Ecoinformatics (Big Data) for Agricultural Entomology: Pitfalls, Progress, and Promise

Jay A. Rosenheim[1,2,*,†] and Claudio Gratton[3,†]

[1]Department of Entomology and Nematology, University of California, Davis, California 95616; email: jarosenheim@ucdavis.edu

[2]Center for Population Biology, University of California, Davis, California 95616

[3]Department of Entomology, University of Wisconsin, Madison, Wisconsin 53706

Annu. Rev. Entomol. 2017. 62:399–417

First published online as a Review in Advance on November 28, 2016

The *Annual Review of Entomology* is online at ento.annualreviews.org

This article's doi:
10.1146/annurev-ento-031616-035444

[*] Corresponding author

[†] Both authors contributed equally to this review.

Keywords

observational studies, large scale, data analysis, spurious correlations

Abstract

Ecoinformatics, as defined in this review, is the use of preexisting data sets to address questions in ecology. We provide the first review of ecoinformatics methods in agricultural entomology. Ecoinformatics methods have been used to address the full range of questions studied by agricultural entomologists, enabled by the special opportunities associated with data sets, nearly all of which have been observational, that are larger and more diverse and that embrace larger spatial and temporal scales than most experimental studies do. We argue that ecoinformatics research methods and traditional, experimental research methods have strengths and weaknesses that are largely complementary. We address the important interpretational challenges associated with observational data sets, highlight common pitfalls, and propose some best practices for researchers using these methods. Ecoinformatics methods hold great promise as a vehicle for capitalizing on the explosion of data emanating from farmers, researchers, and the public, as novel sampling and sensing techniques are developed and digital data sharing becomes more widespread.

1. INTRODUCTION

Data velocity: data collection per unit time; often high (e.g., real time) in Big Data approaches

The era of big data is here. Our ability to collect vast quantities of data, store them, and retrieve them digitally, along with advances in computational power and sophisticated algorithms for the analysis of large data sets, is offering new opportunities for understanding and predicting the behavior of complex natural systems (17, 87). If investment in big data by large agricultural corporations and the rise of startups that handle ecological data and analytics are any indication, there is great promise in what big data can provide society (33, 70, 88). For researchers to contribute to realizing that promise, however, we need to develop research methods that respect the differences between big data and traditional experimental data sets.

Ecoinformatics (see the sidebar titled A Working Definition of Ecoinformatics) is the application of big data methods in ecology (see also 12, 53, 76, 93, 122). Ecologists and entomologists have been using approaches we would now refer to as ecoinformatics for at least 75 years. For example, Waloff (137) and Carpenter (23) used data from historical texts, maps, and museum records to reconstruct migratory locust (*Locusta migratoria*) outbreak patterns and pathways of colonization into Europe using records dating back to 300 CE. Thus, it is fair to ask, is big data in entomological research just old wine in new bottles, or are there new features of ecoinformatics in the digital age that warrant more careful examination? Here we provide the first review of ecoinformatics studies in agricultural entomology, as we attempt to answer this question.

We focus on applications of ecoinformatics methods to questions in agricultural entomology, occasionally drawing from related disciplines (forest entomology, insect conservation, medical entomology, and plant pathology) to identify additional opportunities and techniques. We place ecoinformatics in the context of more traditional experimental entomological science (so-called small data) (74) and highlight the possible pitfalls associated with ecoinformatics methods. As in other fields (39, 72, 82, 94, 122), the use of ecoinformatics in entomology will likely only expand. Innovations in big data technology invite entomologists to adapt their research approaches, acquire

A WORKING DEFINITION OF ECOINFORMATICS

Ecoinformatics as defined here refers to ecological studies that use preexisting data (12, 72, 122). Ecoinformatics data sets used in agricultural entomology have, thus far, been almost entirely observational, and thus our review addresses observational data sets as a key element of current ecoinformatics methods. In the future, however, ecoinformatics built on composite experimental data sets should grow in importance (74). Ecoinformatics is an offshoot of big data research methods in that ecoinformatics data sets are characterized by high data volume, high data variety, and, often, high data velocity [see Ekbia et al. (39) for an excellent discussion of definitions of big data]. Many ecoinformatics studies achieve their central insights by integrating multiple data streams to create a composite data set for analysis. By using preexisting data, and thus freeing researchers from the burden of generating every datum with their own labor, ecoinformatics creates opportunities for working with data sets that are substantially larger than those generated by experimentation (larger number of observations) and that can embrace a larger number of potential predictor or response variables (more diverse data sets) (94). For this reason, ecoinformatics methods are particularly attractive when researchers wish to investigate ecological processes that occur at spatial or temporal scales that are too large to be addressed easily with experimentation, or when larger data sets are needed to resolve small, but still economically important, effect sizes. Ecoinformatics methods often require careful data management and statistical analysis because of the large and heterogeneous data streams and the serious interpretational pitfalls associated with observational data sets. Ecoinformatics methods are often best used in combination with experimentation, which has unique strengths in identifying causal relationships.

new skills, and increase collaborations with experts in areas outside of traditional entomology, including biostatisticians and computer scientists.

Citizen science: research collaborations between the public and scientists to collect, explore, and analyze data on the natural world

Crowdsourcing: participatory activities that use groups of people to solve problems or carry out tasks; recently facilitated by the Internet

Passive surveillance: monitoring of a community or process by analyzing unsolicited reports or collections of information

2. ECOINFORMATICS IN AGRICULTURAL ENTOMOLOGY

We review studies that use ecoinformatics methods (preexisting data sets that are almost invariably observational; see the sidebar titled A Working Definition of Ecoinformatics) to address questions in agricultural entomology. We surveyed the published literature, attempting to review key examples of ecoinformatics methods applied to the full diversity of questions in agricultural entomology. No simple search terms were available to uncover the relevant literature, and we apologize to authors whose works were omitted inadvertently. To characterize the ecoinformatics literature with some simple quantitative metrics, for each reviewed study, we recorded (*a*) the data set size (number of replicates included in the core statistical analysis); (*b*) the temporal scale of the data set (number of years covered); and (*c*) the diversity of the data set (number of explanatory variables included). Few authors quantified the spatial extent of their data sets, so we did not attempt to record the spatial scale of the studies.

Our survey of the literature highlights several important features of ecoinformatics studies (**Table 1**). One theme of our review is that ecoinformatics methods and traditional, experimental methods each have advantages and disadvantages, with the strengths of one largely complementing the weaknesses of the other. Thus, these two methods can gainfully be used together.

2.1. Sources of Data

Studies obtained data from a range of old and new sources of information, including public and private data archives, citizen science and crowdsourcing, social media, and distributed and remote sensing technologies (114). Data can be collected intentionally to answer a specific entomological question, or they may be the by-product of other sampling programs, with data being repurposed to answer new questions.

2.1.1. Federal, state, and private data repositories. International, national, state, and local agencies regularly collect and store a diversity of environmental, agricultural, and entomological data (e.g., **http://traps.ncipmc.org/**, **http://sba.ipmpipe.org/cgi-bin/sbr/public.cgi**, **https://datcpservices.wisconsin.gov/pb**). For example, monitoring programs often map the occurrence of insect pests and attempts to control or eradicate them (97, 123, 133). Insects are also intercepted through port-of-entry monitoring and inspection of produce (5) [e.g., Port Information Network (52, 81)], with samples often deposited in entomological collections for identification (130). Surveys of agricultural producers, including beekeepers, provide additional means of tracking population trends or occurrence of alien species, pests, or pathogens (21, 136). In addition, agricultural cooperatives and private pest management consultants are regularly engaged in insect sampling as part of integrated pest management programs (36, 45, 46, 91, 112, 120). As data collection itself becomes increasingly digital (132), we expect the accessibility of data on insects in agriculture to expand dramatically.

