POLICY CAPTURING USING DECISION TREES:
AN ANALYSIS OF EPA RULE-MAKING

Aparajita (Molly) Bhattacharyya, Ph.D candidate
Department of Agricultural and Consumer Economics,
University of Illinois at Urbana-Champaign

Contact Address:
# 3205, 10 East Ontario
Chicago, IL 60611

Phone: 312-867-0469
Email:tahbildar@yahoo.com

Annual Meetings, August 8-11th, 1999
American Agricultural Economics Association

POLICY CAPTURING USING DECISION TREES:
AN ANALYSIS OF EPA RULE-MAKING

Aparajita (Molly) Bhattacharyya,
Department of Agricultural and Consumer Economics,
University of Illinois at Urbana-Champaign

A major task of economics and policy economics is to explain the pattern of government behavior in its interventions in the economy. Most empirical evaluations of government behavior have used the "Revealed Preference" approach. This is seen both in the analysis of trade policies (Oehmke and Yao 1990, Gordon and Freebairn, 1975, among others) as well as government intervention in such areas as environmental regulation (Gupta et al. 1993, Cropper et al 1992, Magat, Harrington and Krupnik, 1986), railroad closures (Weingast and Moran, 1983), freeway siting (D..McFadden, 1976), and product safety issues (G.L.Thomas, 1985). Discrete decisions such as whether or not to regulate a commodity, as well as continuous choices such as the allowable pollution levels have been analyzed using this approach. Multivariate regression, including binary and multinomial logit and probit are the usual tools employed.

This paper introduces the use of the Decision Tree methodology to help capture policy making behavior and provide extended insights into the decision making process. Decision Tree is a flexible, non-parametric, data-driven procedure, which has been in use since the early 1970's, both as a classification and as a prediction tool; its use in economics has so far been rare, and in modeling regulatory agency behavior in decision-making has been virtually non-existent. Use of this approach not only reveal the significant conditions driving a phenomenon, but also provide a ranking of the conditions

- information not readily available from the regression methods (logit, probit, discriminant analysis) currently used for discrete choice issues. Given a resource-constrained environment, this is expected to be a significant insight with respect to allocation of scarce resources. Variable importance as well as the threshold values at which they are informative is revealed by this procedure.

EPA's rulings on pesticides which came up for Special Review between 1975 and 1995 are analyzed using a decision tree methodology (CART). Section II briefly describes the decision tree methodology. Section III lays out the application domain. Section IV discusses the analysis and results. Section V concludes the discussion.

II. DECISION TREES

The decision tree approach is a statistical approach that creates rules and rule-trees by searching through data for statistical patterns and relationships, and uses information about the distribution of the data to split the data into finer and finer subsets with each subset being progressively "purer" than the previous. The process is called 'Recursive Partitioning'. At each node of the tree, the search process searches through the variables (x_1, x_2, … x_m), one by one. For each variable it finds the best split. Then it compares the m best splits and selects the best of the best. This is done over and over again on each resulting subset or cluster, iteratively, until a tree is built. Each descendent cluster in the tree is 'purer' than the parent cluster , impurity being highest when all the classes are equally mixed in the cluster and lowest when all the data in the cluster is of a single class.

Several measures of impurity have been developed for use. For classification problems [1], the three available measures [2] are:-(i). the Entropy criterion, derived from the mathematics of Information Theory; (ii). the Gini criterion, based on the Fisher's Chi-square; (iii). the Twoing criterion. Choice of the measure usually depends on which measure gives the highest classification accuracy for the particular problem. The the mathematics of the process is well documented and well accepted [3], and is not repeated here. The objective of the process is to maximize the gain in purity from each successive split or alternatively to minimize the overall tree impurity (Breiman, et al, 1984).

Decision trees are particularly useful in bringing out non-homogeneity in the data set, that is, it reveals the different relationships that hold between variables in different parts of the measurement space. As such, it provides clues to the structure of the problem not apparent in regression analysis, and is therefore especially relevant to any attempt to understanding rule-making influences.

Also, the manner in which tree structured analysis handles outliers and mislabeled cases is very relevant in problems where the information set is small, as is usually the case in regulatory decision making. In such situations, every case in the data set is important in the insights it offers. Elimination of even a single case in order to correct for the problem of outliers could result in the loss of valuable information. Decision trees are extremely robust with respect to outliers and mislabeled cases. Outliers are treated in a way that both minimizes their impact, and signals their presence. Each case is weighed as only one

[1] Impurity criterion for continuous variable problems differ from discrete or classification problems.

