GIS

Introduction

The development of GIS applications to this type of research included of several projects.

Stage I. Initial work involved drawing relationships between OSCaR data and GIS shapefiles. OSCaR data was obtained and put into ArcView. Then links were established between counties, cities/towns, and zip code boundaries to map out the distribution of the Oregon cases included in the OSCaR Data. This stage in the work also allowed for exploration of such issues as developing statistical profiles using GIS and the development of a methods for including SIC data into this work, with the hope of relating the spatial distribution of Oregon cases to different types of businesses (as defined by SIC values), and in turn to the Toxic Release Inventory (TRI) and Environmental Clean-up Site Inventory (ECSI) data.

During the initial phase of this work, ECSI and TRI data were found to be similar if not identical (for web-related sources see APPENDIX C; note: changes in these basic website have since merged together and reformatted this ECSI and TRI data). For this reason, during the earliest stage of this research (including the 2000 data obtained prior to the grant proposal), ECSI and TRI data were analyzed separately, initially as ECSI and later as TRI. Therfore, both data sets are referred to in this study. This stage in the analysis had no effect on how the next part of this research was handled—the selection of Oregon Confirmed Release Inventory (CRI) sites to be analyzed.

Later in this phase of the work, an evaluation of the potential use for SIC data was carried out. Since SIC codes are part of a standard database designed for use by businesses, economists, various government agencies and political interest groups, it was realized that such a dataset might also be used to relate industry type to toxic release, in particular regarding the chemical profiles given by the site reports. This led the following research question to be posed: could SIC data be used to perform GIS-related Environmental Health-related work? In essence, this study suggests that the answer to this question is "yes' although considerable effort has to be made to further develop the method proposed by this study for integrating SIC data into future research projects involving chemical exposure-related risk analysis for the different forms of cancer.

Stage II. The second stage of this work involved detailed analyses of Confirmed Release Inventories for all sites in Oregon. These analyses consisted of a statistical profiling of the chemistry for each site based on chemical lists provided as part of the CRI reports made available to the general public on the web. This stage in the work resulted in the production of more than two dozen databases. These databases were of several forms:

1. A number of very brief databases were produced in order to facilitate the cross-linking of datasets from one large database to the next in GIS. Examples of such two column databases include those linking one type of SIC to a reclassification system designed for this work. Another two column database defined Chemical Abstract Series (CAS) numbers for given chemical names (In the end, this database was not effective in cross-linking due to the

inconsistent use of dashes (-) in the CAS numbers, and the inconsistency in which chemical names use commas and numerical prefixes (i.e. the use of "1,2,3-trichloroethylene" versus "trichlorethylene, 1,2,3-".))

2. Lengthy databases defining important aspects of the spatial items being studied. In particular, SIC data was extensive, and entailed redeveloping the data to include a detailed review of shared chemical features (i.e. the similarities in chemistry that exist between certain wood industries, whereas not in others, i.e. wood products development, versus paper manufacturing, or creosoting), and non-SIC defined chemical features (i.e. Petroleum spills occur regardless of SIC). Another lengthy database entailed a full description of the chemicals involved in the exposures listed by CRI. This included information related to a) NOISH/OSHA defined chemical halflives in the environment, b) descriptions of the various forms and levels of toxicity related to these chemicals (including measurements for eye, skin and oral surface toxic levels, and the results of human versus mouse studies), c) data regarding levels and forms of toxicity as defined by the medical field and EPA, and d) the use of MEDLINE data and NIH-defined indices regarding level and type of toxicity, reprotoxicity, mutagenicity and carcinogenicity. This database also included enumeration of Oregon sites, numbers of articles published on the release as noted in Sax's Environmental Chemical dictionary, and numbers of reports filed for CRI related to each toxic chemical. In sum, much of this toxicity data had to be developed from scratch, and formed several databases very useful for later review of similar topics. Other lengthy databases used for this study include a) the OSCaR data, b) the listing of Oregon TRIs (ranging from 1500 to 2000 sites depending on the focus) and c) the listing of all chemical reports filed by Oregon 540 CRI sites review (ca. 4500 reports total).

