10. Creating and Maintaining Geographic Databases

Geographic Information Systems and Science
SECOND EDITION
Paul A. Longley, Michael F. Goodchild, David J. Maguire, David W. Rhind

Outline – Chapter 10, Longley et al.

- Definitions
- Characteristics of DBMS
  - Types of database
  - Geographic Extensions
- Storing data in DBMS tables
- SQL
- Geographic databases types and functions
- Geographic database design
- Structuring geographic Information
  - Topology
  - Indexing methods
- Editing and Data Maintenance
- Multi-user Editing
- Conclusions

Learning objectives

- **By the end of this section students should:**
  - Understand the role of database management systems in GIS;
  - Recognize structured query language (SQL) statements;
  - Understand the key geographic database data types and functions;
  - Be familiar with the stages of geographic database design;
  - Understand the key techniques for structuring geographic information, specifically creating topology and indexing;
  - Understand the issues associated with multi-user editing and versioning.

Keywords and concepts

- DBMS; (RDBMS, ODBMS, ORDBMS);
- parsers; middleware;
- object classes;
- database design; database tables; keys; normal forms;
- SQL; SQL/MM;
- indexes; B-tree indexes; grid indexes; quadtree indexes; R-tree indexes;
- database editing and update; transactions; long transactions; versioning

Overview

- After people, the database is the most important part of a GIS
  - Costs of collection and maintenance
  - Basis of all queries, analysis, and decision making.
- Virtually all large GIS implementations store data in a database management system (DBMS)
  - software designed to handle multi-user access to an integrated set of data.
- Databases need to be designed with great care
  - structured and indexed to provide efficient query and transaction performance.
- Comprehensive security and transactional access
  - to ensure that multiple users can access the database at the same time.
- On-going maintenance is also an essential, but very resource-intensive, activity.

Definitions

- Database – an integrated set of data on a particular subject
- Geographic (=spatial) database - database containing geographic data of a particular subject for a particular area
- Database Management System (DBMS) – software to create, maintain and access databases

10.2 Database management systems

- A DBMS is a software application designed to organize the efficient and effective storage and access of data
  - Separates the physical form of the data from any applications
  - Applications may be developed without changing the data structure
- Capabilities of DBMS:
  - include a data model, a data load capability, indexes, a query language, security, controlled update, backup and recovery, database administration tools, applications and APIs.
- This list of DBMS capabilities is very attractive to GIS users and so, not surprisingly, virtually all large GIS databases are based on DBMS technology.

Advantages of Databases over Files

- Avoids redundancy and duplication
- Reduces data maintenance costs
- Applications are separated from the data
  - Applications persist over time
  - Support multiple concurrent applications
- Better data sharing
- Security and standards can be defined and enforced

Disadvantages of Databases over Files

- Expense
- Complexity
- Performance – especially complex data types
- Integration with other systems can be difficult

Types of DBMS Model

- Hierarchical
- Network
- Relational - RDBMS
- Object-oriented - OODBMS
- Object-relational - ORDBMS

10.2.1 Types of DBMS

- Three main types of DBMS are available to GIS users today:
  - Relational (RDBMS), Object (ODBMS),and Object-relational (ORDBMS).
- A relational database comprises a set of tables, each a two-dimensional list (or array) of records containing attributes about the objects under study.
- Object database management systems (ODBMS) were initially designed to address the weaknesses of RDBMS, including the inability to store complete objects directly in the database (both object state and behavior) and poor performance for many types of geographic query.

Hybrid object-relational DBMS

- ODBMS have not proven to be as commercially successful as some predicted because of the massive installed base of RDBMS. Thus appeared…
- Hybrid object-relational DBMS (ORDBMS). These can be thought of as an RDBMS engine with an extensibility framework for handling objects.
- The ideal geographic ORDBMS is one that has been extended to support geographic object types and functions through the addition of a geographic query parser, a geographic query optimizer, a geographic query language, multidimensional indexing services, storage management for large files, long transaction services, replication services.