2.1.2. Indirect sampling and passive surveillance. Insects can be monitored indirectly by tracking the consequences of their activities, such as damaging plants, and this gives researchers a method for inferring insect abundances, distributions, or phenologies. For example, the presence of insect-vectored pathogens has been used to infer aphid activity (126). This passive surveillance approach is most commonly used with insects of medical importance via detection of disease

Table 1 Complementary strengths and weaknesses of traditional, manipulative experimentation and ecoinformatics methods for agricultural pest management research

Project stage	Attribute	Experimental approaches	Ecoinformatics approaches
Building a data set	Data uniformity, completeness, quality	**Researcher has direct control of data collection**	Data collection is decentralized, biases may be present in population sampled
	Flexibility	**Any variable the researcher can manipulate can be examined, creating opportunities to evaluate novel conditions not manifest in the field**	Only conditions already present in the field can be evaluated
	Privacy concerns	**Data are often collected on research farms, sidestepping privacy issues**	Farmers or other data holders may be unwilling to share data
	Spatial and temporal scale of the data	Often much smaller than the scale of farming or other process being examined	**Often matches the actual scale of farming or process being examined**
	Size of resulting data sets	Data sets are often too small to resolve the effects that dictate farmer decisions (e.g., yield)	**Data sets may be 10–100 times larger than experimental data sets, boosting statistical power**
	Cost efficiency	Labor costs of data collection are high	**Preexisting data can be used at small cost**
Analysis and inference	Ability to evaluate many variables	Experiments rarely examine more than a few variables at once	**May be particularly valuable when many variables must be screened**
	Between-replicate variation	**Reduced between-replicate variation increases statistical power**	Data sets are often noisy, decreasing statistical power
	Causal inference	**Stronger**	Weaker
From research results to adoption of new recommendations by farmers	Ease of extending research results to implementation by farmers	Experimental research is often conducted in off-farm settings, divorcing researchers from farmers	**When data come from farmers, farmers can be involved from the start**
		Results may only be valid under conditions of the experiment	**Data sets can embrace the full range of farming conditions**
		Farmers are sometimes skeptical of results conducted in small, experimental research plots	**Farmers may have greater confidence in recommendations emerging from their own data**

Within each row, the entry in boldface has the more desired characteristics.

occurrences recorded by health agencies (15, 68) or, more recently, through Internet search behaviors (49). Pesticide-use patterns obtained from farmer surveys can also be used as a proxy for pest activity (78, 79, 89, 90).

2.1.3. Citizen science. The Internet has greatly facilitated connections between scientists and amateur naturalists, generating data on insect occurrences at broad spatial and temporal scales (18, 37). Citizen-based monitoring can work well for easily identifiable insects such as butterflies and bees (**http://www.naba.org/butter_counts.html**) (58, 86, 96, 108), lady beetles (47), and moths (44). A recent development is Internet-enabled portals for data collection that make it easy for amateur naturalists to upload images or observations on plants and animals (**http://bugguide.net**, **http://www.bumblebeewatch.org**, **http://iNaturalist.org**, **http://www.projectnoah.org**, **https://www.usanpn.org**) (142). Another form of citizen-driven entomological data collection is the crowdsourcing of transcription or interpretation of archived records such as museum labels (**https://www.zooniverse.org**) (57).

2.1.4. Academic data. An important additional source of information is data already published by other researchers. Until recently, many data sets analyzed in published papers were difficult to obtain in their raw form, but this is rapidly changing (e.g., **https://www.idigbio.org**, **https://www.dataone.org**, **http://vegbank.org**, **http://datadryad.org**, **http://www.gbif.org**). Calls by academics, professional societies, journals, and funding agencies to improve data sharing will ultimately make data more available for reuse (**http://www.nature.com/sdata**, **http://esapubs.org/archive**) (53, 59). This should expand the availability of experimental, rather than observational, data sets for ecoinformatics analyses.

2.2. Building a Data Set

Ecoinformatics data sets often differ in several key respects from the more familiar data sets that researchers gather with their own hands.

2.2.1. Data quality. Given the diversity of entomological data sources and collection methods, it is not surprising that ecoinformatics data sets may be highly heterogeneous. Some ecoinformatics data sets may not reach the quality standards expected by researchers (29, 47, 77, 124). Ecoinformatics data sets that combine multiple sources of data that use different sampling methods can create especially difficult problems (109). An obvious advantage of researcher-led studies is that there is better control over data collection, and thus, one can achieve high standards of data completeness, uniformity, and quality, while implementing protocols that minimize biases.

2.2.2. Flexibility. Experimental methods also have a key advantage in their tremendous flexibility: As long as the experimentalist can implement the manipulations needed to create a condition of interest, any situation can be studied. In contrast, observational data sets are limited to studying current farming practices and simply cannot address novel, not-yet-adopted methods or ranges of variation not commonly observed in commercial settings (80, 111).

2.2.3. Privacy concerns. Experimental methods are less affected by a problem that can be paramount in gathering ecoinformatics data: data privacy. Publication of data gathered for ecoinformatics, if not done thoughtfully, could impinge on the personal privacy of farmers or other agricultural professionals who may not have known their information was being collected. In addition, farmers may not be eager to share information on yield or details of crop or pest management, as such information may be viewed as strictly proprietary (29).

2.2.4. Size of data sets. Ecoinformatics data sets are often quite large, placing ecoinformatics at least partially under the umbrella of big data research methods. Our survey of the agricultural entomology ecoinformatics literature revealed that data sets average nearly 10,000 replicate observations {**Figure 1*a***; mean number of observations = 9,934 ± 3,795 [standard error (SE)] (range: 20–290,101)}. Although we did not attempt to produce a comparable sample of experimental studies, it is clear that ecoinformatics data sets are orders of magnitude larger than typical experimental data sets (e.g., see review in 112). Use of preexisting data offers substantial cost

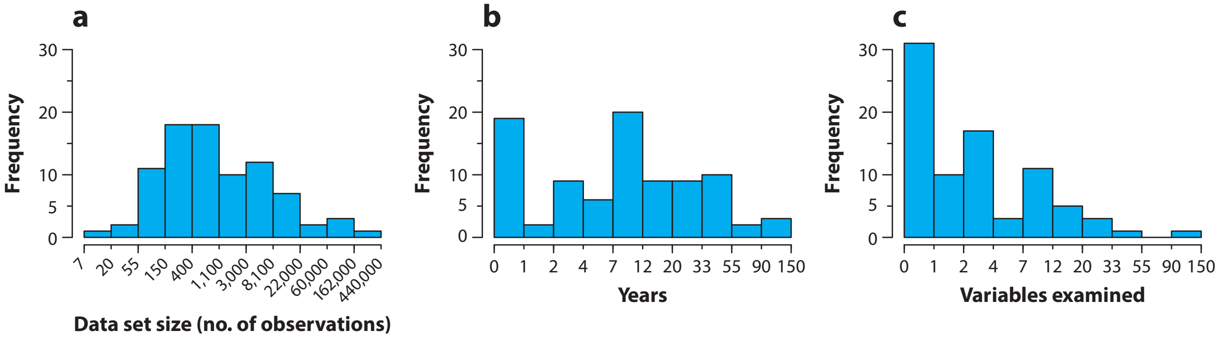

Figure 1

Survey of published studies using ecoinformatics methods in agricultural entomology. Shown are (*a*) the size of the data set assembled (i.e., the number of replicate observations included in the main data set), (*b*) the number of years for which observations were available, and (*c*) the number of explanatory factors examined for possible influences on the response variable of interest (note log scales on the x axes).

savings, and it is these cost savings that make it possible to assemble larger data sets (29, 47, 66, 131). Although experimentation can, in principle, produce data sets of any size, large data sets are costly; for example, a rare experimental study that achieved levels of replication and spatial coverage that approached that seen in ecoinformatics studies was achieved with a budget of £6 million (≈US$8.8 million) (129).

2.2.5. Temporal and spatial scales of data sets. Some processes, such as the influence of the landscape on colonization of crops, the effects of multiyear crop rotations on pest densities, or effects of climate, cannot be studied effectively in small-scale or short-term experimental plots. Ecoinformatics data sets cover an average of 20.5 (mean) ± 2.8 (SE) years (range: 1–140) (**Figure 1*b***), including several multidecadal data sets. These are, in general, much longer than typical researcher-led data sets in agricultural entomology or plant-herbivore-predator interactions more generally (112, 117). Ecoinformatics studies may also reach regional (89), continental, or even global (10, 48) spatial scales.

