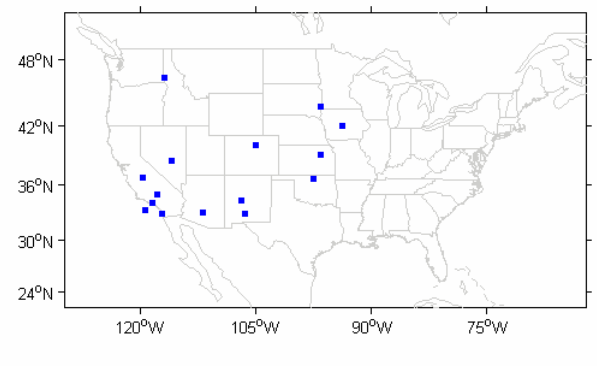

Figure 1. Locations of 15 AERONET sites

spatially (within 0.15°) and temporally (within 90′) collocated observations from MODIS and AERONET were available for our study. More specifically, every of the 1,722 observations corresponded to a spatial block of dimension 0.30°×0.30° surrounding an AERONET site. The observation was counted if the block contained at least one non-cloud pixel, if MODIS AOT retrieval was available, and if at least one AERONET measurement was available within 90 minutes of the satellite overflight. For each observation we collected the following information:

1. Average and minimum radiances over the cloud-free pixels within a block for the 7 lowest wavelengths (0.47 – 2.1μm). We note that the 7 lowest wavelengths are also used in the MODIS operational aerosol retrieval algorithm [6];
2. Average radiance uncertainties for the 7 lowest wavelengths over the cloud-free pixels;
3. 5 Angles (Solar zenith angle, Solar azimuth, Sensor zenith, Sensor azimuth, Scattering angle);
4. Fraction of cloud-free pixels within the block;
5. Fraction of cloud-free pixels over water, land, desert areas;
6. Average AERONET AOT retrieved within 90 minutes of the satellite overflight. Since AERONET does not provide AOT retrieval at 470nm wavelength, it was estimated by logarithmic interpolation of AERONET AOTs at 440nm and 675nm.

TABLE I. SUMMARY OF THE 15 AERONET SITES

	Location Name	Lat	Lon	Elevation (m)	# Points
1	Fresno	36.78	-119.77	0	118
2	San_Nicolas	33.25	-119.48	133	10
3	UCLA	34.07	-118.45	131	42
4	Rogers_Dry_Lake	34.92	-117.88	680	18
5	La_Jolla	32.87	-117.25	115	56
6	Rimrock	46.48	-116.99	824	106
7	Railroad_Valley	38.50	-115.96	1435	93
8	Maricopa	33.06	-111.97	360	144
9	Sevilleta	34.35	-106.88	1477	299
10	White_Sands	32.91	-106.35	1197	31
11	Boulder	40.04	-105.00	1604	303
12	Cart_Site	36.60	-97.48	318	153
13	Sioux_Falls	43.73	-96.62	500	40
14	KONZA_EDC	39.10	-96.61	341	287
15	SMEX	41.93	-93.66	1030	22

B. *Data-Driven AOT Prediction*

The basic assumption in this study was that AERONET AOT retrievals are highly accurate and can be used as a proxy to the ground truth. This assumption is supported by the previous results that show that AERONET retrievals are up to 5 times more accurate than MODIS retrievals [7]. Here, we constructed neural networks that predict AERONET AOT at 470 nm (listed as 6. above) based on the MODIS attributes (listed as 1. – 5. above). The constructed attributes contain basically the same information as the MODIS operational retrieval algorithm. With such attribute choice we were able to get an objective estimate of the possible improvements that could be achieved by modifications of the existing algorithm.

In our experiments, we used feedforward neural networks with a single hidden layer with 10 sigmoid neurons and a linear neuron at the output. The accuracy was estimated by 3-cross validation: in each of the 3 rounds of cross validation neural network was being trained on data from two years and tested on data from the remaining year; the procedure was repeated 3 times, each time using different year as the test set. Prediction accuracy was evaluated by the correlation coefficient between predictions and AERONET AOT observations (CORR), and by the root mean square error (RMS).

C. *Decision Tree Analysis of MODIS Retrieval Errors*

Given the neural network predictions on test data, we compared retrieval errors of neural networks and MODIS algorithm in an unbiased fashion. Specifically, we labeled the data where neural network predictions are significantly more accurate than those of MODIS algorithm as positives, and the remaining data as negatives. If positives and negatives could be discriminated, this would indicate that accuracy of MODIS algorithm could be improved. In this study, we trained decision trees for the discrimination task. The advantage of decision trees is that by analysis of rules that they generate, we could be able to understand what retrieval scenarios lead to the highest improvements of the retrieval accuracy.

We used the following 14 attributes in decision tree training: average MODIS AOT uncertainty at 470nm (A1), average retrieved MODIS AOT at 470nm and 660nm (A2-A3), minimum radiances for the 7 lowest wavelengths (A4-A10), fraction of non-cloud pixels (A11), fraction of pixels over three different surface types (Water, Land, Desert) (A12-A14). Since MODIS provides one retrieval for each 10km×10km region, and there are several such regions within a block, the MODIS AOT was calculated as the average of all MODIS AOT retrievals within the block.

