esa ECOSPHERE

A data-driven, machine learning framework for optimal pest management in cotton

Matthew H. Meisner,[1],† Jay A. Rosenheim,[2] and Ilias Tagkopoulos[3]

[1]*Department of Statistics and Center for Population Biology, University of California-Davis, Davis, California 95616 USA*
[2]*Department of Entomology and Nematology, University of California-Davis, Davis, California 95616 USA*
[3]*Department of Computer Science, University of California-Davis, Davis, California 95616 USA*

Citation: Meisner, M. H., J. A. Rosenheim, and I. Tagkopoulos. 2016. A data-driven, machine learning framework for optimal pest management in cotton. Ecosphere 7(3):e01263. 10.1002/ecs2.1263

Abstract. Despite the significant effects of agricultural pest management on crop yield, profit, environmental quality, and sustainability, farmers oftentimes lack data-driven decision support to help optimize pest management strategies. To address this need, we curated a comprehensive data set that consists of pest, pest management, and yield information from 1498 commercial cotton crops in California's San Joaquin Valley between 1997 and 2008. Using this data set, we built a Markov decision process model to identify the optimal management policy of a key cotton pest, *Lygus hesperus*, that balances the tradeoff between yield loss and the cost of pesticide applications. Our results show that pesticide applications targeting *L. hesperus* are only economically optimal during the first 2 weeks of June, and pesticide applications were associated with increased risk of an unprofitable harvest. About 46% of the observations in our data set involved at least one pesticide application outside of this optimal window, demonstrating the need for a data-driven approach to crop management. Sensitivity analyses on parameter perturbations and reduced data set sizes suggest that our methodology provides a robust policy-making tool, even in noisy data sets.

Key words: agricultural informatics; applied pest management; Markov decision processes; optimal pest management.

Received 3 July 2015; accepted 13 July 2015. Corresponding Editor: D. P. C. Peters.

† **E-mail:** mhmeisner@ucdavis.edu

Introduction

Maximizing crop yield in commercial agriculture is highly desirable for several reasons. First, the world's growing population is generating an increased demand for agricultural products (Godfray et al. 2010). Expanding the land area devoted to agriculture is often not feasible or not desirable, so increasing the yield generated from existing farmland may be the only viable way to meet this demand. Second, agriculture is critical to the global economy. Despite this significance, the profit margins in commercial agriculture are often very small. Increasing crop yield can generate increased revenue for farmers, facilitating the continued economic viability of agriculture.

However, yield maximization is not the only consideration that is relevant to a commercial farmer. Another important consideration for farmers is the cost of various agricultural inputs, including pesticides. Inputs may increase yield, but they are costly, and these costs must be considered by a farmer who wishes to maximize profits. Furthermore, indiscriminate pesticide use has other detrimental effects that are more difficult to quantify. First, excessive pesticide use may accelerate the evolution of pesticide resistance (Mallet 1989), which reduces our capacity to control pests and necessitates the development of more potent pesticides with unknown environmental impact (Roush 1987). Second, pesticide applications can adversely affect nontarget species. When pesticide applications depress populations of beneficial species, such as pest predators, these applications can contribute to secondary pest outbreaks and therefore augment future crop damage and pesticide costs (Gross and Rosenheim 2011). Finally, there is abundant evidence that pesticides are detrimental to both human health and ecosystem health (Rosner and Markowitz 2013), creating a strong incentive to avoid using these chemicals unless they are absolutely necessary.

In this study, we focus specifically on the immediate costs of pesticide applications (i.e., the amount of money that it costs a farmer to apply a pesticide), and examine the tradeoff between these immediate costs and the costs incurred from yield loss due to pest damage. While the less immediate costs of pesticide use (e.g., secondary pest outbreaks, human health) are important, we have excluded them from our analysis for two reasons. First, these costs are more difficult to quantify objectively than are the immediate costs. Second, costs such as the evolution of pesticide resistance and environmental damage from pesticides are not as immediately relevant to farmers, many of whom may be focused on short-term profit maximization, and we seek to provide management recommendations relevant to farmers' actual decision-making.

Currently, farmers lack the computational tools needed to identify management strategies that optimally navigate the tradeoff between minimizing yield loss due to pest damage and minimizing the cost of pesticide applications. Without access to data-driven pest management recommendations, farmers often rely on intuition and personal experience to guide their crop management decisions, and this may lead to suboptimal decision-making that reduces yield, increases pesticide costs, and exacerbates the deleterious effects of pesticides on human health and the environment. These problems are especially severe when pesticides are inexpensive relative to the value of the crop, as this incentivizes growers to apply them prophylactically.

In some cases, farmers may base their pest management decisions on "economic injury levels"—pest densities at which a profit-maximizing farmer is supposed to apply a pesticide. These levels are typically derived from experimental studies. However, experimental studies are often unable to resolve small effects of pests on yield, even though these small yield declines can be of substantial economic significance to commercial farmers (Rosenheim et al. 2011). In addition, growers are often provided with a single economic threshold for the entire growing season, which neglects the possibility that a crop's susceptibility to pest damage may vary throughout the growing season.

Here, we identified the optimal pest management strategies for the pest *Lygus hesperus* in commercial cotton fields. The plant bug *L. hesperus* is one of the most damaging pests of cotton, and a frequent target of insecticide applications (Rosenheim et al. 2006, Godfrey et al. 2013). We determined at which pest densities and at which times in the growing season profit-maximizing farmers should apply pesticides to treat *L. hesperus*, and we quantified how close current farmer behavior is to this optimal policy. To explore these questions, we took an "ecoinformatics" approach in which we analyzed historical records of commercial cotton production from 566 fields in California's San Joaquin Valley. By aggregating historical records from hundreds of fields and various growers, we compiled a robust data set with the power to quantify how pest densities and different pest management strategies affect yield. Ecoinformatics analyses generally involve harnessing the power of large, preexisting, observational data sets to address important questions in environmental biology, especially ones that may be difficult to study experimentally (Rosenheim et al. 2011).

