# Heterogeneous Agricultural Research Via Interactive, Scalable Technology (HARVIST)

**ESTO Final Report:** October 1, 2004 – October 27, 2006

Submitted by Kiri L. Wagstaff, kiri.wagstaff@jpl.nasa.gov, 818-393-6393
Co-investigators: Dominic Mazzoni (Google, Inc.) and Stephan Sain (University of Colorado, Denver, and the National Center for Atmospheric Research)
Contributors: Kurt Cordle (University of Colorado), Michael Kocurek (California Institute of Technology), Lucas Scharenbroich (JPL, University of California, Irvine), and Tim Stough (JPL)

## Project Overview

This project commenced on October 1, 2004. Our goal was 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 focused on relationships between weather and crop yield, but the system we have created is 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 provides multiple machine learning and data analysis algorithms that can be applied to the data. Specifically, we include support vector machines (SVMs; classification), clustering (discovery), and multivariate spatial modeling (regression and prediction) methods. In addition, we have greatly improved the efficiency of the component methods.

## Summary of Accomplishments

1. **Milestone 1:** We 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 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). We achieved an 8x speedup using our previously developed Reduced Set SVM method. This milestone signaled a system advance to TRL 5.
3. **Milestone 3:** We generated crop yield predictions for two crops (corn and wheat) across the state of Kansas (102 counties). The lowest error in predicted yield that we obtained was 20% for corn and 18% for wheat. This milestone involved the use of an SVM to make the predictions. Since this milestone was achieved, we have developed two additional methods, which include a method for modeling spatial dependencies (MSM) and a spatiotemporal model (MSTM). We found that the SVM tended to have the lowest prediction error, while the other models produced smoother yield estimates. All methods are likely to benefit from additional data and observations.

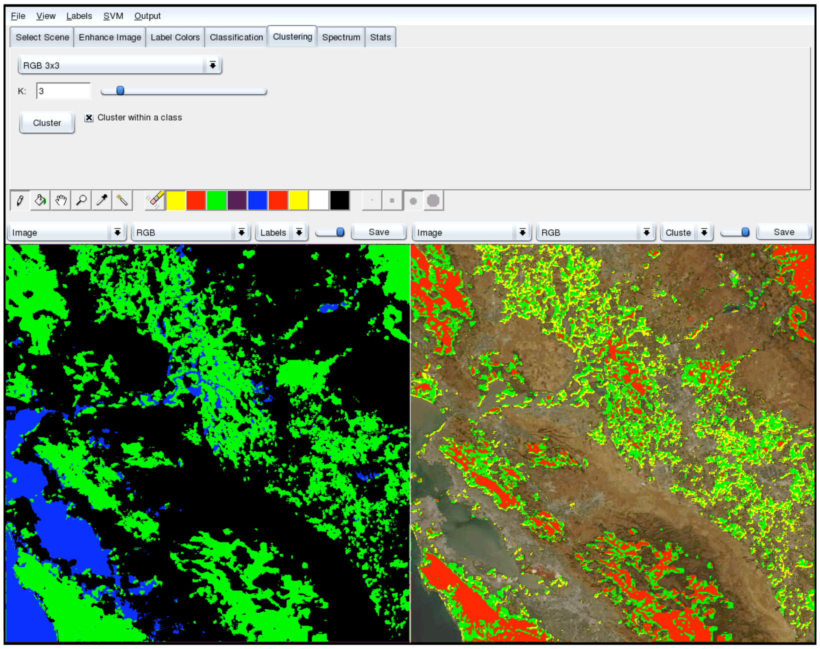
4. **Milestone 4:** We integrated weather data (temperature and precipitation) as an additional data source into our crop yield predictions. We found that the additional information sometimes improved our estimates and sometimes did not.
5. *Field study*. In August, 2005, we conducted a field study to collect ground truth about crop types grown in central California, near Bakersfield, obtaining 384 labeled fields.
6. *Progressive Rendering for SVMs*. We developed a new classification method that allows the user to directly control the tradeoff between computational time invested and accuracy obtained, for support vector machines.
7. *Dissemination of results*. We have published three papers at conferences, established a project website (http://harvist.jpl.nasa.gov/), and have an upcoming invited talk and poster presentation planned. In addition, we have developed a software library for use by the research community.
8. *Student involvement*. We have worked with two undergraduate students (Kurt Cordle, University of Colorado, Denver, and Mike Kocurek, California Institute of Technology) and a graduate student (Yongxia Kuang, University of Colorado, Denver). They each contributed significantly to the project accomplishments.

