# Active Selection of Sensor Sites in Remote Sensing Applications

Debasish Das, Zoran Obradovic, Slobodan Vucetic
Center for Information Science and Technology
Temple University
Philadelphia, USA
debasish.das@temple.edu, zoran@ist.temple.edu, vucetic@ist.temple.edu

*Abstract*— **In a data-mining approach, a model for estimation of Aerosol Optical Depth (AOD) from satellite observations is learned using collocated satellite and ground-based observations. For accurate learning of such a spatio-temporal model, it is important to collect ground-based data from a large number of sites. The objective of this project is to determine appropriate locations for the next set of ground-based data collection sites to maximize accuracy of AOD estimation. Ideally, a new site should capture the most significant unseen aerosol patterns and should be the least correlated with the previously observed patterns. We propose achieving this aim by selecting the locations on which the existing prediction model is the most uncertain. Several criteria were considered for site selection, including uncertainty, spatial diversity, similarity in temporal pattern, and their combination. Extensive experiments on globally distributed data over 90 AERONET sites from the years 2005 and 2006 provide strong evidence that sites selected using the proposed algorithms improve the overall AOD prediction accuracy at a faster rate than those selected randomly or based on spatial diversity among sites.**

***Keywords-aerosol estimation, sensor site selection, uncertainty sampling, active learning***

## I. INTRODUCTION

In many remote sensing applications, especially those involving learning of some spatio-temporal phenomena, it is important to have sensor sites in places covering the major distribution patterns of the observed phenomenon. A question of fundamental importance is where to place these sensors given a set of possible locations. An application of this kind considered in this study is estimation, or retrieval, of Aerosol Optical Depth (AOD), which is a measure of the amount of sunlight absorbed by the atmospheric aerosols. Aerosols are small solid or liquid particles suspended in air, emanating from natural or man-made sources. One of the biggest challenges of today's climate research is to characterize and quantify the effect of aerosols on Earth's radiation budget.

Two types of sensors which collect data about aerosols are used to estimate AOD. They are satellite-based instruments such as MODIS and MISR [8] and ground-based instruments represented by the AErosol RObotic NETwork (AERONET) [3]. The former has higher spatial but lower temporal coverage and results in moderate AOD retrieval accuracy, whereas the later provides highly accurate retrieval but has very limited spatial coverage.

Traditional knowledge-based methods for AOD estimation from satellite observations are developed according to an understanding of physical properties of the aerosol followed by validation according to the ground-based observations [4]. Neural networks trained on satellite observations spatially-temporally merged with AERONET retrievals resulted in significantly improved accuracy of AOD retrievals as compared to the knowledge-based approach [7]. This improvement comes from the utilization of highly accurate ground-based measurements as targets during training of the neural network retrieval model. So, one can argue that adding more AERONET sensors will increase their retrieval accuracy further.

Intuitively, this improvement will be maximized if the new sensors are placed to capture the dominant AOD patterns unobserved by the existing set of AERONET sensors. This means that a pattern observed by the new sensor should be the least correlated with the already seen patterns. Given a set of AERONET sites already in place, the problem being addressed in this paper is to select a predetermined number of new sites from a fixed set of possible alternatives so as to maximize the improvement in retrieval accuracy.

Selection of a new site that is the least correlated with the set of existing sensors could be described as selection of a site where the predictor trained on the data generated from the current sensors is the most uncertain. Due to the similarity of the site selection process with active learning methods, we will call it *active site selection*. However, in conventional active learning, we are allowed to select for labeling any of the unlabeled points, whereas for active site selection, if a particular site is chosen, all satellite observations spatio-temporally collocated with that site will be added to the existing training set. So, the conventional point-by-point active learning algorithm has to be extended here to select the site having maximum average uncertainty.