Geographic DBMS extensions

- Two of the commercial DBMS vendors have released spatial database extensions to their standard ORDBMS products:
- **IBM –**
  - **DB2 Spatial Extender and Informix Spatial Datablade**
- **ORACLE - Oracle Spatial**
  - Neither is a complete GIS software
- Focus is on data storage retrieval and management;
  - provides basic functions to store, manage, and query geographic objects;
  - no real capabilities for geographic editing, mapping, and analysis.

- **Geographic middleware extensions**
- An alternative to extending the DBMS software kernel to manage geographic data is to buildsupport for spatial data types and functions into a middle-tier (or middleware) application server.
  - ESRI's Spatial Data Engine (SDE)

Characteristics of DBMS (1)

- Data model support for multiple data types
  - e.g MS Access: Text, Memo, Number, Date/Time, Currency, AutoNumber, Yes/No, OLE Object, Hyperlink, Lookup Wizard
- Load data from files, databases and other applications
- Index for rapid retrieval

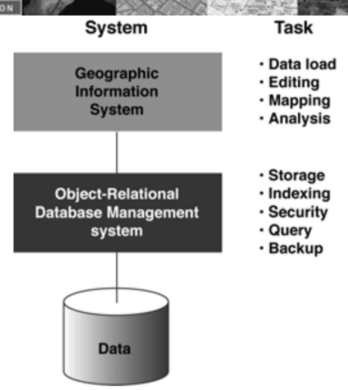
Characteristics of DBMS (2)

- Query language – SQL
- Security – controlled access to data
  - Multi-level groups
- Controlled update using a transaction manager
- Backup and recovery
- DBA tools
  - Configuration, tuning

Characteristics of DBMS (3)

- Applications
  - CASE tools
  - Forms builder
  - Reportwriter
  - Internet Application Server
- Programmable API

Role of DBMS



Relational DBMS (1)

- Data stored as tuples (tup-el), conceptualized as tables
- Table – data about a class of objects
  - Two-dimensional list (array)
  - Rows = objects
  - Columns = object states (properties, attributes)

Relational DBMS (2)

- Most popular type of DBMS
  - Over 95% of data in DBMS is in RDBMS
- Commercial systems
  - IBM DB2
  - Informix
  - Microsoft Access
  - Microsoft SQL Server
  - Oracle
  - Sybase

Table

Column = property

Table = Object Class

Row = object

Object Classes with Geometry called Feature Classes

Attributes of STATES

FID	Shape*	AREA	STATE_NAME	STATE_FIPS
41	Polygon	51715.656	Alabama	01
49	Polygon	576556.687	Alaska	02
35	Polygon	113711.523	Arizona	04
45	Polygon	52912.797	Arkansas	05
23	Polygon	157774.187	California	06
30	Polygon	104099.109	Colorado	08
17	Polygon	4976.434	Connecticut	09
27	Polygon	2054.506	Delaware	10
26	Polygon	66.063	District of Columbia	11
47	Polygon	55815.051	Florida	12
43	Polygon	58629.195	Georgia	13
48	Polygon	6381.435	Hawaii	15
7	Polygon	83340.594	Idaho	16
25	Polygon	56297.953	Illinois	17
20	Polygon	36399.516	Indiana	18
12	Polygon	56257.219	Iowa	19
32	Polygon	82195.437	Kansas	20
31	Polygon	40318.777	Kentucky	21
46	Polygon	45835.898	Louisiana	22
2	Polygon	32161.664	Maine	23
29	Polygon	9739.753	Maryland	24
13	Polygon	8172.482	Massachusetts	25
50	Polygon	57898.367	Michigan	26
9	Polygon	84517.469	Minnesota	27
42	Polygon	47618.723	Mississippi	28
34	Polygon	69831.625	Missouri	29
1	Polygon	147236.031	Montana	30

Record: 0 Show: All Selected Records (0 out of 51 Selected.)

