25th ANNUAL NATIONAL CONFERENCE ON MANAGING ENVIRONMENTAL QUALITY SYSTEMS

APRIL 24-27, 2006

Marriott Renaissance, Austin, Texas

Technical Papers

Statistical Issues for Health and the Environment

- R. Sastry, Statistical Methods to Analyze Occupational Safety Data of DOE Facilities - 10:30 AM
- H. Kahn, Statistical Issues in the Analysis of the Carcinogenic Risk of Ethylene Oxide – 11:00 AM
- M.Nash, Partial Least Squares Regression for Small Sample - 11:30 AM

STATISTICAL METHODS TO ANALYZE THE OCCUPATIONAL SAFETY AND HEALTH DATA OF THE DEPARTMENT OF ENERGY FACILITIES

M. Rama Sastry, PhD
Office of Quality Assurance Programs
Office of Environment, Safety and Health
U.S. Department of Energy
Washington, D.C.20585

(Paper prepared for presentation at the EPA Quality Management Conference, Austin, Texas, April 24-28, 2006. The views expressed in this paper are personal and do not represent the Department of Energy's position.)

1. INTRODUCTION

The Department of Energy (DOE) operates many nuclear and non-nuclear facilities, and National Research laboratories located throughout the United States. Approximately 130,000 employees of various contractors work at these facilities. The DOE is responsible to protect health and safety of the employees and conduct work in an environmentally safe manner. The DOE complies with the Environmental Protection Agency (EPA) regulations for environmental management and the Occupational Safety and Health Administration (OSHA) regulations for worker protection. The DOE contractors record and report incidents and accidents related to occupational injuries and illnesses in accordance with 29 CFR regulations. Such data are collected and maintained by a centralized data base called "Computerized Accident/Incident Reporting System (CAIRS)", which is the main source of information for the statistical methods shown by this paper.

The following statistical methods were considered to analyze the occupational safety and health data:

1. Exploratory Data Analysis (Box Plots)
2. Data Visualization, Data Images or Color Histograms
3. Clustering Analysis (Hierarchical clustering with Complete linkage)
4. Trend Analysis (Exponential Smoothing, Kalman Filtering)
5. Advanced Methods (Discriminanat Analysis)

The above methods are some of the possible techniques useful for the analysis and do not represent a comprehensive or unique list. The Office of Environment, Safety and Health uses a wide variety of statistical method to conduct analysis of environmental data. For example, see the publications by Richard Gilbert and others at the Pacific Northwest National Laboratory (PNNL). The CAIRS data are published on a quarterly basis for all DOE sites and facilities by contractors, by Field offices/Operations Offices, and by the DOE Program Offices. The CAIRS data are validated periodically for data quality and accuracy by reviewing the OSHA 200/300 logs maintained by the contractors. Historical data are available beginning 1980's, however more recent data were considered by the analysis to avoid the many changes occurred in organizations and the DOE mission. For example, after the end of Cold War, the mission of the agency shifted from production to environmental remediation and waste management, and recently additional emphasis being placed in conducting basic research in science and technology at the National Laboratories.

The Department of Labor, Bureau of Labor Statistics (BLS) compiles occupational safety data for private industry within the United States and that data was used by DOE to compare safety performance of DOE contractors. Since the inception of OSHA in 1971, the safety performance of private industries has improved and the same pattern occurred in DOE. However, in general the recordable injury rates at DOE sites are usually lower than private industry. The DOE has also adopted the OSHA's Voluntary protection Program (VPP) to promote safety and health excellence through cooperative efforts between labor and management. During 1994-2005, approximately 25 sites were recognized by the DOE VPP as STAR sites, and several other sites are in the process of obtaining such status. The impact of the VPP and the Value Added by the program are described in the DOE reports cited in the References (Section 4.) of this paper.

