Hanford Waste Tank Grouping Study

KM Remund
Pacific Northwest National Laboratory

BC Simpson
Westinghouse Hanford Company

Prepared for Westinghouse Hanford Company

September 30, 1996

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible electronic image products. Images are produced from the best available original document.
Executive Summary

This letter report discusses the progress and accomplishments of the Tank Grouping Study in FY96. Forty-one single-shell tanks (SSTs) were included in the FY96 study, which is twenty-six more tanks than were included in FY95. In FY96, technical enhancements were also made to data transformations and tank grouping methods. The first focus of the FY96 effort was a general tank grouping study in which the 41 SSTs were grouped into classes with similar waste properties. The second FY96 focus was a demonstration of how multivariate statistical methods can be used to help resolve tank safety issues. The highlights of these two studies are summarized below.

General Tank Grouping Study – In this study, the 41 tanks were grouped based on concentrations of 14 chemical and radiological constituents (selected from the 25 variables studied). These groups of tanks were compared to a grouping of tanks based on historical information. The following are the most significant results of this analysis:

1. Potential Cost Reduction for Sampling Tanks with 224 Waste – The group of tanks containing 224 waste (comprised of Tanks B-201, B-202, B-203, B-204, and T-111) showed excellent agreement with the historical grouping. This finding, coupled with others discussed in the results section of this report, indicate a good potential for saving sampling costs on the remaining unsampled tanks containing 224 waste (Tanks T-201, T-202, T-203 and T-204). The data from the sampled 224 waste tanks may be pooled and used to help make inferences regarding the unsampled tanks.

2. Coherent Historical & Sampling Groups – There were six groups of tanks identified in the sampling data based on the 14 variables. Each of these six sampling data tanks groups matched up reasonably (better than 50% overlap) with a historical tank group. Two of the six groups contain saltcake tanks (B1/T1 and S1/S2 saltcakes). The four remaining groups contain sludge tanks holding 224 waste, ferrocyanide waste, 1C/2C waste, and R1/CWP wastes.

3. Additional Anticipated Benefits – Other tank groups identified in this year’s study can be further investigated for possible sampling savings for unsampled tanks. The 1C/2C waste tank group is a good candidate for investigation. The findings of this research effort will be combined with historic modeling efforts to improve the quality of the model and the Standard Inventory Estimate. The present tank groups should become better defined as tanks continue to be added into the study and as the research techniques are refined in FY97.

Multivariate Classification of Tanks for Safety Issues – At present tanks are classified as watchlist or non-watchlist tanks for a particular safety issue by considering the variables of interest one-at-a-time. This approach can be supplemented by a multivariate approach which considers all variables concurrently. A multivariate approach takes into consideration the correlations that exist between variables. The tank characterization pro-
gram directed us to demonstrate this approach with the ferrocyanide tank safety issue. This demonstration showed the utility of such an approach for:

1. **Identifying Potential Tanks for Watchlist Inclusion or Exclusion** – The misclassified tanks from the multivariate classification approach are candidates for watchlist additions or watchlist exclusions. From the ferrocyanide illustrative example, Tank T-107 was identified as a candidate for exclusion from the ferrocyanide watch list.

2. **Reducing the Number of Variables** – Using information about interrelationships between variables for a particular safety issue, a reduced set of variables can be derived which retains only those variables that contribute significantly to the tank classification. This focuses analytical research in a direction dictated by the data and enables a better and more in-depth understanding of the safety issue. Additionally, it may become possible to evaluate tanks with missing data (which are not supported by the traditional watchlist evaluation process) In the ferrocyanide issue example, the number of variables was reduced from 8 to 3 without losing any group discriminating power.

**Recommendations**

1. **Investigate Cost Savings** – Using the results of this study and others, investigate the feasibility of using the present 224 waste tank group data to make inferences regarding the unsampled tanks in this group. Additional investigations would include cross-tank ANOVA models and cluster analysis of individual analytical results for the sampled 224 waste tanks.

2. **Continue General Tank Grouping Study** – Continue the general tank grouping task in the following areas:
   
   (a) Conduct cross-tank grouping studies using segment and sub-segment sample results rather than tank means. This study would show how robust the present tank groupings are when considering spatial variability within the tanks and analytical variability.

   (b) Continue cross-tank grouping studies which group tanks according to sampling means, Historical Tank Content Estimates (HTCEs), and Hanford Engineering Analysis Estimates.

   (c) Make technical enhancements to data transformations (e.g., an empirical cumulative density function approach) and clustering methods (e.g., optimization clustering methods – see Reference [12]).

   (d) Add more tanks to the cluster analyses as sampling data becomes available.

   (e) Search for additional cost savings associated with other “well-defined” tank groups.

3. **Safety Issues Research** – Apply the multivariate classification methods demonstrated in this paper to other tank safety issues such as Flammable Gas.
Contents

List of Terms v

1 Introduction 1

2 Tank Grouping Data 2

3 General Tank Grouping Study Results 2
   3.1 Grouping Strategy 2
   3.2 Analysis Variable List 3
   3.3 Cluster Analysis Results 3
   3.4 ‘224 Waste’ Tanks Cluster 10

4 Multivariate Approach to Tank Safety Issues 11
   4.1 Ferrocyanide Issue Variables and Tank Lists 12
   4.2 Multivariate Ferrocyanide Watchlist Classification 13

5 Conclusions and Recommendations 15

6 References 18

A Tank Grouping Variables 22

B Data Preparation 24
   B.1 Handling Missing Values 24
   B.2 Transformations 25

C Boxplots of Tank Means by Cluster 27

D ‘224 Waste’ Cluster Confidence Interval Plots 33
List of Figures

1  Cluster Analysis Results ................................................. 5
2  Group Separation ......................................................... 9
3  Nickel, Calcium and Cesium-137 Tank Means 3D Plot ............. 16
4  Data Transformations ..................................................... 26
5  Boxplots of Tank Means by Cluster ................................... 28
6  Boxplots of Tank Means by Cluster (continued) .................. 29
7  Boxplots of Tank Means by Cluster (continued) .................. 30
8  Boxplots of Tank Means by Cluster (continued) .................. 31
9  Boxplots of Tank Means by Cluster (continued) .................. 32
10 Confidence Interval Plots for 224 Waste Cluster .................. 34
11 Confidence Interval Plots for 224 Waste Cluster (continued) ... 35
12 Confidence Interval Plots for 224 Waste Cluster (continued) ... 36
13 Confidence Interval Plots for 224 Waste Cluster (continued) ... 37