An additional challenge in AOD estimation due to inherent properties of the satellite-based instruments is that the observation noise variance is dependent on the spatial-temporal properties of aerosols and land cover. However, it will be argued in the next section that this input-dependent noise variance does not affect the computation of pure model uncertainty using an ensemble of neural networks. The uncertainty at a particular point is calculated as the variance among the predictions from the members of the ensemble [9] [5]. In this project, we have explored several methods for selecting sites using this uncertainty information and taking into consideration spatial and temporal variability of the AOD values. To start with, we have applied two simple methods that select sites (1) randomly and (2) by maximizing spatial diversity. The alternative methods utilize the model uncertainty information to select sites having maximum uncertainty. We have also considered hybrid approaches that combine an uncertainty based selection with selection aimed to minimize the spatial, temporal and spatio-temporal correlations among selected sites, respectively.

In a previous related study on active sensor placement for river monitoring [2], a Gaussian Process model was assumed to describe a spatial variable and was used to derive an active sensor placement strategy to learn the underlying spatial distribution. However, the problem addressed in the monitoring problem is different from the current problem in two ways. First, in the monitoring problem, placing a sensor at a location returns a single observation of the measured variable whereas in our case, an AERONET site returns multiple AOD retrievals corresponding to different time instances. Second, in the river monitoring problem, the spatial variable (the pH value of the water) is predicted solely based on the location, while AOD is predicted from satellite-observed reflectances and other environmental attributes described later.

In summary, contributions of this study consist of:

1. Extending the standard active learning algorithms to active selection of AERONET sites.

2. Developing selection algorithms that take into account spatial and temporal correlations in aerosol data.

## II. Estimation of uncertainty

Several techniques have been proposed for uncertainty estimation in active learning. One of the most popular, due to its simplicity and robust performance, is the query by committee approach [9] and its variants. We summarize this technique in the following. Let us suppose a labeled data set is given as $D=\{(x_n, y_n),\ n = 1\ldots N\}$, where $x_n$ is a multi-dimensional attribute vector and $y_n$ is its label. The first step in uncertainty estimation is to build a committee of $K$ predictors, $f_i(x)$, $i = 1\ldots K$, trained on bootstrap replicates [4] (obtained using sampling with replacement) of the original data set $D$. Given the ensemble, its uncertainty $u(x)$ on an unlabeled point $x$, is obtained as the variance of its members,

$$u(x) = \frac{1}{(K-1)}\sum_{i=1}^{K}(f_i(x) - f(x))^2$$

where $f(x)$ is the average prediction of $K$ ensemble members. In this paper, we will be using feed-forward neural networks as the ensemble components.

In a typical active learning scenario, an unlabeled point with the largest uncertainty would be selected for labeling first. This criterion is modified for the active site selection, as explained in Section III. Before moving there, let us mention an issue that can be of practical importance. In many applications, including AOD retrieval, noise variance is not constant. Noise variance can be estimated by training an additional neural network on the squared errors of the ensemble predictor [1]. An interesting research question is whether and how should the noise variance information be incorporated in the active learning or active site selection. Our preliminary results (not shown in this paper) indicate that it could lead to the improvements in the active site selection, and it will be a subject of our future work.

## III. Active Site Selection

Given a set of $N_l$ locations where sensors are already installed and running, $L = \{L_i,\ i = 1\ldots N_l\}$ and a set of $N_s$ available locations for sensor installation, $S = \{S_i,\ i = 1\ldots N_s\}$, the objective is to install $N < N_s$ new sensors such that the benefit of installing them is maximized. The benefit is defined in terms of the improvement of prediction accuracy when labeled data from the new sites are added to training data. The methods proposed in this section will be compared to the random selection method that picks $N$ sites randomly from the available $N_s$ locations.

### A. Selection based on spatial diversity

Following the observation that neighboring locations are highly correlated, one approach is to install new sensors at spatially diverse locations. Specifically, in this approach we are selecting sites that are farthest away from the existing sites. The procedure described in Table I is to select sites in an iterative fashion where the first site is selected at location that is the most distant from $L$. The *distance of candidate site* $S_i$ is defined as the minimal distance between $S_i$ and any point from $L$, i.e. $d(S_i) = min_j\, dist(S_i, L_j)$. Upon adding selected location to $L$, the selection procedure continues until $N$ sites are selected.

