HARVIST: A System for Agricultural and Weather Studies Using Advanced Statistical Methods

Kiri L. Wagstaff, Dominic Mazzoni
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109–8099

Stephan R. Sain
University of Colorado, Denver
PO Box 173364
Denver, CO 80217–3364

***Abstract*— Remote sensing instruments in Earth orbit provide a rich source of information about current agricultural conditions. Observed over time, patterns emerge that can assist in the prediction of future conditions, such as the yield expected for a given crop at the end of the growing season. It is suspected that these predictions can be made more accurate by incorporating other sources of information, such as weather conditions from ground stations, soil properties, etc. The tools required to access and combine large amounts of data from multiple sources, at different spatial resolutions, are not readily available. The HARVIST (Hetereogeneous Agricultural Research Via Interactive, Scalable Technology) project seeks to address this lack by demonstrating the technology required to perform large-scale studies of the interactions between agriculture and climate. Previously, we have developed successful software tools for multispectral pixel classification using support vector machines, and multispectral image pixel clustering using constrained k-means, which we are leveraging in this effort. To date, we have developed a graphical interface that allows users to interactively run automatic classification and clustering algorithms on multispectral remote-sensing data. We have incorporated technical advances that exploit the spatial nature of the data to greatly increase classification efficiency. Our next goal is to incorporate a predictive component to support applications such as crop yield prediction.**

I. Introduction

Remote sensing instruments in Earth orbit provide a rich source of information about current agricultural conditions. Observed over time, patterns emerge that can assist in the prediction of future conditions, such as the yield expected for a given crop at the end of the growing season [1]. Accurate predictions can aid farmers in making decisions about which crops to plant and what farming techniques should be employed ("precision agriculture"). Previous work has focused on identifying linear correlations between indices such as NDVI (Normalized Difference Vegetation Index) and yield for corn [2], rice [3], or cotton and soybeans [4]. While these predictions are not as accurate as those obtained from direct measurements of crop health, they provide more spatial coverage and are significantly cheaper than *in situ* surveys [5].

However, there are two significant limitations to existing approaches to this problem. First, they have focused on generating yield predictions from a single data source, such as NDVI from remote sensing in the above cases, or temperature and precipitation data as in the Large Area Crop Inventory Experiment (LACIE) [6]. Scientists have identified the need for incorporating data from multiple sources simultaneously, such as remote sensing and weather data [3], but so far the tools necessary for a large-scale analysis of this nature have not been readily available. Second, these studies are also limited in scope; they tend to focus on specific regions and only incorporate tens of data points.

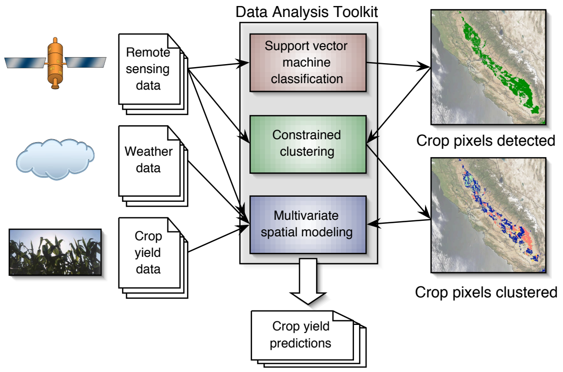
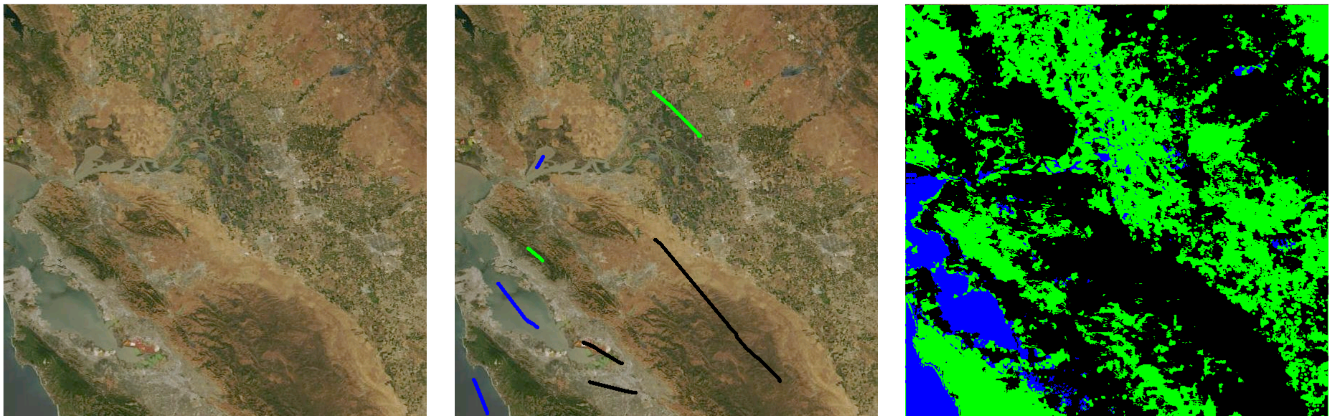


