DOE-EH/0696

Statistical Analysis of Occupational Safety Data of Voluntary Protection Program (VPP) and Non-VPP Sites

U.S. Department of Energy
Office of Environment, Safety and Health
Office of Corporate Performance Assessment

Office of Quality Assurance Programs

April 2005

DOE-EH/0696

Statistical Analysis of Occupational Safety Data of Voluntary Protection Program (VPP) and Non-VPP Sites

U.S. Department of Energy
Office of Environment, Safety and Health
Office of Corporate Performance Assessment

Office of Quality Assurance Programs

April 2005

Statistical Analysis of Occupational Safety Data of Voluntary Protection Program (VPP) and Non-VPP Sites

by

Rama Sastry
Office of Quality Assurance Programs
U.S. Department of Energy
Washington, DC 20585

Holger Schwender
Collaborative Research Center 475
Department of Statistics
University of Dortmund
Dortmund, Germany

The authors would like to thank Prof. Dr. Hermann Bolt, Director, Institute for Occupational Safety and Health (IfADo), University of Dortmund, and Prof. Dr. Katja Ickstadt and Professor. Dr. Ursula Gather of the Department of Statistics at the University of Dortmund, Germany, for their support and encouragement. Holger Schwender would like to thank the German Research Foundation (DFG; SFB 475, “Reduction of complexity in multivariate data structures”) for financial support. Dr. Sastry would also like to thank Chip Lagdon, Director, Office of QA Programs (EH-31), Frank Russo, Deputy Assistant Secretary for Corporate Performance Assessment (EH-3), and David Smith, Team Leader, DOE VPP, for their encouragement.

Table of Contents

Figures

Tables

1. Introduction

The Voluntary Protection Program (VPP) was originally developed by Occupational Safety and Health Administration (OSHA) in 1982 to foster greater ownership of safety and health in the workplace. The Department of Energy (DOE) adopted VPP in 1992; currently 23 sites across the DOE complex participate in the program. As its name implies, it is a voluntary program; i.e. not required by laws or regulations. The DOE VPP encourages DOE contractors to seek excellence by surpassing compliance in implementing health and safety programs, empowering workers, and involving managers in a cooperative effort to protect workers from accidents and occupational injuries or illnesses. The purpose of this study is to compare the safety performance of VPP sites against those who are not VPP members using statistical methods. These methods will be explained in Section 2 of this report.

The Department of Energy is responsible for research and production of special nuclear materials and for manufacturing and testing of nuclear weapons. The Department continues to work in national defense activities including basic and applied science research at the national laboratories and in decontaminating and decommissioning of former production sites. The Department also operates non-nuclear facilities such as the Strategic Petroleum Reserves and Solar Energy Research Institute. During World War II and the Cold War era, more than 200,000 workers were employed by DOE contractors at various locations in the United States. Large sites such as Savannah River, Hanford, and the national laboratories such as Los Alamos, Livermore, and others, were involved in these activities. After the Cold War ended, the mission of the Department changed from production of nuclear weapons to environmental remediation and waste management. During this period, employment levels decreased to 130,000 workers. However, the risks to workers from accidents and injuries or illnesses remain high due to legacy chemicals, radiological contamination, and other industrial safety hazards that are encountered during cleanup work.

Department of Energy facilities and contractors are required to maintain Occupational Safety and Health Administration (OSHA) 300 logs and to record and report all injury and illness data to DOE in accordance with the regulations specified in 29 CFR Part 1904, *Recording and reporting occupational injuries and illnesses*. The Department collects injury and illness information in the Computerized Accident/Incident Information System (CAIRS) database (https://www.eh.doe.gov/cairs), which contains records dating from 1975. In addition to occupational safety data, CAIRS contains information on property damage and vehicle accidents. The CAIRS database is a useful tool for analyzing safety performance, identifying causes of accidents, and establishing priorities to improve safety and health performance.

