Modeling and Monitoring Crop Disease in Developing Countries

John A. Quinn
Department of Computer Science,
Makerere University, Uganda
jquinn@cit.mak.ac.ug

Kevin Leyton-Brown
Department of Computer Science,
University of British Columbia, Canada
kevinlb@cs.ubc.ca

Ernest Mwebaze
Department of Computer Science,
Makerere University, Uganda
emwebaze@cit.mak.ac.ug

Abstract

Information about the spread of crop disease is vital in developing countries, and as a result the governments of such countries devote scarce resources to gathering such data. Unfortunately, current surveys tend to be slow and expensive, and hence also tend to gather insufficient quantities of data. In this work we describe three general methods for improving the use of survey resources by performing data collection with mobile devices and by directing survey progress through the application of AI techniques. First, we describe a spatial disease density model based on Gaussian process ordinal regression, which offers a better representation of the disease level distribution, as compared to the statistical approaches typically applied. Second, we show how this model can be used to dynamically route survey teams to obtain the most valuable survey possible given a fixed budget. Third, we demonstrate that the diagnosis of plant disease can be automated using images taken by a camera phone, enabling data collection by survey workers with only basic training. We have applied our methods to the specific challenge of viral cassava disease monitoring in Uganda, for which we have implemented a real-time mobile survey system that will soon see practical use.

The economies of many developing countries are dominated by an agricultural sector in which small-scale and subsistence farmers are responsible for most production, utilizing relatively low levels of agricultural technology. As a result, disease among staple crops presents a serious risk, with the potential for devastating consequences. It is therefore critical to monitor the spread of crop disease, allowing targeted interventions and foreknowledge of famine risk. Currently, teams of trained agriculturalists are sent to visit areas of cultivation across the country and make assessments of crop health. A combination of factors conspire to make this process expensive, untimely and inadequate, including the scarcity of suitably trained staff, the logistical difficulty of transport, and the time required to coordinate paper reports.

Although computers remain a rarity in much of the developing world, the near-ubiquity of mobile telephony has brought low-cost and reliable wireless internet services to broad regions that still lack electricity, running water and paved roads. Among other benefits, the prevalence of mobile computing devices at last offers a feasible alternative to paper-based data gathering.

This paper describes three innovations. First, we show how accurate response surface models of disease incidence and prevalence can be built using limited data, collected according to existing survey techniques. One challenge arising from the need to cohere with existing practices is that the levels of disease severity across the spatial field must be expressed as ordinal values, requiring the use of ordinal regression techniques. Second, we show how these models can be used in an active learning framework, determining in real time where survey workers should gather their next samples. This approach has workers gather data non-uniformly in order to maximize the value of the information gathered, as measured by a utility function elicited from domain experts. Because workers follow fixed circuits, our active learning task is an online optimization problem: each field must either be surveyed immediately or passed by. Finally, we present computer vision techniques for using camera-enabled mobile devices to make disease diagnoses directly, allowing reliance on survey workers with lower levels of training, and hence reducing survey costs. Specifically, given expert-annotated images of single cassava leaves, we demonstrate classification based on color and shape information.

We have applied these ideas to the domain of viral cassava disease monitoring in Uganda. Cassava is the third largest source of carbohydrates for human consumption worldwide, providing more food calories per cultivated acre than any other staple crop. It is an extremely robust plant which tolerates drought and low quality soil. The foremost cause of yield loss for this crop is viral disease (Otim-Nape, Alicai, and Thresh 2005), a major factor keeping East African farmers trapped in poverty (The Economist, 2011). We have developed a mobile survey system (see screenshot in Figure 3) which is currently being field trialled in partnership with Uganda's National Crops Resources Research Institute (NACRRI), and expect this to be used in their upcoming crop survey. Source code and survey data are available at http://cropmonitoring.appspot.com.

