USE OF MACHINE LEARNING ALGORITHMS TO PREDICT THE INCIDENCE OF LEAD EXPOSURE IN GOLDEN EAGLES

ERICA H. CRAIG[1], TIM H. CRAIG[2], FALK HUETTMANN[3], MARK R. FULLER[4]

[1]*Aquila Environmental, P.O. Box 81291, Fairbanks, AK 99708, USA.*
E-mail: tecraig@acsalaska.net

[2]*Bureau of Land Management, Central Yukon Field Office, 1150 University Avenue, Fairbanks, AK 99709, USA.*

[3]*EWHALE lab, Biology and Wildlife Department, Institute of Arctic Biology, 419 IRVING I, University of Alaska-Fairbanks, Fairbanks, AK 99775-7000, USA.*

[4]*US Geological Survey, Forest and Rangeland Ecosystem Science Center, 970 Lusk St., Boise, ID 83706, USA.*

ABSTRACT.—Quantitative models can be used to predict the occurrence of wildlife relative to certain environmental conditions. Resolving the impacts of environmental contaminants on wildlife often involves complex data sets suitable for analysis with quantitative models; yet despite their potential, such models are not commonly used. In this paper, we use data collected from wintering Golden Eagles *(Aquila chrysaetos)* and GIS-based models to demonstrate the use of stochastic gradient boosting, a machine learning algorithm, to examine factors most likely to influence the incidence of elevated blood lead levels. This fast, data-mining algorithm is capable of constructing predictive but sensitive and generalized models from complex contaminants datasets, and preliminary results suggest it accurately identified patterns that clarified and extended results of analyses performed using traditional statistical techniques. The management implications of using these models are far-reaching in their potential for identifying members of a population most at risk to contaminants, factors most likely to influence the incidence of lead contamination in a population, and potential sources of lead in the landscape. *Received 1 July 2008, accepted 4 December 2008.*

CRAIG, E. H., T. H. CRAIG, F. HUETTMANN, AND M. R. FULLER. 2009. Use of machine learning algorithms to predict the incidence of lead exposure in Golden Eagles. *In* R. T. Watson, M. Fuller, M. Pokras, and W. G. Hunt (Eds.). Ingestion of Lead from Spent Ammunition: Implications for Wildlife and Humans. The Peregrine Fund, Boise, Idaho, USA. DOI 10.4080/ilsa.2009.0303

Key words: stochastic gradient boosting, machine learning algorithm, modeling, lead, Golden Eagles, contaminants, lead contamination, data mining, predictive modeling, TreeNet®, machine learning.

THE INVESTIGATION OF ENVIRONMENTAL CONTAMINANTS is challenging because the variables are many, complex, and difficult to interpret. The sheer complexity of such datasets can affect the ability of researchers to obtain accurate results using traditional statistical approaches and to find the best possible solutions for making sustainable decisions (Craig and Huettmann 2009). Such obstacles can result in under-utilization of available information because of the inability to identify meaningful biological patterns in data with multi-dimensional input variables.

Recently, the incidence of lead in the environment as a result of lead bullet fragments in hunter killed game (Hunt et al. 2006), and its potential effect on the population viability of raptors and scavenger species such as the California Condor (*Gymnogyps californianus*), has received considerable attention (Church et al. 2006, Cade 2007, Johnson et al. 2007, Craighead and Bedrosian 2008). The Golden Eagle (*Aquila chrysaetos)* is known to scavenge carrion, including the remains of gun-killed animals, and it is thought that elevated lead levels in the species are related to the ingestion of ammunition lead remaining in these animals (Craig et al. 1990, Pattee et al. 1990, Wayland and Bollinger 1999, Fisher et al. 2006). Reports from long-term studies of Golden Eagles in Idaho, Colorado, and California suggest declines in productivity (Leslie 1992, Steenhoff et al. 1997, Kochert and Steenhoff 2002, D. Bittner and J. Oakley in Kochert et al. 2002). Declines in counts of migrant raptors in some parts of the western United States (Hoffman and Smith 2003) are coincidental with ongoing environmental changes resulting from wildfires, invasive plants, oil and gas development, urban sprawl, and other factors (Leslie 1992, Kochert and Steenhoff 2002, Knick et al. 2003). Stress from sublethal lead contaminant loads has the potential to further affect survival and reproduction (Fisher et al. 2006, Craighead and Bedrosian 2008). However, the issue of environmental contaminants in wildlife populations and how they affect a species is complex. Sources of lead can vary spatially and temporally and at different scales, and may differentially impact segments of the population; numerous factors could potentially influence the occurrence of elevated lead in individuals and populations.

