CONF-9506241--1

LA-UR- 95-2438

Title: DATA MANAGEMENT AND STATISTICAL ANALYSIS FOR ENVIRONMENTAL ASSESSMENT

Author(s): J. R. Wendelberger, T. I. McVittie

Submitted to: Interface '95/Computing Science & Statistics
Pittsburgh, PA
June 21-24, 1995

Los Alamos
NATIONAL LABORATORY

DISCLAIMER

LA-UR- 95-2438

DATA MANAGEMENT AND STATISTICAL ANALYSIS FOR ENVIRONMENTAL ASSESSMENT

Joanne R. Wendelberger and Thomas I. McVittie
Los Alamos National Laboratory, MS-F600, Los Alamos, NM 87545

1. INTRODUCTION

Data management and statistical analysis for environmental assessment are important issues on the interface of computer science and statistics. Data collection for environmental decision making can generate large quantities of various types of data. This data may need to be shared by many different individuals including analysts, managers, regulators and interested members of the public. A database/GIS system developed at Los Alamos National Laboratory provides efficient data storage as well as visualization tools which may be integrated into the data analysis process.

2. BACKGROUND

The Environmental Restoration Program at Los Alamos National Laboratory has been charged with the mission of investigating over 2000 areas where contamination may have occurred from past activities associated with laboratory operations. The sites represent a large and varied group of contaminants of potential concern including organic and inorganic chemicals as well as radionuclides.

Many different types of questions are raised during the process of conducting environmental assessments. Does a site contain chemicals of potential concern? What is the spatial distribution of chemicals of potential concern? Does a site pose a risk to human health or the environment? What type of remediation or site controls should be implemented? Has remediation been successful? In order to answer these questions, existing information on a site is compiled, and if further information is needed, a sampling plan is designed to collect additional data. The sample collection process generates many different types of information. In addition to laboratory measurements made on soil, air or water samples, data on sample locations, sample identifiers, regulatory values, field screening measurements, geological profiles, and other sample information must also be stored and maintained.

3. DATABASE DESCRIPTION

The data collected by the Los Alamos National Laboratory Environmental Restoration Program is stored in a database system known as the Facility for Information Management, Analysis and Display (FIMAD). As described by system documentation, "FIMAD provides the expertise, visualization tools, database support and systems necessary to support the large quantities of spatial and tabular data collected and analyzed as part of the Environmental Restoration Program at Los Alamos National Laboratory."

FIMAD provides both a database and a graphical information system (GIS). FIMAD employs a relational database consisting of approximately 60 tables. As of May, 1995, the database contained information on over 1.6 million analytical records, plus additional information collected during the sample collection and analysis process. The GIS system provides presentation tools and may be used to produce maps for viewing location information associated with the various sites and sampling activities. This geographic information may be augmented with additional quantitative information associated with samples taken at the site.

Several different types of data associated with the environmental data collection process are collected. Approximately 90% of the data currently stored in FIMAD consists of information pertaining to the results of laboratory chemical analyses. Variables containing analytical information include the analytical value, the units, the type of analysis, the identifier for the sample, the sample matrix and comments.

An increasingly important family of data is obtained directly from screening measurements made in the field. These quick and relatively inexpensive methods provide rapid turnaround and can be used to guide the sampling process as well as to supplement the final analytical values for samples sent to the laboratories.

Sample information includes information about the physical collection of the sample, such as sample location and depth, the technique used to collect the sample and how samples were combined in the case of composite samples.

Another type of data stored in the system is regulatory data. Regulatory values are used in conducting environmental assessments. They include screening action levels which are used as reference values for screening assessments and preliminary remediation goals which are used in deciding whether a remediation should be implemented and what type of remediation techniques may be considered.

Geological information is recorded and stored in the system. The geological record provides important site specific information about the geological features of the site. This information is useful in the interpretation of the analytical data, in the assessment of whether remediation is required, and in the selection of appropriate remediation measures.

Data for the GIS system is stored as an identifier with an x and y coordinate. The GIS system uses this information to generate a variety of map features including site boundaries, buildings, utilities, sampling points, roads, fences and other features of interest.

Photographic images may be stored in FIMAD to provide a visual record of sites and sampling activities. The photographs provide a historical record and also provide additional site information which can be helpful for examining physical characteristics of the site and for selection and transport of appropriate equipment.

4. DATA STORAGE ISSUES

Data Organization

With the diversity of data types used by the Environmental Restoration Program, careful organization of the data is essential. A relational database was designed to handle the large quantities of interrelated data in an efficient manner. All data is stored in tables with each entry in a given table having identifiers which can be used to relate the entries to data stored in other tables. The relational database technology allows the user the ability to combine related tables as needed without having to repetitively store information which may be related to several different entries in a given table. For example, hundreds of analytical records may be associated with a particular soil sample. Using a relational database, the sample location and other information associated with how the sample was collected may be stored as a single entry in the location table instead of repeating this information for every individual analysis in the analytical information table. Approximately sixty different tables are used in the database, with the bulk of the information stored in the table of analytical information.

Data Entry

In the early stages of the development of the FIMAD system, data entry was tedious and error prone. Although analytical information was channeled through a single group responsible for chemical analysis data, other types of data were generated and stored individually by the teams performing the sample collection activities. No standard format was used to store information. In some cases, information about the samples was simply written down by hand in logbooks and had to be entered manually at a later date if the information was deemed important. With all the different systems for data collection, different personnel working at different sites, changing personnel at specific sites and reorganization of the environmental restoration program, a uniform system for data entry became essential. Protocols were developed for automatic generation of barcode labels with identification numbers for samples. A system was developed for immediate electronic entry of on-site information. Information coming in from the sites is now uploaded electronically in specified formats to update the relational database.

Data Verification

Once the data is uploaded to the main system, verification checks are performed to examine the integrity of the data being loaded into the system. Data with obvious problems such as duplicate records, values beyond reasonable ranges, or values with unusual units are rejected and subjected to further scrutiny. This verification stage is intended to perform routine simple checks to eliminate common problems which were identified by users during the initial usage of the system.

Data Validation

Following verification, individual site teams are responsible for performing data validation. Data validation provides a check on whether the data that was requested is what came back and whether the values obtained make sense in light of site-specific information. Data validation also includes routine quality assurance checks performed by analytical chemists prior to the release of analytical data into the database. The database is designed to track which data has been validated and may be released to individuals beyond the site investigation team.