Spatial density estimation

In a crop disease survey, each plant is assigned a disease level $y_i \in \{d_1, ..., d_D\}$, usually by visual inspection of the aerial parts of the plant. A two-class survey might be done, where d_1 and d_2 represent healthy and diseased plants respectively; though often for cassava, disease levels are assigned categories from d_1 (entirely healthy) to d_5 (very severe disease,

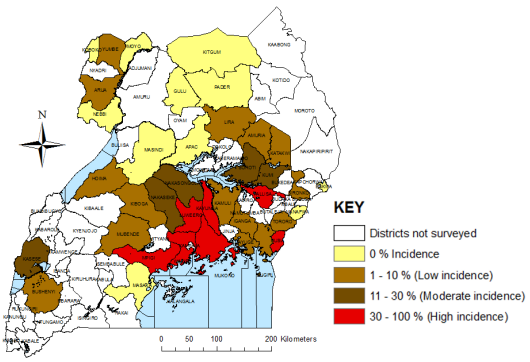

Figure 1: Incidence map of cassava brown streak disease in Uganda, 2009 (NACCRI).

causing dieback), where intermediate categories are characterised by the extent of disease on different parts of the plant. Note that without further information we should deal with these as ordinal categories, $d_1 < d_2 < \cdots < d_D$, and not code them as numbers: we do not have a principled way of mapping these categories to the real line, for example. Surveyors deal with single fields of plants at a time, and report two statistics, *severity* and *incidence* of disease. Incidence is the proportion of infected plants $P(y_i > 1|x)$, where $x \in \mathbb{R}^2$ is the location. Severity is the mean disease level of any non-healthy plants found in the field. Figure 1 shows an example of an incidence map for cassava brown streak disease in Uganda, produced by NACRRI to summarize the findings of its 2009 survey.

Models of crop disease are used for understanding the spread or severity of an epidemic, predicting the future spread of infection, and choosing disease management strategies. Common to all of these problems is the notion of spatial interpolation. Observations are made at a few sample sites, and from these we infer the distribution across the entire spatial field of interest. Standard approaches to this problem (reviewed in (van Maanen and Xu 2003)) include the use of differential equations, spatial autocorrelation, or kriging (Gaussian process regression) (Nelson, Orum, and Jaime-Garcia 1999). Under these schemes, summary statistics such as incidence or severity are interpolated. Note that the observations contain more information than these models use; under an incidence regression, we simplify the raw observation data to an aggregate across some area. Here we describe a way to adapt the geospatial statistical approach in order to directly model the quantity which is being surveyed: the distribution over disease levels, given observations of individual plants.

Given a number of point observations, we would like to predict disease characteristics across the map. The observed data is of the form $\mathcal{D} = \{x_i, y_i | i = 1, \ldots, N\}$ such that each individual observation consists of a location within a spatial region of interest $\mathcal{S}$, with $x_i \in \mathcal{S} \in \mathbb{R}^2$, and a disease level $y_i \in \{d_1, ..., d_D\}$. We take d_1 to represent an entirely healthy plant, and d_D to represent a plant with the highest category of disease damage. We would like to infer $P(y^*|x^*, \mathcal{D})$, the expected distribution over disease levels at a location x^* given a set of labelled observations $\mathcal{D}$. We use a Gaussian process (GP) approach for this ordinal regression, using the formulation of Chu and Ghahramani (2005). Ordinal regression is more appropriate than classification, because the latter treats all classes equally; in our case we know that d_1 is closer to d_2 than d_5, even though we lack a quantitative cost matrix. We did not use ordinary (i.e., real-valued) regression for two reasons. First, although we could represent the categories with real numbers $d_1 = 1, d_2 = 2$ etc., we have no basis for making an assumption about the relative similarities of different classes (e.g., it may be that d_1 and d_2 are extremely different, but d_2 and d_3 are relatively similar). Second, we have limits on the domain ($d_1 \leq y_i \leq d_D$), whereas most regression models cannot be restricted to an interval.