Fig. 1. The HARVIST System Architecture.

The HARVIST (Hetereogeneous Agricultural Research Via Interactive, Scalable Technology) project seeks to address both shortcomings by demonstrating the technology required to perform large-scale studies of the interactions between agriculture and climate. As shown in Figure 1, the HARVIST system will incorporate data from remote sensing instruments, weather ground stations, and historical crop yield databases to generate highly accurate predictions. Using classification, clustering, and prediction methods specifically optimized for spatial data, users can quickly and interactively obtain results over large areas. In addition to remote sensing and weather data, we also propose the use of additional data sources, such as soil properties and land cover databases, to further refine the predictive accuracy of the system.

Predicting crop yield is just one application of the technology in the HARVIST system. It will also be possible for scientists to conduct hypothetical "what-if" experiments to yield better understanding of the interactions between variables, such as temperature and crop yield.

The key innovations of this project are to (1) enhance the



(a) True-color image of central California, including the San Francisco Bay and central valley, acquired by MODIS/Terra, June 2, 2004, 19:00 GMT.

(b) Labels: green = vegetation; blue = water; black = land.

(c) Classification results obtained from training an SVM on the labels in part (b).

Fig. 2. MODIS/Terra data (courtesy Goddard Earth Sciences Data and Information Services Center), with training labels and SVM classification results.

scalability of data analysis methods (for very large, spatial data sets), (2) integrate **heterogeneous data** with different spatial and temporal characteristics, and (3) to provide an **interactive interface** that allows for easy hypothesis generation and testing. To date, we have developed an interactive, graphical interface that allows users to label, classify, and cluster remote sensing data. We have incorporated technical advances that exploit the spatial nature of the data to greatly increase classification efficiency. This paper describes the current system's capabilities and results. Our next goal is to incorporate a predictive component to support applications such as crop yield prediction.

II. HARVIST Analysis Methods

The HARVIST system now encompasses two data analysis methods: support vector machines and clustering. Both algorithms are able to take advantage of multispectral data from remote-sensing images, allowing them to find ways to discriminate between subtly different classes that are hard to distinguish using only red-green-blue (human visible) images. In addition, these methods can incorporate information from neighboring pixels and texture features to aid in distinguishing regions that are characterized more by shape or structure than color.

A. Pixel Classification using Support Vector Machines

Support vector machines (SVMs) are useful when the user has several specific classes of interest and can provide examples of each one [7], [8]. The goal is to build a classifier that learns, from the examples provided, to automatically classify new data in the same way. Figure 2(a) shows a sample data set, which is an image collected from Earth orbit by MODIS (the MODerate resolution Imaging Spectroradiometer). The pixel labels identified by a user are shown in in Figure 2(b), and after training an SVM on this small collection of labeled pixels, we obtain the classification results shown in Figure 2(c). Vegetation, land, and water are clearly distinguished and correspond to visually reasonable areas of the image.

B. Pixel Clustering using k-means

In contrast, clustering methods are useful when the classes of interest are not known, or the user wishes to identify overall trends present in the data set. Instead of providing labeled examples, the user indicates only how many clusters (groups of similar pixels) should be identified. This value, k, functions as a scale parameter, dictating how fine or coarse the inter-cluster resolution will be. We have incorporated the k-means clustering algorithm [9] into the HARVIST system. The results of clustering with $k = 3$ are shown in Figure 4(a). Here, the colors are not associated with any interpretation in terms of surface composition; they simply indicate distinct clusters. Eventually, we will also include more advanced methods for incorporating domain knowledge such as a bias towards spatially contiguous clusters [10] or "seeding" the cluster centers with surface types known to be present in the image [11]. We have demonstrated the ability to classify or cluster a given image with equal ease, by clicking the appropriate button in the graphical HARVIST interface.

