Aerosol Optical Thickness Retrieval from Satellite Observation Using Support Vector Regression

Thi Nhat Thanh Nguyen[1,2], Simone Mantovani[2,3], Piero Campalani[1,2], Mario Cavicchi[2], and Maurizio Bottoni[3]

[1] University of Ferrara, Via Saragat 1, 44122, Ferrara, Italy
[2] MEEO S.r.l, Via Saragat 9, 44122, Ferrara, Italy
[3] SISTEMA GmbH, Währingerstrasse 61, A-1090, Vienna, Austria
{thanhntn,mantovani}@meeo.it, cmppri@unife.it,
cavicchi@meeo.it, bottoni@sistema.at

Abstract. Processing of data recorded by the MODIS sensors on board the Terra and Aqua satellites has provided AOT maps that in some cases show low correlations with ground-based data recorded by the AERONET. Application of SVR techniques to MODIS data is a promising, though yet poorly explored, method of enhancing the correlations between satellite data and ground measurements. The article explains how satellite data recorded over three years on central Europe are correlated in space and time with ground based data and then shows results of the application of the SVR technique which somewhat improves previously computed correlations. Hints about future work in testing different SVR variants and methodologies are inferred from the analysis of the results thus far obtained.

Keywords: MODIS, Aerosol Optical Thickness, Earth Observation, Remote Sensing, Support Vector Regression.

1 Introduction

Remote Sensing allows measuring physical properties of distant objects often on dangerous or inaccessible areas where ground-based measurements are unfeasible. Using devices installed on board aircrafts or satellites, Remote Sensing applied to the Earth Observation makes it possible to monitor the Earth-Atmosphere system through the analysis of the interaction of radiation with matter. The signal received from the sensors is the sum of several contributions due to scattering, absorption, reflection and emission processes. Image processing techniques and specific algorithms allow extracting (direct measurement) or estimating (indirect measurement) the environmental parameters and their characteristics. Active and passive sensors with spectral capabilities ranging from visible to thermal infrared wavelengths are used for a large variety of applications for Earth Observation: Agriculture, Atmosphere, Forestry, Geology, Land Cover and Land Use, Ocean and Coastal monitoring.

Aerosol Optical Thickness (AOT) is representative of the amount of particulates presented in a vertical column of the Earth's atmosphere. AOT is largely used in air

I. Bloch and R.M. Cesar, Jr. (Eds.): CIARP 2010, LNCS 6419, pp. 492–499, 2010.

pollution monitoring applications because Particulate Matter (PM) concentration, one of the major pollutants that affect air quality, can be derived from it. Based on developments of observation technology, nowadays aerosol concentration can be predicted by elaboration of data recorded by satellite-based sensors, airborne instruments or ground-based measurements. Satellites provide monitoring at global scale, but with low temporal frequency and moderated retrieval accuracy. Conversely, ground-based measurements have higher precision and temporal frequency, but limits in spatial coverage.

MODerate resolution Imaging Spectrometer (MODIS) is onboard two polar orbiting satellites Terra and Aqua, launched in 1999 and 2002, respectively and operated by the National Aeronautic and Space Administration (NASA). The methodology for deriving AOT information from MODIS data consists of two main algorithms separated for land and ocean [1][2]. Validation of MODIS aerosol retrieval has shown that these algorithms perform on ocean better than on land [3]. In theory, this limitation occurs over bright surfaces and cloud-contaminated scenes as a result of the reflectance variability of clouds and different land surfaces. These factors strongly affect the Dense Dark Vegetation (DDV) approach used to estimate AOT on dark areas which usually correspond to pixels of vegetation and bare soil. This limitation was also shown in [4] in which AOT maps were validated over different land surfaces to point out the impact of the land cover types.