## Detailed Progress Description

### *1. **Milestone 1:** Integration of SVMs and clustering into the HARVIST graphical system*

One of our initial goals was to incorporate SVMs and clustering methods into the HARVIST graphical system. 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.

However, we aimed for more than just the ability to run one algorithm or the other on a given data set. We worked to enable the algorithms to leverage each other's strengths by exchanging data and results. Specifically, we 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.



RGB 3x3

K: 3

☒ Cluster within a class

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).

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 milestone was achieved via a 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. The remote sensing data 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.

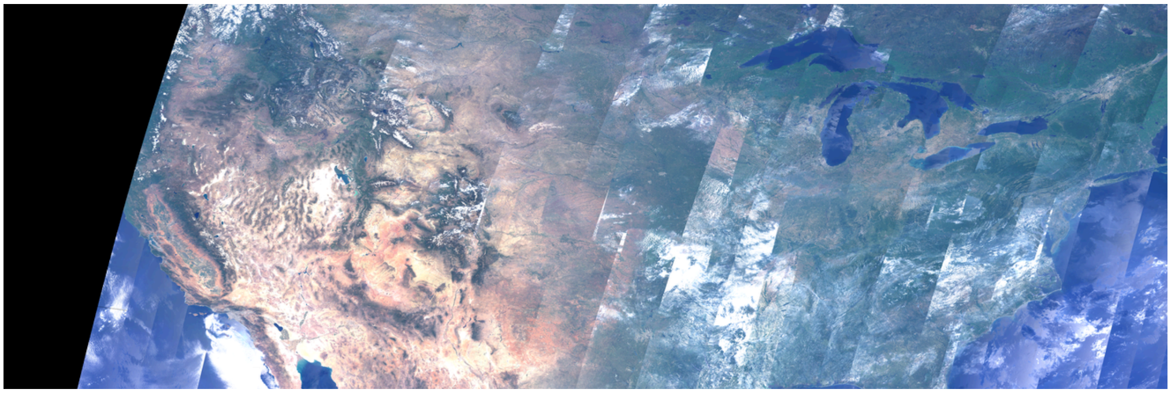


Figure 2. MISR mosaic of data from July, 2005, across the continental U.S.

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% |

## 3. ***Milestone 3:*** *Crop yield predictions across Kansas*

One of the major goals of this project has been to demonstrate the ability to provide a useful analysis product for the agricultural science community. Our proof of concept is the generation of crop yield predictions, in terms of expected bushels per acre. We have evaluated our ability to provide predictions for the state of Kansas, at the county level, for two crops (corn and wheat, the two most prevalent crops in the state). We chose Kansas due to the large amount of cropland present, as a fraction of total state area.

We developed three different approaches to this problem, each of which have their own strengths. The first method is an SVM, which is completely data-driven (no external models are used). The second method, MSM, incorporates spatial dependencies via a Hierarchical Bayesian model. The third method, MSTM, incorporates both spatial and temporal dependencies via a Markov random field.

