

GIS-EpiLink: A Spatial Search Tool for Linking Environmental and Health Data

F. Benjamin Zhan · Jean D. Brender · Yaowen Han ·
Lucina Suarez · Peter H. Langlois

Received: 12 December 2005 / Accepted: 9 March 2006 / Published online: 12 September 2006
© Springer Science+Business Media, Inc. 2006

Abstract One inherent characteristic of both environmental data and health data is that they have a location component. This characteristic makes Geographic Information Systems (GIS) an ideal and sometimes indispensable tool for analyzing environmental and health data. Indeed, the past decade witnessed significant efforts in developing GIS tools for supporting epidemiologic research. Despite these efforts, the availability of accessible GIS tools that can be easily used by epidemiologists to link environmental and health data has remained a problem. We present a simple spatial search tool—GIS-EpiLink—that can be used to link environmental and health data when distance between an environmental site and the location of the maternal address of a case or control is used as a proxy for exposure. The tool was used in a research

project and it successfully provided the necessary data for epidemiological analyses. This tool should be very useful to epidemiologists in environmental health research.

Keywords Geographic information systems · Epidemiology · Environmental exposure assessment · Spatial search

Introduction

Many researchers recognize the importance and potential of Geographic Information System (GIS) technologies and spatial analysis methods for environmental public health tracking which aims to integrate environmental hazards monitoring, exposure assessment, and health effect surveillance for studying the health effects of environmental hazards [1–5]. Indeed, the past decade witnessed a rising interest in using GIS for environmental health research by researchers from different communities in different countries. Among published works, Nobre and his colleagues [6] described a GIS tool that can be used to represent and analyze epidemiologic data based on case studies from Brazil. Vine and other researchers [7] recognized the potential and limitations of GIS in enhancing epidemiologic research and provided discussions about how some of the GIS functions may be used in such research.

O'Dwyer and Burton [8] summarized some of the problems that researchers may have to overcome in realizing the full potentials of GIS for health related research. Boulos and his colleagues [9] and Kistemann and other researchers [10] reviewed the potentials and barriers of using GIS in research areas such as disease ecology and the delivery of health services. Rushton [11] reviewed how GIS had been used for organizing geospatially-referenced health data and how

F. B. Zhan (✉)

Texas Center for Geographic Information Science, Department of Geography, Texas State University, San Marcos, Texas 78666
e-mail: zhan@txstate.edu

J. D. Brender

Department of Epidemiology and Biostatistics, Texas A&M School of Rural Public Health, College Station, Texas 77843
e-mail: jdbrender@srph.tamhsc.edu

Y. Han

Texas State University, San Marcos, Texas 78756
e-mail: ywhan@txstate.edu

L. Suarez

Epidemiology and Disease Surveillance Unit
Texas Department of State Health Services, Austin, Texas 78756
e-mail: lucina.suarez@dshs.state.tx.us

P. H. Langlois

Birth Defects Epidemiology and Surveillance Branch, Texas Department of State Health Services, Austin, Texas 78756
e-mail: peter.langlois@dshs.state.tx.us

GIS-based spatial analysis tools had been used to support health related research activities and applications. He pointed out that future development in this area will likely integrate GIS functions and geo-information processing functions of Geographic Information Science to facilitate the suite of specific requirements in health related applications.

More recently, Kaminska and other researchers [12] reviewed some functions of GIS and discussed how GIS was used for assessing the impact of pesticide pollution on public health. Nuckols and other researchers [5] discussed the use of GIS for research in environmental science and epidemiology and concluded that GIS can indeed be used to enhance environmental epidemiology studies in exposure assessment and to help provide a better understanding of the association between environmental contamination and disease concentration in some geographic areas. Researchers have also recognized GIS technology as among a number of promising technologies for assessing environmental exposure at the individual level [13].

Despite these efforts, the availability of accessible GIS tools that can be easily used by epidemiologists and other public health professionals to link environmental and health data has remained a problem. In this article, we present a simple GIS tool—GIS-EpiLink—that was designed and implemented to help epidemiologists facilitate their research in linking environmental and health data. The tool can be used to search for any pair of environmental sites and cases or controls based on different search criteria when distance is used as a proxy for exposure. For example, the location of an environmental site, the location of the maternal address of a case or control, the environmental hazardous materials (e.g., different chemicals) in question, and a threshold distance between the location of an environmental site and the location of a maternal address of a case or control can be combined into different search criteria. The search results then can be used for subsequent epidemiological analyses.