Data mining approach has recently been investigated to improve quality of aerosol monitoring. The application types ranged from classification, forecasting to estimation of aerosol content and properties from different sensors. Aerosol was distinguished from cloud in CALIPSO data by using Support Vector Machine (SVM) [5]. A series of data mining techniques was applied to analyze aerosol into chemical components and then processed their streams to understand aerosol dynamics in [6]. Besides, many studies emphasized on processing time-serial data to give prediction of air pollutants by using improved BP neural network [7], SVM [8], ensemble of SVM [9], or SVM and wavelet decomposition [10]. In aerosol estimation field, various applications of Neural Network (NN) were also considered [11][12].

Following this trend, in order to improve the traditional MODIS aerosol retrieval, many works proposed the application of data mining techniques on data collected by different instruments. Firstly, integrations of ground-based measurements AERONET (AErosol RObotic NETwork) and satellite data (MISR and MODIS [13], MODIS [14]) were made. Then, NNs techniques were applied on merged data to derive aerosol content and properties. This method proved efficiency in dealing with data uncertainties and in improving estimation accuracy which became comparable with that of results obtained at ground level. The same approach was mentioned in [15] to correct the bias of MODIS Aerosol Optical Depth (AOD) over different land covers by using both NNs and Support Vector Regression (SVR). In this work, SVR presented more advantages in performance than NNs.

In this paper we propose a driven-data approach that applies SVR, firstly introduced by Vapnik [16], on MODIS and AERONET data for AOT retrievals. This proposal is motivated by the better performance of SVR with respect to NNs in finding a global solution instead of a local one, and in coping with huge and high dimensional satellite data. Some similarities can be found in the work done by Vucetic et al. [14] in which NNs were applied on AERONET and MODIS (Collection 004) data

covering the U.S. continental area, recorded between 2002 and 2004. In our approach, we investigated using different prediction methods applied to a different data set: (i) the area of interest covered Europe instead of U.S. continental area and (ii) the improved MODIS Collection 005 products [1] were collected from 2006 to 2008. Moreover, AOT for each MODIS pixel sized 10 km is predicted instead of AOT for a box of 30x30 km^2. This approach is more appropriate for air pollution monitoring over urban areas where the assumption of aerosol stableness in an area of 50x50 km^2 [17] is less appropriate and higher spatial resolution is desirable.

The data fusion methodology and some details of SVR application are presented in section 2. Numerical experiments and their results are discussed in section 3. Finally, conclusions are given in section 4, together with hints about future works.

2 Methodology

The methodology applied for AOT retrievals based on SVR technique consists of three main steps: (i) collecting and processing satellite-based data (MODIS) and ground-based sensor measurements (AERONET) over Europe for a period of three years, (ii) integrating and combining data from two sources having different temporal and spatial resolutions, and (iii) applying SVR technique in form of "instance SVR" and "aggregate SVR" [18] for aerosol estimation.

2.1 Data Sets

AERONET is a global system of ground-based remote sensing aerosol network established by NASA and PHOTONS (Univ. of Lille 1, CNES, and CNRS-INSU) [19]. It uses CIMEL Electronique 318A spectral radiometers, sun and sky scanning sun photometers, to provide AOT retrievals in various wavelengths: 0.340, 0.380, 0.440, 0.500, 0.675, 0.870, 0.940, and 1.020 μm, in intervals of 15 minutes. Because of high accuracy, AERONET data are often used to validate satellite AOT retrievals.

AERONET data level 2.0, cloud-screened and quality-assured, of 105 sites distributed in Europe, in 2006, 2007, and 2008, were collected. AOT at 0.500 μm, the closest to MODIS AOT at 0.550 μm, was used to create SVR aerosol retrievals.

MODIS provides Level 1B Calibrated Geolocation Data Set, presenting a spectrum region from 0.415 to 14.235 μm, separated into 36 bands at 1 km, 500 m, and 250 m resolutions at nadir. Original MODIS data are pre-elaborated by a software package, the most recent version of which is known as "Collection 005" described in detail in [1]. One of the most important products of the MODIS Atmosphere algorithms applied in Collection 005 is the retrieval of aerosol MOD04. It is based on data from Terra platform and supports the monitoring of the ambient aerosol optical thickness over oceans globally and over the continents. MOD04 products consist of AOT maps at seven wavelengths over ocean (0.470, 0.550, 0.670, 0.870, 1.240, 1.630, and 2.130 μm) and three wavelengths over land (0.470, 0.550, and 0.670 μm). All maps have the same spatial resolution of 10x10 km^2. Geometry information such as solar zenith angle, solar azimuth angle, sensor zenith angle, sensor azimuth angle, and scattering angle are also provided in this product.