First, we quantified the yield loss associated with particular pest densities at different times in the growing season. We then used a finite time horizon Markov decision process (Bauerle and Rieder 2011) to identify the optimal pest management strategy for each pest density at each time point. Finally, we assessed how closely aligned the observed grower pest management policies are to the optimal policy.

Materials and methods

Data set

The data set was constructed by collecting existing crop records from commercial cotton fields in California's San Joaquin Valley. The data were shared by both growers and pest

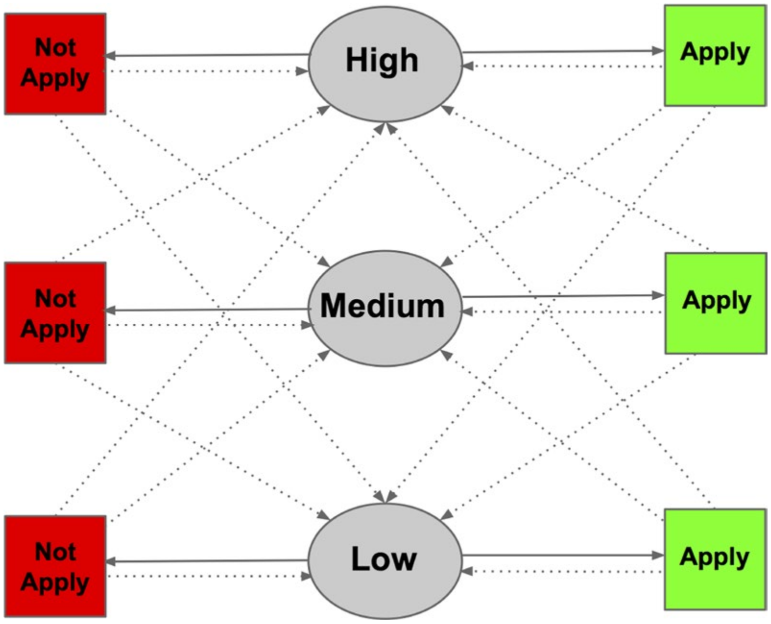

Fig. 1. Network of MDP states and actions. In each state (high, medium, or low pest density), a farmer can choose to apply or not apply a pesticide targeting *L. hesperus*. Solid lines indicate actions, and dashed lines indicate transitions.

control advisors, professional consultants hired to monitor field conditions and recommend crop management strategies. The data set contains records of 1498 unique field-year instances from 566 unique fields, ranging from 1997 to 2008. The following variables were used in our analyses:

1. Cotton yield. Measured once for each field-year instance, cotton lint yield was measured in bales/acre and recorded for 1240 of the 1498 total records.
2. Cotton species. The database consisted of records of two different cotton species: *Gossypium barbadense* L. ("Pima cotton") and *Gossypium hirsutum* L. ("Acala/upland cotton").
3. *Lygus hesperus* densities. Pest control advisors measured *L. hesperus* densities approximately weekly, primarily during June and July. The pest control advisors' sampling procedure consisted of 50 swings of a sweep net across the top of the plant canopy (for the remainder of the manuscript, we use the term "sweep" to refer to this standard, 50-swing sample). As not all pest control advisors sampled on the same days or at exactly the same intervals for all fields, we transformed successive samples into mean *L. hesperus* density estimates (insects/50-swing sweep sample) by calculating the area under the linear curve of *L. hesperus* density vs. time and dividing by the number of days in the sampling interval.
4. Pest management practices. Every time that a pesticide was applied to a field, the date, chemical, and target pest was recorded.

Markov decision process model

Model structure.—To identify optimal crop management strategies, and to quantify the differences in farmer profit under alternate strategies, we modeled farmer decision-making as a discrete time Markov decision process (MDP) with a finite time horizon (Fig. 1). MDPs provide a framework for modeling a system that can be in different states at each discrete point in time. The transitions between states over time are partially under the control of a decision maker, who can choose different actions at each time point, but are also partially stochastic. Transitions are modeled probabilistically, and the transition probabilities between states depend upon the current state and the action taken. MDPs also involve rewards or costs, which are incurred when a particular action is taken or a particular transition occurs. The MDP framework allows us to determine the action in each state at each time that is expected to maximize the sum of rewards over a specified time horizon (Puterman 2005, Bauerle and Rieder 2011).

For our specific MDP application, modeling farmer decision-making about pest management, we divided the growing season into $T = 8$ week-long intervals spanning 5 June through 30 July. While cotton is typically planted in March or April and harvested in October, consistent measurements of the main cotton pest, *L. hesperus*, are only taken during June and July; consequently, we limited our model to this period of the growing season. We used eight intervals to balance (1) the desire to provide management recommendations with fine temporal resolution (as the optimal pest management strategy may differ throughout the growing season), and (2) the inability to reliably estimate a large number of parameters.

States.—At each time point, we classified each field as being in one of three states $s \in S$: low, medium, or high pest density. The exact thresholds for the states were selected so that, aggregated over all eight time points, there was an approximately equal number of observations from each state. The specific boundaries that led to this balance, all in units of insects/sweep, were (0, 0.66) for the low state, (0.66, 1.73) for the medium state, and greater than 1.73 for the high state.