Design and Implementation of GIS-EpiLink

For a GIS tool to be useful to epidemiologists, the tool must be designed in such a way that it closely follows the procedures in which epidemiologists link environmental and health data. Based on this observation, we held several meetings among a group of epidemiologists and GIS specialists. The group discussed the general procedures of linking environmental and health data from the perspectives of an epidemiologist.

In environmental health research, epidemiologists typically use case-control studies to evaluate the health effect of environmental exposure. Cases are subjects with an observed condition, whereas controls are subjects without the condition. Both cases and controls can be exposed to an environmental agent (e.g., a specific chemical) that is suspected

Table 1 An epidemiologist's view of the basic steps in linking environmental and health data

Step	Summary of procedures
1	Specify an exposure of interest—A specific substance (e.g., lead)
2	Search all environmental sites in the geographic area in question and select those sites with the exposure of interest specified in Step 1
3	Specify a search threshold distance between the environmental site and the location of the maternal address of a case or control
4	Identify the pool of cases and controls of interest
5	Perform a spatial search to select the cases and controls from the pool identified in Step 4 that are within the threshold distance of an environmental site
6	Generate an output file containing all possible combinations of the specified chemical, an environmental site, and a case or control for further analysis

of being responsible for causing the condition in question. To determine whether or not a case or a control is exposed to the environmental agent, a specific method is often used to assess environmental exposure. Once the number of cases or controls that were exposed or not exposed to an environmental agent is determined by the environmental exposure assessment method, a ratio between the odds of exposure of cases and odds of exposure of controls is calculated and a confidence interval is determined [14]. This odds ratio and confidence interval are then used to evaluate whether or not the environmental agent can be considered to be associated with the condition [14].

After various discussions, the group came up with the procedures for linking environmental hazard and health data as summarized in Table 1. This set of procedures forms the foundation for the design of the GIS-EpiLink spatial search tool. We developed GIS-EpiLink using Visual Basic Application (VBA) within the ArcGIS application development environment. In developing the tool, we combined Visual Basic and ArcObject programming to utilize the rich set of GIS functions in ArcGIS. The tool must be used as a plug-in within the ArcGIS environment. Therefore, basic familiarity with the ArcGIS user interface is necessary in order to use the tool. The current version of the tool can be obtained from the first author. Users of GIS-EpiLink may modify the source codes to change the tool to suit their specific needs if they are knowledgeable in both Visual Basic and ArcObject Programming. The tool is a pop-up menu-driven and interactive software package that can be used to identify the cases and controls within a given distance of an environmental site containing a specified substance (e.g., a chemical). Figure 1 is a flowchart illustrating the flow of functions in GIS-EpiLink and the process in which a user interacts with the tool.

As shown in Fig. 1, after starting the tool, a user is first prompted to select a database from the list of chemical attribute databases of environmental sites. Each of the chemical attribute databases of environmental sites contains a list of substances that may be of interest to an epidemiologist for an environmental health research project. Once chemical attribute database of environmental sites is selected, the tool displays a list of attributes from the database showing different categories of substances (e.g., metals). When a specific attribute in the chemical attribute database is specified by the user, the tool exhibits a list of possible values (e.g., lead) associated with that attribute. The user then proceeds to con-

struct a query expression using the interactive query form provided by the tool to retrieve a specific substance in the specified category of substance from the chemical attribute database.

The tool then prompts the user to select the attributes from the selected chemical attribute database to be included in the output database—a Dbase file. In the next step, the tool asks the user to select a field in the chemical attribute database of environmental sites as the matching key. Once the matching key is selected, GIS-EpiLink asks the user to select the corresponding geodatabase of environmental sites and a field in the geodatabase as the key to match

Fig. 1 Flowchart illustrating the interactive spatial search process using GIS-EpiLink

