REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

Public Reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comment regarding this burden estimates or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188,) Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 18, 2009	Final progress (Feb 15, 2005 - Feb 14, 2009)

4. TITLE AND SUBTITLE
Tree-structured methods for prediction and data visualization

5. FUNDING NUMBERS
W911NF-05-1-0047

6. AUTHOR(S)
Wei-Yin Loh

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Wisconsin, 1300 University Avenue, Madison, WI 53706

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
45847-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The aim of the research is to develop the GUIDE algorithm into a fast, powerful, and comprehensive procedure for tree-structured prediction and data visualization. During this reporting period, the regression component was enhanced by the addition of least squares regression through the origin, best simple analysis of covariance, all subsets regression, and least median of squares regression. An option to truncate the predicted values also was added. A preliminary classification tree component included kernel and nearest-neighbor node modeling. Numerous improvements were made to the algorithms for split and variable selection and importance scoring. GUIDE now supersedes the older CRUISE and QUEST algorithms. The GUIDE computer program had three major revisions and continues to be distributed for free over the Internet. Two PhDs were graduated and fifteen papers published or accepted for publication during this period.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Statistics, classification, regression, decision trees	16
	16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	**UNCLASSIFIED**	**UNCLASSIFIED**	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev.2-89)
Prescribed by ANSI Std. 239-18
298-102

Report Title

Final Progress Report

ABSTRACT

The aim of the research is to develop the GUIDE algorithm into a fast, powerful, and comprehensive procedure for tree-structured prediction and data visualization. During this reporting period, the regression component was enhanced by the addition of least squares regression through the origin, best simple analysis of covariance, all subsets regression, and least median of squares regression. An option to truncate the predicted values also was added. A preliminary classification tree component included kernel and nearest-neighbor node modeling. Numerous improvements were made to the algorithms for split and variable selection and importance scoring. GUIDE now supersedes the older CRUISE and QUEST algorithms. The GUIDE computer program had three major revisions and continues to be distributed for free over the Internet. Two PhDs were graduated and fifteen papers published or accepted for publication.

List of papers submitted or published that acknowledge ARO support during this reporting period. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Kara, A. B., Wallcraft, A. J., Hurlburt, H. E., and Loh, W.-Y. (2009). Which surface atmospheric variable drives the seasonal cycle of sea surface temperature over the global ocean? Journal of Geophysical Research, vol. 114.

Kara, A. B., Wallcraft, A. J., Hurlburt, H. E., and Loh, W.-Y. (2009). Quantifying SST errors from an OGCM in relation to atmospheric forcing variables. Ocean Modeling, in press.

Number of Papers published in peer-reviewed journals: 2.00

(b) Papers published in non-peer-reviewed journals or in conference proceedings (N/A for none)

Number of Papers published in non peer-reviewed journals: 0.00

(c) Presentations

New Developments in Classification Trees. Presented at the 2008 Army Conference in Applied Statistics, Virginia Military Institute.

Number of Presentations: 1.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts): 0

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Loh, W. and Zheng, W. (2009). On bootstrap tests of hypotheses. Proceedings of the Third Erich L. Lehmann Symposium, IMS Lecture Notes-Monograph Series, in press.

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts): 1

(d) Manuscripts

Loh, W.-Y. (2008). Improving the precision of classification trees. Submitted to Annals of Applied Statistics.

Piper, M. E., Loh, W.-Y., Smith, S. S., Japuntich, S. J., and Baker, T. B. (2008). Using decision tree analysis to identify risk factors for relapse to smoking. Submitted to Experimental and Clinical Psychopharmacology.

Barron, C. N., Kara, A. B., Rowley, C., Gentemann, C. L., and Loh, W.-Y. (2008). Time scales of SST variability over the global ocean. Submitted to Journal of Climate.

Gunduz, M., Kara, A. B., Barron, C. N., and Loh, W.-Y. (2008). The link between climate indices and SST over the Caspian Sea. Submitted to Journal of Climate.