2.3. Statistical Considerations, Inference, and Pitfalls

It is easy for observational studies with large sample sizes to create the illusion of power and validity when, in fact, errors in measurement, selection bias, and unexplained confounding factors can undermine interpretations (19, 125). These lessons have been hard learned in medical epidemiology (134), where analysis of large observational data sets has a longer history than in agricultural entomology. Similarly, ecoinformatics methods primarily use observational data; consequently, observed associations between any two variables need to be interpreted cautiously. The axiom that "correlation does not imply causation" does not disappear just because a data set is large (17). Here we give examples of potential pitfalls in working with ecoinformatics data to demonstrate the level of vigilance that is required and to show that, in some cases, statistical remedies exist for these problems (see also 125).

2.3.1. Statistical power. One of the advantages of using larger data sets is the opportunity to detect small effect sizes even when the underlying data are noisy. Statistical power analyses show that pest management research often demands surprisingly large sample sizes to resolve key effects successfully, even when effect sizes are large (135). Because insecticides are generally inexpensive relative to crop value, profit-maximizing farmers are often motivated to suppress pest populations when only very small yield losses (typically ≈2%) are threatened. Such small effects generally cannot be resolved with traditional experimentation (112) but can be characterized using larger ecoinformatics data sets (111).

Nevertheless, caution must be taken in using small *p* values alone as sufficient evidence to reject a null hypothesis and establish a biologically meaningful finding (103). Rather, use of multiple diagnostics, including confidence intervals and estimates of effect size, will give a more robust sense of the importance of the findings (103, 119). Incorporating expert information on prior expectations could also be a valuable way to improve our understanding of patterns in the data (67).

2.3.2. Bias. One of the most common sources of interpretational difficulty comes when samples are selected nonrandomly from a larger population, creating a potential for bias in the response being studied. Selection bias may frequently occur when a pest control method (e.g., cultural, chemical, or biological control) is implemented nonrandomly across a population of crop fields or farmers (e.g., 98). For example, Kathage & Qaim (71) found that more progressive cotton farmers were more apt to adopt new *Bt* cultivars. Such farmers, however, might also be those likely to produce higher yields, even without any boost conferred by the *Bt* cultivar, potentially creating a spurious association between *Bt* cotton and high yields. This makes it difficult to understand the true causal influence of genetically modified crops on yield. Another common example is when the pests themselves express bias in their use of plants, making it difficult to understand the average effect of pests on the entire population of plants. For example, pests may prefer high-vigor, high-yielding host plants, which could result in a spurious association between pests and high yield and an incorrect inference that plants overcompensate for herbivory (see discussion in 146). In general, these biases, when not accounted for, could give an erroneous picture of the average response of the population to a set of influencing variables and can significantly affect the interpretation of a study, especially when the effect sizes are small (119).

It is possible to remedy such situations by matching populations of like groups and comparing them only with respect to characteristics of interest (75; see also 134) or by including additional factors as covariates in multiple regression models (see also the instrumental variable approach used in 75, 82).

2.3.3. Number of factors examined. Ecoinformatics methods, and observational methods in general, have a particular advantage in permitting researchers to explore associations between many potentially influential factors and key response variables, in contrast to experiments, which are typically limited to one or few manipulated factors.

Whereas many ecoinformatics studies are still narrowly focused on just a single variable, several studies captured data on substantially larger sets of variables (**Figure 1*c***; mean number of explanatory variables examined = 7.9 ± 1.8; range: 1–132). In the early phase of a research program that is examining a poorly defined problem, it can be especially helpful to include many possible explanatory factors. The resulting exploratory analyses can help to formulate more focused hypotheses that can then be examined in follow-up studies, including with experimentation (27, 29, 80, 91, 98, 120, 131). Data sets with many variables provide opportunities to explore potentially important interactions that are difficult to implement in manipulative studies. How best to explore these multidimensional data sets is an active area of inquiry (11, 50, 64, 91, 92, 101, 131).

There are challenges, however, in including a large number of potential response and predictor variables in statistical models. Studies that include multiple variables that are themselves highly correlated (multicollinearity) can create severe problems of interpretation, which are difficult or impossible to resolve through statistical means alone (41).

Spurious correlations can emerge from many other sources, recognized and unrecognized. For example, variable weather can often drive changes in both pest densities and crop performance, creating almost ubiquitous opportunities for spurious correlations. Inclusion of key weather variables as covariates can insulate the researcher against these problems. In some cases, important unmeasured variables can generate patterns of spatial autocorrelation that can distort the results of statistical modeling (9, 38); in these cases, analyses that model the spatial structure of the data are needed (e.g., 89). Unrecognized (and therefore uncorrected) sources of spurious correlations are the worst enemies of ecoinformatics methods, as they can create serious errors of interpretation.

2.3.4. Correlation and causation. Scientists have long debated how best to draw inferences of causality (39, 73). In some disciplines such as physics, medicine, oceanography, and astronomy, observational approaches play a central role in research (113, 114). In agricultural entomology, however, an important goal is to develop research-driven recommendations that allow farmers to implement management actions that result in desired outcomes (e.g., pest suppression). Only knowledge of causal relationships can inform the farmer of the likely consequences of a particular action. The power of well-designed manipulative experiments lies in their ability to circumvent the pitfalls associated with observational studies (82). Thus, we reject the suggestion of the most avid proponents of big data methodologies that knowledge of correlation alone can fully replace knowledge of causation as our primary research goal (3, 87). As emphasized by Harford (55), analysis based on pure correlations is "fragile," because we often cannot anticipate what might cause the correlation to break down and because spurious correlations always lurk as a threat (29). For this reason, above all others, ecoinformatics will always be most valuable when used in tight partnership with experimentation.

2.4. From Research Results to Adoption

When stakeholders, such as farmers or independent consultants, are the sources of ecoinformatics data sets, they can be engaged in the research endeavor from the very outset of a project, facilitating the integration of research and outreach. Farmers may have greater confidence in recommendations that emerge from analyses of their own data, rather than small-plot research conducted at a university farm (29). In addition, whereas experimental research is often performed in narrowly controlled and agronomically optimal settings, ecoinformatics research can embrace the full range of commercial farming conditions (141). Data sets that purposefully encompass a range of heterogeneity also create opportunities to produce site-specific recommendations (67).

3. CONTRIBUTIONS OF ECOINFORMATICS STUDIES TO AGRICULTURAL ENTOMOLOGY

Here we highlight research that has contributed to agricultural entomology using ecoinformatics approaches. The ecoinformatics literature in agricultural entomology is diffuse, making it difficult to do a systematic review. This is due in part to the long history of these research approaches, the diversity of questions that researchers have addressed (and subdisciplines in which they occur), and the creativity of approaches that have been used to explore observational data sets. Because the use of ecoinformatics methods has been conducted without a common methodological framework, and with minimal sharing of key methodological lessons learned by different research groups, researchers have in many cases been forced to reinvent key techniques. We hope that this review will help to integrate the field, accelerating progress.

3.1. Documenting Pest and Disease Patterns

Ecoinformatics approaches have been especially fruitful in studying pest outbreaks, which often occur at irregular intervals and are also heterogeneous in space. Collection of data over larger temporal and spatial extents is often key for understanding underlying factors that drive pest and disease dynamics. Coordinated data collection has been conducted by academic or federal and state agencies such as the National Science Foundation–sponsored Long-Term Ecological Research (LTER) sites (e.g., 6) and, more recently, the National Ecological Observatory Network (69). The Rothamsted suction trap network has operated continuously since 1964 throughout the United Kingdom to help understand relationships between long-term climate oscillations and land cover/land use changes and the abundances and flight activity periods of aphids (see also 13, 14, 31). This sampling approach has been replicated in the midwestern United States (116) and in mainland Europe, where Steinger et al. (127) related increases in potato virus Y incidence in potatoes to aphid flight activity patterns. Ewald et al. (40) used long-term and highly replicated Game and Wildlife Conservation Trust sampling in the southern United Kingdom to show how insect populations quickly rebounded in grain fields in years after extreme weather events.

Ecoinformatics methods have been used to examine spatial and temporal scales of population fluctuations (20, 45), long-term trends in pest densities (138), and localized and region-wide synchrony in pest populations (36). Information on forest insect pests includes some of the longest time series and spatially extensive insect information available. Aerial surveys or detailed inventories of forest damage have been used to study landscape, climatic, and environmental correlates of forest pest outbreaks; long-term population trends and cyclic dynamics; and associated economic losses (2, 32, 54, 105, 143, 144).

