Machine Learning in Ecosystem Informatics

Thomas G. Dietterich

Oregon State University, Corvallis, Oregon, USA
tgd@eecs.oregonstate.edu,
WWW home page: http://web.engr.oregonstate.edu/~tgd

Abstract. The emerging field of Ecosystem Informatics applies methods from computer science and mathematics to address fundamental and applied problems in the ecosystem sciences. The ecosystem sciences are in the midst of a revolution driven by a combination of emerging technologies for improved sensing and the critical need for better science to help manage global climate change. This paper describes several initiatives at Oregon State University in ecosystem informatics. At the level of sensor technologies, this paper describes two projects: (a) wireless, battery-free sensor networks for forests and (b) rapid throughput automated arthropod population counting. At the level of data preparation and data cleaning, this paper describes the application of linear gaussian dynamic Bayesian networks to automated anomaly detection in temperature data streams. Finally, the paper describes two educational activities: (a) a summer institute in ecosystem informatics and (b) an interdisciplinary Ph.D. program in Ecosystem Informatics for mathematics, computer science, and the ecosystem sciences.

1 Introduction

The late Jim Gray (Gray & Szalay, 2003) describes four general approaches to scientific research:

- Observational science, in which scientists make direct observations,
- Analytical science, in which scientists develop analytical models capable of making predictions,
- Computational science, in which scientists employ massive computing power to study the behavior of analytical models and to make predictions at much wider scales of time and space, and
- Data exploration science, in which massive amounts of data are automatically collected from sensors, and scientists employ data mining and statistical learning methods to build models and test hypotheses.

The ecosystem sciences currently employ analytical and computational methods as illustrated, for example, by the extensive work on coupled ocean-atmosphere climate models. However, with the exception of data collected via remote sensing, the ecosystem sciences do not yet have large networks of sensors that automatically collect massive data sets.

Three steps are required to enable ecological research to become a data exploration science. First, sensors that can measure ecologically-important quantities must be developed and deployed in sensor networks. Second, methods for automatically managing and cleaning the resulting data must be developed. Third, data mining and machine learning algorithms must be applied to generate, refine, and test ecological hypotheses.

This paper briefly reviews work at Oregon State University on each of these three steps. Oregon State University has a long history of excellence in the ecosystem sciences. It includes world-leading research groups in forestry, oceanography, and atmospheric sciences, as well as strong teams in machine learning, data mining, and ecological engineering. The campus leadership has made a significant investment in new faculty positions in mathematics, computer science, and forestry with the goal of developing strong interdisciplinary education and research programs in ecosystem informatics.

This paper is organized as follows. The paper begins with a discussion of two sensor development projects, one in wireless sensor networks for plant physiology and the other on computer vision for automated population counting. Then the paper discusses work on automated data cleaning. Finally, the paper briefly describes two educational initiatives aimed at preparing computer scientists, mathematicians, and ecologists to work together in interdisciplinary teams to address the important scientific problems confronting the ecosystem sciences.

2 New Sensor Technologies for Ecology

The study of complex ecosystems is limited by the kinds of data that can be reliably and feasibly collected. Two recent US National Science Board studies (NSB, 2000; NSB, 2002) emphasize the importance of developing new instrumentation technologies for ecological research. At Oregon State, we are pursuing several projects include the following two: (a) wireless, battery-free temperature sensors for forest physiology and (b) computer vision for rapid throughput arthropod population counting.

2.1 Battery-Free Forest Sensors

Forests play an important role in absorbing carbon dioxide and producing oxygen. A central challenge in the study of forest physiology is to understand the exchange of these gasses between the forest and the atmosphere. Existing models of this exchange only capture vertical interactions, under the simplifying assumption that the forest can be modeled as a planar array of trees. But real forests are often on mountain slopes where breezes tend to move up the slope during the day and down the slope at night. Hence, to obtain a more realistic understanding of forest-atmosphere gas exchange, we need to measure and model these lateral winds as well.

Many research groups around the world have developed wireless sensor networks that rely on on-board batteries to provide electric power (Kahn et al.,

1999; Elson & Estrin, 2004). Unfortunately, these batteries typically contain toxic chemicals, which means that these sensors must be retrieved after the batteries have run down. This can be impractical in ecologically-sensitive and inaccessible locations, and it also limits the period of time that the sensor network can be collecting data.

This was the motivation for a team consisting of Barbara Bond (Forest Science), and Terri Fiez, Karti Mayaram, Huaping Liu, and Thinh Nguyen (Electrical Engineering), and Mike Unsworth (Atmospheric Sciences) to develop battery-free sensors for use in the forests of the Pacific Northwest.

The basic design concept is to have a base station that is connected to standard electric power. This base station broadcasts radio frequency energy across the RF spectrum. This energy is harvested by ultra-low power sensor units. They store the energy in a capacitor and use it to make sensor readings and to receive data from and transmit data to other sensors. The data is relayed from the peripheral sensors to the central base station in a series of hops (see Figure 1).