The GP approach to ordinal regression can be summarised as follows. A function $f(x_i) \in \mathbb{R}$ is introduced as an intermediate quantity to relate x_i and y_i. This function is taken to be a zero-mean Gaussian process with covariance $\mathrm{Cov}\left(f(x_i), f(x_j)\right)$ specified by a Gaussian kernel

$$K(x_i, x_j) = \exp -\frac{\kappa}{2}\left((x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2\right) \quad (1)$$

with parameter $\kappa > 0$. To relate $f(x_i)$ and y_i, we split the real line up into D intervals given the thresholds $b_0 = -\infty, b_D = \infty, b_{i|0<i<D} \in \mathbb{R}$, then formulate a likelihood $P(y_i|f(x_i))$, which is high when $b_{y_i-1} < f(x_i) < b_{y_i}$ and low otherwise. Our goal is then to integrate out f to obtain our estimate of y^* at a novel location x^*,

$$P(y^*|x^*, \mathcal{D}) = \int P\left(y^*|f\left(x^*\right), \hat{\theta}\right) P\left(f\left(x^*\right)|D, \hat{\theta}\right) df\left(x^*\right)$$

where $\hat{\theta}$ is a set of optimal parameters learnt from the data. Learning the parameters and calculating this distribution exactly is intractable in general, but following Chu and Ghahramani (2005), predictions can be obtained by applying the Laplace approximation (Rasmussen and Williams 2006, §3.4), yielding

$$\begin{aligned} P(y^*|x^*, \mathcal{D}) \approx \quad & \Theta\left(\tfrac{b_y - \mu}{\sqrt{\sigma^2 + \sigma_x^2}}\right) - \Theta\left(\tfrac{b_{y-1} - \mu}{\sqrt{\sigma^2 + \sigma_x^2}}\right), \quad (2) \\ \Theta(z) = \quad & \textstyle\int_{-\infty}^{z} (2\pi)^{\frac{1}{2}} \exp\left(-\frac{1}{2}x^2\right) dx, \\ \mu = \quad & k(x)^\top \Sigma^{-1} f_{\mathrm{MAP}}, \\ \sigma_x^2 = \quad & K(x,x) - k^\top \left(\Sigma + \Lambda_{\mathrm{MAP}}^{-1}\right)^{-1} k(x), \end{aligned}$$

where Σ is an $N \times N$ covariance matrix with its ijth element defined as in Eq. (1), $k(x)$ is a vector $[K(x, x_1), \ldots, K(x, x_D)]^\top$, the quantity f_{MAP} is the value of $f = \{f(x_i)|i = 1, \ldots, N\}$ which maximises $P(f|\mathcal{D})$, found with a suitable numerical optimisation method, and Λ_{MAP} is a diagonal matrix of second derivatives of the log likelihood, with the ith diagonal element given by $\frac{\partial(-\ln P(y_i|f_{\mathrm{MAP}}(x_i)))}{\partial^2 f_{\mathrm{MAP}}(x_i)}$.

The posterior distribution of disease levels over space can be updated whenever new observations are received. Figure 2 shows an example of inference in this model. Given observations of the disease levels of individual plants, we are able to infer distributions of disease levels at any location. Figure 3 illustrates the variance on these estimates; note low variance where observations have been made, as expected.

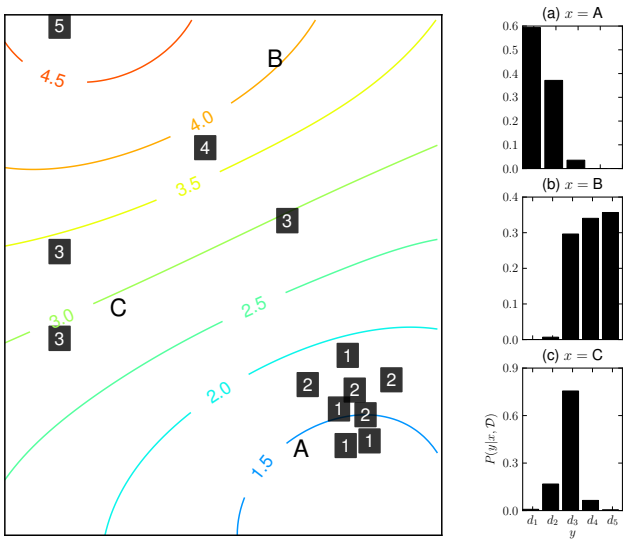

Figure 2: Example ordinal regression of disease levels across a spatial field given sample observations. The panel on the left shows example disease level observations (numbers with dark backgrounds), and contour lines indicate the mean predicted disease level. The three insets show the predicted distribution of disease levels at locations A, B and C respectively.