3.2. Efficacy of Transgenic Crops for Pest Control

A large body of published studies in the entomology, agronomy, and economics literatures have used ecoinformatics methods to quantify the consequences of transgenic crops (rice, cotton, corn) on agricultural systems. The effect of *Bt* crops in particular on target and nontarget pests, as well as pesticide use, labor costs of pest control, crop yield, crop profitability, the incidence of acute pesticide poisonings, and other externalities, including predator densities, farmland biodiversity, and soil and groundwater quality, have been examined using observational data (26, 27, 61, 75, 84, 85, 104, 106, 107, 118, 139). Analysis of long-term and spatially extensive data sets revealed that large-scale adoption of transgenic *Bt* crops produced dramatic regional suppression of pest populations: *Pectinophora gossypiella* in the United States (25) and China (140), *Helicoverpa armigera* in China (145) and India (71), and *Ostrinia nubilalis* in the United States (62).

3.3. Landscape Context Effects on Crop Colonization by Pests

Ecoinformatics studies have been used to identify how the surrounding landscape influences pest colonization of crops (123). Higbee & Siegel (56) identified pistachio orchards as a key source of *Amyelois transitella* moths colonizing almonds, and Parsa et al. (98, 99) identified potato storage units as the primary source of *Premnotrypes* spp. weevils colonizing potatoes in the Andes. Using crop scout–provided data for the Central Valley of California, Sivakoff et al. (120) and Meisner et al. (92) documented associations between specific surrounding crops and densities of *Lygus hesperus* colonizing a focal cotton field.

Machine learning: type of artificial intelligence that uses algorithms that allow computer programs to iteratively learn from data

3.4. Pest Impact on Crop Yield and Patterns in Pesticide Use

Spatially extensive data obtained from pest management consultants have been used to examine the associations between cotton pests and yield losses. Results revealed that California farmers were managing *L. hesperus* in cotton suboptimally: Growers were sustaining yield losses during the early season by not suppressing pest populations sufficiently but overapplying pesticides during the mid-season, when cotton could compensate fully for damage (92, 110, 111).

Pesticide-use data can also be an indirect indicator of pest activity in agricultural landscapes. Ecoinformatics approaches have been used to evaluate the efficacy of traditional pesticides applied under field conditions (98, 99). Farm-scale data have also been used to characterize the likelihood that applications of broad-spectrum insecticides will trigger secondary pest outbreaks in California cotton (51) and walnuts (128). In the midwestern United States, it has been shown that at the county level, replacement of natural plant communities with cropland (i.e., landscape simplification) is positively correlated with pest aphid abundance in suction traps and with increased pesticide use (90)—a pattern that has also been borne out at the national level (79, 89, 147). Several studies have also shown that differences between farmers are a key source of variation in overall pesticide use, rather than regional or between-year effects (4, 89). Nevertheless, the ability to resolve these associations consistently requires careful statistical corrections for issues of spatial autocorrelation that can otherwise obscure the true effects of explanatory variables (89).

3.5. Beneficial Insects

Insects beneficial to agriculture have been the subjects of long-term monitoring. Records of unmanaged bees derived primarily from historical museum collections have been used to demonstrate the declines of particular groups (8, 22, 115). Smyth et al. (121) harnessed the power of distributed continent-wide citizen science to describe the range restriction of an exotic coccinellid, and Bahlai and colleagues (6, 7) used a 24-year data set of lady beetle collections to evaluate how changing agricultural practices influence species turnover patterns. All of these studies gained insights on long-term dynamics of communities and on abundance and distribution of species from archived, long-term data collected by others.

Contributions of beneficial insects to agricultural production have also been explored. Using a grower survey of crop yields for an 11-year period, Gaines-Day & Gratton (46) studied the relationship between honey bee stocking density and farm-level cranberry yield. Data from the United Nations Food and Agriculture Organization and other governmental sources have been used to evaluate honey bee colony population dynamics (95) to quantify the reliance of global food production on pollinators (1) and to link dependency on pollinators with shortfalls of crop yield improvements, elevated variance in yield, and reduced yield responses to agricultural intensification (35, 48).

3.6. Food Webs

Understanding the complexities of trophic interactions among organisms in diverse ecosystems has always been a challenge. Bohan et al. (16) and Tamaddoni-Nezhad et al. (131) described an innovative approach to automating the construction of agricultural food webs. Applying machine learning methods to a previously collected data set of predator and prey densities and using automated text mining of the published literature to corroborate proposed trophic linkages, they built a trophic web with 72 nodes and 407 links. Bohan et al. (16) suggested that the primary value of this analysis was in generating hypotheses of novel, unsuspected trophic linkages. Indeed, in their food web, the most striking novel links involving intraguild predation between carabid beetles and spiders were later confirmed with DNA analysis of beetle stomach contents (34). This approach illustrates the way in which insights from observational data can stimulate experimentation to test specific proposed hypotheses.

3.7. Efficacy of Cultural Controls and Host-Plant Resistance

Management of insect pests using cultural practices and host-plant resistance has also been studied using ecoinformatics. For example, Parsa et al. (100) synthesized decades of field trials of cassava genotypes to identify traits, including leaf pubescence and root hydrogen cyanide, shaping resistance to three cassava pests. Studies of the effects of crop rotation on insect pests demonstrated the yield-enhancing effects of single- and multiyear crop rotation and also revealed that the exact identity of the rotation crop matters (27, 91, 118). Carrière et al. (24) mined pink bollworm pheromone trapping data to model how planting date could be adjusted to protect cotton from early-season attack, and Higbee & Siegel (56) used data on almond infestation to calculate levels of orchard sanitation (removal of mummy nuts harboring overwintering *A. transitella* larvae) required to keep nut damage below an economic threshold.

3.8. Farmer Decision Making

In addition to examining the effects of management or environmental factors on pest and beneficial insects, ecoinformatics methods have also been used to explore factors shaping farmers' pest management decisions. For example, studies in China found that pesticide use was strongly influenced by farmers' anticipation of pest losses, preferences for pesticides that are cheaper and less likely to poison workers, and risk aversion (60, 61, 75, 83, 139). Outreach by extension service agents was generally unimportant in decisions to apply pesticides. The failure of extension service agents to reduce pesticide use was hypothesized to stem from incentives provided to extension agents, whose salaries are augmented by commissions earned on pesticide sales.

4. TEN-POINT CHECKLIST FOR ECOINFORMATICS STUDIES

The use of ecoinformatics in agricultural entomology suggests some clear advantages of these approaches but also highlights some important challenges. In summarizing the key themes found in ecoinformatics studies in entomology, we propose ten best research practices to help avoid common pitfalls and to capitalize fully on the potential of ecoinformatics-based research. Although many of these guidelines are useful for any study that uses observational and correlative approaches, several data-related issues (practices 2–6) are especially relevant to ecoinformatics studies (**Table 2**).

5. RESEARCH NEEDS AND FUTURE OPPORTUNITIES

A broad community of researchers has used ecoinformatics methods to address a diverse array of questions in agricultural entomology. Some research questions, where the spatial or temporal extents of the underlying processes make them experimentally intractable or where effect sizes are expected to be small, have been particularly amenable to investigation through ecoinformatics

Table 2 Ten-point best-practices checklist for ecoinformatics studies

Point	Topic	Actions and issues confronted
1	Identify a research question	Determine whether study is exploratory or has well-defined hypotheses. Open-ended questions may result in a high number of variables or models examined, increasing the chance of spurious relationships being detected, but offer opportunities for hypothesis generation
2	Identify the primary preexisting data set(s)	Data owners may be more likely to share data, even if privacy is a concern, if they see clear value in answering the focal research questions
3	Seek out complementary data streams	The value of ecoinformatics studies often emerges from integrating multiple, disparate sources of data
4	Assess data privacy issues	Obtain appropriate permissions and establish protocols for maintaining data anonymity
5	Understand the data set and evaluate potential biases	Understand how data were obtained. If possible, design a sampling approach for data gathering, including randomized subsampling or stratification, such that the data set is representative of the target population
6	Design data management and workflows	Create system for data entry, quality control, standardization, error detection, data exploration, and creation of summary statistics that can be automated. Deposit data sets in data repositories using standardized metadata
7	Clearly define response and explanatory variables and evaluate potential confounding influences	Carefully consider the potential importance of confounding variables and include measurements of possible confounders in statistical modeling. The importance of this point cannot be overstated
8	Select an appropriate framework for statistical analysis	Data exploration, formal modeling, and hypothesis testing should be consistent with the question being addressed and the nature of the underlying data structure
9	Conclusions and inferences	Frame conclusions carefully, including a critical assessment of competing interpretations and the influence of any suspected confounders that could not be measured. Expert opinion can be valuable in this step
10	Integrate findings with other studies	Whenever possible, integrate correlational analyses of ecoinformatics data sets with experimental studies to establish support for causal relationships

approaches. In combination with expert opinion, ecoinformatics tools can give insights into relationships that were not conceivable at the onset of studies (e.g., 127). Although farmers may be unusually receptive to research-driven recommendations derived from data gathered from the true setting of commercial agriculture, most farmers are also entirely unfamiliar with ecoinformatics research methods, creating challenges for outreach. Extension specialists will need to explain ecoinformatics research and create outreach tools that maximize the utility of ecoinformatics analyses—for example, in producing site-specific recommendations.