3. Medium length/size databases were developed as a result of the above two processes. Several examples of these and what they consist of are provided in the Appendix (**APPENDIX A**). These databases were developed so they could be linked or joined with various forms of GIS related data. Also during this time, a final database consisting of Oregon Leukemia and Lymphoma cases was acquired from the State, and, with the assistance of the Population Center, was geocoded based on address location, resulting in a shapefile and related *.dbf derived from the data contained in OSCaR.

Stage III. Once databases and shapefiles were developed for GIS use, attempts were made to map out this information. In particular, Oregon Leukemia and Lymphoma point maps were produced, and these layout compared with locations produced manually in GIS for site the following types of toxic release locations: Superfund Sites, Superfund applicants, and "Top 22 High Risk CRI sites" as defined by methods detailed in Chapters 5, 6, and 7 of previous writings. Any remaining statistical work was then carried out using Excel, Access, Crystal Reports, ArcView GIS, and SPSS.

RESULTS

Oregon State Cancer Registry (OSCaR) provided the case data used for this study, a brief description for which is found in **APPENDIX B**.

During the initial phase of this part of the research, basemaps were produced detailing the spatial distribution of leukemia in Oregon. Some of the first maps relied on linking cases to point and area shapefiles produced by ESRI and provided as part of the ArcView GIS program OSCaR information enabled maps to be produced using zip code and town or city location (See **APPENDICES C – ELECTRONIC INFORMATION SOURCES** and **D – EARLY GIS MAPS**).

Case Data Issues

Assigning addresses to the OSCaR data resulted in fairly poor results. Addressing the case data is based on address information; this portion of the work was carried out primarily with the assistance of Thomas J. Kimpel and a student assistant employed by the Oregon Population Research Center located on campus. Unfortunately, only about 50% of the total number of Leukemia and Lymphoma cases could be addressed. These fair to poor results were due to the following:

- data entry methods (methods of defining an address, P.O. versus PO)
- inconsistent use of abbreviations (N and No for north)
- use of addresses that don't exist or are not included in the base maps for this program (i.e. unofficial private street addresses)
- the use of rural route delivery codes
- the use of mileage values along a given road (i.e. 1.3 miles along County Road #)
- the use of PO boxes or other form of enumerated mailbox to define location (i.e. trailer park, apartment number).
- lack of any useable address data
- difficulties imposed by the use of terms such as "end of *** street" and "corner of . . . "

In addition, since OSCaR provides use information provided about the patient by physicians, this information is considered second hand and may therefore be even less reliable. In sum, ca. 370 out of 4047 leukemia cases (9%) could not be assigned addresses due simply to the form this data was in; for lymphoma cases, 225 out of 2807 lymphoma cases could be geocoded (a loss of 8%). When address matching was atttempted, only 50% of the total cases could be effectively geocoded. In some cases (i.e. "end of . . . " and street corners), point lcoations were manually assigned. Addresses with private building-related apartment box numbers could not be assigned repcise locations, nor could addresses which appeared to be trailer park locations.

This information was used to produce maps of county, township and zipcode related features (**APPENDIX D**).

TRI Reports, Townships, Zip Codes and SIC

Zipcode related analysis was of limited value. Zipcode areas themselves have to be further defined demographically in order to best relate incidence with local population, economic, business (i.e. SIC), and toxic release data. Most importantly, the relationship between zip code tract and such issues as socioeconomics, place of business, and association with the SIC-related features focused on for this work have not yet truly explored, even though data resulting from valuable demographic studies on this relationship do exist (i.e. Harris InfoSource publications sponsored by Oregon Economic & Community Development Department in Salem, including such publications as the *2001 Oregon Manufacturers Register*). Zipcode-related case analysis was no longer carried out following the initial stage in this work, and most certainly provides an avenue for further exploration of this method of GIS analysis of public health-related spatial data

