Data Mining Support for Aerosol Retrieval and Analysis – Project Summary*

Zoran Obradovic[1], Bo Han[1], Qifang Xu[1], Yong Li[1], Amy Braverman[2], Zhanqing Li[3], Slobodan Vucetic[1]

[1]Center for Information Science and Technology, Temple University, 1805 N. Broad St, Philadelphia, PA 19122

[2]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr, Pasadena, CA 91109

[3]Dept. of Meteorology, University of Maryland, 2207 Computer & Space Sciences Bldg, College Park, MD 20742

1 Introduction

The Earth observations colleted by satellite and ground-based instruments are used to estimate important geophysical parameters such as atmospheric temperature profiles, cloud/aerosol properties, snow/ice cover, or vegetation cover. This process is called *the retrieval*. The retrieved parameters are then used in applications ranging from natural resource monitoring to the development of general circulation models for climate modeling. Achieving accurate and timely retrievals is a critical requirement for the success of ensuing analyses. Our project is focused at facilitating retrievals of aerosol aimed to characterize and quantify its effect on the Earth's radiation budget.

Aerosols are small particles emanating from natural and man-made sources that both reflect and absorb incoming solar radiation. We consider aerosol-related data collected by MODIS and MISR instruments aboard TERRA and AQUA satellites and also by ground based AERONET instruments. MODIS (Moderate Resolution Imaging Spectrometer) instrument observes reflected solar radiation at every point on Earth through 36 spectral bands between 0.41μm and 14μm with a spatial resolution of 250 m - 1 km and a temporal resolution of about 1-2 days [King 1992]. MODIS swaths are 2330 km wide and data is collected in daylight and also at night [Salomonson 1989]. Such a resolution results in the daily collection of about 1.5GB of raw observation data and in the generation of a much larger amount of processed data than previously collected. MISR (the Moderate Resolution Imaging Spectrometer) instrument measures reflected solar radiation from nine view angles along the direction of flight (along-track), and in four spectral bands at each angle. Its spatial resolution is 275 m or 1.1 km depending on band and angle. On each orbit, MISR sweeps out a 360 km wide swath of data from north to south while in daylight. Since MISR does not collect data at night, consecutive swaths are separated geographically resulting in 14 or 15 evenly spaced swaths per day. MODIS and MISR's ground footprint repeats nearly exactly every 16 days (1 *cycle*), which is the time it takes TERRA to fly 233 distinct orbital paths. Finally, AERONET (the Aerosol Robotic Network) is a global remote sensing network of about 180 operational sun/sky radiometers that provides aerosol information from the ground [Holben 1998]. AERONET radiometers measure AOT (Aerosol Optical Thickness) in 10 spectral bands between 340nm and 1640nm. AERONET has relatively high accuracy and precision [6] and is widely used in validation of satellite-based AOT retrievals [Chu 2002; Liu 2004].

2 Specific Aims

Most operational algorithms that retrieve aerosol from remote sensing data are *deterministic*. They are constructed as inverse operators of high-dimensional non-linear functions derived from forward-simulation models according to the domain knowledge of aerosol physical properties. Development of an inversion algorithm is often a compromise between the retrieval accuracy, speed and requirement for a prior knowledge. Due to the modeling complexity and computational constraints, the conventional

*Corresponding author: Zoran Obradovic, Professor and Director, Center for Information Science and Technology, Temple University. Tel: 1-215-2046265, Fax: 1-215-2045082, Email: zoran@ist.temple.edu deterministic retrieval approaches are based on lookup tables representing the most common conditions. As a result, the retrieval accuracy is suboptimal due to inability to use all the available information. In this regard, two substantial improvements are yet to be exploited, namely, the use of multi-sensor data and ever increasing ground-truth information. To date, high-quality ground-based and air-borne observations have been employed chiefly to validate the retrieval results, rather than to directly aid the retrieval. After major sources of retrieval error are detected, this information is used to manually readjust the operational deterministic algorithm. This highly subjective and time-consuming process is not likely to optimally utilize the available satellite and ground-based data.

Our overall objective is to achieve significant advances in retrieval algorithm development, application and modification through the development of a data mining methodology that utilizes multi-sensor satellite data of varying quality together with increasing amounts of ground-based measurements. We aim at improving the quality of single-sensor and multi-sensor aerosol retrievals by increasing efficiency of deterministic retrievals and by developing statistical algorithms that rely on the principles of inductive learning. Experiments summarized in this report were performed to explore:

1. If data-driven statistical AOT retrievals can serve as a practically useful complement to traditional deterministic retrieval methods;
2. If data mining can help understanding the major sources of correctable retrieval errors of deterministic retrievals; and
3. If data mining can facilitate development of joint-sensor retrieval algorithms that take advantage of aerosol observation from multiple instruments.

Our approaches for exploring these questions are summarized in the Methodology section, followed by the findings summarized in the Results section.