2. STATISTICAL METHODS

Two measures of occupational safety performance considered for the analysis area are as follows:
(a) Total Recordable Case Rate (OSHA Recordable injury/illness Case rates), and
(b) Days Away form work, Restricted, and Transfer Case Rate (DART Case rate) formerly known as the Lost Work Day Case Rate., as defined by the Bureau of Labor Statistics
For the sake of illustration, annual data for the years 1996-2005 related to TRC Rate and DART Rates at major DOE Program Offices were retrieved from CAIRS. The Program Offices selected for this analysis are:

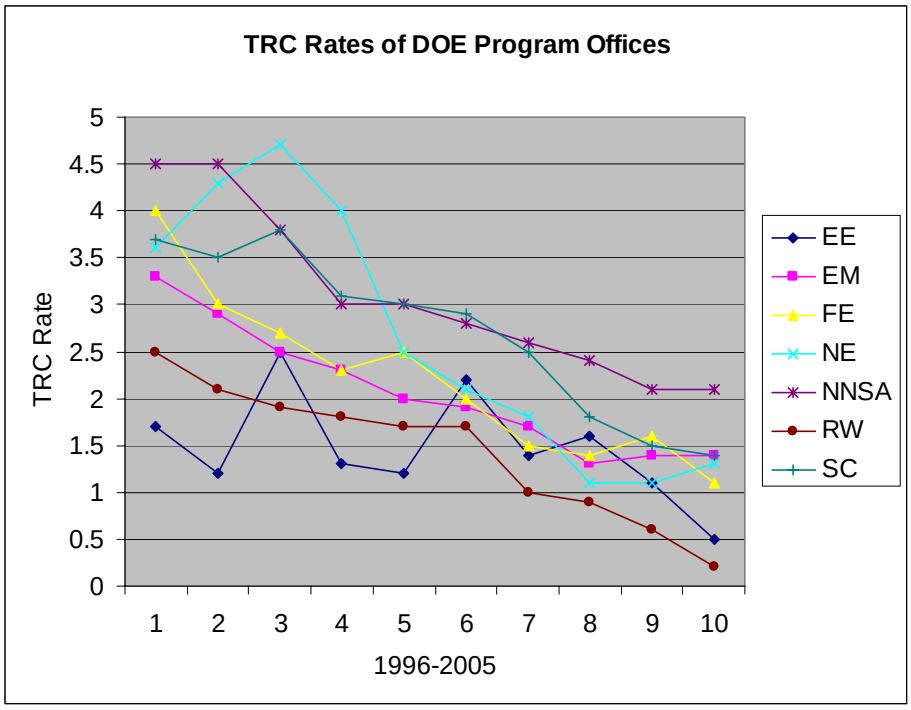
- Energy Efficiency and Renewable Energy (EE)
- Environmental Management (EM)
- Fossil Energy (FE)
- Nuclear Energy (NE)

- National Nuclear Security Administration (NNSA)
- Fossil Energy (FE)
- Science (SC)

Figure 1. below shows the TRC rates during the past ten years (1996-2005) for the seven DOE Program Offices. Three Program Offices, NNSA, EM, and SC employ almost 75% of the contractor work force in DOE, and the occupational hazards of the operations at the sites such as Pantex, Los Alamos, Hanford, Rocky Flats, and the Science Laboratories may be higher than the FE or EE facilities. However, Figure 1. indicates that their injury illness rates are not necessarily higher. For example, the TRC rates at EM facilities are lower than the FE rates in most of the years during 1996-2005.

TRC Rates at the DOE Program Offices (1996-2005)

	EE	EM	FE	NE	NNSA	RW	SC
1996	1.7	3.3	4	3.6	4.5	2.5	3.7
1997	1.2	2.9	3	4.3	4.5	2.1	3.5
1998	2.5	2.5	2.7	4.7	3.8	1.9	3.8
1999	1.3	2.3	2.3	4	3	1.8	3.1
2000	1.2	2	2.5	2.5	3	1.7	3
2001	2.2	1.9	2	2.1	2.8	1.7	2.9
2002	1.4	1.7	1.5	1.8	2.6	1	2.5
2003	1.6	1.3	1.4	1.1	2.4	0.9	1.8
2004	1.1	1.4	1.6	1.1	2.1	0.6	1.5
2005	0.5	1.4	1.1	1.3	2.1	0.2	1.4



Further analysis of the data was conducted by using Box-Whisker plots (See John Tukey). Figure 2. Indicates that the variability of the TRC data at EM facilities to be lower than the variability of the data at FE facilities during the same period. The same chart suggests that the variability at EE to be the smallest and NE to be the highest. Also from Figure 2. we observe that the median of the TRC rates at SC to be the largest and the median of the EE sites to be the smallest among the seven Program offices considered by this analysis.