Data Updating

The database must be updated as new information becomes available. Data comes from a variety of sources and may come in a few records at a time as chemical analysis results arrive from the analytical laboratories. New data is added to the system as it becomes available. In some cases, errors are detected or changes occur which require updating of individual records. In these cases, the original records are supplemented with a replacement record which provides an identifier to reference the user back to the original record. This system allows the analyst to keep track of what data is available at different time points, allowing a return to earlier versions of the data even after updating has occurred. This updating feature is particularly important for regulatory values which may change over time. For example, Screening Action Levels are risk-based decision values which are calculated using toxicological information from an EPA database. As additional toxicological studies are conducted, the toxicology database is updated and screening action levels must be recalculated.

Data Queries

In order to access data from FIMAD, the database must be queried to obtain the desired information from one or more tables. Explicit queries may be given by the user using SQL data query commands. However, this method is rather cumbersome for the end user who will generally not be familiar with the contents of all of the data tables. Generic data views have been created which link information from different tables which are commonly needed by database users. These views significantly reduce the effort required to access typical information needed for environmental assessment by the site teams. A menu driven interface to the system was developed to provide some simple query capabilities and limited analysis. However, problems were encountered with network limitations and the heterogeneous hardware and software found in the user community. Current philosophy is to treat the database more as a repository for data which can be accessed and downloaded to individual users' workstations for analysis and display. Various commercial software packages are available to facilitate the browsing and export of data.

5. STATISTICAL ANALYSIS

Data Summary

Sampling activities for a single site can result in chemical analyses for hundreds of different analytes in a single sample and thousands of analytical values for the overall site. Data summary is an important first step in the process of evaluating these large datasets. Tables of summary statistics are prepared which include the minimum and maximum values and sample size for each analyte. Frequently values are below the detection limits for the chemical analysis methods. Values below detection limits are entered in the database as the detection limits and are flagged as being less than the detection limits. Separate summary statistics are computed and tabulated for values above and below detection limits for each analyte.

Screening Assessment

The summary tables are used to perform a preliminary screening assessment. For each analyte, the maximum values above and below the detection limits are compared to risk-based screening action levels. This process categorizes the analytes into different groups depending upon where the maximum values are relative to the screening action levels. In some cases, screening action levels may be unavailable due to a lack of toxicological data. In other cases, because of below detection limit data, it may be impossible to determine whether the screening action level is exceeded. These indeterminate situations require additional study using other analysis protocols such as comparison to levels for related chemicals, examination of the frequency of values above detection limits and incorporation of other external information. The screening assessment leads to a list of analytes which have values above screening action levels or which can not be removed from the list of contaminants of potential concern using other investigative techniques. For some analytes, comparisons to appropriate background concentration distributions are required to account for naturally occurring levels. Additional analyses are conducted to examine multiple constituents in cases where analyses have been conducted for several related chemicals or radionuclides.

Risk Assessment

For sites where screening assessment identifies the presence of contaminants of potential concern, risk assessment is performed to determine whether any further action is required. The complexity of the risk assessment varies widely and depends on specific site information. The risk assessment models require estimates of several different site specific parameters. Among the key inputs are the mean concentrations of the contaminants averaged over areas called exposure units which depend on the potential future land use of the site. Typically, a 95maximum value in the risk calculations. Determining appropriate estimates of the mean and 95limit can be a challenge in some cases, particularly when spatial correlation and colocated contaminants are present.

Values Below Detection Limits

As mentioned above, many of the analytical values are below detection limits. These values pose additional challenges for data storage, analysis and display. Two general approaches which may be followed are to replace the below detection limit values by appropriate proxy values or to use analysis methods which can handle censored data.

Spatial Analysis

Spatial analysis techniques are useful for determining the spatial distribution of chemicals of potential concern and for identifying regions for additional investigation. Care must be taken to avoid overinterpretation of contours from these surfaces when data is sparse. These techniques are most useful when available data is dense and when auxiliary information about the site can be used to augment the analysis results. Figure 1 shows contours from a kriging model used to examine a site where residual contamination remained after a previous cleanup activity. The locations of the plumes identified by the model correspond to locations of discharges from former operations.

6. VISUALIZATION TOOLS

FIMAD's database and GIS capabilities provide the opportunity for the use of visualization tools as an integral part of the analysis process.

Map Displays

Extensive mapping data is available which includes laboratory technical areas as well as nearby townsite areas. Digital orthophotos and elevation data are also available. The GIS system can extract a desired region and generate a variety of map displays. Site maps generated by FIMAD are widely use by the site teams throughout their investigations. Custom maps can be drawn which show site boundaries, current and former building locations, discharge outfalls, disposal areas, utilities, roads and other site features. These maps provide an important foundation for the planning and implementation of sample collection and analysis.

Data Retrieval

Once analytical results are obtained, these values may be accessed from the database and displayed along with the GIS data. Figure 2 illustrates the way data from the system can be used with commercial software, Arcview2, to view data at the desktop of the investigator. Data is retrieved from the database and symbolized. The investigator can select an area of investigation, pull up a table of sample values, and display the associated sample identification numbers on the map. Outputs are suitable for reports and may be exported in Adobe Illustrator format.

Classification

Figure 3 show a site map which has been augmented with colored dots to show gross alpha field screening values categorized into different groups. Gross alpha field readings were retrieved from the database and displayed using user defined color coding of intervals. Clicking on any of the sample points on the screen will display underlying data such as sample number, location and the value of the reading obtained. The elevation contours show how the values change as the samples go down the drainage and indicate higher values below the level of clean fill which was used during an earler remediation. The figure also contains an aerial photograph of the area indicating where the site lies in relation to nearby townsite landmarks.

Gridding and Classification of Counts

Figure 4 displays frequency data after gridding and classification into ranges to show the areas where shrapnel fragments were found during a recent shrapnel cleanup operation. The intensity of the color shaded squares vary with the frequency of fragments found within the local region.

Comparison to Reference Values

Figure 5 shows a site map augmented with spikes to show sample values compared to reference values. The height of the spikes show the magnitude of the analytical values. Different colors are used to indicate portions of the spikes which are above and below the relevant screening action level.

Isopleths

Isopleths, such as those shown in Figure 6, can be used to generate lines of constant value using a simple routine for data retrieved from the database. Contours based on more complex models, such as kriging models, may be generated external to the system and fed back into the system for display. As noted earlier, care must be taken to discourage indiscriminate use of contouring routines for sparse data.

Multivariate Display

In some cases, physical aspects of the former operations and contaminant dispersal processes lead to colocated contaminants. Multivariate pie displays of the type shown in Figure 7 provide a mechanism for viewing the values of related contaminants on the same map figure. Data for the different contaminants is retrieved form the database and displayed on the map using pie shapes. Each slice position corresponds to a particular contaminant. Different colors are used to indicate the magnitude of the values.