Number of Manuscripts: 4.00

Number of Inventions:

Graduate Students

NAME	PERCENT_SUPPORTED
Wei-Yin Loh	0.00
Xu He	0.14
Chien-Wei Chen	0.04
Chia-Chieh Lin	0.17
Wei Zheng	0.11
FTE Equivalent:	**0.46**
Total Number:	**5**

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Wei-Yin Loh	0.00	No
FTE Equivalent:	**0.00**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period:	0.00
The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....	0.00
The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....	0.00
Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):	0.00
Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:	0.00
The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense	0.00
The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:	0.00

Names of Personnel receiving masters degrees

NAME
Wei Zheng
Total Number: 1

Names of personnel receiving PHDs

NAME
Youngjae Chang
Chien-Wei Chen
Total Number: 2

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

1 Scientific progress and accomplishments

1.1 Regression trees

The following improvements and extensions were added to the regression tree component of the GUIDE algorithm.

Least median of squares (LMS) regression. Possessing the highest possible breakdown point, this method was originally used by Rousseeuw (1984) to fit a plane to a set of data. The option gives GUIDE the capability of producing highly robust piecewise LMS regression models.

Regression without intercept. An option was added to allow the fitting of piecewise multiple linear least squares models without intercept terms. This was motivated by an application involving global weather data where a physical model requires a zero intercept term in the model. The paper is published in Kara et al. (2007).

Control of extrapolation errors. With the exception of piecewise constant models, all regression models can potentially produce large prediction errors if an observation falls outside the range of the training sample. The effect of such errors is magnified if prediction accuracy is measured in terms of squared error. To control the effect, GUIDE now offers five options for extrapolation: truncation within the range of the training sample in a node, truncation within ten percent of the range in a node, truncation within the range of the whole training sample, and, one- and two-sided Winsorization. The results are published in Loh et al. (2007).

Simplification of polynomial models. For piecewise polynomial modeling, GUIDE now fits the smallest statistically significant polynomial, with the significance level being user-specified. Thus, if a piecewise cubic polynomial is desired but the cubic term in a node is not significant, a quadratic is fitted in its place.

Simple ANCOVA models. When there are categorical predictor variables in the data, GUIDE can fit an analysis of covariance (ANCOVA) model with stepwise variable selection of the dummy variables, in each node. This allows the effects of categorical predictors to be modeled within each node using only the important dummy variables. A desirable

side effect is the potential for a simpler and more interpretable tree structure.

All-subsets regression. This option was added to stepwise variable selection for least squares modeling. Now GUIDE can perform stepwise variable selection via forward only selection, forward-and-backward selection, and all-subsets selection. Empirical evidence indicates that the all-subsets option yields slightly better prediction accuracy.

1.2 Classification trees

A classification tree capability was added to the GUIDE algorithm. As a result, the computer source code for the classification and regression algorithms is under the complete control of the PI. Previously, the PI controlled the code for regression while two of his former students controlled the code for the CRUISE (Kim and Loh, 2001, 2003) and QUEST (Loh and Shih, 1997) classification tree algorithms. That arrangement made it very difficult to make changes to the classification algorithms.

The task of expanding GUIDE to handle classification and regression problems was a major effort, because it required modifications to much of the existing code (about 50,000 lines) and new routines specific to classification to be added. But the opportunity also allowed numerous improvements to be made to the CRUISE and QUEST algorithms. As a result, the new GUIDE algorithm is more versatile than CRUISE and it renders QUEST obsolete.

The major features and properties of the GUIDE classification tree algorithm are as follows.

Split selection. GUIDE's split selection strategy is more intelligent than that of previous classification tree algorithms. It is practically unbiased like CRUISE and QUEST, but the splits are designed to be more accurate when local interaction effects are strong and marginal effects are weak. In such situations, all other algorithms fail to split correctly.

Predictive accuracy. Increased predictive accuracy is a direct beneficiary of the improved split selection strategy.

Kernel and nearest-neighbor models. To further increase accuracy, an option to GUIDE was added to fit kernel or nearest-neighbor models to the nodes of the tree.

Table 1: Abbreviated algorithm names used in tables and plots

C45	C4.5
C2d	CRUISE with simple node models
C2v	CRUISE with linear discriminant node models
Qu	QUEST with univariate splits
Ql	QUEST with linear splits
Rp	RPART
Ct	CTree
S	GUIDE with simple node models
K	GUIDE with kernel node models
N	GUIDE with nearest-neighbor node models

Computational speed. While improvements in the split selection strategy required additional computation, opportunities for short-cuts were found so that the average computational speed of GUIDE is better than the algorithms it replaces.