In addition to the mandatory safety programs such as Integrated Safety Management (ISM), many EM sites have adopted the Voluntary Protection Program (VPP) to improve safety performance. For example, Fernald, Hanford, West Valley, WIPP, etc, are VPP STAR sites. The safety performance at any DOE site or facility should not be judged on the basis of one indicator such as TRC or DART. Further analysis is necessary to understand the differences in the operational risks and the safety performance.

The next statistical method used in this paper is related to VPP data, in particular to conduct cluster analysis using TRC rates of VPPsites and Non-VPP sites in DOE. For more details of this method, see Sastry and Schwender (2005), and for a theoretical description of the methods see Trevor Hastie et al (2001), and W.N. Venables and B.D. Ripley (2002). The computer software used was S-Plus and R originally developed by Bell laboratories. The methodology in particular that was used is called “hierarchical clustering with complete linkage “. The primary objective of this method is to generate clusters of data and identify similar patterns. Figure 3. (Dendogram) indicates that most of the VPP sites (labeled in green) clustered into one group and most of the Non-VPP sites (labeled in red) into another group. Only one or two sites or facilities were clustered into a wrong group or miss specified.

3. ADVANCED METHODS

In addition to the Clustering methods, other sorting procedures such as the Principle Component Analysis and Single value decomposition, or the classification / Classification and Regression Trees (CART) may be applied to perform the necessary analysis. Also Discriminant Analysis (Linear or Quadratic) is useful to classify the safety performance of VPP sites and Non-VPP sites. In addition, the distance between the VPP sites and Non-VPP sites can be estimated on the basis of Mahalonobis D-Square statistic.

Quadratic Discriminant Analysis:

- Suppose the distribution for class C is multivariate normal with mean μc and covariance, then the Bayes Rule minimizes a quadratic function:

$$Q_C = D^2 + \log|\sum_C| - 2\log\pi_C$$

where π_C is the prior probability of class C

- D = Mahalanobis Distance
- $D^2 = (x_i - \bar{x})'\bar{s}(x_i - \bar{x})$, where $\bar{x}$ is the sample mean, and *s* is the variance

4. CONCLUSIONS

Classical statistical methods supplemented by Data Mining, Visualization and Graphics can enhance the analysis capability. The results of the analysis should be useful for management decision making and for continuous improvement.

5. REFERENCES

1. National Research Council, " Beyond Productivity , Information Technology, Innovation and Creativity", National Academy Press, Washington, DC
2. William S. Cleveland, Visualizing Data, A T&T Bell Laboratories, Murray Hill, N.J. , 1993
3. Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The elements of Statistical Learning: Data Mining, Inference, and Prediction , Springer Publications, 2001

4. W.N. Venables and B.D. Ripley Modern Applied Statistics with S, 4th Edition, Springer Publications, 2002
5. Rama Sastry and Holger Schwender, Statistical Analysis of Occupational data of VPP and Non-VPP sites, DOE-EH/0696, April, 2005
6. Rama Sastry , Rex Bowser and David Smith, The Value Added of the DOE VPP , 2004 Update, DOE/EH-0690, December 2004
7. Rama Sastry and Carlos Coffman, Safety Performance of Security Forces at the DOE facilities, DOE/EH-0705, December 2005
8. John Tukey, Exploratory Data Analysis, Addison-Wesley Publishing Co, 1977
9. Leo Breiman, et al, Classification and Regression Trees, Wadsworth Publishing Co, 1984
10. R.Gilbert, J.E. Wilson and B.A. Pulsipher and others , "Visual Sample Plan " (various guides and research reports), DOE's Pacific Northwest National Laboratory, Richland, WA

Statistical Issues in the Analysis of the Carcinogenic Risk of Ethylene Oxide

Henry D. Kahn and Jennifer Jinot
National Center for Environmental Assessment
Office of Research and Development

Ethylene oxide (EtO) is a gas at room temperature that is manufactured from ethylene and used primarily as an intermediate in the manufacture of ethylene glycol. It is also used as a sterilizing agent for medical equipment and as a fumigating agent for spices. Human exposure to EtO occurs in manufacturing plants and in hospitals and other facilities where medical equipment is sterilized. EtO can also be inhaled by residents living near production or sterilizing/fumigating facilities. In humans employed in EtO-manufacturing facilities and in sterilizing facilities, the greatest evidence of a cancer risk from exposure is for cancer of the lymphohematopoietic system. Increases in the risk of lymphohematopoietic cancer have been seen in several studies, manifested as an increase either in leukemia or in cancer of the lymphoid tissue. In one large epidemiologic study of sterilizer workers that had a well-defined exposure assessment for individuals, positive exposure-response trends for lymphohematopoietic cancer mortality in males and for breast cancer mortality in females were reported (Steenland et al., 2004). The positive exposure-response trend for female breast cancer was confirmed in an incidence study based on the same worker cohort (Steenland et al., 2003). This presentation will focus on the statistical analysis of human epidemiological data that may be used to estimate the cancer inhalation risk due to exposure to ethylene oxide. Statistical modeling of the data and the methodology for derivation of inhalation unit risk estimates for cancer mortality and incidence will be discussed.