[2] The interested reader may look up Classification, Analysis and Regression Trees by Breiman, et al.

among n different cases and results are therefore not unduly affected by a few cases. Outliers are isolated into very small nodes. To this extent, results are less subject to distortions from them than standard regressions. And information is not lost.

These features of such an analysis, as also the ease of interpreting the results obtained, motivate the use of this methodology for analyzing the problem detailed in the next section.

III. PROBLEM DESCRIPTION

Overview of the Special Review process: The Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) provides for the Special Review (SR) of pesticides, which may be initiated if the Environmental Protection Agency (EPA) has reasons to believe that the use of a pesticide may result in unreasonable adverse effects to people or the environment[4]. The process generally begins with an internal agency review (often with inputs from the registrant), culminating in a determination that the product has met a 'risk criterion' and that a SR should be initiated. The SR process then proceeds to determine whether risks incurred, justify the benefits derived from use of the chemical. The review process may be concluded in various ways depending on the outcome of EPA's risk-benefit assessment. In determining whether risks from use of the pesticide is greater than its benefits, EPA considers possible changes to the terms and conditions of registration which can reduce risks to a level where benefits outweigh risks. Alternatively, EPA may

[3] See Breimen et al, 1984 for details of the mathematical formulations.

[4] 40 CFR Parrt 154.7 of the FIFRA provides details on the criteria which need to be met to initiate SR proceedings

determine that no changes in the terms and conditions of a registration will adequately assure that the use of the pesticide will not pose any unreasonable adverse effects. If EPA makes such a determination , it may require a cancellation or even a suspension. Adversely affected parties, including registrants and applicants for registration, may request further hearings on the suspension or cancellation of a specified use or registration.

Thus, in order a make a decision regarding a pesticide under SR, EPA collects, and makes publicly available, information regarding the potential risks and benefits associated with a pesticide use, and receives comments from concerned parties in this respect. Since 1975, 117 pesticides have come under the SR process. Of these, as of date, 12 pesticides are still in process, 45 have been completed , 20 have been deferred to re-registration , 35 were voluntarily cancelled by the registrants and 6 were cancelled prior to the initiation of the SR process. This study uses data generated by the SR process. Only those chemicals on which the SR process has been completed and on which a final decision has been taken, and those chemicals posing carcinogenic risks are considered for the purpose of this study.

IV. DESCRIPTION OF THE DATA

The data used for this study was obtained from the Public Documents (PD) 1, PD 2/3 and PD 4, often published in the Federal Register by the United States government and always announced in the Federal register. The period covered is 1975-1996.

Benefits: Quantitative benefits estimates are available in only 52 percent of the cases considered. In many cases, a descriptive evaluation of benefits was given. The description was stated as negligible, minor, insignificant, major or significant. Since no reliable means could be determined to translate these to quantitative estimates or vice versa, benefits with such descriptive values were categorized as missing values. A dummy variable is created to account for the missing information and is created as follows (i) to represent benefits estimates missing but yield losses expected and (ii) to represent benefits estimates missing and no expected yield losses. Benefits are expressed in millions of dollars and range from a low of a loss of $2.13 million if continued to a high of $ 227 million lost if cancelled.

Risks: For cancer risks this is expressed as the probability of cancer from a lifetime of exposure to the chemical. It has median value of 4.9 x 10^{-9} and a maximum value of 4.2 x 10^{-3} for diet risks, a maximum value of 10^{-2} for applicator risks and mixer risks. It must be noted that risk estimates are missing in 31 percent of the cases for diet risks, 51 percent of cases for applicator risks and 70 percent of cases for mixer risks. To account for this missing information, dummy variables are created to represent whether or not the respective quantitative risk estimate is available. Among the other risks, wildlife risks and reproductive risks are included in the study, also as dummy variables to indicate the presence or otherwise of these risks. These two risks are included since an examination of the FR documents reflect these as concerns often commented on by environmentalists and academicians and are therefore expected to influence decision-making. Reproductive risks were reported for 33 percent of cases and wildlife risks for 58 percent of the cases analyzed.

Comments: For the purpose of this study, comments are represented as dummy variables to indicate the presence or absence of comments. Since the same party may have commented on several occasions and on several issues, and this information is not easily extractable from the public documents, no reliable representation could be made of the frequency of commenting. Only comments from farmers, environmentalists and academicians are represented here. SAP's comments are typically scientific critique and do not represent public participation. USDA's comments relate closely to farmers' comments. Hence, only farmers comments are considered here. Comments from consumer groups are extremely few and far between and therefore are not statistically viable for inclusion in the study. Of the comments included, environmentalists commented in 39 percent of cases, farmers in 16 percent of cases and academics in 25 percent of cases that came up for Special review.