List of Tables

1  General Tank Grouping Study Analyte List .......................... 4
2  Sampling Data Cluster Analysis Tank Groups ....................... 6
3  Tank Cluster Means ...................................................... 8
4  Ferrocanide Safety Issue Tank Grouping Analyte List .............. 13
5  Comparison of Classification Methods Results ..................... 14
6  Tanks Classified Differently by Two Methods ........................ 14
7  FY96 Analytes for SSTs .................................................. 23
### List of Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1C/2C</td>
<td>Bismuth Phosphate Process 1st and 2nd Cycle Decontamination Waste</td>
</tr>
<tr>
<td>224</td>
<td>Lanthanum Fluoride Process Waste from the 224 Building</td>
</tr>
<tr>
<td>A1/A2</td>
<td>Saltcake wastes from the A1/A2 Campaigns</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>B1/T1</td>
<td>Saltcake wastes from the B1/T1 Campaigns</td>
</tr>
<tr>
<td>BY</td>
<td>Saltcake waste from the BY Evaporator Campaign</td>
</tr>
<tr>
<td>CWP</td>
<td>PUREX Coating Waste</td>
</tr>
<tr>
<td>EM Algorithm</td>
<td>Expectation Maximization Algorithm</td>
</tr>
<tr>
<td>FeCN</td>
<td>Ferrocyanide Waste</td>
</tr>
<tr>
<td>FY</td>
<td>Fiscal Year</td>
</tr>
<tr>
<td>HCA</td>
<td>Hierarchical Cluster Analysis</td>
</tr>
<tr>
<td>HTCE</td>
<td>Historical Tank Content Estimate</td>
</tr>
<tr>
<td>KNN</td>
<td>Kth Nearest Neighbor Classification Method</td>
</tr>
<tr>
<td>LANL</td>
<td>Los Alamos National Laboratory</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>MANOVA</td>
<td>Multivariate Analysis of Variance</td>
</tr>
<tr>
<td>PNNL</td>
<td>Pacific Northwest National Laboratory</td>
</tr>
<tr>
<td>PUREX</td>
<td>Plutonium and Uranium Extraction Process</td>
</tr>
<tr>
<td>R1</td>
<td>REDOX Waste Sludge</td>
</tr>
<tr>
<td>REDOX</td>
<td>Reduction Oxidation Process</td>
</tr>
<tr>
<td>S1/S2</td>
<td>Saltcake wastes from the S1/S2 Campaigns</td>
</tr>
<tr>
<td>SST</td>
<td>Single-Shell Tank</td>
</tr>
<tr>
<td>TCR</td>
<td>Tank Characterization Report</td>
</tr>
<tr>
<td>TIC</td>
<td>Total Inorganic Carbon</td>
</tr>
<tr>
<td>TOC</td>
<td>Total Organic Carbon</td>
</tr>
<tr>
<td>TWINS</td>
<td>Tank Waste Information Network System</td>
</tr>
<tr>
<td>UR</td>
<td>Uranium Recovery Process Waste</td>
</tr>
<tr>
<td>WHC</td>
<td>Westinghouse Hanford Company</td>
</tr>
</tbody>
</table>
Acknowledgments

We gratefully acknowledge the contribution of Ken Higbee, Don Daly and Tom Ferryman. They have provided valuable technical input and review of this work.
1 Introduction

This end-of-year report is being written for the Tank Characterization Program to summarize the Hanford Tank Grouping Study efforts in FY96. Results of a general tank grouping study and a second tank grouping study that focuses specifically on the ferrocyanide safety issue are given. The general tank grouping study is expected to continue as more tank data become available. Other safety issue tank grouping studies, similar to the ferrocyanide issue study, are expected to continue also.

This introductory section contains background information on the need for tank grouping and gives a brief description of FY95 tank grouping efforts. This section also discusses the study's objective and scope. Section 2 briefly describes the data that was used for tank grouping. Section 3 contains the results of the general tank grouping study while Section 4 contains the results of the ferrocyanide safety issue tank grouping study. Section 5 contains some conclusions and suggestions for future tank grouping work. The appendices contain detailed information about the data gathering effort and data preparation steps that were taken before statistical analysis.

Throughout this report the words 'group', 'class' and 'cluster' will be used interchangeably to refer to tank groups.

The primary objective of this tank grouping study is to group tanks into classes with similar waste properties. Meeting this objective would reduce costs by reducing tank sampling requirements and would allow tanks in each class to be treated similarly in mitigation and remediation activities. That is, safety issues could be resolved and tank waste could be processed in part on a tank group level rather than solely on an individual tank level. Also, due to tank grouping efforts, only a subset of the tanks within a group may need to be sampled, or fewer samples may need to be taken from all tanks within the group. Reference [31] describes how sampling and historical grouping models could be used to reduce tank sampling requirements for a group of tanks containing first and second cycle decontamination waste from the bismuth phosphate process.

This study groups tanks by two independent efforts, using a set of waste type discriminating variables (variable selection is discussed in Section 2). The first approach uses sampling data; the second approach uses historical model information (see Reference [1]) on the tanks and variables of interest. Grouping results from these two independent models can be compared to identify meaningful tank groups. The example from Reference [31] mentioned above and another example discussed in Section 3 of this report show how the results of these two independent tank grouping efforts are used together.

Stated concisely, accomplishing the study objective will benefit the characterization process by:

1. Reducing Sampling Requirements - Rather than sampling every tank in a group with similar waste properties, sample only a subset of the tanks in the group or take fewer samples from each of the tanks in the group.

1The authors recognize that the ferrocyanide issue will soon be removed from the list of tank safety issues. This issue is used to demonstrate the utility of a tank grouping approach to help in the resolution of safety issues in general.
2. Optimizing the Sampling Strategy - Allocate limited resources to sample those tanks that do not belong to a “well-defined” tank group.

3. Improving Uncertainty Estimates - The sampling data across all tanks within an identified group can be pooled. This pooling of data would yield improved estimates of uncertainty for inventory estimates.

4. Safety Issue Resolution - Tank watchlists can be refined by considering correlations that exist between variables important to the safety issues.

2 Tank Grouping Data

Sampling estimates were gathered for 25 chemical and radiological waste constituents (see Appendix A) on each of 41 single-shell tanks. Many of the mean concentration estimates were obtained from tank characterization reports (TCR) while others were obtained from analytical data taken from the Tank Waste Information Network System (TWINS). In the cases where the data were taken from TWINS, analysis of variance (ANOVA) models similar to those used in the TCRs were fit to the data to get mean concentration and spatial variability estimates.

Once the sampling data was gathered, it had to be prepared for statistical analysis. Two data preparation steps were taken before analysis:

1. Missing Data Values - At the conclusion of the data gathering effort, it was noted that there were several “holes” in the dataset (i.e., missing values). To do the tank grouping analyses, these missing values had to be estimated without introducing any changes to the covariance structure of the multivariate data. The approach used to estimate these missing values is described in Appendix B.1.

2. Data Transformations - Due to varying concentration ranges of the waste constituents and the asymmetric distributions of the constituents, the data had to be transformed so that meaningful results could be obtained from the tank grouping analysis. A discussion of the data transformations and the basis for them is given in Appendix B.2.