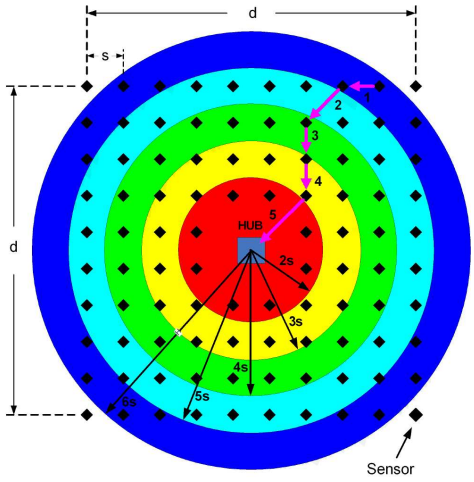


Fig. 1. Spatial layout of battery-free sensor network with powered base station at center.

The development of such passively-powered sensor nodes requires that all components of the sensor employ ultra-low power methods. The initial design includes a temperature sensor, an RF energy harvesting circuit, a binary frequency shift keying (BFSK) receiver, and a BFSK transmitter. The receiver and transmitter share a single antenna. Figure 2 shows the layout of the current prototype sensor.

Note that this prototype contains only a temperature sensor. While it will be easy to add other sensors to the chip, it turns out that by measuring temper-

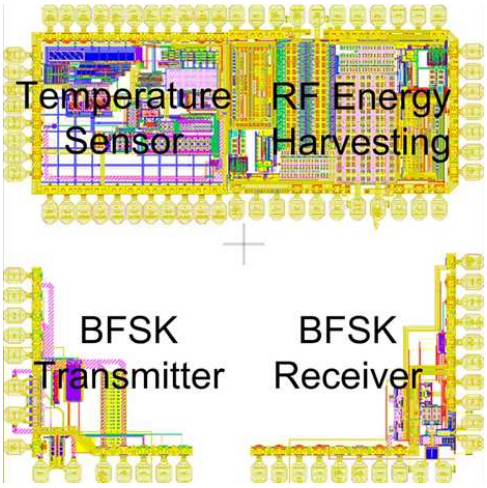


Fig. 2. Layout of prototype battery-free temperature sensor chip

atures, it is possible to infer the lateral winds. So this initial sensor chip will be sufficient to address the forest physiology question that motivated the project.

The ultra-low power temperature sensor measures the outside temperature from -10 to 40 degrees Celsius with an accuracy of ± 0.5 degrees. It is able to achieve this accuracy while consuming only 1nJ per measurement, which is a factor of 85 less energy than is required by state-of-the-art sensors.

The energy harvesting circuit employs a 36-stage "floating gate" design (Le et al., 2006). It is able to harvest energy up to a distance of 15 meters, which is substantially better than the best previously-reported method which only works out to 4.5 meters. Hence, the maximum size of the sensor network region will be approximately 30 meters in diameter.

The transceiver consumes the largest amount of power in the sensor. A low power super-regenerative design based on binary frequency shift keying is employed in the prototype. Experiments in the Oregon coastal mountains with a separate test platform show that even when the sensors are only 10cm above the ground, this design should be able to transmit 10 meters with a raw bit error rate of 10^{-4} (see Figure 3). By applying error-correcting coding, the effective bit error rate will be much lower.

The first version of the chip will be fabricated in summer 2007, while will make it possible to test the complete sensor network design, including energy harvesting and communications protocols.

2.2 Rapid-Throughput Arthropod Population Counting

Two central questions in ecology are (a) to explain the observed distribution of species around the world and (b) to understand the role of biodiversity in maintaining the health and stability of ecosystems. The key data necessary to

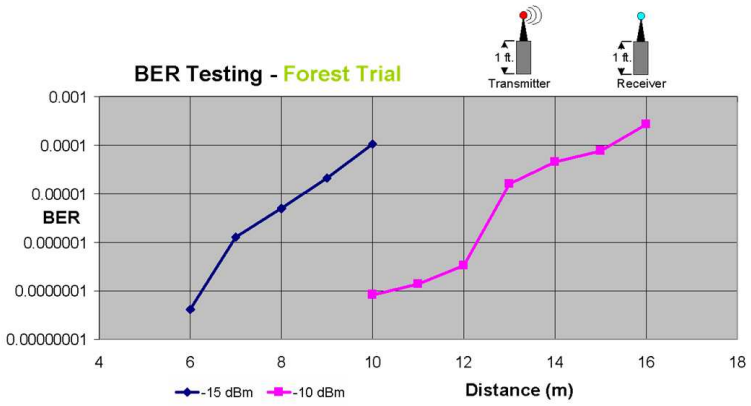


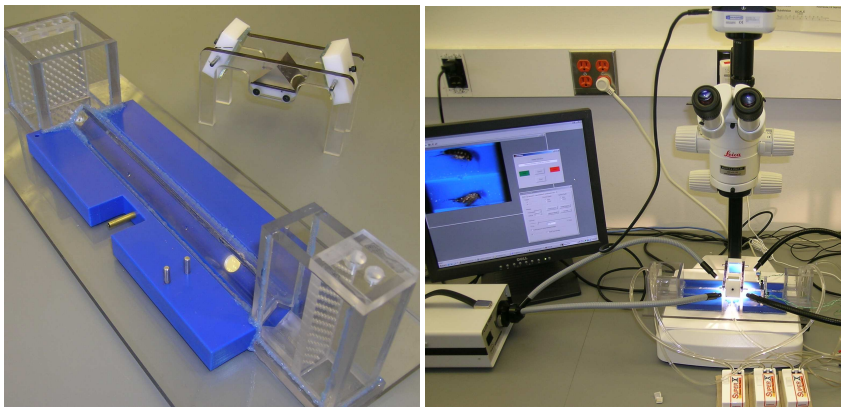
Fig. 3. Bit error rate experiments at two different transmission power levels.

study these questions consists of counting the number of individuals belonging to each species at many different sites.

