Heterogeneous Agricultural Research Via Interactive, Scalable Technology (HARVIST)

ESTO Annual Report: October 1, 2004 – September 30, 2005

Submitted by Kiri L. Wagstaff, kiri.wagstaff@jpl.nasa.gov, 818-393-6393
Contributors: Dominic Mazzoni and Stephan R. Sain

Project Overview

This two-year project commenced on October 1, 2004. Our goal is to integrate multiple Earth Science data sources into a single graphical user interface that allows for the investigation of connections between different variables. In particular, we focus on relationships between weather and crop yield, but the system we are creating will be capable of integrating data for other studies as well. The data sources are heterogeneous in that they contain information at different spatial, spectral, and temporal resolutions. The HARVIST (Heterogeneous Agricultural Research Via Interactive, Scalable Technology) system will provide multiple machine learning and data analysis algorithms that can be applied to the data. Specifically, we aim to include support vector machines (SVMs; classification), clustering (discovery), and multivariate spatial modeling (regression and prediction) methods. In addition, we aim to greatly improve the efficiency of the component methods.

Summary of Accomplishments

Our accomplishments in the first year of the project reflect significant progress towards our goals, including a transition from TRL 4 to TRL 5. We achieved both of our Year 1 milestones on schedule, and we have completed several additional accomplishments. The following list summarizes our Year 1 major accomplishments:

1. **Milestone 1:** We have integrated SVM and clustering methods into the HARVIST graphical interface, enabling fast experimentation with either method. An important consequence of having both methods in the same toolkit is that they can also share results. For example, we have demonstrated the ability to cluster pixels belonging to a specific class as identified by an SVM. This allows exploratory analysis (clustering) that focuses on a class of interest to the user (such as vegetation).
2. **Milestone 2:** We have demonstrated the ability to apply our classification methods to continental-scale data sets by training and applying an SVM to all of North America at 275-m resolution (1.2 gigabytes of data). Demonstrated an 8x speedup using our previously developed Reduced Set SVM method. This milestone signals a system advance to TRL 5.
3. *Efficiency improvements.* We have incorporated several efficiency improvements to both clustering and classification methods into the HARVIST system. These advances make it possible to apply our analysis methods to continental and global-scale data. Overall speedups for each type of algorithm are ~10x, with little or no decrease in accuracy.
4. *Outlier detection.* Designed three outlier detection methods (spatial, offline spatio-temporal, and online spatio-temporal) to allow us to automatically detect and exclude, for example, cloudy pixels or data gaps.

5. *Crop shape modeling*. We have accomplished an initial modeling of crop growth profiles ("crop shape models") that allow us to model the impact of weather on crop growth and to separate out distinct growth profiles (such as different crop types).
6. *Field study*. In August, 2005, we conducted a field study to collect ground truth about crop types grown in central California, near Bakersfield, obtaining 384 labeled fields.
7. *Dissemination of results*. We published a paper at the 2005 Earth-Sun System Technology Conference describing the results of the first six months of this project, and we established a project website at http://harvist.jpl.nasa.gov/.
8. *Student involvement*. Worked with a Ph.D. student (Lucas Scharenbroich, University of California, Irvine) and two undergraduates (Kurt Cordle, University of Colorado, Denver, and Mike Kocurek, California Institute of Technology) over the summer. They each contributed significantly to the project accomplishments.

Current Progress Description

In the first year, our overall goal has been to prepare our data analysis toolkit to reach TRL 5, by demonstrating its operation in a realistic environment. In this case, that means demonstrating the ability to apply data analysis methods such as clustering and classification to remote sensing observations that cover the entire United States. To achieve that goal, we have also worked to obtain faster, more efficient versions of the data analysis techniques. These advances were obtained with the assistance of two undergraduate summer interns, Mike Kocurek of the California Institute of Technology and Kurt Cordle of the University of Colorado at Denver, and a graduate student, Lucas Scharenbroich of the University of California, Irvine. We also developed methods for automatically detecting outliers in the data and for producing crop shape models that will aid in identifying different crops in remote sensing data. Finally, we conducted a field study in August to collect ground truth about crops in central California, which allows us to validate analysis results obtained from orbital data. In this report, we describe each of these accomplishments in more detail. We conclude with a discussion of the project Technology Readiness Level (TRL) and our plans for the second year of the project.

*1. **Milestone 1:** Incorporation of SVM and clustering methods into the HARVIST graphical interface.*

The HARVIST system encompasses two data analysis methods: SVMs and clustering. SVMs are useful when the user has several specific classes of interest and can provide examples of each one. The goal is to build a classifier that learns, from the examples provided, to automatically classify new data in the same way. 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 items) should be identified. This value, k, functions as a scale parameter, dictating how fine or coarse the inter-cluster resolution will be. Both methods are now available inside the HARVIST toolkit.