Adaptive selection of survey sites

Besides summarizing the findings of a survey, our models can be used to guide surveyors to collect more valuable data, holding fixed their budgeted number of samples. Based on our interviews with surveyors from NACRRI, we assume that survey teams follow a network of roads to conduct their surveys, and to take measurements from farms near the road. In rural parts of the developing world, the road network is extremely sparse. This makes it reasonable to assume that survey teams will follow a set route, corresponding to a one dimensional manifold $\mathcal{R}$ within the spatial field. NACRRI's current survey methodology is to sample uniformly along $\mathcal{R}$. With a survey budget allowing k stops, we are interested in finding a set of points along $\mathcal{R}$ that maximises the expected utility of the survey. (For now, we simply presume the existence of some utility function; we provide such a function for our cassava domain in the Experimental Evaluation section below.) Because we propose the collection of data with mobile devices, we can calculate in real time the best place for a survey team to go next, every time a new observation is made.

The simplest survey scheme would be for the surveyors to complete a single tour, returning to their headquarters afterwards. However, it can still be practical for a survey team to complete their tour more than once. In such a scenario, the first circuit can be used to explore disease levels across the region, and subsequent circuits can use the remaining budget to focus on the areas of greatest interest. We assume that each survey follows a tour that can be repeated t times ($t \geq 0$; in practice, we expect $t \leq 3$), that there is a limit of N samples which can be taken in total by any survey team,

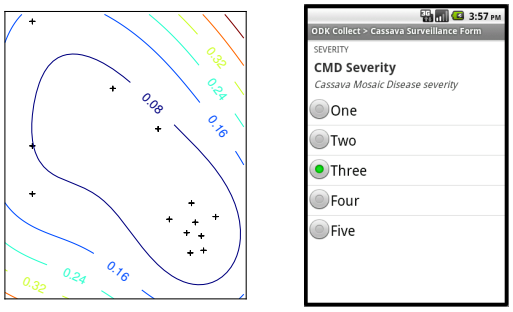

Figure 3: Variance of the predictions across the spatial field from Figure 2; crosses indicate the location of observations (left). Screenshot of our mobile data collection system (right). For each plant being surveyed, the system collects an expert diagnosis, leaf image, and GPS coordinates.

and that $p_i\%$ of the budget must be used on the i^{th} circuit.

Our goal is to use our model to determine the locations that a survey team should visit, in a way that outperforms the simple uniform sampling strategy currently used by NACRRI. Such an "active learning" problem is straightforward when there is no constraint on the order in which observations can be obtained. In our setting, however, we face the constraint that the tour can be repeated at most t times, and hence will often prefer to sample a nearby point before sampling a distant point, even if the latter is expected to be more informative. The intuition behind our algorithm is to leverage a traditional, unconstrained active learning framework to identify a distribution of k-tuples of points that are expected to yield high-value observations (where on the i^{th} tour, k is initialized to $\lfloor p_i N \rfloor$), and then to select the nearest point (formally, the first order statistic) from this distribution. Then we obtain an observation at that point, update the model, and repeat the process with $k \leftarrow k - 1$.

More formally, Algorithm 1 gives the details of our sampling strategy. At each point it greedily selects the point along the current route with the largest expected utility, and then repeatedly samples from the model at that point to "hallucinate" different potential observations that could be obtained. We update the model based on these hallucinated observations, and then look for the next most utility-increasing point. The function $ExpectedUtilityGain(r, u_y, P(y|x, \mathcal{D}^*))$ quantifies the amount by which the utility u_y of the current density model $P(y|x, \mathcal{D}*)$ would be expected to increase if observations were provided at location r. If the utility is neutral with respect to disease levels, this is the standard active learning setting where we choose the point with the highest uncertainty in the current model. If the utility is weighted according to the true disease level (see next section for an example), then we can weight the uncertainty to account for this, according to the current disease level estimates. This procedure results in a set of k points along the current route (represented as the distances along the route from the last stopping location). We can repeat this many times and take the mean of the first order statistic of these sets. We take this mean as our next survey location.