Draping of Orthophotos

Figure 8 shows a 2 1/2 dimensional representation of a site which was generated using an aerial photograph obtained during a 1991 flyover of the laboratory. A digital elevation model was generated for the entire laboratory on a one foot grid. The photo was "draped" over the digital elevation model. Then other features were plotted on top. These views are useful for placing the data in the context of the land features and are helpful for planning equipment movement and sample collection activities.

7. CONCLUDING REMARKS

FIMAD is a living database and GIS system. The system has changed and developed over time to meet the needs of the Los Alamos National Laboratory Restoration Program. The system provides a repository for data which may be accessed by different individuals for different purposes. The database structure is driven by the large amount and varied types of data required for environmental assessment. The integration of the database with the GIS system provides the foundation for powerful visualization and analysis capabilities. Further information on FIMAD is available on the World-wide Web at the following URL address: http://ees-www.lanl.gov/EES5/fimad/fimad_main.html.

Figure 1

TA-10, Central Area, Depth=17ft, beta

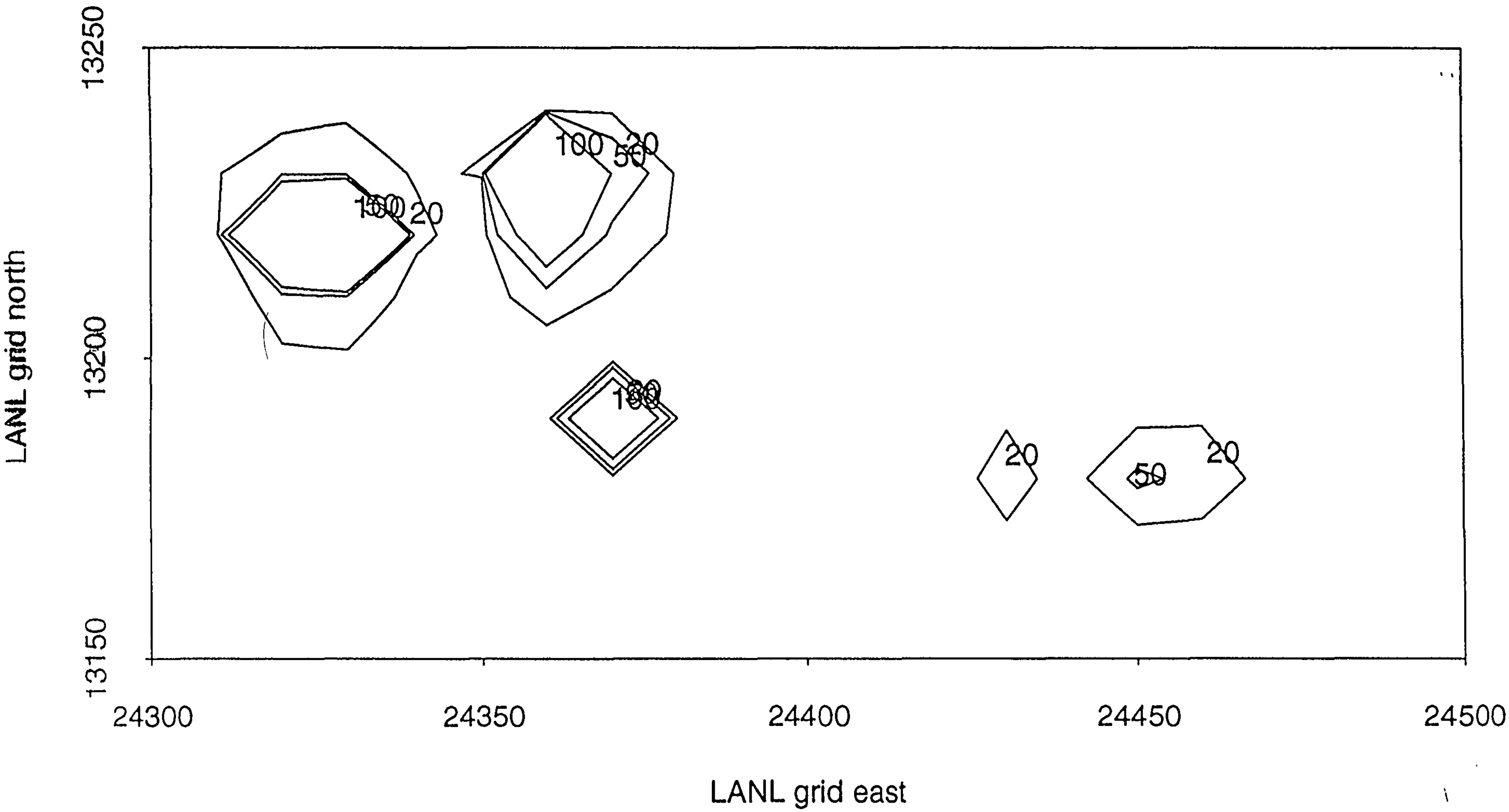
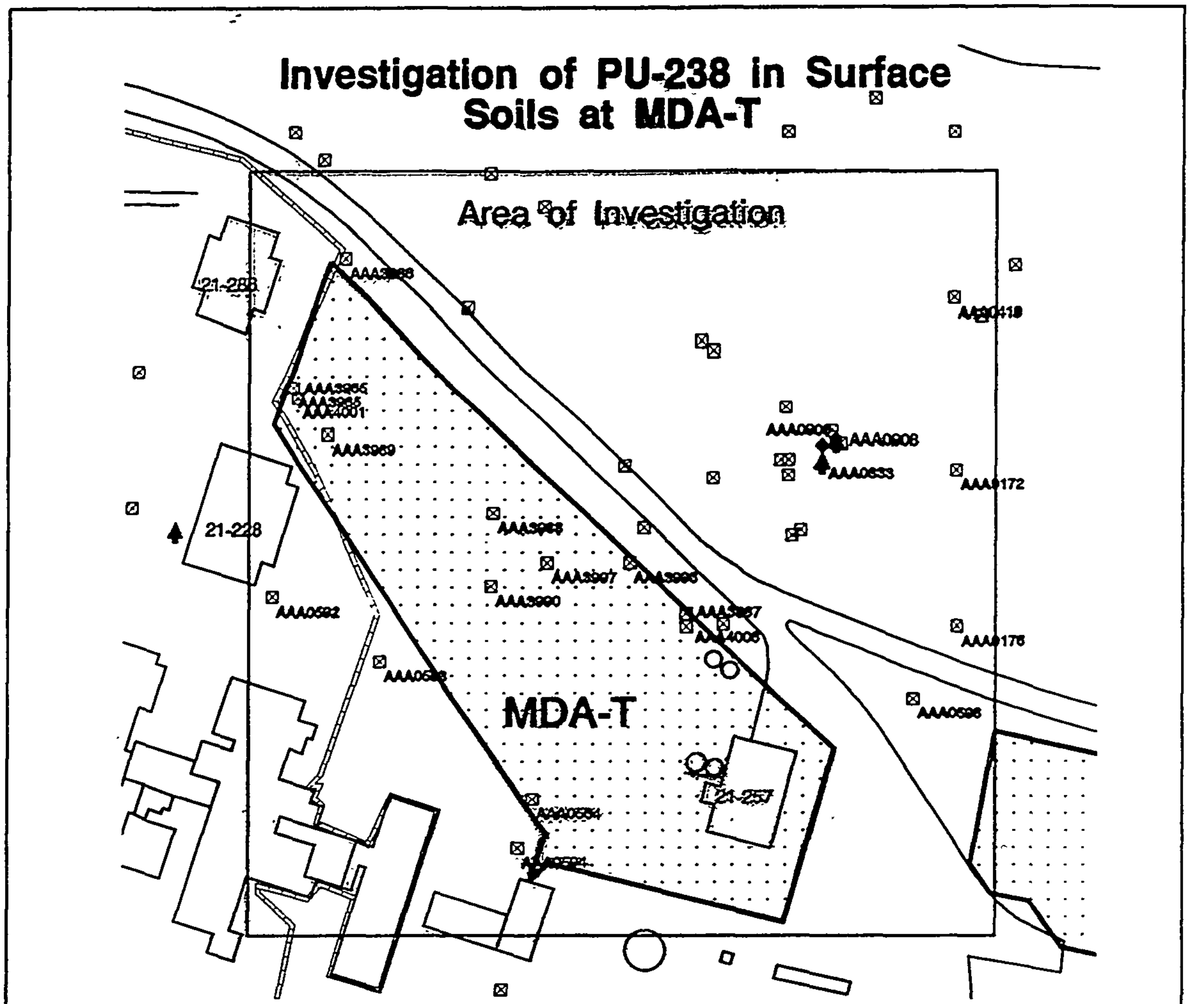
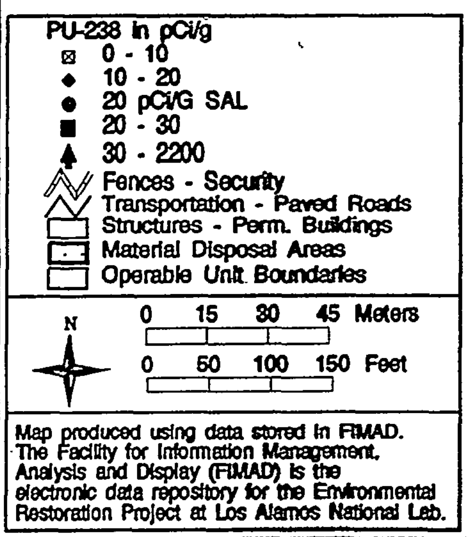