2. Statistical Methods

One of the goals of the statistical analysis is to find groups of DOE laboratories or sites that show a similar pattern. The DOE sites or laboratories within a group should thus have similar Total Recordable Case (TRC) rates or Days Away, Restricted, or Job Transfer (DART) rates (Note: prior to January 2004, DART was known as Lost Workday Case or LWC). These values should differ between the groups. Such analysis may identify two groups, one containing the VPP sites and the other consisting of the non-VPP sites. Statistical methods that identify such groups are the so called *clustering procedures*. See Hastie, Tibshrani, and Friedman (2001) (pages 453-480) for a description of cluster analysis including various algorithms developed by statisticians. One of the most widely used methods of clustering analysis is *hierarchical clustering* with *complete linkage* which is explained below with the help of some examples.

The validity of clustering analysis techniques depends on the degree of similarity or dissimilarity between individual objects being clustered. Therefore, we need to define the *measure of the dissimilarity* of two objects; e.g., DOE sites or laboratories. Again, we will use one of the most widely used distance measures: the *Euclidean distance*. Given a set of n values measured on both of the two observations X and Y, the squared Euclidean distance is computed by

$$d_{XY}^2 = \sum_{i=1}^{n} (x_i - y_i)^2 .$$

Table 1 illustrates an example by comparing the average daily Internet use of 5 people over 4 consecutive months. From looking at this table, we could say that Persons 1 and 3 and Persons 2, 4, and 5 show similar patterns in their average daily Internet use.

By performing a cluster analysis, we can confirm this impression. First, the Euclidean distances for all pairs of observations are computed (see Table 2). For example, the Euclidean distance between Person 1 and Person 2 is calculated by the square root of

$$d^2 = (1.2 - 4.3)^2 + (0.9 - 5.2)^2 + (2.1 - 5.2)^2 + (1.1 - 4.4)^2 = 48.6,$$

i.e., $d = 6.97$.

Table 1: Average Daily Internet Usage

	Month 1	Month 2	Month 3	Month 4
Person 1	1.2	0.9	2.1	1.1
Person 2	4.3	5.2	5.2	4.4
Person 3	0.6	2.1	1.4	1.3
Person 4	4.1	3.9	4.4	4.2
Person 5	5.1	4.6	4.7	4.3

Table 2: Euclidean Distances between Each Pair of Observations

	Person 1	Person 2	Person 3	Person 4	Person 5
Person 1	0.00	6.97	1.53	5.68	6.77
Person 2	6.97	0.00	6.88	1.55	1.12
Person 3	1.53	6.88	0.00	5.74	6.81
Person 4	5.68	1.55	5.74	0.00	1.26
Person 5	6.77	1.12	6.81	1.26	0.00

The *distance matrix* presented by Table 2 can now be used to perform hierarchical clustering. We start with searching for the smallest distance between two observations, which in this example, is the Euclidean distance of 1.12 between Person 2 and Person 5 (shown in Table 2). These two observations are merged into one observation. Let's call this observation Person 2/5. This merging presents a problem in that we now have two distances between Person 2/5 and other observations. For example, Person 2/5 has a distance of 6.97 to Person 1 (coming from Person 2) and a distance of 6.77 (coming from Person 5) to Person 1. These two distances must be combined into one distance.

This can be done in several ways: by taking the minimum of the two distances (*single linkage*), the maximum (*complete linkage*), the average (*average linkage*), or an alternative. As mentioned above, we will perform the hierarchical clustering using a complete linkage ,so we take the maximum of the two distances. From our example above, this means that the distance between Person 2/5 and Person 1 is given by

$$d = \max \{6.97, 6.77\} = 6.97.$$

Table 3 shows the other combined distances.

Table 3: Distance Matrix after the First Step of Hierarchical Clustering with Complete Linkage

	Person 1	Person 2/5	Person 3	Person 4
Person 1	0.00	6.97	1.53	5.68
Person 2/5	6.97	0.00	6.88	1.55
Person 3	1.53	6.88	0.00	5.74
Person 4	5.68	1.55	5.74	0.00

This procedure is repeated until a single observation remains. In the next step, Person 1 and Person 3 are merged into Person 1/3, then Person 2/5 and Person 4 are merged into Person 2/5/4, and finally Person 1/3 and Person 2/5/4 are merged into Person All. While doing this, we keep in mind the Euclidean distances between the Persons as we merge them:

- between Person 2 and Person 5 is 1.12,
- between Person 1 and 3 is 1.53,

- between Person 2/5 and Person 4 is 1.55, and
- between Person 1/3 and Person 2/5/4 is 6.97.