Algorithm 1: Survey route optimisation.

Input: $\mathcal{S} \in \mathbb{R}^2$, spatial field of interest.
$\mathcal{R}$, the remaining route, a 1D manifold in $\mathcal{S}$.
$\mathcal{D} = \{x^{(i)} \in \mathcal{S}, y^{(i)} | i = 1, \ldots, N\}$, previous samples.
k, the number of stops remaining.
Output: d_{next}, distance to travel along $\mathcal{R}$ before the next stop.
$S \longleftarrow \emptyset$
 # Sample many sets of stopping points
for $i = 1 : N_{\text{samples}}$ **do**
  $A \longleftarrow \emptyset$
  $D^* \longleftarrow \mathcal{D}$
  for $j = 1 : k$ **do**
    $x \longleftarrow \arg\max_{r \in \mathcal{R}}$ $ExpectedUtilGain(r, u_y, P(y|x, \mathcal{D}^*))$
    Append $DistanceAlongRoute(x, \mathcal{R})$ to A
    Sample observations from $P(y|x, \mathcal{D}^*)$
    Append sampled observations to $\mathcal{D}^*$
  end
  Append $A_{(1)}$ to S
end
$d_{\text{next}} = \frac{1}{N_{\text{samples}}} \sum_i S_i$

Figure 4: Examples of healthy leaves (left) and those infected with cassava mosaic disease (right).

Image-based diagnosis of crop disease

One reason that crop disease surveys in the developing world tend to be under-resourced is a scarcity of expert surveyors. Reliable automatic methods for performing surveys therefore offer the possibility of extending the scope of disease surveillance. The ubiquity of camera phones in even the most rural parts of many developing countries introduces the possibility of survey by non-expert workers, who submit images of crops that are then automatically classified.

We consider two types of features from these images, following the approach taken in previous work on visual cassava disease diagnosis (Aduwo, Mwebaze, and Quinn 2010). The first is normalised histograms of hue (within the yellow/green range) using 50 histogram bins. The second is local image gradient information, represented using Scale Invariant Feature Transform (SIFT) descriptors (Lowe 2004). Whereas SIFT features are often used in object detection, we used them here to create a texture model of each image. Specifically, we calculate SIFT descriptors across each image (each descriptor being a 128-dimensional vector), summarizing each with a mean vector μ_i and a covariance Σ_i.

To classify images, we use simple k-nearest-neighbor classification, using Kullback-Leibler divergence as a measure of distance between images for each of the features. For the hue histogram features, this is simply the weighted log difference of the histogram bins,

$$D_{KL}(h_1, h_2) = \sum_i h_{1,i} \log \frac{h_{1,i}}{h_{2,i}} .$$

To conveniently calculate a distance between images given first and second moments of the SIFT descriptors in those images, we assume they follow a multivariate Gaussian distribution. KL divergence with Gaussian densities is

$$D_{KL}(\mu_1, \Sigma_1, \mu_2, \Sigma_2) = \frac{1}{2} \text{trace} \Big\{ \Sigma_1 \Sigma_2^{-1} + \Sigma_2 \Sigma_1^{-1} - 2I + \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right) (\mu_1 - \mu_2) (\mu_1 - \mu_2)^\top \Big\} .$$

We use a combination of the two features for classification, implemented by rank-weighting under the two distance measures to get a single set of k neighbours.

Experimental Evaluation

We expect that our system will ultimately be used in an upcoming crop survey by NACRRI. In the absence of the real data that we will obtain through such a survey, we conducted computational experiments with artificial data to demonstrate the promise of our approach.