Figure 10.2b

Attributes of POPULATION

STATE_FIPS	SUB_REGION	STATE_ABBR	POP1990	POP1996
53	Pacific	WA	4866692	5629613
30	Mtn	MT	799065	885762
23	N Eng	ME	1227928	1254465
38	W N Cen	ND	638800	633534
46	W N Cen	SD	696004	721374
56	Mtn	WY	453588	487142
55	E N Cen	WI	4891769	5144123
16	Mtn	ID	1006749	1201327
50	N Eng	VT	562758	587726
27	W N Cen	MN	4375099	4639933
41	Pacific	OR	2842321	3203820
33	N Eng	NH	1109252	1156932
19	W N Cen	IA	2776755	2831890
25	N Eng	MA	6016425	6066573
31	W N Cen	NE	1578385	1622272
36	Mid Atl	NY	17990455	18293435
42	Mid Atl	PA	11881643	12077607
09	N Eng	CT	3287116	3287604
44	N Eng	RI	1003464	993306
34	Mid Atl	NJ	7730188	7956917
18	E N Cen	IN	5544159	5801023
32	Mtn	NV	1201833	1532295
49	Mtn	UT	1722850	2000630
06	Pacific	CA	29760021	32218713
39	E N Cen	OH	10847115	11123416
17	E N Cen	IL	11430602	11731783
11	S Atl	DC	606900	550076

Record: 0 Show: All Selected Records (0 out of 51 Sele

Attributes of COMBINED STATES and POPULATION

FID	Shape*	AREA	STATE_NAME	STATE_FIPS	SUB_REGION	STATE_ABBR	POP1990	POP1996
0	Polygon	67286.875	Washington	53	Pacific	WA	4866692	5629613
1	Polygon	147236.031	Montana	30	Mtn	MT	799065	885762
2	Polygon	32161.664	Maine	23	N Eng	ME	1227928	1254465
3	Polygon	70810.156	North Dakota	38	W N Cen	ND	638800	633534
4	Polygon	77193.625	South Dakota	46	W N Cen	SD	696004	721374
5	Polygon	97799.492	Wyoming	56	Mtn	WY	453588	487142
6	Polygon	56088.066	Wisconsin	55	E N Cen	WI	4891769	5144123
7	Polygon	83340.594	Idaho	16	Mtn	ID	1006749	1201327
8	Polygon	9603.218	Vermont	50	N Eng	VT	562758	587726
9	Polygon	84517.469	Minnesota	27	W N Cen	MN	4375099	4639933
10	Polygon	97070.750	Oregon	41	Pacific	OR	2842321	3203820
11	Polygon	9259.514	New Hampshire	33	N Eng	NH	1109252	1156932
12	Polygon	56257.219	Iowa	19	W N Cen	IA	2776755	2831890
13	Polygon	8172.482	Massachusetts	25	N Eng	MA	6016425	6066573
14	Polygon	77328.336	Nebraska	31	W N Cen	NE	1578385	1622272
15	Polygon	48560.578	New York	36	Mid Atl	NY	17990455	18293435
16	Polygon	45359.238	Pennsylvania	42	Mid Atl	PA	11881643	12077607
17	Polygon	4976.434	Connecticut	09	N Eng	CT	3287116	3287604
18	Polygon	1044.850	Rhode Island	44	N Eng	RI	1003464	993306
19	Polygon	7507.302	New Jersey	34	Mid Atl	NJ	7730188	7956917
20	Polygon	36399.516	Indiana	18	E N Cen	IN	5544159	5801023
21	Polygon	110667.297	Nevada	32	Mtn	NV	1201833	1532295
22	Polygon	84870.187	Utah	49	Mtn	UT	1722850	2000630
23	Polygon	157774.187	California	06	Pacific	CA	29760021	32218713
24	Polygon	41192.863	Ohio	39	E N Cen	OH	10847115	11123416
25	Polygon	56297.953	Illinois	17	E N Cen	IL	11430602	11731783
26	Polygon	66.063	District of Columbia	11	S Atl	DC	606900	550076
27	Polygon	2054.506	Delaware	10	S Atl	DE	666168	724890