We collected MODIS data in corresponding period of the retrieved AERONET data. Reflectance of seven bands, geometry information, and aerosol concentration are considered at 10x10 km^2 spatial resolution.

2.2 AERONET-MODIS Combination

AERONET and MODIS data are products of separate sensors, which causes problems of temporal and spatial resolution differences. Data combination aims at obtaining data collocated in space and synchronized in time. MODIS data are considered if their distances from AERONET sites are within a radius of 30 km, while the contemporaneous measurements of AERONET instruments are selected and averaged within a temporal window of 60 minutes around the satellite overpasses.

AERONET-MODIS combinations are separated into two sets: *instance data set* and *aggregate data set*. The first one consists of 66,225 samples, each of which is a combination of measurements on a single MODIS pixel with an averaged AERONET AOT value satisfying collocation and synchronization conditions. One sample is presented as a vector including AERONET AOT at 0.500 μm, MODIS geometric data (solar zenith angle, solar azimuth angle, sensor zenith angle, sensor azimuth angle, scattering angle) and seven MODIS reflectances (0.646, 0.855, 0.466, 0.553, 1.243, 1.632, and 2.119 μm). The aggregate data set contains 5,289 samples that are combinations of an AERONET AOT, averaged MODIS geometric data and averaged MODIS reflectances calculated on all cloud-free pixels around this AERONET site. These vectors are stored in the same format as ones in the instance data set.

2.3 Support Vector Regression

SVR was applied to instance data set and aggregate data set in order to create different data models for AOT retrievals, called *instance SVR* and *aggregate SVR* respectively. SVR with epsilon loss function and Radial Basic Function (RBF) kernel provided by LIBSVM [20] was used. The accuracy was measured on three year data cross-validation in which we repeated selections of two year data for training and one year data for testing. Root Mean Square (RMS) error and correlation coefficient (CORR) were calculated from SVR AOT prediction and AERONET AOT data. SVR regularizations were searched in appropriate range with exponentially growing sequences. For each case, cross-validation was applied on a training data set and the best accuracy was picked. At the end of searching process, the chosen regularizations minimized mean square error in the training phase.

Both instance and aggregate SVR were used to bring out data models for AOT prediction at pixels of 10x10 km^2. We made experiments on them to investigate their accuracy and consuming time. Besides, SVRs were applied separately on different land cover types in order to investigate the effect of surface reflectance on aerosol retrievals. Concerning the land cover analysis, a spectral rule-based software system, called SOIL MAPPER [21], were used to distinguish surface types. This software uses reflectances in eight wavelengths (0.66, 0.87, 0.47, 0.55, 1.64, 2.13, 11.03, and 12.02 μm) to identify 57 different classes, out of which 40 refer to different land types. In our experiments a compact classification mode with 12 land cover classes was used. Cloud, snow, and unclassified pixels were discarded, whereas the nine

remaining classes (see Table 4) were utilized to evaluate the SVR prediction model on a land cover basis. A land cover class for each pixel sized 10x10 km^2 was determined as result of application of the classification system on reflectances averaged from all cloudy-free pixels of 1x1 km^2 available in this area.

3 Experimental Results

Our experiments focused on assessing accuracy of the SVRs' AOT in comparison with AERONET AOT. We applied and considered results obtained by aggregate SVR, instance SVR, and MODIS aerosol algorithm at different conditions: by year, by season, and by surface type.