Precision of tree structures. The area where the new GUIDE truly excels is in the compactness of the tree structures. This is very important for two reasons: (i) a compact tree is more comprehensible, and (ii) a non-compact tree often contains irrelevant predictor variables that are inevitably mistaken to be important. The tendency of many well-known algorithms to produce overly large trees is the main reason their use is not more widespread.

The plots in Figure 1 compare GUIDE against CRUISE, QUEST, C4.5 (Quinlan, 1993), CTree (Hothorn et al., 2006), and RPART (Atkinson and Therneau, 2000) in terms of predictive accuracy, compactness of the trees, and computational time on 46 data sets. A legend for the plot symbols is given in Table 1. The best algorithms in terms of accuracy and tree size are the four in the bottom-left corner of the left panel in the figure. Three of them are from GUIDE (K, N, and S). The other algorithm is QUEST with linear splits, but the right panel shows that it takes twice as long on average as GUIDE to execute. Besides, linear splits are much harder to interpret than univariate splits. More details are provided in Loh (2008), which is submitted for publication.

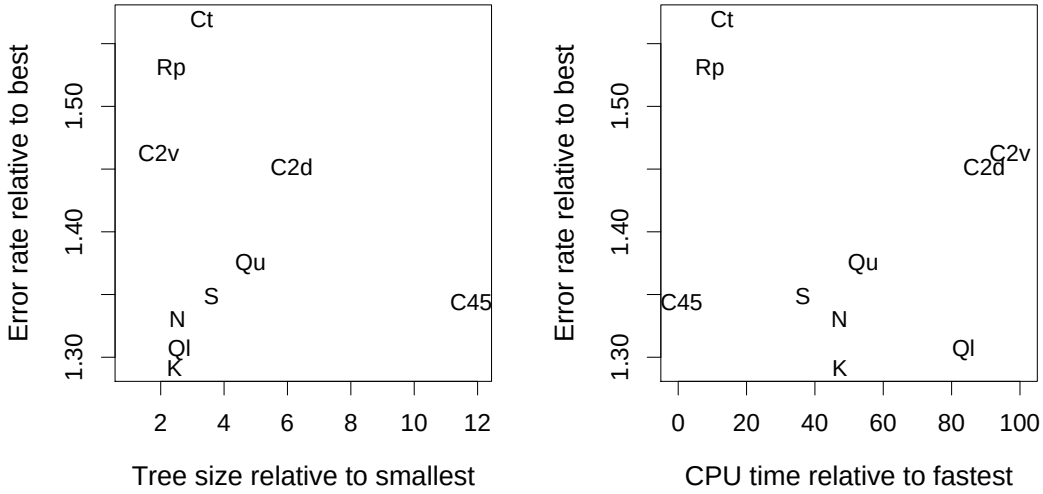

Figure 1: Plots of average relative error rates versus average relative tree sizes and average relative error rates versus average relative computational times. Relative error rates are ratios of error rates to the smallest error rate for each data set. Relative tree size is the ratio of the number of leaf nodes of an algorithm to the smallest number of leaf nodes for each data set. Relative computational time is the ratio of computational time to the smallest computational time for each data set.

1.3 Split and subset selection, and importance ranking

The following improvements apply to both the classification and regression components of GUIDE.

Split variable selection. In the old version of the GUIDE algorithm, two sets of hypothesis tests were carried out on the residuals in each node. One set (called curvature tests) tests for association between the signs of the residuals and the values of each predictor variable. The other set (called interaction tests) tests for association between the signs of the residuals and value-pairs of each pair of predictor variables. If there are k predictor variables, the number of curvature tests is k and the number of interaction tests is $k(k-1)/2$. This creates a bias in favor of a interaction test being found to be most significant. The bias is now eliminated by using a two-stage procedure as follows. Given α, perform the curvature tests and see whether the smallest significance probability is less than the Bonferroni-corrected level of α/k. If it is, choose the variable with the smallest probability and do not perform the interaction tests. Otherwise, perform the interaction tests and see if its smallest significance probability is less than $2\alpha/[k(k-1)]$. If it is, choose the split variable from the pair that has the smallest probability. Otherwise, choose the variable with the smallest probability from the curvature tests. This technique eliminates the bias as well as significantly increases the computational speed of the algorithm.