### B. Site selection based on uncertainty

The traditional approach in active learning is to label the most uncertain examples. But in our case, instead of selecting an individual example, we have to select a site that will produce multiple labeled examples. Therefore, we define uncertainty $u(S_i)$ of site $S_i$ to be the average uncertainty over all unlabeled examples available for that site. Following the query by committee procedure outlined in Section II, we train $K$ neural networks on labeled data obtained from the existing AERONET sites. Then we measure the average uncertainty for each candidate site using historical unlabeled data from that site. The selected sites are those with the highest estimated site uncertainty.

TABLE I. SITE SELECTION BASED ON SPATIAL DIVERSITY

```
$L = \{L_i, i = 1 \ldots N_l\}$ := Set of existing sites.
$S = \{S_j, j = 1 \ldots N_s\}$ := Set of possible new sites
$N$ := Number of new sites to be chosen.

repeat $N$ times
    for each $S_i$ in $S$
        $d(S_i) = dist(S_i, L)$;
    end
    $i^* = arg\ max_i\ d(S_i)$;
    $L = L + S_{i^*}$;
    $S = S - S_{i^*}$;
end
```

TABLE II. UNCERTAINTY BASED SITE SELECTION

```
Given $L, S, N$ (defined in Table I)

1. for each $S_j$ in $S$
     Compute average uncertainty $u(S_i)$ of site $S_i$;
   end
2. Pick $N$ sites with the largest average
   uncertainty from $S$ and add them to $L$;
```

### C. Site selection based on temporal correlation

One drawback of the site selection method described in III.B is that a global measure such as average uncertainty might fail to account for the similarity in temporal variation of the uncertainty among sites. Each of the candidate sites can be regarded as a time-series of uncertainty values over a year (there can be at most one valid observation per day). In our approach, each of these daily time-series is converted into monthly time-series by averaging the uncertainties over each month in order to resolve the missing values problem. The underlying assumption is that sites which have similar temporal uncertainty patterns are redundant. The algorithm summarized in Table III is preventing selection of sites with similar temporal uncertainty patterns. We denoted $du(S_i)$ as *the temporal similarity* of candidate site $S_i$, and defined it as the minimum Euclidean distance between its monthly uncertainties and those of the existing sites in $L$.

### D. Site selection based on of the composite uncertainty

Algorithms in Tables I, II, and III focus on a single measure for site selection. Each of them might favor different sites. For example, we observed that AOD patterns of East U.S. are similar to those of Europe while they differ significantly from West U.S. [8]. So, if the learner is trained on sites from East U.S., it will have low uncertainty when applied on sites from Europe, but relatively high uncertainty in West U.S. This, the uncertainty-based selection will favor sites from West U.S., whereas spatial selection will favor sites from Europe. Furthermore, $N$ sites having the highest uncertainties might be spatially and temporally correlated. In order to combine all of the above metrics together we define a new distance measure for the candidate sites as the weighted sum of spatial and temporal distance from the nearest existing site and uncertainty.

We define *the composite spatio-temporal uncertainty* $ust(S_i)$ of candidate site $S_i$ as

$$ust(\mathrm{S_i}) = \alpha \cdot d(S_i) + \beta \cdot u(S_i) + \gamma \cdot du(S_i),$$

where $\alpha \cdot d(S_i) + \beta \cdot u(S_i) + \gamma \cdot du(S_i)$ are defined in III.A, III.B, and III.C, respectively. To simplify interpretation, all three quantities are scaled between 0 and 1. The resulting active site selecting procedure is shown in Table IV. In the experiments, we used several combinations of $\alpha$, $\beta$, $\gamma$.