Actions.—In our model, we considered two different actions $a \in A$: either applying a pesticide or not applying a pesticide that a farmer can take. For each field at each time point, we determined if a pesticide application for L. hesperus had occurred by seeing if an application was recorded for which L. hesperus was listed as one of the target pests.

Transition probabilities.—Transition probabilities from state s to state s′ are denoted as $Q_{a,t}$ (s, s′); they depend on the action a, time t, origin state s, and destination state s′. Transition probabilities were modeled using a Bayesian ordered logistic regression, so that we could quantify and account for uncertainty in our estimates of these quantities. Ordered logistic regression was selected because the response variable of interest, the identity of the destination state, is a categorical variable with an ordered structure (low, medium, or high) (Agresti 2010). For every unique combination of origin state, time, and action, we examined the records in the data set which described fields in that state at that time where that action was taken. We then fit an ordered logistic regression with destination state (low, medium, or high) as the ordered categorical response variable. As there are three possible categories, the ordered logistic regression involves two intercept terms, both of which were given uninformative N(0, 100) priors. No predictor variables were included in the model, as the purpose of this regression was only to quantify uncertainty in the estimated transition probabilities, not to draw inferences about what affects these probabilities across fields in the same state where a given action is taken at a given time.

Costs.—The final component of the MDP model involves costs, $c_{a,t}$ (s′), which depend on the action taken, the time, and the destination state. To calculate these costs, we considered both the cost of pesticide application and the cost due to yield loss from L. hesperus damage. First, there is a cost associated with a pesticide application; this cost encompasses both the cost to purchase the product and the fuel and labor costs of applying the product. Using estimates provided by cooperating growers and crop consultants, we estimated the cost of a pesticide application targeting L. hesperus to be $20 per acre. Estimating this cost exactly is challenging for several reasons. First, growers use a variety of different pesticide products to suppress L. hesperus, and the prices of these different products vary substantially. Second, different growers may not pay the same price for the same product, as growers often negotiate discounts with chemical suppliers. Finally, some pesticide applications for L. hesperus may occur concurrently with other applications to a field; in this case, the additional cost of adding on the treatment for L. hesperus may be significantly lower than that of a dedicated application solely for L. hesperus. Despite these challenges, we feel that our estimate of $20 per acre represents a typical cost of a pesticide application targeting L. hesperus.

The second component of the cost term involves the cost associated with yield loss due to damage from *L. hesperus*. We used a hierarchical Bayesian linear mixed model (Gelman and Hill

2009) to estimate the change in yield associated with being in the medium or high pest density state, compared to being in the low pest density state, at each of the eight-time intervals. Our model included cotton yield as the response variable, a random effect for field to control for field-specific differences in yield potential, a random effect for year to control for annual fluctuations in yield, a fixed effect identifying the cotton species to control for variable yield potential between the two species, and fixed effect indicator variables indicating which state each field was in at each time interval. Uninformative N(0, 100) priors were used for all fixed effects, and uninformative InvGamma(0.001, 0.001) priors were used for all variance components. We considered the low state to be a baseline level of insect pressure, and calculated the costs of being in the medium or high state, relative to the low state. As a statistical model can only identifiably estimate two indicator variables for a categorical predictor variable with three levels, we selected one state as a baseline and only considered yield differences between the other states and that baseline state.

We quantified how yield differences between the high and low state, and yield differences between the medium and low state, changed over the growing season using a Bayesian linear regression of the estimated yield differences vs. time. As the intervals are equally spaced, we labeled the time intervals as $t = 1, \ldots, 8$ and regressed the yield differences against these numeric values. We performed a separate regression of yield difference vs. time for the yield differences between the high and low state and the yield differences between the medium and low state. We used uninformative N(0, 100) priors for the intercept and slope, and an uninformative InvGamma(0.001, 0.001) prior for the variance. To account for uncertainty in the estimates of yield differences, we repeated the regressions of yield difference vs. time, 10 000 times, each time using estimated yield differences sampled from the posterior of the mixed model in which these differences were estimated. We obtained 10 000 posterior samples from each individual linear regression, and then analyzed all 10 000 × 10 000 posterior samples in order to perform inference about the slope of the regression.

After calculating, for each of the eight time intervals, the difference in yield between fields in the medium state and the low state, and the difference in yield between fields in the high state and the low state, we converted these differences in yield to differences in farmer revenue, in dollars, using the most recent estimates of the value of cotton per pound published by the UC-Davis Department of Agricultural and Resource Economics: $0.90 per pound for Acala cotton (Hutmacher 2012a), and $1.30 per pound for Pima cotton (Hutmacher 2012b). As the difference in value of the two cotton species is substantial, we considered Acala and Pima cotton separately when converting the high vs. low and medium vs. low yield differences into revenue.

The total cost for a particular transition at time t to state s' at $t + 1$, denoted as $c_{a,t}\ (s')$, was the sum of the $20 cost of pesticide application, if an application occurred, and the cost of the yield loss due to *L. hesperus* damage if a transition led to either the medium or high state at the next time point. Costs for transitions to the low state only involved the cost of a pesticide application, if it occurred. If we expected yield to be lost in the medium or high states, compared to the low state, the cost was considered to be a negative value.

All statistical models were fit using Markov Chain Monte Carlo, implemented in the Stan probabilistic programming language for Bayesian inference, and accessed through the RStan package in the R language for statistical computing (Stan Development Team, 2013). For each model, we performed two independent Markov chain simulations of length 10 000, and discarded the first 5000 samples of each as burn-in. We verified that our MCMC simulations had converged by ensuring that $\widehat{R}$, a measure of expected posterior scale reduction if sampling were to be continued indefinitely, was near one (Gelman and Hill 2009). All code for our MDP analyses was written in the R programming language, without the use of any MDP-related packages.