There are many thousands of species of arthropods. They populate many different habitats including freshwater streams, lakes, soils, and the oceans. They are also generally easy to collect. Despite all of these advantages, the great drawback of using arthropod population data is the tedious and time-consuming process of manually classifying each specimen to the genus and species level. At Oregon State, a team consisting of Tom Dietterich, Eric Mortensen (Computer Science), Robert Paasch (Mechanical Engineering), Andrew Moldenke (Botany and Plant Pathology), David Lytle (Zoology) along with Linda Shapiro (Computer Science) from the University of Washington is developing a rapid-throughput system that combines robotic manipulation with computer vision to automatically classify and count arthropod specimens.

The first application project has been to classify stonefly larvae that live in the substrate of freshwater streams. Stoneflies are an excellent indicator of stream health. They are highly sensitive to pollution, and, because they live in the stream, they provide a more reliable measurement than a single-point-in-time chemical assay. Figure 4 shows the mechanical apparatus that we have developed. In the left image, each individual stonefly specimen is dropped into the plastic reservoir in the lower right part of the image. This reservoir (and the rest of the apparatus) contains alcohol, and the specimen is manipulated via pumps and alcohol jets. The blue part of the apparatus contains a diamond-shaped channel that is covered with transparent plastic. The specimen is pumped into this tube. Infrared detectors (not shown, but located at the two vertical posts and the circular mirror) detect the specimen, cut off the main pump, and turn on a side jet (see the small metal tube emerging from the left side of the blue base). This side jet "captures" the specimen within the field of the microscope (see image (b)). When the side jet is turned off, the specimen falls to the bottom of the channel and a photo is taken. Then the side jet is turned on, which causes the

specimen to rotate rapidly. The jet is again turned off, and another picture taken. This continues until a good image of the back (dorsal) side of the specimen is obtained. The pictures are taken through a mirror apparatus (upper right of (a)), which allows us to capture two views of the specimen with each photo of the camera. This increases the likelihood of capturing a good dorsal view.



(*a*) (*b*)

Fig. 4. (a) Prototype mirror and transportation apparatus. (b) Entire stonefly transportation and imaging setup (with microscope and attached digital camera, light boxes, and computer controlled pumps for transporting and rotating the specimen.

Figure 5 shows example images captured by the apparatus for four different taxa. Notice the large variation in size, pose, and coloration.

The next step in the process is to apply a learned visual classifier to classify the dorsal views into the class. To do this, we employ a variation on the bag-of-interest-points approach to generic object recognition. This approach consists of the following steps:

1. Apply region detectors to the image to find "interesting" regions. We apply three different detectors: The Hessian Affine detector (Mikolajczyk & Schmid, 2004), the Kadir Entropy detector (Kadir & Brady, 2001), and our own PCBR detector (Deng et al., 2007). Figure 6 shows examples of the detected regions.
2. Represent each detected region as a 128-element SIFT vector (Lowe, 2004). The SIFT descriptor vector is a set of histograms of the local intensity gradient direction. Although SIFT was originally developed for object tracking, it has been found to work well for object recognition.
3. Compute a feature vector from the set of detected SIFT vectors. Let $D : R^{128} \mapsto \{1, \ldots, N_D\}$ be a visual dictionary that maps each SIFT vector into an integer between 1 and N_D (N_D varied from 65 to 90 in our experiments).

Fig. 5. Example images of different stonefly larvae species. (a) *Calineuria*, (b) *Doroneuria*, (c) *Hesperoperla* and (d) *Yoraperla*.

The visual dictionary is constructed by fitting a gaussian mixture model with N_D components to the SIFT vectors observed on a separate "clustering" data set. The function D takes a SIFT vector and maps it to the gaussian mixture component most likely to have generated that vector.
Given the visual dictionary, the set of SIFT vectors computed from the image is converted into a feature vector $\mathbf{x}$ such that $\mathbf{x}[i]$ is the number of SIFT vectors $\mathbf{v}$ in the image such that $D(\mathbf{v}) = i$. In effect, $\mathbf{x}$ is a histogram where the ith element counts the number of SIFT vectors that matched the ith dictionary entry.

4. Apply a learned classifier to map $\mathbf{x}$ to one of the K possible taxa.

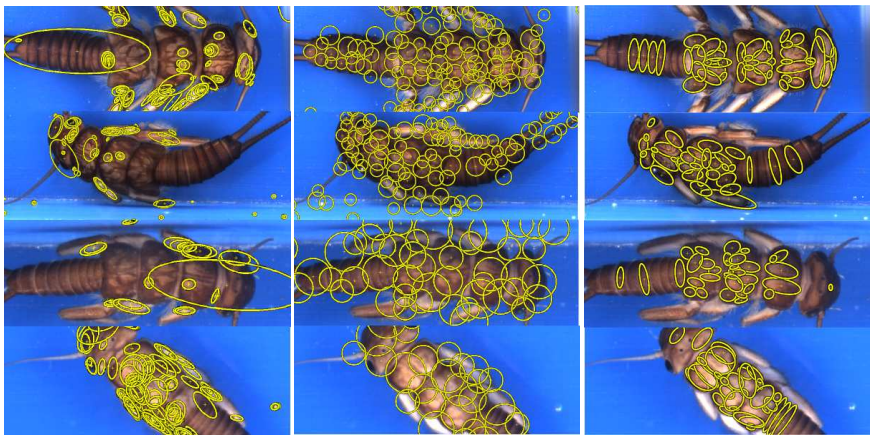


Fig. 6. Visual Comparison of the regions output by the three detectors on three *Calineuria* specimens. (a) Hessian-affine, (b) Kadir Entropy, (c) PCBR