In a final series of early reviews of this data, information concerning TRI or CRI-related Wood and Lumber industries was gathered and related to case data. The results included the development of a map of "relative risk" for leukemia. This "relative risk" was assigned to given township based on their SIC-derived business data and OSCAR-derived case data. Relying on economic data available in *Ward's Business Directories*, SIC24, SIC25, SIC27, and SIC28 business series were mapped and numbers of cases in the same town related to the given towns or villages these industries operated in. These "Relative Rates" (i.e. figures in **APPENDIX D**) were defined as incidence of Leukemia in relation to town or population size, based on the "normalization" techniques available in the ArcView GIS classification program. This work entailed normalizing the numbers of cases to the demographic data included in the related ESRI databases for this project. It is important to note however that this normalization process is distinctly different from age-adjusted rate calculations; therefore, this type of review may have limited applications until proper adjustments are made of the final data. This analysis served mostly as a preliminary step in this work.

Example of SIC Use (**APPENDIX E**)

To provide an example of how GIS may be used to research a topic of public concern or interest, an industry often pointed to as a source for potential exposure and toxic release-related carcinogenesis was evaluated--Creosoting (SIC 2491). Unlike other SIC24-defined Wood Industries in Oregon, SIC 2491 is unique in its chemical production. Particular to this industry is the production of potentially carcinogenic agent related to cresol-- meta-Cresol and para-Cresol (see **APPENDIX E, FIGURE E-1**). With a reclassification of SIC data based on chemical profiling of TRI and CRI sites, a series of Creosoting industries were identified and grouped together for analysis of data derived from the *Ward Business Directories*. This information next led to research focused on SIC-related chemical data for given areas and toxic waste sites related to Creosoting.

In theory, this project served as a brief exercise in carrying out a site chemistry analysis based on SIC data. It may also be used to demonstrate several issues that may erupt in such analyses.

Even though one's first impression might be that chemicals common to the creosoting industry such as cresols, pentachlorophenols, arsenic and heavy metals, are the main reason for local cases clusters of childhood leukemia, this research shows that careful profiling of chemical industries is warranted before blaming such industries for these "disease aggregates". The steps taken in developing a base map and buffering the "high risk" townships where creosoting business is the mainstay of the local economy demonstrated a number of possibly conflicting findings (see **E-15**). In particular, a number of chemical similarities and differences exist between Creosoting industry, wood industries, and other industries that use pentachlorophenol, arsenic and two most common heavy metals tested for at Oregon CRI sites (**E-12**).

Most importantly, areas where pentachlorophenol could be blamed for carcinogenesis are more likely to rely upon the paper industry than the creosoting companies (see **E-12** in appendix). Moreover, if it was publicly decided that cresol in fact to blame, then this fairly scarce chemical should not be present at other sites used as "controls" in the final statistical analysis. Only when it is decided that cresol is in fact to blame, its chief producer may probably also be blamed. However, to blame a single type of business on the impact of a single chemical it released assumes such an impact is derived from just one-chemical. TRI sites rarely have just one chemical released into the environment (recall Chapter 6 results). In addition, such blame might prevent recognition of other more toxic by-products released at the same site or elsewhere. In this case, cause and effect is evident, although not proven or proveable. This use GIS serves more in improving public awareness of the issue at stake in its entirety, without allowing the public to focus on just one narrow issue or concern.

In essence, this exercise in the use of GIS represented just what types of information may be assembled in fairly short time using several of the database systems (GIS, EXCEL, ACCESS, CRYSTAL REPORTS and SPSS). A number of maps may be drawn up in a day's worth of time, and a better understanding developed of the issues at hand. In this case of review of individuals with "cancer" around wood industry cities, although no statistical was carried out, the end final map could be used to select for and begin to design a research plan for engaging in more detailed approach to this public health issue. In this case, 115 cases could be found in the immediate vicinity of creosoting companies in twelve towns. Moreover, an enlargement of the buffer zone chosen to do this form of case identification (ca. 20 miles), resulted in the production of ca 450 cases (three times as many) by tripling the buffer region (See last map in **APPENDIX E**).

Aside from statistical profiling, this application of GIS is helpful in identifying new clusters and, more importantly, eliminating cause. It can be used to provide valuable public health/community relationship data and resources. More importantly, this detailed evaluation of SIC-defined locations demonstrates a number of issues attached to assigning blame to just a single chemical or single type of industry.