Figure 2

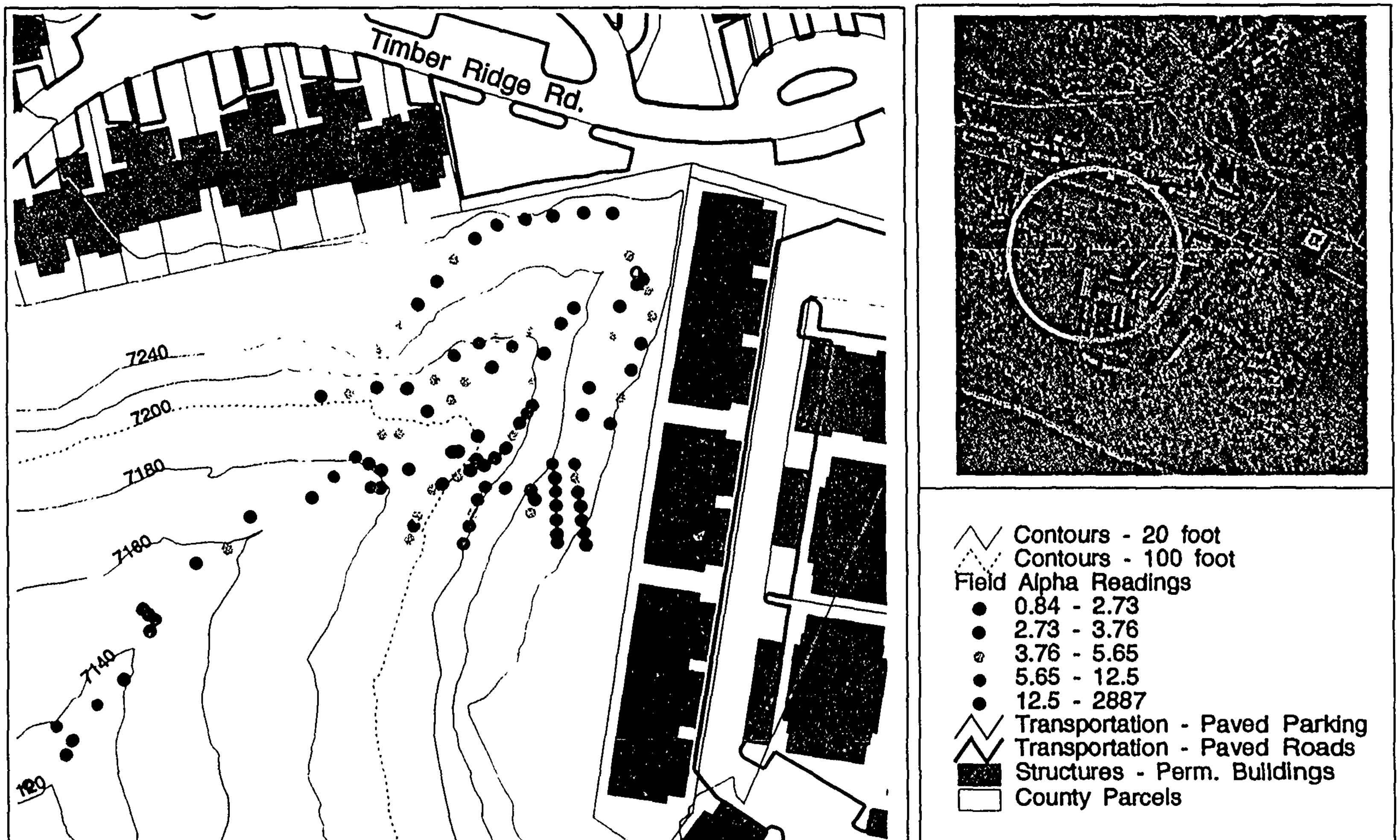




[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Start Depth	End Depth	[illegible]
AAA0564	SS	6.97000	PCI/G	PU-238	0.000	1.000	in.
AAA3987	SS	6.85100	PCI/G	PU-238	0.000	6.000	in.
AAA4001	SS	6.62500	PCI/G	PU-238	0.000	6.000	in.
AAA4002	SS	6.53200	PCI/G	PU-238	0.000	6.000	in.
AAA3965	SS	3.63900	PCI/G	PU-238	0.000	6.000	in.
AAA0908	SS	2.80000	PCI/G	PU-238	0.000	6.000	in.
AAA4012	SS	1.75100	PCI/G	PU-238	0.000	6.000	in.
AAA0592	SS	0.87800	PCI/G	PU-238	0.000	6.000	in.
AAA4010	SS	0.87700	PCI/G	PU-238	0.000	6.000	in.
AAA4009	SS	0.65900	PCI/G	PU-238	0.000	6.000	in.
AAA0594	SS	0.62500	PCI/G	PU-238	0.000	1.000	in.
AAA0593	SS	0.52200	PCI/G	PU-238	0.000	1.000	in.
AAA4006	SS	0.51100	PCI/G	PU-238	0.000	6.000	in.