Solving for optimal policies.—The MDP modeling framework allows us to identify the optimal policy that a farmer should follow in order to maximize profits over the course of the growing season. For each state at each time, we determined whether or not a profit-maximizing farmer should apply pesticides to suppress L. hesperus. The optimal policy, a function of state and time, is given by Eq. 1:

$$\pi_t(s) = \underset{a}{\operatorname{argmax}} \left\{ \sum_{s' \in S} Q_{a,t}(s,s') \left[c_{a,t}(s') + V_{t+1}(s') \right] \right\} t \in [1,7]$$

And the "value" of each state at each time is given by Eq. 2:

$$V_t(s) = \sum_{s' \in S} Q_{\pi_t(s),t}(s,s') \left[c_{\pi_t(s),t}(s') + V_{t+1}(s') \right] t \in [1,8]$$

To solve for the optimal policy, one needs estimates of the transition probabilities, costs, and the terminal value, i.e., $V_8(s)\ s \in\ S$. Using these estimates, we can work backwards using stochastic dynamic programming to solve for the optimal policy in each state at each time (Puterman 2005). Due to the different values of the two cotton species, we obtained the optimal policy separately for both Acala and Pima cotton.

Parameter estimation under uncertainty.—As both the transition probabilities and yield declines due to L. hesperus are quantities that we had to estimate from the data, there is uncertainty associated with these estimates. To propagate this uncertainty, we obtained, for each time-state combination, 10 000 posterior "samples" of $\sum_{s' \in S} Q_{a,t}(s,s') \left[c_{a,t}(s') + V_{t+1}(s') \right]$ for both possible actions, and selected the action with highest posterior mean as the optimal action. The samples were obtained by considering uncertainty in all three components of the value function: the costs, transition probabilities, and the value function at the next time step. We obtained 10 000 posterior samples of each of these three components. Specifically, we obtained 10 000 posterior samples of the costs by sampling from the posterior of the mixed model in which we estimated the yield decline associated with being in the high or medium state, compared to the low state, at each time interval. We obtained 10 000 posterior samples of the transition probabilities by sampling from the posterior of the ordered logistic regression with which we estimated the transition probabilities corresponding to that time-state-action combination. Our parameterization included the cost of yield loss in the destination state s′ during the next time interval in the $c_{a,t}$ (s′) term. This means that the cost of yield loss due to L. hesperus at t = 8 is included in the cost term that is considered when selecting the optimal action at t = 7; therefore, we set $V_8(s) = 0\ s \in \mathrm{S}$ and worked backwards using dynamic programming to determine the optimal policy for each state at each time. We stored 10 000 posterior samples of the value of each state at each time, and used these stored samples when accounting for uncertainty in, $V_{t+1}(s')$ the value at the next time step, which is a component of $\sum_{s' \in S} Q_{a,t}(s,s') \left[c_{a,t}(s') + V_{t+1}(s') \right]$ Our reasons for accounting for uncertainty in the model parameters, instead of using point estimates, are twofold. First, it is misleading to consider the costs and transition probabilities as known, as we used statistical models to estimate these parameters from the data. The values of these parameters can substantially affect the recommended policy. Unless every posterior distribution is normally distributed, which is unlikely, using maximum likelihood estimates to obtain point estimates of the value function may lead to a different optimal policy than the one obtained using our approach of accounting for uncertainty in all of the model parameters to obtain posterior samples of the value function. To provide growers with useful recommendations, we need to quantify the statistical uncertainty about the values to quantify our confidence in the optimal policy.

Second, propagating uncertainty in all the model parameters provided us a more in-depth quantification of the value differences between the two actions than we would have otherwise obtained. If we had instead used point estimates for the costs and transition probabilities, we would have only obtained point estimates for the value of each action at each time and state, and, therefore, only a point estimate of the difference in value between the two actions. With our Bayesian method, we obtained posterior distributions of the difference in value between the two actions. These distributions facilitate the exploration of a comprehensive set of questions about the differences between the two actions, such as which action leads to the lowest variance in value, which minimizes extreme losses, and which leads to the highest median value.

Optimality criteria.—Different optimality criteria can be considered when choosing the optimal action. To determine the profit-maximizing strategy, we selected the action that maximized the posterior mean of the value function (Eq. 1). In addition, we determined the lowest risk strategy by exploring the likelihood, under each possible action, of net revenue (income from crop yield minus pesticide costs) falling below the minimum revenue required for a profitable harvest. We determined the minimum revenue required for a farmer to break even using the most recent estimates of total cotton production costs published by the UC-Davis Department of Agricultural and Resource Economics: $1800 per acre for both Acala and Pima cotton (Hutmacher 2012a,b). When selecting the optimal action for each state at each time, we selected the action that minimized the likelihood of revenue falling below this threshold. Again, separate policies were identified for both Acala and Pima cotton due to the substantial difference in value of the species. This criterion can be mathematically formulated as follows, where r_{crit} is the critical revenue threshold:

$$\pi_t(s) = \underset{a}{\operatorname{argmin}} \left\{ P\left(\sum_{s' \in S} Q_{a,t}(s,s')[C_{a,t}(s') + V_{t+1}(s')] < r_{\text{crit}} \right) \right\} t \in [1,7]$$

Sensitivity analyses

We performed three different sensitivity analyses to quantify the robustness of our results.

Missing data.—First, we randomly divided the data set into three subsets with approximately the same number of observations. One at a time, we removed one-third of the data set, repeated the MDP-solving procedure (using the posterior mean to determine optimal policy), and calculated the proportion of time-state combinations in which the optimal policy was the same as when we used the full data set.