Traditional statistical approaches to such problems usually require assumptions about the distribution of variables, independence of variables, and linearity of the data, and typically restrict the number of predictor variables. Such methods are often labor intensive and thus, costly. They also require *a priori* knowledge and decisions about the data relevance and applicability that might introduce bias in the resultant models. For these reasons, researchers are increasingly using powerful and flexible data mining techniques for exploratory analysis of complex ecological questions (Hochachka et al. 2007). They are finding that such alternative approaches greatly outperform traditional modeling methodologies with non-linear and linear data (Olden and Jackson 2002, Prasad et al. 2006).

We demonstrate the advantages of using a data mining algorithm known as stochastic gradient boosting (SGB) to identify meaningful patterns and relationships (Friedman 2001, 2002) in the investigation of contaminants in wildlife. We use blood lead levels (BLL) from 323 wintering Golden Eagles in Idaho, and satellite telemetry locations from six of those eagles (Craig and Craig 1998) in a GIS environment in a sample application of this method.

METHODOLOGY AND DISCUSSION

Machine learning is a rapidly advancing field of artificial intelligence whereby computer programs strategize in response to diverse data input. Stochastic gradient boosting is a refinement of classification and regression tree analysis (Breiman et al. 1984). It is a hybrid "boosting and bagging" algorithm that randomly selects subsets of the sample data (without replacement) and uses them to fit a series of very simple decision trees in sequence. Each successive tree is built from the prediction residuals of the preceding tree(s) and the final model is a summation of the best partitioning of the data (Friedman 2002). Models are developed using the software known as TreeNet® (Salford Systems, Inc.). Model construction is fast, requires no *a priori* assumptions about the relationship between the response and predictor variables, does not limit the number of predictor variables, and is capable of uncovering the underlying structure in data that are non-additive or hierarchical in nature (Prasad et al. 2006). Resultant models are robust with a high degree of predictive accuracy (Hochachka et al. 2007). The user can define an unlimited number of predictor variables that might potentially influence the occurrence of the target variable (in the case of our example, blood lead levels in Golden Eagles).

These features are highly applicable to the field of applied ecology (Olden et al. 2008) and critical for evaluating the validity of models. With the rapidly changing environment and numerous threats to biological diversity, such accurate predictive capability is increasingly important. With SGB, evaluation can be accomplished by: 1) applying the model to

an independent dataset; 2) withholding a randomly selected portion of the dataset and using it as "testing" data (we used this method for our example dataset, withholding 20% of the data to be used for independent testing); or 3) in the case of very small datasets, through k-fold cross-validation (Kohavi 1995). The software produces a confusion matrix of the percent of eagles (in the case of our example model) correctly predicted with the presence or absence of elevated BLL and a Receiver Operating Characteristic Curve (ROC) that describes the sensitivity and specificity of the model (Collinson 1998). To further aid in interpretation of the models, and the biological significance of the results, TreeNet® produces an index ranking the relative importance of the predictor variables. The top predictor variable is given a score of 100 and then all other predictors are ranked in descending order, based on their influence upon the target variable, in comparison with the top predictor. To further aid in the interpretation of the data, graphic displays of dependence plots for single variables and two way interactions among variables are produced (Friedman 2001, 2002). This information, combined with expert opinion about a species, can be a valuable tool in the development of models that have ecological validity.

We used six predictors to develop our models (age and sex of the eagle, year and month in which it was captured, location of capture, and time of day the bird was captured). We defined the number of nodes in each tree and the number of samples allowed in each terminal node; we used six-node trees (the default) and 10 samples in each terminal node. The final number of trees for best model performance was determined by the software, which indicated the point at which no further variation in the data was explained by the formation of additional trees. We used the binary logistic option for model development, with presence or absence of BLL above background ($\geq$0.20 ppm, Kramer and Redig 1997) as our binary response.