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

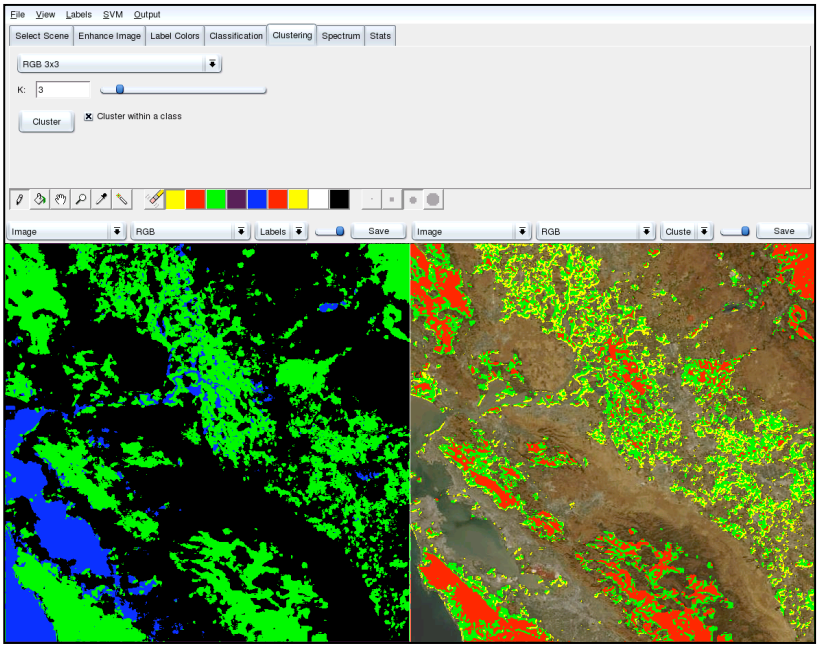

Figure 1. The HARVIST graphical interface. The left panel shows the output from a support vector machine trained to identify land (black), water (blue), and vegetation (green) pixels in California's Central Valley (MODIS image). The right panel shows the result of applying clustering to divide just the vegetation class into three further sub-clusters (colors indicate distinct clusters but do not have any assigned meaning).

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 1 shows this scenario in action. The left panel of the graphical interface shows the result of classifying pixels from a MODerate resolution Imaging Spectroradiometer (MODIS) image into three classes: land (black), water (blue), and vegetation (green). The right panel shows the result of applying clustering only to the vegetation class. 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.

2. ***Milestone 2:*** *Advance to TRL 5: Demonstration over the Continental United States.*

Our second major achievement for this year is the demonstration that our data analysis toolkit can successfully be applied to data sets on the continental scale. Specifically, we trained and applied an SVM classifier to a 1.2-gigabyte remote sensing data set and evaluated computational speed and classification accuracy.

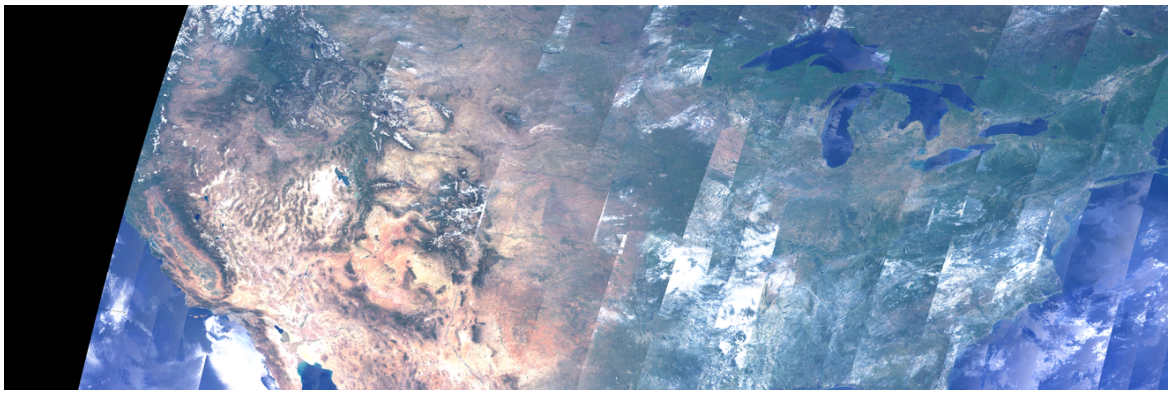

Figure 2. MISR mosaic of data from July, 2005, across the continental U.S.