At the same time, we need to acknowledge the potential pitfalls associated with correlative analyses and, in particular, the difficulties of distinguishing associations that reflect true causal relationships from associations that merely reflect spurious correlations. Ecoinformatics methods used uncritically could easily do more harm than good in entomological research. Ecoinformatics methods can be used to generate hypotheses that can then be tested with focused experimentation, combining the best of both worlds. Another valuable approach will be to take researcher-generated (small data) data sets, and their inherent advantages, and link them together to achieve the advantages of large data sets (74).

Ecoinformatics and big data approaches in applied entomological research are not new, will not go away, and will continue to improve over time. The entomologist of the future working on applied questions will have to be skilled at designing and implementing experimental studies as well as have training in quantitative methods needed to work with observational data. Enabling the data revolution in entomological research will require that we embrace a culture of data sharing, understand the limitations of observational research, and collaborate with others in diverse areas such as computer science, statistics, and engineering in order to understand the causes and consequences of insect pests and how best to manage them.

Bioinformatics: field that combines elements of biology and computer science to understand and organize large volumes of macromolecule data

FUTURE ISSUES

1. Embrace the advantages of ecoinformatics-based insights by combining ecoinformatics methods with experimentation that tests mechanistic hypotheses and establishes causal relationships (29, 84).
2. Collaborate with biostatisticians to address analytical challenges and strengthen the interpretation of observational data (63, 103, 125).
3. Develop data sources that take advantage of citizen science and crowd-sourced data collection to increase the rate and spatial extent of data gathering (18, 37, 42).
4. Work with engineers and computer scientists to create novel ways of automating the detection and identification of arthropods and their activities in agriculture (28, 65, 102).
5. Develop mobile platforms operating pest management software applications for use in agriculture that facilitate rapid data digitization, uploading to centralized databases, and availability for ecoinformatics analyses and in-season management recommendations (30, 43).
6. Borrow and adapt approaches used in bioinformatics research to create a cyberinfrastructure to store, retrieve, and share ecoinformatics data (53, 59, 101).
7. Work with stakeholders in entomological subdisciplines to create data collection protocols and platforms that enable researcher-developed data sets to be standardized and collated into exchangeable forms (74, 93).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank Bodil Cass, James Cock, Murray Clayton, Tania Kim, Tim Meehan, Nancy Schellhorn, and Brian Spiesman for valuable feedback on an earlier draft of this review, and Karen Dunn for help in tracking down older references. This work was supported in part by the US Department of Agriculture National Institute of Food and Agriculture grants 2011-67009-30022 and 2015-06432, McMaster Fellowship (CSIRO Australia), US DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494), and the DOE Office of Energy Efficiency and Renewable Energy (DE-AC05-76RL01830).

LITERATURE CITED

1. Aizen MA, Garibaldi LA, Cunningham SA, Klein AM. 2009. How much does agriculture depend on pollinators? Lessons from long-term trends in crop production. *Ann. Bot.* 103(9):1579–88

2. Allstadt AJ, Haynes KJ, Liebhold AM, Johnson DM. 2013. Long-term shifts in the cyclicity of outbreaks of a forest-defoliating insect. *Oecologia* 172(1):141–51
3. Anderson C. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired*, June 23. **https://www.wired.com/2008/06/pb-theory/**
4. Andert S, Bürger J, Gerowitt B. 2015. On-farm pesticide use in four Northern German regions as influenced by farm and production conditions. *Crop Prot*. 75:1–10
5. Bacon SJ, Bacher S, Aebi A. 2012. Gaps in border controls are related to quarantine alien insect invasions in Europe. *PLOS ONE* 7(10):e47689
6. Bahlai CA, Colunga-Garcia M, Gage SH, Landis DA. 2014. The role of exotic ladybeetles in the decline of native ladybeetle populations: evidence from long-term monitoring. *Biol. Invasions* 17(4):1005–24
7. Bahlai CA, van der Werf W, O'Neal M, Hemerik L, Landis DA. 2015. Shifts in dynamic regime of an invasive lady beetle are linked to the invasion and insecticidal management of its prey. *Ecol. Appl.* 25(7):1807–18
8. Bartomeus I, Ascher JS, Gibbs J, Danforth BN, Wagner DL, et al. 2013. Historical changes in northeastern US bee pollinators related to shared ecological traits. *PNAS* 110(12):4656–60
9. Beale CM, Lennon JJ, Yearsley JM, Brewer MJ, Elston DA. 2010. Regression analysis of spatial data. *Ecol. Lett.* 13(2):246–64
10. **Bebber DP, Ramotowski M, Gurr SJ. 2013. Crop pests and pathogens move polewards in a warming world. *Nat. Clim. Change* 3(11):985–88**

10. Large, decadal, global database used to examine geographic range expansion by insect and fungal pests.

11. Behmann J, Mahlein A-K, Rumpf T, Römer C, Plümer L. 2014. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precis. Agric.* 16(3):239–60
12. Bekker RM, van der Maarel E, Bruelheide H, Woods K. 2007. Long-term datasets: from descriptive to predictive data using ecoinformatics. *J. Veg. Sci.* 18(4):458–62
13. Bell JR, Alderson L, Izera D, Kruger T, Parker S, et al. 2015. Long-term phenological trends, species accumulation rates, aphid traits and climate: five decades of change in migrating aphids. *J. Anim. Ecol.* 84(1):21–34
14. Benton T, Bryant D, Cole L, Crick H. 2002. Linking agricultural practice to insect and bird populations: a historical study over three decades. *J. Appl. Ecol.* 39(4):673–87
15. Bisanzio D, Mutuku F, LaBeaud AD, Mungai PL, Muinde J, et al. 2015. Use of prospective hospital surveillance data to define spatiotemporal heterogeneity of malaria risk in coastal Kenya. *Malar. J.* 14(1):482
16. Bohan DA, Caron-Lormier G, Muggleton S, Raybould A, Tamaddoni-Nezhad A. 2011. Automated discovery of food webs from ecological data using logic-based machine learning. *PLOS ONE* 6(12):e29028
17. Bollier D. 2010. *The Promise and Peril of Big Data*. Washington, DC: Aspen Inst. Commun. Soc. Progr.
18. Bonney R, Cooper CB, Dickinson J, Kelling S, Phillips T, et al. 2009. Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience* 59(11):977–84
19. boyd d, Crawford K. 2012. Critical questions for big data. *Inform. Comm. Soc.* 15(5):662–79
20. Brooks D, Perry JN, Clark S, Heard M, Firbank L, et al. 2008. National scale metacommunity dynamics of carabid beetles in UK farmland. *J. Anim. Ecol.* 77:265–74
21. Brown M, Pharo H, Hendrikx P, Ribière-Chabert M, Chauzat MP, Marris G. 2014. Epidemiological surveillance and control of bee diseases. In *Bee Health and Veterinarians*, ed. W Ritter, pp. 193–213. Paris: OIE
22. Cameron SA, Lozier JD, Strange JP, Koch JB, Cordes N, et al. 2011. Patterns of widespread decline in North American bumble bees. *PNAS* 108(2):662–67
23. Carpenter JR. 1940. Insect outbreaks in Europe. *J. Anim. Ecol.* 9(1):108–47
24. Carrière Y, Ellers-Kirk C, Pedersen B, Haller S, Antilla L. 2001. Predicting spring moth emergence in the pink bollworm (Lepidoptera: Gelechiidae): implications for managing resistance to transgenic cotton. *J. Econ. Entomol.* 94(5):1012–21
25. Carrière Y, Ellers-Kirk C, Sisterson M, Antilla L, Whitlow M, et al. 2003. Long-term regional suppression of pink bollworm by *Bacillus thuringiensis* cotton. *PNAS* 100(4):1519–23
26. Cattaneo MG, Yafuso C, Schmidt C, Huang C, Rahman M, et al. 2006. Farm-scale evaluation of the impacts of transgenic cotton on biodiversity, pesticide use, and yield. *PNAS* 103(20):7571–76