AAA0591	SS	0.47400	PCI/G	PU-238	0.000	1.000	in.
AAA3992	SS	0.42600	PCI/G	PU-238	0.000	6.000	in.
AAA4007	SS	0.38900	PCI/G	PU-238	0.000	6.000	in.
AAA4015	SS	0.27500	PCI/G	PU-238	0.000	6.000	in.
AAA3989	SS	0.27300	PCI/G	PU-238	0.000	6.000	in.
AAA3991	SS	0.25600	PCI/G	PU-238	0.000	6.000	in.
AAA0172	SS	0.23800	PCI/G	PU-238	0.000	1.000	in.

Figure 13.4 MDA-T

Figure 3



Field Alpha Readings Townsite - Los Alamos

Map produced using data stored in FIMAD. The Facility for Information Management, Analysis and Display (FIMAD) is the electronic data repository for the Environmental Restoration Project at Los Alamos National Lab.

Figure 4

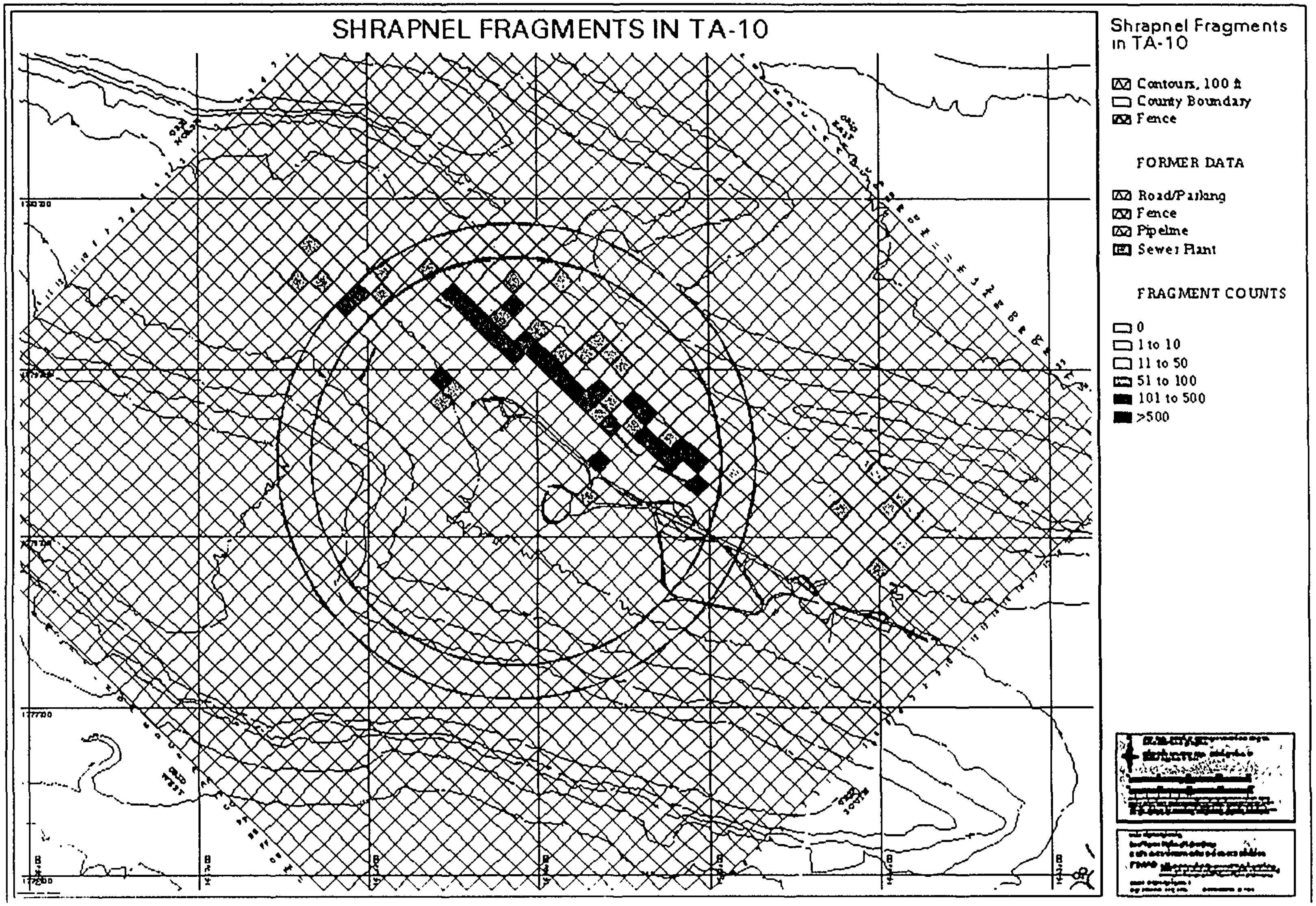
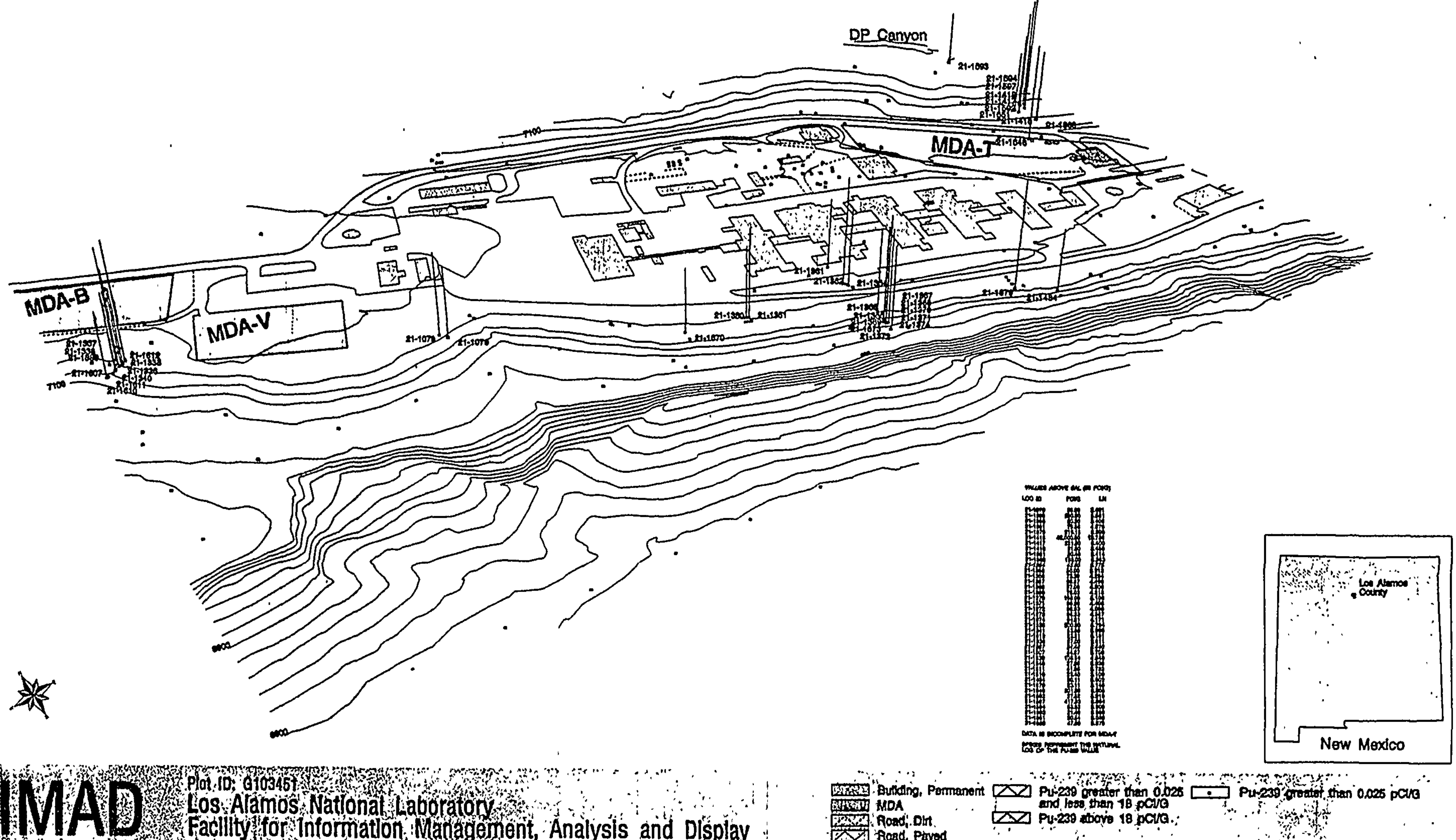