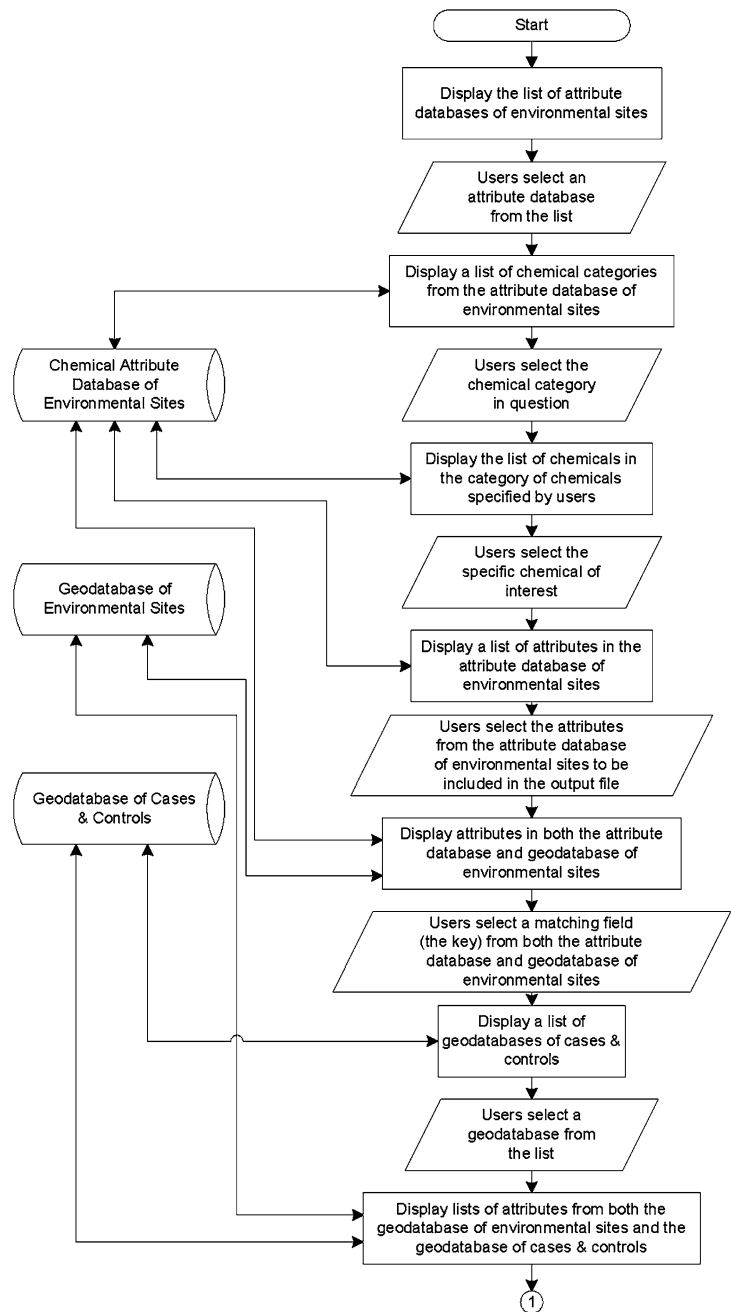
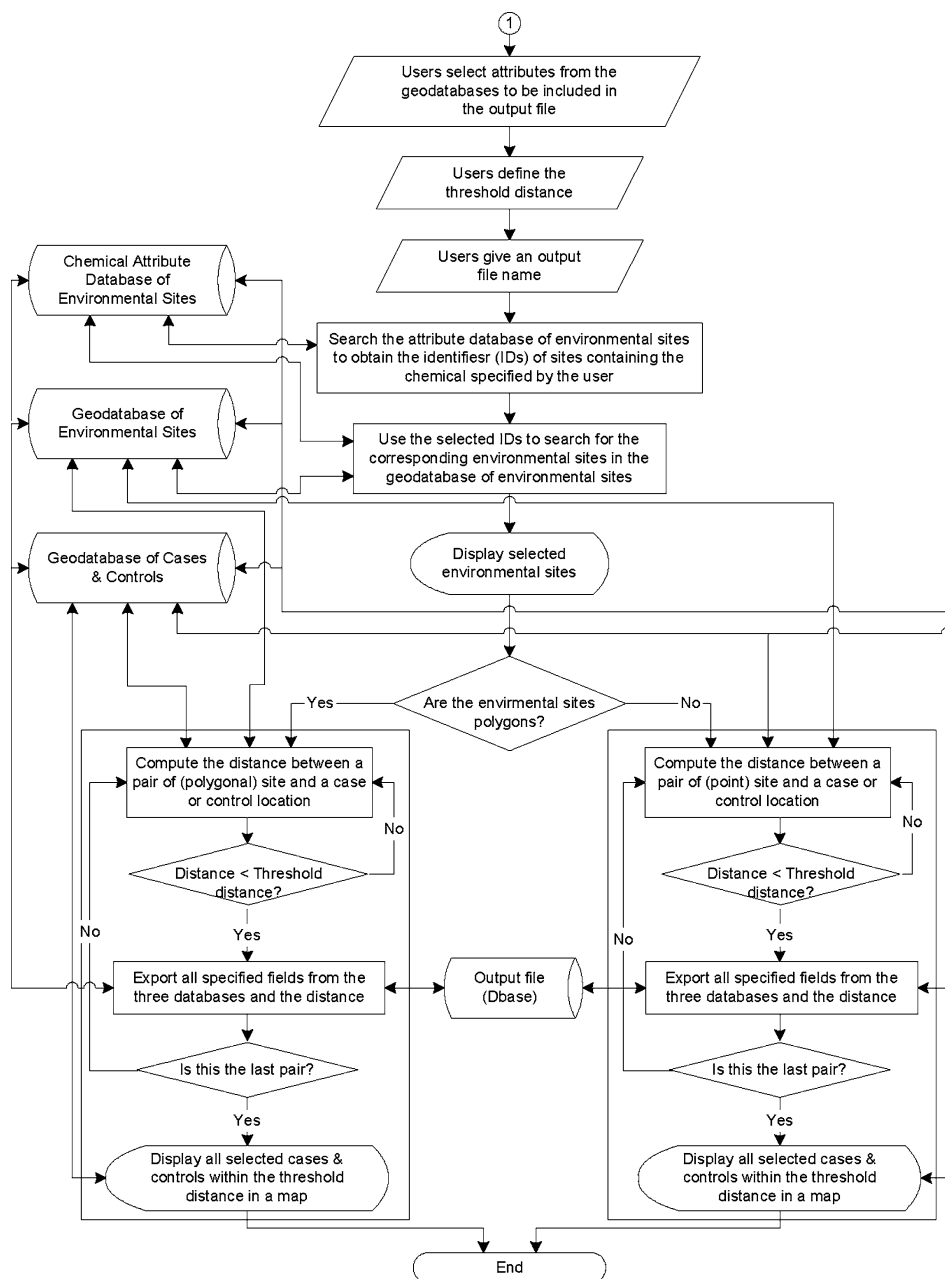