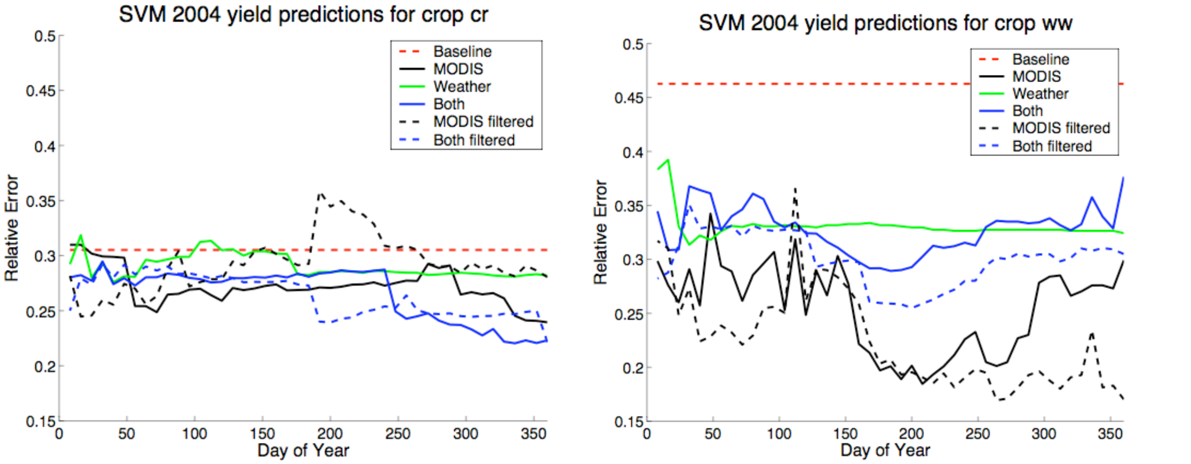


Figure 3. Relative error in corn (left) and wheat (right) yield predictions for 2004 obtained using a support vector machine trained on MODIS time series data.

**Support Vector Machine (SVM).** We were particularly interested in determining how early in the year we could make reliable yield predictions. We want to generate predictions before the harvest occurs, which is at a different time in the year for each crop. For example, corn is harvested in Kansas in mid-September, while the wheat we are studying is winter wheat and therefore harvested in mid-June. Therefore, we trained 46 different support vector machines (SVMs) using different subsets of the time series available to us. The first SVM had access only to the first observation, made on January 1 of each year. The final SVM had access to all 46 observations throughout the year, from January 1 to December 26.

Figure 3 shows the relative error in our corn and wheat yield predictions for 2004 for SVMs that were trained on observations in 2001 and 2002. All results are averages across all 102 counties. The baseline that we used for comparison (red dashed line) is the relative error obtained if we simply predicted that 2004 would have the same yield as was observed for 2003. We see that this baseline error is much higher for wheat than for corn. We also plot the error obtained when predicting yield based solely on MODIS data (black line) and when filtering the input data (black

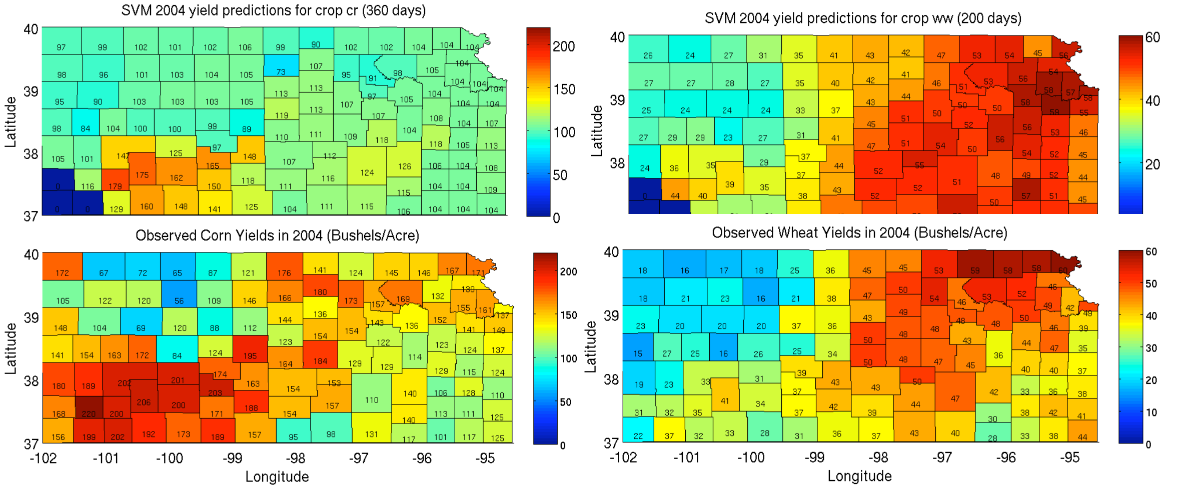
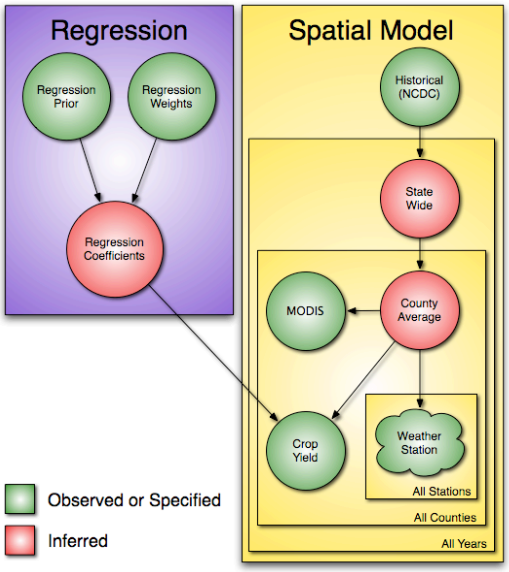