Noise.—Second, we added stochastic noise to all of the variables in the model (the pest densities at each time interval, and yield), repeated the MDP-solving procedure (again using the posterior mean to select the optimal policy), and calculated the proportion of time-state combinations where the optimal policy was the same as that when we used the real data to fit the model. To each variable, we added normally distributed random noise with mean zero, and standard deviation of 0.1, 0.5, 1, and 2 times the sample standard deviation of the original values for that variable.

Mislabeling.—Finally, we assessed our model's robustness to errors in the records of actions (applying or not applying pesticides) at each time point. We randomly selected a certain percentage of observed actions in each time interval, and switched the recorded action (i.e., we labeled applications as no applications, and no applications as applications). We repeated this for various percentages, refit the model, and calculated the proportion of time-state combinations in which the optimal policy was the same as the one selected using the original data.

Results

Effects of *L. hesperus* on yield

We used a Bayesian linear mixed model to estimate the change in yield associated with being in the high and medium pest density states vs. being in the low state, for all eight time intervals (Fig. 2). There was a trend for the difference in yield between fields in the high vs. low state to increase throughout the growing season; a similar trend was noticed for the difference in yield between fields in the medium and low state. We quantified this trend with a Bayesian linear regression of the yield differences between the high and low

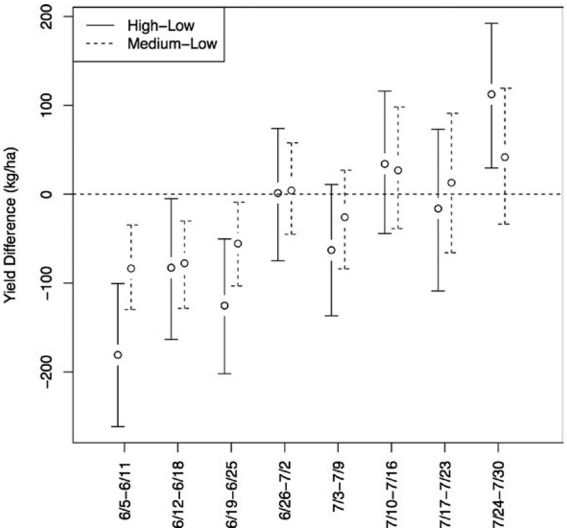

Fig. 2. Means and 95% highest posterior density intervals for the difference in yield between fields in the high vs. low, and medium vs. low, pest density states during 8-week-long intervals.

state, and between the medium and low state, vs. time. The slope when regressing the yield difference between the high and low state vs. time had a posterior mean of 30.3 kg/(hectare × week), with a 95% HPDI of (11.2 kg/(hectare × week), 55.7 kg/(hectare × week)); the slope when regressing the yield difference between the medium and low state vs. time had a posterior mean of 19.0 kg/(hectare × week) with a 95% HPDI of (5.6 kg/(hectare × week), 32.9 kg/(hectare × week)). These entirely positive 95% HPDIs suggest that cotton is more susceptible to yield loss from *L. hesperus* herbivory early in the growing season. Interestingly, the posterior means for the yield differences between high and low and between medium and low states were positive for the 24 July–30 July time interval (the final time interval in our analysis), and the 95% HPDI for the difference in yield between the high and low state was entirely positive.

Optimal policies

Highest mean.—When the optimal policy is determined by selecting the action resulting in the highest posterior mean of the value function, the optimal policies for Pima and Acala cotton are presented in Tables 1 and 2, respectively. For Pima cotton, the optimal policy only called for a pesticide application to target L. hesperus in the medium and high states during week 1, and the high state during week 2. The optimal policy for Acala cotton only called for pesticide application in the medium and high states during week 1. Tables 3 and 4 display the difference in posterior mean of the value function between the optimal and suboptimal action, for Pima and Acala cotton, respectively. These values can be interpreted as the expected increase in profit when the optimal policy is followed compared to when the suboptimal policy (i.e., applying a pesticide when not recommended or not applying a pesticide when recommended) is followed (assuming the optimal policy is followed at all subsequent time steps).

Lowest risk.—When the optimal policy is determined by selecting the action resulting in the lowest likelihood of yield falling below the threshold required for a profitable harvest, the optimal policy for Pima and Acala cotton is presented in Table 5. Coincidentally, when the optimal policy is defined in this way, the optimal policy is the same for both species of cotton. In contrast to the mean-maximizing policy, the risk-minimizing policy never calls for pesticide applications.

Sensitivity analyses

The optimal policy when using two-thirds of the data was the same as the policy when using the entire data set for 87% of state-time combinations, indicating that our results are robust to data removal. Adding noise to the variables

Table 1. The optimal L. hesperus management policy in Pima cotton, for each state at each time. The optimal policy is defined by the management decision, either apply (A) or not apply (NA) pesticides, with the lowest posterior mean of expected long-term costs.

	5–11 Jun	12–18 Jun	19–25 Jun	26 Jun–2 Jul	3–9 Jul	10–16 Jul	17–23 Jul
Low	NA	NA	NA	NA	NA	NA	NA
Medium	A	NA	NA	NA	NA	NA	NA
High	A	A	NA	NA	NA	NA	NA

Table 2. The optimal L. hesperus management policy in Acala cotton, for each state at each time. The optimal policy is defined by the management decision, either apply (A) or not apply (NA) pesticides, with the lowest posterior mean of expected long-term costs.