Chemical Databases

The final part of this work focused on the development of databases for defining high risk sites in Oregon based on CRI chemical reports. In total, nearly 4500 reports were databased and analyzed. This led to the identification of particular site features which in turn were related to reclassified SIC data for Oregon CRI sites. This led to the definition of chemicals profiles for given classes of SICs (Chapter 6), and suggested that in some cases this kind of approach may be used to predict or explain local environmental chemistry features. Such a model was shown to be most effective for such industries as dry cleaning, companies with limited chemical use.

This kind of detailed study of chemicals is often excluded from evaluations of large numbers of toxic release sites due to data form and management issues, especially regarding the amount of work involved with data production and improvement. For this study, chemical complexity was maintained as smaller numbers and groups of sites became the focus. To carry out this research of Oregon's 540 CRI sites, just a few would be selected for and labeled "high risk sites." These sites were evaluated in addition to the already identified superfund and superfund applicant sites.

Data Development. To define Oregon's High Risk sites, the following databases were developed and then tested for this application.

These databases focused on (see **APPENDIX C** for listing of source):

- Oregon TRI/ECSI site reports (downloaded from EPA sites)
- CRI Reports (one report per chemical, per test, i.e. air, soil and water tests)
- Superfund application site information
- Superfund sites information
- High Risk site information (defined as part of this study)
- Chemical lists (for describing the 250 chemicals found at Oregon CRI sites)
- Chemical Class/Group information (esp. amounts of each according to related Site Id)
- Oregon TRI/CRI Company Business or SIC-related data
- CAS – Chemical name cross referencing
- SIC-Reclassified SIC cross referencing
- Chemical carcinogenicity and toxicity information (two different types of databases)
- Related NIOSH and OSHA carcinogenicity and environmental halflife data.

Following completion of these databases, extensive testing and redesigning of the data was necessary, including exporting/importing the data back and forth between EXCEL, ACCESS and GIS to test out this type of use of the information. In general, these databases are fairly stable. In a few cases, data did not transfer due to glitches in formatting, for examples, 1) the ACCESS program misread data which was primarily numerical as presumably solely numerical data and at times failed in accepting all of the data, or truncating the import process early on, and 2) OSCaR data could not be fully utilized due to an inability to join/link the *.dbf-derived data related to the shapefile in order to map *.dbf data (i.e. Age and Sex); this complication was a result of actions taken in the address coding process and could not be corrected for in time.

Database Management. For ArcView GIS to read this data, it was exported from EXCEL in either DB III format or txt format. The DB III format worked reliably, although it truncated chemical names due to limitations in datum size (generally ca 9 characters). This problem also impacted addresses for companies and long company names. In spite of these early problems experienced, much of the data DB III format worked nicely due to the focus on Site Identification numbers provided with all CRI datasets (these databases are reviewed in **APPENDIX A**). This reliance on SITEID allowed for an analyses of chemicals related to given site to be used for defining the sites considered to be of highest risk. In later work, when lengthier forms of datum were imported in GIS, a delimited text format (.txt) was applied instead.

Depending on the project, counts of data extracted from TRI and ECSI gave Oregon TRI sites ranged from 2487 to nearly 3000, representing the same number of TRI/ECSI sites. An overview of this data is given in the following figures (**FIGURE 8-1, 8-2**).