3 General Tank Grouping Study Results

This section contains the analysis approach and results of a general tank grouping study. There are 41 single-shell tanks included in the present study compared to the 15 that were included in the FY95 effort.

3.1 Grouping Strategy

The basic strategy for grouping tanks for this general tank grouping study is as follows:

1. Group tanks into classes using waste constituent concentrations from sampling results only. There are 41 SSTs represented in this sampling data grouping effort at present.
2. Compare the tank groups from Step 1 with the historical tank groups derived from process history information provided by LANL\textsuperscript{2}. All 149 SSTs are represented in this historical grouping of tanks. These two grouping efforts are obviously independent of each other.

3. The final step in this grouping strategy is to team with the technical staff at WHC and PNNL (chemists and chemical engineers) to interpret the results of this comparison. At this step, if reasonable agreement is found between the sampling and historical tank groupings, information from sampled tanks in a group may be used to make inferences about unsampled tanks in a group. This result could save sampling resources by reducing tank sampling requirements.

At present, steps 1 and 2 have been carried out. With the delivery of this report to the tank characterization program, step 3 can be considered as well.

### 3.2 Analysis Variable List

At the beginning of FY96, the authors solicited input from technical staff at WHC and PNNL to determine which variables should be included as grouping criteria for this general tank grouping study. The analytes in Table 1 were approved by this group as a list of waste properties of interest to a wide number of waste safety and disposal programs. The analytes in the table which are marked with an asterisk are considered to be waste type discriminators. These analytes were used as variables to group tanks. Further details of this analyte selection effort are contained in Appendix A.

### 3.3 Cluster Analysis Results

We begin this section with a conceptual description of the grouping methodology that is used in the data analyses. Each tank can be thought of as a single point in multi-dimensional space. Each measured tank variable (e.g., cesium-137 or aluminum concentration) represents a single dimension in this multi-dimensional space. If we were only interested in three measured variables (waste constituents), the tanks would be positioned like balls hanging from a ceiling of a room at different heights and horizontal locations. The distance between tank points can be measured in three dimensions. This distance is called Euclidean distance. Tank points that are close together relative to other tank-points are called groups or clusters (i.e., tanks with similar waste properties). Although there are 14 dimensions in this tank grouping exercise, the simplicity of the three-dimensional example given can be thought of as the tank grouping results are presented.

A technique called Hierarchical Cluster Analysis (HCA) using Complete Linkage (see Reference [12]) was used as the primary method to group tanks into classes with similar

\textsuperscript{2}This historical grouping of tanks is a qualitative grouping of the tanks. The grouping criteria that was used to group tanks for this effort was that for a tank to belong to a given group, it had to contain at least 40\% of the waste type that identifies the group. This historical grouping effort is closely tied to concurrent efforts at LANL to provide HTCEs.
Table 1: General Tank Grouping Study Analyte List

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Analyte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aluminum*</td>
<td>Americum-241</td>
</tr>
<tr>
<td>Bismuth*</td>
<td>Calcium*</td>
</tr>
<tr>
<td>Cesium-137*</td>
<td>Chloride</td>
</tr>
<tr>
<td>Chromium*</td>
<td>Fluoride*</td>
</tr>
<tr>
<td>Iron*</td>
<td>Lanthanum*</td>
</tr>
<tr>
<td>Lead</td>
<td>Manganese*</td>
</tr>
<tr>
<td>Nickel*</td>
<td>Nitrate</td>
</tr>
<tr>
<td>Phosphorus*</td>
<td>Plutonium-239/240</td>
</tr>
<tr>
<td>Silicon</td>
<td>Sodium</td>
</tr>
<tr>
<td>Strontium-90*</td>
<td>Sulfate</td>
</tr>
<tr>
<td>Uranium*</td>
<td>TIC (Carbonate)</td>
</tr>
<tr>
<td>TOC</td>
<td>Water</td>
</tr>
<tr>
<td>Nitrite*</td>
<td></td>
</tr>
</tbody>
</table>

waste properties. This method was supplemented by other statistical techniques some of which will be mentioned later in this section.

Figure 1 is a summary of the results from a HCA on the tank sampling data available by September 1, 1996. There is a branch on the inverted tree for each of the 41 tanks along the right side of the figure. The name of each tank is listed as well as the historical group to which the tank belongs. Along the bottom of the figure (the y-axis) is a measure of overall “distance” between tanks in 14 dimensions (14 waste constituents). The tanks that are “closest” together are grouped together first at the bottom of the tree. For example, Tanks B-203 and B-204 are grouped together first since they are the closest together, or otherwise stated, since they have very similar waste properties. Proceeding upward along the y or distance axis, these two tanks are followed by the grouping of Tanks BX-107 and C-110 and then the grouping of C-109 and C-112. If one continues up the tree from there, the single tanks and groups of tanks continue to be joined until all tanks are in one super group at the top of the inverted tree.

Another way of describing the grouping process is as follows. The process of grouping the tanks starts with 41 groups where each group consists of one tank. After the first step in the grouping algorithm there are 40 tank groups, 39 of which have one tank and one group that has two tanks (Tanks B-203 and B-204). At step two, there are 39 groups, 37 of which have one tank and two groups that have two tanks each (Tanks B-203 and B-204 in one group and Tanks BX-107 and C-110 in the other). This process continues until the final step when all tanks are in one “super” group.

Figure 1 can be quite handy in seeing the pattern of tank similarities based on the selected 14 variables. It is important to remember that if other measured tank variables are used, a different tank grouping may result. In this report, we work with 6 tank groups, since this is the optimal stopping point as indicated by the stopping rule described next.

The two dashed lines drawn through Figure 1 indicate two possible cuts in the tree to
Figure 1: Cluster Analysis Results
break the tanks into 5 or 6 groups. The Calinski-Harabaz stopping rule (see Reference [12]) was used to determine these optimal places to cut the tree. The stopping rule seeks to find the cut in the tree (i.e., number of clusters) where the between-group differences are maximized and the within-group differences are minimized. Based on the stopping rule, the cut at 6 groups is the best, followed closely by the cut at 5 groups. Both the 5 and 6 tank group cuts have good physical meaning.

The six boxed-in areas along the right-side of Figure 1 are the six tank groups. These tank groups are enumerated in Table 2. The Table also lists the historical group labels for each tank, and shows the agreement between the historical grouping and the sampling data tank grouping. For example, Tanks B-201, B-202, B-203, and B-204 are the only tanks in this study that are from the 224 historical group. These tanks are all in the same group based on the sampling data cluster analysis, showing agreement between the sampling and historical groupings. Tank T-111, on the other hand, is a member of this sampling group, but is not boxed-in with the other tanks in Figure 1 since it does not belong to the same historical group as the other tanks mentioned above. Discrepancies between the groupings, like Tank T-111, are indicated in Figure 1 as "non-boxed" in tanks and in Table 2 with asterisks.