Our models indicated that the contaminants dataset for the wintering Golden Eagles was not simply linear in nature and further clarified results of analyses performed using traditional statistical techniques (Craig and Craig 1998). Similar to other studies, our preliminary model indicated that females were more likely to have elevated BLL than males (Craighead and Bedrosian 2008, Pain et al. 1993), and there existed a positive relationship between eagles exposed to lead and the month of December, when offal from hunter-killed game associated with gun hunting seasons was most often available (Pattee et al. 1990, Church et al. 2006, Craighead and Bedrosian 2008). Our models also allowed us to examine the interactions among variables, particularly in regard to differences observed in BLL among years and valleys. This information, in conjunction with our geospatial analysis, allowed us to examine the distribution of lead contaminated eagles relative to the location of historical lead mines within the study area and identified an area for additional research. We will further investigate the incidence and sources of lead for this wintering population by examining stable lead isotopes signatures among a sample of the wintering eagles.

CONCLUSION

Preliminary results suggest that SGB accurately and quickly produced a model that independently detected patterns of lead contamination in wintering Golden Eagles corroborative with other studies using more traditional analytical methods and those of our previous research on this population. Based on our model, it is likely that lead contamination in hunter-killed game carcasses is the principal source of lead exposure to the wintering population of Golden Eagles in this study, but differences among years, study sites, and timing of capture raise additional questions that need to be clarified. For example, we must examine the possible role of contamination from local historical lead mining sites.

The management implications of using predictive modeling to investigate the effects of contaminants in wildlife are potentially far-reaching. Our preliminary results suggest that machine learning algorithms, such as SGB, are particularly appropriate management tools for identifying patterns among data to generate hypotheses for further research. SGB also very quickly created structured predictive models as stand-alone products useful for identification of factors influencing the incidence of elevated lead levels in Golden Eagles, of possible sources of contamination, and of segments of the population most at risk. SGB is appropriate for

approaching any similar problem, and we believe that linking such computational software with wildlife ecology and conservation management in an interdisciplinary framework is crucial for timely responses to critical ecological questions, and is relevant for species sustainability (Chernetsov and Huettmann 2005, Olden et al. 2008).

ACKNOWLEDGMENTS

Many individuals have provided technical support for this project. We would particularly like to acknowledge the contributions of R. Craig, J. Craig, and H. Craig. Funding was provided by the US Bureau of Land Management, US Geological Survey, University of Alaska Fairbanks, E-WHALE Lab, Western Ecological Studies Team, Idaho Department of Fish and Game, Golden Eagle Chapter of the Audubon Society and The Idaho Wildlife Society. We thank W. G. Hunt and C. J. Henny for reviewing our manuscript. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the US Government.

LITERATURE CITED

BREIMAN, L., J. H. FRIEDMAN, R. A. OLSHEN, AND C. J STONE. 1984. Classification and regression trees. CRC Press. Boca Raton, Florida, USA.

CADE, T. J. 2007. Exposure of California Condors to lead from spent ammunition. Journal of Wildlife Management 71:2125–2133.

CHERNETSOV, N., AND F. HUETTMANN. 2005. Linking global climate grid surfaces with local long-term migration monitoring data: Spatial computations for the Pied Flycatcher to assess climate-related population dynamics on a continental scale. Pages 133–142 *in* Lecture Notes in Computer Science (LNCS) 3482, International Conference on Computational Science and its Applications (ICCSA), Springer Berlin. Heidelberg, Germany.

CHURCH M. E., R. GWIAZDA, R. W. RISEBROUGH, K. SORENSON, C. P. CHAMBERLAIN, S. FARRY, W. HEINRICH, B. A. RIDEOUT, AND D. R. SMITH. 2006. Ammunition is the principal source of lead accumulated by California Condors reintroduced to the wild. Environmental Science and Technology 40:6143–50.

COLLINSON, P. 1998. Of bombers, radiologists, and cardiologists: time to ROC. Heart 80:215–217.

CRAIG, E., AND F. HUETTMANN. 2009. Using "blackbox" algorithms such as TreeNet and Random Forests for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data: An overview, an example using Golden Eagle satellite data and an outlook for a promising future. *In* Hsiao-fan Wang (Ed.). Intelligent Data Analysis: Developing New Methodologies through Pattern Discovery and Recovery. IGI Global, Hershey, Pennsylvania, USA.

CRAIG, E. H., AND T. H. CRAIG. 1998. Lead and mercury levels in Golden and Bald Eagles and annual movements of Golden Eagles wintering in east-central Idaho: 1990–1997. Idaho BLM Technical Bulletin No. 98–12.

CRAIG, T. H., J. W. CONNELLY, E. H. CRAIG, AND T. PARKER. 1990. Lead concentrations in Golden and Bald Eagles. Wilson Bulletin 102:130–133.