FIGURE 8-1

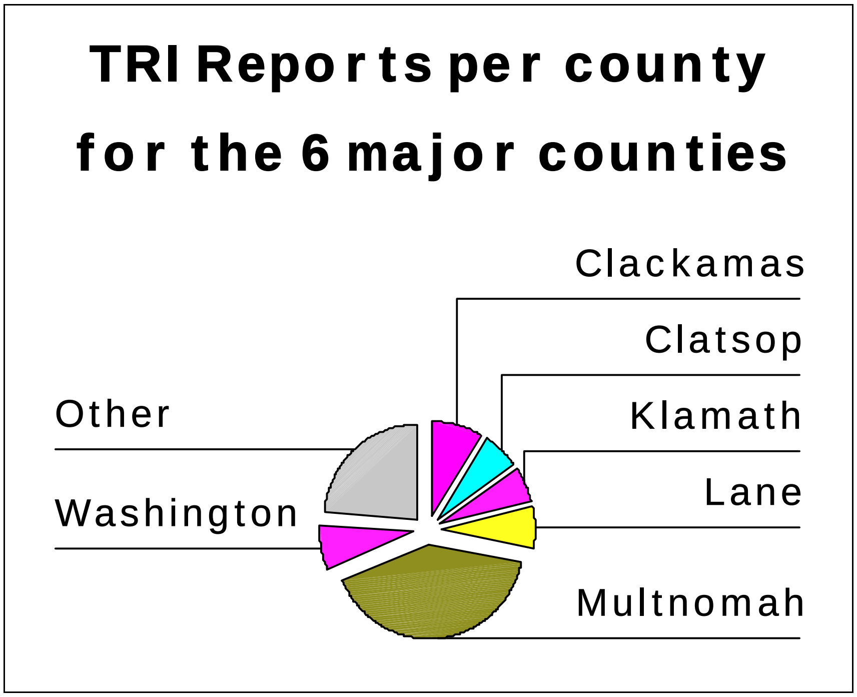
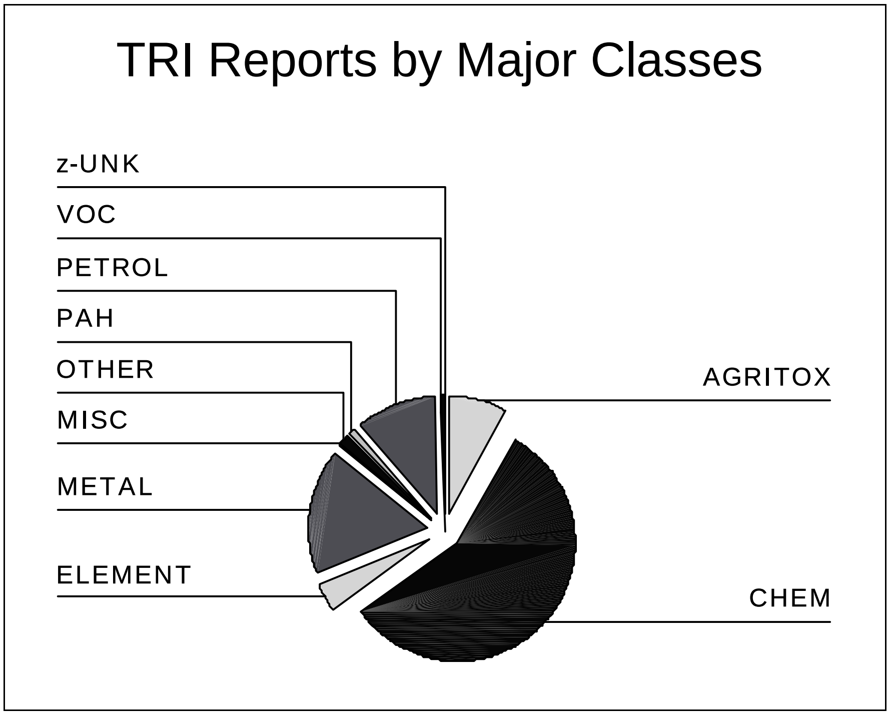