Split value selection. In the old version, if an interaction test is found significant, the split variable is chosen from the pair of variables involved by separately searching for the best split point on each. This can yield poor splits because it does not take advantage of the interaction effect between the two variables. In the new version, the best split is found using a two-step look-ahead strategy by partitioning the data into four subsets: first split the data into two subsets on one variable and then split each subset into two more by splitting on the other variable. This technique is highly compute-intensive if both variables are categorical with many categories each. An approximate solution employing a computational trick is used instead. Simulation experiments show that the approximate method is highly effective.

Importance ranking and subset selection. If the number of variables k is larger than the number of data points n, many methods, including

multiple linear regression, cannot be applied. Further, empirical results show that the prediction accuracy of even those algorithms than have built-in variable selection (such as GUIDE, MARS, and Random forest) deteriorates as the number of irrelevant variables (i.e., variables that have no effect on the regression function) increases. One way to slow the deterioration rate is to carry out a preliminary variable selection step before application of the respective method. The GUIDE algorithm was modified to perform this task. The idea is to convert the significance probabilities of the above-mentioned curvature and interaction tests into chi-squared values and then take a weighted sum (with weights being square roots of node sample sizes) of the chi-squared values over the intermediate nodes of the tree. This yields a ranking of the variables according to their "importance." To find a threshold for separating the truly important variables from the irrelevant ones, a critical value for the distribution of the weighted sum of chi-squares is obtained using the Satterthwaite approximation. Experimental results based on simulated data show that this approach is very effective in maintaining the prediction accuracy of GUIDE, MARS and Random forest for k as large as five times n, in a variety of models.

1.4 Applications

The PI was invited to use his algorithms to assist three teams of researchers. The first is a team of medical faculty from the University of Wisconsin, the second is a team of atmospheric scientists from the NASA Stennis Space Center, and the third is a team of civil engineers from the Wisconsin Department of Transportation.

Smoking cessation. This application involved the analysis of a set of clinical trial data on smoking cessation. About 900 smokers were randomized to receive a nicotine treatment or a placebo and their smoking status was recorded at three time points: one week after start of treatment, at the end of treatment, and six months later. More than 70 variables were recorded for each smoker, including prior smoking habits, attempts at quitting, medical and psychological health scores, income, gender, race, and marital status.

The traditional approach to modeling the probability of smoking abstinence for such data is logistic regression. There being $70 \times 69/2 = 2415$

two-factor interactions, it is impossible to fit a full second-order logistic model. Higher-order interactions are also out of the question. Stepwise logistic regression is not helpful because the resulting regression coefficients are biased and are tricky to interpret. Even if a logistic regression model is fitted, it cannot explain which variables are most important for predicting smoking abstinence.

GUIDE, on the other hand, can model interactions of any order. Further, it directly addresses the investigators' goal, since it optimizes prediction accuracy, whereas stepwise logistic regression optimizes a penalized likelihood function. GUIDE finds that degree of smoking dependence, marital status, and age at which smoking began are the most important variables. Figure 2 shows the GUIDE tree for predicting one-week abstinence. The complete results are reported in Piper et al. (2009), which is submitted for publication.

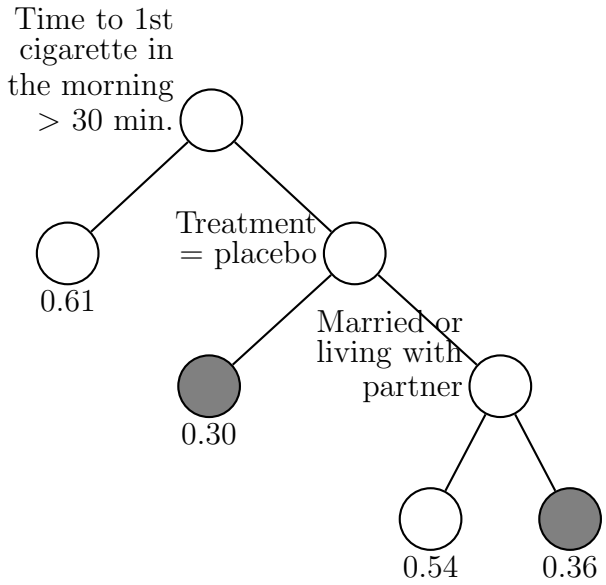

Figure 2: GUIDE classification tree for predicting smoking abstinence after one week of treatment. An observation goes to the left node if and only if the stated condition is satisfied. The number beneath a leaf node is the estimated probability of abstinence. Nodes with estimated probabilities less than the overall probability of 0.46 are shaded gray.

