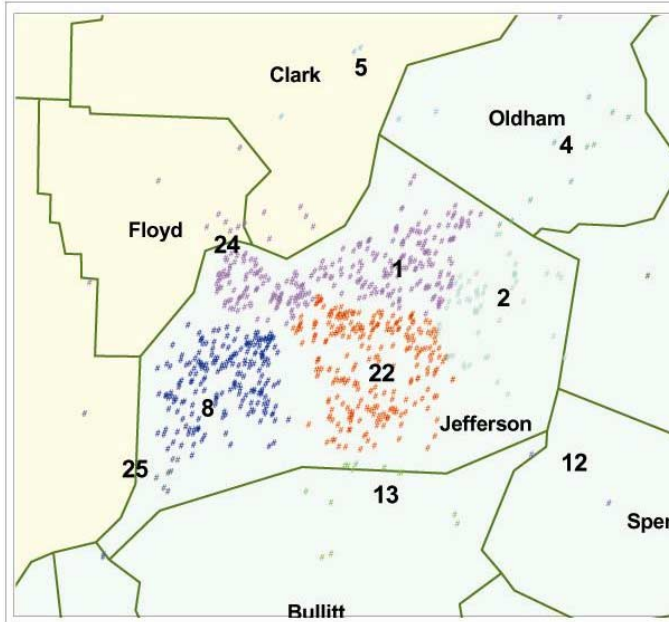




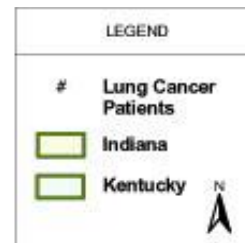
A Medical Geography Applications Area in the Department of Mathematics

Patricia B. Cerrito
Department of Mathematics
Jewish Hospital Center for Advanced Medicine
pcerrito@louisville.edu

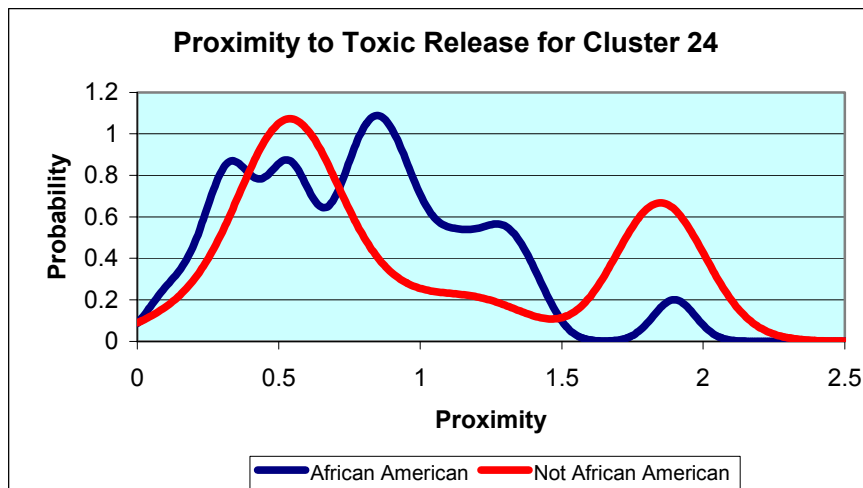
Patient Locations from Screening for Lung Cancer Study



With the same data, the subjects were clustered by locality. Cluster 24 in the northwestern corner of Jefferson County, with the greatest proportion of African Americans in the Louisville area had a higher rate of smoking than any of the other clusters. However, the African Americans in cluster 24 reported a lower rate of current smoking (58.49%) than the did the general Caucasian population in the same cluster.



Using a statistical measure, the actual proximity to hazardous waste was identified:



Medical Geography Applications Area Courses

- **Math 665 Advanced Linear Statistical Models.** Distribution of quadratic forms, estimation and hypothesis testing in the general linear model, special linear models, applications.
- **Math 667 Methods of Classification.** Classification methods used in the industry to handle large databases. Logistic regression, structural equation modeling, multivariate analysis, data mining.
- **GEOG 521: Medical Geography.** Introduction to concepts, methods and tools used to investigate geographic aspects of health and disease. Application of concepts and methods through analysis of health, population and environmental data.
- **GEOG 522: GIS and Public Health.** Application of tools and methods of analysis in geographic information systems (GIS) to public health. Use of ArcGIS software to manage and analyze health, census and spatial data.
- **GEOG 557: Advanced Geographic Information Systems.** Application of advanced GIS concepts to real-world projects. Will focus on development and implementation of a digital geo-spatial database. The project will be carried from the design phase through completion.
- **GEOG 583: Spatial and Non-spatial Database Management.** Provide students with "hands-on" experience in development, management and integration of spatial and non-spatial databases, using GIS and database management software.
- **GEOG 656: Spatial Statistics.** The analysis of spatial patterns and processes through the use of spatially based statistics.

Other elective courses are also available.

New Courses Offered

During the summer of 2004, a 2-course sequence is proposed that will integrate medical geography with statistics. The proposed syllabus is listed below.