	5–11 Jun	12–18 Jun	19–25 Jun	26 Jun–2 Jul	3–9 Jul	10–16 Jul	17–23 Jul
Low	NA	NA	NA	NA	NA	NA	NA
Medium	A	NA	NA	NA	NA	NA	NA
High	A	NA	NA	NA	NA	NA	NA

Table 3. The mean difference (dollars/acre) of the posterior means of the value function evaluated at the optimal policy and the suboptimal policy for each state and time, for Pima cotton.

	5–11 Jun	12–18 Jun	19–25 Jun	26 Jun–2 Jul	3–9 Jul	10–16 Jul	17–23 Jul
Low	$14.50	$17.51	$20.81	$19.79	$4.04	$54.70	$35.50
Medium	$45.59	$2.79	$18.75	$18.55	$22.46	$17.95	$45.61
High	$33.81	$2.64	$20.06	$12.96	$23.89	$19.99	$42.99

Table 4. The mean difference (dollars/acre) of the posterior means of the value function evaluated at the optimal policy and the suboptimal policy for each state and time, for Acala cotton.

	5–11 Jun	12–18 Jun	19–25 Jun	26 Jun–2 Jul	3–9 Jul	10–16 Jul	17–23 Jul
Low	$16.18	$18.15	$20.36	$20.18	$3.26	$44.23	$30.52
Medium	$25.43	$8.45	$19.53	$19.58	$21.73	$18.49	$37.86
High	$18.04	$4.43	$20.11	$15.36	$22.69	$19.99	$35.75

Table 5. The optimal L. hesperus management policy, for each state at each time, when the optimal policy is defined by the management decision, either apply (A) or not apply (NA) pesticides, that minimizes the likelihood of an unprofitable harvest. This policy was, coincidentally, the same for both Acala and Pima cotton.

	5–11 Jun	12–18 Jun	19–25 Jun	26 Jun–2 Jul	3–9 Jul	10–16 Jul	17–23 Jul
Low	NA	NA	NA	NA	NA	NA	NA
Medium	NA	NA	NA	NA	NA	NA	NA
High	NA	NA	NA	NA	NA	NA	NA

with the same standard deviation as the original data resulted in the same optimal policy in more than 80% of time-state combinations (Fig. 3A), suggesting that our model is robust to moderate measurement error in the underlying data. Even when 25% of the actions were mislabeled, the optimal policy remained unchanged in more than 80% of state-action combinations, suggesting that our model is robust to errors in the records of pest management practices (Fig. 3B).

Current practices

For every state at each time interval, we examined the instances in our data set that were in that state that time, and calculated the proportion of those instances in which the optimal policy was followed (Tables 6 and 7). While growers generally followed the optimal policy during the final weeks of the season, only 2% and 6% of Pima instances followed the optimal policy during week 1 in the medium and high states, respectively. Only 8% and 18% of Acala instances followed the optimal policy during week 1 in the medium and high states, respectively. In other words, most farmers we observed did not apply pesticides targeting *L. hesperus* in this period, despite evidence that *L. hesperus* during this period is associated with decreased yield. On average, farmers followed the optimal policy 92% of the time when a pesticide application was not recommended, but only 10% of the time when an application was recommended.

Aggregated across both species and all states and times, there were an average of 0.52 pesticide applications per field when an application was not recommended, compared to 0.36 failures to apply pesticides when an application was recommended. So, on average, after considering the fact that applications were only infrequently recommended, the cotton farmers we observed applied pesticides when not recommended more frequently than they failed to apply pesticides when applications were recommended. About 46% of the observations in our data set involved

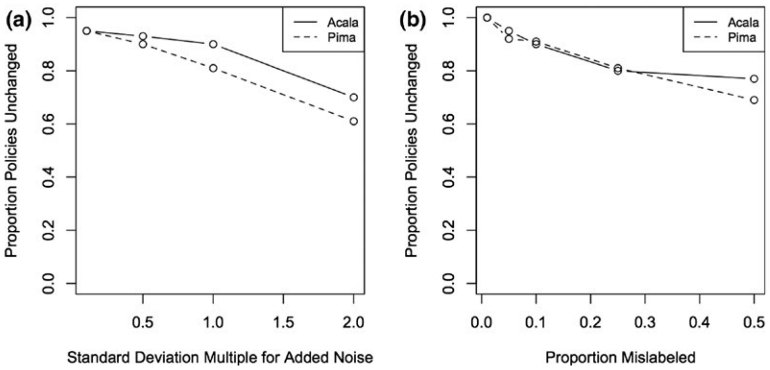

Fig. 3. The proportion of time-state instances in which the optimal policy remained unchanged, compared to using the original data, when all the variables used in the model fitting were perturbed with the addition of stochastic variation that was normally distributed with mean set to 0 and standard deviation set to various multiples of the sample standard deviation for each variable (a); the proportion of time-state instances in which the optimal policy remained unchanged when various proportions of the actions were mislabeled (b).

Table 6. The proportion of Pima cotton instances in the database that followed the optimal policy (defined by maximizing the posterior mean), for each state at each time.

	5–11 Jun	12–18 Jun	19–25 Jun	26 Jun–2 Jul	3–9 Jul	10–16 Jul	17–23 Jul
Low	0.98	0.97	0.98	1.00	0.94	1.00	0.85
Medium	0.02	0.86	0.95	0.91	0.98	0.94	0.94
High	0.06	0.21	0.67	0.83	0.81	0.79	0.80

Table 7. The proportion of Acala cotton instances in the database that followed the optimal policy (defined by maximizing the posterior mean), for each state at each time.

	5–11 Jun	12–18 Jun	19–25 Jun	26 Jun–2 Jul	3–9 Jul	10–16 Jul	17–23 Jul
Low	0.97	0.98	0.95	0.97	0.99	0.99	1.00
Medium	0.08	0.89	0.89	0.96	0.98	0.97	0.97
High	0.18	0.67	0.74	0.83	0.88	0.90	0.87

at least one pesticide application that was not recommended.

