Embedding and Extending GIS for Exploratory Analysis of Large-Scale Species Distribution Data

Jianting Zhang
Department of Computer Science
The City College of the City University of New York
New York, NY, 10031
jzhang@cs.ccny.cuny.edu

Le Gruenwald
School of Computer Science
University of Oklahoma
Norman, OK 73071
ggruenwald@ou.edu

ABSTRACT
Exploratory analysis of large-scale species distribution data is essential to gain information and knowledge, stimulating hypotheses and seeking possible explanations of species distribution patterns. Geographical Information System (GIS) has played an important role in modeling and visualizing species distribution patterns for a single or a limited number of species. However, traditional GIS models do not take taxonomic components of species distribution data into consideration and are neither effective nor efficient in managing large-scale species distribution data.

In this study, we propose to embed and extend GIS for large scale species distribution data analysis. We provide an integrated data model that seamlessly links geographical, taxonomic and environmental data related to species distribution data analysis. We then present LEEASP (a Linked Environment for Exploratory Analysis of large-scale Species Distribution data), a prototype that has been developed based on the integrated data model. LEEASP utilizes the state-of-the-art advanced visualization techniques and multiple view coordination techniques to visualize different data sources that are relevant to species distribution data analysis. The North America tree species distribution data and other related data are used as an example to demonstrate the feasibility of the realization of the proposed integrated data model and how LEEASP can help users explore the geographical-taxonomic-environmental relationships

Keywords
Species Distribution, Data Modeling, Visualization, Exploratory Analysis

1. INTRODUCTION
Quantifying species-environment relationships, i.e., analyzing how plants and animals are distributed on the Earth in space, has been one of the important questions studied by biogeographers and ecologists. While traditionally species distribution data are limited and mostly descriptive and spatially inaccurate, the availability of species distribution and the associated environmental data has increased significantly in recent years due to technological advances. Examples of such technologies include GPS technology in modern field survey and geo-referring technology in transforming descriptive museum records to geographical coordinates (Wieczorek et al, 2004). In addition, spatial databases and GIS technologies make it easier to manage and analyze species distribution data and the Internet and the cyber-infrastructure greatly reduce the barriers for distributed data access (Bisby, 2000, Laihonen et al 2004, Guralnick et al 2007, Sarkar 2007). More than a million of species have been recorded in several repositories, such as the repositories provided by the "Catalogue of Life" project (COL, 2007) and the uBio project (uBio, 2007). While most of the current studies on species distribution modeling and prediction focus on a single species or a small number of selected species (Guisan and Zimmermann 2000, Guisan and Thuiller 2005), the capabilities of exploratory analysis of large-scale species distribution data are essential to gain information and knowledge, stimulate hypotheses and seek possible explanations of species distribution patterns.

There are several technical challenges in integrating disparate data sources that are related to species distribution analysis, such as individual species range maps, taxonomic categorization of species, different types of environmental data in the study area and various regionalization schemes (i.e., ecoregions, Loveland and Merchant 2004) based on human experts (Olson et al 2001) or computer programs (Hargrove and Hoffman et al 2004). First, we believe an integrated data model that provides a holistic view is crucial to the exploratory analysis from data management perspective. Second, visualization techniques are needed to present the data to users in a vivid and understandable manner based on the integrated data model. Third, the visualization components should be coordinated so that the selection of a subset of data in one view can be easily highlighted in other views to help users identify the relationships among the different types of data represented by the views.

In this study, we aim at developing an integrated data model that seamlessly links geographical, taxonomic and environmental data. We utilize state-of-the-art visualization techniques, such as embedding GIS for visualizing geographical maps, graph/tree visualization for taxonomic trees and ecoregion hierarchies, and, sortable table and Parallel Coordinate Plot (PCP, see Edsall 2003 for details) for multivariate environmental data. Finally we design and implement the prototype, called LEEASP (a Linked Environment for Exploratory Analysis of large-scale Species Distribution data), based on the integrated data model. We use the North America tree species distribution data as an example to demonstrate the utilization of the data model and the visualization system. The rest of the paper is arranged as follows. Section 2 provides some background on species distribution data analysis and formulates the research problem. Section 3 presents the integrated data model and the logical operations based on the data model. Section 4 provides the implementation details of LEEASP using the tree species distribution data in North America as a case study. Section 5 is the discussions of the related work and Section 6 is the summary and conclusion.

2. BACKGROUND AND PROBLEM FORMULATION

The problem of exploring the relationships among species distributions and the environment can be schematically illustrated in Fig 1. At the left side is the targeted species that can be either individual species or a group of species under a certain taxonomic rank. At the right side are the possible relationships to explore, such as species-area, species-water/energy and species-productivity. The species-latitude and species-altitude relationships are interesting mostly because the water/energy and productivity change along with latitude and altitude. Thus they can be categorized as the special cases of species-water/energy and/or species-productivity relationships. The relationships can be studied at the different ecological scales, from community to ecosystem and to global biomes.

Given a set of species in a particular region to study a certain type of relationships between species distribution data and the environment data, the primary task of data modeling is to represent the relevant data sources in a cohesive model to facilitate data manipulations. To this end, we categorize the different data types involved in the species distribution data analysis into three categories: geographical, taxonomic and environmental. The geographical data defines the spatial configurations of how the taxonomic data and the environmental data are observed/measured, which can be based on either the vector polygonal or the raster grid tessellations. The relationships among the three data types are illustrated in Fig. 2.

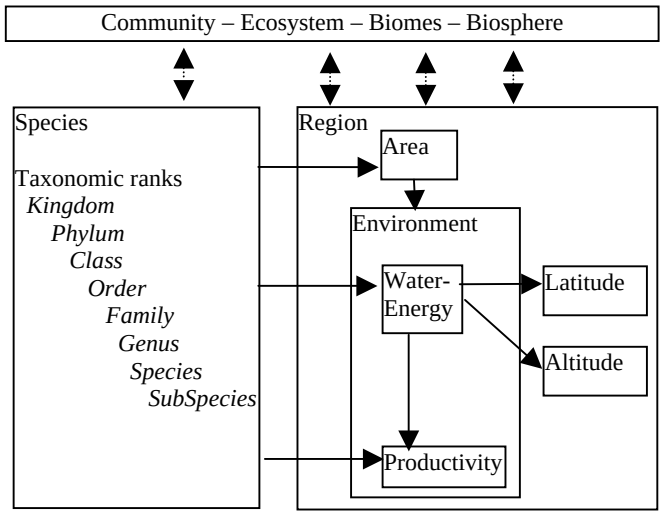

Fig. 1 Large-Scale Species Distribution Data Analysis Problems

The geographical distributions of species can be given either in the form of a set of individual species range maps or in the form of lists of species associated with a set of predefined regions. The former is more accurate but is difficult to obtain. While the "Catalogue of Life" project has recorded more than a million species (COL 2007), only a small portion has accurate range maps. The geographical distributions of species in the second form are more common for a large number of taxonomic groups at a larger geographical extent. For example, the World Wild Fund (WWF)'s WildFinder database has nearly 30,000 species in more than 800 hundred ecoregions at the global scale (WWF 2006). On the other hand, it is possible to derive the lists of species associated with the regions or cells from individual species range maps through zonal statistics or rasterization. However, the reverse process is generally not possible.