Table 2: Sampling Data Cluster Analysis Tank Groups

<table>
<thead>
<tr>
<th></th>
<th>R1/CWP Cluster</th>
<th>FeCN Cluster</th>
<th>1C/2C Cluster</th>
<th>224 Cluster</th>
<th>B1/T1 Cluster</th>
<th>SALT Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-103</td>
<td>CWP</td>
<td>C-109</td>
<td>FeCN</td>
<td>BX-109</td>
<td>UR</td>
<td>B-203</td>
</tr>
<tr>
<td>T-105</td>
<td>1C/2C</td>
<td>C-112</td>
<td>FeCN</td>
<td>TY-106</td>
<td>UR</td>
<td>B-204</td>
</tr>
<tr>
<td>A-102</td>
<td>A1/A2</td>
<td>C-105</td>
<td>CWp</td>
<td>BX-105*</td>
<td>MW</td>
<td>T-111*</td>
</tr>
<tr>
<td>BX-104</td>
<td>CWP</td>
<td>BY-108</td>
<td>FeCN</td>
<td>BY-104*</td>
<td>BY</td>
<td>B-201</td>
</tr>
<tr>
<td>T-102</td>
<td>CWP</td>
<td>BY-110</td>
<td>FeCN</td>
<td>BX-112</td>
<td>1C/2C</td>
<td>B-202</td>
</tr>
<tr>
<td>S-101</td>
<td>R1</td>
<td>C-111</td>
<td>FeCN</td>
<td>B-110</td>
<td>1C/2C</td>
<td></td>
</tr>
<tr>
<td>S-104</td>
<td>R1</td>
<td>TY-104*</td>
<td>UR</td>
<td>B-111</td>
<td>1C/2C</td>
<td></td>
</tr>
<tr>
<td>SX-108</td>
<td>R1</td>
<td>U-110*</td>
<td>1C/2C</td>
<td>B-106*</td>
<td>B1/T1</td>
<td></td>
</tr>
</tbody>
</table>

* Differences between sampling data grouping and historical grouping

Appendix C contains tank mean boxplots that show the separation between the sampling data tank groups for each analyte. These boxplots were used to determine the distinguishing cluster characteristics. The following is a discussion of the characteristics of each of the groups in Figure 1 and Table 2 (from left to right):

1. R1/CWP Cluster - This cluster is primarily comprised of tanks from PUREX Coating Waste (CWP) and REDOX waste (R1) historical groups. These two historical groups are expected to share several characteristics based on waste stream source terms common to both. The characteristics that distinguish this cluster from the other 5 clusters are high levels of aluminum, chromium, strontium-90, and low levels of bismuth, fluo-
ride, and phosphorus. Tanks T-105 and A-102 are part of this cluster but are considered "discrepancies" since they belong to historical groups other than R1 and CWP.

2. Ferrocyanide Cluster - The second group from the left in Figure 1 and Table 2 includes eight tanks, five of which are in the ferrocyanide historical group. These five tanks and Tank TY-104 (belonging to the UR historical waste group) are also on the ferrocyanide watchlist. Tanks marked with an asterisk in Figure 1 are on the ferrocyanide watchlist. The remaining two tanks in the group (Tanks C-105 and U-110) are labeled as "discrepancies" since the historical records do not indicate that they contain ferrocyanide wastes. The characteristics that distinguish this group of tanks from others are medium levels of aluminum, and high calcium, cesium-137, iron, nickel, strontium-90, and uranium.

3. 1C/2C Cluster - The third tank cluster is identified as the 1C/2C group, since 9 of the 12 tanks in this cluster are from either the 1C/2C or UR historical groups. The tanks containing uranium recovery waste (UR) are expected to have similar characteristics to the 1C/2C tanks, since UR waste is 1C/2C waste that has been reprocessed to recover uranium from it. Tanks BX-105, BY-104 and B-106 are labeled as "discrepancies" since they are from historical groups other than 1C/2C and UR. The characteristics that distinguish this tank group from others are the low levels of aluminum, cesium-137 and strontium-90 and relatively high levels of bismuth and phosphorus. We should note that Tanks BY-104 and T-107 are on the ferrocyanide watchlist.

4. '224' Cluster - This cluster of five tanks shows the strongest agreement with the historical tank grouping. All four 224 historical group tanks represented in this study are included in this cluster. The fifth tank (T-111) is known to have both 224 and 1C/2C waste in it. Due to this result, it may be reasonable to put this tank in the 224 historical group rather than the 1C/2C historical group. This cluster has characteristically high levels of bismuth, manganese and lanthanum and low levels of radionuclides, aluminum and nitrite. This cluster is discussed in more detail in the next section.

5. Saltcake Clusters - Both clusters 5 and 6 have tanks that contain almost solely saltcake. If the clustering tree is cut at five clusters rather than six, these two clusters merge together, as can be seen in Figure 1. The two saltcake clusters are obviously higher than the other clusters in sodium and nitrate, which were not included in this clustering exercise since these two constituents are not considered waste type discriminators. Both of these clusters are also low in calcium, bismuth, iron, and strontium-90. The B1/T1 saltcake cluster is also very high in phosphorus and fluoride and has low levels of chloride and chromium. According to the historical record, Tank S-107 primarily contains R1 sludge. However, the historical record also reports that this tank contains a large amount of saltcake.

Figure 2 shows the separation between the 6 tank clusters in the two new dimensions\(^3\) from the original 14 that provide the greatest separation between clusters while minimizing

\(^3\)These new dimensions are linear combinations of the original 14 variables. These linear combinations were determined using linear discriminant analysis (LDA).
the differences between tanks within a cluster. Our analysis results show that statistical differences\(^4\) exist between groups. Note that the 224 waste cluster (Tanks B-201, B-202, B-203, B-204 and T-111) is separated from the other clusters by the largest margin. This difference between the 224 cluster and the other clusters was also found to be significant at the 95\% confidence level. The 224 cluster is discussed further in the next section. Table 3 contains the cluster means for each analyte. The cluster differences indicated in the figures and by the statistical tests can also be seen in this tabulation.