FIGURE 8-2



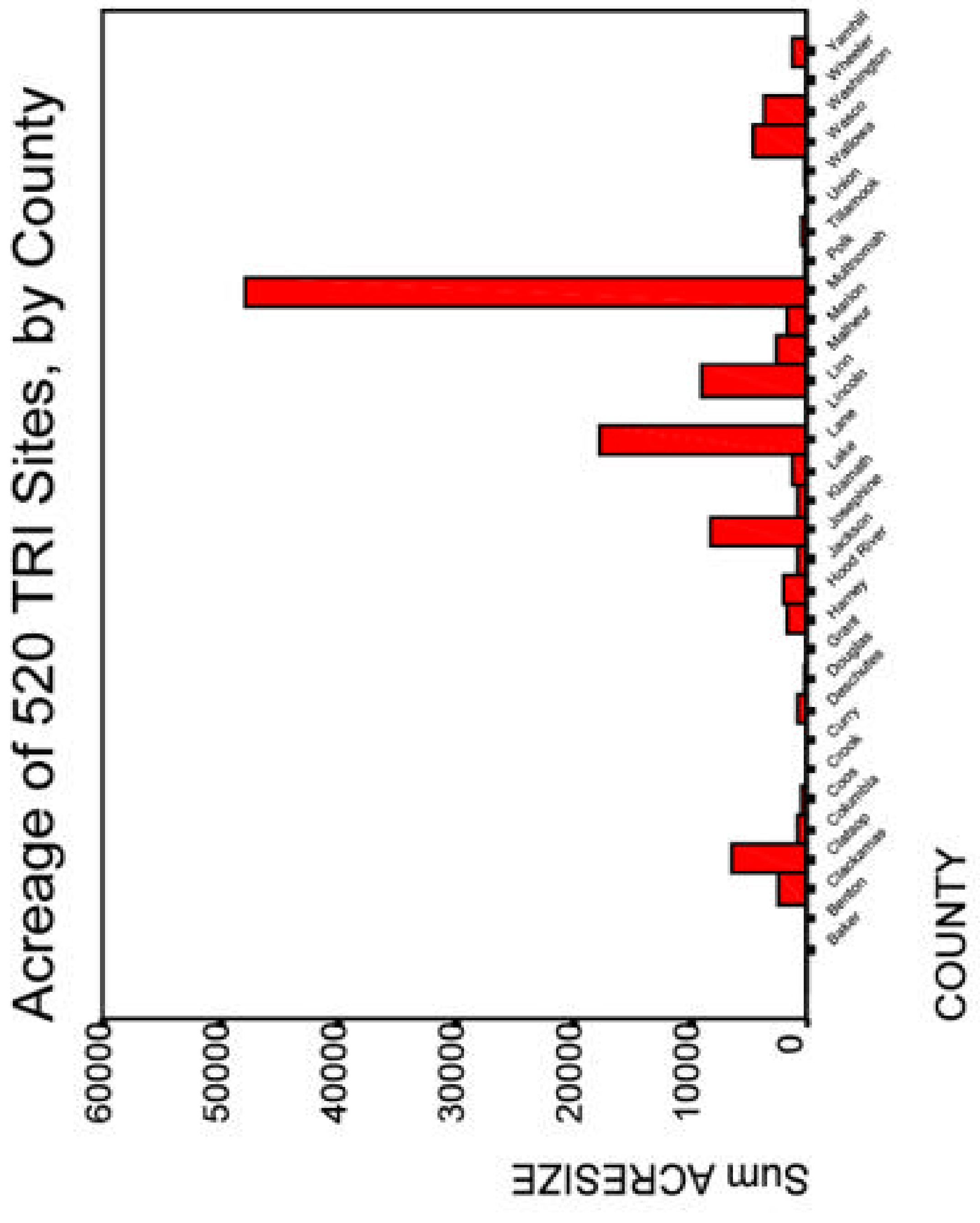
As described in Chapter 5, CRI data was reviewed extensively to produce chemical profiles for the bulk of the sites. Sites not profiled are detailed in a lengthy table included at the end of this chapter (**TABLE 8-2**). In brief, these include reports concerning fuel products and home/industrial products such as paints.

A more detailed database was then produced using Oregon CRI site data. Oregon CRI sites numbered 540, twenty of which were excluded due to insufficient data, i.e.:

- Important data missing, such as Address
- No chemical reports were filed, although chemicals are listed in detail in text boxes for this company (early on, a few examples were included in the total assessment based on information extracted from these text boxes)

The following information was allowed to be missing: SIC, zip, years in operation.

Depending on the analysis, the initial extraction of dataset due to missing data varied. In an analysis of age-related data, 41 sites (7.5%) were excluded.   For the analysis of chemical groupings, data from 527 sites (2.4%) was used.   All in all, the data provided by CRI served this project very well. The acreage of these sites per county is detailed in the following figure (**FIGURE 8-3**), information later used for assessing #spills/TRI/CRI reports per acre to better chemical risk for particular regions.