While the species list is sufficient in some applications, it is desirable to organize species based on a taxonomic nomenclature in large-scale species distribution data exploration. Currently there are a few repositories, such as the "Catalogue of Life" (COL, 2007), the Integrated Taxonomic Information System (ITIS, 2007) and uBio (uBio, 2007), that provide services to find the taxonomic hierarchy based on the common name or scientific name of a species. In our previous study (Zhang et al 2007), we have introduced the concept of Taxonomic Tree from data management perspective and defined operations on taxonomic trees. A taxonomic tree can be constructed from the species list distributed in a geographical unit (a region in the vector model or a cell in the raster model) using a certain taxonomic monoculture that complies with a tree structure. Representing a list of species as a taxonomic tree has the following advantages. First, it helps users understand the taxonomic relationships among the species. Second, it can be used to explore the species distribution patterns at higher taxonomic ranks that might not exist at the lower taxonomic ranks.

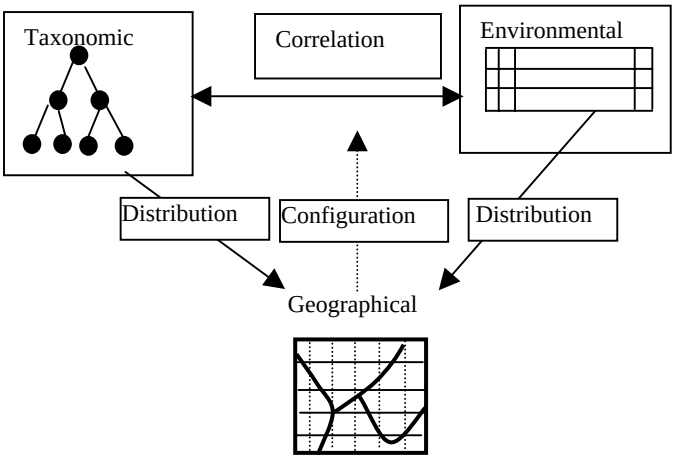

Fig.2 Illustration of the Three Types of Data in Species Distribution Exploration

In species distribution data analysis, very often the environmental data involved are the derived statistics from long term measurements. The examples of these are the monthly average temperature and precipitation and the bioclimate variables such as precipitation of the coldest quarter. While traditionally most environmental data come from ground observations, satellite derived products, such as the Normalized Vegetation Index (NDVI), the Enhanced Vegetation Index (EVI) and the Gross Primary Productivity (GPP), are increasingly used in the analysis due to their broad spatial coverage and regular temporal coverage (Pettorelli et al 2005). The ground point-based meteorological data are often interpolated into raster grids according to different requirements for providing continuous spatial coverage (Hijmans et al 2005). The satellite data are provided as raster datasets after certain types of preprocessing and, very often, they involve temporal aggregation. We next introduce our integrated data

model to provide a holistic view of different data types that are related to species distribution data analysis and facilitate data query and visualization.

3. THE INTEGRATED DATA MODEL

Modern GIS have well-established data models to associate spatial data with tabular data for both vector and raster based spatial tessellations. A natural way to use GIS for species distribution data analysis is to represent the distributions of multiple species as separate layers and associate the environmental data with the geographical units as their attribute tables. Mappings between spatial to tabular and from tabular to spatial in a single layer are natively supported by most GIS. The approach is illustrated in Fig. 3. However, there are three disadvantages with this approach.

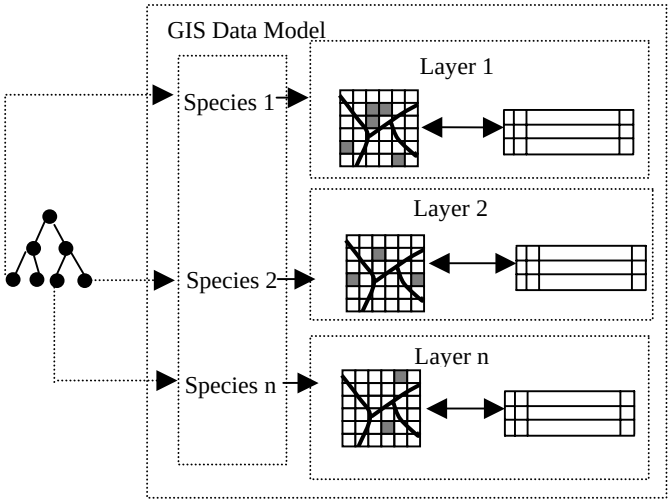

Fig. 3 Managing Large Scale Species Distribution Data Using Traditional GIS Data Model

First, the relationships among the geographical units in different layers are not a part of the traditional GIS data models. This makes answering cross-layer queries that are important in species distribution data analysis inefficient if not impossible. For example, querying the distributions of a group of species would require a union of the query results for all the individual species and a generation of a new layer for the union. Given the large number of species and their possible combinations, the number of layers that need to be generated is virtually countless, which will impose significant overheads for both GIS and users to manage the derived layers. Second, to use the layer-based GIS data model for managing multiple species distribution data, the geographical and the environmental data need to be joined for each layer, either permanently or dynamically. This is conceptually cumbersome and operationally inefficient. While it may not be a serious problem when only a few species are involved, our experiences show that the performance of ESRI ArcGIS decreases significantly when it was used to manage hundreds of species distribution data as separate layers. It may not be possible to use current GIS for species distribution data analysis when thousands or more species are involved. Third, while it is possible to arrange the species layers into groups in modern GIS (such as ArcGIS) to mimic the taxonomic hierarchy, it is difficult to identify/visualize query results that involve multiple layers back in the layer list. Unlike the geometric data and tabular data to which quite a few advanced visualization techniques can be applied, it is difficult to visualize the structures of layers in most existing GIS.

