# Induction for Ecological Hypothesis Evaluation: A Case Study

L. Karl Branting[1], William A. Reiners[2], and Yulan Wei[3]

**Abstract**
An important objective of ecology is identification of the processes that produce patterns of life. This project explores how machine learning techniques can help evaluate ecological hypotheses. Seven features (six topographical features and mean wind velocity) were derived from GIS coverages of a portion of the Wyoming Snowy Range. The ability of these features to predict occurrence of trees was tested using various induction algorithms. Wind velocity was found to be a weaker determinant of tree occurrence than topographical features. However, position and tree occurrence in surrounding pixels were found to be even stronger predictors, indicating that topological and wind-flow features are insufficient to account for the observed tree distribution.

## 1 Introduction

One of the objectives of ecology is identification of the processes that underlie patterns of species distributions, and resulting physical structuring of environmental space. Geographic Information Systems (GIS) provide a tool for deriving features predictive of life patterns. Machine learning techniques can, in turn, help identify the combinations of features that are most relevant to the patterns of interest.

The paradigm underlying the research described in this paper is as follows:

- Derive features of possible ecological relevance from GIS coverages and models.
- Evaluate the accuracy of classifiers induced from these features.
- Until a sufficiently predictive set of features is found, revise the feature set or model parameters and repeat.

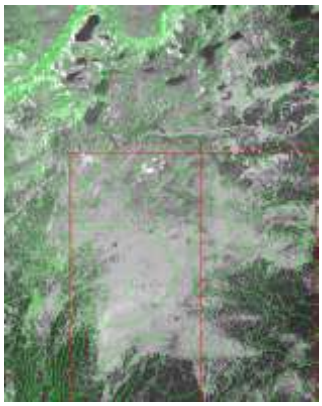
We applied this procedure to the domain task of identifying factors determining the distribution of trees near timberline in the Snowy Range. We hypothesized that the primary determinant of tree distribution above a critical elevation is mean wind velocity, but that topographical features such as altitude, slope, and aspect would have similar predictiveness. More specifically, we hypothesized that terrain positions in which wind velocity was enhanced (windward or convex features) would lead to lesser snow deposition (the main source of moisture for plants of this environment) and more mechanical damage by wind itself and its entrained abrasive ice crystals.

## 2. Feature Derivation

### 2.1 Study Area

The study area was a region of southeastern Wyoming in the Medicine Bow National Forest between 2864 m and 3653 m. The most dominant feature in the area is the Snowy Range, which lies obliquely across the spine of the Medicine Bow Mountains. Libby Flats, a broad ridge crest of the Medicine Bow Mountains themselves, is oriented in northwest to southeast direction. A large portion of the area is above treeline. Coniferous forest occurs primarily at lower elevations in the watershed and in more protected sites at intermediate elevations. Timberline is at approximately 3230 m. Above this elevation, coniferous forest is replaced by alpine tundra vegetation, bare rock and snow.

**Figure 1:** Orthophoto image of the Snowy Range study area.



### 2.2 Data Sources

USDA Forest Service aerial photos were digitized into two data products. One was an ERDAS-formatted, gray scale, orthophoto image with a horizontal resolution of 1 m, shown in Figure 1. The dark, textured regions in the periphery of Figure 1 consist of trees. The smooth-edged regions in the upper-right of Figure 1 are lakes.

The second was an x,y,z digital elevation data set. Both data sets comprised an area of 52,000,000 $m^2$. The orthophoto image was converted to a 1 m resolution grid of 256 integer numbers, with low values for

[1] Department of Computer Science, University of Wyoming , Laramie, WY 82071, USA, email: karl@uwyo.edu.
[2] Department of Botany, University of Wyoming, Laramie, WY 82071, USA, email: reiners@uwyo.edu.
[3] CSI Technology Group, USA, email: victoria@csitech.com.

low surface reflectance and high number for high surface reflectance. The x, y, z data were processed in ARC/INFO into a digital elevation model (DEM) with a grid resolution (x,y) of 25 m and a vertical (z) resolution of 1 m. The 25 m DEM was resampled into a 1 m grid.