The accuracies of both instance and aggregate SVR estimators are slightly better than those of the MODIS algorithm, as summarized in Table 1. Based on RMS error and correlation coefficient between predicted AOT and AERONET AOT measurements, averaged in 3 year data, instance SVR achieves the highest accuracy, then aggregate SVR follows and finally the MODIS algorithm is. This order is justified by the increase of RMS errors (0.077, 0.084, and 0.090, respectively) and the decrease of correlation coefficients (0.835, 0.812, and 0.807, respectively). The MODIS and SVRs AOT data in 2008 seem to have low quality as shown by the lowest correlation with AERONET AOT. However, instance SVR, in this case, still outperforms (CORR=0.802) the aggregate SVR (CORR=0.758) and MODIS algorithm (CORR=0.764).

Table 1. MODIS algorithm, Aggregate SVR, and Instance SVR accuracy by year

Year	Obs.	MODIS		Aggregate SVR		Instance SVR	
		RMS	CORR	RMS	CORR	RMS	CORR
2006	21,555	0.095	0.831	0.087	0.847	0.086	0.850
2007	24,251	0.087	0.827	0.081	0.831	0.074	0.853
2008	20,455	0.087	0.764	0.084	0.758	0.072	0.802
Total	**66,225**	**0.090**	**0.807**	**0.084**	**0.812**	**0.077**	**0.835**

Table 2 shows in detail the consuming time of aggregate SVR and instance SVR for the above experiment. Executions are tested on a computer with Intel (R) Core(TM)2 CPU 6400 @2.13 GHz, 2Gb RAM and Ubuntu 8.10 platform. Instance SVR spends about 240 seconds to predict 66,255 data, while aggregate SVR uses much smaller amount of time, 26 seconds. This difference is mainly due to the number of aggregate data set used for training in aggregate SVR less than instance data set used in instance SVR (132,522 data compared to 10,778), which induces data models with different sizes. The performance time will be meaningful for further SVR applications that aim at increasing spatial resolution of aerosol retrievals. In fact, with 10x10 km^2 spatial resolution, each MODIS image consists of 135x203 pixels. Increasing spatial resolution up to 1x1 km^2, more than 2 million pixels in an image would need to be processed. Also, the slow performance of instance SVRs hints at the need for further investigations of data selection and application of pruning techniques in the training phase.

Table 2. Aggregate SVR vs. Instance SVR in consuming time performance

Year	Obs.	Aggregate SVR		Instance SVR	
		Training Data	Time (s)	Training Data	Time (s)
2006	21,555	3,549	7.4563	44,706	66.59
2007	24,251	3,378	8.9790	42,010	106.69
2008	20,455	3,851	9.4843	45,806	70.94
Total	**66,225**	**10,778**	**25.9196**	**132,522**	**244.22**

We carried out the same further experiments on data sets separated by seasons and surface types to consider effects of meteorological conditions and surface reflectance on aerosol retrieval. Data in pairs of years were used for training SVRs, while data on the remaining year were classified by seasons and surface classes for testing purposes.

In autumn period (Oct.-Dec.), aerosol retrieval has the lowest accuracy obtained in all algorithms. As shown in Table 3, instance SVR has the most competitive accuracies that are better than those of MODIS algorithm in spring (Apr.-Jun.), summer (Jul.-Sep.) and autumn (Oct.-Dec.) and slightly worse in winter (Jan.-Mar.). The aggregate SVR presents its weakness for AOT estimation in winter (CORR=0.799). The RMS errors of all retrieval algorithms, which are higher in spring and summer, reflect the fact that larger AOT values are observed during these periods [2].

Table 3. MODIS algorithm, Aggregate SVR, and Instance SVR accuracy by season

Season	Obs.	MODIS		Aggregated SVR		Instance SVR	
		RMS	CORR	RMS	CORR	RMS	CORR
Jan. - Mar.	9,014	0.074	0.828	0.073	0.799	0.066	0.819
Apr. - Jun.	21,885	0.094	0.824	0.088	0.814	0.082	0.837
Jul. - Sep.	24,465	0.096	0.791	0.089	0.798	0.081	0.825
Oct. - Dec.	10,452	0.079	0.728	0.073	0.733	0.071	0.742