These distances are then displayed in a diagram called *dendrogram* (see Figure 1).

We can now identify groups of observations with a similar pattern by drawing a line on the dendrogram at a specific height and examining the dendrogram below the line. All of the observations that are still connected form a group. In Figure 2, for example, we have drawn a line on the dendrogram at a height of 3. The two resulting groups are composed of Persons 1 and 3 and Persons 2, 4, and 5.

Figure 1: Dendrogram

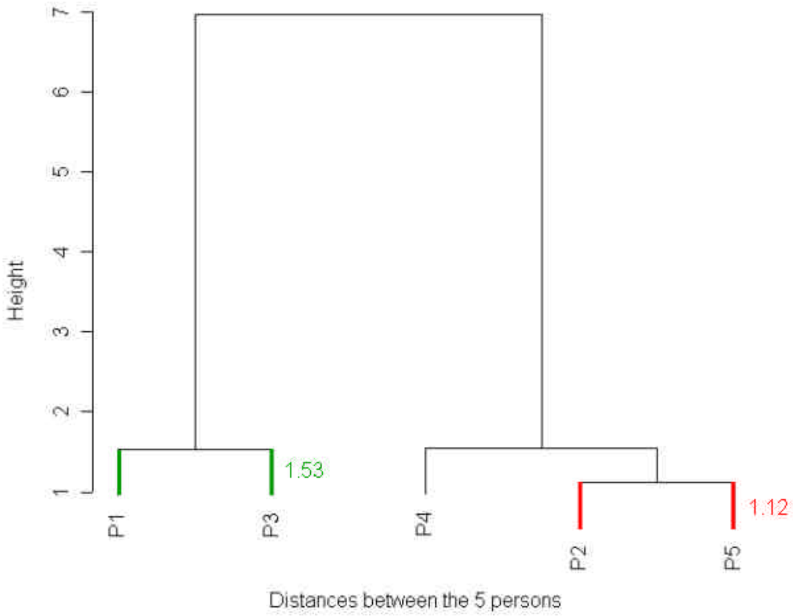
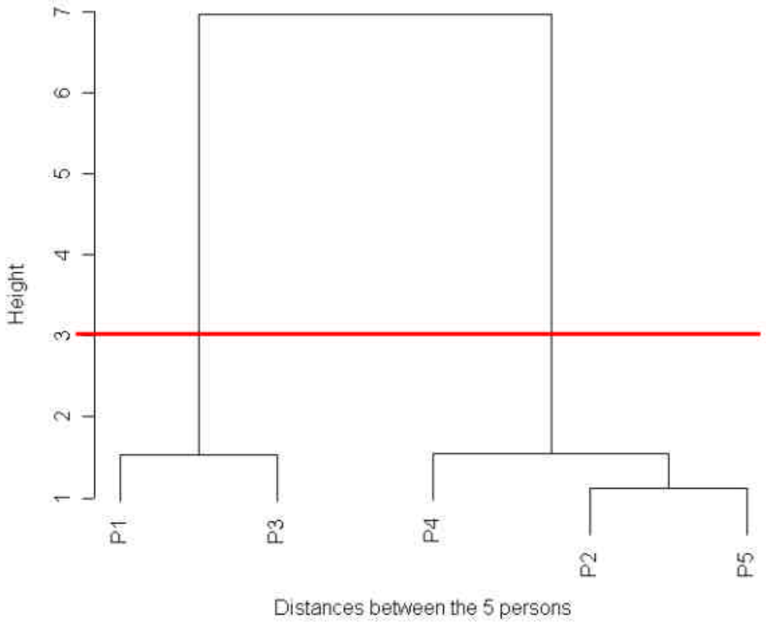