TABLE III. SELECTION BASED ON UNCERTAINTY AND TEMPORAL CORRELATION

```
Given $L, S, N$ (See Table I)

for each $S_j$ in $L$
    $um(L_i)$ := vector of monthly uncertainties;
end
for each $S_j$ in $S$
    $um(S_i)$ := vector of monthly uncertainties;
end
repeat $N$ times
    for each $S_i$ in $S$
        $du(S_i) = min_j\ dist(um(S_i), um(L_j))$;
    end
    $i^* = arg\ max_i\ du(S_i)$;
    $L = L + S_{i^*}$;
    $S = S - S_{i^*}$;
end
```

TABLE IV. SELECTION BASED ON THE COMPOSITE UNCERTAINTY

```
Given $L, S, N$ (See Table I)

repeat $N$ times
    for each $S_i$ in $S$
        calculate $ust(S_i)$;
    end
    $i^* = arg\ max_i\ ust(S_i)$;
    $L = L + S_{i^*}$;
    $S = S - S_{i^*}$;
end
```

## IV. Experimental Results

### A. Dataset

The experimental dataset was created by collocating spatially and temporally labels collected by 171 AERONET sites and attributes observed by MODIS instrument on the TERRA satellite. The data spans the entire world and cover period between January of 2005 and December of 2006. The resulting data set has 28,418 labeled 13-dimensional examples. The attributes used are MODIS solar and sensor zenith and azimuth angles, scattering angle of radiation, mean and standard deviation of MODIS radiances at 4 different wavelengths, and AERONET site elevations.

Training dataset was created by using data points from 2005 and test dataset was created from 2006. To maintain interpretability of the results, 70 examples were chosen from each training site and 50 examples were chosen from each test site. Sites which had less than the specified number of examples were removed. After the removal, there were 90 training sites and 70 test sites. Figure 1 shows locations of the training sites. Figure 2 shows AOD time series at 10 AERONET sites during 2005.

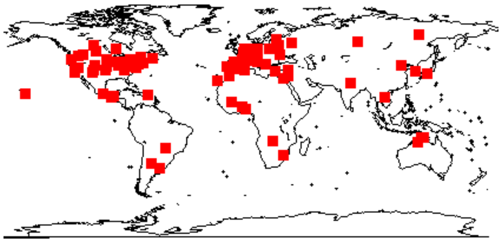

Figure 1. All training sites

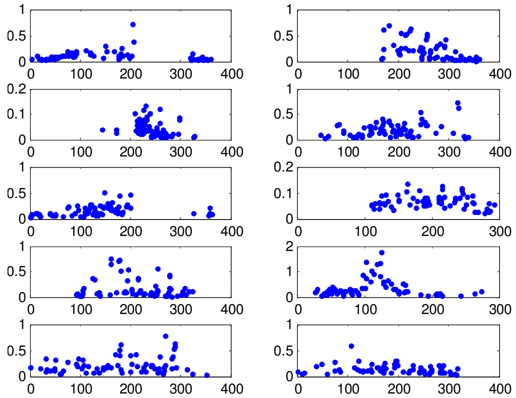

Figure 2. AOD distribution at 10 sites

### B. Results

In our experiments we used feed-forward neural networks to estimate AOD. We trained a committee of 20 neural networks, each having 10 hidden nodes. AOD is estimated as the average of the committee predictions. Initial training set was created with 700 examples collected during 2005 from 10 randomly selected AERONET sites. We treated the remaining AERONET sites as the candidate sites. We evaluated 5 different site selection algorithms. For each algorithm, we measured the $R^2$ accuracy of the committee after adding $t$ candidate sites to the training set. For each $t = \{1, 2 \ldots 20\}$ we ran a separate experiment.

The entire set of experiments was repeated 10 times, each time starting with a different set of initial sites. We report average accuracies obtained from those experiments. From Figure 3 it can be seen that uncertainty-based site selection gives significantly higher accuracy than random site selection or selection based on spatial diversity (Table I). The difference is particularly large when only a few candidate sites are selected ($t < 5$). The average uncertainty (Table II) and the temporal uncertainty (Table III) result in similar accuracies. However, there are some interesting differences, as will be illustrated in IV.C and IV.D. The composite uncertainty (Table IV) with α=1, β=1, γ=1, is the best overall, although the difference is not large.