Table 3: Tank Cluster Means

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Units</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R1/CWP</td>
</tr>
<tr>
<td>Al</td>
<td>µg/g</td>
<td>1.02e+05</td>
</tr>
<tr>
<td>Am-241</td>
<td>µCi/g</td>
<td>4.19e-01</td>
</tr>
<tr>
<td>Bi</td>
<td>µg/g</td>
<td>3.42e+02</td>
</tr>
<tr>
<td>Ca</td>
<td>µg/g</td>
<td>1.57e+03</td>
</tr>
<tr>
<td>Cs-137</td>
<td>µCi/g</td>
<td>9.96e+01</td>
</tr>
<tr>
<td>Cl</td>
<td>µg/g</td>
<td>2.35e+03</td>
</tr>
<tr>
<td>Cr</td>
<td>µg/g</td>
<td>4.35e+03</td>
</tr>
<tr>
<td>F</td>
<td>µg/g</td>
<td>2.20e+02</td>
</tr>
<tr>
<td>Fe</td>
<td>µg/g</td>
<td>1.31e+04</td>
</tr>
<tr>
<td>La</td>
<td>µg/g</td>
<td>1.85e+01</td>
</tr>
<tr>
<td>Pb</td>
<td>µg/g</td>
<td>4.43e+02</td>
</tr>
<tr>
<td>Mn</td>
<td>µg/g</td>
<td>2.88e+03</td>
</tr>
<tr>
<td>Ni</td>
<td>µg/g</td>
<td>4.25e+03</td>
</tr>
<tr>
<td>NO2</td>
<td>µg/g</td>
<td>3.17e+04</td>
</tr>
<tr>
<td>NO3</td>
<td>µg/g</td>
<td>1.13e+05</td>
</tr>
<tr>
<td>P</td>
<td>µg/g</td>
<td>1.52e+03</td>
</tr>
<tr>
<td>Pu-239/40</td>
<td>µCi/g</td>
<td>2.03e+00</td>
</tr>
<tr>
<td>Si</td>
<td>µg/g</td>
<td>2.74e+03</td>
</tr>
<tr>
<td>Na</td>
<td>µg/g</td>
<td>9.70e+04</td>
</tr>
<tr>
<td>Sr-90</td>
<td>µCi/g</td>
<td>1.15e+03</td>
</tr>
<tr>
<td>SO4</td>
<td>µg/g</td>
<td>4.34e+03</td>
</tr>
<tr>
<td>TIC</td>
<td>µg/g</td>
<td>4.67e+03</td>
</tr>
<tr>
<td>TOC</td>
<td>µg/g</td>
<td>4.71e+03</td>
</tr>
<tr>
<td>U</td>
<td>µg/g</td>
<td>1.39e+04</td>
</tr>
<tr>
<td>H2O</td>
<td>µg/g</td>
<td>4.00e+05</td>
</tr>
</tbody>
</table>

\(^4\)A multivariate analysis of variance along with contrasts was used to test the statistical significance of the grouping.
Figure 2: Group Separation

Two Dimensional Group Separation Plot
3.4 ‘224 Waste’ Tanks Cluster

As noted in the last section, the 224 waste tanks show good agreement between the sampling and historical groupings and also show a high level of separation from the other tank clusters. Because of these findings, this group might be especially useful to aid in future sampling activities.

The 224 waste historical group consists of eight tanks. Four of these tanks (B-201, B-202, B-203, and B-204) have been sampled and are included in the cluster analyses described in this report. The other four tanks in this historical group ( Tanks T-201, T-202, T-203 and T-204) have not been sampled to date. There is a good possibility that pooled sampling data from the B-Farm tanks mentioned above can be used to help make inferences regarding the four unsampled tanks if the following conditions hold:

1. The historical records (e.g., tank transfer records) for the eight 224 waste tanks are reasonably complete and the transfer histories are uncomplicated. Otherwise stated, there is good reason to believe that the T-Farm tanks and the B-Farm tanks have very similar waste process and tank transfer histories.

2. There is reasonably good agreement between the sampling and the historical grouping of the 224 waste type tanks.

3. The waste property differences (practical and statistical) among the five sampled 224 waste group tanks need to be within acceptable bounds as determined by the tank characterization program and other waste programs of interest.

Condition 2 has been satisfied by the analysis and results reported here. Conditions 1 and 3 are discussed below.

Reference [1] documents the recorded transfers made to the B and T farm 200-series tanks. These tanks received one waste type according to the available sources, lanthanum fluoride waste from the 224 Building (thus the designation 224 waste). The transfer histories for these tanks are relatively uncomplicated. They did not interact with the tank farm at large, therefore the spatial variability in these tanks is expected to be relatively small. Some heterogeneity is expected from the variability of the process and tank filling order. However, the spatial heterogeneity for the B and T 200 series tanks should be approximately the same magnitude. This can be investigated for the B 200 series tanks by clustering individual analytical results across these four tanks and noting the grouping patterns.

Appendix D contains plots of estimated means and 90% confidence intervals on the means for Tanks B-201, B-202, B-203, B-204 and T-111, and for the analytes in Table 1 for which data were readily available. For each tank, the dot in the center of the line is the mean. The arrow tips on each line are the upper and lower 90% confidence limits of the mean.

The plots in Appendix D can be used to help determine waste property differences among the 224 waste tanks (i.e., help determine if Condition 3 is met). If there is no overlap of confidence intervals for different tanks, we infer there is a difference between the tank means. Conversely, if there is an overlap of the confidence intervals for two tanks, then the tank means did not distinguish themselves from each other. For example, looking at the
plots, the chromium mean for Tank T-111 is not within the confidence interval of any of the other chromium means. Since the intervals overlap for all of the B-Farm tanks, it is not apparent that there are differences in the chromium means for these tanks. Usually to test the statistical difference between more than two means, one does an analysis of variance (ANOVA) on a cross-tank model. This was not done here, due to limited time and resources. This ANOVA model should be fit in the future FY97 effort to more adequately assess the similarity between the tank means and to get pooled mean and variance estimates across the 224 waste tanks. The Appendix D plots are given here to show roughly the level of agreement among Tanks B-201, B-202, B-203, B-204 and T-111, until the ANOVA model can be fit next year.

A review of the confidence interval plots in Appendix D shows that at present we cannot statistically distinguish between the 224 waste tanks means for most analytes. The plots show that Tanks T-111 and B-201 are occasionally different from the other tanks for some analytes. This result is a good preliminary indication of agreement between the 224 waste tank means.

The following possibilities suggest how data from the sampled 224 waste tanks may be used to help reduce costs in the characterization of the unsampled 224 waste tanks:

1. Do no further sampling of the 224 waste group tanks. Use the sampling data from Tanks B-201, B-202, B-203, B-204 and possibly T-111 to obtain pooled mean concentration and uncertainty estimates for unsampled Tanks T-201, T-202, T-203 and T-204. This option would result in a savings of eight core samples.

2. Take one sample from each of two of the T Farm 200 series tanks. A check could be made to see that the analytical results from these cores grouped reasonably with the other sampling data presently available for this group. The pooled estimates would then be used for the unsampled tanks. This option would result in a savings of six core samples.

3. Take one core sample from each of the remaining 224 waste group tanks that have not yet been sampled. Perform the same grouping check and then use the pooled estimate as in option 2. This would save four core samples.

It is envisioned that these results can be used in FY97 tank sampling plans. The promising results of this section should encourage further discussions and data analyses to determine the feasibility of reducing sampling requirements for Tanks T-201, T-202, T-203 and T-204.