SVM 2004 yield predictions for crop cr (360 days)
SVM 2004 yield predictions for crop ww (200 days)
Observed Corn Yields in 2004 (Bushels/Acre)
Observed Wheat Yields in 2004 (Bushels/Acre)
Latitude
Longitude
37
38
39
40
-102
-101
-100
-99
-98
-97
-96
-95
0
50
100
150
200
0
20
40
60
0
10
20
30
40
50
60

Figure 4. SVM predictions (top) and true yield values (bottom) for Kansas in 2004. Corn yields are shown on the left, and wheat yields are shown on the right.

dashed line) to exclude observations with very low NDVI (Normalized Difference Vegetation Index) values.  Filtering is more helpful for wheat predictions than for corn.  The remaining lines will be discussed in the next section (Milestone 4).

Figure 4 shows example yield predictions produced by the SVM for 2004, for both corn and wheat.  While individual county estimates are not exactly correct, the SVM is able to identify the overall patterns present (e.g., corn is largely produced in southwestern Kansas, and wheat is largely produced in northeastern Kansas).  The three dark blue counties shown in the SVM predictions are counties for which we had no MODIS data available.  Hence, all averages are reported over 102, rather than 105, counties in Kansas.

**Multivariate Spatial Modeling (MSM).** The data that we are analyzing has a known spatial component, in that neighboring observations are likely to be correlated.  We developed a Hierarchical Bayesian model (see diagram at right) to explicitly model dependencies between counties (for yields and county-level weather averages) as well as within counties (for observations from weather stations).  The spatial model in the right part of this diagram shows how historical averages connect to state-wide averages, which influence county averages, which are connected to the actual observations: MODIS data, weather station data, and crop yield data.  The regression model on the left part of the diagram illustrates the model for crop yields.  The yields are predicted by a linear regression on the county-level data.  We fixed the correlations between neighboring counties to be proportional to the average distance between all pair-wise points in each county.  Correlation among each pair of weather stations was set to be proportional to their pairwise distance.  To obtain a yield prediction, we generated several thousand samples from the posterior distribution of the county yield and took the expected value as our prediction.



Each arrow in the model represents a *conditional distribution*.  All of the information necessary to evaluate a given node is provided by the node's parents, its children, and its children's parents.  As is typical in Bayesian models, we specify the parametric form of each conditional distribution according to our modeling choices and in order to make the problem tractable.

**Spatial Model.** The spatial component of the model is a Normal-Normal hierarchical model, which means that each level of the model (stations, counties and state) is modeled using a Normal (Gaussian) distribution.  This ensures that the model is *conjugate*, which implies that every posterior conditional distribution in the model is also a Normal distribution and the parameters of the distribution, the mean and covariance, have a closed form solution.  This analytic tractability is an important detail that decreases the computational cost of the model so that it can scale up to model all of the counties in a state simultaneously.