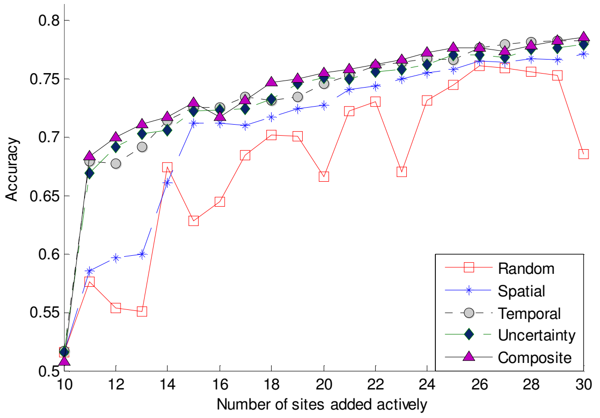

Figure 3. Comparisons of accuracies obtained from different types of site selection method

### C. Selected Sites

In Figures 4, 5, 6 and 8, actively selected sites based on different selection criteria are plotted to better illustrate the differences among them. We illustrate the outcome of a single experiment (because each of the 10 different experiments started from a different set of initial sites). The locations of the initial 10 sites are shown as blue dots in the figures, and their AOD time series are shown in Figure 2. As expected, sites selected based on spatial distance are quite sparse and cover almost the entire world. None of the sites in Europe and North America were selected. Unlike spatial selection, uncertainty-based selection picks almost all sites from Europe and North America. It is worth noting that some of the sites selected based on uncertainty are quite close to each other spatially. Additionally, average uncertainty and temporal uncertainty selection algorithms end up selecting similar set of sites. Mostly, this is the result of the fact that they use the same underlying information about prediction uncertainty. The only difference in the selected sites is site ‘Mongu’ located in southern Africa (see Figure 6). The AOD distribution of ‘Mongu’ shown in Figure 7 reveals why it was not selected by the temporal uncertainty criterion. Its AOD time series is very similar with two of the initial sites (1st and 3rd sites in the first column of Figure 2).

From Figure 8, it can be seen that the sites selected by combining uncertainty and spatial distance are different from those selected by both uncertainty and spatial correlation only. The selection is still spatially diverse, but the algorithm selected few sites from North America due to their high uncertainty. The two selected sites are over desert areas that are known to be very challenging surfaces for AOD estimation.

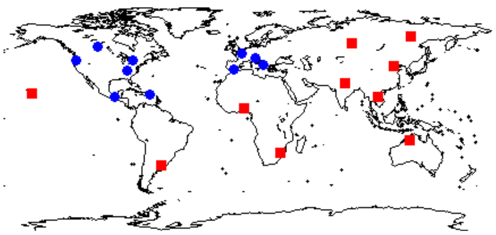

Figure 4. Sites selected based on spatial diversity. (Circle = initial sites; Square = chosen actively)