Discussion

Using a large observational data set from commercial cotton fields, we used a Markov decision process to model the pest management decisions made by farmers throughout eight key weeks of the growing season. Using this model, we were able to determine, for each pest density at each time interval, whether or not a farmer should apply pesticides to treat *L. hesperus* in order to maximize his/her expected profits over the course of the growing season.

Our observation that *L. hesperus* was only associated with decreased yield during the early part of the growing season, and was in fact associated with increased yield during the last week of July, highlights the limitations of adopting a pest management strategy that involves pest suppression when pests exceed a certain threshold. Our results suggest that using a single threshold for *L. hesperus* management throughout the season is undesirable, as the sensitivity of cotton yield to *L. hesperus* pressure decreased steadily throughout the growing season. Based on our results, cotton growers should focus their efforts toward detection and suppression of *L. hesperus* on the earlier part of the growing season (the first 2 weeks of June, in particular). Furthermore, they should avoid pesticide applications later in the growing season, where the yield loss due to *L. hesperus* herbivory (which in some cases was actually a yield increase) does not justify the costs of these applications.

One limitation of the Markov decision process model we implemented is that it involved discretizing pest density, which is a continuous variable, into a finite state space with three states. Doing so implicitly assumes that the value function is a constant value across all pest densities comprising a single state. This is unlikely to be exactly the case, especially for the high state, which had an infinite upper boundary. Despite this limitation, it was necessary to discretize the state space into discrete states in order to implement this model. With more data, it would be possible to accurately estimate yield effects associated with more, finer resolution states.

A second limitation of our approach stems from the fact that our work with observational data restricts our analysis to the existing variation in the data. As our data set did not contain any observations of extremely high *L. hesperus* densities, our recommendation of not applying pesticides during most weeks, even when in the high pest density state, may not actually be optimal in a field with a severe *L. hersperus* outbreak at densities beyond the range of densities that we analyzed. Extremely high *L. hesperus* densities can inflict crop damage so severe that a harvest is abandoned altogether, so it is likely that pesticide applications actually are economically optimal during these severe infestations. However, without data under such conditions, our analysis was unable to consider this possibility.

Interestingly, we found that the policy that minimizes the likelihood of yield falling below the level required for farmers to be financially profitable never involved the application of pesticides. Commercial farmers, particularly when operating under narrow profit margins, may tend to be financially risk averse; some growers may find it more desirable to minimize the risk of losing money rather than to maximize expected profits. We are not sure why pesticide applications were associated with increased risk for unprofitable harvests. However, one possible explanation is that pesticide applications may increase the risk of secondary pest outbreaks, possibly due to their detrimental effects on nontarget, beneficial species (Gross and Rosenheim 2011).

As noted in the results section, during the 24 July–30 July time interval fields in the high and medium pest density states were associated with higher yield than were fields in the low pest density state. While we do not have data to specifically test this hypothesis, we hypothesize that the increased yield observed in fields with higher *L. hesperus* densities in late July results from *L. hesperus* preferentially attacking young flower buds, which, this late in the season, may be too late to mature before harvest. Herbivory on these young flower buds may result in the cotton plant devoting more resources to the more mature, existing flower buds, which do contribute to yield. So, by removing young flower buds that will not contribute to yield and focusing the plant's resources on harvestable fruits, higher *L. hesperus* densities in late July may contribute to increased yield.

Our finding that high and medium *L. hesperus* densities are associated with the largest yield decline at the beginning of June, as well as our related finding that the optimal policy only involves suppression of *L. hesperus* at the beginning of June, are both consistent with previous research suggesting that cotton is more susceptible to *L. hesperus* damage early in the growing season. Simulation models (Gutierrez et al. 1975, Mangel et al. 1985), experimental studies (Falcon et al. 1971), and a previous analysis of this data set (that was focused only on estimating yield loss) (Rosenheim and Meisner 2013) have all provided evidence for increased cotton sensitivity to *L. hesperus* herbivory early in the growing season.

Our current study extends on this existing body of work in several ways. First, we have quantified how much yield loss can be expected from different pest densities at different times of the growing season. Second, we have quantified the trend for cotton sensitivity to *L. hesperus* herbivory to decrease over the growing season. Third, we have quantified how pesticide applications affect the likelihood of pest suppression at different times of the growing season. Fourth, we have integrated information on crop yield loss due to herbivory, the economic cost of that yield loss, pesticide efficacy, and the cost of pesticide applications in order to construct a holistic model of farmer pest management decision-making. This MDP model allowed us to determine what pest management decision a profit-maximizing farmer should make, based on the time in the growing season and the density of *L. hesperus* in a field, and it allowed us to quantify exactly how much increase in profit a farmer should expect when following the optimal, compared to the suboptimal, policy.

This study serves as an example of how data-driven decision-making in agriculture can confer significant advantages. Only 5% of farmers were following the optimal policy when in the medium pest state during week one, so, in some situations, current grower behavior appears to be suboptimal. Following the optimal policy has the potential to enhance commercial cotton production in several ways. First, by more actively suppressing *L. hesperus* in early June, farmers could help reduce yield loss associated with *L. hesperus* at this time. This could increase yield and farmer profits, both of which are desirable in order to sustain the economic vitality of agriculture and generate sufficient resources for a growing population. Second, by eliminating unnecessary pesticide applications in July, farmers could save money and help avoid the adverse effects of pesticides on the environment and human health.