All the above three disadvantages are related to the fact that the taxonomic data, which plays an important role in species distribution data analysis, are not treated as the first-class data type in traditional GIS data models. In this study, we propose an integrated data model by extending the data model proposed in our previous study (Zhang et al 2007) to unify taxonomic, geographical and environmental data. The integrated data model is schematically illustrated in Fig. 4. In the data model, geographical data consist of a set of geographical units which can be either polygons or cells which we call the basic geographical units. The basic geographical units form a complete tessellation of the study area and each basic geographical unit is linked to a list of environmental values. It is clear that the geographical-environmental part of the integrated data model utilizes the traditional GIS data model but only a single layer is involved. To associate the geographical data with the taxonomic data, we represent the species distributed in a basic geographical unit as a taxonomic tree, which obviously is a subtree of the taxonomic tree for the whole study area. In addition to the links between the basic geographical units and their environmental value lists, the links between the geographical units and their taxonomic trees are also needed in the data model. While it is possible to physically build the taxonomic trees for all the basic geographical units as implemented in our previous study (Zhang et al 2007), an alternative approach is to use a bit vector of 0s and 1s to represent the presence/absence of all species distributed in a basic geographical unit. The bit vector can be used to identify the branches of the taxonomic tree for the whole study area and then construct a taxonomic tree for the basic geographical unit dynamically. As both the storage and operations on bit vectors are very efficient, the alternative approach might be more efficient for managing a large number of species. The explanations will be further detailed in the next section. However, the proposed data model is flexible enough to allow diverse implementations of the linking mechanisms between the geographical units and their taxonomic trees.

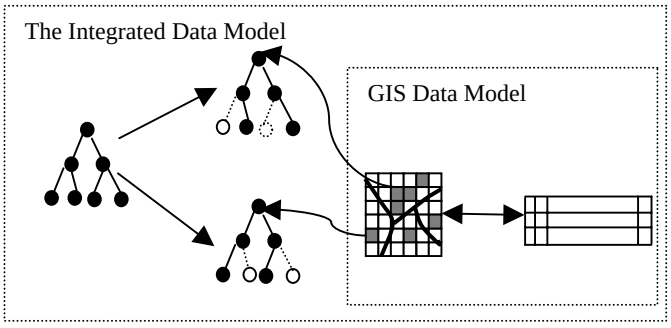

Fig. 4 The Integrated Data Model

The integrated data model supports the following five operations among the three data types to explore the species distributions and their relationships with the environments:

a. Geographical to Taxonomic (G->T): select a group of basic geographical units and identify the species distributed in the units and their taxonomic hierarchies. As a special case, the data model allows to count the number of species at different taxonomic levels, i.e., alpha diversities at the different taxonomic

ranks, for the selected units or the whole study area. The G->T operation and its derived statistics can be used in species-area analysis.

b. Geographical to Environmental (G->E): select a group of basic geographical units and compute the distributions of the corresponding environmental variable values recorded by the units. As shown in Fig. 4, this part can be realized by the traditional GIS data model and, subsequently, various statistics can be generated for each of the environmental variables (columns in the attribute table).

c. Taxonomic to Geographical/Environmental (T->G+E): pick one or more species or species groups at different taxonomic ranks and examine their geographical distributions and the value distributions of environmental variables. An application of the operation is to examine whether two species that have a close taxonomic relationship have similar or very different geographical distributions and value distributions of environmental variables.

d. Environmental to Geographical/Taxonomic (E->G+T): specify the minimum and maximum values of a subset of environmental variables and use their combinations as the criteria to identify basic geographical units that satisfy the criteria. The species distributed in the units can be identified by following the G->T operation. This operation is very useful in examining how species distributions change as the environment changes in different scenarios. For example, assume a set of possible value combinations of environmental variables and examine species and their geographical distributions for each of the combinations. Another example is more related to sensitivity analysis, i.e., using the environmental value ranges of a selected group of species (or niches) and/or a selected set of geographical units as the initial conditions, change the range of an environmental variable one by one and see how the species corresponding to the niches and their geographical distribution changes.

e. Taxonomic and Environmental to Geographical (T+E->G): this operation combines the taxonomic and the environmental criteria and maps the basic geographical units that satisfy such criteria. An example would be identifying the units that have a particular species or species group distributed and satisfy additional value ranges of environmental variables. T+E->G can be implemented as the combinations of T->G+E and E->G+T if the results of the previous operation can be kept and intersected with the results of the later operation. We make it a separate operation in case the combination of separate results is not possible as restricted by the underlying implementations.

All the above five operations involve geographical data. The geographical data type plays a central role in the integrated data model which we deem appropriate for the following considerations. First, as shown in Fig. 2, geographical data are the bridge between taxonomic data and environmental data. Even exploring the taxonomic-environmental relationship does not explicitly involve individual basic geographical units. The taxonomic data and the environmental data must be collected in a certain geographical area, which is the union of the individual basic geographical units in the area. Second, instead of maintaining pair wise associations (links) among the three data types, only the associations between geographical and taxonomic and between geographical and environmental data types are needed in the data model, which greatly reduces the complexity of the data model. This is especially true when a data type has multiple data sources, for example, bioclimate variables and satellite derived variables for environmental data.

Compared with the traditional GIS data model based approach, the integrated data model has the following advantages. First, taxonomic data are natively supported. Operations among the taxonomic data and operations among taxonomic, geographical and environmental data can be defined. Second, rather than duplicating geographical and environmental data to form layers for all the species to fit the traditional GIS data model, the three data types are now independent dimensions in the integrated data model. Conceptually, the integrated data model is much simpler when comparing Fig. 4 with Fig. 3. Finally, the independence among the three types of data in the integrated data model allows us to separate the visualization of individual data types from the coordination of multiple data types. This often leads to an easier implementation. In LEEASP, we have implemented a few visualization components for the different data types by combining a variety of open source packages that were originally designed for different purposes. The relatively inexpensive development cost of LEEASP supports the usefulness and effectiveness of the integrated data model.

The integrated data model is an extension to our previous research (Zhang et al 2007). While the data model proposed in the previous study is based on vector (region) GIS models, the integrated data model proposed in this study also allows raster tessellation of geographical data. The new data model also handles environmental data, which was left untouched in the previous study. The new data model provides more comprehensive and realistic supports to explore taxonomic-geographical-environmental relationships in large-scale species distribution data analysis.

4. ADVANCED VISUALIZATION FOR EXPLORATION

Visualization plays an important role in exploratory data analysis (Maceachren et al 1999). Exploratory Spatial Data Analysis (ESDA) techniques have been successfully applied to many GIS-centric applications (Haining et al 1998). We next show how the advanced visualization techniques can be applied to visualize taxonomic, geographical and environmental data and how to build visual components (or data views) for them. These data views can be coordinated to facilitate exploratory species distribution analysis. In particular, we demonstrate how GIS can be embedded into the larger exploratory analysis application and coordinate with other components. In this section, we first introduce the system and the data that are used as a test bed to realize the integrated data model and experiment on various visualization and coordination techniques. The design and implementation details of the data views and their coordination are elaborated subsequently.