Partial Least Squares (PLS) Regression for Small Sample with Collinear Predictors in Landscape Ecology.

Maliha S. Nash and Ricardo Lopez*
US EPA, PO Box 93478, Las Vegas NV 89193-3478.
E-mail: nash.maliha@epa.gov

(Notice: Although this work was reviewed by EPA and approved for publication, it may not necessarily reflect official Agency policy.)

1. Introduction

Investigation of associations among constituents of surface water and landscapes involves statistical analyses of fundamentally different data sets. Data on surface water conditions are generally obtained through field sampling programs and field/analysis programs are expensive and labor intensive; consequently, the total number of sample sites is usually small. The data set may contain missing values due to the realities of sampling or cost. Landscape data, however, is derived from remote sensing platforms, thereby permitting wall-to-wall coverage. The landscape data sets may contain a very large number of variables, although many of these are not wholly independent (i.e., they may be collinear). Single- and multiple-regression analysis has frequently been used to relate water nutrient concentrations to selected landscape variables are sensitive to missing values and dependence of predictors (landscape variables). Reliable statistically significant results generally cannot be obtained unless the total number of samples greatly exceeds the number of variables. Partial least squares (PLS) analysis offers a number of advantages over the more traditionally used regression analyses. It has been found to be useful both for providing accurate predictions and for interpreting relationships between data sets containing a high degree of collinearity (see references in Nash et al., 2005). Additionally, the prediction error in PLS is smaller than in other multivariate methods (for references and more see Nash et al., 2005).

2. Data Description

The study area is the Upper White River study area is in the Ozarks of Missouri and Arkansas, where 244 water-quality sampling locations were sampled and used as 'pour-points. For each of the 244 sites, the watershed support area was delineated and a suite of landscape variables was calculated. It is important to understand that some of the 244 subwatersheds are nested completely within other larger subwatersheds. Total of 46 landscape metrics were generated per each watershed. Measured total phosphorous (TP), total ammonia (TAM), and *E. coli* were only existed in 18, 6, and 15 sites, respectively. For the purpose of this paper, we used non nested (0 level) sub-watersheds representing first order streams, hence eliminating nested watershed. Sample size, therefore, was 5, 6, and 5 for TP, TAM and *E. coli,* respectively that used for building the PLS models. Landscape metrics were for year 2000 and surface water constituents were averaged over a period of 1997-2002. Prediction of the surface water constituents for the remaining from the 244 sites were made using the PLS models above.

3. STATISTICAL METHODOLOGY

PLS is a multivariate analysis technique permits analysis and prediction for data sets with missing values, with collinearity and with a relatively small number of observations (see references in Nash et al., 2005). In the PLS analyses, both response and predictor data sets (e.g. water and landscape variables) are first centered and scaled. A linear combination is composed on the independent variables ($T = L_o W$; T is the score and W is weight) forming a number of orthogonal latent variables [T] that are less in number (dimensions) than that of the original landscape variables. The linear combination in [T] is formed so that the covariance between [T] and the linear composition of the dependent variables are maximized (T& U; $U = B_o V$; U is the score and V is weight). Prediction of both water and landscape data will be via regression on the common latent variables (T). Modeling and prediction in PLS, therefore, is not solely based on the conditional distribution of the predictors (water variables) in the presence of independent variables (landscape variables), instead it accounts for both landscape and water together through [T] (see references in Nash et al., 2005).