4 Multivariate Approach to Tank Safety Issues

At present, tanks are being classified as watchlist or non-watchlist tanks for specific safety issues considering variables of interest one-at-a-time. This means that specific variable thresholds are considered independently of other variable thresholds. This approach is reasonable, but it may be complemented nicely by a multivariate analysis of the data that considers all variables of interest concurrently. This approach allows one to consider the correlations
between the variables of interest. This section explains a possible multivariate approach to safety issue resolution. The ferrocyanide issue is used as an example to demonstrate this classification technique, because much is already known about ferrocyanide as a result of earlier work. We recognize that the ferrocyanide issue will soon be resolved. However, the ferrocyanide issue provides a good example of how this multivariate approach can be used to address other tank safety and waste processing issues.

The statistical analysis methods referred to in this section are similar to those discussed in Section 3 in that both sets of methods are multivariate (i.e., take into consideration the correlations that exist between the variables). However, statistical analyses in this section differ from those in the previous section in the following two ways:

1. The HCA methods discussed in the previous section involves no preliminary information as to which group a tank should belong to. This is contrasted with the multivariate discrimination method used in this section, which is based on a training set of data which contains group membership information of all tanks in the training set. The method then finds a discrimination rule that can classify new tanks into the right group. The aim of this part of the study is to classify tanks as watchlist or non-watchlist tanks and therefore the multivariate discrimination method is of primary interest.

2. Different lists of variables are used in the analyses of each section. This may cause differences in the results.

The remainder of this section discusses the waste property variables and tanks that are represented in this Ferrocyanide Safety Issue example, followed by a description of the multivariate methods used and their results.

4.1 Ferrocyanide Issue Variables and Tank Lists

Table 4 lists the constituents that are considered important to the Ferrocyanide Safety Issue. Several of these constituents were checked in suspect tanks to determine if a tank should be on the ferrocyanide watchlist (i.e., if permissible thresholds were exceeded). All of these analytes are included in the statistical analysis of this section, with the exception of total cyanide and elemental strontium. These two analytes were excluded because measurements were missing for a large number of the sampled tanks.

The 41 tanks that were considered in the general tank grouping study were also used in the analysis of this section with a few exceptions. Tank T-109 was excluded from this analysis since it was missing a large proportion of measurements for the constituents listed in Table 4. Tank C-108 was added to this analysis since measurements were available for most constituents in Table 4. Tanks BY-104, BY-108, BY-110, C-108, C-109, C-111, C-112, T-107, and TY-104 are the ferrocyanide watchlist tanks represented in this study. This list of tanks represents half of the tanks that are on the ferrocyanide watchlist.
Table 4: Ferrocanide Safety Issue Tank Grouping Analyte List

<table>
<thead>
<tr>
<th>Analyte</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aluminum*</td>
<td>Calcium*</td>
</tr>
<tr>
<td>Cesium-137*</td>
<td>Nickel*</td>
</tr>
<tr>
<td>Strontium</td>
<td>Strontium-90*</td>
</tr>
<tr>
<td>Total Cyanide</td>
<td>TOC*</td>
</tr>
<tr>
<td>Uranium*</td>
<td></td>
</tr>
</tbody>
</table>

* Included in analysis

4.2 Multivariate Ferrocanide Watchlist Classification

A multivariate approach to classifying tanks as watchlist or non-watchlist tanks for various safety issues would complement the current classification approaches in the following ways:

1. The current tank classification rule that classifies tanks as watchlist or non-watchlist can be checked by considering the data multivariately.

2. The tanks that are classified differently by the two different classification methods can be considered as candidates for exclusion from or inclusion on a watchlist.

3. Correlations likely exist between the variables on the initial list for a safety issue. The list of variables may be reduced to help simplify the classification of tanks. Reducing the number of variables could also help in classifying tanks that lack a full data set.

All three of these potential complementary roles that multivariate analysis can play in this classification effort will be illustrated for the ferrocyanide example.

A classification method called 'Kth Nearest Neighbor' (KNN) was used as the multivariate classification method (KNN is described in Reference [28]). This method was selected because it is robust to varying distributions found in tank data, and the algorithm is easy to apply and understand. In KNN, the tank points are plotted in multidimensional space, as discussed in Section 3. In the ferrocyanide example, there are seven dimensions – the seven constituents of interest. Each tank in the study is identified as watchlist or non-watchlist, based on the current classification. When measurements for a new (unclassified) tank are available for the seven variables of interest, the plotted points representing the nearest neighboring classified tanks can be found. If the majority of these tanks have non-watchlist classifications, then the new tank is classified as a non-watchlist tank as well. If the majority of the neighboring tanks have watchlist classifications, then the new tank is classified as a watchlist tank. The number of neighbors to be considered for this analysis can be specified by the user. For the Ferrocyanide Safety issue example, the KNN method was tried using $K = 1, 2, \ldots, 7$. The best results (lowest classification error) were obtained when $K$ was equal to 5 (i.e., the five closest neighbors were polled for class membership) and this was the value of $K$ that was used in the results that follow.

To check the performance of the KNN method on the ferrocyanide set of tank data, a cross-validation exercise was conducted. In this exercise, each tank is left out in turn, and
reclassified by the remaining tanks. Tanks classified differently by the two methods from a cross-validation exercise can be considered as candidates for inclusion on or exclusion from a tank safety issue watchlist. Table 5 contains the KNN method cross-validation results for the Ferrocyanide Safety Issue. This method classified 30 out of the 32 non-watchlist tanks the same as the current method. It classified 7 out of 9 watchlist tanks the same as the current method. Overall, the multivariate and the current classification approaches classified 90% of the tanks the same.

Table 5: Comparison of Classification Methods Results

<table>
<thead>
<tr>
<th>Current Class.</th>
<th>Multivariate Class.</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Watchlist</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Watchlist</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Watchlist</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Totals</td>
<td>32</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 6 presents the four tanks where the two approaches classified differently. This table lists each tanks name, its class membership based on the current classification and the multivariate classification, and the number out of the five closest neighboring tanks (in the KNN multivariate analysis) that are currently classified as non-watchlist and watchlist tanks. Tanks BX-105 and BY-104 have several constituent concentrations that are close to the boundaries that separate the non-watchlist and watchlist tanks. These two tanks did not group with the ferrocyanide waste type tanks in Section 3. Tank C-105 is also near the boundary between the two classes except its constituent concentrations resemble a ferrocyanide watchlist tank much closely than the two tanks just mentioned. Tank C-105 also was a part of the ferrocyanide cluster in the general tank grouping study discussed in Section 3. For this reason, Tank C-105 could be a candidate for inclusion on the ferrocyanide watchlist. Tank T-107's four closest neighboring tanks are non-watchlist tanks. For this reason, Tank T-107 may be a good candidate for exclusion from the ferrocyanide watchlist. Referring to the clustering results from Section 3, Tank T-107 ended up in the middle of the 1C/2C waste cluster rather than the ferrocyanide cluster.