4.1 Prototype System and Example Data

We have developed a system called Linked Environment for Exploratory Analysis of Large-Scale Species Distribution Data (LEEASP) as a test bed to realize the integrated data model and experiment on various visualization and coordination techniques. LEEASP is not intended to be the only or the best implementation of the integrated data model; rather, the purpose was to provide a concrete example to demonstrate the feasibility of building a large-scale species distribution system based on the proposed data model and the state-of-the-art visualization techniques. LEEASP extends and complements the GBD-Explorer prototype system presented in our previous study (Zhang et al 2007). While LEEASP supports using both vector

polygons and the raster cells as the basic geographical units based on the integrated data model, we use the raster model in this study to further distinguish it from the work presented in (Zhang et al 2007). Since LEEASP for the North America tree species distribution data is based on the raster model, the basic geographical units in the integrated data model are mapped to the raster grids and we thus use the basic geographical units and the cells interchangeably.

Large scale species distribution data are becoming increasingly available, for example, the digital distribution maps of the birds of the western hemisphere from NatureServe (NatureServe, 2007) cover 4273 species along with the distribution maps of the world's amphibians (5743 species) and the distribution maps of mammals of the western hemisphere (1786 species). For demonstration purposes, we use the North America tree species distribution data as an example. First, range maps of 679 tree species in ESRI shapefile format are downloaded from USGS (http://esp.cr.usgs.gov/data/atlas/little/) and imported to ArcGIS 9.0. The list of species, shapefile names and their Catalog of Life (COL, 2007) classifications are manually compiled. The shapefile data are rasterized at 0.5 by 0.5 degree resolution in LEEASP; however, finer resolution can be used at the costs of higher storage and computation requirements. Tree species whose ranges are less than a single cell are excluded and there are 606 tree species used in LEEASP. Second, global datasets of altitude and 19 bioclimate variables at 10 arc-minutes resolution are downloaded from http://www.worldclim.org/current.htm. The methods to generate these bioclimate datasets can be found in (Hijmans et al 2005). The datasets are further subsetted to the northwest hemisphere and downscaled to 0.5 by 0.5 degree resolution. Normally the bioclimate data are only available in land surfaces and cells without valid bioclimate data are excluded. Third, the North America Ecoregion data (up to Level III) are obtained from the EPA website http://www.epa.gov/wed/pages/ecoregions/ecoregions.htm. The Ecoregion dataset was rasterized at the same 0.5 by 0.5 degree spatial resolution and the Ecoregion hierarchy was constructed.

4.2 Data Views

Each data source can be implemented as a data view by applying appropriate visualization techniques. Data sources of the same data type can be visualized differently to facilitate the understanding of the relationships among the relevant data sources. LEEASP has implemented four data views, namely the basic geographical data view (Geographical View for short), the additional Ecorgion data view for geographical data (Ecoregion View for short), the taxonomic data view (Taxonomic View for short) and the environmental data view (Environmental View for short). The four views in LEEASP are shown in Fig. 5 and they will be introduced in the next subsections.

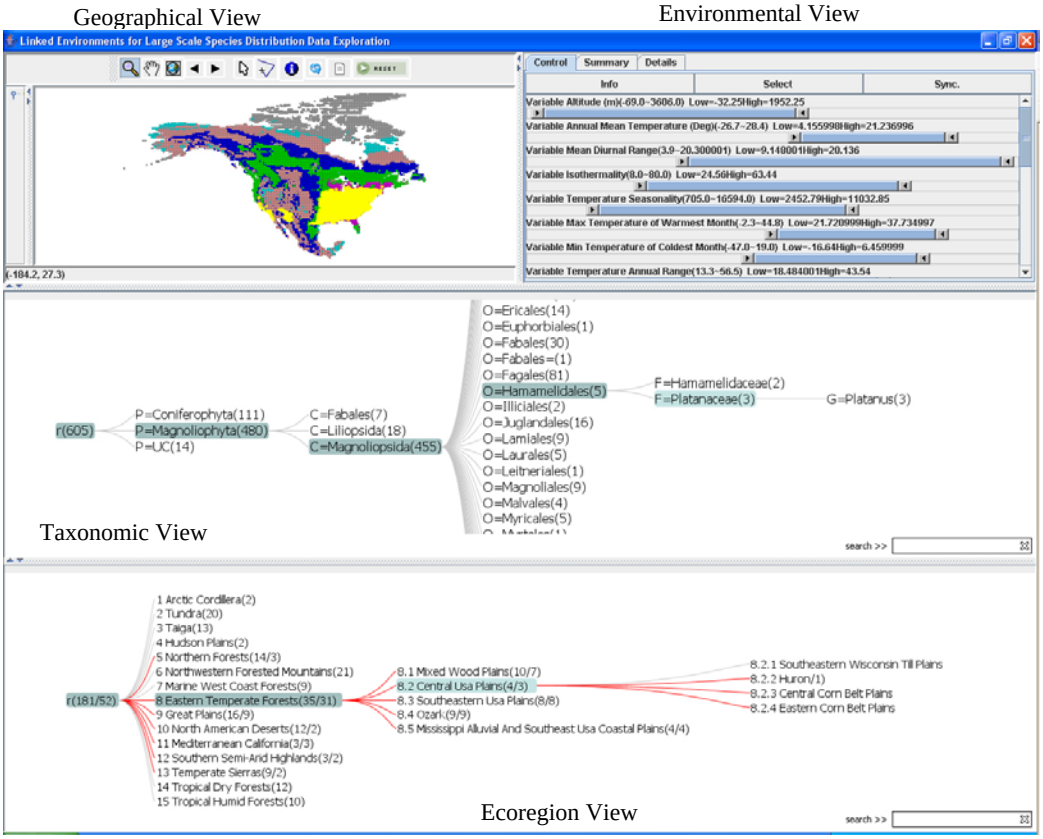

Fig. 5 The Four Views in LEEASP: Geographical (Top-Left), Environmental (Top-Right), Taxonomic (Middle) and Ecoregion (Bottom)

4.2.1 Geographical View

The basic geographical units can be visualized in an embedded GIS that provides standard GIS functions, such as zoom, pan, and selection. In species distribution data analysis, there are some predefined regions, such as biodiversity hotspots (Myers et al 2000, Willis et al 2007) and environmental transects and gradients (e.g., Willig et al 2003, Godefroid et al 2006) that are of particular interests to biogeographical and ecological research. The links between these regions and the basic geographical units should be pre-computed and stored. In addition, the basic geographical data view should allow users to select basic units to form their "regions of interests". In LEEASP, we have extended an open source GIS called JUMP (Vividsolutions 2004) to visualize the basic geographical data and link the basic geographical data with other data sources. LEEASP allow users to select and deselect a subset of basic geographical units in three ways: (1) select units that intersect with the current geographical location with a certain distance tolerance; (2) select units that intersect with a rectangular region that is interactively specified by a user; and (3) select units that intersect with a polyline that is interactively drawn by a user.