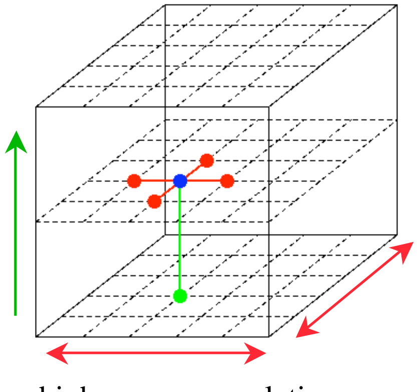
Another key modeling innovation in our model is that we impose a shared structure among the counties. Abstractly, the full dependence structure requires tracking a number of variables equal to the number of observations per county times the number of counties. In the case of Kansas, there are 105 counties with 184 features (46 temperature, 46 precipitation and 92 MODIS), which would require almost 1.5 gigabytes of memory to model. By imposing a blocked structure, we are able to reduce the amount of memory needed by several orders of magnitude.

In addition to reducing the amount of space needed for the MSM, our factorization also enables all of the necessary sampling operations to be evaluated efficiently, resulting in another order of magnitude savings in computational time.

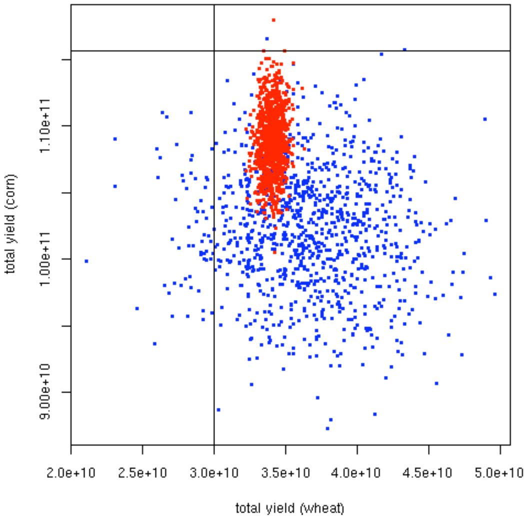
**Regression Model.** The regression model is set up as a simple linear regression model where the observed county yields are regressed against the estimated county-level aggregated feature vector. Notice that the regression coefficients lie outside of the nested boxes of the spatial model. This indicates that there is a single set of regression coefficients that are shared across all counties over all years. The motivation for this structure is that we expect the observed features to have a deterministic (up to uncertainty) relationship to the crop yield. Succinctly, if we observe identical data in two different years, we expect the crop yields to be identical (up to uncertainty).

When we apply this model to the observed data from 2001 and 2002 to predict 2004 yields, we find that predictions made by this model do not improve as more observations are available over the year. Prediction error remains relatively constant. This may be due to the linear regression technique used, which has more difficulty fitting data with higher dimensionality. However, a regular linear regression applied to the same data experiences a dramatic increase in error due to this dimensionality effect. The spatial priors may compensate for this effect.

**Multivariate SpatioTemporal Modeling (MSTM).** There is a third dimension that is relevant in our data. In addition to spatial dependencies, there also exist temporal dependencies. We developed a model that can encode these dependencies as well, connecting observations from this year with those observed at the same location last year (see diagram at right; time moves from bottom to top). These relationships are specified with a Markov Random Field. Given the spatiotemporal model, we again perform a linear regression and sample from the posterior to obtain a conditional mean estimate. This particular MSTM permit us to jointly estimate corn and wheat yield simultaneously.



We find that predictions made using this model also tend to have a higher *average* relative error. However, they also produce a wider spread in terms of possible predicted values. For example, in the figure at right (next page), total corn yield is shown on the y-axis and total wheat yield is shown on the x-axis (total yield values are the total estimated crop, in bushels, rather than the yield per acre). The solid lines show the true yields observed for each crop, in 2004. The red dots specify the yield predictions obtained when using a regular regression, with no spatial or temporal dependencies specified ("independent model"). The blue dots indicate predictions obtained from the spatiotemporal model. We observe that the independent model produces more tightly clustered predictions (lower variance), but that it rarely (for corn) and never (for wheat) gets close to the true yield values. In contrast, the MSTM predictions are more widely dispersed, but a small (for corn) and a significant (for wheat) fraction of them occur near the true yield values. This suggests that the MSTM, because it has access to more background knowledge, is less susceptible to overfitting the few years of data we have to train from. The MSM exhibits similar behavior.