3 Methodology

Statistical retrievals. Our statistical AOT retrieval approach is based on developing regression methods for learning a mapping from observation attributes to the corresponding parameters [Han 2005a; 2005b]. As a global model we considered a global neural network trained on data from the entire domain. Local neural networks and local spatial interpolation models were developed using data from a limited region (single orbit in our experiments). Hybrid statistical models were constructed by combining predictions obtained from component neural networks and spatial interpolation models. These components were integrated using weighted averaging with weights optimized to minimize the mean squared prediction error over a specific region. In such an approach a global component is dominant at regions statistically similar to global data while region specific components dominate elsewhere. Another kind of hybrid is also considered where spatial interpolation models were trained to correct prediction error of global neural networks. This approach is based on our observation that error in global statistical models has a strong spatial component that can be significant over distances larger than 100km. Therefore, if a deterministic algorithm is overestimating over a given location, it is highly likely that it will overestimate over neighboring locations up to 100km away and this is where spatial interpolation can help correct the error.

Understanding retrieval errors. Our approach to discovering sources of retrieval errors in deterministic retrievals [Han, in review] has two main components: 1) use of neural networks to learn relationships between observations and AOT; 2) use of decision trees to detect conditions when the neural network is more accurate than model-driven deterministic retrievals. The drawback of data-driven neural network retrieval is that it can be accurate only over the conditions similar to those represented by training data. As such, if neural networks can achieve higher retrieval accuracy over the selected set of conditions, this provides a clear signal that accuracy of MODIS algorithm can be further improved. Decision trees were developed in the second step to help in identification of such conditions.

Development of joint-sensor retrievals algorithms. We used multi-source aerosol data consisting of attributes derived from MISR and MODIS observations and of a target attribute representing AOT retrieved by AERONET instruments. Given such data, the problem of AOT retrieval was treated as nonlinear regression. In our approach [Xu, 2005], from the nonlinear regression perspective, the only difference between single-sensor and multi-sensor retrieval is the number of attributes used in regression. This is in contrast with challenges that occur in the development of joint-sensor retrieval using deterministic algorithms.

4 Results

Statistical AOT retrievals [Han 2005a; b]. We have performed statistical retrieval experiment using MISR aerosol data collected in 2002 over the entire continental USA during four 16 day cycles of non-cloud points over the 17.6km×17.6km grid. The obtained AOT retrieval results suggest that both global and local statistical approaches can serve as a practically useful complement to traditional deterministic retrieval methods. We found that statistical prediction of AOT for orbits in the western US is more difficult than such prediction in the east. The results also showed that the hybrid approaches (averaged ensembles and spatial models used to correct errors of global neural networks) achieved higher overall accuracy than either local or global models did alone. Although the best overall results were obtained through weighted averaging of global and local neural networks, replacement of local neural networks by spatial interpolation models achieved comparable accuracy, and had the additional benefit of running faster. The benefits of the hybrid statistical approaches were particularly clear when smaller fraction of deterministic AOT retrievals were used for training. This suggests that integrating statistical and deterministic AOT retrievals may be useful for obtaining high quality AOT retrievals at higher resolutions, without significant additional computational burden.

Understanding retrieval errors [Han, in review]. To test if MODIS aerosol retrieval can be improved, we collected 1,722 spatially (within 0.15°) and temporally (within 90 minutes) collocated observations from MODIS and AERONET, which cover 15 AERONET sites at the West of the Continental U.S. during the three-year period between 2002 and 2004. Attributes for training of neural networks were derived from MODIS radiance, cloud, and aerosol products where we made sure that neural network did not use any additional attributes as compared to the MODIS algorithm. The target attribute was taken as AERONET AOT retrieval at 470nm. A neural network with 10 hidden nodes was used and its accuracy was estimated by 3-cross validation, where it was being trained on data from two years, and tested on data from the remaining year. The accuracy results show that the overall 0.10 root mean squared error (RMS) of neural network is significantly lower than the 0.20 error achieved by MODIS retrieval. This result is a strong indicator that the MODIS algorithm can be further improved. The question we posed next was: "could we explain situations where ANN is significantly more accurate than MODIS algorithm"? To answer this question, we labeled the data where "ANN is 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. A decision tree classifier was constructed on such data, and the result was that it can discriminate between positives and negatives with 72% accuracy, which was above 57% accuracy of a trivial predictor (since there were 57% of positives). This result indicates that it is possible to obtain a partial understanding of conditions where MODIS algorithm can be improved. Analysis of the decision tree reveals that MODIS 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.

*Joint-sensor retrieval*s [Xu, 2005]. We collected 75GB of MISR, and 750 GB of MODIS data covering the continental United States over three 16-day cycles in 2002. MISR and MODIS data were joined when both retrievals were available within 30 km of the AERONET location during the day on which daily AERONET AOT retrieval was available. This resulted in 118 data points collected over 34 AERONET locations. Rows 1 and 4 in Table 1 show R^2 accuracies of operational MISR and MODIS retrieval algorithms evaluated at the 118 AERONET points. The remaining rows are out-of-sample accuracies of neural networks. For example, Row 2 corresponds to statistical retrieval using MISR AOT and MISR data as training attributes, while Row 3 is an experiment performed training neural networks based on MISR data alone (MISR AOT was not used).