C. Prediction: Multivariate Spatial Models

Our plan is to also incorporate predictive methods into the system, to provide the ability to predict crop yield given specific remote sensing, weather, and other observations. In particular, statistical models that incorporate spatial dependencies can provide more accurate predictions than those that assume that samples are independent [12]. The techniques that we will use can model non-linear relationships, predict values for multiple response variables simultaneously, and provide a straightforward method for estimating the uncertainty associated with each prediction [13].

TABLE I

Sample 2003 corn yield prediction results for 12 California counties. The best prediction for each county is in bold.

		Maximum	Avg. monthly	Error in bushels (rate)		
County	**NDVI**	**temp. (F)**	**precip. (in.)**	**NDVI**	**temp.+precip.**	**NDVI+temp.+precip.**
Butte	0.348	105.1	0.41	80.0 (44.4%)	18.7 (10.4%)	**11.6 (6.4%)**
Fresno	0.575	107.6	2.02	6.2 (3.6%)	**1.6 (0.9%)**	3.8 (2.2%)
Kern	0.557	106.0	2.39	13.7 (7.9%)	5.4 (3.1%)	**1.6 (0.9%)**
Kings	0.463	107.6	2.07	53.2 (28.6%)	16.4 (8.8%)	**2.7 (1.4%)**
Madera	0.584	107.1	1.88	**28.7 (20.6%)**	29.5 (21.2%)	34.8 (25.0%)
Merced	0.578	106.0	1.64	**32.1 (24.0%)**	32.2 (24.0%)	38.7 (28.9%)
Sacramento	0.719	105.1	1.05	63.6 (44.4%)	20.0 (14.0%)	**18.6 (13.0%)**
San Joaquin	0.641	106.0	1.36	22.0 (13.6%)	**3.2 (2.0%)**	6.5 (4.0%)
Solano	0.674	109.0	1.28	21.6 (12.6%)	2.4 (1.4%)	**0.5 (0.3%)**
Stanislaus	0.663	107.1	1.44	15.9 (9.1%)	7.3 (4.1%)	**5.5 (3.2%)**
Tulare	0.632	105.8	2.33	**0.6 (0.3%)**	13.1 (7.2%)	14.7 (8.1%)
Yuba	0.650	108.0	0.71	11.8 (8.2%)	8.4 (5.9%)	**1.1 (0.6%)**
Average				29.1 (17.5%)	13.2 (7.9%)	**11.7 (7.0%)**

D. A Preliminary Study

In a preliminary study, we explored the ability to combine support vector machine classification with crop yield prediction on a small-scale problem. First, we trained an SVM to automatically identify all of the cropland pixels in a larger MODIS image that covers California's central valley. After training on a random subset of 3000 labeled pixels, the SVM classified a disjoint random subset, also of size 3000, with 99.6% accuracy.

Next, we analyzed summary statistics for 12 California counties and used least-squares linear regression to predict corn yield. We calculated NDVI from the MODIS data, obtained weather data (maximum temperature and average monthly precipitation from May to October) from the NCDC, and obtained historical corn yield data from the USDA. We computed the regression over data from 2001 and 2002, then used the model coefficients to predict yield for 2003. If we only used the observed NDVI to predict yield, the average prediction error was 18%. If we used only weather data, we observed an error of 8%. However, when we combined data from both sources, the error dropped to 7%. Results for all twelve counties are shown in Table I. As expected, predictions that incorporate multiple data sources tend to result in increased accuracy. Despite the simplicity of this quick study, we achieved results comparable to the state of the art in crop yield prediction, e.g. 2-14% error in rice yield prediction [3].

These results support our claim that analyses combining input from multiple sources can achieve higher accuracy, motivating the need for a system such as HARVIST that can provide the integrated data interface. Eventually, we plan to use the full HARVIST system to generate crop yield predictions across the full United States.