Our general conclusion from these experiments is that the SVM can provide the best predictions when the amount of training data is limited. We would like to perform the same experiments when additional years of data are available. However, the MSM and MSTM offer a significant advantage in that they provide uncertainty estimates in their predictions, which is critical for any fielded use of these predictions.

## 4. ***Milestone 4:*** *Incorporation of weather data into crop yield predictions*

Our goal for this milestone was to integrate the weather observations with the MODIS remote sensing data. We collected weather observations from all stations in California and Kansas, from 2000-2005, from the National Climatic Data Center (NCDC). The data set contains daily observations of temperature, pressure, precipitation, etc., for 106 stations in California and 28 stations in Kansas. We need some information (an observation or estimate) of the weather values in each county, but weather stations are not distributed evenly across the state, and there are several counties that have no observations at all. Therefore, we used kriging to interpolate the observations across each state, using a Gaussian covariance prior. Figure 5 shows the kriged weather data for Kansas on day 178 of 2000. The smoothed weather field permits us to obtain weather estimates across the entire state for each day, with an associated uncertainty estimate, shown in the standard deviation plots. The uncertainties are largest for areas with few true observations, and more confident for areas with a higher density of observations.

The results obtained by the SVM, when incorporating the additional weather source (temperature and precipitation) are shown in Figure 3. We observe that using both weather and remote sensing data (blue line) tends to provide better estimates than when using weather data alone (green line). The fact that the green line is so flat over time suggests that those inputs may not be as relevant for yield prediction. We find that using both inputs can be helpful for corn yield predictions, but it is a detractor for wheat yield predictions. This may be a result of different farming practices; if wheat is primarily watered via irrigation, then precipitation is far less relevant to its growth (and eventual harvest).

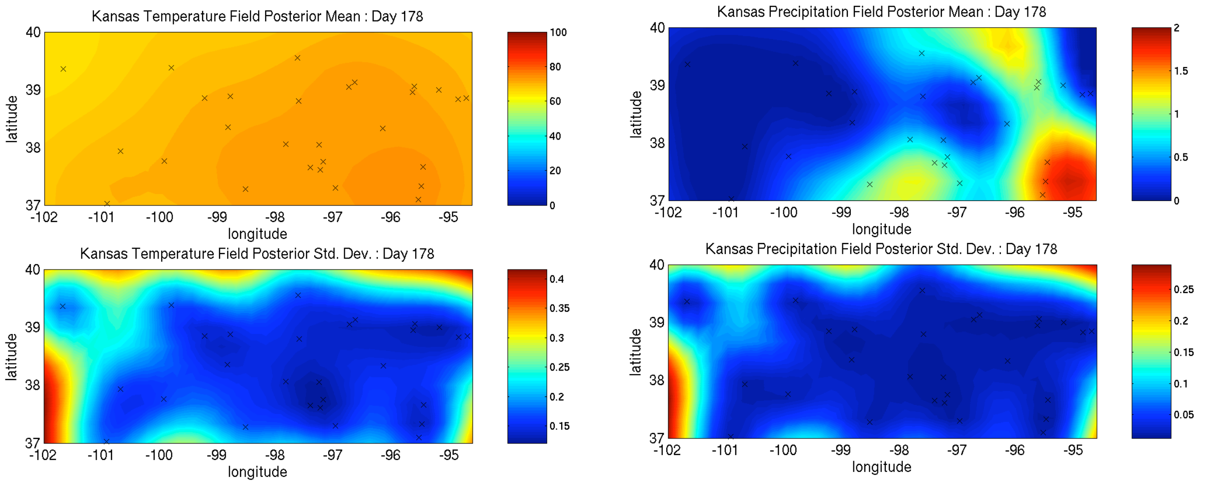

Kansas Temperature Field Posterior Mean : Day 178
Kansas Precipitation Field Posterior Mean : Day 178
Kansas Temperature Field Posterior Std. Dev. : Day 178
Kansas Precipitation Field Posterior Std. Dev. : Day 178
latitude
longitude
40
39
38
37
-102
-101
-100
-99
-98
-97
-96
-95
100
80
60
40
20
0
2
1.5
1
0.5
0
0.4
0.35
0.3
0.25
0.2
0.15
0.25
0.2
0.15
0.1
0.05