Record: 0 Show: All Selected Records (0 out of 51 Selected.) Options

Figure 10.2c

Figure 10.3A

ParcelNumb	OwnerNam	OwnerAddress	PostalCode	ZoningCode	ZoningType	Date / AssessedValue
673/100	Jeff Peters	10 Railway Cuttings	114390	2	Residential	2002 220000
673-101	Joel Campbell	1115 Center Place	114390	2	Residential	2003 545500
674-100	Dave Widseler		114391	3	Commercial	99 249000
674-100		452 Diamond Plaza	114391	3	Commercial	2000 275500
674 100	D Widseler	452 Diamond Plaza	114391	3	Commercial	2001 290000
670-231	Sam Camarata	19 Big Bend Bld	114391	2	Residential	2004 450575
674-112	Chris Capelli	Hastings Barracks	114392	2	Residential	2004 350000
674-113	Sheila Sullivan	10034 Endin Mansions	114391	2	Residential	02 1005425

Figure 10.3B

Attributes of Tab10_2

OBJECTID*	ParcelNumb	OwnerNam	OwnerAddress	PostalCode	ZoningCode	ZoningType	DateAssessed	AssessedValue
1	673-100	Jeff Peters	10 Railway Cuttings	114390	2	Residential	2002	220000
2	673-101	Joel Campbell	1115 Center Place	114390	2	Residential	2003	545500
3	674-100	Dave Widseler	452 Diamond Plaza	114391	3	Commercial	1999	249000
4	674-100	Dave Widseler	452 Diamond Plaza	114391	3	Commercial	2000	275500
5	674-100	Dave Widseler	452 Diamond Plaza	114391	3	Commercial	2001	290000
6	670-231	Sam Camarata	19 Big Bend Bld	114391	2	Residential	2004	450575
7	674-112	Chris Capelli	Hastings Barracks	114392	2	Residential	2004	350000
8	674-113	Sheila Sullivan	10034 Endin Mansions	114391	2	Residential	2002	1005425

Record: 0 Show: All Selected Records (0 out of 8 Selected.) Options

Figure 10.3C

Attributes of Tab10_3a

OBJECTID*	ParcelNumb	ZoningCode	DateAssessed	AssessedValue	OwnersName
1	673-100	2	2002	222000	Jeff Peters
2	673-101	2	2003	545500	Joel Campbell
3	674-100	3	1999	249000	Dave Widseler
4	674-100	3	2000	275500	Dave Widseler
5	674-100	3	2001	290000	Dave Widseler
6	670-231	2	2004	450575	Sam Camarata
7	674-112	2	2004	350000	Chris Capelli
8	674-113	2	2002	1005425	Sheila Sullivan

Record: 8 Show: All Selected Records (0 out of 8 Selected.) Options

Attributes of Tab10_3b

OBJECTID*	ZoningCode	ZoningType
1	2	Residential
2	3	Commercial

Record: 1 Show: All Selected Records (0 out of 2 Selected.)

Attributes of Tab10_3c

OBJECTID*	OwnerName*	Address	PostalCode
2	Jeff Peters	10 Railway Cuttings	114390
3	Joel Campbell	1115 Center Place	114390
4	Dave Widseler	452 Diamond Plaza	114391
5	Sam Camarata	19 Big Bend Bld	114391
6	Chris Capelli	Hastings Barracks	114392
7	Sheila Sullivan	10034 Endin Mansions	114391

Record: 0 Show: All Selected Records (0 out of 6 Selected.)