The remote sensing data we used for this study was collected by the Multi-angle Imaging SpectroRadiometer (MISR). We assembled a mosaic that covers the continental United States, ranging from 70 to 130°W and 30 to 50°N, at 275 meters per pixel (see Figure 2). The resulting data set contains 232 million pixels, each represented by four features (red, green, blue, and near infrared values), for a total size of 1.2 gigabytes.

To conduct the demonstration, we labeled 100,000 pixels into one of five classes: water, cloud, land, cropland, or non-crop vegetation. We identified a random subset of 5,000 labeled pixels for training and a disjoint subset of ~47,000 pixels to assess the accuracy of the trained classifier. Each pixel was represented by the four feature values for all pixels in a 5x5 neighborhood; this helps capture useful contextual information. As a result, each item possessed 100 features.

The following table summarizes the results of this evaluation. Using a Reduced Set SVM, we were able to reduce the number of support vectors in the resulting classifier from 944 to 36 and to obtain an 8x speedup in classification time. The time required to classify pixels for the entire United States dropped from almost two days to just over five hours, with only a 2.8% decrease in accuracy.

	Preprocessing (find reduced set)	Number of support vectors	Classify 5x5 degree tile	**Classify entire U.S.**	**Speedup**	Accuracy
Regular SVM	N/A	944	59 minutes	47 hours	N/A	88.9%
Reduced Set SVM	13 minutes	34	7 minutes	5.5 hours	8x	86.1%

We have further analyzed the accuracy rates for the different classes in this data set through the use of *confusion matrices*. A confusion matrix shows the number of items belonging to each of the five classes (first column) that were assigned by the SVM classifier to each of the different classifications (first row). For example, the following table shows the confusion matrix for the regular SVM:

	Water	Vegetation	Land	Crop	Cloud
Water	**18,823**	16	0	232	0
Vegetation	13	**9,388**	233	3,335	9
Land	97	0	**9,065**	379	0
Crop	4	392	118	**2,477**	6
Cloud	308	0	0	0	**1,231**

Values in the diagonal entries indicate the number of pixels correctly classified into each class, and the sum of these values, divided by the total number of labeled pixels (46,126), yields the overall accuracy figure in the previous table. However, the confusion matrix also allows us to determine which classes are most difficult to distinguish. For example, a large number of "vegetation" pixels were incorrectly classified as "crop" (3,335 pixels). This is the most difficult distinction to make. In contrast, just 15 pixels were incorrectly assigned to the "cloud" class, because it is relatively easy to distinguish from the other classes.

The next table shows the confusion matrix for the Reduced Set SVM. Overall accuracy has decreased slightly, and we note that the number of "vegetation" pixels incorrectly classified as "crop" has risen (to 4,578 pixels). However, the number of "crop" pixels correctly classified as such has also risen (from 2,477 to 2,674). This indicates an increase in *recall* (number of crop pixels found) but a decrease in *precision* (because more pixels from the "vegetation" and "land" classes are mistakenly classified as "crop"). As expected, the Reduced Set SVM provided a large increase in speed (8x) while imposing a slight drop in accuracy.

	Water	Vegetation	Land	Crops	Cloud
Water	**18,409**	295	0	0	367
Vegetation	11	**8,335**	15	4,578	39
Land	9	59	**8,766**	486	221
Crops	0	160	104	**2,674**	59
Cloud	0	0	0	0	**1,539**

3. Efficiency Improvements

To enable the application of our data analysis methods to very large data sets, it is essential that they be highly efficient. In particular, we have developed several efficiency improvements that leverage the fact that remote sensing data exhibits a spatial bias towards spatial contiguity. This section discusses efficiency improvements we have developed in this project for k-means clustering and support vector machine (SVM) classifiers.

Clustering methods are untrained data analysis methods in that they do not require any labeled data from the user. Instead, the k-means algorithm divides the data set into k distinct, well separated groups, called clusters. Each cluster is represented by its mean, which is represented by feature values that are averaged across the items assigned to that cluster. Starting from a random "guess" at what the k means should be, the algorithm proceeds by iteratively refining the means to better fit the observed data until no reassignment of data points would improve the solution. In our work, a data set consists of the pixels in a given remote sensing (here, MISR) image.