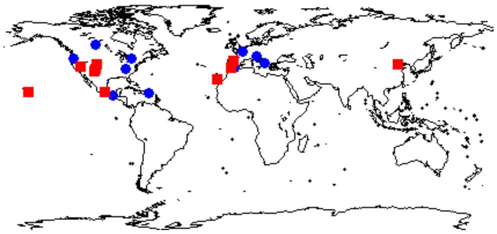

Figure 5. Sites selected based on temporal uncertainty

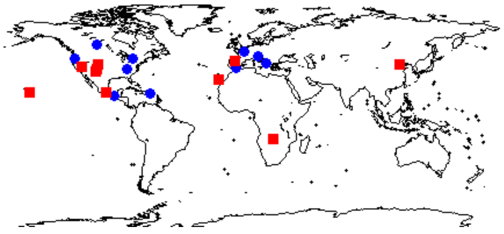

Figure 6. Sites selected using uncertainty

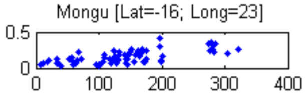

Figure 7. AOD time series at the ‘Mongu’ site in Southern Africa

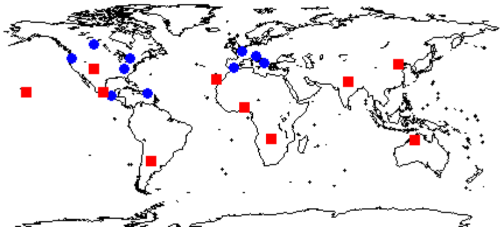

Figure 8. Sites selected based on composite measure

### D. *Continent-Wise Accuracies*

In all the results described in the previous figures the reported $R^2$-accuracy is average over all test locations. To get a better insight, the test data was divided into different continents and accuracy was measured on each continent separately.

The results are shown in Figure 9. It can be seen that different continents favor different selection algorithms. For Europe, random selection works quite well because a large fraction of candidate sites are from Europe. A predictor which is trained on a large number of European sites is expected to perform well over Europe. In North America, however, this logic does not work because despite a large number of candidate sites, random selection does not work too well. A possible explanation is because North America has a large mountainous desert-like region in West that is very difficult for AOD estimation. While random selection can occasionally pick some sites from that region, the uncertainty-based selection is focusing on these sites and thus improves the overall North America accuracy.

In Asia (Figure 9.c), Africa (Figure 9.d), and South America (not shown) spatial selection does a better job than other methods. The reason is that AERONET sites are underrepresented over these continents. While random selection is biased to Europe and North America, and uncertainty selection is biased to bright desert surfaces, spatial selection is biased to underrepresented spatial regions and continents. This clearly results in improved accuracy over underrepresented areas.

## V. Conclusion and Future Work

The novelty of this project lies in the nature of the problem being addressed here. This paper reports the first systematic study on active selection of future AERONET sites. The nature of active selection problem addressed here is different from conventional active learning problems. It can be regarded as batch selection because selecting one site actually adds multiple training points. We proposed and evaluated several site selection strategies. All proposed methods worked better than a random site selection. Selection based on prediction uncertainty resulted in high accuracy gains. Selection based on spatial diversity was somewhat less successful with respect to improvements in overall accuracy.