SHRAPNEL FRAGMENTS IN TA-10
Shrapnel Fragments in TA-10
Contours, 100 ft
County Boundary
Fence
FORMER DATA
Road/Parking
Fence
Pipeline
Sewer Plant
FRAGMENT COUNTS
0
1 to 10
11 to 50
51 to 100
101 to 500
>500

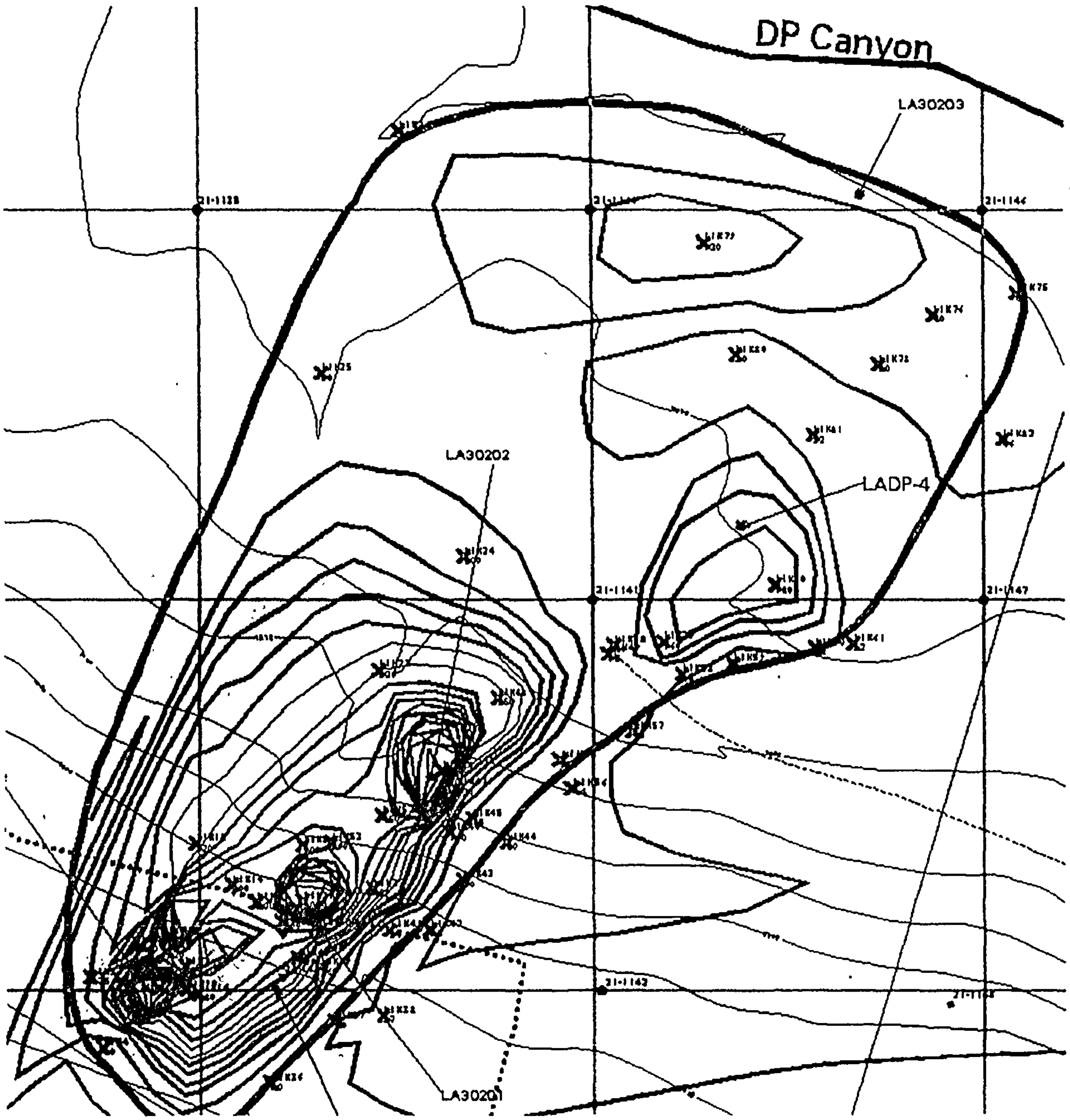
Figure 5

Sitewide Pu-239 Characterization at TA 21



FIMAD Plot ID: G103451
Los Alamos National Laboratory
Facility for Information Management, Analysis and Display

SURVEY RESULTS FROM 1 m ABOVE SURFACE (micro-rem)
OUTFALL SWMU 21-011(k)



Rad Survey Results
Outfall SWMU 21-01

LEGEND

- Boundaries, Operable Unit
- Boundaries, Technical Area
- Contours, 100 ft
- Contours, 10 ft
- Fence, Industrial
- SWMU Boundary, Temporary Fence
- Road, Dirt
- Road, Paved
- Road/Trail
- Sample Grid
- Storm Drain
- Water Line
- Material Disposal Areas
- Permanent Buildings

FIRST ROUND

- 0-1" Deposition-Layer Soil Sample (Grid)
- 0-6" Surface Soil Sample (Grid) and 0-1" Deposition-Layer Soil Sample (Grid)

SECOND ROUND

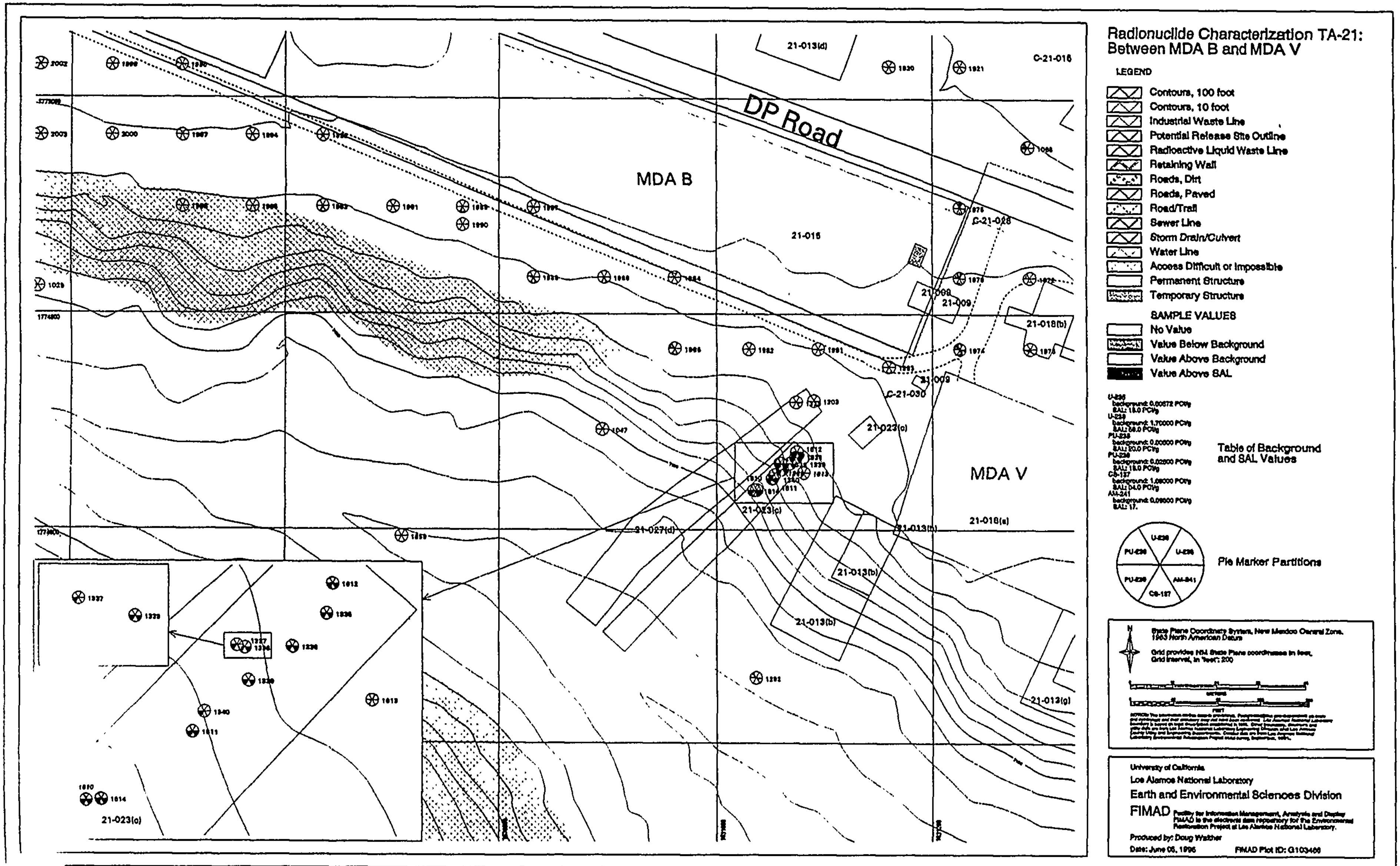
- 0-1" Deposition-Layer Soil Sample (Grid)
- 0-6" Surface Soil Sample (Grid) and 0-1" Deposition-Layer Soil Sample (Grid)

