Tradeoff Studies about Storage and Retrieval Efficiency of Boundary Data Representations for LLS, TIGER and DLG Data Structures

David Clutter and Peter Bajcsy
National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign (UIUC)

Presenter:
Peter Bajcsy, Ph.D.
Research Scientist, NCSA
Adjunct Assistant Professor, CS and ECE Departments, UIUC

NCSA

Outline

- Motivation for the Tradeoff Studies
- Goals and Problem Statement
- Boundary Storage Schemes and Their Evaluations
- DLG File Format
- TIGER File Format
- ESRI Shapefile File Format
- Experimental Evaluations
- Conclusions and Future Directions

NCSA

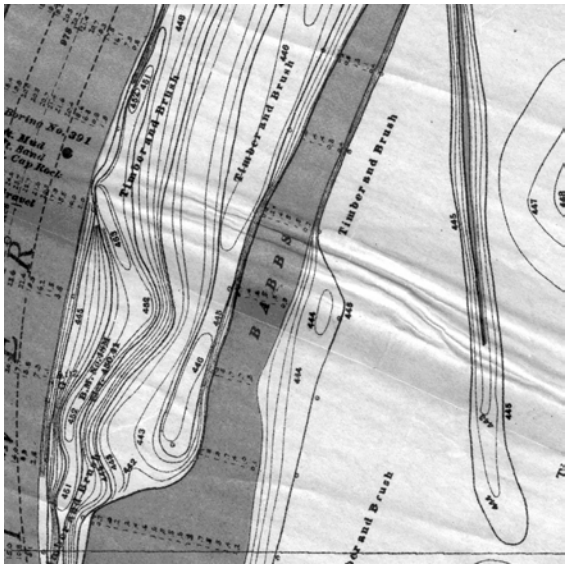
Motivation for the Tradeoff Studies

- **Initial Map Analysis Problem**
  - Processing a large volume of historical maps on hard copy materials
  - Illinois Waste Management and Environmental Protection Agency (EPA)
- **Consequential Information Archival Problem**
  - Preserving and retrieving geo-spatial information extracted from hard copy materials
  - National Archive and Record Administration (NARA), and other institutions providing geo-spatial information

NCSA

Initial Map Analysis Problem

- **Application**
  - Backwater Restoration Opportunities: Illinois River
- **Challenges**
  - Flooded Areas Might Contain Objects That Destroy Expensive Dredging Equipment, e.g., Tree Stumps
  - How Much Silt Should Be Removed to Restore Landscape
- **Approach: Analyze Historical Maps To Recover Information About**
  - Historical Land Cover, e.g. Trees, Shrubs
  - Historical Elevation to Estimate Depth of Deposited Silt Over Time

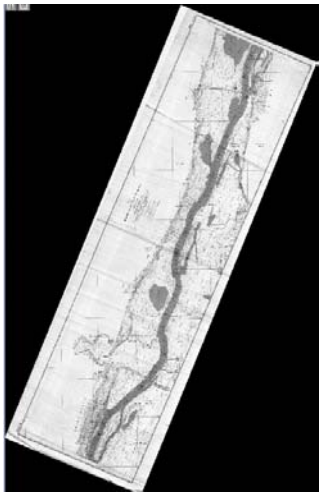


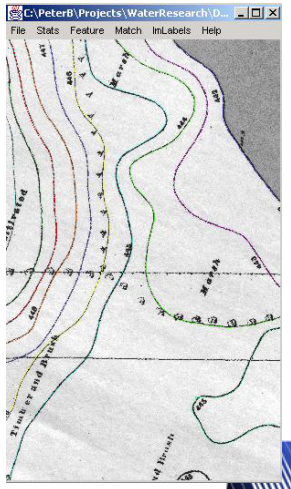
Isocontours are horizontal cross sections of 3D terrain at equal elevation