There are several reasons why our ecoinformatics approach, which involved collecting large amounts of historical data from commercial farms, is a valuable approach for helping derive insights into optimal pest management policy. First, collecting data sets of this nature can be an affordable and efficient way to obtain data sets much larger than those that could be obtained experimentally. Replicating an experiment at the scale of a commercial plot hundreds of times requires resources rarely, if ever, available to agricultural researchers. Yield losses of just 1 or 2% can be economically significant to farmers, and it is very difficult to resolve yield effects of this magnitude without a large amount of data. Second, as our data set consists of data collected from commercial farms, it captured the true spatial and temporal scale of commercial agriculture. Experimental studies often rely on small plots, which may not be of sufficient size to provide a realistic picture of commercial cotton fields, which can be hundreds of acres in size. In particular, small plots may fail to capture the spatial dynamics of highly mobile pests. Third, collecting data from actual farms allowed us to quantify how closely, and at which times of the growing season, current farmer decision-making matches the optimal decisions; this information could be useful in guiding outreach and extension efforts that will have the most beneficial impact of cotton production.

While our ecoinformatics approach has several advantages, there are limitations to this approach. First, as our data set consists of observational data, inferences about causal relationships are not possible without making strong and untestable assumptions. When a trend is observed in a controlled experiment, one can be confident that the treatment manipulated by the researcher caused the observed change in response variable; however, in an observational data set, it is impossible to prove that some external factor did not affect both the treatment assignment and the response variable, thus spuriously suggesting a treatment effect. Second, our approach limits us to analyzing the range of variation already present in the data set. For example, we are not able to estimate the yield loss due to extremely high pest densities, as none of the farmers from whom we gathered data allowed the pest densities to reach these levels. However, in an experimental study, the researcher can decide which levels of a treatment he or she wishes to investigate.

Thus, ecoinformatics will not replace or diminish the value of experimental research; rather, it will serve as a complementary approach for generating data-driven crop management recommendations and promising hypotheses for future experimental investigation.

Acknowledgments

We sincerely thank the growers and crop consultants who generously donated their data and time to make this research possible. We also thank the California State Support Committee of Cotton Incorporated, University of California Statewide IPM Program, USDA-NRICGP (Grant 2006-01761), California Department of Pesticide Regulation, UC-Davis Opportunity Fund (Grant 07427) and National Science Foundation GRFP (Grant DGE-1148897) for financial support.

Literature Cited

Agresti, A. 2010. Analysis of ordinal categorical data. John Wiley & Sons, New York, New York, USA.

Bauerle, R., and U. Rieder. 2011. Markov decision processes with applications to finance. Springer, New York, New York, USA.

Falcon, L. A., R. van den Bosch, J. Gallagher, and A. Davidson. 1971. Investigation of the pest status of Lygus hesperus in cotton in central California. Journal of Economic Entomology 64:56–61.

Gelman, A., and J. Hill. 2009. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge, UK.

Godfray, H. C. J., J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas, and C. Toulmin. 2010. Food security: the challenge of feeding 9 billion people. Science 327:812–818.

Godfrey, L. D., P. B. Goodell, E. T. Natwick, D. R. Haviland and V. M. Barlow. 2013. UC IPM pest management guidelines: cotton. University of California Division of Agriculture and Natural Resources. http://www.ipm.ucdavis.edu/PMG/selectnewpest.cotton.html

Gross, K., and J. A. Rosenheim. 2011. Quantifying secondary pest outbreaks in cotton and their monetary cost with causal-inference statistics. Ecological Applications 21:2770–2780.

Gutierrez, A. P., L. A. Falcon, W. Loew, P. A. Leipzig, and R. van den Bosch. 1975. An analysis of cotton production in california: a model for acala cotton and the effects of defoliators on its yields. Environmental Entomology 4:125–146.

Hutmacher, R. B. 2012a. Sample costs to produce cotton: acala variety. University of California Cooperative Extension. http://cottoninfo.ucdavis.edu/files/150401.pdf

Hutmacher, R. B. 2012b. Sample costs to produce cotton: pima variety. University of California Cooperative Extension. http://cottoninfo.ucdavis.edu/files/150403.pdf

Mallet, J. 1989. The evolution of insecticide resistance: Have the insects won? Trends in Ecology and Evolution 4:336–340.

Mangel, M., S. E. Stefanou, and J. E. Wilen. 1985. Modeling *Lygus hesperus* injury to cotton yields. Journal of Economic Entomology 78:1009–1014.

Puterman, M. L. 2005. Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons, New York, New York, USA.

Rosenheim, J. A., and M. H. Meisner. 2013. Ecoinformatics can reveal yield gaps associated with crop-pest interactions: a proof-of-concept. PLoS ONE 8:e80518.

Rosenheim, J. A., K. Steinmann, G. Langellotto, and A. Zink. 2006. Estimating the impact of *Lygus hesperus* on cotton: the insect, plant, and human observer as sources of variability. Environmental Entomology 35:1141–1153.

Rosenheim, J. A., S. Parsa, A. A. Forbes, W. A. Krimmel, Y. H. Law, M. Segoli, M. Segoli, F. S. Sivakoff, T. Zaviezo and K. Gross. 2011. Ecoinformatics for integrated pest management: expanding the applied insect ecologist's tool-kit. Journal of Economic Entomology 104:331–342.

Rosner, D., and G. Markowitz. 2013. Persistent pollutants: a brief history of the discovery of the widespread toxicity of chlorinated hydrocarbons. Environmental Research 120:126–133.

Roush, R. T. 1987. Ecological genetics of insecticide and acaricide resistance. Annual Review of Entomology 32:361–380.

Stan Development Team. 2013. Stan: A C++ library for probability and sampling, Version 1.3.