Figure 10.3D Joined table

Attributes of Tab10_3a_Tab10_3b_Tab10_3c

OBJECTID	ParcelNumb	DateAssessed	AssessedValue	ZoningType	ZoingCode	OwnerName	OwnerAddress	PostalCode
1	673-100	2002	222000	Residential	2	Jeff Peters	10 Railway Cuttings	114390
2	673-101	2003	545500	Residential	2	Joel Campbell	1115 Center Place	114390
3	674-100	1999	249000	Commercial	3	Dave Widseler	452 Diamond Plaza	114391
4	674-100	2000	275500	Commercial	3	Dave Widseler	452 Diamond Plaza	114391
5	674-102	2004	290000	Residential	2	Dave Widseler	452 Diamond Plaza	114391
6	670-231	2004	450575	Residential	2	Sam Camarata	19 Big Bend Bld	114391
7	674-112	2004	350000	Residential	2	Chris Capelli	Hastings Barracks	114392
8	674-113	2002	1005425	Residential	2	Sheila Sullivan	10034 Endin Mansions	114391

Record: 0 Show: All Selected Records (0 out of 8 Selected.) Options

Fig 10.4

Join3a3b3c : Select Query

ParcelNumb	Address	AssessedValue
673-101	1115 Center Place	545500
670-231	19 Big Bend Bld	450575
674-112	Hastings Barracks	350000
674-113	10034 Endin Mansions	1005425

Record: 1 of 4

Relation Rules (Codd, 1970)

- Only one value in each cell (intersection of row and column)
- All values in a column are about the same subject
- Each row is unique
- No significance in column sequence
- No significance in row sequence

Normalization

- Process of converting tables to conform to Codd's relational rules
- Split tables into new tables that can be joined at query time
  - The relational join
- Several levels of normalization
  - Forms: 1NF, 2NF, 3NF, etc.
- Normalization creates many expensive joins
- De-normalization is OK for performance optimization

Relational Join

- Fundamental query operation
- Occurs because
  - Normalization
  - Data created/maintained by different users, but integration needed for queries
- Table joins use common keys (column values)
- Table (attribute) join concept has been extended to geographic case

SQL

- Structured (Standard) Query Language – (pronounced SEQUEL)
- Developed by IBM in 1970s
- Now *de facto* and *de jure* standard for accessing relational databases
- Three types of usage
  - Stand alone queries
  - High level programming
  - Embedded in other applications

Types of SQL Statements

- Data Definition Language (DDL)
  - Create, alter and delete data
  - CREATE TABLE, CREATE INDEX
- Data Manipulation Language (DML)
  - Retrieve and manipulate data
  - SELECT, UPDATE, DELETE, INSERT
- Data Control Languages (DCL)
  - Control security of data
  - GRANT, CREATE USER, DROP USER

Geometry Class Hierarchy

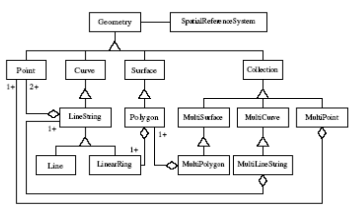


Figure 1: OpenGIS Geometry Class Hierarchy

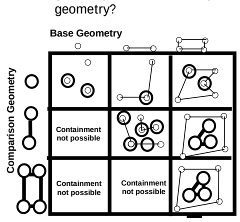
OpenGIS Geometry Class Hierarchy p. 226

Spatial Relations

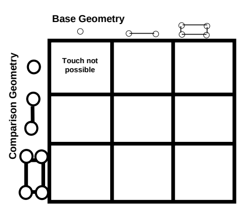
- Equals – same geometries
- Disjoint – geometries share common point
- Intersects – geometries intersect
- Touches – geometries intersect at common boundary
- Crosses – geometries overlap
- Within– geometry within
- Contains – geometry completely contains
- Overlaps – geometries of same dimension overlap
- Relate – intersection between interior, boundary or exterior

Two Possible Spatial Relations for Geographic Databases (Fig. 10.6 Text p. 227)

- CONTAINS: Does the base geometry contain the comparison geometry?