Figure 2: Dendrogram with a Line at a Height of 3



Another statistical technique related to cluster analysis but not used in this report is classification tree analysis. Traditional methods such as discriminant analysis are based on more stringent theoretical and distributional assumptions and have less flexibility than classification methods. See Leo Breiman et al. (1984) for a description of this method including the Classification and Regression Tree (CART) algorithm. The recursive and hierarchical nature of the decision trees is the main feature of this method that enables decision makers to reach the prediction based on the available information. The goal of classification analysis is to predict or explain responses of a categorical dependable variable based on predictor variables that may be continuous or categorical. Classification methods have been used in many fields of application, for example, to identify the risk of breast cancer in women or prison inmates who are likely to engage in serious misconduct while incarcerated. At DOE, we can use this type of analysis to identify sites with high injury and illness rates so that management can act to reduce them.

The goal of the classification rule is to classify new observations into one of the groups. To build a classification rule, we used a so-called training set; i.e., a set of observations from which the observations were known to belong. In our DOE VPP model, the objective was to find a rule that correctly classified whether or not the sites we studied were members of the VPP; the best case being that all VPP sites were classified as one group, and all non-VPP sites were classified as another group).

If we could divide the DOE sites into two groups – those who belong to the VPP and those who do not – by using a clustering method, this would have a higher explanatory power than if we found such a clear distinction using a classification rule.

Two major issues involved in classification trees are related to: 1. determining the optimum size of the tree or the number of splits of all leaves, and 2. calculating the misclassification error

(MCR). The limitations of CART such as the instability of the trees, lack of smoothness, and other refinements to CART such as Multivariate Adaptive Regression Spline (MARS), Support Vector Machines (SVM), and k-Nearest Neighbor classifiers are presented by Hastie, Tibshirani, and Friedman (2001). A more recent application of these methods to molecular epidemiological data can be found in the paper by Schwender et al (2004).

The statistical software used for performing cluster analysis of this report is called "R," which is derived from the original programs "S" and "S-Plus" developed by Bell Laboratories in New Jersey. See Richard A. Becker, John Chambers, and Allan Wilks (1988), and W.N. Venables et al (2002) for more details of this software. R is free software that can be run on Windows and some UNIX and Linux operating systems.

3. Input Data

We selected a subset of the CAIRS database to analyze the injury and illness rates of VPP sites and non-VPP sites. The selected data correspond to theTRC rates and DART case rates of the following sites from the first quarter of 1995 through the second quarter of 2004. There are no missing values in the selected data subset of CAIRS. This gives 38 observations (time-series data) on TRC and DART rates for 10 VPP sites and 14 non-VPP sites (see Table 4).

Table 4: VPP and Non-VPP Sites or Laboratories

VPP Sites/Facilities or Laboratories	Non-VPP Sites
1. Fernald (FEMP)	1. Ames Laboratory
2. Hanford (includes all VPP facilities at this site; i.e., data are combined)	2. Argonne National Laboratory – East (ANL-East) 3. ANL- West
3. Idaho National Laboratory (INL)	4. Brookhaven National Laboratory (BNL)
4. Kansas City Plant (KCP)	5. Fermi National Accelerator Laboratory
5. Oak Ridge Institute for Science and Education (ORISE)	6. Lawrence Berkeley National Laboratory (LBNL)
6. Pacific Northwest National Laboratory (PNNL)	7. Lawrence Livermore National Laboratory (LLNL)
7. Savannah River (SR) Site	8. Los Alamos National Lab (LANL)
8. West Valley	9. Oak Ridge National Laboratory (ORNL)
	10. Sandia National Laboratory (SNL)
9. Waste Isolation Pilot Project (WIPP)	11. Stanford Linear Accelerator (SLAC) 12. Princeton Plasma Physics Laboratory (PPPL)
10. Yucca Mountain Project (YMP)	13. Thomas Jefferson Laboratory (TJL) 14. Y-12

4. Trend Analysis

As a preliminary analysis, we plotted raw time-series data of the 10 VPP sites and 14 non-VPP sites on a graph to visualize the trend. Because the resulting graph with 24 curves is very crowded, it is not shown here. To alleviate this problem, we averaged the data for the VPP and non-VPP sites and plotted graphs to show the trend of TRC and DART rates. A technique called Kalman filtering is used to smooth the time series. Figure 3 shows the TRC rate trend and Figure 4 shows the trend of DART rates from the first quarter of 1995 through the second quarter of 2004.