Table 6: Tanks Classified Differently by Two Methods

<table>
<thead>
<tr>
<th>Tank</th>
<th>Current Classification</th>
<th>Multivariate Classification</th>
<th># Neighbors on watchlist</th>
<th># Neighbors on non-watchlist</th>
</tr>
</thead>
<tbody>
<tr>
<td>BX-105</td>
<td>non-watchlist</td>
<td>watchlist</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>C-105</td>
<td>non-watchlist</td>
<td>watchlist</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>BY-104</td>
<td>watchlist</td>
<td>non-watchlist</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>T-107</td>
<td>watchlist</td>
<td>non-watchlist</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
A variable selection technique was then applied to look at all possible subsets of the seven constituents used in this exercise and find the subset of variables that contributes most to the proper classification of tanks for the ferrocyanide issue. The results showed that nickel, calcium, and cesium-137 contributed the most to this classification effort. The KNN classification algorithm was run again using only these three variables. The cross-validation results were identical to the results in Table 5 and Table 6. This result shows that these three constituents primarily drive the classification effort. Now that the number of variables has been reduced to three, a visual interpretation of the separation between classes can be obtained.

Figure 3 is a two-dimensional representation of the three-dimensional plot of the data. The ferrocyanide watchlist tanks are denoted by ‘x’ and the non-watchlist tanks are denoted by ‘o’. In the figure, calcium is the dimension or axis that is going up. The cesium-137 axis is coming out of the page and down slightly to the right. The nickel axis is going into the page and up to the right slightly. Note that most of the ferrocyanide watchlist tanks are in the upper right part of the figure, with the exception of Tank T-107, which is down near the origin (in this plot the origin is the mean of the three variables). The two non-watchlist tanks that are closest to the watchlist tanks are Tanks C-105 and BX-105, which is consistent with the results reported earlier.

This multivariate approach to tank safety watchlist classification cannot initially replace the current methods, partly because an initial set of tanks must be previously classified into non-watchlist and watchlist classes before using of multivariate technique. Once this has been done, however, the multivariate approach can be used to refine the classification rules for future tanks.

5 Conclusions and Recommendations

Significant progress was made on the Tank Grouping Study in FY96. Twenty-six additional single-shell tanks were added to the study to bring the total number of tanks included in the study up to 41. Several technical enhancements were also made in FY96. The effort has progressed to the point that results from this effort can be used to conserve and better allocate the limited sampling resources. As technical enhancements and tank data continues to be added to this grouping study in the future, we expect that the results will continue to support the effort to effectively allocate the available sampling resources.

In the general tank grouping study, 41 tanks were grouped based on concentrations of 14 chemical and radiological constituents. There were six groups of tanks identified in the sampling data based on the 14 variables. Each of these six sampling data tank groups matched up reasonably well (more than 50% overlap) with a historical tank group. Two related groups contained saltcake tanks (B1/T1 and S1/S2 saltcakes). The four remaining groups, all containing sludge tanks consisted of 224 waste, FeCN waste, 1C/2C waste, and R1/CWP wastes. The 224 waste tanks group (comprising Tanks B-201, B-202, B-203, B-204, and T-111) showed excellent agreement with the historical grouping. This finding, coupled with others discussed in the results section of this report, indicate that there is a good potential for saving sampling costs on the remaining unsampled 224 waste type tanks.
Figure 3: Nickel, Calcium and Cesium-137 Tank Means 3D Plot
(Tanks T-201, T-202, T-203 and T-204). The data from the sampled 224 waste tanks may be pooled and used to help make inferences regarding the unsampled tanks.

The multivariate KNN classification method was proposed to complement the current classification method for safety issue tank classification. The misclassified tanks from the multivariate classification approach are considered as candidates for watchlist additions or exclusions. From the ferrocyanide illustrative example, Tank T-107 was identified as a candidate for exclusion from the ferrocyanide watch list. Tank C-105 was identified as a candidate for ferrocyanide watchlist inclusion.

We recommend the following:

1. Using the results of this study and others, investigate the feasibility of using the present 224 waste tank group data to make inferences regarding the unsampled tanks in this group. Additional investigations would include cross-tank ANOVA models and cluster analysis of individual analytical results for the sampled 224 waste tanks.

2. Continue the general tank grouping task in the following areas:

   (a) Conduct cross-tank grouping studies using segment and sub-segment sample results rather than tank means. This study would show how robust present tank groupings are when considering tank spatial and analytical variability.

   (b) Continue cross-tank grouping studies which group tanks using tank sampling means, Historical Tank Content Estimates, and Hanford Engineering Estimates.

   (c) Make technical enhancements to data transformations (e.g., Empirical Cumulative Density Function Approach) and clustering methods (e.g., optimization clustering methods in Reference [12]).

   (d) Add more tanks to the cluster analyses as sampling data becomes available.

   (e) Search for additional cost savings associated with other “well-defined” tank groups.

3. Apply the multivariate classification methods demonstrated in this paper to other tank safety issues such as Flammable Gas.
6 References

References


A Tank Grouping Variables

The variables for the general tank grouping study were selected to discriminate between groups of tanks with similar waste properties. These variables should be measured on the tank level rather than on a subtank or multiple tank level (i.e., an individual tank is considered the sampling unit). To meet these needs, mean concentration estimates from sampling data and HTCEs were obtained for various chemical and radiological constituents identified below. Spatial variability estimates were also obtained when available to see how tanks group based on the spatial distribution of the waste.

The constituents were selected for the FY96 tank grouping effort for SSTs based on the following criteria:

1. Waste Type Discriminating Analytes - These analytes are found at high or low levels in specific waste types and therefore can be used to discriminate among waste types. The major waste components for various Hanford chemical processes are listed in Reference [6] and were used to identify the waste type discriminating analytes.

2. Large Inventory Analytes - These constituents have tank inventories of more than 100,000 moles across all Hanford tanks. These constituents were identified using the Hanford Tank Waste Chemical Inventory from Reference [6].

3. Program Interest Analytes - Each of these analytes is important to several programs such as safety, pretreatment, treatment, disposal, etc. This information was obtained from a matrix which identifies program needs, assembled by NG Colton of PNNL.

Some constituents satisfy more than one of these criteria. In fact, some constituents satisfy all three criteria. Table 7 contains the proposed list of constituents for the grouping study for single-shell tanks along with the criteria each constituent meets. This general tank grouping study variable list was reviewed by WHC and PNNL technical staff.