In our work, we learn a separate dictionary $D_{s,d}$ for each species s and each detector d. Consequently, we compute a separate histogram vector $\mathbf{x}_{s,d}$ for each dictionary. In our case, we have 3 detectors and 4 species, so we compute 12 dictionaries and 12 histograms. We then concatenate all of these feature vectors to obtain one very long feature vector which is processed by the learned classifier.

Table 1. Specimens and images employed in the study

Taxon	Specimens	Images
Calineuria	85	400
Doroneuria	91	463
Hesperoperla	58	253
Yoraperla	29	124

To train the system, our entomology collaborators (Lytle and Moldenke) collected and independently classified 263 stonefly specimens. These were then photographed resulting in the data summarized in Table 1. These data were then randomly partitioned into 3 folds (stratifying by specimen and by class), and a 3-fold cross-validation was performed. In each iteration, one fold of the data was employed to learn the visual dictionaries, one fold to train the classifier, and one fold to evaluate the results.

We employed bagged logistic model trees as implemented in the WEKA system (Landwehr et al., 2005) as the classifier (with 20 iterations of bagging). Table 2 shows the results. Overall, the classifier correctly classifies 82.4% of the images (with a 95% confidence interval of $\pm 2.1\%$). The distinction between *Calineuria* and *Doroneuria* is the most challenging. Separate experiments have shown that our accuracy on this 2-class problem is statistically indistinguishable from human performance, when humans are given the same whole-specimen images that our program observes.

We have recently extended this work to apply to 9 stonefly taxa, with an overall accuracy of 85%. This level of accuracy is more than sufficient for use in routine biomonitoring tasks. Consequently, we are planning a trial with standard field samples later this year. More details on this work can be found in Larios et al. (Larios et al., In Press).

Table 2. Confusion matrix of the combined Kadir, Hessian-affine and PCBR detectors

predicted as ⇒	*Cal.*	*Dor.*	*Hes.*	*Yor.*
Calineuria	315	79	6	0
Doroneuria	80	381	2	0
Hesperoperla	24	22	203	4
Yoraperla	1	0	0	123

We have now begun working on a new apparatus and algorithms for recognizing and classifying soil mesofauna and freshwater zooplankton. We anticipate that this apparatus will have a broader range of applications in ecological studies of biodiversity.

3 Automated Data Cleaning for Sensor Networks

As sensors collect data, various things can go wrong. First, the sensors can fail. Second, the data recording process (e.g., the network connection) can fail. Third, the semantic connection between the sensor and the environment can be broken. For example, a thermometer measuring stream water temperature will change to measuring air temperature if the water level falls too low.

To catch these errors, we need methods for automated data cleaning. These methods can be applied to automatically flag data values so that scientists using this data can take appropriate steps to avoid propagating errors into their model building and testing.

Ethan Dereszynski, a doctoral student at Oregon State, has developed an automated data cleaning system for identifying anomalies in temperature data collected at the H. J. Andrews Experimental Forest, which is one of the NSF-funded Long Term Ecological Research (LTER) sites. In this forest, there are three major meteorological stations at three different altitudes. At each station, there is a tower with four temperature sensors which measure and report temperature every 15 minutes. Hence, for this simple sensor network, there are 12 parallel data streams, one for each thermometer.

This data is collected and posted on a web site in raw form. At regular intervals, the LTER staff manually inspect the data to find and remove errors. They then post a clean version of the data, which is the version intended for use by scientists around the world. Our goal is to replace this human data cleaning with an automated process. But a nice side effect of the existing practice is that we have several years of supervised training data for constructing and testing data cleaning methods.

We have adopted a density estimation approach to anomaly detection. Our goal is to develop a model that can evaluate the probability of a new sensor reading given past sensor readings. If the new reading is highly unlikely, it is marked as an anomaly, and it is not used in making subsequent probability estimates. In our work to date, we have focused only on anomaly detection for a single sensor data stream. In future work, we will study simultaneous anomaly detection over the 12 parallel data streams.

Figure 7 shows typical temperature readings as a function of time for the 2.5m sensor at the Central Meteorological station. Observe that there are seasonal effects (it is colder in the winter and warmer in the summer), diurnal (daily) effects (colder at night; warmer in the day), and weather system effects. The weather system effects are the hardest to model. They generally cause the temperature to be systematically warmer or colder than normal over a period of 3-10 consecutive days.