The charts below illustrate that both TRC and DART rates at DOE VPP sites are substantially below those of the non-VPP sites during the period studied. The plots also show that injury and illness rates at VPP and the non-VPP sites trended downward; suggesting that overall safety performance at the DOE complex has improved during this period. This is consistent with the general trend of private industry in the United States, according to the Bureau of Labor Statistics (BLS), even though DOE averages are usually lower than those of private industry. In 1995, the Department of Energy initiated a safety program called the Integrated Safety Management System (ISMS) and mandated all DOE contractors to adopt it. Another program, the Behavior-Based Safety (BBS) program was adopted by several DOE contractors. With the exception of two VPP contractors (those at the SPR and SR sites), none of the VPP sites practiced BBS. In any case, the second tenet of VPP, “worker participation” bears similarity to BBS and has the same objective, although the process of observing worker behavior in BBS is not part of the VPP. Since ISMS is common to all DOE contractors, (whether VPP or non-VPP), we concluded that the reduction of injury rates shown in Figures 3 and 4 can be attributed to VPP.

Figure 3: Mean TRC Time Series

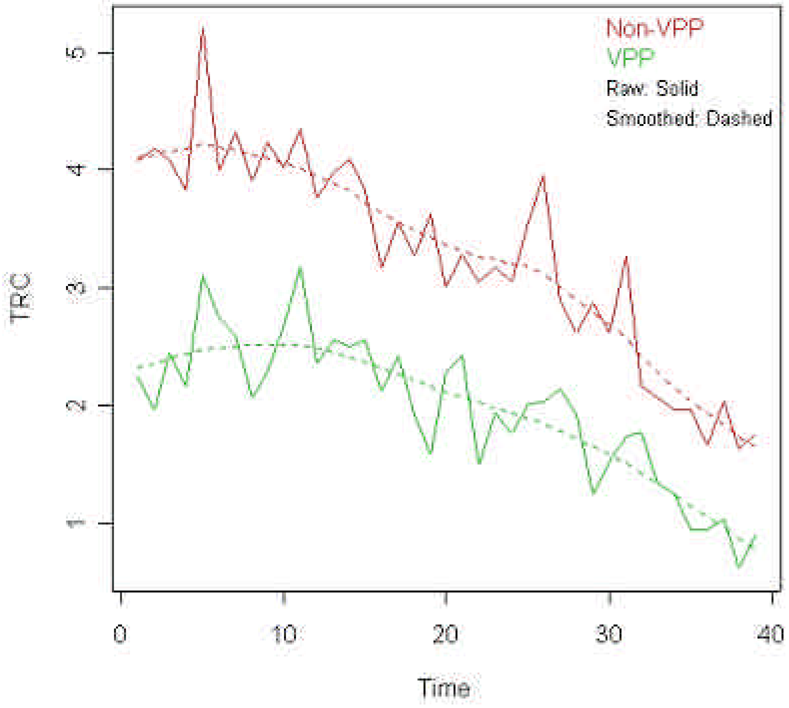
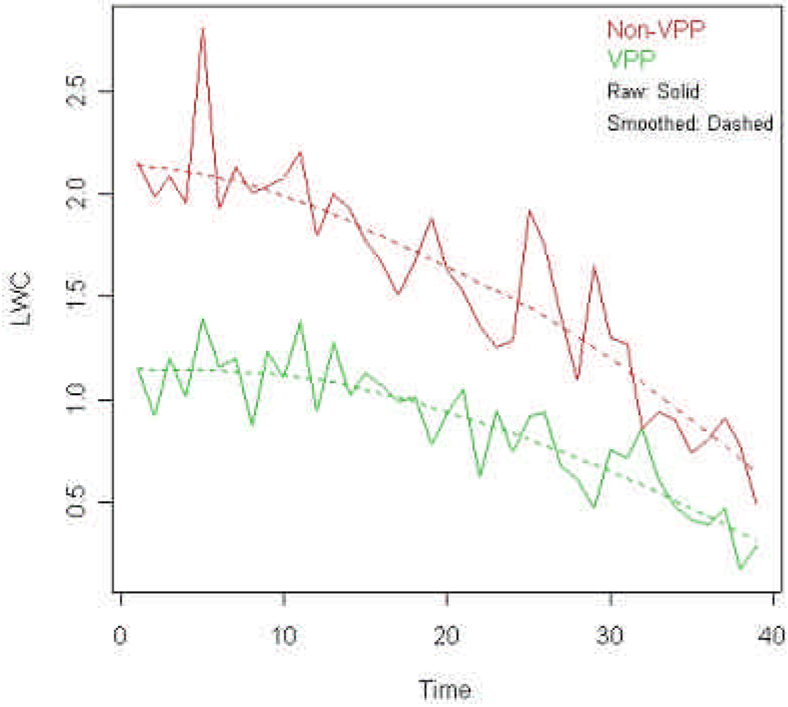