NCSA

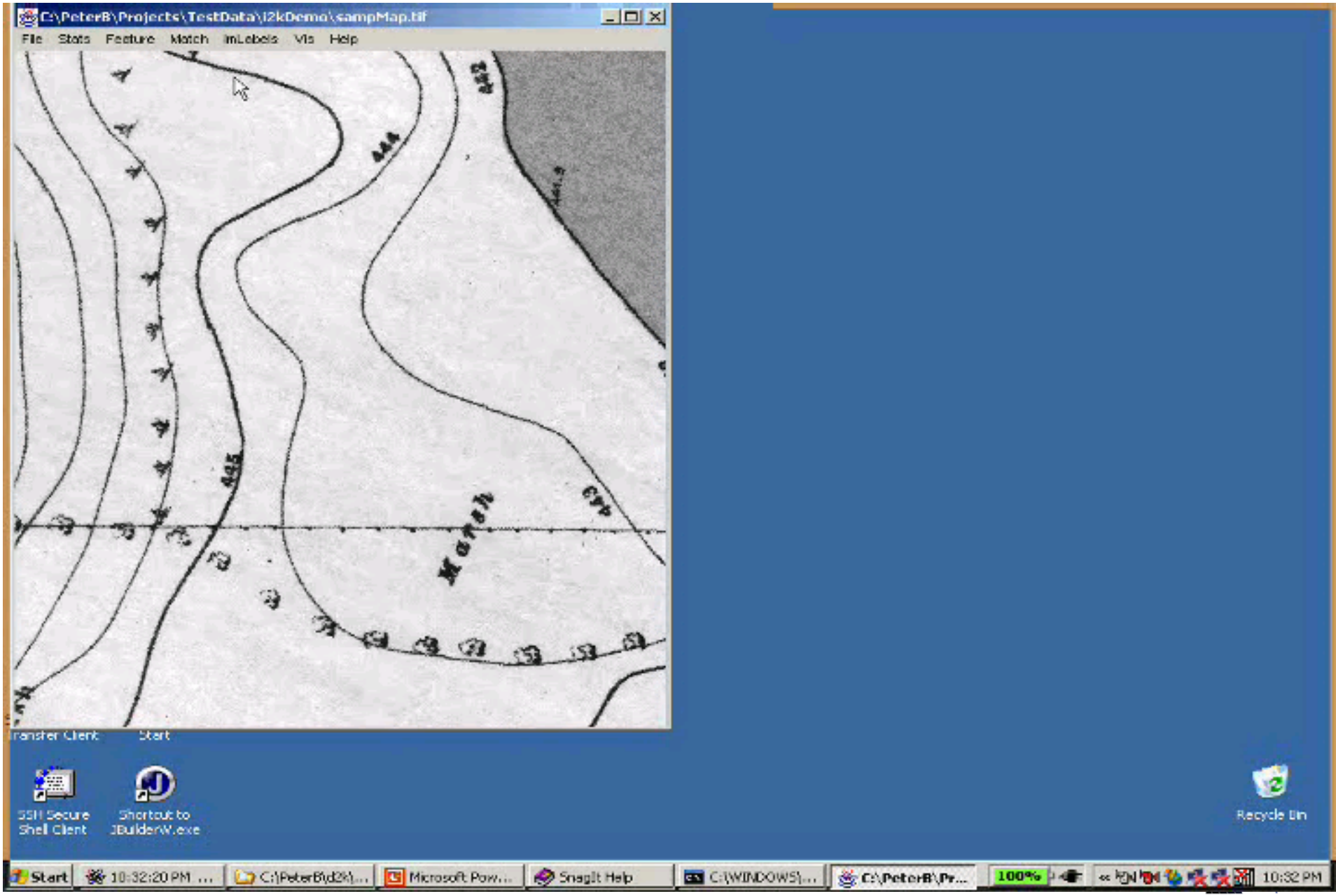
Initial Map Analysis Problem

- **Input Material: Historical Maps from 1900 (Woerman Maps)**
- **Digital Input: Large Images in Tiff File Format (255MB a sheet)**
- **Needed: Software Tools to Extract Information From Maps in Digital Format**
- **Motivation:**
  - There Are No Commercial GIS Software Tools For Automatic Isocontour Extraction
  - Manual Isocontour Tracing in Adobe Photoshop is Laborious and Time-Consuming
- **Publication**
  - Bajcsy P., “Automatic Extraction Of Isocontours From Historical Maps,” Proc. of the 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003), Orlando, Florida, July 27-30, 2003.





NCSA


C:\PeterB\Projects\TestData\i2kDemo\sampMap.tif
File Stats Feature Match imLabels Vis Help
Transfer Client
Start
SSH Secure Shell Client
Shortcut to JBuilderW.exe
Recycle Bin
Start
10:32:20 PM ...
C:\PeterB\i2k\...
Microsoft Pow...
SnagIt Help
C:\WINDOWS\...
C:\PeterB\Pr...
100%
10:32 PM

NCSA

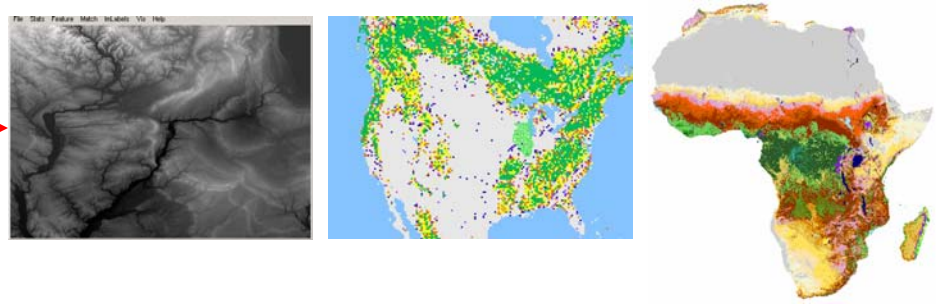
Consequential Information Archival And Retrieval Problem

- **Application:**
  - Information archival and retrieval
- **Challenges**
  - Raster (Image) geo-spatial information:
    - Grid-based information, e.g., from satellite or air-borne sensors, scans of historical maps
  - Point geo-spatial information
    - Geographical point information, e.g., locations of water gauges, oil wells, houses (address database)
  - Boundary geo-spatial information
    - Man-made boundaries, e.g., Counties, US Census Bureau Territories
    - Boundaries defined by environmental characteristics, e.g., eco-regions, soil boundaries or iso-contours
  - Attributes of geo-spatial locations
    - Categorical and continuous attributes, e.g., average elevation per county, forest label type, soil type

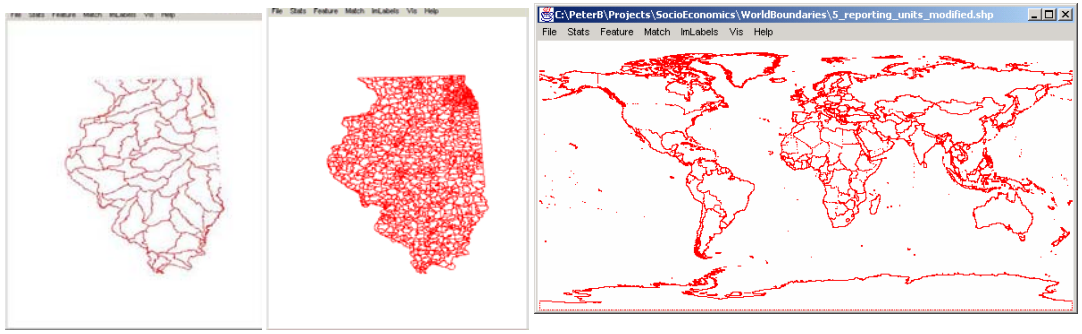
NCSA

Geo-Spatial Information Heterogeneity

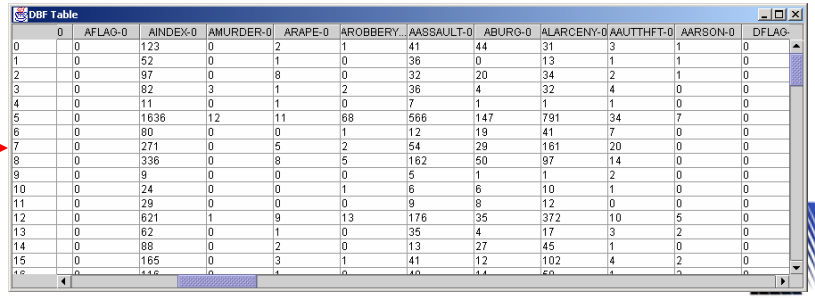
- Raster Information: GeoImage Object



- Boundary Information: Shape Object



- Tabular Information: Table Object
- Neighborhood Information: NBH Object

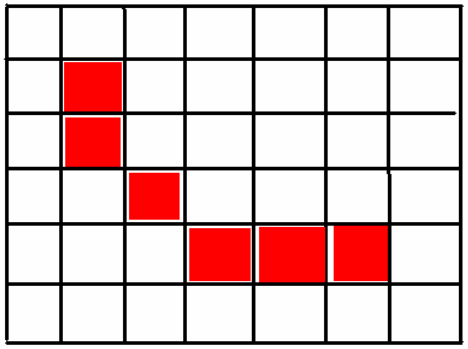


	0	AFLAG-0	AINDEX-0	AMURDER-0	ARAPE-0	AROBBERY...	AASSAULT-0	ABURG-0	ALARCENY-0	AAUTTHFT-0	AARSON-0	DFLAG-
0		0	123	0	2	1	41	44	31	3	1	0
1		0	52	0	1	0	36	0	13	1	1	0
2		0	97	0	8	0	32	20	34	2	1	0
3		0	82	3	1	2	36	4	32	4	0	0
4		0	11	0	1	0	7	1	1	1	0	0
5		0	1636	12	11	68	566	147	791	34	7	0
6		0	80	0	0	1	12	19	41	7	0	0
7		0	271	0	5	2	54	29	161	20	0	0
8		0	336	0	8	5	162	50	97	14	0	0
9		0	9	0	0	0	5	1	1	2	0	0
10		0	24	0	0	1	6	6	10	1	0	0
11		0	29	0	0	0	9	8	12	0	0	0
12		0	621	1	9	13	176	35	372	10	5	0
13		0	62	0	1	0	35	4	17	3	2	0
14		0	88	0	2	0	13	27	45	1	0	0
15		0	165	0	3	1	41	12	102	4	2	0