Fig. 1 Continued.



records in the chemical attribute database. The two selected fields mentioned above are used to link the chemical attribute database and the geodatabase.

In subsequent steps, the tool shows a list of geodatabases of cases and controls and asks the user to select one from the list. The user selects a geodatabase of cases and controls from the list. The tool displays the list of attributes in the selected geodatabase of environmental sites and the geodatabase of cases and controls. The user then selects the attributes to be included in the output database from both geodatabases. The user proceeds to enter the threshold distance and provide a name for the output database based on prompts from the tool.

In the final steps, the tool searches the selected chemical attribute database of environmental sites to obtain the identifiers (IDs) of all sites that contain the specified substance. The tool then uses the IDs to retrieve the corresponding sites in the geodatabase of environmental sites. The tool displays all retrieved sites in a map form. GIS-EpiLink then checks to see if the environmental sites are represented in the geodatabase as polygons or as points. If they are represented as polygons, then for each pair of environmental sites and cases/controls, the tool computes the distance from the location (a point) of a case/control to the nearest boundary of the environmental site in question. If the environmental sites

are represented as points, then the distances are computed between two point locations for each pair of environmental sites and cases/controls.

As can be seen from the last part of Fig. 1, the tool checks if the computed distance is less than the user specified threshold distance after computing the distance between a pair of environmental site and case/control. If it is, then the tool selects the case or control and exports the values of all attributes selected for output in the corresponding records of the three databases to a Dbase file. If it is not, then the tool goes to the next pair of environmental site and case/control. This process is repeated until all possible pairs of environmental site and case/control are exhausted. In the final step, GIS-EpiLink displays all selected cases and controls along the selected environmental sites in a map form.