Table 1. Accuracy comparison. MISR (MODIS) Data denotes radiances and ancillary attributes from MISR (MODIS). MISR (MODIS) AOT denotes retrievals by the operational algorithms. Sign +/– denotes whether MISR/MODIS AOT and/or MISR/MODIS Data were used as inputs for AERONET AOT prediction

Experiment	MISR AOT	MISR Data	MODIS AOT	MODIS Data	R^2 on AERONET AOT
1	+	–	–	–	0.593
2	+	+	–	–	0.632 ± 0.065
3	–	+	–	–	0.651 ± 0.056
4	–	–	+	–	0.390
5	–	–	+	+	0.545 ± 0.076
6	–	–	–	+	0.312 ± 0.061
7	–	+	–	+	0.700 ± 0.065
8	+	+	+	+	0.684 ± 0.047

The results show that:

- Completely data-driven single-sensor statistical retrieval was more accurate than the operational deterministic retrieval for MISR (row 3 vs. 1), while it was less accurate for MODIS (row 6 vs. row 4). This indicates that statistical retrieval is a promising approach and that it is a viable alternative/complement to deterministic retrieval.
- Using aerosol deterministic retrieval as an additional input to the neural network significantly improved accuracy for MODIS (row 5 vs. 6), while it did not significantly reduce accuracy for MISR (row 2 vs. row 3). This indicates that combining data-driven and physical modeling approaches can improve the retrieval.
- Joint-source statistical retrieval (rows 7 and 8) was significantly more accurate than the single-sensor alternatives (rows 1-6). This illustrates the potential benefits of using observations from multiple instruments to improve retrieval quality.

Conclusion

The reported results are encouraging and provide conceptual proof for the explored tasks. The ideas of this study will be expanded by using more sophisticated data mining techniques and evaluated on much larger and more representative aerosol data sets.

References

1. Chu, D. A., Kaufman, Y. J., Ichoku, C. , Remer, L. A. , Tanre´, D. and Holben, B. N. "Validation of MODIS aerosol optical depth retrieval over land," *Geophys. Res. Lett.*, 29(12): 8007, 2002.
2. Han, B., Obradovic, Z, Li, Z. and Vucetic, S., (in review) "Data Mining Support for Improvement of MODIS Aerosol Retrievals."
3. Han, B., Vucetic, S., Braverman, A. and Obradovic, Z "Integration of Deterministic and Statistical Algorithms for Aerosol Retrieval," *Proc. International Conference on Novel Applications of Neural Networks in Engineering,* Lillie, France, pp. 85-92, 2005a.
4. Han, B., Vucetic, S., Braverman, A. and Obradovic, Z. "Construction of an accurate geospatial predictor by fusion of global and local models," *Proc. IEEE 8th International Conference on Information Fusion*, Philadelphia, PA, B.11.2 pp. 1-8, 2005b.
5. Herring, D.D., King, M.D., "Space-based observations of the Earth," Encyclopedia of Astronomy and Astrophysics, Murdin, Ed., Institute of Physics Publishing, 2959-2962, 2000.
6. Holben, B. N. , Eck, T. F., Slutsker, I. , Tanre, T. ,. Buis, J. P , Setzer, A. , Vermote, E. , Reagan, J. A., Kaufman, Y. J., Nakajima, T., Lavenu, F. , Jankowiak, I. and Smirnov, A. "AERONET: a federated instrument network and data archive for aerosol characterization," *Remote Sens. Environ.*, 37 : 2403 – 2412, 1998
7. King, M.D., Kaufman, Y.J., Menzel, W.P., Tanreacute, D., "Remote sensing of cloud, aerosol, and water vapor properties from the Moderate Resolution Imaging Spectrometer (MODIS)," *IEEE Tran. Geosci. Rem. Sen*., 30, 2-27, 1992.
8. Liu, Y. A. Sarnat, J.A., Coull, B.A., Koutrakis, P. and Jacob, D.J., "Validation of Multiangle Imaging Spectroradiometer (MISR) aerosol optical thickness measurements using Aerosol Robotic Network (AERONET) observations over the contiguous United States," *Journal of Geophysical Research*, 109, D06205, doi: 10.1029/2003JD003981, 2004.
9. Salomonson, V.V.. Barnes, W. L., Maymon, P.W., Montgomery, H.E. and Ostrow, H.. MODIS Advanced facility instrument for studies of the earthy as a system. *IEEE Trans. on Geoscience and Remote Sensing*, 27(2) : 145-153, 1989.
10. Xu, Q., Han, B., Li, Y., Braverman, A., Obradovic, Z. and Vucetic, S. "Improving aerosol retrieval performance by integrating AERONET, MISR, and MODIS data products," *Proc. IEEE 8th International Conference on Information Fusion*, Philadelphia, PA, B.11.3 pp. 1-8, 2005.