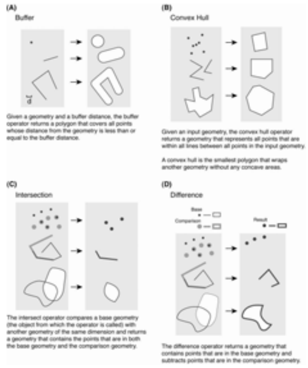
- TOUCHES: Does the base geometry touch the comparison geometry?



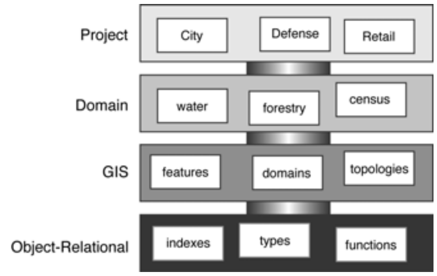
Spatial Methods

- Distance – shortest distance
- Buffer – geometric buffer
- ConvexHull – smallest convex polygon geometry
- Intersection – points common to two geometries
- Union – all points in geometries
- Difference – points different between two geometries
- SymDifference – points in either, but not both of input geometries

Examples of spatial analysis methods on geometries



Four levels of data model available for use in GIS projects.



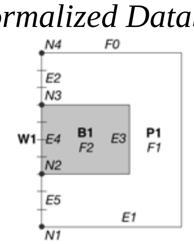
Stages in Database Design

- Reality
- Conceptual
- Logical
- Physical

Topology

- Two main database approaches
  - Normalized
    - Arc-node primitives
  - Physical
    - Simple features + rules

Normalized Database Topology Model



Parcels

ID
P1

Buildings

ID
B1

Walls

ID
W1

Parcel x Face

Parcel	Face
P1	F1
P1	F2

Wall x Edge

Wall	Edge	Order	Orientation
W1	E2	1	+
W1	E4	2	-
W1	E5	3	+

Building x Face

Parcel	Face
B1	F2

Nodes

ID
N1
N2
N3
N4

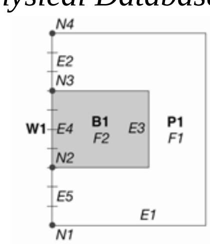
Edges

ID	Vertices	From	To	Left	Right
E1	(0,10),(8,10),(8,0),(0,0)	N4	N1	F0	F1
E2	(0,10),(0,7)	N4	N3	F1	F0
E3	(0,7),(5,7),(5,3),(0,3)	N3	N2	F1	F2
E4	(0,3),(0,7)	N2	N3	F0	F2
E5	(0,3),(0,0)	N2	N1	F1	F0

Faces

ID
F0
F1
F2

Physical Database Topology Model



Parcels

ID	Vertices
P1	(0,0),(0,3),(0,7),(0,10),(8,10),(8,0),(0,0)

Buildings

ID	Vertices
B1	(0,3),(0,7),(5,7),(5,3),(0,3)

Walls

ID	Vertices
W1	(0,10),(0,7),(0,3),(0,0)

Topology Rules

ID	Rule Type
R1	Pacels no overlap
R2	Buildings no overlap

Topology Errors

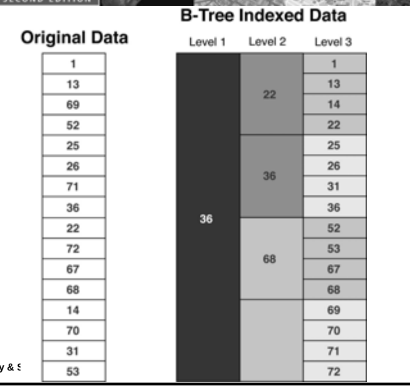
ID	Vertices	Rule Type	Feature	IsException
E1	...	-	-	-

Dirty Area

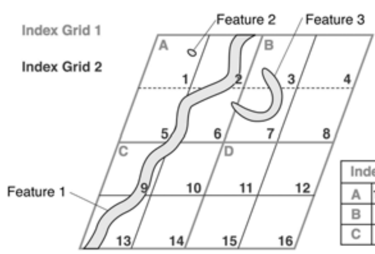
Vertices
...