III. RESULTS

In Table 2 we compare the accuracies of MODIS algorithm and neural network predictors after the 3-cross-validation experiment. As seen, based on the root mean squared error (RMS) accuracy measure, neural networks are almost twice more accurate than the MODIS algorithm. By careful inspection of scatter plots in Figure 2.a (AERONET AOT vs. MODIS AOT) we can observe a strong bias in MODIS algorithm that tends to overestimate the AOT values. On the other hand, neural network predictions appear less biased and

tend to underestimate the AOT (see Figure 2.b – AERONET AOT vs. NN Predictions).

The correlation coefficient (CORR) accuracy measure is neglecting the bias term. This measure is reasonable since the bias in MODIS retrievals could be easily corrected. The overall CORR accuracy of neural network predictors (CORR = 0.662) is slightly better than that of the MODIS algorithm (CORR = 0.640). Both RMS and CORR results indicate that MODIS algorithm could be further improved.

It is interesting to observe the year-by-year fluctuations in the accuracy of both retrieval algorithms. While these fluctuations are quite significant, the difference between MODIS algorithm and neural networks is quite consistent.

In Table 3 we summarize seasonal variations in prediction accuracy. The largest difference between MODIS algorithms and neural networks occurred between January and March, while in the remaining seasons the difference in CORR accuracy was rather small. It is also interesting to observe that the CORR accuracy of both retrieval algorithms was lower during winter months. On the contrary, the RMS measure indicates the decreased accuracy during the summer months. This phenomenon is explained by the known result that AOT values are significantly larger during summer months.

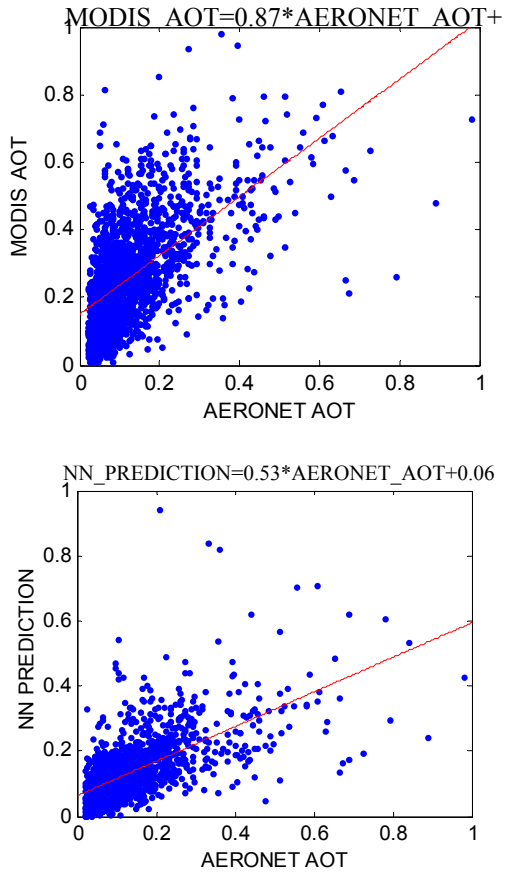

Figure 2. a). AERONET AOT vs. MODIS AOT, b) AERONET AOT vs. neural network prediction

Finally, we explored influence of surface types on the retrieval accuracy. The results indicate that neural networks are the most benefitial over the desert surface type. This results is explained by the increased surface reflectance of desert areas. From previous studies, it is known that MODIS algorithm is the least successful over the bright areas. Our results suggest that it should be possible to significantly improve MODIS retrieval accuracy over bright desert areas.

TABLE II. ACCURACY COMPARISON BY YEAR

Year	# Points	MODIS Retrieval		NN Prediction	
		RMS	CORR	RMS	CORR
2002	588	0.185	0.755	0.099	0.785
2003	576	0.212	0.578	0.107	0.568
2004	558	0.192	0.523	0.069	0.600
Overall	1722	0.197	0.640	0.094	0.662

TABLE III. ACCURACY COMPARISON BY SEASON

Season	# Points	MODIS Retrieval		NN Prediction	
		RMS	CORR	RMS	CORR
Jan-Mar	358	0.171	0.422	0.077	0.553
Apr-Jun	652	0.221	0.636	0.104	0.642
Jul-Sep	378	0.213	0.625	0.101	0.659
Oct-Dec	334	0.124	0.580	0.080	0.582

TABLE IV. ACCURACY COMPARISON BY LAND TYPE

Surface Type	# Points	MODIS Retrieval		NN Prediction	
		RMS	CORR	RMS	CORR
Water	28	0.209	0.705	0.057	0.718
Coast	3	0.135	-0.142	0.054	0.612
Desert	577	0.247	0.595	0.087	0.638
Land	1088	0.154	0.733	0.098	0.673

The question we posed next was: "could we explain situations where neural networks are significantly more accurate than MODIS algorithm"? To answer this question, we labeled the data where "neural networks are at least 3 times more accurate than MODIS algorithm AND error of MODIS retrieval is larger than 0.05" as positives, and the remaining data as negatives. Decision trees are a supervised learning technique that learns non-linear relationships between attributes and classification targets that could be easily represented as a set of classification rules. A decision tree classifier was constructed on such data, and the result was (Table 5) that it can discriminate between positives and negatives with 73.1% accuracy, which was above 57.6% accuracy of a trivial predictor (since there were 57.6% of positives).