### 2.3 Target Category

The category of interest in this research is the presence of coniferous vegetation in either an upright tree form or in a shrub-like ("krummholz") form. Coniferous plants were mainly Englemann spruce (Picea englemannii) but included some subalpine fir (Abies lasiocarpa). Non-coniferous features of the study area included alpine meadows, willow thickets, bare rocks, and lakes. Different reflectances on the surface reflectance grid in the orthophoto were differentiated by 256 consecutive integer numbers from O to 255. The coniferous class was separated from non-coniferous classes by screening for low reflectance values. The threshold number for distinguishing the coniferous class was determined with the aid of aerial photos. A reflectance rate of 99 was found to distinguish coniferous class from non-coniferous class except for the lakes, which were relabeled manually.

### 2.4 Feature Derivation

For each 1 $m^2$ pixel, the following predictive features were derived from the GIS topographic coverage:

- Altitude
- Slope
- Distance to global high point
- Angle to global high point
- Distance to local high point
- Angle to local high point
- Mean wind velocity

Altitude has a clear effect on vegetation. We hypothesized that slope would also affect vegetation by determining the rate of water flow resulting from melting snow and, to a lesser degree, summer rains. We therefore calculated the slope at a pixel as the steepest elevation change relative to the pixel's neighbors. This was calculated by finding the greatest elevation decrease among the pixel's eight neighbors.

#### 2.4.1 Global Distance and Angle

Wind velocity and snow deposition at a point are affected by the proximity of that point relative to the ridge with respect to prevailing wind direction. The prevailing wind direction in the study area is westerly (and therefore parallel to lines of latitude). We defined a pixel's *global distance* to be the pixel' s relative distance from the pixel with highest altitude in the same latitude. A pixel to the east of highest pixel was defined as positive, and a pixel to the west was negative. A pixel's *global angle* was defined by the angle relative to the highest pixel.

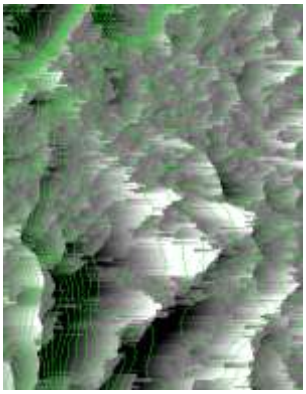
A GIS coverage for global distance in which values are displayed with brightness proportional to the global distance modulo 256 is shown in Figure 2.

**Figure 2:** A GIS coverage for global distance. Brightness is proportional to distance modulo 256.

#### 2.4.2 Local Distance and Angle

If there are multiple ridges of similar altitude along the direction of the prevailing wind, each ridge may have a separate effect on vegetation. **Local distance** and **local angle** were calculated the same way as global features, but relative to the closest local maximum rather than to the global maximum. Figure 3 shows a GIS coverage for local distance

**Figure 3**: A GIS coverage for local distance, with brightness proportional to distance.



#### 2.4.3 Wind Velocity

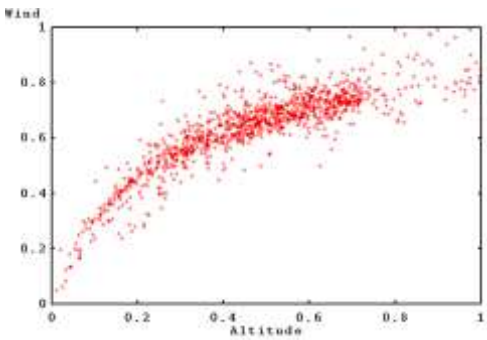
An understanding of air flow over topography has many applications, including predicting vegetation distribution, local pollution transport, and wind loading on buildings. However, accurate measurements of fine resolution wind flow over remote mountain are normally impossible. As a result, numerical wind flow models must be used instead.

Numerous wind-flow models have been developed for various applications. In general, surface curvature has a significant effect on wind flow near the ground (Coppin et al., 1986). Typically, concave curvature destabilizes and convex curvature stabilizes the wind-

flow field. Since the study area has a complex surface, a model that includes an explicit curvature term is necessary.