Examples of Using GIS-EpiLink

We used GIS-EpiLink to retrieve the necessary environmental exposure information based on datasets from the State of Texas in the United States that were compiled for a research project titled “Residential proximity to environmental hazards and congenital malformations in offspring” [15]. Objectives of this project included: (1) an examination of the association between maternal residence near NPL (National Priority List) sites, state superfund sites, or other waste sites and risk of congenital malformations in offspring; (2) an investigation of the association between maternal residence near TRI (Toxic Release Inventory) sites with industrial emissions of solvents, metals, and other potential teratogens and risk of congenital anomalies in offspring; (3) an analysis of the association between maternal residence near petroleum refineries, chemical industries, or primary metal industries and risk of congenital malformations in offspring. Additional information about the methods of this project can be found in a recent article by Brender and her colleagues [15].

In order to use GIS-EpiLink, three separate databases must be prepared for each year in the study period: (1) A chemical attribute database of environmental sites containing the identifier of every environmental site, a list of chemicals or other hazardous materials observed at each site, and other necessary information associated with each site; (2) a GIS geodatabase consisting of the location information and identifiers of all environmental sites; (3) a GIS geodatabase containing the location information and identifiers of all cases and controls as well as all other necessary attributes associated with each case and control. All GIS geodatabases mentioned in this article were projected to the Texas Centric Mapping System based on the 1983 North American Datum. This projection is necessary to reduce the effects of curvature of the Earth when computing the distance between an environmental site and the location of a maternal address.

We did not have the resources to thoroughly check whether all geocoded TRI and superfund sites are the “true sources of contaminant” during the course of this study. But we made every effort to achieve the best results when constructing the databases and checked the results from GIS-EpiLink to make sure that computed distances and retrieved attributes were correct. More details about the construction of the databases and investigation of the positional accuracy of geocoded addresses are reported in other publications [15,18].

Spatial Search to Determine Exposure of a Case or Control to Air Emissions From a Point Type Environmental Site—The Case of Toxic Release Inventory Facilities

To build the environmental databases, we obtained data regarding air emissions of chemicals from Texas industrial facilities. The data were collected and maintained by the Toxic Release Inventory (TRI) program of the United States Environmental Protection Agency (USEPA) [16]. Based on mandates specified in Section 313 of the Emergency Planning and Community Right-to-Know Act (EPCRA) [17], a company is required to report to USEPA if the company is in a certain type of industry as specified by a given standard industrial code (SIC), has 10 or more employees, and it manufactures, imports, processes, or otherwise uses any of the 650+ EPCRA Section 313 chemicals in amounts greater than the specified threshold quantities. The reported data are maintained in the USEPA TRI databases. We geocoded the address of each geocodable TRI facility in Texas and obtained the geographic location (latitude and longitude) of each TRI facility to achieve a high positional accuracy of these facilities in the GIS databases [15,18]. We geocoded 7197 (87.2%) of the accumulated 8215 industrial facility records with reported air emissions from 1996 to 2000.

We then constructed the first two databases—the chemical attribute database of TRI facilities and the GIS geodatabases of TRI facilities—for each year of the study period from 1996 to 2000 using the TRI databases from USEPA and the geocoded TRI data. The chemical attribute database contains the identifier of every TRI facility, year of report, a list of chemicals released by a TRI facility, and other information associated with each TRI facility (e.g., SIC code). The GIS geodatabase consists of the location and contact information as well as identifiers of all TRI facilities.

We constructed the third database—GIS geodatabase of cases and controls—using geocoded cases and controls from the project mentioned above. This GIS geodatabase contained the location information and identifiers of all cases and controls as well as all other necessary attributes associated with each case and control. In building this

GIS geodatabase, we obtained birth defect data for births occurring 1996–2000 from the Texas Birth Defects Registry (TBDR) at the Texas Department of State Health Services (DSHS). Control births were randomly selected and frequency matched to case births by year of birth (1996–2000) and public health region of maternal residence as recorded on the birth certificate. There are a total of 11 public health regions in Texas.

All control births were selected from births with no documented defects as determined by the Texas Birth Defects Registry (TBDR). Case births with congenital malformations of interest were first linked to their respective birth and fetal death records, and these data were then merged with the control birth file to form the complete data set for case and control births. We obtained maternal addresses of cases from vital records. For cases with no maternal address in vital records, we obtained addresses from the TBDR which abstracts maternal addresses from medical records as part of Registry activities. The only addresses available for controls were from those on vital records. We geocoded maternal addresses and used the resulting coordinates to represent the geographic locations of the maternal addresses of all cases and controls. A total of 5391 (89.1%) maternal addresses of case births and 4368 (88.0%) addresses of control births were successfully geocoded.