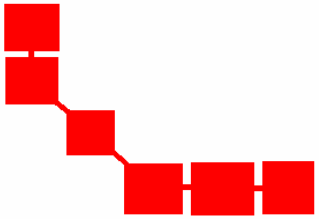
NCSA

Raster & Boundary & Point Data

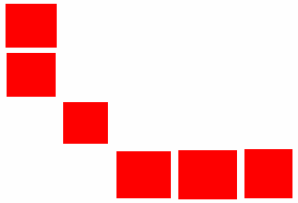
Image



Boundary



Points



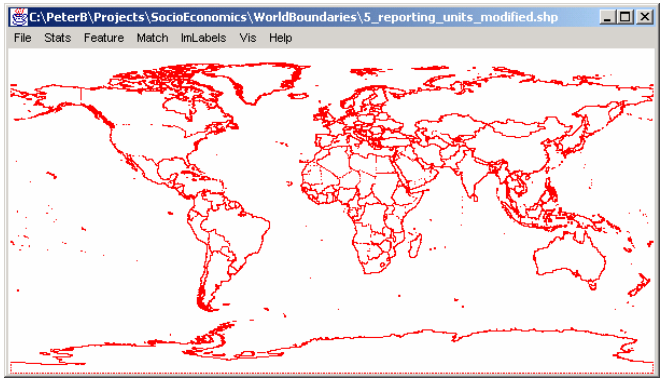
Efficient and Accurate Representation of Geo-spatial Information?

NCSA

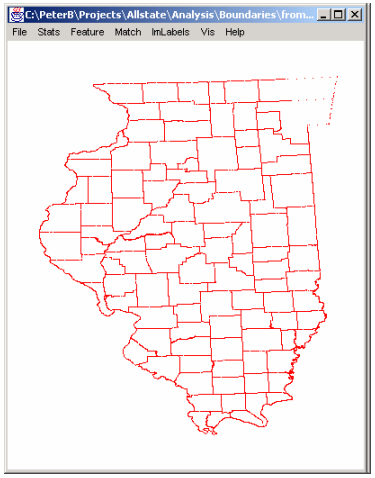
Focus on Boundaries

Boundary Information

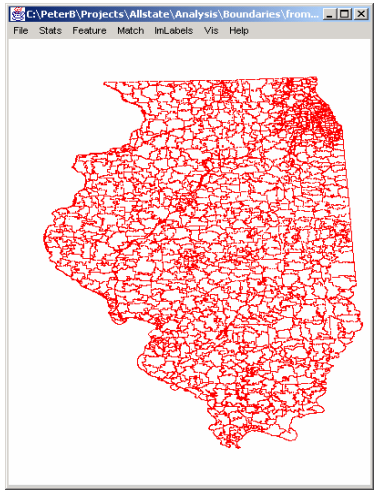
Watersheds



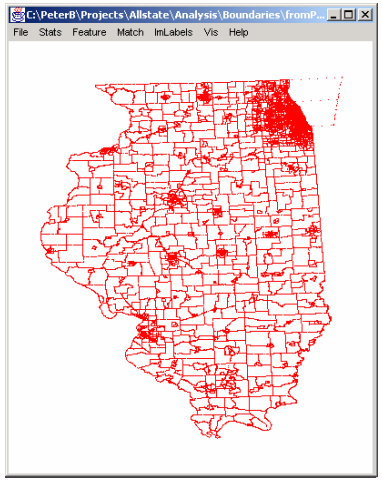
Countries



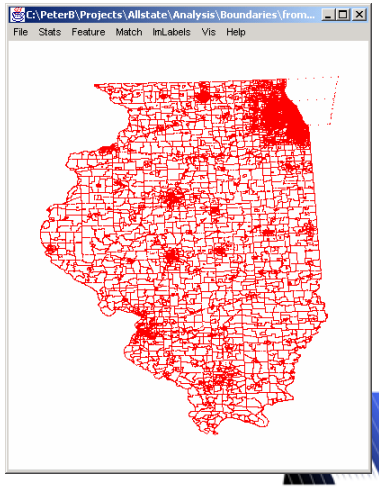
Counties



Zip Codes



Census Tracks



Census Blocks

NCSA

Goals and Motivation

- ***Boundary information is viewed as an efficient representation of image documents describing borders of spatial regions.***

- ***Goals of our work:***
  - *to study the impacts of choosing boundary information representation on document image management and information retrieval*
  - *to improve our understanding of the processing noise introduced during representation conversions.*
- ***Motivation:***
  - *Provide quantitative support for institutional document image management decisions.*

NCSA

Problem Statement

- ***Known:***
  - *Boundary data types are preferred over image data types when it comes to storing boundary information*
  - *There are multiple memory storage schemes and data representation for geo-spatial boundary information*
- ***Unknown:***
  - *How to choose a storage scheme and data representation for boundary information that meets institutional needs of archiving and retrieving geospatial boundary information?*
- ***Challenges***
  - *Choosing the storage scheme that minimizes memory requirements might have a detrimental impact on boundary information retrieval efficiency*
  - *Choosing the storage scheme that maximizes information retrieval efficiency might have a detrimental impact on memory requirements*
  - *Converting data from one representation to another introduces processing noise*

NCSA

Boundary Information Storage Schemes

- ***Most Frequent Storage Schemes***
  - *location list data structure (LLS),*
  - *point dictionary structure (PDS),*
  - *dual independent map encoding structure (DIME),*
  - *chain file structure (CFS),*
  - *digital line graphs (DLGs) and*
  - *topologically integrated geographic encoding and referencing (TIGER) files.*

NCSA

Boundary Information Storage Schemes of Interest

- ***Schemes of Interest***
  - *location list data structure (LLS)*
  - *digital line graphs (DLGs)*
  - *topologically integrated geographic encoding and referencing (TIGER) data organization*
- ***Usage in GIS domain***
  - *ESRI Shapefiles (LLS)*
  - *SSURGO DLG-3 soil files (DLG)*
  - *U.S. Census Bureau 2000 TIGER/Line files (TIGER).*

NCSA

SSURGO DLG File Format Description

NCSA

DLG File Format

- **The Soil Survey Geographic (SSURGO) Digital Line Graphs (DLG) files provide geographical information on the boundaries of soil types**
- **Downloadable files for each county**
  - dlg.zip (digital line graph)
  - tab.zip (ASCII attribute data available in Microsoft Access 97 or later template database).
- **SSURGO Illinois county files contain soil boundaries of 18,000 soil series recognized in the United States.**
- **Information from the DLG-3 documentation:**
  - (a) file naming convention,
  - (b) spatial resolution,
  - (c) spatial accuracy,
  - (d) geographic coordinate system and
  - (e) storage format.