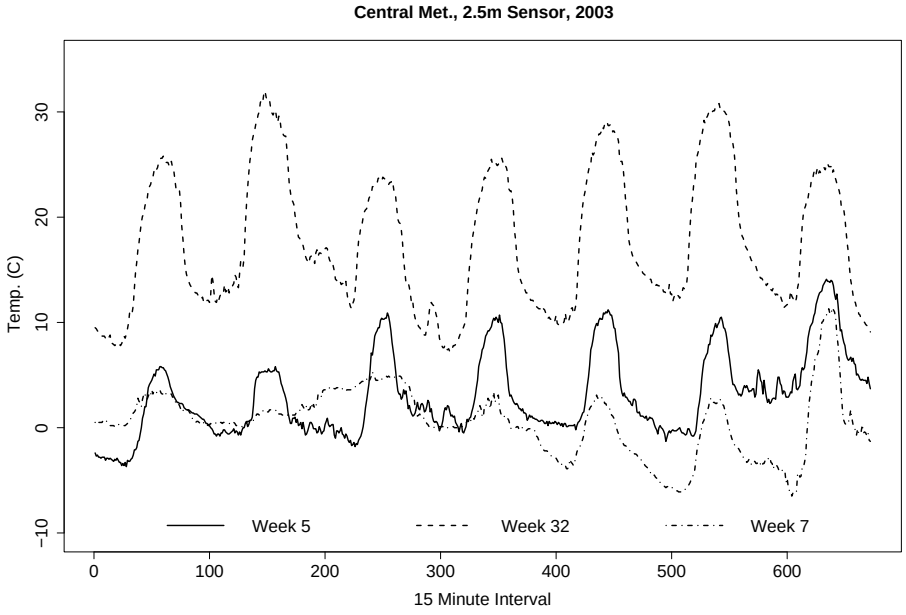


Fig. 7. Seasonal, Diurnal, and Weather effects

Anomalies can be divided into easy, medium, and hard cases. The easy cases are things such as the failure of the connection between the sensor and the data logger. If the data logger loses contact with the sensor, it records a fixed value of -53.3. Similarly, if the data logger receives an input voltage outside the legal bounds, it records a fixed value of -6999. Obviously, these anomalous values are easy to detect.

Medium anomalies can be detected from a single sensor, but they require more subtle analysis. Figure 8 (top) shows a case in which the heat shield on a sensor has been damaged. This causes the sensor to warm up too quickly, measure incorrectly high readings in the hottest part of the day, and then cool down too quickly in the evening. Figure 8(bottom) shows what happens when snow buries the 1.5m and 2.5m sensors. The 1.5m sensor records a steady value of zero (the freezing point), while the 2.5m sensor's readings are damped toward zero. As the snow melts, first the 2.5m sensor recovers and then the 1.5m sensor recovers.

Hard anomalies require the analysis of multiple data streams. One of the most interesting anomalies arose when the cables for two of the sensors were interchanged during maintenance. Normally, the 1.5m, 2.5m, 3.5m, and 4.5m sensors exhibit a monotonic temperature ordering. At night, the 1.5m sensor is warmest, because it is closest to the warm soil. In the day time, the 4.5m sensor is warmest and the 1.5m sensor is coldest. To detect the cable-swap anomaly,

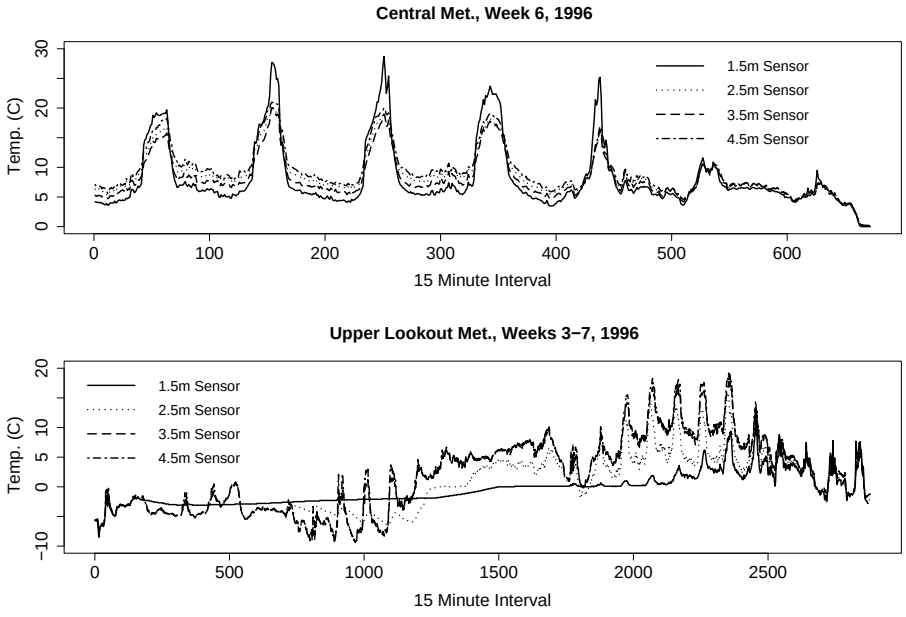


Fig. 8. Top: Broken Sun Shield, Bottom: 1.5m Sensor buried under snowpack, 2.5m Sensor dampened

we need to model the joint distribution of the four sensors and detect that this monotonic relationship is violated. As indicated above, this will be a topic of our future work.