TABLE V. DECISION TREE ACCURACY

	Positives	Negatives	Accuracy	Sensitivity	Specificity
2002	315	266	0.7418	0.7619	0.7180
2003	354	259	0.7080	0.7825	0.6062
2004	330	209	0.7477	0.8152	0.6411

This result indicates that it is possible to obtain a partial understanding of conditions where MODIS algorithm can be improved. In Figure 3, we show an initial portion of the resulting decision tree that had a total of 104 nodes. Analysis of the initial portion of the decision tree reveals that MODIS

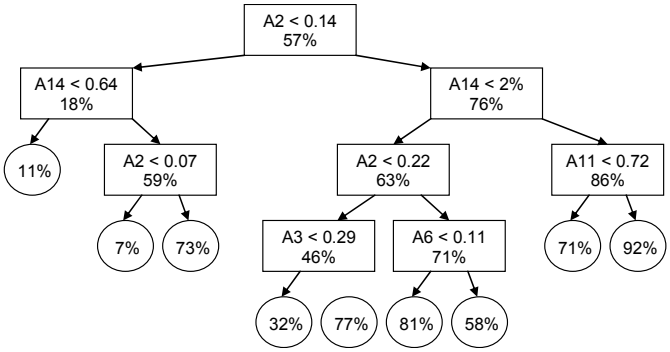

Figure 3. Decision Tree. Boxes show the splitting rule and fraction of positives. The circles represent fraction of positives in tree leafs. Attribute decription: A2 - MODIS AOT at 470nm, A3 - MODIS AOT at 660nm, A6 - AOT radiance at 660nm, A11 - fraction of clouds, A14 - fraction of desert sites.

algorithm can be improved in cases when MODIS AOT retrieval has high values, when Angstrom exponent is large, over areas contaminated with clouds, and over desert areas.

IV. CONCLUSIONS

We proposed a data mining method to help aerosol scientists in improvement of MODIS operational retrieval algorithm. The results over West of Continental U.S. during the period 2002-2004 strongly indicate that MODIS aerosol retrieval accuracy could be significantly improved. Moreover, experiments with decision trees provided valuable clues about the conditions where potential for the improvement is the largest.

ACKNOWLEDGMENT

We thank Amy Braverman for many useful discussions about the project. We thank Atmospheric Sciences Data Center at NASA Goddard Space Flight Center for their help in collection of MODIS data. We also thank the AERONET PIs for providing the ground-based aerosol data.

REFERENCES

[1] Y. J. Kaufman, D. Tanre, and O. Boucher, "A satellite view of aerosols in the climate system," Nature, 419: 215-223, 2002.

[2] M. D. King, Y. J. Kaufman, P. W. Menzel, and D. Tanreacute, "Remote sensing of cloud, aerosol, and water vapor properties from the Moderate Resolution Imaging Spectrometer (MODIS)," IEEE Trans. on Geoscience and Remote Sensing, 30: 2-27, 1992.

[3] B. N. Holben, T. F. Eck, I. Slutsker, T. Tanre, J. P. Buis, A. Setzer, E. Vermote, J. A. Reagan, Y. J. Kaufman, T. Nakajima, F. Lavenu, I. Jankowiak, and A. Smirnov, "AERONET: a federated instrument network and data archive for aerosol characterization," Remote Sens. Environ., 37 : 2403 – 2412, 1998.

[4] W. A. Abdou, D. J. Diner, J. V. Martonchik, C. J. Bruegge, R. A. Kahn, B. J. Gaitley, K. A. Crean, L. A. Remer, and B. Holben, "Comparison of coincident MISR and MODIS aerosol optical depths over land and oceans scenes containing AERONET sites," Journal of Geophysical Research, 110: D10S07, doi: 10.1029/2004JD004693, 2005.

[5] V. V. Salomonson, W. L. Barnes, P. W. Maymon, H. E. Montgomery, and H. Ostrow, "MODIS Advanced facility instrument for studies of the earthy as a system," IEEE Trans. ON Geoscience and Remote Sensing, 27(2) : 145-153, 1989.

[6] Y. J. Kaufman and D. Tanré, "MODIS ATB: Algorithm for Remote Sensing of Tropospheric Aerosol From MODIS," (http://modis.gsfc.nasa.gov/data/atbd /atbd_mod02.pdf), 1998.

[7] D. A. Chu, Y. J. Kaufman, C. Ichoku, L. A. Remer, D. Tanre´, and B. N. Holben, "Validation of MODIS aerosol optical depth retrieval over land," Geophys. Res. Lett., 29(12) : 8007, 2002.