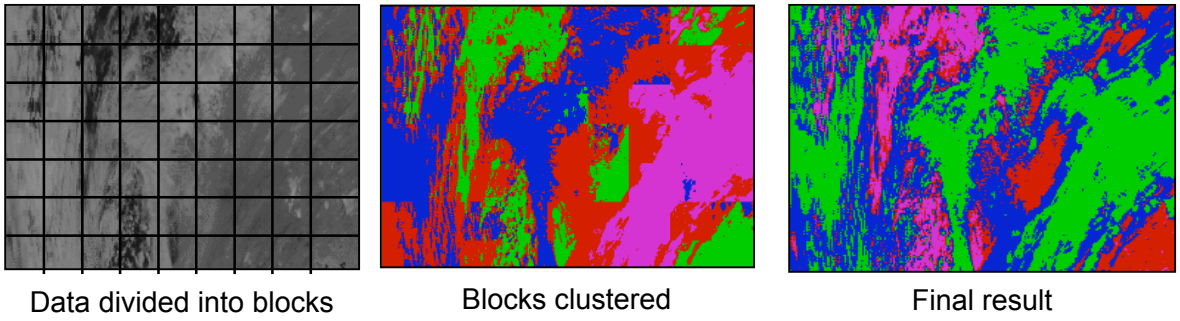

Figure 3. Block clustering results on a small subset of a MISR image of the northern California coast.

Efficiency Improvement 1: Random-Sample Seeded (RSS) K-means.
It is well known that the performance of the k-means algorithm is heavily dependent on the quality of the initial set of cluster means. Instead of selecting these randomly, as is commonly done, we selected a subset (~5%) of the data points and clustered those to obtain an approximate set of cluster means. Then, using these means as a starting point, we clustered the entire data set. We found that runtime dropped by 4-10x overall when clustering the MISR images described above. This is not a completely novel idea; variations of this method have been described in the academic literature before. Our contribution was to work out implementation details and investigate its performance on remote-sensing image data.

Efficiency Improvement 2: Block Clustering.
In addition to improving the starting point for the k-means algorithm, we also developed a method for leveraging the strong spatial contiguity of the remote sensing images. To do this, we divided each image into a set of smaller blocks, clustered them independently, and then merged all of the block clusters into a full set of *k* clusters in a final step (see Figure 3). This process yielded different speedups depending on the data set, number of blocks, and number of clusters used for the individual block clustering step. Similar approaches have been previously suggested in the literature, but again, they have not yet been evaluated on remote-sensing data. Further work is needed to better understand under what conditions this approach is most fruitful.

Finally, with the help of Lucas Scharenbroich, we incorporated three dimensionality reduction methods (simple projection, Principal Components Analysis, and Linear Discriminant Analysis) into the HARVIST graphical system. These methods further increase clustering speed by reducing the number of dimensions (features) that must be processed to just one dimension, and then applying a special scalar k-means method to generate the clusters. However, since information is discarded in the process, accuracy may be degraded. Depending on the application, this may be tolerable; for example, these approaches can provide an initial approximation of the clustering results without requiring an exhaustive analysis.

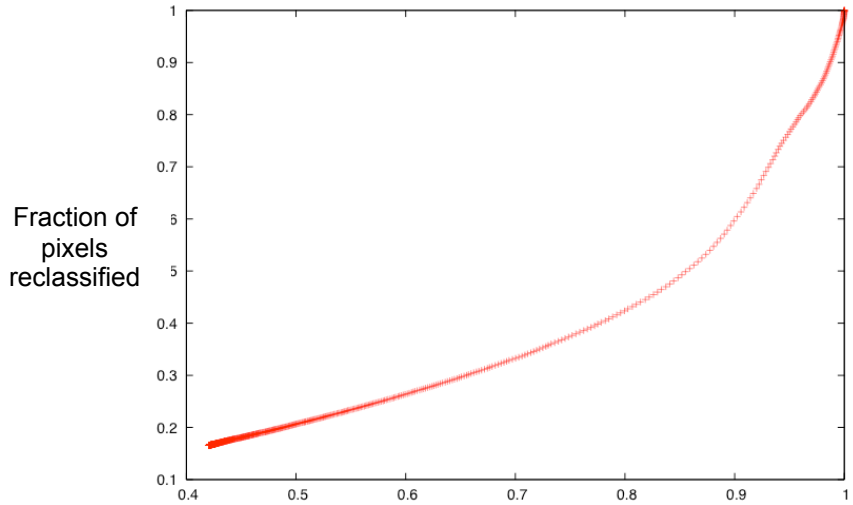

Figure 4. Plot showing the fraction of pixels that must be reclassified to correct a given fraction of the incorrectly classified pixels. The user can specify where on this sliding scale the algorithm should be, from the extreme on the left (fast but less accurate) to the right (slow but highly accurate).