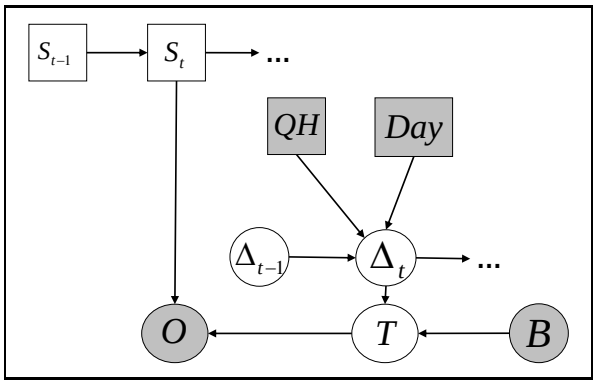


Fig. 9. Dynamic Bayesian network for anomaly detection. Square nodes denote discrete variables; circular nodes denote normally-distributed variables. Grey nodes are observed in the data.

Figure 9 shows our dynamic Bayesian network for anomaly detection. The heart of the model consists of three variables O (the observed temperature), T (the predicted temperature), and S_t (the state of the sensor). The state of the sensor is quantized into four levels ("very good", "good", "bad", and "very bad"). If the sensor is "very good", then O should be equal to T with some slight variation. This is captured by asserting that

$$P(O|T) = Norm(T, 1.0).$$

That is, the mean value of O is T with a standard deviation of 1.0. If S_t is "good", then the standard deviation is 5.0. If S_t is "bad", the standard deviation is 10.0, and if S_t is "very bad", the standard deviation is 100,000 (i.e., effectively infinite).

In practice, we observe O and, based on previously-observed values, compute the probability distribution of T. Then the most likely value of S_t is determined by how different O and T are.

The key to good anomaly detection in this model is therefore to make good predictions for T. To do this, we need to capture the seasonal, diurnal, and weather system variation in temperature. We capture the first two via a "baseline" temperature B. The weather system variation is captured by a first-order Markov variable Δ.

Conceptually, B is the average temperature reading that would be expected for this particular quarter hour and day of the year ignoring short-term changes due to weather systems. However, we have only four years of training data, so if we average only the four readings for the specific time of day and day of year, we

will get a very poor estimate for B. To overcome this problem, we combine the observed values from the 5 temperature readings before and after the particular quarter hour and the 3 days before and after the target day. The local trend within each day and across the 7 days is computed and removed and then the de-trended temperature values are averaged across the years in the training data.

The Δ variable attempts to capture the local departure from the baseline caused by weather systems. It is modeled as a first-order Markov process:

$$P(\Delta_t|QH, D, \Delta_{t-1}) = Norm(\mu_{QH,D} + \Delta_{t-1}, \sigma^2_{QH,D}).$$

QH denotes the quarter hour of each measurement (1, ..., 96); Day (or D) denotes the day of the year (1, ..., 365). The main idea is that Δ_t is approximately equal to Δ_{t-1} but with a slight offset $\mu_{QH,D}$ that depends on the time of day and the day of the year and a variance that similarly depends on the time of day and the day of the year. A warm spell is represented by $\Delta_t > 0$, and a cold period by $\Delta_t < 0$. If $\Delta_t > 0$, then it will tend to stay > 0 for a while, and similarly if $\Delta_t < 0$, it will tend to stay < 0 for a while.

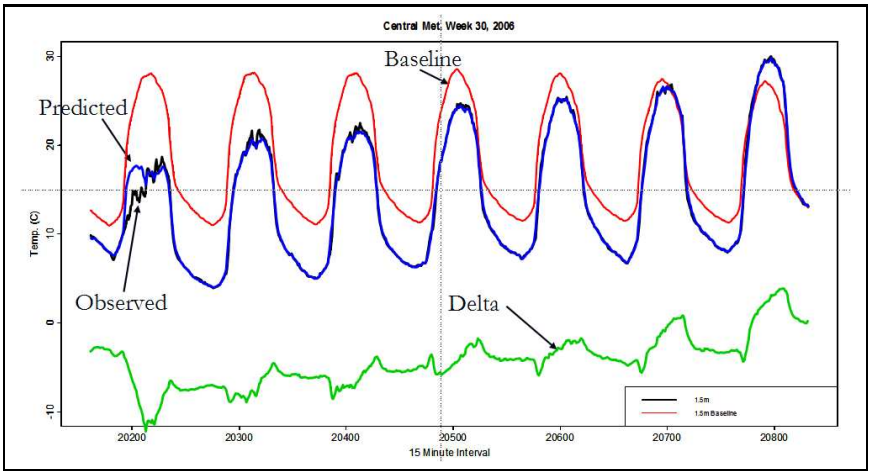


Fig. 10. Relationship between the baseline, Δ, and the observed and predicted temperatures. Note that the baseline curve captures the diurnal variation. It is also slowly dropping, which captures the gradual seasonal change. The Δ curve starts out negative and then gradually increases so that the sum of the baseline plus Δ, which gives the predicted temperature T almost exactly matches the observed temperature O. Where these two curves differ, the model will declare anomalies.

Figure 10 illustrates the relationship between the baseline B, the Δ process, and the observed and predicted temperatures. The fact that Δ varies somewhat erratically reveals that the model still has room for improvement, since ideally, it would be a fairly smooth curve.