The geographical view plays a unique role among the four views in the sense that it is designed to present the distribution information and it displays all cells at the same time to provide users an overview. The subset of cells identified by the operations in other views can be highlighted and contrasted with the rest easily. Furthermore, the geographical view can combine the cells identified by multiple selections, which is to say that the geographical view is "stateful". On the contrary, a new selection clears up a previous selection in the other three views, i.e., they are "stateless". The purpose of the design is to use the geographical view as the "context" and the other views as the "focus" under the "Focus+Context" visualization framework (Ivan et al 2000) with respect to species distribution explorations.

4.2.2 Environmental View

The environmental data are multivariate tabular data. Essentially any multivariate visualization techniques can be applied to the environmental data, such as the sortable tables, histograms, graphs and the Parallel Coordinate Plots (PCP, Edsall 2003).

LEEASP has implemented three components to visualize the subset of environmental data corresponding to a subset of selected basic geographical units in the geographical view. The three components, namely *Summary, Control and Details*, are implemented as tab pages (referred to as panels for short hereafter) in LEEASP (Fig. 6). The *Summary* panel presents the information in text (key-value pairs) format. The *Control* panel visualizes the same information using sliders and will be introduced shortly. The *Details* panel shows the values of the subset of environmental data in two forms: the sortable table and PCP. Users can sort the table based on one or more columns in the sortable table. The two visualization components are linked to provide better understanding, i.e., any selected rows in the sortable table will be highlighted in the PCP and the rest is displayed as background information. If we treat the geographical view as the "Context", then the *Details* panel showing the detailed information of the selected basic geographical units is the "Focus". The design falls nicely in the "Focus+Context" visualization framework (Ivan et al 2000). In addition, the *Summary and* the *Control* panels provide overview information while the detailed information for the environmental data are shown in the *Details* panel; thus the design also utilizes the "Overview+Detail" visualization principle (Ivan et al 2000).

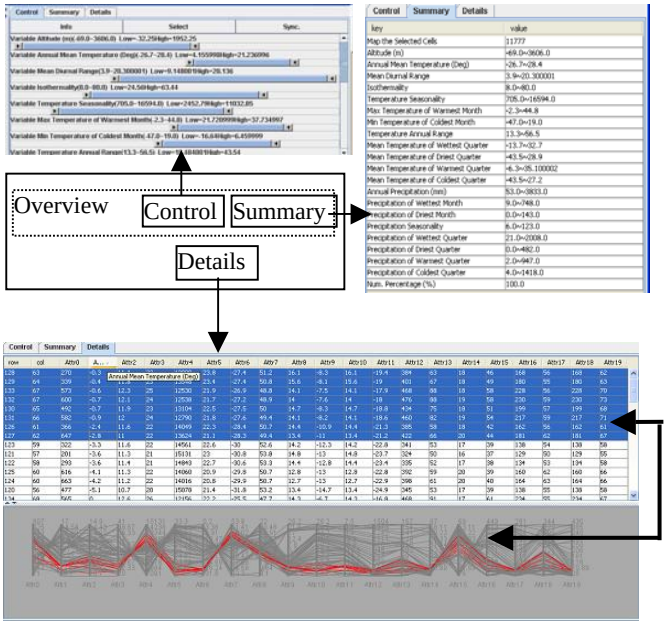

Fig. 6 The Visualization Components and Coordination in the Environmental Data View

The *Control* panel consists of a number of sliders representing the set of environmental variables used in an environmental dataset. A slider range represents the minimum and maximum values of a specific environmental variable of the whole dataset (global min/max) and the low/high marks in the slider represent the minimum and maximum values of the environmental variable of the selected basic geographical units (local min/max). The sliders and their corresponding labels display the environmental envelopes in a more vivid manner compared with the text information displayed in the *Summary* panel. The sliders in LEEASP are used as a compact form of the histograms for the environmental variables. While not as informative as a histogram with sophisticated controls, the slider, as a visualization component, may be more preferable when there are a large number of environmental variables to handle but only a limited space is available for visualization. The *Control* panel is designed to allow two-way operations. Besides visualizing the min/max values of the environmental variables corresponding to the subset of selected geographical units, the sliders in the *Control* panel can be used to specify the min/max values of the environmental variables that a user might be interested in and perform the E->T+G operations to identify the species, their taxonomic hierarchies and geographical distributions that satisfy the environmental constraints.

4.2.3 Taxonomic View

The integrated data model represents the taxonomic data distributed in a basic geographic cell as a tree. Tree visualization, as a special type of graph visualization, has been well studied and quite a few efficient tree visualization techniques are available (Ivan et al 2000). While algorithms for graph layout are generally very expensive, there are efficient tree layout algorithms that can handle millions of nodes in a tree (Christoph et al 2002). This

seems to be sufficient to handle all the species with available range maps in the near future.

In the current implementation of LEEASP, we use the following eight levels of taxonomy: Kingdom/ Phylum/ Class/ Order/ Family /Genus/ Species/ SubSpecies. Hereafter we will refer to these eight levels of taxonomy as taxonomic ranks and taxon names at all taxonomic ranks as taxa. LEEASP uses Prefuse (Jeffrey et al, 2005) to visualize the taxonomic trees. Prefuse has been successfully used to visualize trees with more than 600,000 nodes and demonstrated excellent scalability in tree visualization (Jeffrey et al, 2005). In LEEASP, in addition to use the taxa names as the tree node labels, we also put the tree sizes (corresponding to the numbers of species) under non-leaf tree nodes (representing higher rank taxa) in the labels. This is to give users immediate access to the numbers of species of the whole study area for all taxa with different taxonomic ranks. We have added to the Taxonomic view a few typical operations supported by Prefuse for trees, such as zooming in/out of the canvas, animation when a tree node is expanded and highlighting the nodes along the path from the root to the node that users are currently exploring (Fig. 5). When a tree node is expanded by clicking on the node (showing the details of the node), other nodes that are the decedents of the sibling nodes of the chosen node will collapse. However, the nodes in the path from the root to the node being chosen and the siblings of the nodes in the path (context) will be kept. The advanced tree visualization functionality provided by Prefuse makes it more preferable compared with the standard Java Swing based tree visualization.