NCSA

Spatial Accuracy

- **Spatial resolution of DLG files**
  - Large-scale DLG data is digitized from 1:24,000-scale USGS topographic quadrangles (Soil Survey Geographic = SSURGO).
  - Intermediate-scale DLG data is digitized from 1:100,000-scale USGS quadrangles (State Soil Geographic = STATSGO).
  - Small-scale DLG data is digitized from 1:2,000,000-scale sectional maps (National Soil Geographic = NATSGO).
- **DLG Levels of details (number of attributes)**
  - Highest number of attributes to be encoded (Level 3).

NCSA

DLG Georeferencing Information

- **SSURGO DLG-3 data**
  - normally reported in the Universal Transverse Mercator (UTM) system.
- **NATSGO & STATSGO DLG data**
  - reported using the Albers Equal-Area Conic projection.
- **3D Model: North American Datum of 1983 reference system**
  - based upon the Geodetic Reference System of 1980.

NCSA

DLG Data Description

- **DLG data are reported as nodes, lines, and areas.**
  - Lines are composed of a series of nodes
  - Areas are composed of lists of lines (or optionally nodes).
- **A node is a coordinate on a map. Nodes define the points of each line and are encoded with (1) a unique identifier and (2) the coordinates that the node represents.**
- **Lines are a series of nodes. Each line is encoded with a unique identifier, as well as its starting node and ending node.**
- **An area is an enclosed section. Areas can be encoded as either a sequence of lines or a sequence of nodes. Areas are specified in a clockwise direction around the perimeter of the area, and islands are specified in a counter-clockwise direction.**

NCSA

Software Development for SSURGO DLG-3 Files

- **Software Functionality**
  - Loader for SSURGO DLG-3 files
  - 2D visualization of SSURGO DLG-3 files
  - Conversion function from SSURGO DLG-3 data structure to ESRI Shapefile (LLS) data structure
- **Boundary information retrieval from DLG-3 file format**
  - Read all the defined lines first. The lines are kept in a lookup table, and indexed by their unique identifier for later use. The size of this structure is directly proportional to the number of lines.
  - Areas are retrieved by populating our internal ShapeObject data structure for boundary information. In the ShapeObject, an area has a list of the coordinates that make up its boundary. This list is dynamically constructed when reading an area. Areas that share a boundary will have copies of the common coordinates.
  - Once all areas have been read and processed, the lookup table containing the lines can be safely discarded. The coordinates for the areas are copied into a ShapeObject.

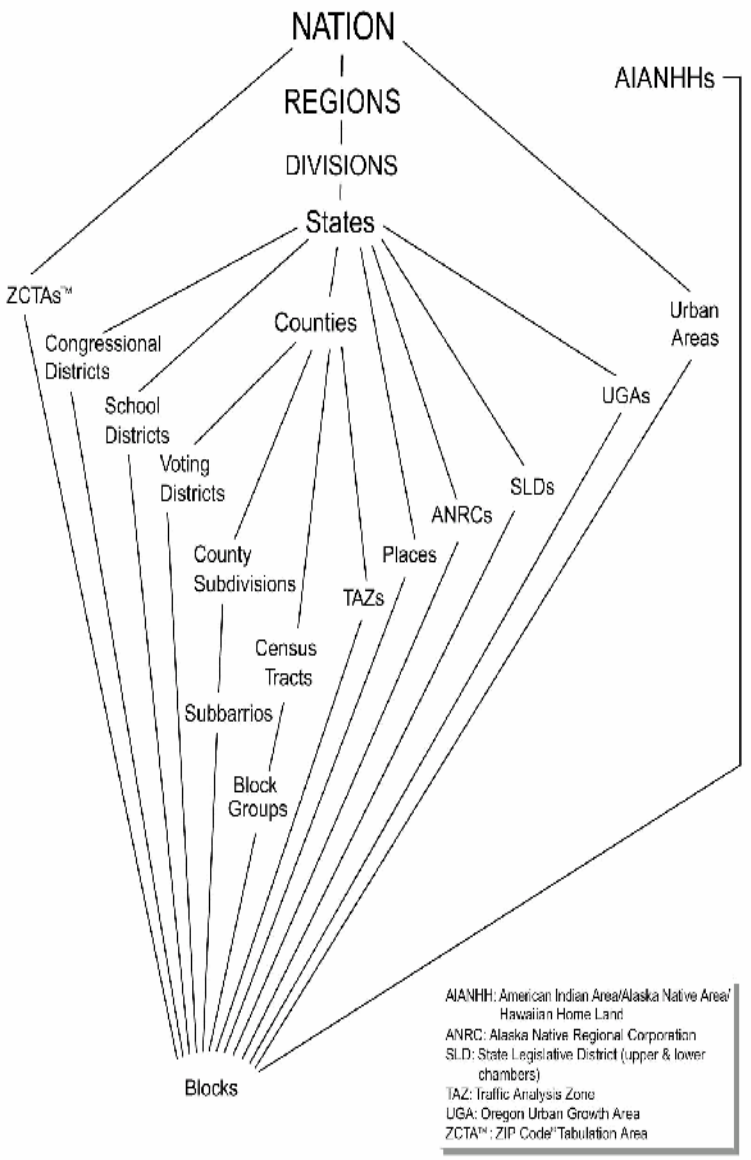
NCSA

Census 2000 TIGER/Line File Format Description

NCSA

Census 2000 TIGER/Line Files

- **The TIGER files provide geographical information on the boundaries of counties, zip codes, voting districts, and a geographic hierarchy of census relevant territories, e.g., census tracts that are composed of block groups, which are in turn composed of blocks.**
- **It also contains information on roads, rivers, landmarks, airports, etc, including both latitude/longitude coordinates and corresponding addresses.**



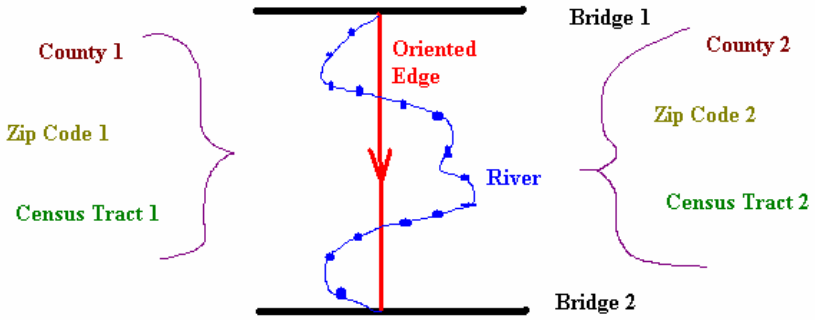
US Census Bureau TIGER Files

- **Downloadable files for each county**
  - Each type of GIS information is self-contained in a subset of files.
  - As a result users can process only the desired information by loading a selected subset of relevant files.
  - Each primary region (county) is fully represented by a maximum of 17 files.
- **Files relevant to boundary point extraction:**
  - Record Type 1: Edge ID (TLID), Lat/Long of End Points
  - Record Type 2: TLID, Shape Points
  - Record Type I: TLID, Polygon ID Left, Polygon ID Right
  - Record Type S: Polygon ID, Zip Code, County, Census Tract, Block Group, etc.
  - Record Type P: Polygon ID, Internal Point (Lat/Long).