Figure 4: Mean LWC Time Series

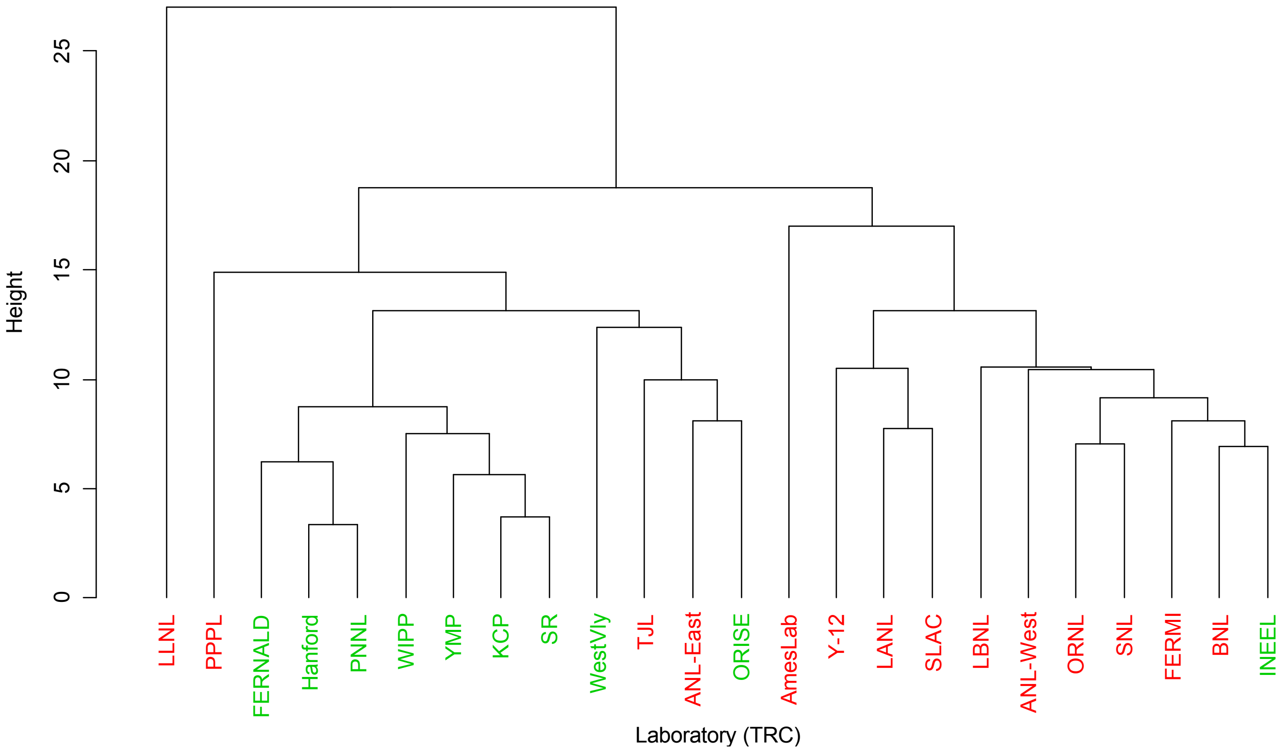


Some non-VPP sites may have satisfied the VPP criterion that three-year averages of injury and illness rates should be below those of comparable private industry. This criterion, however, is only one of the five major tenets of VPP: Management Leadership, Employee Involvement, Hazard Prevention and Control, Work Site Analysis, and Training. Previous reports by Sastry, Bowser, and Smith (2002), (2004) considered the benefits of VPP (value added).