The model is applied one temperature reading at a time. First the observed temperature O, and the QH and D are asserted as evidence. Then probabilistic reasoning is performed to compute updated probability distributions for Δ_t and T and the most likely value of S_t. The data point is tagged with this most likely value. If the most likely value is "very bad", then the observed temperature is removed as evidence, and the value of Δ_t is recomputed. Also, the variance $\sigma^2_{QH,D}$ is set to a small value, so that the distribution of Δ_t remains concentrated near the value of Δ_{t-1}. Then the next data point is processed and tagged.

The model was trained using four years of data and then evaluated on the remaining three years. The model correctly detects all of the easy anomalies. Quantitative evaluation of the medium anomalies is more difficult, because the domain expert tended to mark long contiguous intervals of time as anomalous when there was a problem, whereas the model is more selective. For example, when a sun shield was missing, the expert would label whole days as incorrect, whereas the model only marks the afternoon temperatures as bad, because the sensor is still measuring the correct temperature at night. Figure 11 shows the performance of the model in this case. Notice that it not only detects that the peak temperatures are too high but also that the temperature rises and falls too quickly.

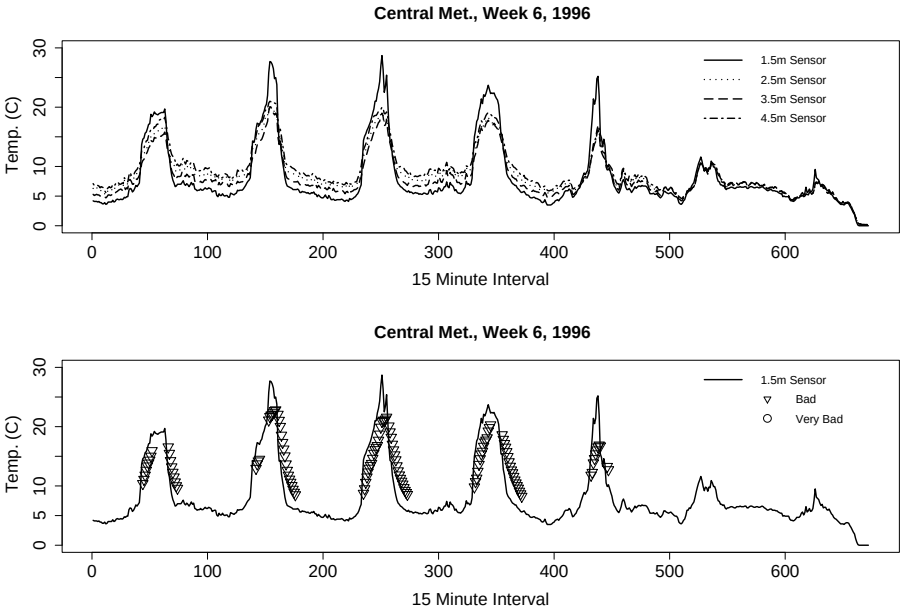


Fig. 11. Top: Lost sun shield in 1.5m sensor. Bottom: Data cleaning applied to 1.5m sensor. Triangles and circles are plotted at points declared to be anomalous. They mark the mean of the predicted temperature distribution.

Our overall assessment is that we are achieving near-100% recall for anomalies but with a false positive rate of roughly 5.3%. This means that we are reducing by over 94% the amount of data that the domain expert must review manually without missing any anomalies. More details are available in Dereszynski and Dietterich (Dereszynski & Dietterich, 2007).

This work shows that carefully-designed dynamic Bayesian networks can do an excellent job of anomaly detection for challenging single-sensor data streams. As more sensor networks are deployed, the need for data cleaning will become much greater, because it will be impossible for human experts to manually inspect and clean the data. We hope that the methods described here will be able to help address this challenge.

4 Education and Training

Ecosystem informatics is inherently an interdisciplinary research area that addresses the scientific problems that arise in various ecological sciences (botany, zoology, population genetics, forest science, natural resource management, earth sciences, etc.) with the modeling and computational methods of mathematics, computer science, and statistics. At Oregon State University, we have developed two educational programs to prepare students for research careers in ecosystem informatics.

4.1 Summer Institute in Ecoinformatics

Under funding from the US National Science Foundation, Professor Desiree Tullos leads a 10-week summer institute in ecosystem informatics for advanced undergraduate and first-year graduate students. Students spend the summer in residence at the Andrews Experimental forest. For the first 3 weeks, they attend an intensive course in ecosystem informatics that introduces them to the scientific problems, research methods, and the terminology of ecosystem informatics. The next 6 weeks involves working on a research project supervised by faculty and doctoral students. This typically involves a mix of field work, data analysis, and mathematical modeling. The final week consists of a series of oral presentations of the results of their research projects.

4.2 Graduate Program in Ecosystem Informatics

The second educational program is a Ph.D. minor in Ecosystem Informatics. This was initiated by a five-year IGERT grant (Julia Jones, Principal Investigator) from the US National Science Foundation that provides graduate fellowship support for students in the program. This was complemented by the hiring of four new faculty members to teach and lead research in this program.