We derived mean wind velocity at each point of the study area from surface curvature using the streamline coordinate system proposed in Finnigan (1988). The prevailing wind is from the west (Hiemstra 1999) with a default velocity of 7.4 m/s at 3400 meters. A complete description of our wind-flow model is set forth in Wei (1997). Figure 4 shows a scatter plot of predicted velocity as a function of altitude across the study area.

**Figure 4:** Predicted wind velocity as a function of altitude.




## 4 Learning Experiments

### 4.1 Experimental Design

The ability of the seven features described above to predict the distribution of trees was tested using 3-fold cross-validation in four trials. Each instance was representted as a feature vector in terms of the various combinations of the seven predictive features, normalized onto the [0..1] interval together with the target classification. Accuracy was measured by mean percentage of correct classifications.

The feature sets used in the experiments were as follows:

- Altitude only
- Wind only
- Altitude and slope
- Altitude, slope, and wind
- Altitude, slope, wind, global distance and angle
- Altitude, slope, wind, local distance and angle
- All features (altitude, slope, wind, global distance and angle, local distance and angle)
- All but wind features (altitude, slope, global distance and angle, local distance and angle)

To determine whether other, unidentified factors play an important role in tree growth, several experiments used as additional features (1) the <x,y> coordinates of each pixel or (2) the classification of the 8 surrounding pixels. Clearly, these features have no ecological meaning. However, to the extent that they are predictive, they indicate that tree distribution is not entirely random.

Four different learning methods were applied in each experiment to reduce the possible bias from any given learning method.

1. KNN (k=5)
2. ID3
3. Perceptron
4. Back-propagation

See (Russell & Norvig , 1995; Mitchell 1997) for a description of these learning methods.

### 4.2 Experiment 1

In the first experiment, the four machine learning algorithms were applied to data sets derived from the entire study area. Since the grid resolution was 1 meter, there were about 52,000,000 pixels in the study area. 945 instances were selected from these pixels and represented in terms of each of the 8 feature sets shown above. In the sample set 26.4% of the pixels were members of the coniferous class and the remaining 71.6% were in the nonconiferous class. The results of the experiment are shown in Table 1.

**Table 1:** Results of experiments on full study area.

| Data sets | KNN | ID3 | Percep. | Back prop |
|---|---|---|---|---|
| Altitude only | 66.23 | 63.45 | 54.34 | 71.53 |
| Wind only | 67.32 | 62.31 | 54.18 | 66.41 |
| Altitude slope | 67.81 | 61.34 | 67.66 | 71.59 |
| Altitude slope wind | 67.91 | 62.58 | 60.66 | 69.92 |
| Alt slope wind global | 72.87 | 69.52 | 64.65 | 67.60 |
| Alt slope wind local | 68.84 | 63.63 | 62.55 | 65.03 |
| All | 73.67 | 70.76 | 66.32 | 69.73 |
| All but wind | 71.38 | 69.54 | 63.10 | 72.96 |

Surprisingly, the accuracy for most feature sets was less than the 71.6% accuracy of the majority classification rule. To determine whether tree distribution was predictable at all in the study area, the experiments were repeated using <x,y> coordinates and the classification of the 8 adjacent pixels as features. As shown in Table 2, these features produced much better results. These results indicate that (1) the position of pixels is predictive of coniferous vegetation and (2) coniferous vegetation is highly aggregated. However, the topographic features appear to capture little of the factors responsible for tree distribution in this data set.

**Table 2**: Classification accuracy using non-topographic features

| Data sets | KNN | ID3 | Percep. | Back prop |
|---|---|---|---|---|
| 8 adjacent pixels | 95.50 | 95.31 | 95.50 | 95.47 |
| x, y | 86.12 | 84.35 | 70.40 | 83.20 |

### 4.3 Experiment 2

A careful examination on the study area revealed that the southwest-to-northeast oriented Snowy Range lies across the northwest corner of the study area. We hypothesized that the huge elevation variation of the Snowy Range causes inaccuracy in the wind simulation and some other derived features. To exclude the huge terrain variation of the Snowy Range portion, we restricted our analysis area to the lower right comer of the original study area, the Libby Flats area. 1120 pixels were sampled from 23,000,000 $m^2$ Libby Flats area, which was 35.2% coniferous. The accuracy was significantly higher (74.99% for back-propagation using all 6 topographical features plus wind velocity), although still not satisfactory.