The comment variable is incorporated in the study to analyze the impact of factors other than risks and benefits in the decision making process. An earlier study by Cropper et al, 1992, included a particular EPA administrator – Burford, as a separate variable in the study. Since the variable was found significant in that study, it is also included in this study. However, since this study covers a longer time period and therefore more political change, two other variables are used to reflect political and ideological influence. The first represents whether the ruling party is Republican or otherwise (GOP). The second represents presidents in power (Prez1-4). Burford was the EPA administrator for 75 of the 371 cases analyzed. Presidents Carter, Reagan, Bush and Clinton were presidents for 18, 62, 15 and 5 percent of total cases analyzed.

The data set consists of 371 decisions, of which 159 were to cancel and 212 were to continue. A list if the variables used in this study is provided in the appendix.

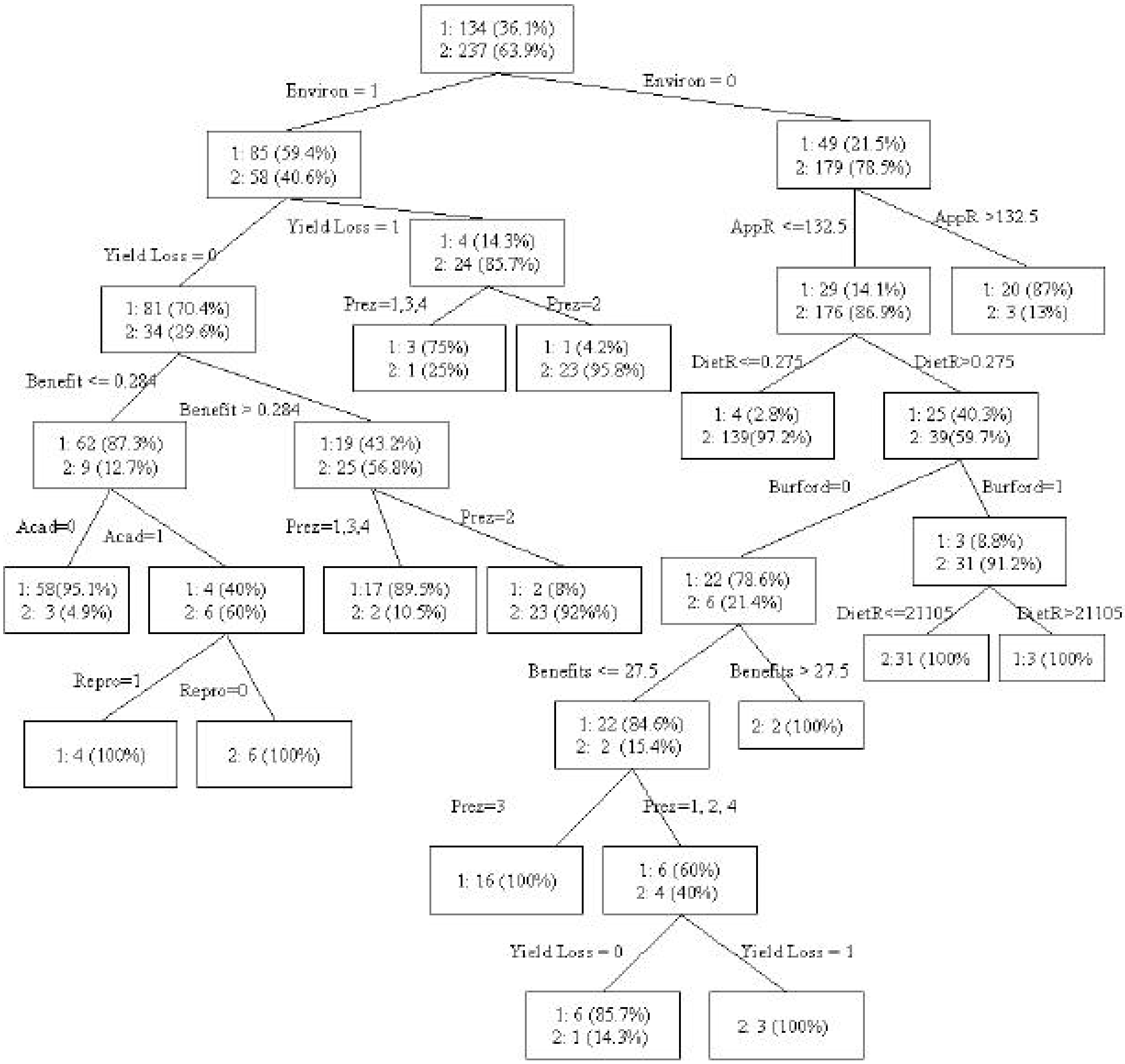
V. EMPIRICAL ANALYSIS OF EPA'S DECISIONS

FIFRA's mandate requires that EPA conduct a cost-benefit analysis prior to a decision being made on the use or cancellation of a chemical. A larger study conducted by the author, as well as a previous study by Cropper et al, concluded that risks and benefits are both significant determinants of the decision choice. All the political variables: participation by academics and environmentalists (but not farmers), Burford, political party in power and president in power are significant – pointing to the relevance of interest group influence as well as ideological influence.