A support vector machine (SVM) is a trained classifier that learns from labeled data provided by the user. After analyzing a set of labeled items, the SVM is constructed from a subset of the items that are identified as *support vectors*. These items are used to classify new data into one of the user-specified classes. The cost of classifying a new item (pixel) depends on the number of support vectors used by the SVM.

Previous work funded by other projects has resulted in two relevant methods for increasing the efficiency of SVM classification: the **Reduced Set** method, which identifies an approximate set of support vectors that is smaller than the "true" set of support vectors, and the **Nearest Support Vector** method, 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. In addition to conducting experiments with these methods, we developed an additional improvement that provides a user with the ability to obtain performance anywhere desired on the spectrum between "slow but accurate" and "fast but less accurate".

Efficiency Improvement 3: Selective Approximation SVM.
It is often the case that some items are easier than others to classify. If we could selectively apply a "fast but less accurate" SVM to the easy items, and a "slow but accurate" SVM to the harder ones, we could better allocate computational effort where it is most needed. In this SVM variant, we use two such SVMs and a confidence metric to achieve this goal. First, we use the quick SVM to classify the entire data set. Then, we analyze each of the classifications to determine where the quick SVM is likely to have been wrong. Those items are then reclassified using the slow (but more accurate) SVM. The hybrid result has a speed and accuracy somewhere between the two SVMs, depending on how many items were reclassified.

For this approach to be effective, the confidence metric that determines which items will be reclassified must be very reliable. In this case, we used a third (meta) SVM, which was trained on examples that were classified by the quick and slow SVMs and whether or not they agreed. The trained meta-SVM we used was about 80% accurate in making this determination.

In evaluating this approach on MISR data, we found that, for example, by reclassifying one third of the pixels in an image, we could correct approximately 65% of the errors made by the quick SVM (see Figure 4). This figure illustrates clearly that the user can achieve results anywhere on the spectrum from fast (less accurate) to slow (more accurate), based on how much computational time they are willing to devote to the reclassification process (or alternatively, how much error they are willing to tolerate).

4. Outlier Detection Methods

Outlier detection is an important component for any system that aims to analyze real-world data. For remote sensing data, cloudy pixels are a common data defect that, if not identified and eliminated, can significantly skew any subsequent analysis. Remote sensing, weather, and other data also suffer from occasional data gaps. These phenomena stand out as outliers due to how greatly they differ from "regular" observations. We are currently investigating several methods for detecting these outliers.

Our outlier detection scheme involves identifying a region of interest, fitting the parameters of a spatial or a spatio-temporal model based on data from the region, and then using a type of cross-validation procedure to compare observed pixel values with predictions based on the model. Observed pixel values that are inconsistent with the model predictions are flagged as potential outliers.

Predictions based on such models are actually linear combinations of the neighboring pixels (in space and/or time). The weights are determined by the nature of the spatial and/or temporal dependence, but essentially pixels closer in space and/or time are given more weight. Three different regimes for incorporating space and time are being investigated:

1. A purely spatial model where only spatial neighbors at a fixed time are used.
2. An "off-line" spatio-temporal model that uses spatially neighboring pixels from the past and the future. This model is likely to be the most effective, as it has access to the full temporal perspective, but it can only be applied to archival data.
3. An "on-line" spatio-temporal model that uses spatially neighboring pixels from the past only. This model does not have the advantage of seeing future pixels, but it can be used to identify outliers immediately, instead of waiting until data from additional time steps has been collected.

5. Crop Shape Modeling

The Normalized Difference Vegetation Index (NDVI) is a "greenness" index, derived from the red and infrared frequency bands, that measures the amount and condition of vegetation. NDVI has a temporal profile that is generally low in the winter months and higher during the growing

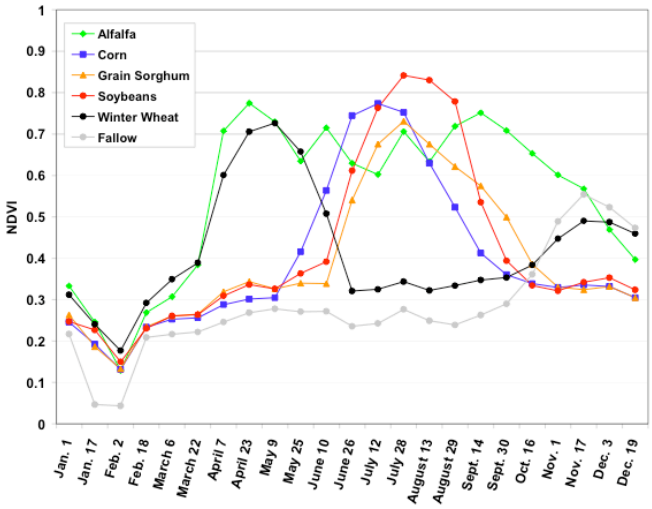

Figure 5. NDVI profiles for different crop types in Kansas, based on MODIS observations (Brian Wardlow, University of Kansas).

season. The peak height and duration as well as the temporal location of the peak differ across the different crops (see Figure 5), and these features can also be affected by weather, soil, farming practices, etc.