CRAIGHEAD, D., AND B. BEDROSIAN. 2008. Blood lead levels of Common Ravens with access to big-game offal. Journal of Wildlife Management 73:240–245.

FISHER, I. J., D. J. PAIN, AND V. G. THOMAS. 2006. A review of lead poisoning from ammunition sources in terrestrial birds. Biological Conservation 131:421–432.

FRIEDMAN, J. H. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 2:1189–1232.

FRIEDMAN, J. H. 2002. Stochastic gradient boosting: Nonlinear methods and data mining. Computational Statistics and Data Analysis 38:367–378.

HOCHACHKA, W. M., R. CARUANA, D. FINK, A. MUNSON, M. RIEDEWALD, D. SOROKINA, AND S. KELLING. 2007. Data-mining discovery of pattern and process in ecological systems. Journal of Wildlife Management 71:2427–2437.

HOFFMAN, S. W., AND J. P. SMITH. 2003. Population trends of migratory raptors in western North America, 1977–2001. Condor 105:397–419.

HUNT, W. G., W. BURNHAM, C. N. PARISH, K. K. BURNHAM, B. MUTCH, AND J. L. OAKS. 2006. Bullet fragments in deer remains: Implications for lead exposure in avian scavengers. Wildlife Society Bulletin 34:67–170.

JOHNSON, C. K., T. VODOVOZ, W. M. BOYCE, AND J. A. K. MAZET. 2007. Lead exposure in California Condors and sentinel species in California. Report prepared for the California Fish and Game Commission [Online.] Available at http://www.dfg.ca.gov/habcon/info/bm_research/bm_pdfrpts/WHCleadcondorsfinal.pdf

KNICK, S. T., D. S. DOBKIN, J. T. ROTENBERRY, M. A. SCHROEDER, W. M. VANDER HAEGEN, AND C. VAN RIPER III. 2003. Teetering on the edge or too late? Conservation and research issues for avifauna of sagebrush habitats. Condor 105:611–634.

KOCHERT, M. N., AND K. STEENHOF. 2002. Golden Eagles in the US and Canada: Status, trends, and conservation challenges. Journal of Raptor Research 36 (1 Suppl.): 32–40.

KOCHERT, M.N., K. STEENHOF, C.L. MCINTYRE, AND E.H. CRAIG. 2002. Golden Eagle (*Aquila chrysaetos*). *In* A. Poole and F. Gill (Eds.). The Birds of North America, No. 684. The Birds of North America, Inc., Philadelphia, Pennsylvania, USA.

KOHAVI, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. (Pages 1137–1143). *In* Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Mateo, California, USA.

KRAMER, J. L., AND P. T. REDIG. 1997. Sixteen years of lead poisoning in eagles, 1980–1995: An epizootiologic view. Journal of Raptor Research 31:327–332.

LESLIE, D. G. 1992. Population status, habitat and nest-site characteristics of a raptor community in eastern Colorado. Unpubl. Master's Thesis, Colorado State University, Fort Collins, Colorado, USA.

OLDEN, J. D., AND D. A. JACKSON. 2002. A comparison of statistical approaches for modeling fish species distributions. Freshwater Biology 47:1–20.

OLDEN, J. D., J. J. LAWLER, AND N. L. POFF. 2008. Machine learning without tears: A primer for ecologists. The Quarterly Review of Biology 83:171–193.

PAIN, D. J., C. AMIARD-TRIQUET, C. BAVOUX, G. BURNELEAU, L. EON, AND P. NICOLAU-GUILLAUMET. 1993. Lead poisoning in wild populations of Marsh Harriers *Circus aeruginosus* in the Camargue and Charente-Maritime, France. Ibis 135:379–386.

PATTEE, O. H., P. H. BLOOM, J. M. SCOTT, AND M. R. SMITH. 1990. Lead hazards within the range of the California Condor. Condor 92:931–937.

PRASAD, A. M., L. R. IVERSON, AND A. LIAW. 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 9:181–199.

STEENHOF, K., M. N. KOCHERT, AND T. L. MCDONALD. 1997. Interactive effects of prey and weather on Golden Eagle reproduction. Journal of Animal Ecology 66:350–362.

WAYLAND, M., AND T. BOLLINGER. 1999. Lead exposure and poisoning in Bald Eagles and Golden Eagles in the Canadian prairie provinces. Environmental Pollution 104:341–350.