Utility function

To quantitatively evaluate the quality of our models, we need a utility function that scores how effectively a given model captures a true underlying distribution. For our example of cassava disease in Uganda, we interviewed the director of NACRRI to elicit a utility function expressing the relative importance of learning different kinds of information. His responses indicated that: 1) it is better to make small errors (e.g., predicting a disease level of 4 when the true level is 5) than big errors (e.g., predicting 2 when the true level is 5); 2) it is considerably more important to correctly assess whether a region is diseased or not than to correctly determine the specific level of disease in a diseased region; 3) being wrongly optimistic—under-predicting the incidence of disease—is about twice as bad as being wrongly pessimistic. We formalized these principles as follows. First, consider a single point in space; let y denote its true disease level, and let $\hat{y}$ denote its predicted disease level. The utility function is defined as follows, recalling that D denotes the number of disease levels and where, based on our interview with the NACRRI director, we set $\alpha = .5$ and $\beta = 2$:

$$Error(y, \hat{y}) = \frac{|y - \hat{y}|}{D - 1};$$

$$DiseasePenalty(y, \hat{y}) = \begin{cases} \alpha & y = 1 \text{ XOR } \hat{y} = 1 \\ 0 & \text{otherwise;} \end{cases}$$

$$OptimismFactor(y, \hat{y}) = \begin{cases} \beta & y > \hat{y} \\ 1 & \text{otherwise.} \end{cases}$$

Our utility function is $u_y(\hat{y}) = -OptimismFactor(y, \hat{y}) \cdot (Error(y, \hat{y}) + DiseasePenalty(y, \hat{y}))$. The $Error$, $DiseasePenalty$ and $OptimismFactor$ functions capture our expert's responses 1–3 respectively. Observe that $Error$ is a utility penalty for misclassification normalized to the

range $[0, 1]$, that α (intended to be a value from $[0, 1]$) expresses an additional penalty in the same units, and that the total penalty is multiplied by β (intended to be greater than 1) in the event that the prediction is optimistic.

We can determine the expected utility of predicting a distribution over disease levels $\hat{P}(y)$ given a true distribution over disease levels $P(y)$ by taking straightforward expectations, where $Dom(P)$ denotes the domain of P:

$$u_P(\hat{P}) = \sum_{y \in Dom(P)} P(y) \sum_{\hat{y} \in Dom(\hat{P})} \hat{P}(\hat{y}) u_y(\hat{y}).$$

Finally, we can further extend our definition of expected utility to true and predicted stochastic processes M and $\hat{M}$ that predict different distributions at predefined points in a spatial region. Let X denote a set of points of interest, let M_x denote the true distribution over disease levels at point $x \in X$, and let $\hat{M}_x$ denote the predicted distribution at x. Then the expected utility is

$$u_M(\hat{M}) = \frac{1}{|X|} \sum_{x \in X} u_{M_x}(\hat{M}_x).$$

Spatial density estimation

To compare the effectiveness of our density estimation method against the existing aggregation method used to create the map in Figure 1, we simulated a number of ground truth incidence distributions and sets of observations, and evaluated the expected utility achieved by the two methods. Simulation of disease levels was done by sampling a grid of uniform random values across the spatial field, and then convolving this matrix with a Gaussian smoothing kernel. The effect of this convolution was to filter out high-frequency components in the noise, leaving a smooth distribution, as illustrated in Figure 6 (left). The resulting values were normalised to the range [0, 1]. We simulated $I(x)$ 500 times, and from each simulation sampled 150 observations amongst 5 disease levels at uniformly random locations. With these observations, we used both methods to estimate the incidence. To evaluate the aggregative model, we split the spatial field into $n \times n$ cells, and counted the proportion of diseased observations in each cell. (For comparison, the maps currently made of disease in Uganda (as in Figure 1) aggregate at the district level which is around around 80 regions across the country.)

Figure 5 shows the relative utility, varying n, when the two density estimation models are applied to the same set of observations. We show the ratio of utility from the GP inference to the utility from simple aggregation. Using a bootstrap test we find that the mean utility for GP inference is lower than the mean utility for the aggregate model for every n at a significance level of $p = 0.01$. It is therefore clear that it consistently and significantly outperforms the aggregative density model on this utility function.