We are developing an advanced statistical model that links remote sensing and weather observations for each pixel. The model has several components. The first is a parameterization of a common NDVI shape profile. This represents an average NDVI profile that is a function of time and weather (temperature and precipitation). The variation across different crops is modeled through a mixture model. Each component of the mixture allows for a systematic deviation from the common NDVI profile. Finally, a spatial/temporal error process is included to account for smaller scale variation.

The results from a simplified example are displayed in Figures 6 and 7. All of the pixels within a 5-mile radius of the Jetmore 12 NNW weather station in Hodgeman County, Kansas (n=949) were selected. The common shape profile fit to these data is shown in the left frame of Figure 6. The basic structure of the NDVI profile is modeled as a function of temperature (black line) with adjustments due to precipitation (blue line). Three mixture components were also included in this example and the deviations from the common profile are shown in the right frame of Figure 6. One of the components is consistent with the common profile, but the other two show features of early greening and late greening, respectively. These components are likely to signal different crop types being grown as well as other varying factors, such as agricultural practices. We will further investigate the ability to split the observed NDVI into components, and to match these up with individual crop types.

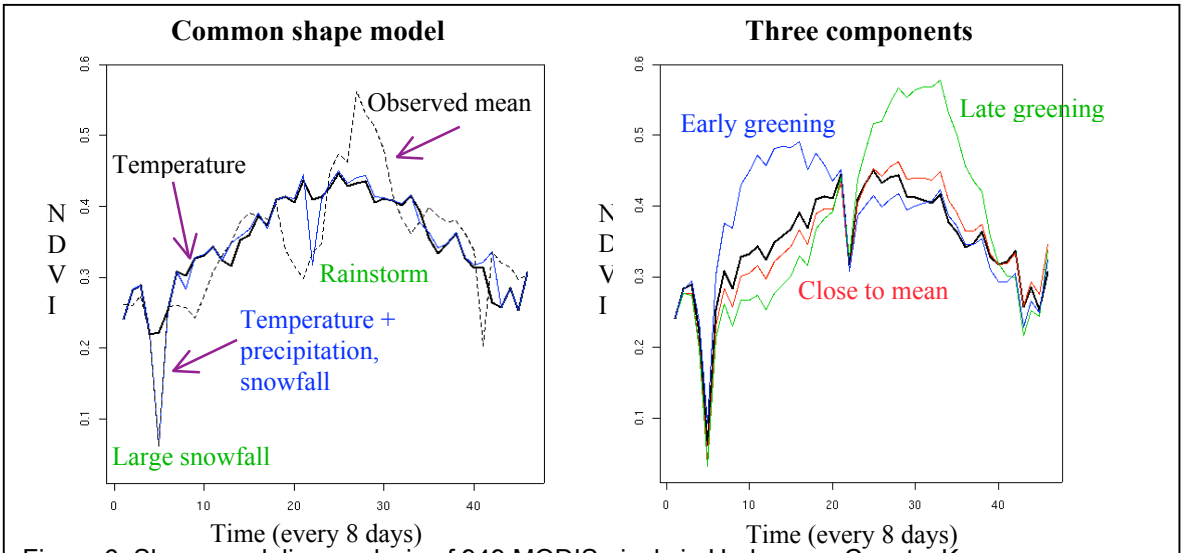

Figure 6. Shape modeling analysis of 949 MODIS pixels in Hodgeman County, Kansas.

Figure 7 shows the spatial component for one of the time slices. The deviations from the basic NDVI profiles show cohesive spatial regions where the NDVI is higher (lower) than the component profiles. These deviations are what remains after assigning pixels to the three components in Figure 6, so they are unlikely to represent different crop types or agricultural practices. Instead, since we are using weather information to generate the NDVI model, these are likely to be regions of local changes from the weather reported at the Jetmore 12 NNW weather station, which is located at the center of this area. Small increases or decreases in temperature and/or precipitation can impact crop growth and therefore NDVI.