Medical Geography Topics:

1. Exploring GIS concepts
2. Displaying digital data
3. Database query
4. Working with spatial data
5. Working with tabular data
6. Editing data
7. Working with georeferenced data
8. Presentation of data

Statistics Topics:

1. Kernel Density Estimation
2. Cluster Analysis
3. Rule Induction and Artificial Neural Networks
4. Association Rules
5. Text mining

Because these courses are new, they will be taught as special topics courses. They will count towards the application area of Medical Geography, or as electives in Data Mining. The courses will be taught by a team of faculty from Geography and from Mathematics.

Use of Statistical Software

Spreadsheets have some statistical tools but are extremely limited and should not be used as statistical packages. Small statistical packages can be purchased for use on 1 desktop at cost < \$500. Their use is limited and can only perform relatively simple, routine statistical methods.

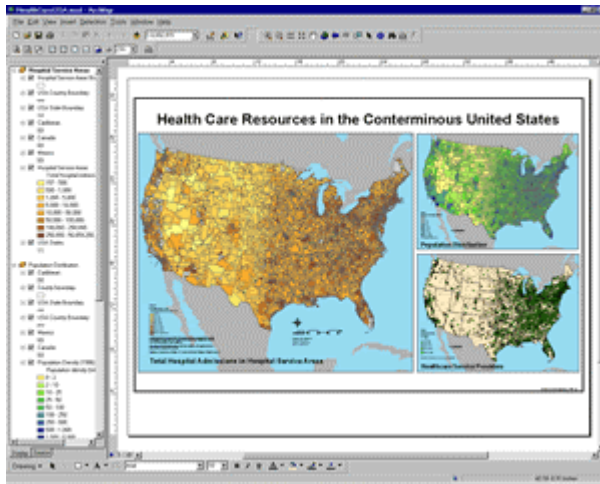
The two main statistical packages used in research are SPSS and SAS. SPSS was developed for use with the social sciences. It still sells primarily to an academic market. SAS now incorporates all the social science methodology of SPSS + database management. It was developed for use in the natural sciences. It is now the primary package used in the business market. **The ability to use SAS has now become a very marketable skill.** Businesses advertise for SAS experience. For example, FDA statistical guidelines were in part derived with SAS language. Pharmaceutical companies all use SAS as the statistical software package. Other statistical packages (such as S-plus) might be used as add-ons to SAS, but SAS remains the primary package.

One of the reasons that SAS remains so popular is that each year in early April, the SAS User's Group International holds an annual conference attracting approximately 4000 attendees. SAS developers interact with SAS users (mostly from the business environment) to improve the product. **Most of the innovations in using statistical methods are now coming from the business world rather than the academic world.**

Use of GIS Software

The primary desktop GIS software used in the world is ArcView, developed by Environmental Systems Research Institute (ESRI) Inc. From the ESRI website (www.esri.com):

“ArcView 8.x is designed with an intuitive Windows user interface and includes Visual Basic for Applications for customization. ArcView consists of three desktop applications: ArcMap, ArcCatalog, and ArcToolbox. ArcMap provides data display, query, and analysis. ArcCatalog provides geographic and tabular data management, creation, and organization. ArcToolbox provides basic data conversion. Using these three applications together, you can perform any GIS task, simple to advanced, including mapping, data management, geographic analysis, data editing, and geoprocessing.



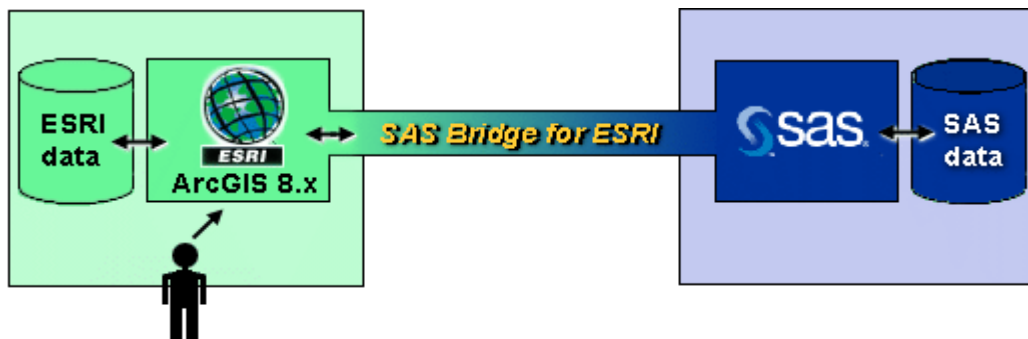
ArcView 8.x offers many exciting capabilities such as expanded symbology, new editing tools, metadata management, and on-the-fly projection.”

SAS Bridge for ESRI

Most GIS software has limited statistical analysis capabilities. In the past, GIS users have exported data into an interchange format, then imported it into a statistical analysis package, such as SAS. Recently, a bridge has been developed to link ESRI's GIS data with SAS's statistical analysis capabilities.