III. Current Results

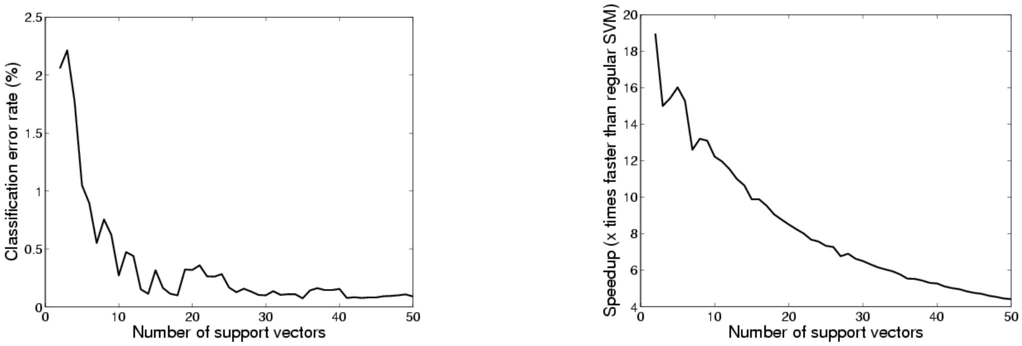
A. SVM Efficiency Improvements

When working with large data sets at the state, country, or even global level, efficiency is critical. We have incorporated two efficiency improvements into the SVM component of the HARVIST system: the Reduced Set method and the Nearest Support Vector method.

During the training phase, an SVM creates a classifier based on a carefully chosen subset of the training vectors (in this case, multispectral MODIS pixels). These vectors become the "support vectors". An image containing millions of pixels may result in thousands of support vectors; while this can provide very high classification accuracy, it comes at the expense of speed. Each new pixel to be classified must be compared to each of the support vectors.

Several approaches exist to improve SVM classification speed. These can be broadly grouped into two categories: those that obtain large speedups but require preprocessing, and those that obtain small speedups but require no preprocessing. We are exploring both approaches. In the first category is the method of Reduced Sets, which finds a smaller set of support vectors with the same relevant mathematical properties as the larger set. We have developed a new variation on this technique, which we call RS+, that achieves much greater speedups than previous published methods. Still, finding a good reduced set can take minutes or hours of computation. In the other category, we previously developed the Nearest Support Vector algorithm [14], which dynamically adapts the classification computation, based on the "difficulty" of each item to be classified, so that easy items can be quickly classified and computation time can be largely devoted to the more difficult items. So far, we observe only a 2x speedup in most real-world cases, but no preprocessing is required.

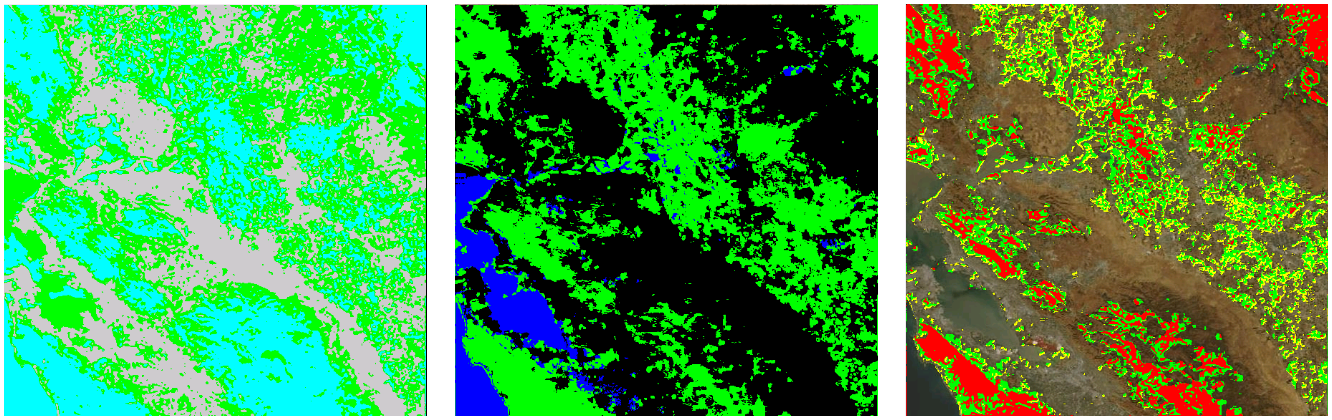
Our eventual goal is to develop a hybrid between these two methods, with virtually no increase in error while still achieving speedups of 10x. We have assessed this hybrid method experimentally, again on the task of recognizing crops in MODIS images. Figure 3 shows plots of the error rate (compared to the full SVM) and the speedups obtained by running a new variant of the Nearest Support Vector method aided by a "quick" reduced set that required very little time to compute. There is a clear tradeoff between efficiency and error rate. As the number of support vectors increases, the error decreases, as does the effective speedup.