5. Clustering

We used the TRC and DART data to generate clusters and to find similar patterns. TRC data produced the dendrogram displayed in Figure 5, which 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 a few laboratories clustered into the wrong group; for example, the VPP member INL clustered into the non-VPP group.

Figure 5: Clustering of TRC Data



As explained in Section 2, we identified site groups or clusters by drawing a line across the dendrogram (sometimes called tree because of its structure) at a reasonable height. The hierarchy of clusters in a dendrogram is usually derived by drawing the line across the tree at different heights. The sites that belong to a subtree; i.e., all sites that are still connected below the line are members of the same group or cluster. For example, when we drew a line at a height of 14, we generated the subtree structure displayed in Figure 6.

Figure 6 illustrates that the DOE sites can be broadly classified into five clusters:

Cluster 1. LLNL
Cluster 2. PPPL
Cluster 3. Fernald, Hanford, ORISE, and others (all VPP sites)
Cluster 4. Ames Lab
Cluster 5. Y-12, LANL, and others (all non-VPP sites except for INL)

A plot of smoothed time series with the above five clusters is shown in Figure 7.

Figure 6: Clustering at Height 14: TRC Rates

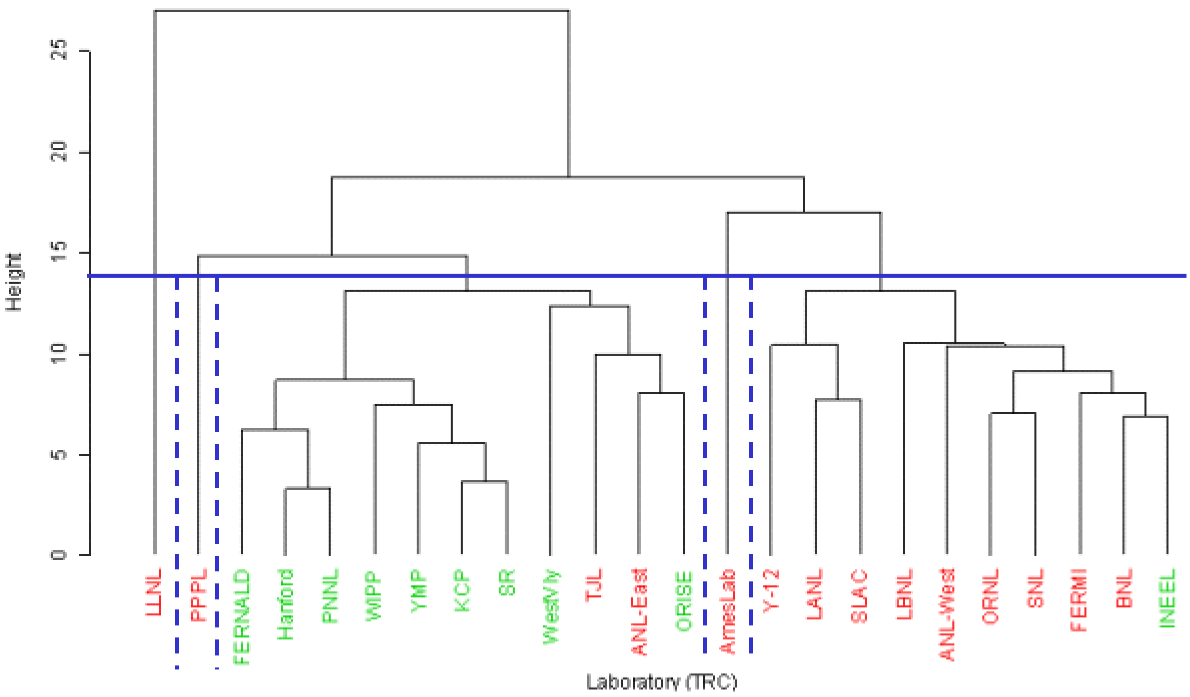


Figure 7: Smoothed Time Series with Clusters of Laboratories

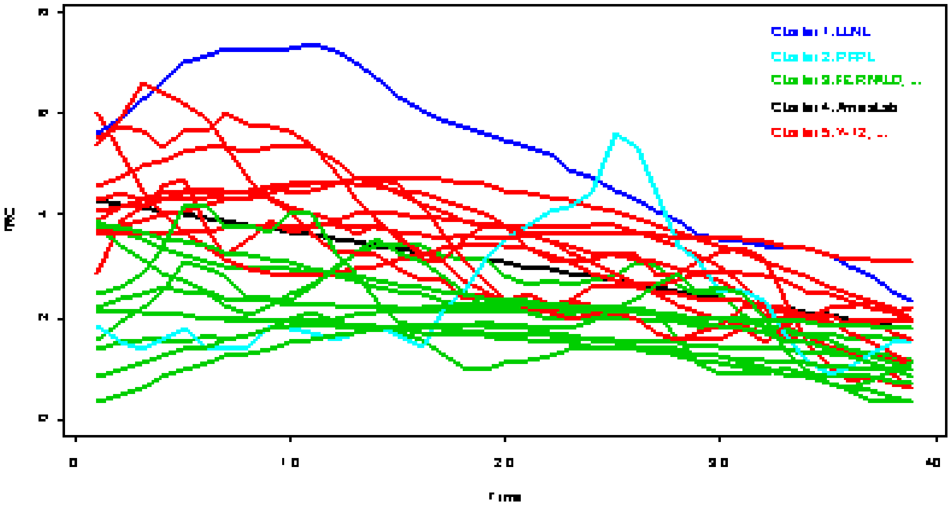


Figure 7 illustrates that sites in cluster 3 (all VPP sites) and cluster 2 had the lowest injury rates. Since these are mostly VPP sites, we concluded that the injury rates of the VPP sites were lower than those of non-VPP sites.

6. Conclusion