Figure 5. Kriged estimates of temperature and precipitation for the state of Kansas (day 178, 2000). Black x's mark actual observations at weather stations; the remaining values are interpolated based on a Gaussian prior. Lower plots show uncertainty associated with each value.

## *5. 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 used this data set to train and validate the crop type classifier that we constructed 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. 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 was sufficient for our crop type classifier development.

## *6. Progressive rendering for support vector machines*

Support vector machines (SVMs) have good accuracy and generalization properties, but they tend to be slow to classify new examples. In contrast to previous work that aims to reduce the time required to fully classify all examples, we developed a method that provides the best-possible classification given a specific amount of computational time. This approach is analogous to the progressive rending of GIF or JPEG images as they are downloaded over the Internet. During preprocessing, we construct two SVM classifiers: a "full" SVM that is optimized for high accuracy, yet may be slow, and a reduced-set SVM that provides extremely fast, but less accurate, classifications. We apply the reduced SVM to the data set, estimate the posterior probability that each classification is correct, and then use the full SVM to reclassify items in order of their likelihood of misclassification. Figure 6 shows an example of this hybrid approach applied to a scene observed by the Multi-angle Imaging Spectroradiometer (MISR).

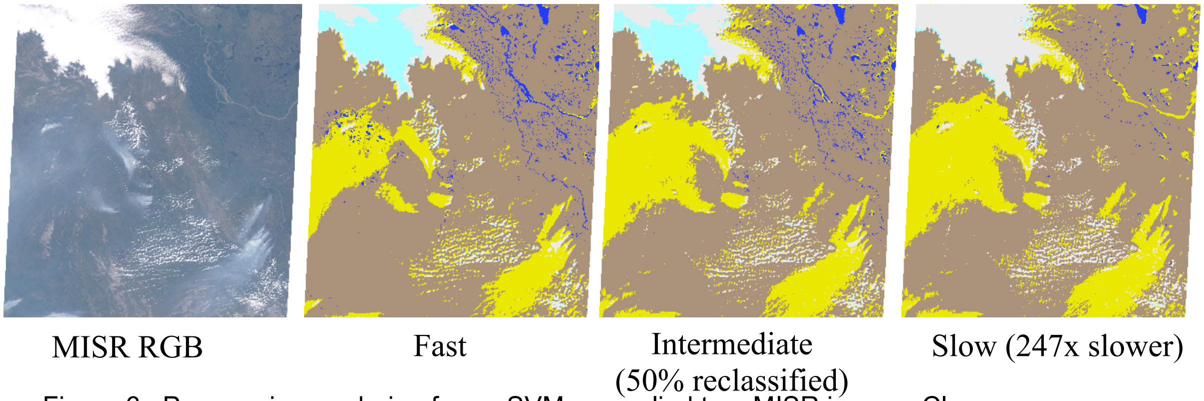


Figure 6. Progressive rendering for an SVM as applied to a MISR image. Classes are smoke (yellow), land (brown), water (blue), cloud (white), and snow/ice (cyan).

The first result contains some errors but is obtained almost instantaneously. The final result requires 247x more processing time.

A major benefit of this advance is that, in interactive applications such as our graphical interface to the HARVIST system, a user can specify what the relative importance of speed and accuracy should be. In an exploratory phase, users can indicate that a fast, coarse estimate of the final output is sufficient. When final, polished results are desired, users can indicate that the SVM should use as much time as is needed to fully classify the image. We have incorporated this new functionality into the graphical system.