NCSA

TIGER Data Description

- TIGER/Line files are based on an elaboration of the chain file structure (CFS), where the primary element of information is an edge.
- Each edge has a unique ID number (TIGER/Line ID or TLID) and is defined by two end points.
- Each edge then has polygons associated with its left and right sides, which in turn are associated with a county, zip code, census tract, etc.
- The edge is also associated with a set of shape points, which provide the actual form an edge takes.



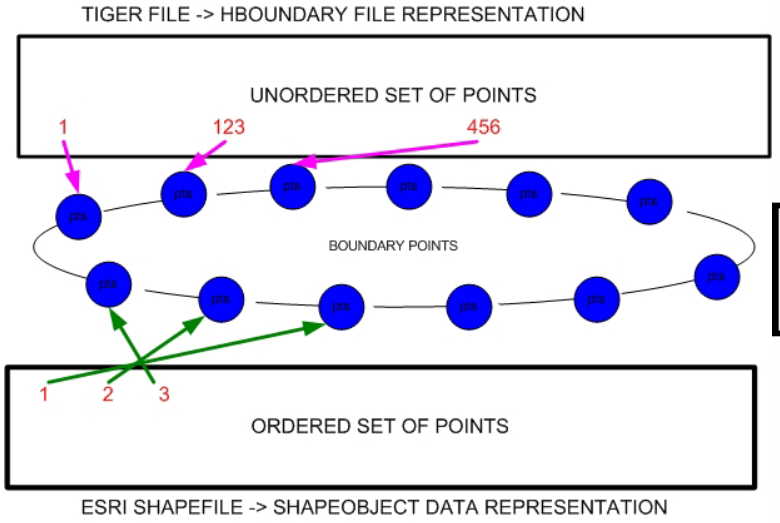
NCSA

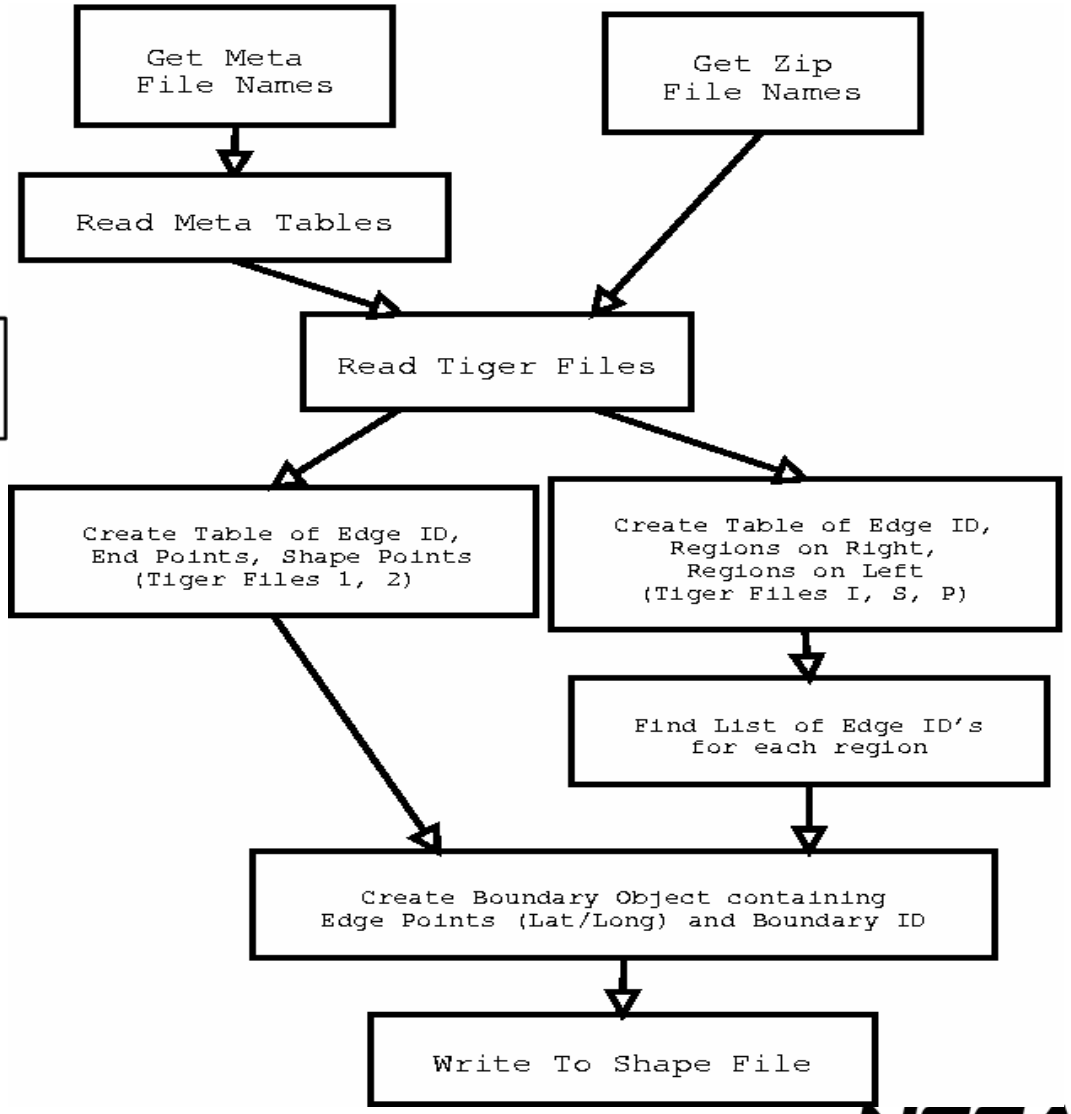
Software Development for TIGER Files

- **Software Functionality**
  - Loader for TIGER files
  - Conversion function from TIGER data structure to ESRI Shapefile (LLS) data structure
  - 2D visualization of TIGER files
- **Boundary information retrieval from TIGER file format**
  - One master list of boundary points that all boundaries reference by pointers (HierarchicalBoundaryObject data structure)
  - The memory savings for each point that is shared by two county, two census tract, and two block group boundaries is 30 bytes. For the state of Illinois, this optimization translated into a 38% reduction in memory usage (16.45 MB versus 26.64 MB)

NCSA

TIGER To ESRI (LLS) Shapefile





NCSA

ESRI Shapefile (LLS) File Format Description

NCSA

ESRI Shape Files

- **ESRI shape files: Environmental Systems Research Institute shape files for vector data**
- **File Configuration:**
  - Main file (.shp): records describing a shape with a list of its vertices
  - Index file (.shx): records describing the offset of the corresponding main file record from the beginning of the main file
  - dBASE table (.dbf): columns with features describing each record in the main file
- **Software for visualization**
  - GIS software, e.g., ArcExplorer (freeware)
  - Excel Spread Sheet for .dbf file