This bridge will link tabular to spatial data. It can use all of the SAS statistical and data mining methods on spatial data developed using ArcView. From www.esri.com/sas:

“To address customer-driven demand, SAS and ESRI are creating an integration product between the companies' software. This joint product development will give organizations the ability to exchange spatial and attribute data and metadata between Arc GIS and the SAS server in a GUI-driven environment. Leveraging the work done to-date, the companies are developing a tight integration between their respective technologies.”



Text Mining

Text mining can be used to investigate coded information. Inventories generally are identified by codes. To investigate customer choices, the standard statistical tool is to identify each inventory code as a separate category. The result is too many categories. In healthcare, patient diagnoses are also categorized by codes that can be analyzed using text mining.

Text mining uses the process of “stemming”. Words that have similar root stems are considered the same. Related inventory items generally have the same code stem. Therefore, similar items can be grouped using text mining to reduce the number of categories to a manageable amount.

Suggested Master's Degree

Course	Description
Math 566 Nonparametric Statistics	Rank tests for comparing two or more treatments or attributes, the one-sample problem, tests of randomness and independence, nonparametric estimation, graphic methods, and computer programs.
Math 661 Probability Theory	A measure-theoretic approach to topics in probability theory; conditional probability, conditioned expectation, types of convergence, strong law of large numbers, characteristic functions, and the central limit theorem.
Math 662 Advanced Mathematical Statistics	Classical theory of statistical inference, asymptotic theory and robustness, Bayesian inference, and statistical decision theory.
Math 665 Advanced Linear Statistical Models	Distribution of quadratic forms, estimation and hypothesis testing in the general linear model, special linear models, applications.
Math 667 Methods of Classification	Classification methods used in the industry to handle large databases. Logistic regression, structural equation modeling, multivariate analysis, and data mining.
Math 699 Special Topics I	New course sequence to integrate the use of GIS with SAS in examining medical geography using spatial statistics.
Special Topics II	
GEOG 521 Medical Geography	Introduction to concepts, methods and tools used to investigate geographic aspects of health and disease. Application of concepts and methods through analysis of health, population and environmental data.
GEOG 557 Advanced Geographic Information Systems	Application of advanced GIS concepts to real-world projects. Will focus on development and implementation of a digital geo-spatial database. The project will be carried from the design phase through completion.
GEOG 656 Spatial Statistics	The analysis of spatial patterns and processes through the use of spatially based statistics.
Math 695	Graduate Thesis

For the PhD program, the following additional courses (after the Master's list) are recommended:

Course	Description
Math 567 Sampling Theory	Random, systematic, stratified, and cluster sampling techniques. Ratio and proportion estimates. Sample size and strata determination.
Math 601 Real Analysis I	Basic set theory and real topology, Lebesgue measure and integration on the real line, differentiation of integrals, $L(p)$ spaces.
Math 602 Real Analysis II	Elementary Halberd space theory, abstract measure spaces and integration, product spaces. Applications to other areas.
Math 681 Combinatorics and Graph Theory I	Fundamental topics in Graph Theory and Combinatorics through Ramsey theory and Polya's theorem respectively. Motivation will be through appropriate applications.
Math 682 Combinatorics and Graph Theory II	Fundamental topics in Graph Theory and Combinatorics through Ramsey theory and Polya's theorem respectively. Motivation will be through appropriate applications.
GEOG 522 GIS and Public Health	Application of tools and methods of analysis in geographic information systems (GIS) to public health. Use of ArcGIS software to manage and analyze health, census and spatial data.
GEOG 583 Spatial and Non-spatial Database Management	Provide students with "hands-on" experience in development, management and integration of spatial and non-spatial databases, using GIS and database management software.

Joint BS/MA with a concentration in data mining:

Course	Description	Number of Hours
Core Requirements	Math 205, 206, 301, 311, 325, 405, 501 or 521	24
BS Option Probability and Statistics Requirements	Math 560, 561, 562, 564	12
Computer Requirements	CECS 121, 230	6
Mathematics Electives	Math 502, 566	6
Geography Applications (Applications Area)	Geog 357, 521, 522, 557, 583	15

With the above courses, students can complete the proposed certificate in data mining. The following schedule is suggested for completing the requirements:

Course	Description	Number of Hours
Year 1	Math 205, 206 CECS 121, 230 16 hours of general education	14
Year 2	Math 301, 311, 325 Geog 357, 521 15 hours of general education, science requirements	15
Year 3	Math 560, 561, 562 Geog 522, 557, 583 12 hours of science and general education	18
Year 4	Math 405, 501-502, 564, 665-667 Thesis	18
Year 5	Math 566, 567, 660-662, 635-536	18

Summary

- Statistics is a marketable and profitable field with a large number of possible fields of specialization.
- The use of GIS in public health has been increasing rapidly in recent years. A background in Medical Geography and knowledge of GIS is in considerable demand.
- There are many opportunities available to analyze geographic health data using data mining tools.
- The Data Mining Applications Area can be a part of the BS/MA, MA, and PhD curricula.