Global ocean temperature. The PI collaborated with a group of geo-

physicists from NASA to model sea-surface temperature over the earth's oceans. Owing to the presence of complex spatial and temporal influences, it is very difficult to find a mathematical model that accurately predicts the temperature. It is even more difficult to infer from the model the relative importance of the predictor variables. By using piecewise linear models, GUIDE can adapt well to local conditions. Besides, the piecewise linear functions are easy to interpret. Further, the GUIDE algorithm can be used repeatedly with one or more predictor variables excluded from the model to study the effects of their exclusion. The collaboration allowed the PI to test the GUIDE computer program on large data sets of the order of several hundred megabytes (the meteorological measurements were recorded at one-degree intervals over the earth's oceans monthly for several years). Two papers have been published (Kara et al., 2007, 2009b), one accepted for publication (Kara et al., 2009a), and and two submitted for publication (Barron et al., 2008; Gunduz et al., 2009).

Snow storm operations. The third collaboration involves using GUIDE to fit regression models to winter driving data. The purpose is to find a suitable measure to evaluate the effectiveness of maintenance operations during snow storms. The data are derived from a sample of 954 winter storm reports in 24 Wisconsin counties over three years. Each data record pertains to one snow storm in one Wisconsin county. It includes information on the duration of the storm, type and amount of precipitation, pavement temperature, starting and ending time of maintenance activities, and the maximum speed reduction during the storm. The GUIDE models confirm previous smaller studies that find vehicle speed to be a good indicator of winter driving conditions during snow storms. Speed recovery duration (the amount of time for average vehicle speed to return to its pre-storm level) is identified as an effective metric for success of winter maintenance operations. The results are published in Lee et al. (2008).

1.5 PhD student research

Two students, Y. Chang and C.-W. Chen, completed their PhD degrees under the PI's supervision during the project period.

Chang (2008) studied the iterative application of GUIDE for solving classification and regression problems when there are many irrelevant variables. At each iteration, the variables selected by GUIDE are removed from the data set. Iteration stops when no more variables are selected. Then the data set is fitted once more with GUIDE, using only the removed variables. Results from real and simulated data show that this approach is more effective than Random Forest and EARTH (Doksum et al., 2008) for classification as well as linear and quantile regression.

Bagging is a technique for increasing the prediction accuracy of a mediocre procedure by applying the latter multiple times using bootstrap samples and then averaging the predictions. It was originally proposed by Breiman (1996), who applied it to piecewise constant classification trees. The technique is now called "random forest" (Breiman, 2001) and it has been extended to regression. Empirical results obtained by the PI shows that a single GUIDE piecewise multiple linear regression tree can possess higher average prediction accuracy than a random forest of 500 piecewise constant trees. Chen (2008) studied the performance of an ensemble of GUIDE trees as well as enhancements to the random forest algorithm, such as the inclusion of linear combination splits.

1.6 Technology transfer

The PI participated in every U.S. Army Annual Statistics Conference during the grant period. He delivered the keynote address at the conference held in Monterey, CA, in October 2005.

He visited Dr. Barry Bodt and his staff at the Tactical Collaboration & Data Fusion Branch of CISD in Aberdeen on 19 July 2006. The PI gave a 3.5-hour talk on data mining with classification and regression trees to 14 staff members in the morning and met individually with Barry Bodt, David Webb, Timothy Hanratty, and others in the afternoon to consult on their projects. According to Bodt, one purpose of the visit was to expose his computer science staff to a more statistical approach to data mining. Bodt expects that the PI's work could assist in the development of several current projects at his lab, including soft target exploitation, data fusion, and network enabled command and control. Staff member Joan Forester started to experiment with version 4.0 of the GUIDE computer program immediately after the PI's visit.