It took about 1000 h for a graduate research assistant to reorganize the data and then geocode the addresses of TRI facilities and the maternal addresses of cases and controls. It should be noted that some cases and controls may have multiple maternal addresses associated with them in an epidemiology study. In the current version of GIS-EpiLink, the tool treats each maternal residential location as a separate record in the GIS geodatabases and computes a distance from each maternal residential location to an environmental site containing a specific chemical in question. These different records are linked through a unique identifier of a mother in the GIS geodatabases. If any one of the maternal residential locations of a case or control is within the threshold distance of exposure, then this mother is considered to be exposed to the chemical in question. This arrangement of the records in the GIS geodatabase also makes it possible for epidemiologists to choose the address among many that is relevant to a specified critical biological time (e.g., at the time of conception of a child).

In structuring the three databases, we separated the chemical attribute database from the GIS geodatabase of TRI facilities. The reason to construct separate databases for the environmental attributes and location data of TRI facilities is that a TRI facility may have air emissions containing many different chemicals. Separate databases avoid the redundancy of repeating the location and contact information for each chemical. Records in the two databases with the same TRI facility are linked through a TRI facility identifier.

We used the GIS-EpiLink spatial search tool to estimate environmental exposure of cases and controls based on data in the three databases described above following the procedure illustrated in Fig. 1. Figure 2 is a map showing counties with a selected sample of cases and controls that were within 16 km of the 2000 Texas Toxic Release Inventory (TRI) facilities with air emissions containing lead. Individual cases or controls are not displayed in the map because of confidentiality of medical information. As can be seen from Fig. 2, the selected cases and controls are concentrated in four areas: El Paso County, Harris County (the greater Houston area), Collin County, and Brazos County.

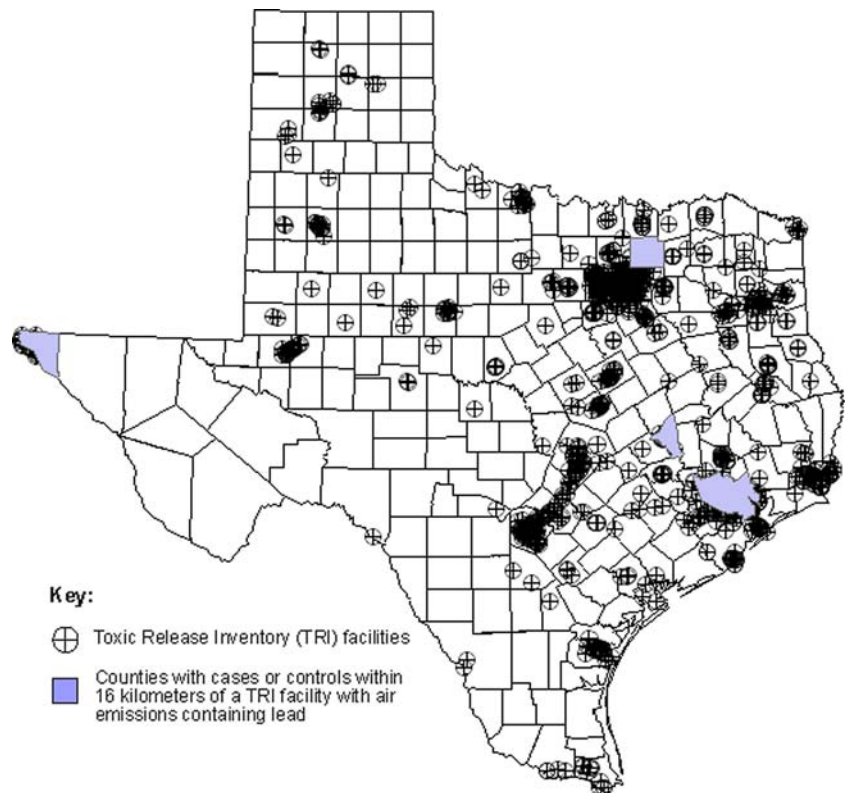
The use of threshold distance of 16 km mentioned above and in the rest of this discussion is for illustration purpose only. The tool can be used to retrieve information about any pair of case or control and environmental site within *any* given distance. The current version of GIS-EpiLink only uses distance as a proxy of exposure. It does not account for other routes of exposure (e.g., water supply).