Indexing

- Used to locate rows quickly
- RDBMS use simple 1-d indexing (R-tree, B-tree, etc.)
- Spatial DBMS need 2-d, hierarchical indexing
  - Grid
  - Quadtree
  - R-tree
  - Others
- Multi-level queries often used for performance (MBR)



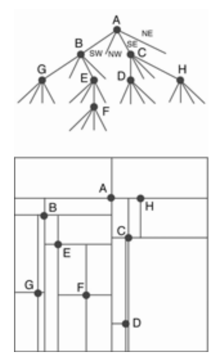
Multi-level Grid Index



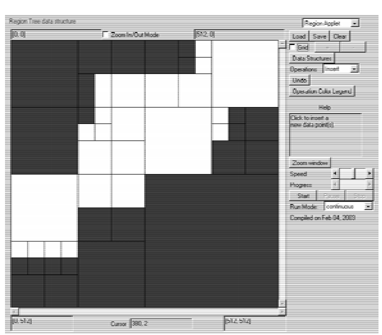
Index 1	
A	1,2,3
B	3
C	1

Index 2	
1	2
2	1
3	3
5	1
6	1,3
7	3
9	1
10	1
13	1

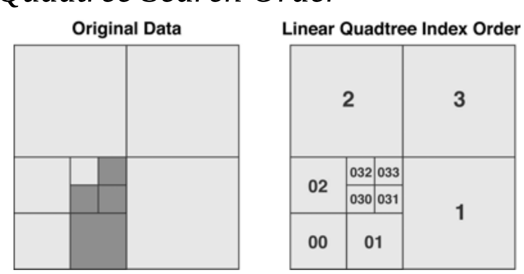
Point Quadtree



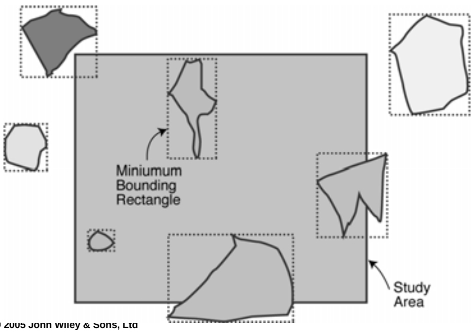
Region Quadtree



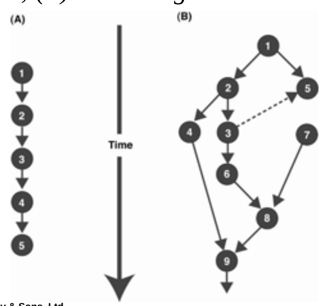
Quadtree Search Order



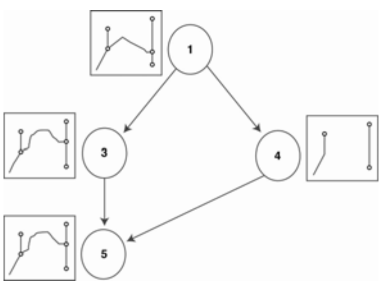
MBR



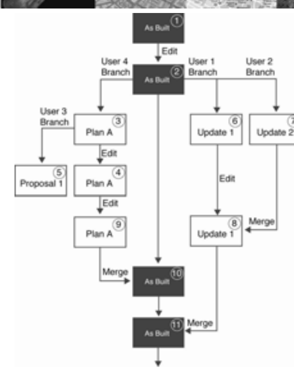
Database transactions: (A) Linear short transactions; (B) Branching version tree



Version Reconciliation



Version tree showing database states for a Main Plant geographic database



Summary

- Database – an integrated set of data on a particular subject
- Databases offer many advantages over files
- Relational databases dominate
- Some limitations for GIS