The PI also participated in a statistics workshop organized by COL Rod-

ney Sturdivant at the U.S. Military Academy in April 2008. He subsequently consulted with COL Sturdivant on the application of GUIDE to a genetics problem. The data come from an experiment where 146 cell cultures are treated with three factors, each at several levels. The objective is to find out how these factors and their levels affect the response on 25,000 genes. This is a formidable problem because of the large number of genes and the small number of samples.

The PI has a continuing collaboration with Dr. A. B. Kara and his colleagues at the Stennis Space Center, Naval Research Laboratory, on statistical modeling of global ocean weather patterns.

The GUIDE computer program went from version 4 through version 7 during the period of the grant. Versions compiled for Linux, Macintosh, and Windows operating systems continue to be distributed for free from the website `www.stat.wisc.edu/~loh/guide.html`, which receives about a hundred hits each week.

A search of the *Web of Knowledge* reveals more than thirty papers using the PI's algorithms have been published by other researchers in the last eight years. The papers and their scientific areas are:

Biology: Jones et al. (2008), Olden and Jackson (2002)

Chemistry: Khlebnlkov et al. (2008), Bertelli et al. (2007)

Climate: Stahl (2005), Connor and Woodcock (2000)

Computer Science: Goddard and MacKinney-Romero (2006), Okura et al. (2002)

Economics: Kannebley et al. (2005), Balaras et al. (2005)

Engineering: Juni et al. (2008), Qin and Han (2008), Adams et al. (2006), Cartmell et al. (2005)

Genomics: Heidema and Nagelkerke (2008), Wei et al. (2008), Cho et al. (2007), Moon et al. (2006)

Geography: Schmidt et al. (2008), Sesnie et al. (2008), Archibald and Scholes (2007), Behrens and Scholten (2006), Sullivan et al. (2006), Pal and Mather (2003)

Medicine: Hoque et al. (2006), Royall et al. (2005), Baeten et al. (2004), Huang et al. (2004), Ibanez et al. (2004), Kedia and Williams (2003), Pryse-Phillips et al. (2002)

References

Adams, T. M., Juni, E., M., S. and Xu, L. (2006). Regression tree models to predict winter storm costs, *Management and Delivery of Maintenance and Operations Services* **18**(1948): 117–124.

Archibald, S. and Scholes, R. J. (2007). Leaf green-up in a semi-arid African savanna — separating tree and grass responses to environmental cues, *Journal of Vegetation Science* **18**: 583–594.

Atkinson, E. J. and Therneau, T. M. (2000). An introduction to recursive partitioning using the RPART routines, *Technical report*, Mayo Foundation.

Baeten, D., Kruithof, E., De Rycke, L., Vandooren, B., Wyns, B., Boullart, L., Hoffman, I. E. A., Boots, A. M., Veys, E. M. and De Keyser, F. (2004). Diagnostic classification of spondylarthropathy and rheumatoid arthritis by synovial histopathology — a prospective study in 154 consecutive patients, *Arthritis and Rheumatism* **50**(9): 2931–2941.

Balaras, C. A., Droutsa, K., Dascalaki, E. and Kontoyiannidis, S. (2005). Deterioration of European apartment buildings, *Energy and Buildings* **37**: 515–527.

Barron, C. N., Kara, A. B., Gentemann, C. L. and Loh, W.-Y. (2008). Time scales for SST over the global ocean, *Journal of Climate* . Submitted for publication.

Behrens, T. and Scholten, T. (2006). Digital soil mapping in Germany – a review, *Journal of Plant Nutrition and Soil Science* **169**: 434–443.

Bertelli, D., Plessi, M., Sabatini, A. G., Lollo, M. and Grillenzoni, F. (2007). Classification of Italian honeys by mid-infrared diffuse reflectance spectroscopy (DRIFTS), *Food Chemistry* **101**: 1582–1587.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

Breiman, L. (2001). Random forests, *Machine Learning* **45**: 5–32.

Cartmell, J., Enoch, S., Krstajic, D. and Leahy, D. E. (2005). Automated QSPR through competitive workflow, *Journal of Computer-Aided Molecular Design* **19**: 821–833.