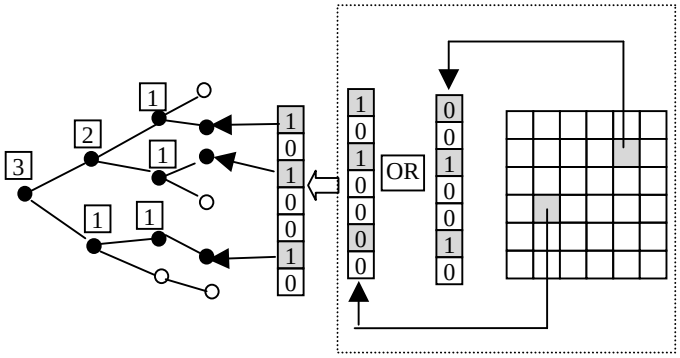

Fig. 7 Implementation of G->T Operation in LEEASP

As discussed in the previous section, the integrated data model allows us to generate taxonomic trees of the basic geographical units dynamically using bit vector representations. The details for implementing the G->T and T->G+E operations are shown in Fig. 7 and Fig. 8, respectively. To identify all the species that are distributed in a subset of selected basic geographical units, the bit vectors representing the presence/absence (1/0) of the cells are first retrieved and then combined using the logic add (OR) operation. The leaf nodes in the taxonomic tree representing the species corresponding to the 1s in the combined bit vector are subsequently retrieved. The leaf nodes then recursively follow the links to their parents until the root of the taxonomic tree is researched. The paths from the identified leaf nodes to the root will form the subtree of the taxonomic tree. Obviously, the subtree is equivalent to the combinations of the taxonomic trees of species distributed in the selected basic geographical units. However, the implementation is more efficient than directly combining the subtrees which involves level wise set operations (Zhang et al 2007). The improved efficiency is more desirable for identifying the subtree resulting from a large number of selected basic geographical units. Traveling the paths from the leaf nodes to the root, the number of species at different taxonomic ranks can be accumulated and displayed as part of the node labels. The node labels along the paths now look like "O=Fabales(30/15)", where O stands for the taxonomic rank *Order* and *Fabales* is the taxa name. The first number in the bracket tells the amount of species (or species richness) under the taxa represented by the node for the whole dataset, and the second number tells the amount of species under the taxa for the selected cells. The ratio of the second number to the first number can be used to tell to what degree the selected cells have the same species richness as the whole study area.

Fig. 8 shows the details of implementing the T->G+E operation of the integrated data model in LEEASP. For any selected taxa, the leaf nodes under the subtree representing the selected taxa are retrieved and stored in a species vector. LEEASP then checks all the basic geographical units and selects the units with at least one of the bit corresponding to the species vector being set to true (1). The selected basic geographical units will be highlighted by the embedded GIS in the geographical view and by the visualization components in the environmental view.

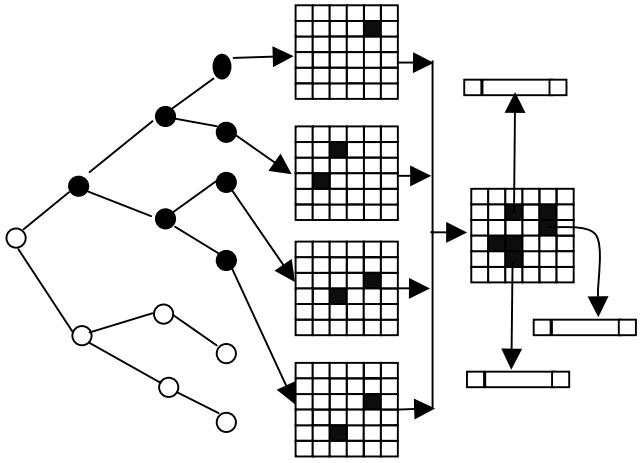

Fig. 8 Implementation of T->G+E Operation in LEEASP

4.2.4 Ecoregion View

The Ecoregion data view implemented in LEEASP serves as a demonstrative example to show how multiple data sources of the same data type can provide useful complementary information in species distribution data analysis. As discussed in the data modeling section, the Ecoregion data is an additional data source to the basic geographical data. One way of visualizing the Ecoregion data is to join the Ecoregion hierarchy with the basic geographical units, make the Ecoregion data a part of the attribute table of the geographical data, and then form relational queries to highlight the basic geographical units belonging to a particular ecoregions group. This is a common practice in current GIS. While the implementation is relatively easy in a GIS, one drawback is that the Ecoregion hierarchy cannot be explicitly visualized and easily interact with users. The Ecoregion hierarchy plays an important role in helping users understand how species compositions and the observed environmental measurements change when the levels of the hierarchy go up or down and how

they change among different Ecoregions. LEEASP applies the similar tree visualization techniques as for the taxonomic data to visualize the Ecoregion hierarchy. Similar to the Taxonomic data view, the Ecoregoin data view allows users to select an ecoregion at any level and the basic geographical units fall in the Ecoregion will be identified and highlighted. This can be further used to explore the relationships among the taxonomic data and the environmental data based on the selected ecoregions. Depending on the configuration, the basic geographical units selected by other data views can be mapped back to the Ecoregion hierarchy, i.e., identifying and highlighting paths in the Ecoregion hierarchy. An application might be to identify the ecoregions that have a particular species distributed, or the ecoregions that satisfy certain value ranges of environmental variables, or their combinations.

4.3 View Coordination

Coordination among multiple views is an important component in an integrated data exploratory analysis system. LEEASP coordinates the four implemented views, namely, the Geographical View, the Taxonomic View, the Environmental View and the Ecoreigon View, based on the integrated data model. We have introduced some of the coordination mechanisms between individual data views when they are introduced and the coordination among different components within a single data view, such as the *Summary, Details and Control* visualization components in the Environmental Data View. In this section, we provide a high level view of the coordination implemented in LEEASP.

The control flows of the coordination among the four data views in LEEASP are shown in Fig. 9. While each data view may have specialized controls to coordinate with other views, LEEASP abstracts three types of operations that are common to all the four views, namely *Info*, *Select* and *Sync*. The *Info* operation shows the min/max values of environmental variables for the basic geographical units associated with the nodes in the taxonomic/ecoregion trees or the currently selected unit(s) in the map (Geographical view). The information will be displayed in the *Control* and the *Summary* panels of the Environmental view. The *Select* operation selects the basic geographical units associated with the nodes in the taxonomic and the ecoregion trees and highlights them in the Geographical view. The *Select* operation does not remove previously selected units in the Geographical view so that users can combine the basic geographical units selected from different sources and perform operations supported by the Geographical view. The *Sync* operation performs both of the *Info* and the *Select* operations except that the previous selected basic geographical units in the Geographical view are cleared before the new selection. In addition, the selected basic geographical units resulting from the *Sync* operation will be automatically propagated to other views and the corresponding values in the views will be highlighted. While the primary goal of distinguishing S*elect+Info* from *Sync* is to allow users to combine subsets of basic geographical units selected from different sources before displaying the information of the units in the appropriate views, it also helps experienced users working on low-end computers to run the prototype smoothly since the *Info/Select* operation is much less computationally intensive than the *Sync* operation.