### 4.4 Experiment 3

A comparison of the vegetation pattern on west and east portions of the Libby Flats area revealed that the east portion shows non-coniferous area seemingly unrelated to terrain and inferred wind velocity. This suggested that the driving mechanism of the vegetation pattern on the east portion is not merely altitude and wind. A possibility is past wildfire. Libby Flats is known to have experienced fire because of fire-killed trees scattered across the landscape. Since possible anomalous areas were located in the eastern half of the restricted Libby Flats area, the third experiment was confined to the western portion of Libby Flats.

**Table 3**: Classification accuracy for the western portion of Libby Flats.

| Data sets | KNN | ID3 | Percep. | Back prop |
|---|---|---|---|---|
| Altitude only | 82.66 | 79.40 | 84.03 | 85.07 |
| Wind only | 84.49 | 77.87 | 50.99 | 85.15 |
| Altitude slope | 83.23 | 81.55 | 84.42 | 84.69 |
| Altitude slope wind | 83.46 | 80.13 | 63.82 | 84.49 |
| Alt slope wind global | 89.93 | 88.52 | 79.67 | 86.91 |
| Alt slope wind local | 87.36 | 86.03 | 82.35 | 85.11 |
| All | 90.02 | 87.22 | 73.18 | 90.12 |
| All but wind | 90.23 | 88.70 | 82.74 | 90.73 |
| x, y | 93.03 | 90.81 | 80.02 | 91.66 |
| x, y and all | 92.84 | 89.74 | 85.72 | 92.00 |

In a test of the western portion of Libby Flats 968 instances were drawn from the 12,000,000 $m^2$ coverage. The instances were 17.7% coniferous. The results are shown in Table 3.

Using back-propagation as the learning method, altitude and wind alone yielded 85.07% and 85.15% accuracy, respectively. When trained with all the features, back-propagation had 90.12% accuracy. This indicates that the features are able to predict vegetation pattern very well where their influence is not superceded by other factors such as fire history.

The accuracy of back-propagation result did not decrease when wind was excluded. This suggests that some features are redundant. The local features, global features, and slope feature might be another means of expressing the wind effect on vegetation pattern. The x and y coordinates as features give 93.03% and 91.66% accuracy from KNN algorithm and back-propagation algorithm, respectively. This accuracy is higher than the accuracy of the entire topographic feature set. This means that a portion of the predictiveness of position was not captured by the topographic features.

## 5 Conclusion

The learning experiments showed that the wind-flow model was only weakly predictive of tree distribution, and that topographical features were much more predictive. However, even in West Libby Flats, the region for which the learning methods yielded the best results, <x,y> position is as predictive as any combination of topographic features. This indicates that other factors not captured by topographic features, such as a history of fire, must play an important role in tree distribution.

These experiments illustrate how machine learning methods can be used to evaluated the predictiveness of proposed ecological features and can guide hypothesis formation and modification.


## Acknowledgments

This research was supported in part by a grant from the Andrew W. Mellon Foundation.


## References


Coppin, P. A., Bradley, E. G., and Katen, P. G., 1986. The evolution of atmospheric turbulence over a two-dimensional ridge. In *Proceedings of the Ninth Australian Fluid Mechanics Conference*., Auckland, New Zealand, 569-572.

Finnigan, J. J. 1988. Air flow over complex terrain. In *Flow and Transport in the Natural Environment: Advances and Applications.* (W. L. Steffen and 0. T. Denmead, Eds.), Springer-Verlag, Berlin, Germany, 183-229.

Hiemstra, C.A. 1999. Wind *Redistribution of Snow at Treeline, Medicine Bow Mountains, Wyoming.* MS Thesis, Department of Botany, University of Wyoming.

Russell, S. and Norvig, P., 1995. Artificial Intelligence: A Modern Approach, Prentice Hall.

Mitchell, T., 1997. Machine Learning, McGraw-Hill.

Wei, Y. 1997. *Integration of Geographical Information Systems (GIS), machine learning, and Wind Flow Modeling for Prediction of Alpine Vegetation Pattern,* MS Thesis, Department of Computer Science University of Wyoming.