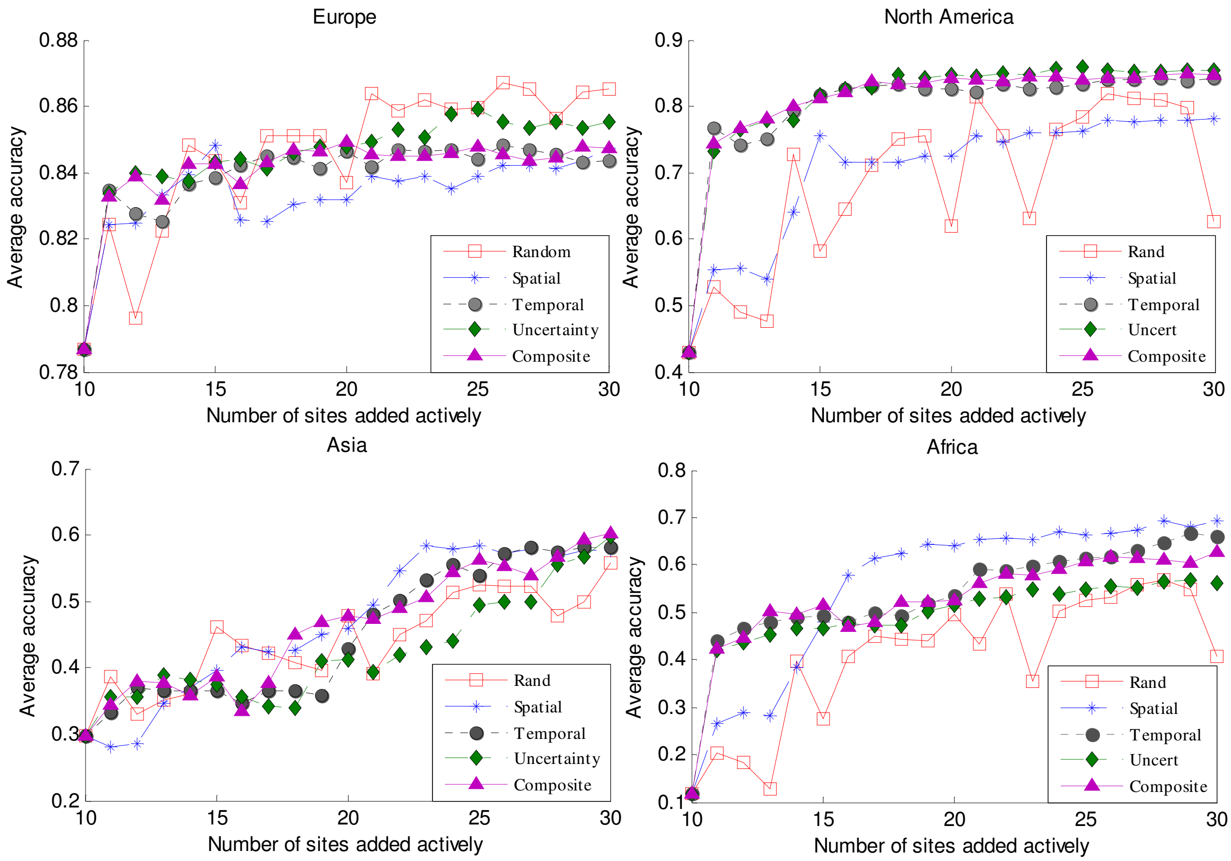

Figure 8. Continent-wise performance of different types of site selection algorithms

However, it can be very useful because it selects sites over underrepresented regions. A composite selection criterion that takes into account both spatial diversity and uncertainty seems to be a good compromise – it gives large overall accuracy gains that are particularly large over underrepresented regions.

We believe that this work can be valuable in future selections of AERONET sites, but also in many related remote sensing applications. As mentioned previously, in the present work we did not consider the effect of heteroscedastic noise variance in the site selection process. This will be a topic of our future research.

## ACKNOWLEDGMENT

This work is funded in part by NSF grant IIS-0612149.

## REFERENCES

[1] Bishop, C. M., Qazaz, C. S., “Regression with input-dependent noise: A Bayesian treatment”, NIPS, 1996.

[2] Guestrin, C., Krause, A., and Singh, A. P., “Near-optimal sensor placements in Gaussian processes”, ICML, 2005.

[3] Holben, B. N. et al., “AERONET: A federated instrument network and data archive for aerosol characterization”, Remote Sens. Environ., 1998.

[4] http://modis.gsfc.nasa.gov/, Official MODIS website.

[5] Krogh, A., Vedelsby, J., “Neural network ensembles, cross validation, and active learning”, NIPS, 1995.

[6] Papadopoulos, G., Edwards, P. J., Murray, A. F., “Confidence estimation methods for neural networks: A Practical Comparison”, IEEE Trans. on Neural Networks 2001.

[7] Radosavljevic, V., Vucetic, S., Obradovic, Z.,“Spatio-Temporal Partitioning for Improving Aerosol Prediction Accuracy”, SIAM Int’l Conf. on Data Mining 2008.

[8] Remer, L. A., Tanré, D., and Kaufman, Y. J., “Algorithm for remote sensing of tropospheric aerosol from MODIS for Collection 005”, 2006. http://modisatmos.gsfc.nasa.gov/atbd02.pdf

[9] Seung, H. S., Opper, M., Sompolinsky, H., “Query by committee”, COLT, 1992.