(a) Classification error rate as a function of the size of the reduced set used.

(b) Speedup (efficiency) obtained as a function of the size of the reduced set.

Fig. 3. The tradeoff between error rate and speedup when using the Reduced Set SVM method.



(a) Clusters (green, cyan, and grey) identified when clustering with $k = 3$.

(b) SVM classification output; vegetation class is marked green. (Same as Fig. 2(c).)

(c) Three clusters (red, yellow, and green) identified within the vegetation class only.

Fig. 4. Clustering results on MODIS data. Each cluster's pixels are represented with a different color; colors themselves have no intrinsic meaning.

B. Integration of Clustering and Classification

One of our primary goals with the HARVIST project is not simply to provide multiple standalone analysis methods, but also to enable them to leverage each other's strengths by exchanging data and results. Therefore, we also added the ability to combine classification and clustering by first classifying an image, then identifying one of those classes as worthy of further exploratory analysis and applying clustering only to the pixels contained in the selected class. No manual intervention is required between these phases; the user simply clicks "classify" and then "cluster" to identify the sub-regions present in the class of interest. This process permits the user to focus the clustering algorithm's attention on specific classes, without needing to analyze the entire image at once. It is thereby possible to identify subtle distinctions within a class that would be swamped by the larger differences between classes when analyzing the entire image.

Figure 4(c) shows this scenario in action. Here, we have restricted clustering to the vegetation class only, as identified in Figure 4(b). As compared to Figure 4(a), we see that finer distinctions are identified, which may correspond to differences in land cover type, moisture in the soil, or other local conditions. A full interpretation of the clusters requires the examination of the cluster centers, which summarize the overall characteristics of the pixels assigned to each cluster. Displaying the cluster centers is one of the next capabilities we plan to provide.

IV. Data Fusion

We have also designed a multi-resolution image mosaic grid, which will allow us to incorporate remote sensing data at multiple spatial and temporal resolutions. Because we plan to incorporate data from multiple sources, with different resolution capabilities, it is essential that we be able to merge them in a principled way. In addition, we want to provide the ability to quickly browse the data at a low spatial resolution, identify regions of interest, and then apply analysis methods to the underlying data at high spatial resolution. We currently plan to approach this problem using a spatiotemporal grid as

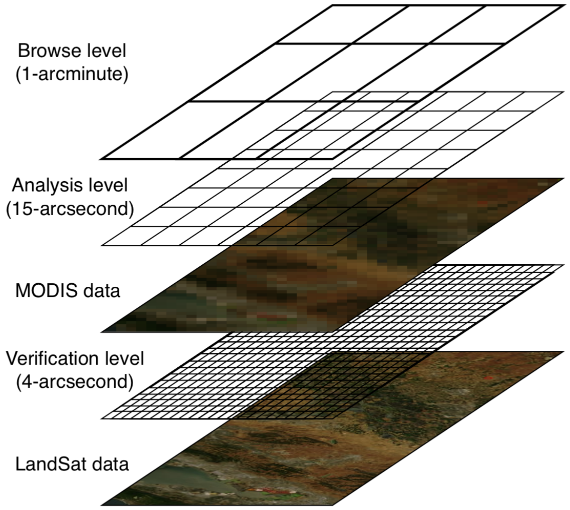


Fig. 5. Proposed multi-resolution image analysis grid.

shown in Figure 5. For clarity, we here show the multiple spatial resolutions, but there is also a time component; we aim to store and provide data at a one-month temporal resolution.