27. Chavas J-P, Shi G, Lauer J. 2014. The effects of GM technology on maize yield. *Crop Sci*. 54(4):1331
28. Chen Y, Why A, Batista G, Mafra-Neto A, Keogh E. 2014. Flying insect classification with inexpensive sensors. *J. Insect Behav*. 27(5):657–77
29. Cock J, Oberthür T, Isaacs C, Läderach PR, Palma A, et al. 2011. Crop management based on field observations: case studies in sugarcane and coffee. *Agric. Syst*. 104(9):755–69
30. Cohen Y, Cohen A, Hetzroni A, Alchanatis V, Broday D, et al. 2008. Spatial decision support system for medfly control in citrus. *Comput. Electron. Agric.* 62(2):107–17
31. Conrad KF, Warren MS, Fox R, Parsons MS, Woiwod IP. 2006. Rapid declines of common, widespread British moths provide evidence of an insect biodiversity crisis. *Biol. Conserv*. 132(3):279–91
32. Cooke BJ, Roland J. 2000. Spatial analysis of large-scale patterns of forest tent caterpillar outbreaks. *Ecoscience* 7(4):410–22
33. Davenport TH, Patil DJ. 2012. Data scientist: the sexiest job of the 21st century. *Harv. Bus. Rev.* October:70–76
34. Davey JS, Vaughan IP, King RA, Bell JR, Bohan DA, et al. 2013. Intraguild predation in winter wheat: prey choice by a common epigeal carabid consuming spiders. *J. Appl. Ecol*. 50(1):271–79
35. Deguines N, Jono C, Baude M, Henry M, Julliard R, Fontaine C. 2014. Large-scale trade-off between agricultural intensification and crop pollination services. *Front. Ecol. Environ.* 12(4):212–17
36. de Valpine P, Scranton K, Ohmart CP. 2010. Synchrony of population dynamics of two vineyard arthropods occurs at multiple spatial and temporal scales. *Ecol. Appl.* 20(7):1926–35
37. Dickinson JL, Shirk J, Bonter D, Bonney R, Crain RL, et al. 2012. The current state of citizen science as a tool for ecological research and public engagement. *Front. Ecol. Environ.* 10(6):291–97
38. Dormann CF, McPherson J, Araújo M, Bivand R, Bolliger J, et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30(5):609–28
39. **Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, et al. 2015. Big data, bigger dilemmas: a critical review. *J. Assoc. Inf. Sci. Technol.* 66(8):1523–45**

39. Excellent overview of epistemological, social, and scientific challenges in Big Data.

40. Ewald JA, Wheatley CJ, Aebischer NJ, Moreby SJ, Duffield SJ, et al. 2015. Influences of extreme weather, climate and pesticide use on invertebrates in cereal fields over 42 years. *Glob. Change Biol.* 21(11):3931–50
41. Fieberg J, Johnson DH. 2015. MMI: multimodel inference or models with management implications? *J. Wildl. Manag.* 79(5):708–18
42. Flockhart DTT, Wassenaar LI, Martin TG, Hobson KA, Wunder MB, Norris DR. 2013. Tracking multi-generational colonization of the breeding grounds by monarch butterflies in eastern North America. *Proc. R. Soc. B* 280(1768):20131087
43. Fountas S, Carli G, Sorensen CG, Tsiropoulos Z, Cavalaris C, et al. 2015. Farm management information systems: current situation and future perspectives. *Comput. Electron. Agric.* 115:40–50
44. Fox R, Randle Z, Hill L, Anders S, Wiffen L, Parsons MS. 2011. Moths count: recording moths for conservation in the UK. *J. Insect Conserv*. 15(1):55–68
45. Frost KE, Esker PD, Van Haren R, Kotolski L, Groves RL. 2013. Factors influencing aster leafhopper (Hemiptera: Cicadellidae) abundance and aster yellows phytoplasma infectivity in Wisconsin carrot fields. *Environ. Entomol.* 42(3):477–90
46. Gaines-Day HR, Gratton C. 2016. Crop yield is correlated with honey bee hive density but not in high-woodland landscapes. *Agric. Ecosyst. Environ.* 218:53–57
47. Gardiner MM, Allee LL, Brown PM, Losey JE, Roy HE, Smyth RR. 2012. Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. *Front. Ecol. Environ.* 10(9):471–76
48. Garibaldi LA, Aizen MA, Klein AM, Cunningham SA, Harder LD. 2011. Global growth and stability of agricultural yield decrease with pollinator dependence. *PNAS* 108(14):5909–14
49. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. 2014. Evaluation of Internet-based dengue query data: Google Dengue Trends. *PLOS Negl. Trop. Dis.* 8(2):e2713
50. Glymour C, Scheines R, Spirtes P. 2014. *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. New York: Academic
51. Gross K, Rosenheim JA. 2011. Quantifying secondary pest outbreaks in cotton and their monetary cost with causal-inference statistics. *Ecol. Appl.* 21(7):2770–80

52. Haack RA. 2006. Exotic bark-and wood-boring Coleoptera in the United States: recent establishments and interceptions. *Can. J. For. Res.* 36(2):269–88

53. Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, et al. 2013. Big data and the future of ecology. *Front. Ecol. Environ.* 11(3):156–62

54. **Hanewinkel M, Breidenbach J, Neeff T, Kublin E. 2008. Seventy-seven years of natural disturbances in a mountain forest area—the influence of storm, snow, and insect damage analysed with a long-term time series. *Can. J. For. Res.* 38(8):2249–61**

54. Long-term forest inventory study uses temporal analyses to understand extent and drivers of insect damage.

55. Harford T. 2014. Big data: a big mistake? *Significance* 11(5):14–19

56. Higbee B, Siegel J. 2009. New navel orangeworm sanitation standards could reduce almond damage. *Calif. Agric.* 63(1):24–28

57. Hill A, Guralnick R, Smith A, Sallans A, Gillespie R, et al. 2012. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. *ZooKeys* 209:219–33

58. Howard E, Davis AK. 2015. Investigating long-term changes in the spring migration of monarch butterflies (Lepidoptera: Nymphalidae) using 18 years of data from Journey North, a citizen science program. *Ann. Entomol. Soc. Am.* 108(5):664–69

59. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, et al. 2008. Big data: the future of biocuration. *Nature* 455(7209):47–50

60. Huang J, Hu R, Pray C, Qiao F, Rozelle S. 2003. Biotechnology as an alternative to chemical pesticides: a case study of Bt cotton in China. *Agric. Econ.* 29(1):55–67

61. Huang J, Hu R, Rozelle S, Pray C. 2005. Insect-resistant GM rice in farmers' fields: assessing productivity and health effects in China. *Science* 308(5722):688–90

62. **Hutchison WD, Burkness EC, Mitchell PD, Moon RD, Leslie TW, et al. 2010. Areawide suppression of European corn borer with Bt maize reaps savings to non-Bt maize growers. *Science* 330(6001):222–25**

62. Elegant example of a long-term and large-scale ecoinformatics data set to study pest suppression.

63. Isaac NJB, van Strien AJ, August TA, de Zeeuw MP, Roy DB. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods Ecol. Evol.* 5(10):1052–60

64. James G, Witten D, Hastie T, Tibshirani R. 2013. *An Introduction to Statistical Learning*. New York: Springer

65. Jiang J-A, Tseng C-L, Lu F-M, Yang E-C, Wu Z-S, et al. 2008. A GSM-based remote wireless automatic monitoring system for field information: a case study for ecological monitoring of the oriental fruit fly, *Bactrocera dorsalis* (Hendel). *Comput. Electron. Agric.* 62(2):243–59