LEEASP also allows users to coordinate its internal views with external views. More specifically, LEEASP associated two links to the Web resources for each of the taxa, one for searching the taxa name in the COL species database and the other for searching the taxa name in the Google Image. The tools have been proved to be handy when more information other than the taxa name is needed during the process of exploring the taxonomic data. The similar tool can be applied to Ecoregion data and environmental data as well, such as retrieving detailed textural descriptions/images for a selected ecoregion.

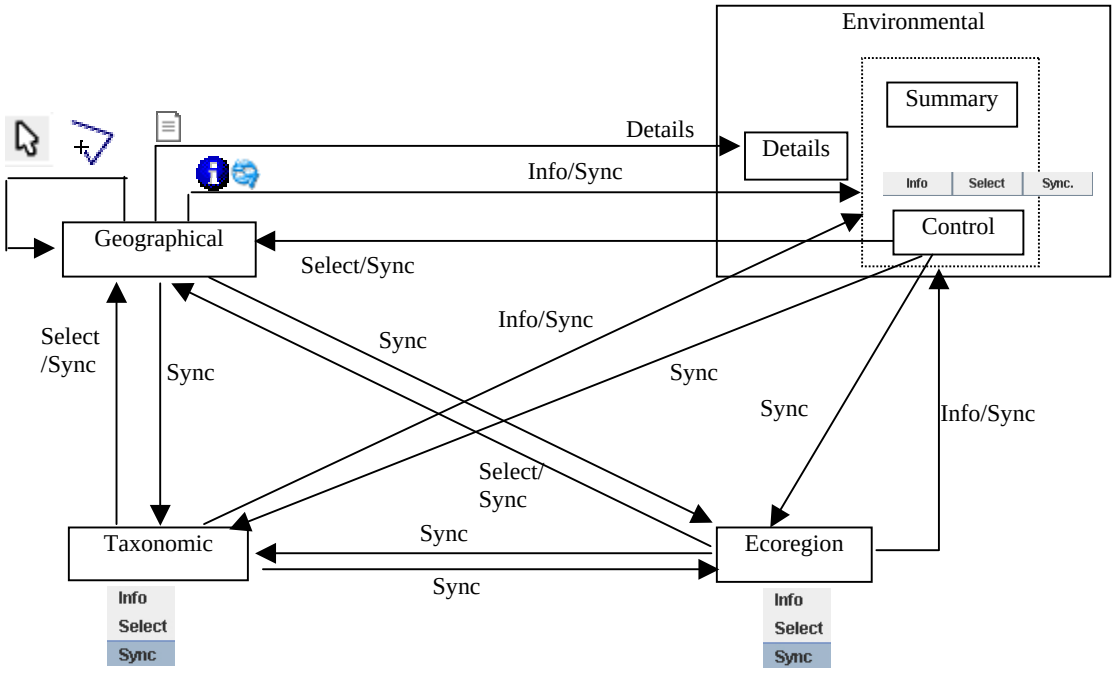

Fig. 9 Control flows for Coordination among the Geographical, Environmental, Taxonomic, and Ecoregion Views in LEEASP

5 DISCUSSIONS

The work presented in this paper is largely motivated by the Climate-Vegetation Atlas of North America project at the USGS in 1990s (USGS, 1999). The atlas presents information on the modern relations between climate and the distributions of 407 plant taxa and biogeographic entities from across North America at the 25 km grid resolution. The work is also related to an ongoing project called A Climate Change Atlas for 80 Forest Tree Species of the Eastern United States (Prasad and Iverson, 1999-ongoing) which delivers tree species distribution maps under different climate change scenarios and the value tables of environmental variables of the species through the Internet. However, the work reported in this paper focuses on exploratory data analysis rather than providing static images/data values.

The research on integrated data modeling benefits from existing works on spatiotemporal data modeling and visualization (Andrienko et al 2003, Guo et al 2006). While both the taxonomic and environmental data discussed in the integrated data model can be treated as the attributes of space in generic spatiotemporal data models (Andrienko et al 2003), no guidelines have been provided on how to extend the generic spatiotemporal data models for domain specific applications. The integrated data model provided in this study can be regarded as a step towards tailoring the generic data models for domain specific needs in species distribution data analysis. On the other hand, in our integrated data model, we treat some temporal information as part of the environmental attributes, which is different from the spatiotemporal data models that always model temporal information as an orthogonal dimension in the space-time cube. This is because we focus on the present day species-environment relationships and treat the temporal variations as part of the environmental data. It may be necessary to model time explicitly and incorporate the spatiotemporal data models when historical species-environment relationships are the primary focuses (Willis et al 2007).

Tree and graph visualization techniques are very useful in exploring species taxonomies (Bongshin et al 2004, Hillis et al 2005, Graham and Kennedy 2005, Parr et al 2007). However, the taxonomic data in the studies are not linked to the geographical distributions and no geospatial exploration is involved. We are working on further extending the integrated data model to allow more sophisticated operations on the taxonomic data, such as exploring the taxonomic relationships between different taxonomic nomenclatures similar to the work presented in (Graham and Kennedy 2005). We also believe that many of the visualization and coordination techniques introduced by the previous works can be incorporated into LEEASP to further enhance its functionality.

LEEASP has been evaluated by a few ecologists and biogeography researchers since it was first released. Their feedbacks have greatly improved the usability of the system. However, the development of LEEASP, in its current form, is largely driven by data modeling and multiple coordinated view research. We plan to conduct more thorough user evaluations by domain scientists to reveal its strengths and weaknesses for future improvements.

LEEASP currently assumes all the relevant data sources are locally available and well-formatted. With the emerging Service-Oriented Architecture (SOA, Erl 2005), more and more species range maps, taxonomic repository and environmental data are becoming programmatically available through the service interfaces (Frehner and Brandli 2006, Graham et al 2008), it is possible to extend LEEASP to support distributed and dynamic data sources. Another technical direction is to experiment on the 'mashup' approach (Wood et al 2007) to specify the visualization and coordination requirements in a standardized way (e.g., XML) and use existing generic visualization tools (e.g., Google Earth) to integrate different data sources, visual components and coordination for user interactions. Under the new architecture, LEEASP is likely to work at the server side and generates visual components upon the requests of the generic visualization tools.

6 SUMMARY AND CONCLUSIONS

In this study, we have proposed an integrated data model that seamlessly links geographical, taxonomic and environmental data that are related to species distribution data analysis. Visualization components for different data sources by applying a variety of advanced visualization techniques are discussed. The LEEASP system that integrates the visualization components through view coordination is used to demonstrate the feasibility and effectiveness of the proposed data model using tree species distribution data in North America. This dataset has more than six hundred species, more than eleven thousands basic geographical units, nearly twenty environmental variables and all the EPA North America ecoregions. We plan to apply the data model and the LEEASP system to more large-scale species distribution data and test its effectiveness.