Chemical Profiling High Risk Sites.

With the databases produced, the focus turned to relating Oregon cases to the Geographic Information System. For this, several databases were improved to serve specifically as links to the ArcView shapefiles. This process allowed these high risk sites to be linked to databases in order that chemical profiles be produced and spatially displayed relative to leukemia and lymphoma case distribution. Final maps were produced based on these efforts (**APPENDIX F**). This enabled maps to be used demonstrating site location and chemical profile in relation to local leukemia and lymphoma case distribution. In theory, a spatial analysis of this work may then be employed, using age-adjusted incidence data to compare local cancer incidence with proximity to Superfund, Superfund applicant and predefined High Risk CRI sites. Given an ample amount of time, similar profiles may be defined for more "High Risk" Sites, increasing the density of the High Risk Site map and more accurately demonstating any high risk for possible exposure to chemical that might exist.

This most important part of the GIS database used the following dataset to assign chemical values to CRI site numbers:

TABLE 8-1

Column#	Datum	Description
1	ID	Ordered from highest to lowest risk
2	SITEID	N defined by EPA
3	NAME	Company Name
4	Plainrank	
5	TotalChems>=30?	Assign 1 for Yes; 0 for No
6	PAHs>8?	Assign 1 for Yes; 0 for No
7	>30 with Aromatic?	Assign 1 for Yes; 0 for No
8	>30 with Metals?	Assign 1 for Yes; 0 for No
9	Petrols, with >30?	Assign 1 for Yes; 0 for No
10	Agrichem with >30?	Assign 1 for Yes; 0 for No
11	Total + Hal Ali >30?	Assign 1 for Yes; 0 for No
12	Total + Hal Aro > 30?	Assign 1 for Yes; 0 for No
13	heteroSUmIndex	Sum of columns 5 through 12.
14	Address	Company Address
15	Lat	degrees minutes seconds
16	Long	degrees minutes seconds
17	City	As given in CRI
18	Zip	As given in CRI
19	County	As given in CRI
20	Acreage	As given in CRI
21	SIC	As given in CRI
22	Yrs1	Year business operation commenced
23	Yrs2	Year business operation ceased.

SUMMARY

In conclusion, this work provided the opportunity to explore new methods of GIS application to epidemiological and environmental chemistry work. This research accomplished the following:

1. It demonstrated the need to improve OSCaR-derived cancer data in order to improve its mappability. In particular, addresses need to be more accurate and should be entered in a standardized fashion. Such a process would not only improve the address matching process, but also increase the number of cases that can be used in any statistical evaluation process.

2. It demonstrated the need to incorporate GPSing into OSCaR datsets. To improve the definition for actual home address location, research results suggest that GPS-derived latitude and longitude data need to be added to OSCaR. Should the researcher chose to GPS these sites, a careful review by IRBs Is required.

3. It demonstrated the value of chemical profiling toxic release sites to spatially depict chemical risks for given contamination sites. Such an approach would improve arguments made regarding actual cause-effect relationships and depicts more accurately the potential exposure risk. One chemical or carcinogen cannot be argued if other remaining risk factors are ignored.

Future projects need to take the following steps to improve results:

1. Increase site sample size to examples beyond CRI-defined locations. This should make SIC-defined groups more consistent in size and thereby improve chemical profiling process, i.e. include TRI site information.

2. SIC reclassification methods should be re-evaluated and other methods of reclassifying tried, for SIC alone doesn't define the contamination that exists.

3. Grouping certain chemicals together reduces the power of this method or reasoning, i.e. toxicity features for one PAH, naphthalene, do not resemble the carcinogenicity and toxicity of other PAHS like benzofluoranthenes.

4. Methods of defining "signatures" may need to be improved, and the assignment of specific chemicals to specific groups in need of an emphasis on degree of toxicity certain chemicals possess. Some chemicals in fact may in fact be excluded from the final cancer risk assessment equation or the process of evaluating sites based on chemical heterogeneity.

5. An evaluation of site chemistry based on the amount actually present in the environment at the time of the sampling needs to be added.