Based on the statistical analysis presented in this report, we concluded that the injury and illness rates of the DOE VPP sites were substantially lower than the rates of non-VPP sites over the time period we studied.

Appendix A: References

1. Leo Breiman, J. Friedman, R. Olshen, and C. Stone, (1984). *Classification and Regression Trees*, Wadsworth Publishing Co., Belmont, CA.

2. Leo Breiman (2001). *Random Forests.* Machine Learning, Volume 45, 5-32.

3. Leo Brieman, *Looking Inside the Black Box.* Wald Lecture II, University of California, Berkeley.

4. Richard A. Becker, John Chambers, and Allan Wilks (1988). *The New S Language*. Chapman & Hall, New York. This book is often called the "Blue Book."

5. John Chambers and Trevor Hastie, editors (1992). *Statistical Models in S*. Chapman & Hall, New York. This book is often called the "White Book."

6. Trevor Hastie, Robert Tibhirani, and Jerome Friedman (2001). *The Elements of Statistical Learning*, Springer Publishing Co, Berlin.

7. Rama Sastry, Rex Bowser, and David Smith (2002). *The Value Added of the Department of Energy Voluntary Protection Program*, DOE/EH-0647, June 2002.

8. Rama Sastry, Rex Bowser, and David Smith (2004). *The Value Added of the Department of Energy VPP, 2004 Update*, DOE/EH-0690, December 2004.

9. Holger Schwender et al. (2004). *A pilot study on the application of statistical classification procedures to molecular epidemiological data*, Toxicology Letters, 151 (2004) 291-299.

10. W.N. Venables and B.D.Ripley (2002). *Modern Applied Statistics with S*, Springer Publishing Co, Berlin.

11. W. N. Venables et al. (2002). *An Introduction to R*, Published by Network Theory Limited.

12. V.N. Vapnik (1988). *Statistical Learning Theory*. John Wiley & Sons, New York.