Spatial Search to Determine Exposure of a Case or Control to Chemicals From a Polygon Type Environmental Site—The Case of Superfund Sites

Because we already had the GIS geodatabases of cases and controls constructed as described in the last subsection, we only had to construct the chemical attribute database of the superfund sites and the GIS geodatabase of superfund sites. Again the reason for separating the two databases was to avoid redundancy in the databases. We constructed these two databases based on the National Priority List (NPL) superfund sites from the Agency for Toxic Substances and Disease Registry (ATSDR) online Hazardous Substances Release/Health Effects Database (HazDat) [19]. Information in the HazDat database included site characteristics and contaminants present by environmental media and maximum concentrations found. Attribute information on NPL and state superfund sites, such as site status (active/deleted), was obtained from the Texas Commission on Environmental Quality (TCEQ) online superfund database [20]. For state superfund sites, study investigators abstracted information about site contaminants and contaminated environmental media from paper and microfilmed files stored at TCEQ in Austin, Texas. The NPL database and the abstracted state superfund data were then merged into a single database containing site and chemical-specific information for NPL and state superfund sites in Texas. A total of 43 NPL sites and 70 state superfund sites were being investigated and/or remediated (active status) during the study period.

Because the TCEQ database only contained geographic coordinates of points representing the geographic locations of the waste sites, we digitized the boundaries of the

Fig. 2 Texas counties with cases and controls that were within 16 km of 2000 Toxic Release Inventory (TRI) facilities with air emissions containing lead



sites based on Digital Orthophoto Quarter Quads (DOQQ) images with a 1-meter resolution from the Texas Natural Resources Information System (TNRIS). Land area of the waste sites ranged from less than 2 to as large as 760 acres. The superfund sites with digitized boundaries were treated as polygon type putative sources. It is worth reiterating the importance of using polygon type of putative sources to represent superfund sites because it reduces misclassification of proximity that would have been introduced by treating large hazardous waste sites as point locations. In constructing the superfund databases, it took about 500 person-hours for the researchers to abstract state superfund records at TCEQ and to enter the data in the same format as the ATSDR HazDat database. It took another 250 person-hours to organize the downloaded online databases. It took a graduate student 160 h to digitize the boundaries of the superfund sites and convert the data into a GIS format.

We were then ready to use GIS-EpiLink to determine exposure of a case or control to chemicals from a hazardous waste site using the chemical attribute database of superfund sites, the GIS geodatabase of superfund sites containing all hazardous waste sites represented by polygons, and the GIS geodatabase of cases and controls described in previous discussions. Cases and controls in a selected sample that were within 16 km of Texas state and federal superfund sites

containing benzene could be easily retrieved and displayed using GIS-EpiLink.

Concluding Remarks

Studying the human health consequences of chemicals in the environment is a complicated and complex endeavor. One of the crucial issues to making advancement in this field is the development of practical tools for epidemiologists to link environmental and health data for subsequent epidemiological analysis. We presented a simple spatial search tool—GIS-EpiLink—that can be used to link environmental and health data when distance between an environmental site and the location of the maternal address of a case or control is used as a proxy for exposure. The tool was used in a research project and successfully produced the necessary data for epidemiological analyses. This tool should be very useful to epidemiologists whose research requires the link of environmental and health data based on distances between environmental sites and the locations of maternal addresses of cases or controls. Such a tool could also be used to address community concerns about perceived excesses of birth defects and other adverse reproductive outcomes around hazardous waste sites and industrial facilities. Readers who are interested in working with this tool should contact the lead author for further information.

Acknowledgements This study was in part supported through cooperative agreement U50/CCU613232 from the Centers for Disease Control and Prevention and contract 7547547549 from the Texas Department of State Health Services Center for Birth Defects Research and Prevention.

The authors thank Ionara Delima, MAG, for her work in geocoding maternal residences and industrial locations; Zunera Gilani, MPH, for her assistance in linking birth files with the congenital anomaly registry files; and Wendy Marckwardt, MS, for her assistance in abstracting information about state superfund sites and in coding parental occupations and industries.