MODIS used two algorithms for land and ocean because of different physical interactions between aerosol and matters. Among all surface types listed in Table 4, only the water class refers to water pixels while remaining surfaces present the land pixels. MODIS ocean algorithm gained high accuracy (RMS=0.067, CORR=0.822), but it can be further improved by instance SVR (RMS=0.062, CORR=0.850). Out of land surface types, four classes Peat Bog, Evergreen Forest, Agricultural Areas and/or Artificial non Agricultural, Areas Scrub/Herbaceous Vegetation have a small number of samples, so their results should not be considered. In all remaining cases, instance SVR is more accurate than the MODIS algorithm. The biggest improvement can be observed at Artificial Surfaces and/or Open Spaces with little or no Vegetation surface, which is consistent with results of previous studies that showed the poor performance of the MODIS algorithm on bright surfaces [4].

Aggregate SVR has the worst accuracy on water pixels. It can be explained as result of the small contribution of water pixels on averaged data used for training aggregate SVR model, that didn't occur with instance SVR. This phenomenon influences pixels belonging to other surface types except Deciduous Forest and/or Agriculture Area class that has a large data set and therefore can be represented well by averaged values.

Table 4. MODIS algorithm, Aggregate SVR, and Instance SVR accuracy by surface

Classes	Obs.	MODIS		Aggregate SVR		Instance SVR	
		RMS	CORR	RMS	CORR	RMS	CORR
Water	2,981	0.067	**0.822**	0.071	**0.799**	0.062	**0.850**
Peat Bogs	91	0.112	**0.622**	0.151	**0.527**	0.129	**0.550**
Deciduous Forest	2,734	0.086	**0.692**	0.072	**0.681**	0.065	**0.700**
Evergreen Forest	19	0.054	**0.489**	0.065	**0.584**	0.053	**0.714**
Deciduous Forest and/or Agricultural Area	34,316	0.080	**0.824**	0.075	**0.824**	0.702	**0.833**
Agricultural Areas and/or Artificial non Agricultural Areas	25	0.103	**0.895**	0.086	**0.926**	0.080	**0.950**
Scrub/Herbaceous Vegetation and/or Agricultural Areas	5,302	0.082	**0.825**	0.083	**0.806**	0.075	**0.829**
Artificial Surfaces and/or Open Spaces with little or no Vegetation	5,961	0.096	**0.746**	0.085	**0.769**	0.078	**0.808**
Scrub/Herbaceous Vegetation	134	0.060	**0.892**	0.075	**0.871**	0.066	**0.882**

4 Conclusion and Future Works

In this paper an application of SVR technique on MODIS and AERONET data to predict AOT information has been presented. Satellite and ground-based data covering the European areas from 2006 to 2008 were considered. Then, SVRs were applied on instance data set and aggregate data set to make different non-linear regressions for aerosol retrievals. The experiment results show that SVR approach is competitive to MODIS algorithm and, especially, can improve prediction accuracy over areas having no or little vegetation. Out of two SVR models, instance SVR outperforms the aggregate SVR, but more improvements should be investigated to deal with training data overload and time execution.

In future, we will investigate the instance SVR more deeply in order to overcome the mentioned disadvantages. Also, this approach will be applied to estimate AOT at 1x1 km^2 spatial resolution, which is suitable for local-scale monitoring applications.

References

1. Remer, L.A., Tanré, D., Kaufman, Y.J.: Algorithm for Remote Sensing of Tropospheric Aerosol From MODIS: Collection 5. MODIS ATBD (2004)
2. Kaufman, Y.J., Tanre, D.: Algorithm for Remote Sensing of Tropospheric Aerosol from MODIS. In: MODIS ATBD (1997)
3. Abdou, W.A., Diner, D.J., Martonchik, J.V., Bruegge, C.J., Kahn, R.A., Gaitley, B.J., Crean, K.A.: Comparison of coincident Multiangle Imaging Spectroradiometer and Moderate Resolution Imaging Spectroradiometer aerosol optical depths over land and ocean scenes containing Aerosol Robotic Network sites. Journal of Geophysical research 110(D10S07), 11967–11976 (2005)
4. Nguyen, T.N.T., Mantovani, S., Bottoni, M.: Estimation of Aerosol and Air Quality Fields with PM MAPPER – An Optical Multispectral Data Processing Package. In: ISPRS Commission VII Symposium, Vienna, Austria (2010)