PLS produces n-1 factors, with each factor containing a pair of scores (T_i, U_i). Linear combinations on each data set are called factors. PLS extracts the second factor using the residuals from the first and finds the linear combinations of both data sets such that their covariance is maximized. This process is repeated by taking residuals from the previous factor, producing n-1 factors, where n is the number of observations. Not all of these factors are significant using the Cross Validation (CV) method; only the significant factors are used in the final model. When applying CV, one data point is held out and the fitted models are tested using the rest of data set and the predicted values are compared with that of observed using PRESS (Predictive Residual Sum of Square) to assess the predictive ability of the model. SAS gives the root means PRESS and its significant level (the lower the value, the better the model is).

After defining the significant PLS factors; scores, weights and VIP (Variable Influence on Projection) are used to examine the strength of the relationship, irregularities and the contribution of the independent variable (landscape) in the model. If VIP for an independent variable is small in value, it implies that variable has a relatively small contribution to the model and may be deleted from the model. It was indicated VIP values of less than 0.8 are considered to be small. The quality of the model was determined by examining the residuals for both the response and the landscape variables. An examination of any possible outliers using residuals was carried out to finalize the fitted PLS model. SAS was used for statistical analyses.

4. RESULTS

TP PLS model resulted in one significant factor explaining 83% of the variability in the TP (see Table). Barren soil had the most significant effect based on the whole watershed and in the riparian zone immediate to the stream. While the stream density relates inversely with TP, percent barren enhances TP in surface water especially in areas adjacent to the stream. The forest- and urban- related variables contribute equally with opposite effect on the TP. Urban enhanced TP whereas forest, especially within the proximity (Rfor0) of the sampling site, depressed the level of TP.

TAM PLS model resulted in one significant factor explaining 93% of the variability in the TAM (see Table). Riparian and natural within all distances have negative effect on TAM, whereas urban has a positive effect. Urban within the riparian zone enhanced the level of TAM but beyond the proximity of stream (i.e. distance of 30 m and more).

E. Coli PLS model resulted in two significant factors explaining 99.7% of the variability in the *E.coli* population (see Table). Elevation has the highest effect on the *E. coli,* the flatter the soil surface the higher abundance of the *E. coli* in surface water. Urban- related variables enhanced the level of the *E. coli* especially in riparian within area of the sampling site.

The prediction of the constituents in the 244 watersheds (from a small filed-based data sample) was used to visualize the joint behavior of the predicted TP, TAM, and *E. coli* in surface water of the Upper White River (Figure 1. Using PLS we determined four distinct surface water conditions among subwatersheds in the Ozarks:
(1) subwatersheds with high concentrations of TAM, high concentrations of TP, and high cell counts of *E. coli*; (2) subwatersheds with high concentrations of TAM, low concentrations of TP, and high cell counts of *E. coli*; (3) subwatersheds with low concentrations of TAM, low concentrations of TP, and high cell counts of *E. coli*; and (4) subwatersheds with moderate concentrations of TAM, TP, and cell counts of *E. coli.*

5. Discussion and Conclusion

The results indicate PLS may prove to be a valuable statistical analysis tool for ecological studies. The PLS methodology is less sensitive to the limitations than other statistical methods. The joint behavior of TP and TAM as related with *E. coli* (Figure 1) was not possible using the measurements from the study area sites (5, 6, and 5 sites for TP, TAM, and *E. coli*, respectively) but it was overcome by prediction form the PLS model for the 244 sites. Hence, further analyses and comparisons within and between the above 4 groups may reveal the spatial characteristics setting for watersheds and their effect on surface water quality.

The model results may help landscape ecologists produce indicators of surface water condition, such that unique combinations of these indicators can be used to infer the potential cause(s) and origin(s) of non-point pollution, which may lead to eutrophication in aquatic ecosystems, the loss of aquatic ecosystem function, and the injury of humans that consume from (or recreate in) the aquatic resources of the Ozarks. Sensitivity analyses for the above model and the PLS results discussed in this presentation are actively being used to prioritized subwatersheds in the Ozarks for watershed management activities.

Reference

Note: The authors would like to thank Ms. Deborah Chaloud, EPA/LEB for valuable input. The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), funded and performed the research described here. Although this work was reviewed by EPA and approved for publication, it may not necessarily reflect official Agency policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Table 1. Coefficients of the non-centered value of landscape metrics to predict the ln(TP), TAM, and ln(*E. coli*). Number of significant PLS factors and percent variation explained by PLS for the responses are in the last two rows.