To obtain further insights into the decision making process, a decision tree was estimated using the classification and regression tree (CART) procedure, popularised by Breimen, Freidman, Ohlsen and Stone. The findings from the tree analysis are as follows:

1. Participation by environmentalists appear to significantly influence the decision making strategy. Rules indicate that when environmentalists participate, the factors influencing decisions are benefits and the proxy for benefit estimates, yield loss, as well as participation by academics and the president in power. If yield losses are expected, chances for cancellation is seriously reduced. With yield losses, only 14 percent of cases are cancelled, the decisions being influenced by the president in power. During the Carter, Bush and Clinton era, 75percent of these decisions are cancelled compared to

Figure 1: Decision tree



Note: 1: cancel decision
2: continue decision

Example: Read the top-most box as: 134 cancel decision decisions and 237 continue decisions which represents 36.1% and 63.9% of the total cases (371) in that cluster.

For definitions and coding of the variables, please see appendix.

only 4 percent during the Reagan era. Without yield losses, 70 percent of cases are cancelled, the decisions being affected by the level of producer benefits, academic participation and the presence or otherwise of reproductive risks. Benefits less than \$284,000 result in a 87 percent cancellation, whereas if benefits are greater than \$284,000the deciding factor becomes then, the president in power. During Carter, Bush and Clinton, all 100 percent (the 2 cases of continue shown on the tree are missclassified) with benefits greater than \$284,000 (and Yield Loss and Environmentalists participation) are cancelled, compared to 100 percent cases being continued during Reagan reign.

When benefits are less than \$284,000, and academics do not participate, 95 percent of cases are cancelled. When academics participate, existence of reproductive risks split the decisions on cancellation 40:60. It may be pointed out that academics frequently comment on yield effects – both from cancellation of a pesticide use or substitution of a pesticide with another. Thus, this rule may be interpreted as academic participation implying higher expected yield losses, wherein then, cancellations are made only if reproductive risks exist.

2. Non-participation by environmentalists appears to lead to a different decision making strategy. When applicator risks are greater than 1.325 x10 $^{-4,}$ 98 percent of cases (2 of the three cases predicted as continue are missclassified) are cancelled. If applicator risks are less than 1.325 x 10 $^{-4}$, diet risks are less than 2.75 x 10 $^{-7}$ lead to a 97 percent cancellation rate$^{.}$ If diet risks are greater than 2.75 x 10 $^{-7}$, Burford is the discriminating factor. If Burford was present and diet risks are greater than 2.75 x 10 $^{-7}$, but less than 2.1 x 10 $^{-10}$, all cases are continued. If diet risks are greater than 2.1 x 10 $^{-10}$, all cases are cancelled.

If Burford was not the EPA administrator, and diet risks are greater than 2.75×10^{-7}, benefits (both producer benefits and yield losses) as well as ideology (presidents in power) become influencing factors. If benefits are greater than $27.5 million all cases are continued. If benefits are less than $27.5 million, and the president in power was Bush, all cases are cancelled. If Bush was not present , cases are decided on the presence or absence of yield losses. If there are no expected yield losses, cases are cancelled. Otherwise, cases are continued.

It thus appears that when environmentalists enter the rule making process, the principal influencing factors are benefits and proxies thereof. Risks seem to appear to be inconsequential. However, if environmentalists do not enter the decision making process, risks and benefits as well as political ideology (Presidents in power, Burford) become significant deciding factors. Overall, the cancellation rate is 60 percent when environmentalists comment and 22 percent otherwise. The overall accuracy (percent of cases correctly predicted) of the optimal tree is 94.61 percent. Stability of the tree was tested using a ten-fold cross-validation procedure.

VI. CONCLUSION

This study confirms earlier conclusions that risks and benefits and political variables, represented by participation variables and heads of state, are important determining factors in the discrete choice regulatory decision on pesticide regulation. It also reveals the implicit decision rules used for trade- offs between the different variables. It further points out the threshold values at which the variables are informative. The major policy implications are that given the importance of the participation variables and risk estimates, greater resources and attention needs to be focussed on better managing these two aspects of the rule-making process.