Adaptive selection of survey sites

We compared the quality of inferences made under our adaptive survey strategy to those made when survey stops were made at regular intervals along each survey route. We did this by randomly generating many survey routes and "ground truth" disease distributions, applying both survey strategies to each problem instance, and then scoring the models obtained by each strategy using our utility function.

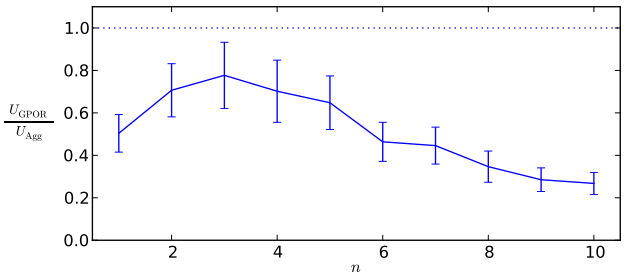

Figure 5: Comparison of GP ordinal regression density model and simple aggregation, over 500 simulations with aggregations made on a $n \times n$ grid. A utilityratio of less than 1 means that the GP model outperformed the aggregation model (since our utilities are negative numbers). The error bars show standard deviations.

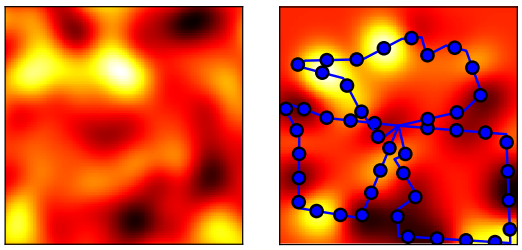

Figure 6: Heatmap showing simulated true disease incidence, where lighter colours denote higher incidence (left); Sampled trajectories of three survey groups, with survey sites at regular distances (right); heatmap shows estimated probability of disease given observations at survey locations.

To sample trajectories amongst a number of survey groups, we uniformly sampled a set of 2D points, then finding the permutation which minimised the total tour distance. A sample of three survey trajectories is shown in Figure 6 (right).

We evaluated the resulting survey utility with both regular survey locations and our adaptive survey location strategy. Figure 7 shows examples of the placement of survey sites along a single tour. When the utility function was modified to be neutral with respect to disease level ($\alpha = 0, \beta = 1$), the effect of the optimisation was to space the sample sites widely apart in areas of maximum uncertainty. When the utility function was set to reward discovery of diseased areas, survey locations were moved closer to areas of high disease incidence. Figure 8 shows survey utility as a function of budget. When the budget was low (very sparse sample sites), optimising survey location made little difference due to lack of information on which to base the adaptation. As the budget increased, the utility achieved with optimisation increased significantly compared to regular sampling.

Image-based diagnosis of crop disease

We collected a set of 469 images of healthy cassava plants (53%) and those with cassava mosaic disease (47%). We cal-

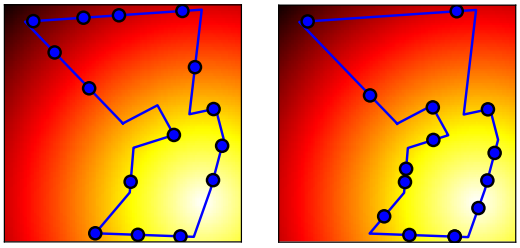

Figure 7: Stopping locations where the utility function is unbiased towards disease incidence (left), linearly correlated with underlying incidence (right). The background heatmap indicates true underlying disease incidence. Where the utility function rewards discovery of diseased regions (right), survey points are more concentrated in the regions of high incidence.