	TP		TAM		**E. coli**	
Landscape Metrics	Coefficient	VIP	Coefficient	VIP	Coefficient	VIP
Intercept	-2.2892		0.07867		-2.20795	
Fdensity	-0.00194	0.12864				
Fedge210	-0.00195	1.03754				
F_mdcp	0.00079	0.97022				
F_plgp					-0.00794	1.015
Pfor	-0.00118	1.01968			0.00736	0.985
Rfor0	-0.00123	1.12558			-0.00023	0.976
Rfor30	-0.00118	1.10126			-0.00065	0.979
Rfor120	-0.00117	1.10279			0.00172	0.977
Rnat0	**-0.00123**	**1.12558**	**-0.00027**	**0.91965**	**-0.000232**	**0.976**
Rnat30	**-0.00118**	**1.10126**	**-0.00025**	**0.92898**	**-0.00065**	**0.979**
Rnat120	**-0.00117**	**1.10279**	**-0.00023**	**0.94123**	**0.00172**	**0.977**
Purb	**0.00102**	**1.06728**	**0.00031**	**1.06739**	**0.00557**	**1.008**
Rurb0	0.00161	1.13694			0.00929	1.014
Rurb30	0.00153	1.1.133	0.00042	1.07047	0.00868	1.014
Rurb120	**0.00141**	**1.16164**	**0.00036**	**1.06935**	**0.00743**	**1.006**
Rhum0	0.00123	1.12558			0.00023	0.976
Rhum30	0.00118	1.10126			0.00065	0.979
Rhum120	0.00117	1.10279			-0.00172	0.977
Pctia_rd	0.00197	1.04816			0.01373	1.021
Rddens	0.01074	1.03470			0.09992	1.051
Pmbar	0.3312	1.21778				
Rmbar0	0.08693	0.50283				
Rmbar30	0.15629	0.50283			0.11702	0.757
Rmbar120	0.38269	1.24011				
Strmdens	-0.04314	1.13699			-0.91794	1.018
Elevmin					0.00024	1.273
Number of Factors	1		1		2	
% Variation	83		93		99.7	

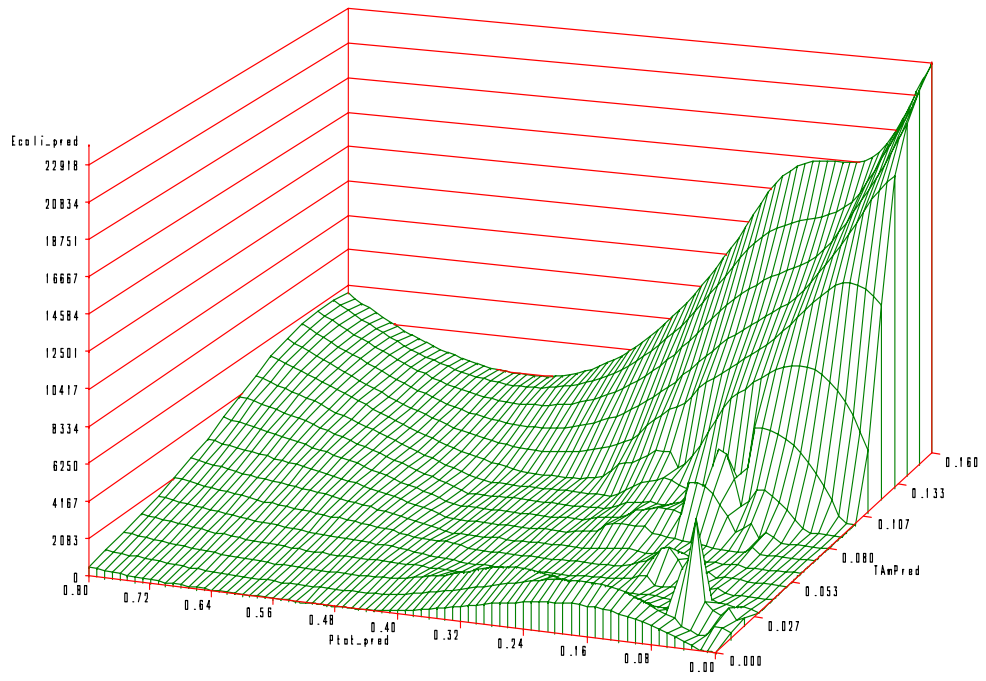


Figure 1 Three-dimensional plot of predicted TAM (x-axis), TP (y-axis), and *E. coli* cell counts (z-axis) among 244 subwatersheds in the Upper White River region of the Ozarks.