References

1. Jerrett, M., Burnett, R. T., Goldberg, M. S., Sears, M., Krewski, D., Catalan, R., Kanaroglou, P., Giovis, C., and Finkelstein N., Spatial analysis for environmental health research: Concepts, methods, and examples. *J. Toxicol. Environ. Health A*. 66(16–19):1783–1810, 2003.
2. Jarup, L., Health and environment information systems for exposure and disease mapping, and risk assessment. *Environ. Health Perspect*. 112:995–997, 2004.
3. Mather, F. J., White, L. E., Langlois, E. C., Shorter, C. F., Swalm, C. M., Shaffer, J. G., Hartley, W. R. Statistical methods for linking health, exposure, and hazards. *Environ. Health Perspect*. 112(14):1440–1445, 2004.
4. McGeehin, M. A., Qualters, J. R., and Niskar, A. S., National environmental public health tracking program: Bridging the information gap. *Environ. Health Perspect*. 112:1409–1413, 2004.
5. Nuckols, J. R., Ward, M. H., and Jarup, L., Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environ. Health Perspect* 112:1007–1015, 2004.
6. Nobre, F. F., Braga, A. L., Pinheiro, R. S., and Lopes, J. A. D. GISEpi: A simple geographical information system to support public health surveillance and epidemiological investigations. *Comput. Methods Programs Biomed*. 53:33–45, 1997.
7. Vine, M. F., Degnan, D., and Hanchette, C. Geographic information systems: Their use in environmental epidemiologic research. *Environ. Health Perspect* 105:598–605, 1997.
8. O'Dwyer, L. A., and Burton, D. L., Potential meets reality: GIS and public health research in Australia. *Aust. N. Z. J. Public Health*. 22:819–823, 1998.
9. Kamel Boulos, M. N., Roudsari, A.V., and Carson, E. R., Health geomatics: An enabling suite of technologies in health and healthcare. *J. Biomed. Inform.* 34:195–219, 2001.
10. Kistemann, T., Dangendorf, F., and Schweikart, J. New perspectives on the use of geographical information systems (GIS) in environmental health sciences. *Int. J. Hyg. Environ. Health* 205:169–181, 2002.
11. Rushton, G., Public health, GIS, and spatial analytic tools. *Annu. Rev. Public Health* 24:43–56, 2003.
12. Kaminska, I. A., Oldak, A., and Turski, W. A., Geographical information system (GIS) as a tool for monitoring and analysing pesticide pollution and its impact on public health. *Ann. Agric. Environ. Med.* 11:181–184, 2004.
13. Weis, B. K., Balshaw, D., Barr J. R., Brown, D., Ellisman, M., Liov, P., Omenn, G., Potter, J. D., Smith, M. T., Sohn, L., Suk, W. A., Sumner, S., Swenberg, J., Walt, D. R., Watkins, S., Thompson, C., and Wilson, S. H. Personalized exposure assessment: Promising approaches for human environmental health research. *Environ. Health Perspect*. 113:840–848, 2005.
14. dos Santos Silva, I., *Cancer Epidemiology: Principles and Methods*, International Agency for Research on Cancer, Lyon, France, 1999.
15. Brender, J. D., Zhan, F. B., Suarez, L., Langlois, P., Gilani, Z., DeLima, I., and Moody, K., Linking environmental Hazards and birth defects data: *Int. J. Occup. Environ. Health* 12(2):126–133, 2006.
16. USEPA—United States Environmental Protection Agency. TRI: State Data Files. Available at: http://www.epa.gov/tri/tridata/state_data_files.htm. (Accessed June 24, 2005).
17. USEPA—United States Environmental Protection Agency. (2001). The Emergency Planning and Community Right-to-Know Act. Section 313 Release and Other Waste Management Reporting Requirements. Available at: www.epa.gov/tri/guide_docs/2001/brochure2000.pdf. (Accessed July 6, 2005).
18. Zhan, F. B., Brender, J. D., DeLima, I., Suarez, L., and Langlois, P., Match Rate and Positional Accuracy of Two Geocoding Methods for Epidemiological Research. *Annals of Epidemiology* (in press), 2006.
19. ATSDR—Agency for Toxic Substances and Disease Registry. (2005). HazDat Database. Available at: <http://www.atsdr.cdc.gov/hazdat.html>. (Accessed June 22, 2005).
20. TCEQ—Texas Commission on Environmental Quality. (2005). TCEQ Site Layers. Available at: <http://www.tnrcc.state.tx.us/gis/sites.html>. (Accessed June 22, 2005).