5. Ma, Y., Gong, W., Zhu, Z., Zhang, L., Li, P.: Cloud Amount and Aerosol Characteristic Research in the Atmosphere over Hubei Province, China, pp. III-631–III-634. IEEE/IGARSS (2009)
6. Ramakrishnan, R., Schauer, J.J., Chen, L., Huang, Z., Shafer, M.M., Gross, D.S., Musicant, D.R.: The EDAM project: Mining atmospheric aerosol datasets: Research Articles. International Journal of Intelligent Systems 20(7), 759–787 (2005)
7. Chen, Q., Shao, Y.: The Application of Improved BP Neural Network Algorithm in Urban Air Quality Prediction: Evidence from China. In: 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, pp. 160–163 (2008)
8. Lu, W., Wang, W., Leung, A.Y.T., Lo, S.M., Yuen, R.K.K., Xu, Z., Fan, H.: Air Pollutant Parameter Forecasting Using Support Vector Machine. In: Proceedings of the 2002 International Joint Conference on Neural Network, pp. 630–635 (2002)
9. Siwek, K., Osowski, S., Garanty, K., Sowinski, M.: Ensemble of Neural Predictors for Forecasting the Atmospheric Pollution. In: Proceedings of IEEE International Joint Conference on Neural Network, pp. 643–648 (2008)
10. Osowski, S., Garanty, K.: Wavelets and Support Vector Machine for Forecasting the Meteorological Pollution. In: Proceedings of the 7th Nordic Signal Processing Symposium 2006, pp. 158–161 (2006)
11. Okada, Y., Mukai, S., Sano, I.: Neural Network Approach for Aerosol Retrieval. IEEE/IGARSS 4, 1716–1718 (2001)
12. Han, B., Vucetic, S., Braverman, A., Obradovic, Z.: A statistical complement to deterministic algorithms for the retrieval of aerosol optical thickness from radiance data. Engineering Applications of Artificial Intelligence 19, 787–795 (2006)
13. Xu, Q., Obradovic, Z., Han, B., Li, Y., Braverman, A., Vucetic, S.: Improving Aerosol Retrieval Accuracy by Integrating AERONET, MISR and MODIS Data. In: 8th Intenational Conference on Information Fusion, Philadelphia, PA (2005)
14. Vucetic, S., Han, B., Mi, W., Li, Z., Obradovic, Z.: A Data-Mining Approach for the Validation of Aerosol Retrievals. IEEE Geoscience and Remote Sensing Letter 5(1), 113–117 (2008)
15. Lary, D.J., Remer, L.A., MacNeill, D., Roscoe, B., Paradise, S.: Machine Learning Bias Correction of MODIS Aerosol Optical Depth. IEEE Geoscience and Remote Sensing Letters 6(4), 694–698 (2009)
16. Vapnik, V.N.: The nature of statistical learning theory. Springer, New York (1995)
17. Ichoku, C., Chu, D.A., Mattoo, S., Kaufman, Y.J., Remer, L.A., Tanre, D.T., Slutsker, I., Holben, B.N.: A spatio-temporal approach for global validation and analysis of MODIS aerosol products. Geophysical Research Letter 29(12), 1–4 (2002)
18. Wang, Z., Radosavljevic, V., Han, B., Obradovic, Z., Vucetic, S.: Aerosol Optical Depth Prediction from Satellite Observations by Multiple Instance Regression. In: SIAM Conference on Data Mining, Atlanta, GA (2008)
19. AERONET - AErosol Robotic Network, http://aeronet.gsfc.nasa.gov/
20. LIBSVM: A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/
21. MEEO: Meteorological Environmental Earth Observation, http://www.meeo.it/