Chang, Y. (2008). *Robustifying regression and classification trees in the presence of irrelevant variables*, PhD thesis, Department of Statistics, University of Wisconsin, Madison.

Chen, C.-W. (2008). *Enhancing the prediction accuracy of regression trees: linear splits and variable selection*, PhD thesis, Department of Statistics, University of Wisconsin, Madison.

Cho, W. C. S., Yip, T. T. C., Ngan, R. K. C., Yip, T.-T., Podust, V. N., Yip, C., Yiu, H. H. Y., Yip, V., Cheng, W.-W., Ma, V. W. S. and Law, S. C. K. (2007). ProteinChip array profiling for identification of disease- and chemotherapy-associated biomarkers of nasopharyngeal carcinoma, *Clinical Chemistry* **53**: 241–250.

Connor, G. J. and Woodcock, F. (2000). The application of synoptic stratification to precipitation forecasting in the trade wind regime, *Weather and Forecasting* **15**(3): 276–297.

Doksum, K., Tang, S. and Tsui, K. (2008). Nonparametric variable selection: the EARTH algorithm, *Journal of the American Statistical Association* **103**: 1609–1620.

Goddard, J. and MacKinney-Romero, R. (2006). Finding Spanish syllabification rules with decision trees, *Lecture Notes in Artificial Intelligence*, Vol. 4139, pp. 333–340.

Gunduz, M., Kara, A. B., Barron, C. N. and Loh, W.-Y. (2009). The link between climate indices and sea-surface temperature over the Caspian Sea, *Journal of Climate* . Submitted.

Heidema, A. G. and Nagelkerke, N. (2008). Developing a discrimination rule between breast cancer patients and controls using proteomics mass spectrometric data: a three-step approach, *Statistical Applications in Genetics and Molecular Biology* **7**(2).

Hoque, M. O., Feng, Q. H., Toure, P., Dem, A., Critchlow, C. W., Hawes, S. E., Wood, T., Jeronimo, C., Rosenbaum, E., Stern, J., Yu, M. J., Trink, B., Kiviat, N. B. and Sidransky, D. (2006). Detection of aberrant methylation of four genes in plasma DNA for the detection of breast cancer, *Journal of Clinical Oncology* **24**: 4262–4269.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework, *Journal of Computational and Graphical Statistics* **15**: 651–674.

Huang, J., Lin, A., Narasimhan, B., Quertermousd, T., Hsiung, C. A., Ho, L.-T., Grove, J. S., Olivier, M., Ranade, K., Risch, N. J. and Olshen, R. A. (2004). Tree-structured supervised learning and the genetics of hypertension, *Proceedings of the National Academy of Sciences of the United States of America* **101**: 10529–10534.

Ibanez, J., Arikan, F., Pedraza, S., Sanchez, E., Poca, M. A., Rodriguez, D. and Rubio, E. (2004). Reliability of clinical guidelines in the detection of patients at risk following mild head injury: results of a prospective study, *Journal of Neurosurgery* **100**: 825–834.

Jones, C. B., Milanov, V. and Hager, R. (2008). Predictors of male residence patterns in groups of black howler monkeys, *Journal of Zoology* **275**(1): 72–78.

Juni, E., Adams, T. M. and Sokolowski, D. (2008). Relating cost to condition in routine highway maintenance, *Transportation Research Record* (2044): 3–10.

Kannebley, S., Porto, G. S. and Pazello, E. T. (2005). Characteristics of Brazilian innovative firms: an empirical analysis based on PINTEC — industrial research on technological innovation, *Research Policy* **34**: 872–893.

Kara, A. B., Hurlburt, H. E. and Loh, W.-Y. (2007). Which near-surface atmospheric variable drives air-sea temperature differences over the global ocean?, *Journal of Geophysical Research* **112**: C05020.

Kara, A. B., Wallcraft, A. J., Hurlburt, H. E. and Loh, W.-Y. (2009a). Quantifying SST errors from an OGCM in relation to atmospheric forcing variables, *Ocean Modelling* . In press.

Kara, A. B., Wallcraft, A. J., Hurlburt, H. E. and Loh, W.-Y. (2009b). Which surface atmospheric variable drives the seasonal cycle of sea surface temperature over the global ocean?, *Journal of Geophysical Research* **114**: D05101.