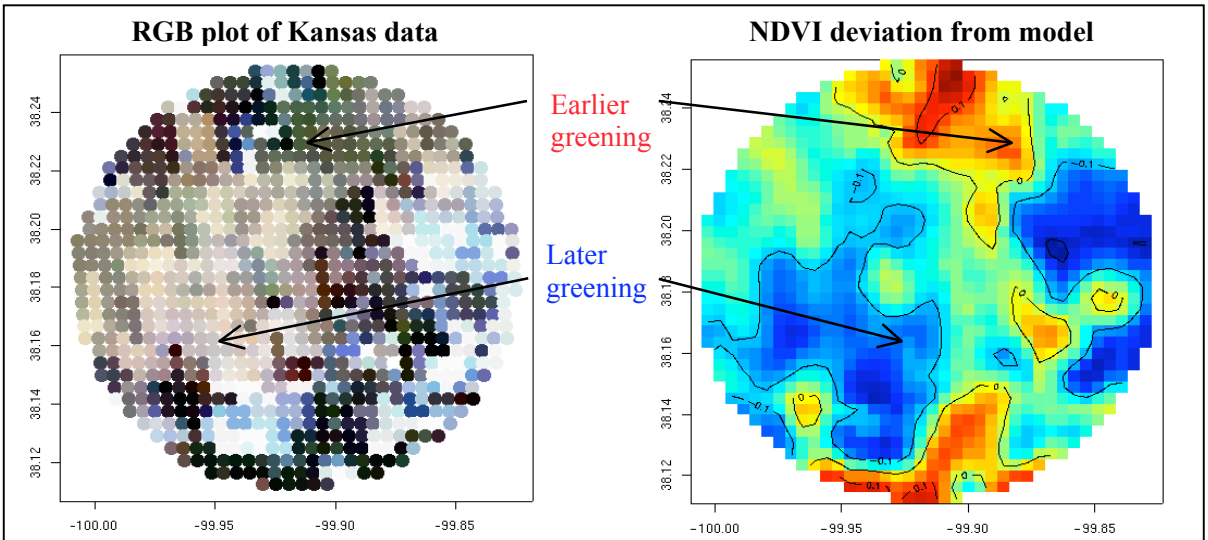

Figure 7. Spatially contiguous features identified in the Kansas data, where crops exhibit greening earlier or later than the model predictions. These deviations are likely to indicate local weather that differs from the regional observations.

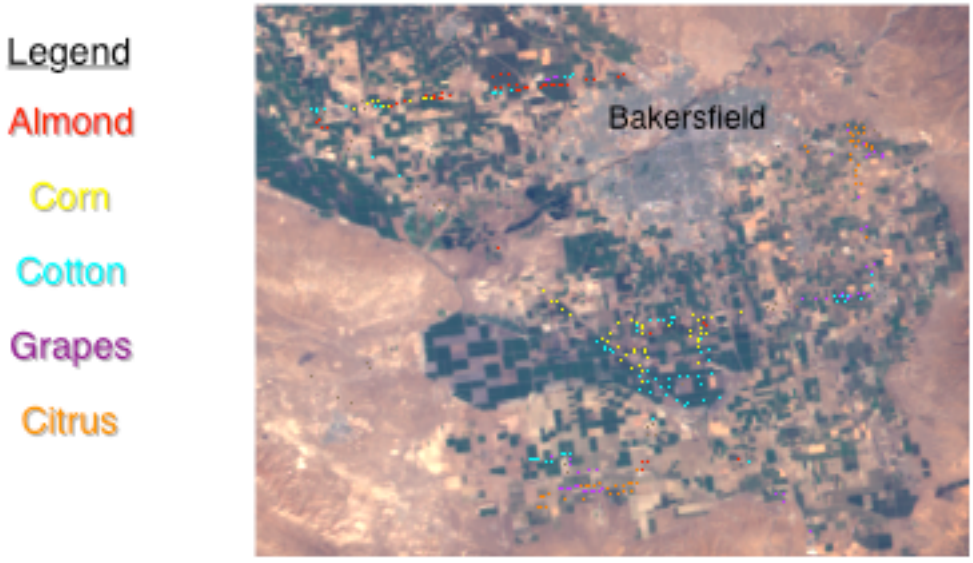

Figure 8. Data points collected for five crops near Bakersfield, CA.

6. Central California Field Study

On August 2, 2005, we conducted a field study in the Bakersfield area of central California. The goal of the field study was to collect information about crops currently being grown that we could match up with concurrent remote sensing data. We plan to use this data set to train and validate the crop type classifier we will be constructing during the first quarter of FY'06.