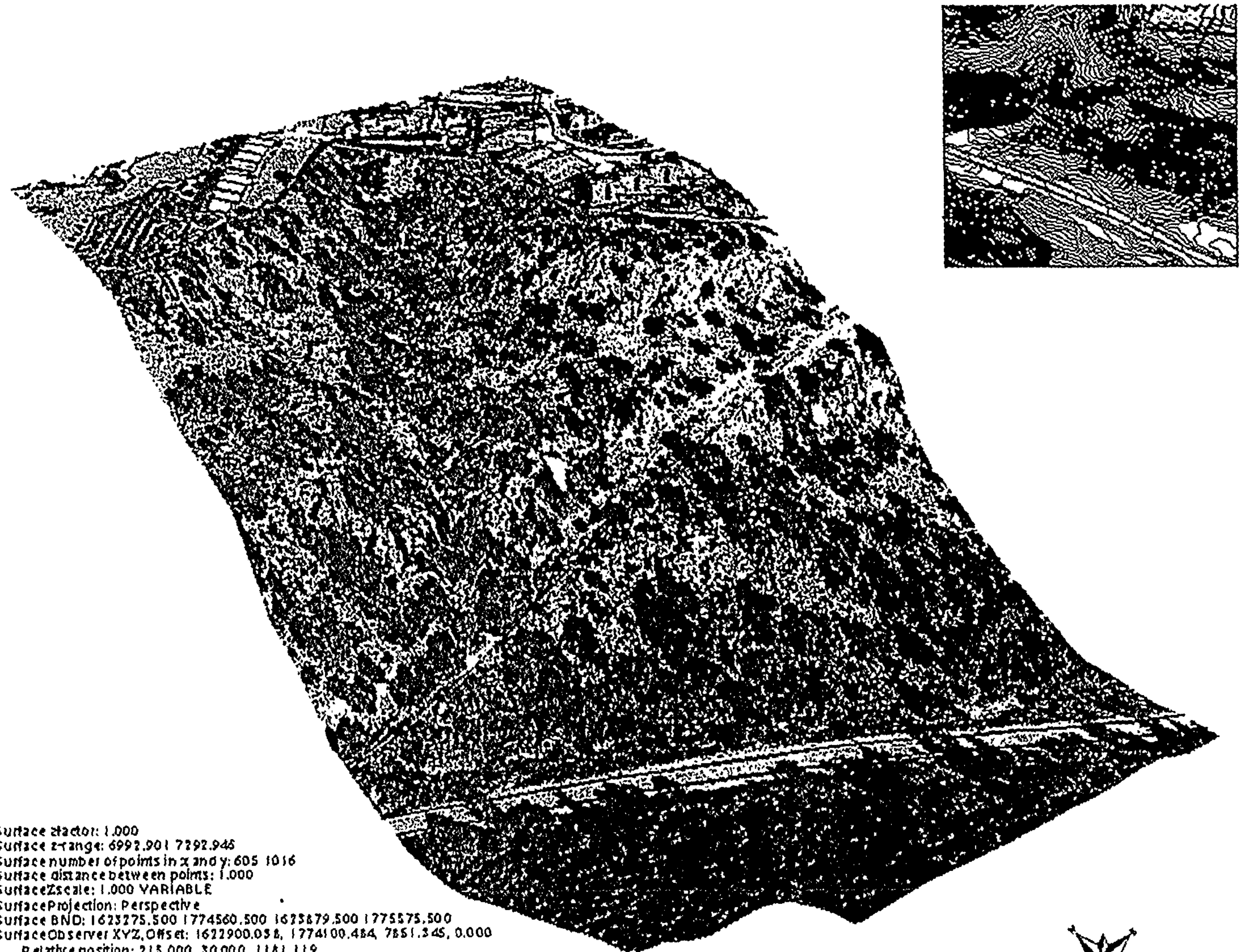
- DOE Environmental Problem #8 Sample Locations (approx. locations)
- Micro-R Samples
- Outfall Sample
- < 140 urem/hr
- < 240 urem/hr
- < 340 urem/hr
- < 440 urem/hr
- < 540 urem/hr
- < 640 urem/hr
- > = 640 urem/hr

Micro-R Sample Units are micro rem taken 1 mete the surface. Isogram of Micro-R readings created a bivariate quintic interpolation which assumes a smooth surface model. The interpolation uses onl

11Kxx = Rad Sample Location ID
\## = Instrument Reading in Micro-R/hour
21-xxxx = RFI Soil Sampling Location ID
21-xxx = SWMU ID

Figure 7


Radionuclide Characterization TA-21:
Between MDA B and MDA V
LEGEND
Contours, 100 foot
Contours, 10 foot
Industrial Waste Line
Potential Release Site Outline
Radioactive Liquid Waste Line
Retaining Wall
Roads, Dirt
Roads, Paved
Road/Trail
Sewer Line
Storm Drain/Culvert
Water Line
Access Difficult or Impossible
Permanent Structure
Temporary Structure
SAMPLE VALUES
No Value
Value Below Background
Value Above Background
Value Above SAL
Table of Background and SAL Values
Pie Marker Partitions
U-238
U-235
AM-241
CS-137
PU-239
PU-238
University of California
Los Alamos National Laboratory
Earth and Environmental Sciences Division
FIMAD
Produced by: Doug Walther
Date: June 05, 1996
FIMAD Plot ID: G103466
DP Road
MDA B
MDA V
21-013(d)
C-21-016
21-016
C-21-028
21-008
21-009
21-018(b)
C-21-030
21-022(c)
21-013(c)
21-027(d)
21-013(h)
21-018(a)
21-013(b)
21-013(g)
21-023(c)



Surface zfactor: 1.000
Surface z-range: 6992.901 7292.946
Surface number of points in x and y: 605 1016
Surface distance between points: 1.000
SurfaceZscale: 1.000 VARIABLE
SurfaceProjection: Perspective
Surface BND: 1623275.500 1774560.500 1623879.500 1775575.500
SurfaceObserver XYZ, Offset: 1622900.038, 1774100.484, 7851.345, 0.000
Relative position: 215.000, 30.000, 1181.119
SurfaceTarget XYZ, Offset: 1623577.500, 1775068.000, 7169.426, 0.000
Relative position: 35.000, -30.000, 1181.119
SurfaceRange: minimum 1.000, maximum 10000000000.000
SurfaceViewField angles: 21.556, 25.391, 15.954, 21.190, AUTO
SurfaceLimits: 0.000, 0.000, 24.000, 24.000, PAGE
SurfaceResolution. 10.255
Number of points in X and Y within SurfaceExtent. 60, 100