Appendix:

The following gives definations of the variables used and the construction of the variables.

1. Diet risk: is expressed as the probability, per million persons, of developing cancer for a lifetime of exposure (70 years)to the chemical .
2. Applicator risk: is expressed as the probability, per million persons, of developing cancer for a lifetime of exposure (35 years)to the chemical.
3. Mixer risk: is expressed as the probability, per million persons, of developing cancer for a lifetime of exposure (35years)to the chemical.
4. Reproductive risk: expressed as a dummy variable representing 1 if reproductive risk is present, 0 otherwise.
5. Wildlife risk: expressed as a dummy variable representing 1 if wildlife risk is present, 0 otherwise.
6. Diet risk missing: equals 1 if missing, 0 otherwise.
7. Applicator risk missing: equals 1 if missing, 0 otherwise.
8. Mixer risk missing: equals 1 if missing, 0 otherwise.
9. Producer benefits: in millions of 1986 dollars.
10. Benefits missing, no yield loss: equals 1 if there are no expected yield lossesand benefits are missing, 0 otherwise.
11. Benefits missing, yield losses: equals 1 if there are no expected yield lossesand benefits are missing, 0 otherwise.
12. Academic comments: equals 1 if comment was made, 0 otherwise.
13. Farmers comments: equals 1 if comment was made, 0 otherwise.
14. Environmentalists comments: equals 1 if comment was made, 0 otherwise.
15. Burford: equals 1 if Burford was EPA administrator, 0 otherwise.
16. GOP: equals 1 if Republican was the ruling party, 2 otherwise.
17. Prez1-4: equals 1 if president was Carter, 2 if president was Reagan, 3 if president was Bush, 4 if president was Clinton.

References

Belsley, D., E. Kuh, R. Welch. *Regression Diagnostics*. New York: Wiley, 1980.

Breiman, Leo, J. Freidman, R. Olsen and C. Stone. *Classification and Regression Trees*. Pacific Grove: Wadsworth, 1984.

Burrows, W., M. Benjamin , S. Beuchamp. Cart Decision Tree Statistical Analysis and Prediction of Summer Season Maximum Surface Ozone Over Vancouver, Montreal and Atlantic Regions of Canada. *Journal of Applied Meteorology*, 34, 1848-1862.

Cropper,M., W.N. Evans, S.Berardi, M. Ducla-Soares and P.Portney (1992). "The Determinants of Pesticide Regulation: A Statistical Analysis of EPA Decision Making". *J. of Political. Economy.* , Vol 100(1) 175-97.

Federal Register, 1975-95. Various issues

Gupta,S., G. VanHoutven and Maureen Cropper, 1993. "Clean-up Decisions under Superfund: Do Benefits and Costs Matter?" *Resources.* Spring 1993

Gordon,R.C. and J.W. Freebairn. "Estimation of Policy Preference Functions: an Application to U.S. Beef Import Quotas." *Review of Economics and Statistics*, 1975, 437- 449.

Harrison, D., and D. Rubinfield. Hedonic Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 5: 81-102, 1978.

Kolluru, Rao, Steven Bartell, Robin Pitbaldo and Scott Stricoff, eds. *Risk Assessment and Management Handbook for Environmental, Health and Safety Professionals*. McGraw Hill, 1996

Magat,W., A.Krupnick and W.Harrington. Rules in the Making: a Statistical Analysis of Regulatory Agency Behavior. Resources for the Future, 1986

McFadden,D. "The Revealed Preferences of a Government Bureaucracy: Theory." *Bell Journal of Economics,* Vol. 6, # 2, Autumn 1975.

"The Revealed Preferences of a Government Bureaucracy: Empirical Evidence." Bell Journal of Economics, Vol. 7, #1, Spring 1976.

Oehmke, J.F. And X. Yao. " A Policy Preference Function for Government Intervention in the Wheat Market." *Amer J. of Ag. Econ*. Aug 1990, 631-640

Steinberg, D., and P. Colla. CART: tree Structured Non-parametric Analysis. San Diego, CA: Sanford Systems, 1995

Thomas, G.L. "Revealed Bureaucratic preferences: Priorities of the Consumer Protection Safety Commission." *Rand J. of Econ.* Vol 19(1) 102-113.

Weingast, B. And M. Moran, 1983. "Bureaucratic Discretion or Congressional Control? Regulatory policy Making by the Federal Trade Commission." *Journal of Political Economy,* Vol. 91, #5.