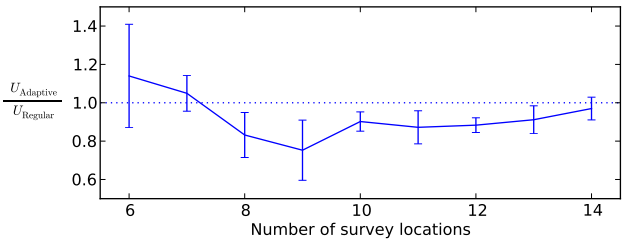

Figure 8: Average utility of simulated survey results, for regular and optimised sampling positions, as a function of the sample budget. Error bars show standard deviations. Beyond the smallest number of survey locations the mean utility using optimisation is significantly better (at $p = 0.05$ for all numbers of survey locations at least 8).

culate classification performance using leave-one-out cross-validation (LOO-CV). Classification with the combination of both features for $6 \leq k \leq 12$ achieved AUC in the range 0.959-0.961, and error rate 0.078-0.083.

In Figure 9, we put together all of the methods described in this paper. Specifically, we compare the quality of the models achieved using adaptive site selection and automatic image classification to the quality of the models achieved using adaptive site selection and expert image labeling. When both methods are given equal amounts of data (leftmost point in Figure 9) we achieve somewhat worse performance with automatic classification, due to classification error. However, our motivation for proposing the automatic method was that it would allow us to collect more data by freeing us from dependence on scarce expert surveyors. As the amount of data collected by the automatic method increases (moving right along the x axis), we observe that its relative performance improves, reaching the same level of performance as expert labeling when 80% more data is collected, and exceeding it beyond that point. Note that in these last experiments we only use two disease levels, corresponding to the image training data we have available.

Conclusions

We have demonstrated three computational techniques for modeling and monitoring crop disease in developing countries: first, a rich and flexible modeling approach that allows spatial interpolation of ordinal response values; second, a method for adaptively deciding which data to collect in order to maximise survey utility; and third, an application of computer vision that reduces the need for skilled experts.

Many interesting avenues of work remain open. One intriguing idea is to solicit ad-hoc data from agricultural extension workers, setting dynamic prices for surveys according to the expected values of different locations. It would also be useful to explicitly incorporate temporal dynamics by using the history of disease density from previous years, and knowledge of the mechanisms of disease spread

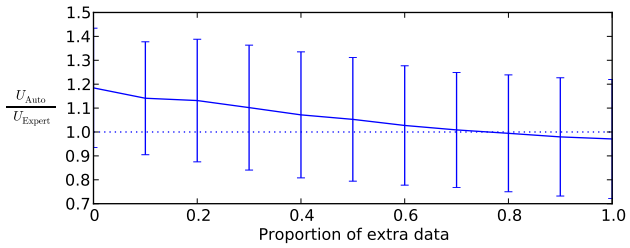

Figure 9: Comparison of utility in "automatic" classification mode (a random 8% of labels are incorrect) and "expert" mode (all labels are correct). The x axis gives the proportion of additional data available to the automatic method. The mean utility of the automated system becomes equivalent to that of the expert system when around 70-80% extra data is available.

Acknowledgements

This work was funded by a Google Research Award. Particular thanks to Titus Alicai at NACRRI. Also thanks to Chris Williams for discussions, Jennifer Aduwo for practical assistance, and to the anonymous reviewers.

References

Aduwo, J.; Mwebaze, E.; and Quinn, J. 2010. Automated vision based diagnosis of cassava mosaic disease. In *Proc ICDM Workshop on Data Mining in Agriculture*.

Anonymous. 2011. Second helpings of tapioca pudding: a crucial crop in new trouble. The Economist. Jan 27th.

Chu, W., and Ghahramani, Z. 2005. Gaussian processes for ordinal regression. *Journal of Machine Learning Research* 6:1019–1041.

Lowe, D. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2):91–110.

Nelson, M. R.; Orum, T. V.; and Jaime-Garcia, R. 1999. Applications of geographic information systems and geostatistics in plant disease epidemiology and management. *Plant Disease* 83:308–319.

Otim-Nape, G.; Alicai, T.; and Thresh, J. 2005. Changes in the incidence and severity of cassava mosaic virus disease, varietal diversity and cassava production in Uganda. *Annals of Applied Biology* 138(3):313–327.

Rasmussen, C., and Williams, C. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

van Maanen, A., and Xu, X.-M. 2003. Modelling plant disease epidemics. *European Journal of Plant Pathology* 109:669–682.