NCSA

ESRI Shapefiles

- **Shapefile Primitives**
  - Arc, Point, PolylineZ, Polygon, ArcM,…
  - Polygons Might Contain Holes
- **Examples**
  - Skyscraper has its own zip code
  - Lake is excluded

NCSA

Software Development for ESRI Files

- **Software Functionality**
  - Loader for ESRI (LLS) files
  - 2D visualization of ESRI (LLS) files
  - Writer for ESRI (LLS) files
- **Boundary information retrieval from ESRI (LLS) file format**
  - Bounding box
  - A sorted list of points forming each boundary

NCSA

Tradeoff Evaluations

NCSA

Evaluations

- **Goals:**
  - to experimentally evaluate the tradeoffs between storage and retrieval efficiency, and
  - to explain the tradeoffs by comparing fundamental format differences.
- ***Tradeoff evaluations***
  - *Theoretical evaluations*
  - *Experimental evaluations*
- ***Evaluation Metrics of Storage and Retrieval Efficiency***
  - *load time*
  - *computer memory*
  - *hard disk space requirements*
- **Comparisons:**
  - DLG & LLS,
  - DLG & TIGER & LLS

NCSA

Test Data

- ***Ideal test data would contain identical boundary information represented by LLS, TIGER and DLG files.***
  - *We were not able to find such files.*
- ***Ideal software tools would convert LLS, TIGER and DLG files from one file format to another with no processing noise.***
  - *LLS formats (ESRI Shapefiles) are supported by most GIS software packages*
  - *There is a very limited support for DLG and TIGER file formats.*
  - *This corresponds to our assessment of the implementation complexity to support loading of TIGER, DLG and LLS formats.*
- ***Test data for experimental evaluations***
  - *SSURGO DLG-3 soil files of* Madison County, IL *for DLG & LLS evaluations*
  - *U.S. Census Bureau 2000 TIGER/Line files* of Illinois counties, zip codes, census block and census tracts *for DLG & LLS & TIGER evaluations*

NCSA

DLG & LLS Evaluation

	Total Load Time (s)		Load RAM Required (MB)	Hard Disk (MB)		Number of Nodes
	Zip	Unzip		Zip	Unzip	
LLS (Shapefile)		41.36	290	65	90	2,787,490
DLG	105.72	103.72	380	23	79	2,787,790

- Test Data Preparation: ***SSURGO DLG-3 soil files of SSURGO Soil Database, Madison County, IL converted to ESRI Shapefile (LLS) format***
- ***Metrics:*** *load time, computer memory, hard disk space requirements*

NCSA

TIGER & LLS & DLG Evaluations

	Total Load Time (s)	Load RAM Required (MB)	Hard Disk (MB)		Number of Nodes
	Unzip		Zip	Unzip	
TIGER	1300.2	200	112*	940 *	2,176,719 *
LLS	12.7	37	27	47	641,955
DLG-3	12.9	52	8	24	457,850

- Test Data Preparation: **TIGER files for the state of Illinois (102 counties) converted to LLS with our software and to DLG with ArcToolbox.**
- ***Metrics:*** **Loading is constrained to block groups, zcta, census tract, and counties. Hard Disk and Number of Nodes measurements for LLS and DLG formats contain only block groups, zcta, census tract, and county boundaries, whereas the same measurements for TIGER format include all types of boundary information for the state of Illinois.**

NCSA

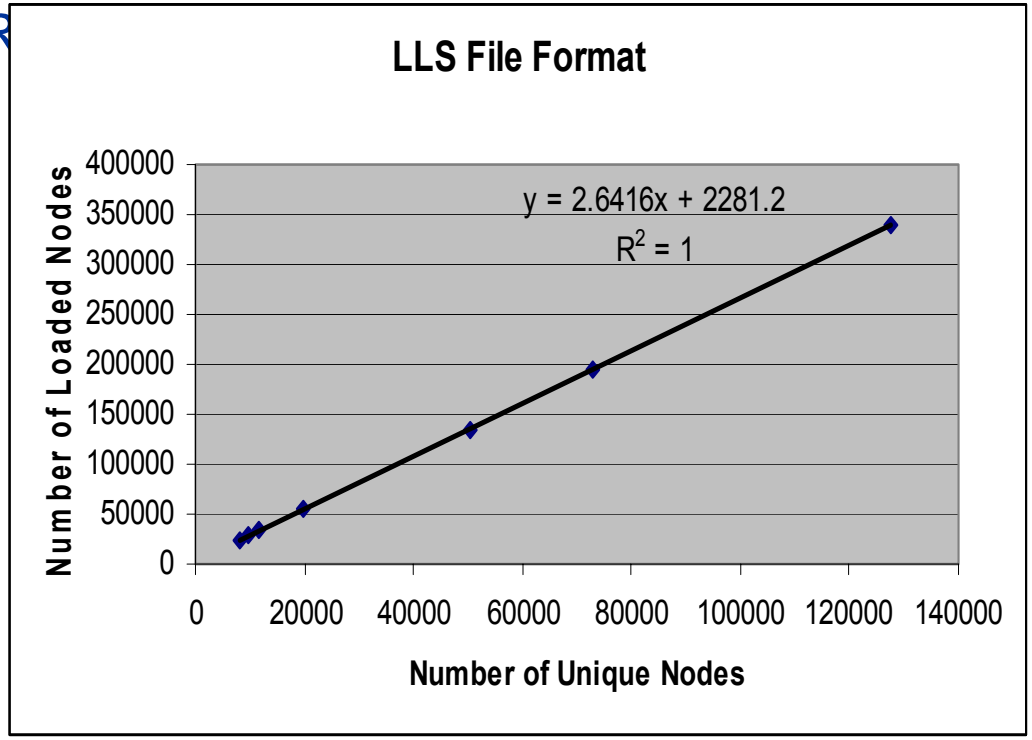
Storage Dependency on Boundary Content?

- **The answer to this question is related to the amount of boundary overlap.**
  - Ideally, one would experiment with sets of boundaries that span cases from a zero overlap (e.g., non-adjacent county boundaries) to an overlapping hierarchy of polygons (census blocks, block groups and tracts).
  - Our data sets represent the cases of partial overlap (SSURGO) and large overlap (TIGER) of boundaries.
- **Experimental results vary as a function of boundary content**
  - the more overlapping boundaries, the smaller hard disk requirements for TIGER format in comparison with DLG and LLS (in this order), and the smaller load RAM requirements for LLS format in comparison with DLG and TIGER.

NCSA

Boundary Overlap

- **Test Data:**
  - Sub-sets of the original TIGER files for the state of Illinois in order to vary the amount of boundary overlap and the number of nodes.
  - A subset is formed by selecting 1, 2, 3, 4, 10, 15, or 24 counties from the original TIGER files, and forming several triplets of test data sets (TIGER, LLS and DLG).
  - Any selected subset of counties always forms a geographically contiguous region so that neighboring counties would have some overlap of boundary points.



LLS Storage Dependency on Boundary Content (Overlap)

NCSA

Boundary Information Retrieval Dependency on Number of Nodes?