Kedia, S. and Williams, C. (2003). Predictors of substance abuse treatment outcomes in Tennessee, *Journal of Drug Education* **33**: 25–47.

Khlebnlkov, A. I., Schepetkin, I. A. and Quinn, M. T. (2008). Structure-activity relationship analysis of N-benzoylpyrazoles for elastase inhibitory activity: a simplified approach using atom pair descriptors, *Bioorganic and Medicinal Chemistry* **16**(6): 2791–2802.

Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits, *Journal of the American Statistical Association* **96**: 589–604.

Kim, H. and Loh, W.-Y. (2003). Classification trees with bivariate linear discriminant node models, *Journal of Computational and Graphical Statistics* **12**: 512–530.

Lee, C., Loh, W.-Y., Qin, X. and Sproul, M. (2008). Development of new performance measure for winter maintenance using vehicle speed data, *Transportation Research Record* **2055**: 89–98.

Loh, W.-Y. (2008). Improving the precision of classification trees, *Annals of Applied Statistics* . Submitted for publication.

Loh, W.-Y., Chen, C.-W. and Zheng, W. (2007). Extrapolation errors in linear model trees, *ACM Trans. Knowl. Discov. Data* **1**(2): 6.

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees, *Statistica Sinica* **7**: 815–840.

Moon, H., Ahn, H., Kodell, R. L., Lin, C. J., Baek, S. and Chen, J. J. (2006). Classification methods for the development of genomic signatures from high-dimensional data, *Genome Biology* **7**(R121).

Okura, Y., Matsumura, Y., Harauchi, H., Sukenobu, Y., Kou, H., Kohyama, S., Yasuda, N., Yamamoto, Y. and Inamura, K. (2002). An inductive method for automatic generation of referring physician prefetch rules for PACS, *Journal of Digital Imaging* **14**: 226–231.

Olden, J. D. and Jackson, D. A. (2002). A comparison of statistical approaches for modelling fish species distributions, *Freshwater Biology* **47**(10): 1976–1995.

Pal, M. and Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification, *Remote Sensing of Environment* **86**: 554–565.

Piper, M. E., Loh, W.-Y., Smith, S. S., Japuntich, S. J. and Baker, T. B. (2009). Using decision tree analysis to identify risk factors for relapse to smoking, *Experimental and Clinical Psychopharmacology* . Submitted.

Pryse-Phillips, W., Gawel, M. A. M., R. Nelson, A. P. and Wilson, K. (2002). A headache diagnosis project, *Headache: The Journal of Head and Face Pain* **42**: 728–737.

Qin, X. and Han, J. (2008). Variable selection issues in tree-based regression models, *Transportation Research Record* (2061): 30–38.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo.

Rousseeuw, P. J. (1984). Least meadian of squares regression, *Journal of the American Statistical Association* **79**: 871–880.

Royall, D. R., Chiodo, L. K. and Polk, M. J. (2005). An empiric approach to level of care determinations: the importance of executive measures, *Journals of Gerontology Series A – Biological Sciences and Medical Sciences* **60**: 1059–1064.

Schmidt, K., Behrens, T. and Scholten, T. (2008). Instance selection and classification tree analysis for large spatial datasets in digital soil mapping, *Geoderma* **146**: 138–146.

Sesnie, S. E., Gessler, P. E., Finegan, B. and Thessler, S. (2008). Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments, *Remote Sensing of Environment* **112**(5): 2145–2159.

Stahl, K. (2005). Influence of hydroclimatology and socioeconomic conditions on water-related international relations, *Water International* **30**: 270–282.

Sullivan, M. S., Jones, M. J., Lee, D. C., Marsden, S. J., Fielding, A. H. and Young, E. V. (2006). A comparison of predictive methods in extinction risk studies: contrasts and decision trees, *Biodiversity and Conservation* **15**: 1977–1991.

Wei, H., Kuan, P. F., Tian, S., Yang, C., Nie, J., Sengupta, S., Ruotti, V., Jonsdottir, G. A., Keles, S., Thomson, J. A. and Stewart, R. (2008). A study of the relationships between oligonucleotide properties and hybridization signal intensities from NimbleGen microarray datasets, *Nucleic Acids Research* **36**(9): 2926–2938.