We surveyed 384 crop fields and, for each one, recorded the latitude, longitude, and crop type being grown. We also took digital pictures of the majority of the fields for later reference. Figure 8 shows a MISR image with fields identified for the top five crops (in terms of number of fields observed) overlaid on it. Crops in our data set with less representation include alfafa, almonds, watermelon, tomatoes, wheat, and red peppers. We also collected observations of uncultivated fields and urban areas. It was not possible to collect an evenly distributed sample of fields, due to fences and warnings against trespassing on private land. Therefore, our sample is biased in that we were restricted to fields that were near roads or highways. We were nevertheless able to collect a diverse sample of crops that we expect to be sufficient for our crop type classifier development.

7. Dissemination of Results

In June, 2005, we gave a presentation at the Earth-Sun System Technology Conference titled "HARVIST: A System for Agricultural and Weather Studies using Advanced Statistical Methods." An accompanying five-page paper with the same title was published as part of the conference proceedings. This paper is available from our project website, at http://harvist.jpl.nasa.gov/, which also includes additional information about project status.

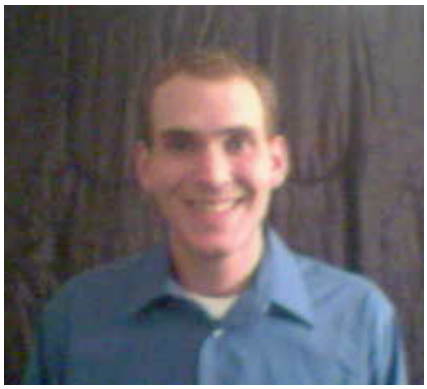

Figure 9. Summer students who contributed to the HARVIST project: Kurt Cordle, Mike Kocurek, and Lucas Scharenbroich.

8. Summer Student Projects

From June through the middle of August, we worked with three students on this project (see Figure 9).

- **Kurt Cordle** is now a senior at the University of Colorado, Denver. He implemented methods for outlier detection that allow us to automatically exclude missing or cloudy pixels in the remote sensing data, or data gaps in weather or other data sources. He also contributed to the analysis of the Kansas data used to derive crop shape models. Kurt will continue working on this project during the upcoming school year.
- **Mike Kocurek** is now a junior in Computer Science at the California Institute of Technology. Over the summer, he implemented and tested several efficiency advances for clustering and support vector machine methods that enable the application of these methods to very large data sets. He also assisted with our crop type data gathering effort in central California.
- **Lucas Scharenbroich** is a graduate student in Information and Computer Sciences at the University of California, Irvine. He implemented three dimensionality reduction methods and incorporated them into the HARVIST system. He also implemented an ENVI data format reader for HARVIST, which is significant due to this data format's wide popularity in a variety of science fields.

Schedule Status

The project is fully on schedule. We have achieved both of our Year 1 milestones as well as several additional accomplishments.

TRL Assessment

This project started out at TRL 4. The HARVIST system achieved TRL 5, as interpreted for software systems, after being tested on "realistic data" in its "final environment". In this case, we applied the trainable classifier algorithm (SVM) in HARVIST to remote sensing data covering the continental U.S. (232 million pixels) and demonstrated several of the efficiency improvements we have developed, for clustering and SVM algorithms.

Goals for Year 2

We have two milestones planned for Year 2:

1. **March 2006:** We will generate crop yield predictions for eight crops, across the United States. To achieve this milestone, we will first develop a crop type classifier that will allow us to specialize our crop yield models for different crops, which mature at different times over the growing season. We will evaluate this classifier using our ground-truth data from the central California area as well as crop-type classifications produced by prior analysis of time series MODIS data (provided by Brian Wardlow of the University of Kansas).
2. **September 2006:** We will demonstrate the ability to incorporate weather, land cover, and soil type data along with our current analysis of remote sensing data. We will conduct comparative experiments to determine which inputs are most useful in predicting crop yield.

In addition, we feel that it is important to evaluate the utility of the HARVIST toolkit from the perspective of agricultural scientists. We will identify one or more agricultural scientists for collaboration and solicit feedback from them on areas where the HARVIST system could be improved.

We hope that our advances will continue to yield benefits to scientists long after funding for this particular project has ended. We will achieve this goal both by working with scientists directly and by publicizing our results and advances. We will publish both our technological and scientific advances in appropriate venues. For example, the INTERFACE 2006 conference focuses on advances in the analysis of massive data sets; our innovations with clustering and SVM algorithms, as applied to gigabyte-scale remote sensing data sets, are very relevant. We also believe that the results of our crop yield prediction studies will be useful to the agricultural science community, and we plan to submit a paper describing our findings to a journal such as *Agronomy*. By publicizing the advances made possible by an analysis toolkit such as HARVIST, and our findings based on crop shape modeling, we will demonstrate the tangible science benefits of this technology.