## *7. Dissemination: Publications and Presentations*

Over the course of this project, we have worked actively to disseminate the results we have obtained over the course of this project. We published three papers at conferences, established a project website (http://harvist.jpl.nasa.gov/), and have an upcoming invited talk and poster presentation planned. An additional paper is currently under review. (See the end of this report for full paper citations.) We also have a journal paper in preparation, titled "Progressive Rendering for Support Vector Machines," which we will submit to the Data Mining and Knowledge Discovery (DMKD) journal. Finally, we have put together a software library that includes all of the multivariate spatiotemporal modeling methods that were developed under the auspices of this project. The software library is in R. It will be made available from the HARVIST website (http://harvist.jpl.nasa.gov/) and the central R community archive ().

## *8. Summer Student Projects*

Over the course of this project, we have worked with four students on this project (see Figure 7).

- **Kurt Cordle** is now a senior in Applied Mathematics 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.
- **Mike Kocurek** is now a senior in Computer Science at the California Institute of Technology. During the summer of 2005, he implemented and tested several efficiency

Figure 7. Summer students who contributed to the HARVIST project:
Mike Kocurek and Lucas Scharenbroich. Not pictured: Kurt Cordle, Yongxie Kuang.

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. He was named a Semi-Finalist in the Doris S. Perpall Speaking Competition on the basis of his talk describing his work with this project. In 2006, he developed the progressive rendering approach described above and integrated it into the HARVIST system. He also implemented an optimal hyperparameter search for the SVM crop yield estimation experiments.

- **Yongxie Kuang** is a graduate student in Applied Mathematics at the University of Colorado, Denver. She is working on her Master's degree. She has investigated spatiotemporal modeling methods for interpolating spatial fields, such as the missing values that arise in remote sensing data when cloudy or snowy pixels are detected and removed.
- **Lucas Scharenbroich** is a graduate student in Information and Computer Sciences at the University of California, Irvine. He implemented the dimensionality reduction methods described above 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 HARVIST project has completed all milestones on or ahead of schedule.

## 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" (September 30, 2005). 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. The project achieved an advance to TRL 6 in October, 2006, when we successfully demonstrated the ability to incorporate multiple data sources (remote sensing and weather station data) into a single prediction effort.

**To achieve TRL 7:** We view TRL 7 as a deployment of the HARVIST system for personal use by scientists. We have an opportunity to do just that: one consequence of our meeting with three soil scientists from the USDA in December, 2006, has been the continued interest on their part in using HARVIST for an upcoming soil salinity study. Successfully infusing the technology demonstrated in this task by enabling them to use it for their own science goals would achieve TRL 7. We estimate that the additional effort required to do this would be approximately 6 months of effort and $75K to support labor.

Publications

1. "HARVIST: A System for Agricultural and Weather Studies using Advanced Statistical Methods," Kiri Wagstaff, Dominic Mazzoni, and Stephan Sain. Earth-Sun Systems Technology Conference (ESTC), June 2005.
2. "Recent HARVIST Results: Classifying Crops from Remote Sensing Data," Kiri Wagstaff and Dominic Mazzoni. NASA Data Mining Workshop, May 23, 2006.
3. "Active Learning with Irrelevant Examples," Dominic Mazzoni, Kiri L. Wagstaff, and Michael C. Burl. In *Proceedings of the Seventeenth European Conference on Machine Learning (ECML)*, p. 695-702, September, 2006.
4. "Fast, Interactive Analysis of Remote Sensing Data with the HARVIST System," Michael J. Kocurek, Kiri L. Wagstaff, Dominic Mazzoni, Stephan Sain, Lucas Scharenbroich, and Timothy M. Stough. To be presented at the Fall Meeting of the American Geophysical Union, December, 2006.
5. "Predicting Crop Yields using Multivariate Markov Random Fields on Remote Sensing Information," Stephan Sain. Invited talk at the American Statistical Association Joint Meeting, July, 2007.
6. "Univariate *k*-means Clustering for Exploratory Data Analysis," Lucas Scharenbroich. Submitted to the SIAM Data Mining Conference.