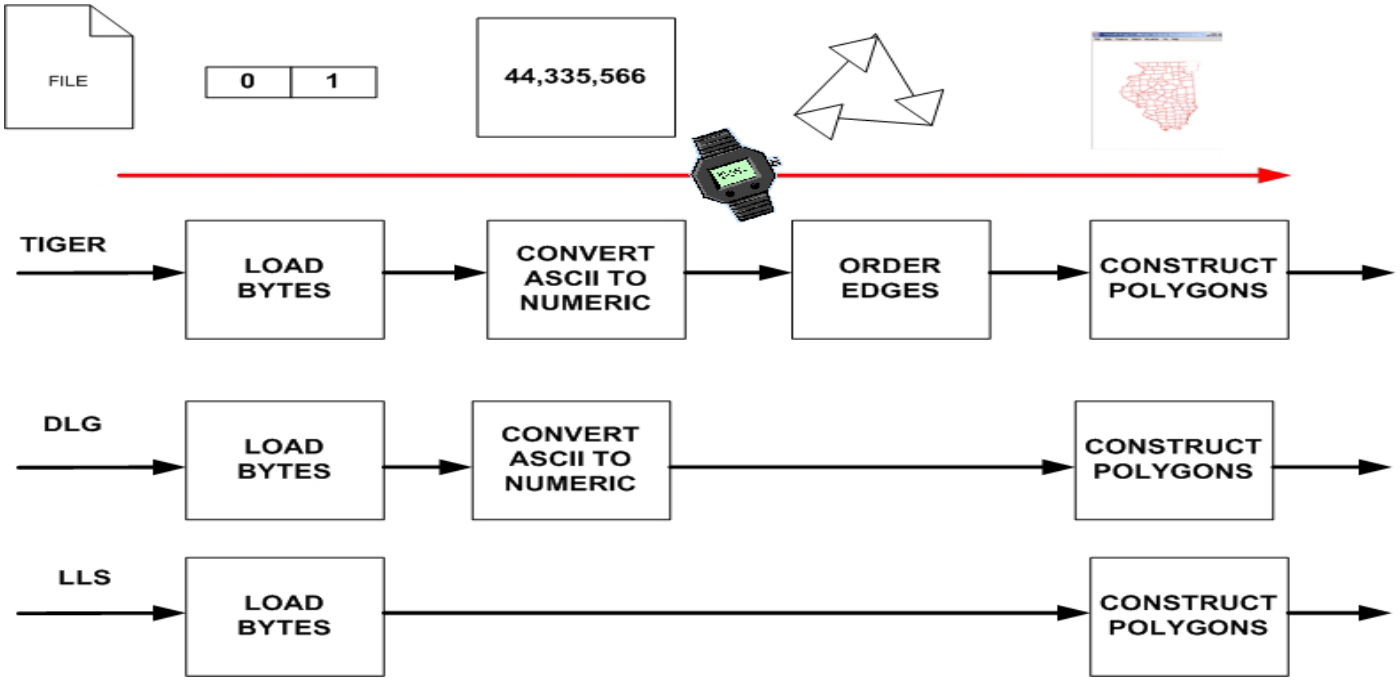
- **Can we predict the total load time as a function of the number of polygons/nodes without exhaustive experimentation? Or in other words, what would be the dependency between boundary information retrieval and the number of retrieved nodes?**
- **Total Load Time is divided into four components:**
  - t1 corresponds to the time to construct polygons from an ordered list of edges.
  - t2 is for the time to create an ordered list of edges from an unordered set of edges
  - t3 represents the time to convert ASCII characters to numeric type values
  - t4 is the time to load any sequence of bytes (ASCII characters or binary values) from a file

$$Total\ Load\ Time = t_1 + t_2 + t_3 + t_4$$

NCSA

Boundary Information Retrieval

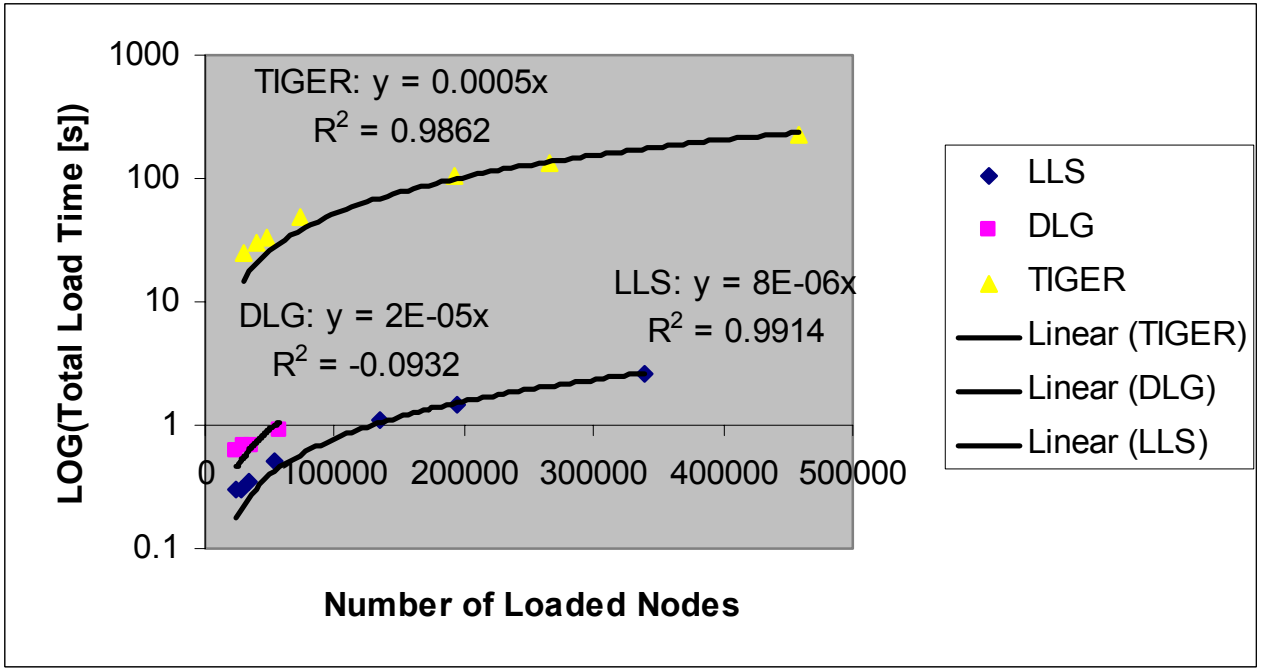
Total Load Time=Sum(t_i)	t_1	t_2	t_3	t_4
LLS	X	0	0	X
DLG	X	0	X	X
TIGER	X	X	X	X