The multiple levels of resolution exist so that we can easily browse the data collection while still applying our analysis methods at the highest reasonable resolution provided by each instrument. As shown, we will provide browse capabilities at the lowest spatial resolution (1 arcminute or 1.8 km per pixel), which is sufficient for the identification of regions of interest, such as agricultural areas. For analysis purposes, we will work with MODIS data at 15 arcseconds or 1.1 km per pixel, which is sufficient for the identification of individual crop fields. We also have access to LandSat data for some regions at a very high spatial resolution (4 arcseconds or 120 m per pixel), permitting the identification of specific crop types. We prefer to use the MODIS data for our actual analysis, as it is freely available and provides better temporal coverage than LandSat can. This is particularly important for tracking the maturation of crops over the growing season. However, LandSat is useful for verification of our results, and it will aid us in training a crop type classifier to further specialize our methods based on the type of crop present in a given area.

V. Conclusions

In this paper, we have presented the HARVIST system, which provides advanced statistical analysis methods that can be applied to data from heterogeneous sources, such as remote sensing and weather data. We have described the current status of the system, which now includes both classification and clustering methods. A next step will be to integrate a predictive component to provide the ability to estimate numeric values associated with spatial locations, such as crop yield for counties across the United States. In a preliminary study, we showed that combining multiple data sources results in higher accuracy for these predictions.

A new contribution of this system is the ease with which users can integrate the results of different analyses. As one example, we showed how classification results could be used to restrict the input to a clustering method, to permit a focus on details only within that class.

Finally, we have presented our ideas for how to address the data fusion problem. We will combine data that has been recorded at different spatial and temporal resolutions by registering it onto a multi-resolution data analysis grid. We expect to refine this data hierarchy as we explore additional data sources.

Acknowledgments

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. The authors would like to thank Amy Braverman for her assistance with this collaboration and for recommending a crop modeling expert. This work was supported by a two-year grant from the NASA Earth-Sun Systems Technology Office, and the MODIS remote-sensing data was obtained from the Goddard Earth Sciences Data and Information Services Center.

References

[1] A. L. Hammond, "Crop forecasting from space: toward a global food watch," *Science*, vol. 188, pp. 434–436, 1975.

[2] J. F. Shanahan, J. S. Schepers, D. D. Francis, G. E. Varvel, W. W. Wilhelm, J. M. Tringe, M. R. Schlemmer, and D. J. Major, "Use of remote-sensing imagery to estimate corn grain yield," *Agronomy Journal*, vol. 93, pp. 583–589, 2001.

[3] N. K. Patel, N. Ravi, R. R. Navalgund, R. N. Dash, K. C. Das, and S. Patnaik, "Estimation of rice yield using IRS-1A digital data in coastal tract of Orissa," *International Journal of Remote Sensing*, vol. 12, no. 11, pp. 2259–2266, 1991.

[4] C. T. Leon, D. R. Shaw, M. S. Cox, M. J. Abshire, B. Ward, M. C. Wardlaw, and C. Watson, "Utility of remote sensing in predicting crop and soil characteristics," *Precision Agriculture*, vol. 4, no. 4, pp. 359–384, 2003.

[5] R. Allen, G. Hanuschak, and M. Craig, "Limited use of remotely sensed data for crop condition monitoring and crop yield forecasting in NASS," http://www.usda.gov/nass/nassinfo/remoteuse.htm, 2002.

[6] R. B. MacDonald and F. G. Hall, "Global crop forecasting," *Science*, vol. 208, pp. 670–679, May 1980.

[7] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, D. Gaussler, Ed., 1992, pp. 144–152.

[8] C. Cortes and V. Vapnik, "Support-vector network," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[9] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Symposium on Math, Statistics, and Probability*, vol. 1. Berkeley, CA: University of California Press, 1967, pp. 281–297.

[10] K. Wagstaff, "Intelligent clustering with instance-level constraints," Ph.D. dissertation, Cornell University, August 2002.

[11] K. L. Wagstaff, H. Shu, D. Mazzoni, and R. Castaño, "Semi-supervised data summarization: Using spectral libraries to improve hyperspectral clustering," *Interplanetary Network Progress Report*, 2005, in preparation.

[12] N. Cressie, *Statistics for Spatial Data*. New York: John Wiley, 1993.

[13] S. R. Sain and D. Nychka, "A multivariate spatial model for soil water profiles," *Journal of Agricultural, Biological, and Environmental Statistics*, 2004, submitted.

[14] D. DeCoste and D. Mazzoni, "Fast query-optimized kernel machine classification via incremental approximate nearest support vectors," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003, pp. 115–122.