Software Availability

The LEEASP prototype system, including documentation, source codes, binary distributions, third-party libraries and data, is publicly available at http://www-cs.ccny.cuny.edu/~jzhang/tech/LEEASPV10.zip. We encourage interested readers to try LEEASP by following an easy-to-install process.

Acknowledgement

This work is supported in part by NSF grant ITR #0225665 SEEK and NSF grant ATM #0619139 while Jianting Zhang was with the University of New Mexico and the University of California at Davis, respectively. We thank Dr. Kate He at the Murray State University, Dr. Weimin Xi at TAMU and Anantha M. Prasad at USDA Forest Service for evaluating the LEEASP prototype and providing constructive suggestions.

References

1. Bisby, F. A., 2000. The quiet revolution: biodiversity informatics and the Internet, Science, 289(5488), pp. 2309-12.
2. Bongshin, L., Parr, C. S., et al., 2004. How users interact with biodiversity information using TaxonTree. Proceedings of the working conference on Advanced Visual Interfaces, pp. 320-327.
3. COL (Catalogue of Life), 2007. http://www.catalogueoflife.org/
4.
5. Edsall, R. M., 2003. The parallel coordinate plot in action: design and use for geographic visualization. Computational Statistics & Data Analysis, 43(4), pp. 605-619.
6. Erl, T., 2005. Service-Oriented Architecture (SOA): Concepts, Technology and Design: Prentice Hall PTR.
7. Frehner, M., Brandli, M., 2006. Virtual database: Spatial analysis in a Web-based data management system for

distributed ecological data. Environmental Modelling & Software, 21(11), 1544-1554.
8. Godefroid, S., Rucquoij, S., et al., 2006. Spatial variability of summer microclimates and plant species response along transects within clearcuts in a beech forest. Plant Ecology 185(1), pp. 107-121.
9. Graham, M., Kennedy, J., 2005. Extending taxonomic visualisation to incorporate synonymy and structural markers. Information Visualization 4(3), pp. 206-223.
10. Graham, J., Simpson, A., Crall, A., Jarnevich, C., Newman, G., Stohlgren, T. J., 2008. Vision of a cyberinfrastructure for nonnative, invasive species management. Bioscience, 58(3), 263-268.
11. Guisan, A., Zimmermann, N. E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135(2-3), pp. 147-186.
12. Guisan A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8(9), pp. 993-1009.
13. Guralnick, R. P., Hill, A. W., Lane, M., 2007. Towards a collaborative, global infrastructure for biodiversity assessment. Ecology Letters, 10(8), 663-672.
14. Haining, R. Wise, S., Ma, J. S. 1998. Exploratory spatial data analysis in a geographic information system environment. Journal of the Royal Statistical Society Series D-the Statistician, 47, 457-469.
15. Hargrove, W. W., Hoffman, F. M., 2004. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. Environmental Management, 34, S39-S60.
16. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G.Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25(15), pp. 1965-1978. Also see http://www.worldclim.org/
17. Hillis, D. M., Heath, T. A., et al., 2005. Analysis and visualization of tree space. Systematic Biology 54(3), pp.471-482.
18. Ivan, H., Guy, M. et al, 2000. Graph Visualization and Navigation in Information Visualization: A Survey. IEEE Transactions on Visualization and Computer Graphics 6(1), pp.24-43.
19. ITIS, 2007. The Integrated Taxonomic Information System. http://www.itis.gov/
20. Jeffrey, H., Stuart, K. C.James, A. L., 2005. Prefuse: a toolkit for interactive information visualization. Proceedings of the SIGCHI conference on Human factors in computing systems. Portland, Oregon, USA, ACM Press. Also see http://www.prefuse.org/
21. Laihonen, P., Kalliola, R., Salo, J., 2004. The biodiversity information clearing-house mechanism (CHM) as a global effort. Environmental Science & Policy, 7(2), 99-108.
22. Loveland, T. R., Merchant, J. M., 2004. Ecoregions and ecoregionalization: Geographical and ecological perspectives. Environmental Management, 34, pp. S1-S13.
23. Maceachren, A. M., Wachowicz, M., Edsall, R., Haug, D., Masters, R., 1999. Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. International Journal of Geographical Information Science, 13(4), 311-334.
24. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature, 403(6772), 853-858.
25. NatureServe, 2007. http://www.natureserve.org/getData/animalData.jsp
26. Olson, D. M., Dinerstein, E., Wikramanayake, et al, 2001. Terrestrial ecoregions of the worlds: A new map of life on Earth. Bioscience, 51(11), 933-938.
27. Parr, C. S., Lee, B., et al, 2007. EcoLens: Integration and interactive visualization of ecological datasets. Ecological Informatics 2(1), pp.61-69.
28. Pettorelli, N., Vik, J. O., et al, 2005. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends in Ecology & Evolution, 20(9), pp. 503-510.
29. Prasad, A. M., Iverson, L. R., 1999-ongoing. A Climate Change Atlas for 80 Forest Tree Species of the Eastern United States [database]. http://www.fs.fed.us/ne/delaware/atlas/index.html, Northeastern Research Station, USDA Forest Service, Delaware, Ohio.
30. Sarkar, I. N., 2007. Biodiversity informatics: organizing and linking information across the spectrum of life. Briefings in Bioinformatics, 8(5), 347-357.
31. uBio, 2007. Universal Biological Indexer and Organizer, http://www.ubio.org/
32. USGS, 1999. Atlas of Relations Between Climatic Parameters and Distributions of Important Trees and Shrubs in North America, http://pubs.usgs.gov/pp/p1650-a/
33. Vivid Solutions, 2004. Unified Mapping Platform (JUMP).http://www.vividsolutions.com/jump/.
34. Wieczorek, J., Guo, Q. G.Hijmans, R. J., 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science, 18(8), pp. 745-767.
35. Willig, M. R., Kaufman, D. M., et al, 2003. Latitudinal gradients of biodiversity: Pattern, process, scale, and synthesis, Annual Review of Ecology Evolution and Systematics 34: 273-309.
36. Willis, K. J., Gillson, L., Knapp, S., 2007. Biodiversity hotspots through time: an introduction. Philosophical Transactions of the Royal Society B-Biological Sciences, 362(1478), 169-174.
37. Wood, J., Dykes, J., Slingsby, A., Clarke, K., 2007. Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. Ieee Transactions on Visualization and Computer Graphics, 13(6), 1176-1183.
38. WWF, 2006. World Wildlife Fund WildFinder: Online database of species distributions, ver. Jan-06. From http://www.worldwildlife.org/WildFinder.
39. Zhang, J., Pennington. D., Liu, X., 2007, GBD-Explorer: Extending Open Source Java GIS for Exploring Ecoregion-Based Biodiversity Data, Ecological Informatics, 2(2), pp.94-102.