66. Jiménez D, Cock J, Jarvis A, Garcia J, Satizábal HF, et al. 2011. Interpretation of commercial production information: a case study of lulo (*Solanum quitoense*), an under-researched Andean fruit. *Agric. Syst.* 104(3):258–70

67. Jiménez D, Dorado H, Cock J, Prager SD, Delerce S, et al. 2016. From observation to information: data-driven understanding of on farm yield variation. *PLOS ONE* 11(3):e0150015

68. Kampen H, Medlock JM, Vaux AGC, Koenraadt CJM, van Vliet AJ, et al. 2015. Approaches to passive mosquito surveillance in the EU. *Parasites Vectors* 8(1):9

69. Kao RH, Gibson CM, Gallery RE, Meier CL, Barnett DT, et al. 2012. NEON terrestrial field observations: designing continental-scale, standardized sampling. *Ecosphere* 3(12):1–17

70. Kaskey J. 2013. Monsanto buying Climate Corp. to add big data for farmers. *Bloomberg*, October 3. **http://www.bloomberg.com/news/articles/2013-10-02/monsanto-to-buy-climate-corp-profit-forecast-trails-estimates**

71. **Kathage J, Qaim M. 2012. Economic impacts and impact dynamics of Bt (*Bacillus thuringiensis*) cotton in India. *PNAS* 109(29):11652–56**

71. Carefully assesses how biases can intrude into ecoinformatics analyses and describes effective statistical solutions.

72. Kelling S, Hochachka WM, Fink D, Riedewald M, Caruana R, et al. 2009. Data-intensive science: a new paradigm for biodiversity studies. *BioScience* 59(7):613–20

73. **Kitchin R. 2014. Big data, new epistemologies and paradigm shifts. *Big Data Soc.* 1(1):1–12**

73. Explores ways in which Big Data is changing the way we do science.

74. Kitchin R, Lauriault TP. 2014. Small data in the era of big data. *GeoJournal* 80(4):463–75

75. Kouser S, Qaim M. 2013. Valuing financial, health, and environmental benefits of Bt cotton in Pakistan. *Agric. Econ.* 44(3):323–35

76. Krapivin VF, Varotsos CA, Soldatov VY. 2015. *New Ecoinformatics Tools in Environmental Science: Applications and Decision-Making*. Berlin: Springer

77. Kremen C, Ullman KS, Thorp RW. 2011. Evaluating the quality of citizen-scientist data on pollinator communities. *Conserv. Biol.* 25(3):607–17

78. Larsen AE. 2013. Agricultural landscape simplification does not consistently drive insecticide use. *PNAS* 110(38):15330–35

79. Larsen AE, Gaines SD, Deschênes O. 2015. Spatiotemporal variation in the relationship between landscape simplification and insecticide use. *Ecol. Appl.* 25(7):1976–83

80. Lawes RA, Lawn RJ. 2005. Applications of industry information in sugarcane production systems. *Field Crops Res*. 92(2–3):353–63

81. Liebhold AM, Work TT, McCullough DG, Cavey JF. 2006. Airline baggage as a pathway for alien insect species invading the United States. *Am. Entomol.* 52:48–54

82. Little RJ, Rubin DB. 2000. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu. Rev. Public Health* 21:121–45

83. Liu EM, Huang J. 2013. Risk preferences and pesticide use by cotton farmers in China. *J. Dev. Econ.* 103:202–15

84. **Lu Y, Wu K, Jiang Y, Guo Y, Desneux N. 2012. Widespread adoption of Bt cotton and insecticide decrease promotes biocontrol services. *Nature* 487(7407):362–65**

84. Ideal integration of large ecoinformatics data set with smaller, researcher-collected observational and experimental data sets.

85. Lu Y, Wu K, Jiang Y, Xia B, Li P, et al. 2010. Mirid bug outbreaks in multiple crops correlated with wide-scale adoption of Bt cotton in China. *Science* 328(5982):1151–54

86. Lye GC, Osborne JL, Park KJ, Goulson D. 2012. Using citizen science to monitor *Bombus* populations in the UK: nesting ecology and relative abundance in the urban environment. *J. Insect Conserv.* 16(5):697–707

87. Mayer-Schonberger V, Cukier K. 2013. *Big Data: A Revolution That Will Change How We Live*. London: John Murray

88. McAfee A, Brynjolfsson E. 2012. Big data: the management revolution. *Harv. Bus. Rev.* 90(10):60–68

89. **Meehan TD, Gratton C. 2015. A consistent positive association between landscape simplification and insecticide use across the Midwestern US from 1997 through 2012. *Environ. Res. Lett.* 10(11):114001**

89. Demonstrates that uncorrected spatial autocorrelation in data sets can create spurious correlations that distort central results.

90. Meehan TD, Werling BP, Landis DA, Gratton C. 2011. Agricultural landscape simplification and insecticide use in the Midwestern United States. *PNAS* 108(28):11500–505

91. Meisner MH, Rosenheim JA. 2014. Ecoinformatics reveals effects of crop rotational histories on cotton yield. *PLOS ONE* 9(1):e85710

92. Meisner MH, Zaviezo T, Rosenheim JA. 2016. Landscape crop composition effects on cotton yield, *Lygus hesperus* densities and pesticide use. *Pest Manag. Sci.* In press. **https://doi.org/10.1002/ps.4290**

93. Michener WK, Jones MB. 2012. Ecoinformatics: supporting ecology as a data-intensive science. *Trends Ecol. Evol.* 27(2):85–93

94. Mooney SJ, Westreich DJ, El-Sayed AM. 2015. Commentary: epidemiology in the era of big data. *Epidemiology* 26(3):390–94

95. Moritz RFA, Erler S. 2016. Lost colonies found in a data mine: global honey trade but not pests or pesticides as a major cause of regional honeybee colony declines. *Agric. Ecosyst. Environ.* 216:44–50

96. O'Brien JM, Thorne JH, Rosenzweig ML, Shapiro AM. 2011. Once-yearly sampling for the detection of trends in biodiversity: the case of Willow Slough, California. *Biol. Conserv.* 144(7):2012–19

97. Papadopoulos NT, Plant RE, Carey JR. 2013. From trickle to flood: the large-scale, cryptic invasion of California by tropical fruit flies. *Proc. R. Soc. B* 280(1768):20131466

98. Parsa S, Ccanto R, Olivera E, Scurrah M, Alcázar J, Rosenheim JA. 2012. Explaining Andean potato weevils in relation to local and landscape features: a facilitated ecoinformatics approach. *PLOS ONE* 7(5):e36533

99. Parsa S, Ccanto R, Rosenheim JA. 2011. Resource concentration dilutes a key pest in indigenous potato agriculture. *Ecol. Appl.* 21(2):539–46

100. Parsa S, Medina C, Rodríguez V. 2015. Sources of pest resistance in cassava. *Crop Prot*. 68:79–84

101. Peters DPC, Havstad KM, Cushing J, Tweedie C, Fuentes O, Villanueva-Rosales N. 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. *Ecosphere* 5(6):67

102. Philimis P, Psimolophitis E, Hadjiyiannis S, Giusti A, Perelló J, et al. 2013. A centralised remote data collection system using automated traps for managing and controlling the population of the Mediterranean (*Ceratitis capitata*) and olive (*Dacus oleae*) fruit flies. *Proc. SPIE 8795: First Int. Conf. Remote Sens. Geoinform. Environ. (RSCy2013)*:87950X

103. Poole C. 2001. Low p-values or narrow confidence intervals: Which are more durable? *Epidemiology* 12(3):291–94

104. Pray CE, Huang J, Hu R, Rozelle S. 2002. Five years of Bt cotton in China – the benefits continue. *Plant J.* 31(4):423–30

105. Preisler HK, Hicke JA, Ager AA, Hayes JL. 2012. Climate and weather influences on spatial temporal patterns of mountain pine beetle populations in Washington and Oregon. *Ecology* 93(11):2421–34

106. Qaim M, Zilberman D. 2003. Yield effects of genetically modified crops in developing countries. *Science* 299(5608):900–2

107. Qiao F. 2015. Fifteen years of Bt cotton in China: the economic impact and its dynamics. *World Dev.* 70:177–85

108. Ries L, Oberhauser K. 2015. A citizen army for science: quantifying the contributions of citizen scientists to our understanding of monarch butterfly biology. *BioScience* 65(4):419–30