One of the challenges of interdisciplinary education is to prepare people to work together across disciplinary lines without requiring them to become experts in multiple fields. To address this challenge, we decided to structure the program

so that students must have a "home" Ph.D. department, and they receive a doctoral degree in their home department. In addition, they receive a Ph.D. minor in Ecosystem Informatics. The minor involves the following:

- Participation in the Ecosystem Informatics "Boot Camp", which is a one week residential course held at the Andrews Experimental Forest prior to the start of classes in the fall.
- Participation in a year-long Introduction to Ecosystem Informatics class. In this class, students are introduced to the problems and terminology of ecosystem informatics, and they work in cross-disciplinary student teams to study emerging problems in ecosystem informatics.
- Participation in a 6-month internship, preferrably at an institution outside the US. The goal of this is to expose students to research questions motivated by ecological problems outside the US and to give them a more global perspective. Often, this results in a published paper or an idea that can form the basis of their doctoral research.
- Inclusion of an ecosystem informatics chapter in the doctoral dissertation. This chapter is devoted to interdisciplinary work, sometimes with another student in the program. The research topic for this chapter sometimes grows out of the year-long class or the internship. In addition, to help students develop these topics, we organize cross-disciplinary brainstorming sessions for each student. The student presents a proposed problem, and faculty members and other students brainstorm ideas for how to formulate and study the problem.

We are now entering the fourth year of this graduate program. One of the biggest benefits so far has been the development of interesting mathematical models for analyzing disturbance in forests and habitats in streams. In addition, the program has served as a nexus for fostering new interdisciplinary projects including the battery-free sensor network program described in this paper.

5 Concluding Remarks

Many of the most important scientific and policy questions facing humanity require major advances in the ecological sciences. Ecology has traditionally been a difficult area to study because of the difficulty of measuring the primary data: the fluxes of chemicals and nutrients and the distribution and interaction of living organisms. Fortunately, we are in the midst of a revolution in sensor technology that is going to make it possible to measure this primary data continuously with dense networks of sensors. This will enable the ecosystem sciences to apply the methods of data exploration science including data mining, machine learning, and statistical model building to make rapid progress.

This paper has briefly described some of the activities in sensors and ecosystem informatics at Oregon State University. At the level of sensor development, we have discussed the development of ultra-low power temperature sensor nodes that can operate by harvesting power from spread-spectrum RF broadcast from

a central powered base station. We have also described our work on applying computer vision and robotics to automatically manipulate and classify arthropod specimens. At the level of data analysis, we have described work on automated data cleaning for temperature data streams collected over a 7-year period at the Andrews Experimental Forest. Finally, we have discussed two new educational programs that seek to train researchers to work in interdisciplinary teams.

Much more research is required in all of these areas. Furthermore, there is a great need for new kinds of data analysis and data management tools. In particular, machine learning and data mining methods must be developed that can deal with spatially explicit models and that can model interactions among hundreds or thousands of species in time and space. I hope this paper will motivate the reader to consider contributing new ideas to this exciting and important research area.

Acknowledgements

The research described in this paper is funded by several grants from the US National Science Foundation. The battery-free sensor network research is funded by NSF grant BDI-0529223 (Barbara Bond, PI). The arthropod classification project is funded by NSF grant IIS-0326052 (Tom Dietterich, PI). The data cleaning project and the graduate fellowship program are funded by an IGERT grant DGE-0333257 (Julia Jones, PI). And the Summer Institute in Ecoinformatics is funded by grant EEC-0609356 (Desiree Tullos, PI). The author gratefully acknowledges the assistance of Barbara Bond, Adam Kennedy, Ethan Dereszynski, Huaping Liu, and Karti Mayaram in preparing this paper.

Bibliography

Deng, H., Zhang, W., Mortensen, E., Dietterich, T., & Shapiro, L. (2007). Principal curvature-based region detector for object recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2007)*.

Dereszynski, E., & Dietterich, T. (2007). Probabilistic models for anomaly detection in remote sensor data streams. *23rd Conference on Uncertainty in Artificial Intelligence (UAI-2007)*.

Elson, J., & Estrin, D. (2004). Wireless sensor networks: A bridge to the physical world. In Raghavendra, Sivalingam and Znati (Eds.), *Wireless sensor networks*. Kluwer.

Gray, J., & Szalay, A. (2003). *Online science: The world-wide telescope as a prototype for the new computational science* (Technical Report Powerpoint Presentation). Microsoft Research.

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *Int. J. Computer Vision*, *45*, 83–105.

Kahn, J. M., Katz, R. H., & Pister, K. S. J. (1999). Next century challenges: Mobile networking for Smart Dust. *Proceedings of the Fifth Annual ACM/IEEE international Conference on Mobile Computing and Networking* (pp. 271–278). ACM.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, *59*, 161–205.

Larios, N., Deng, H., & Zhang, W. (In Press). Automated insect identification through concatenated histograms of local appearance features. *Machine Vision and Applications*, *In Press*.

Le, T., Mayaram, K., & Fiez, T. S. (2006). Efficient far-field radio frequency power conversion system for passively powered sensor networks. *IEEE 2006 Custom Integrated Circuits Conference (CICC 2006)* (pp. 293–296). IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, *60*, 91–110.

Mikolajczyk, K., & Schmid, C. (2004). Scale and affine invariant interest point detectors. *International J. Computer Vision* (pp. 63 – 83).

NSB (2000). *Environmental science and engineering for the 21st century* (Technical Report NSB-00-22). National Science Foundation.

NSB (2002). *Science and engineering infrastructure for the 21st century: The role of the national science foundation* (Technical Report NSF-02-190). National Science Foundation.