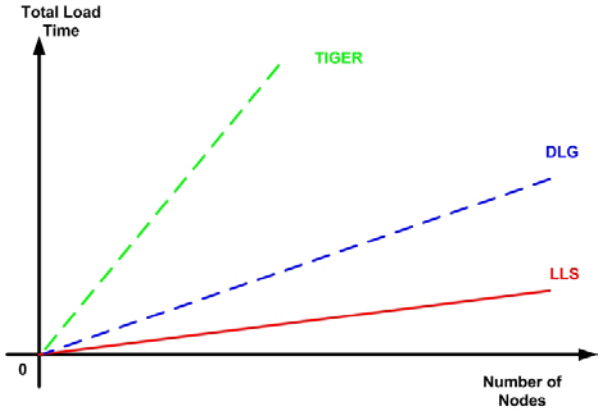


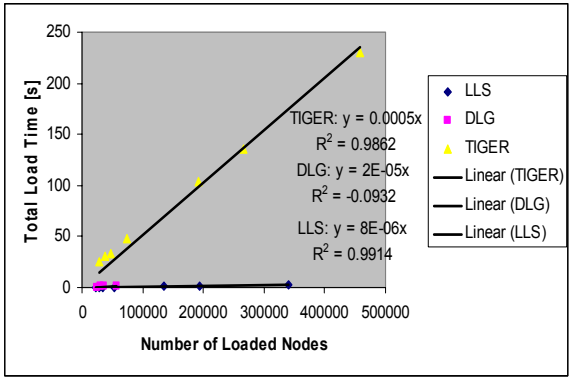
NCSA

Boundary Information Retrieval Comparisons

- **Total Load Time vs. Number of Nodes for 1, 2, 3, 4, 10, 15, and 24 counties with a best-fit line.**







NCSA

Pair-wise Comparisons

DLG and LLS Comparisons

- **DLG: ASCII**
- **DLG format uses nodes, lines, and areas to define its polygons**
- **All coordinates in SSURGO DLG files are stored in UTM format => conversion.**
- **SSURGO DLG files are stored as quarter-quadrangles (7.5 minutes of a degree of longitude and latitude).**
  - It is necessary to load 64 individual files to represent a one degree block.

- **LLS: Binary**
- **LLS format lists the bounding box and the points for each boundary contained within it.**
- **LLS normally contains latitude and longitudinal geo-referencing information.**

NCSA

TIGER and DLG Comparisons

- TIGER's use of an edge with shape points
- TIGER polygon is comprised of a series of edges
- TIGER format groups all edges together regardless of layer. The different metadata files are used to determine which edges to use.
- TIGER boundaries must be found programmatically. Each edge is labeled with the polygons that appear on the left and right of the edge.

- DLG's use of lines and coordinates
- DLG area is made up of a series of lines.
- DLG format typically encodes one layer of data in a file (some redundancy between the layers.
- DLG format specifies the exact boundaries for each polygon

NCSA

Conclusions

- ***Quantitative Support for Institutional Document Image Management Decisions***
- ***Conclusions***
  - *LLS files will provide the fastest boundary retrieval at the price of file size*
    - *Retrieval from LLS files is 40 times faster than from TIGER files and 2.5 times faster than from DLG files*
    - *Storage redundancy for LLS files is between 70% and 180% in our experiments*
  - *DLG format offers a smaller file size, but is less efficient for boundary retrieval.*
  - *TIGER format also offers a compact physical representation, at the cost of more processing for boundary retrievals.*

NCSA

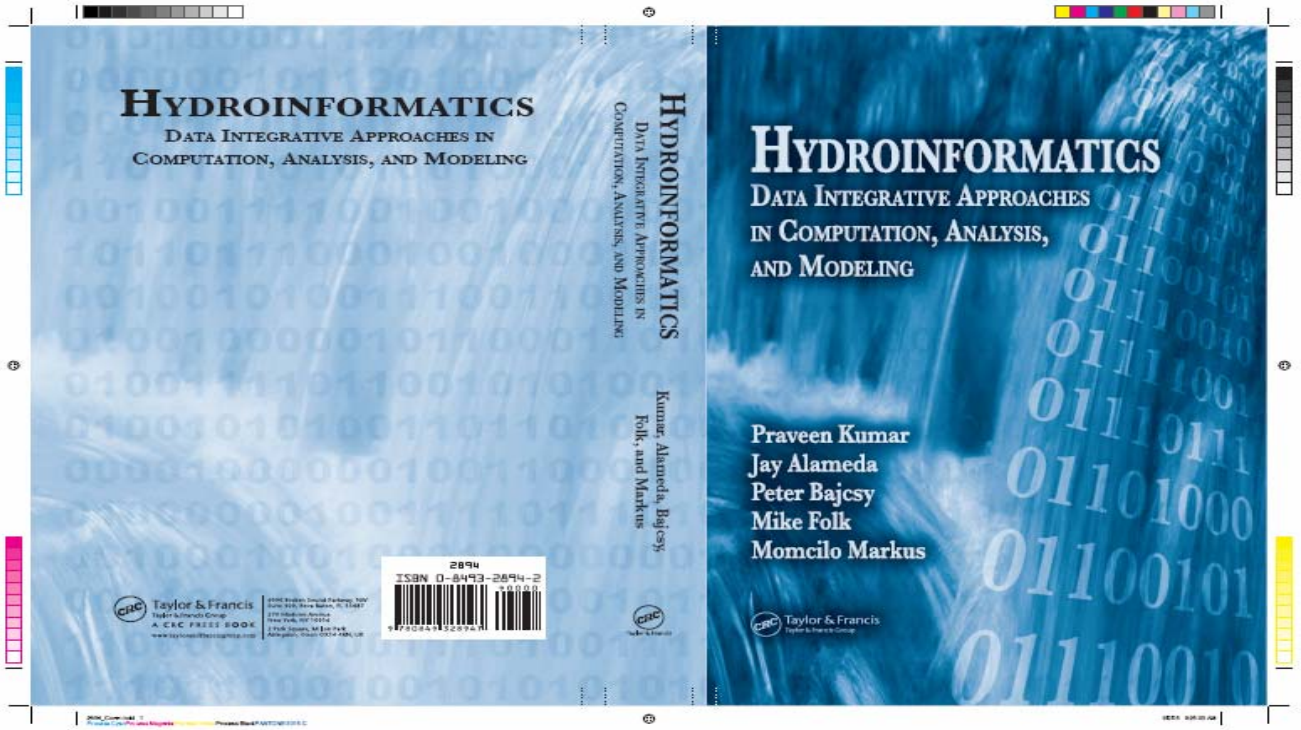
Future Directions

- **Investigate the effect of computer clusters on boundary retrieval efficiency assuming distributed or centralized locations of a large number of boundary files.**
- **Answer questions about mass storage and IO:**
  - Can more efficient I/O schemes be used to improve boundary retrieval?
  - Would message passing interface input/output (MPI-IO) have any effect?
  - What would be the bottlenecks?
- **Understand multiple effects of electronic vector files on the archival process.**
  - vector file format, data organization and representation, algorithmic parallelization, scalability of vector file loading in terms file size and centralized or distributed file locations, software re-usability, computer platform dependency, computer cluster environments, I/O bandwidth and I/O schemes, and mass storage systems.

NCSA

More Information

- **Questions:**
  - Peter Bajcsy
  - email: pbajcsy@ncsa.uiuc.edu



NCSA