109. Rochester WA, Zalucki MP, Ward A, Miles M, Murray DAH. 2002. Testing insect movement theory: empirical analysis of pest data routinely collected from agricultural crops. *Comput. Electron. Agric.* 35(2–3):139–49

110. Rosenheim JA. 2013. Costs of *Lygus* herbivory on cotton associated with farmer decision-making: an ecoinformatics approach. *J. Econ. Entomol.* 106(3):1286–93

111. Rosenheim JA, Meisner MH. 2013. Ecoinformatics can reveal yield gaps associated with crop-pest interactions: a proof-of-concept. *PLOS ONE* 8(11):e80518

112. Rosenheim JA, Parsa S, Forbes AA, Krimmel WA, Law YH, et al. 2011. Ecoinformatics for integrated pest management: expanding the applied insect ecologist's tool-kit. *J. Econ. Entomol.* 104(2):331–42

113. Rothman KJ, Greenland S. 2005. Causation and causal inference in epidemiology. *Am. J. Public Health* 95(S1):S144–50

114. Sagarin R, Pauchard A. 2009. Observational approaches in ecology open new ground in a changing world. *Front. Ecol. Environ.* 8(7):379–86

115. Scheper J, Reemer M, van Kats R, Ozinga WA, van der Linden GTJ, et al. 2014. Museum specimens reveal loss of pollen host plants as key factor driving wild bee decline in The Netherlands. *PNAS* 111(49):17552–57

116. Schmidt NP, O'Neal ME, Anderson PF, Lagos D, Voegtlin D, et al. 2012. Spatial distribution of *Aphis glycines* (Hemiptera: Aphididae): a summary of the suction trap network. *J. Econ. Entomol.* 105(1):259–71

117. Schmitz OJ, Hambäck Beckerman AP. 2000. Trophic cascades in terrestrial systems: a review of the effects of carnivore removals on plants. *Am. Nat.* 155(2):141–53

118. Shi G, Chavas J-P, Lauer J. 2013. Commercialized transgenic traits, maize productivity and yield risk. *Nat. Biotechnol.* 31(2):111–14

119. Siontis GC, Ioannidis JP. 2011. Risk factors and interventions with statistically significant tiny effects. *Int. J. Epidemiol.* 40(5):1292–1307

120. Sivakoff FS, Rosenheim JA, Dutilleul P, Carrière Y. 2013. Influence of the surrounding landscape on crop colonization by a polyphagous insect pest. *Entomol. Exp. Appl.* 149(1):11–21

121. Smyth RR, Allee LL, Losey JE. 2013. The status of *Coccinella undecimpunctata* (L.) (Coleoptera: Coccinellidae) in North America: an updated distribution from citizen science data. *Coleopt. Bull.* 67(4):532–35

122. Soranno PA, Schimel DS. 2014. Macrosystems ecology: big data, big ecology. *Front. Ecol. Environ.* 12(1):3

123. Stack-Whitney K, Meehan TD, Kucharik CJ, Zhu J, Townsend P, et al. 2016. Explicit modeling of abiotic and landscape factors reveals precipitation and forests associated with aphid abundance. *Ecol. Appl.* In press. doi: 10.1002/eap.1418

124. Stafford R, Hart AG, Collins L, Kirkhope CL, Williams RL, et al. 2010. Eu-social science: the role of internet social networks in the collection of bee biodiversity data. *PLOS ONE* 5(12):e14381

125. Steel EA, Kennedy MC, Cunningham PG, Stanovick JS. 2013. Applied statistics in ecology: common pitfalls and simple solutions. *Ecosphere* 4(9):115

126. Steinger T, Gilliand H, Hebeisen T. 2014. Epidemiological analysis of risk factors for the spread of potato viruses in Switzerland. *Ann. Appl. Biol.* 164(2):200–7

127. Steinger T, Goy G, Gilliand H, Hebeisen T, Derron J. 2015. Forecasting virus disease in seed potatoes using flight activity data of aphid vectors. *Ann. Appl. Biol.* 166(3):410–19

127. Multiyear data set that links aphid activity from suction traps with prevalence of plant disease.

128. Steinmann KP, Zhang M, Grant JA. 2011. Does use of pesticides known to harm natural enemies of spider mites (Acari: Tetranychidae) result in increased number of miticide applications? An examination of California walnut orchards. *J. Econ. Entomol.* 104(5):1496–501

129. Storkey J, Bohan DA, Haughton AJ, Champion GT, Perry JN, et al. 2008. Providing the evidence base for environmental risk assessments of novel farm management practices. *Environ. Sci. Policy* 11(7):579–87

130. Suarez AV, Holway DA, Ward PS. 2005. The role of opportunity in the unintentional introduction of nonnative ants. *PNAS* 102(47):17032–35

131. Tamaddoni-Nezhad A, Milani GA, Raybould A, Muggleton S, Bohan DA. 2013. Construction and validation of food webs using logic-based machine learning and text mining. *Adv. Ecol. Res.* 49:225–89

131. Innovative application of machine learning to automate the construction of hypothesized food web linkages.

132. Teacher AGF, Griffiths DJ, Hodgson DJ, Inger R. 2013. Smartphones in ecology and evolution: a guide for the app-rehensive. *Ecol. Evol.* 3(16):5268–78

133. Tobin PC, Kean JM, Suckling DM, McCullough DG, Herms DA, Stringer LD. 2014. Determinants of successful arthropod eradication programs. *Biol. Invasions* 16(2):401–14

134. Vandenbroucke JP. 2004. When are observational studies as credible as randomised trials? *Lancet* 363(9422):1728–31

135. van der Voet H, Goedhart PW. 2015. The power of statistical tests using field trial count data of nontarget organisms in environmental risk assessment of genetically modified plants: power analysis for field trial count data. *Agric. For. Entomol.* 17(2):164–72

136. vanEngelsdorp D, Hayes JH Jr., Underwood RM, Pettis J. 2008. A survey of honey bee colony losses in the U.S., fall 2007 to spring 2008. *PLOS ONE* 3(12):e4071

137. Waloff ZV. 1940. The distribution and migrations of *Locusta* in Europe. *Bull. Entomol. Res.* 31(3):211–46

138. Wang L, Hui C, Sandhu HS, Li Z, Zhao Z. 2015. Population dynamics and associated factors of cereal aphids and armyworms under global change. *Sci. Rep.* 5:18801

139. Wang Z-J, Lin H, Huang J-K, Hu R-F, Rozelle S, Pray C. 2009. *Bt* cotton in China: Are secondary insect infestations offsetting the benefits in farmer fields? *Agric. Sci. China* 8(1):83–90

140. Wan P, Huang Y, Tabashnik BE, Huang M, Wu K. 2012. The halo effect: suppression of pink bollworm on non-Bt cotton by Bt cotton in China. *PLOS ONE* 7(7):e42004

141. Welch JR, Vincent JR, Auffhammer M, Moya PF, Dobermann A, Dawe D. 2010. Rice yields in tropical/subtropical Asia exhibit large but opposing sensitivities to minimum and maximum temperatures. *PNAS* 107(33):14562–67

142. Widenfalk LA, Ahrné K, Berggren Å. 2014. Using citizen-reported data to predict distributions of two non-native insect species in Sweden. *Ecosphere* 5(12):156

143. Williams DW, Liebhold AM. 1995. Herbivorous insects and global change: potential changes in the spatial distribution of forest defoliator outbreaks. *J. Biogeogr.* 22(4/5):665–71

144. Williams DW, Liebhold AM. 2000. Spatial synchrony of spruce budworm outbreaks in eastern North America. *Ecology* 81(10):2753–66

145. Wu K-M, Lu Y-H, Feng H-Q, Jiang Y-Y, Zhao J-Z. 2008. Suppression of cotton bollworm in multiple crops in China in areas with Bt toxin–containing cotton. *Science* 321(5896):1676–78

146. Yang LH, Karban R. 2009. Long-term habitat selection and chronic root herbivory: explaining the relationship between periodical cicada density and tree growth. *Am. Nat.* 173(1):105–12

147. Yang W-R, Grieneisen M, Chen H, Zhang M. 2015. Reduction of crop diversity does not drive insecticide use. *J. Agric. Sci.* 7(10):1