The waste type discriminators identified in Table 7 are used as the fourteen variables for the general tank grouping study described in Section 3. The variable list for the ferrocyanide safety issue tank grouping study is described in Section 4.
Table 7: FY96 Analytes for SSTs

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Justification</th>
<th>Large Inventory</th>
<th>Waste Type Discrim.</th>
<th># Interested Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aluminum</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>5</td>
</tr>
<tr>
<td>Americium-241</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Bismuth</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>2</td>
</tr>
<tr>
<td>Calcium</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>4</td>
</tr>
<tr>
<td>Cesium-137</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>5</td>
</tr>
<tr>
<td>Chloride</td>
<td></td>
<td>✓</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Chromium</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>4</td>
</tr>
<tr>
<td>Fluoride</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>5</td>
</tr>
<tr>
<td>Iron</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>4</td>
</tr>
<tr>
<td>Lanthanum</td>
<td></td>
<td></td>
<td>✓</td>
<td>2</td>
</tr>
<tr>
<td>Lead</td>
<td></td>
<td>✓</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Manganese</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>3</td>
</tr>
<tr>
<td>Nickel</td>
<td></td>
<td></td>
<td>✓</td>
<td>3</td>
</tr>
<tr>
<td>Nitrate</td>
<td></td>
<td>✓</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Nitrite</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>5</td>
</tr>
<tr>
<td>Phosphorus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>4</td>
</tr>
<tr>
<td>Plutonium-239/240</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Silicon</td>
<td></td>
<td>✓</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Sodium</td>
<td></td>
<td>✓</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Strontium-90</td>
<td></td>
<td></td>
<td>✓</td>
<td>5</td>
</tr>
<tr>
<td>Sulfate</td>
<td></td>
<td>✓</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Uranium</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>4</td>
</tr>
<tr>
<td>TIC (Carbonate)</td>
<td></td>
<td>✓</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>TOC</td>
<td></td>
<td>✓</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Water</td>
<td></td>
<td>✓</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>
B  Data Preparation

Once the sampling data was gathered, it had to be prepared for statistical analysis. The following two data preparation steps were taken before analysis:

1. **Missing Data Values** – At the conclusion of the data gathering effort, it was noted that there were several “holes” in the dataset (i.e., missing values). To do the tank grouping analyses, these missing values had to be estimated.

2. **Data Transformations** – Due to varying concentration ranges of the waste constituents and the asymmetric distributions of the constituents, the data needed to be transformed so that meaningful results could be obtained from the tank grouping analysis.

The approach and steps taken to prepare the data for analysis are described in this Appendix.

B.1  Handling Missing Values

Two types of missing values were encountered in the tank data. Some values are missing because the results are below detection limits and others are missing because chemical or radiological analysis of the samples was not completed, due to specific analysis requirements for the tank.

Several methods have been proposed to deal with censored data (results that are below the instrument detection limits). One of the most popular is a maximum likelihood approach, but the approach is difficult for non-statisticians to understand. For this reason, a simpler method was used to estimate these missing values; namely to use the minimum detection limit across tanks for each censored variable. The detection limit for a given tank and variable was not used because in some cases the detection limit for a given tank and variable is at a higher concentration than some quantified results for the same variable in a different tank.

An Expectation-Maximization (EM) Algorithm discussed in Reference [35] was used to estimate the missing values that were due simply to no laboratory analysis being performed. In this approach, one assumes that the missing values are “missing at random”. The EM algorithm follows these steps:

1. Replace the missing values with the median for the particular variable.

2. Next, for each variable \(j\) obtain the least squares fit of

\[
\hat{y}_{ij} = b_0 + b_1 y_{i1} + \cdots + b_{j-1} y_{i(j-1)} + b_{j+1} y_{i(j+1)} + \cdots + b_p y_{ip}.
\]  

3. Replace the missing value estimate from the last iteration with \(\hat{y}_{ij}\).

4. If the convergence criteria is met, STOP; else GOTO step 2 and continue.
B.2 Transformations

The analyte concentration ranges for this study varied widely, as can be seen in Figure 4 left-side. If cluster analysis were run on this data, those constituents with the greatest concentration range would dominate in the grouping of tanks. We wanted each constituent to have "equal weight" in the cluster analysis. For this reason, the data was normalized for each constituent $j$ with the following transformation:

$$
\tilde{y}_{ij} = \frac{y_{ij} - \bar{y}_j}{s_j}
$$

(2)

where

$y_{ij}$ is the concentration estimate for the $i^{th}$ tank and the $j^{th}$ variable,

$\bar{y}_j$ is the mean concentration estimate across all tanks for the $j^{th}$ variable, and

$s_j$ is the standard deviation for the $j^{th}$ variable.

Figure 4 middle shows the normalized data. Note that the variables have much more similar ranges than the raw data (Figure 4 left-side). Each variables distribution is quite asymmetric (skewed positively), however. The hierarchical clustering techniques used to group the tank data are adversely affected by skewed variable distributions particularly the complete linkage algorithm (see Reference [37]). For this reason, the following transformation was made on the data to deal with the asymmetry and the differing ranges of the constituents:

$$
\tilde{x}_{ij} = \log_{10}(y_{ij})
$$

(3)

$$
\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\tilde{s}_j}
$$

(4)

where

$x_{ij}$ is the log to the base 10 of $y_{ij},$

$\bar{x}_j$ is the mean of the log transformed data for the $j^{th}$ variable, and

$\tilde{s}_j$ is the standard deviation of the log transformed data for variable $j.$

This second data transformation is shown on the right side of Figure 4. Note that the transformed data now have similar ranges and their distributions are more symmetric than before.

---

This plot contains boxplots for each variable. The three vertical lines in each boxplot represent the first quartile, the median, and the third quartile. The whiskers coming out of the box represent the minimum and maximum of the data.
Figure 4: Data Transformations
C  Boxplots of Tank Means by Cluster

This appendix contains boxplots of each tank group for all analytes considered in this study. The ends of the boxes are the 25th and 75th quantiles. The line within the boxes is the median of the data and the two whiskers coming out the end of the boxplots mark the minimum and the maximum tank values for the given group and analyte.

Boxplots are presented for the 25 analytes on the general tank grouping variable list and also for cyanide, strontium and phosphate, since the data was readily available for these analytes.
Figure 6: Boxplots of Tank Means by Cluster
Figure 6: Boxplots of Tank Means by Cluster (continued)
Figure 6: Boxplots of Tank Means by Cluster (continued)
Figure 6: Boxplots of Tank Means by Cluster (continued)
Figure 6: Boxplots of Tank Means by Cluster (continued)
D  ‘224 Waste’ Cluster Confidence Interval Plots

This appendix contains 90% confidence intervals on the mean for Tanks B-201, B-202, B-203, B-204 and T-111. The solid circles in the plots represent the estimated means. The arrow points mark the lower and upper 90% confidence limits for the means.

Confidence interval plots are made for all variables on the general tank grouping list with the exception of americium-241 and sulfate since the estimates of uncertainty for these constituents were not readily available.
Figure 7: Confidence Interval Plots for 224 Waste Cluster
Figure 7: Confidence Interval Plots for 224 Waste Cluster (continued)
Figure 7: Confidence Interval Plots for 224 Waste Cluster (continued)
Figure 7: Confidence Interval Plots for 224 Waste Cluster (continued)