Ecological Informatics 6 (2011) 93–110

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

HerbariaViz: A web-based client–server interface for mapping and exploring flora observation data

Tom Auer *,[1], Alan M. MacEachren, Craig McCabe [2], Scott Pezanowski, Michael Stryker

GeoVISTA, Dept. of Geography, Penn State University, University Park, PA 16802 USA

ARTICLE INFO

Article history:
Received 4 March 2010
Received in revised form 4 August 2010
Accepted 13 September 2010
Available online 18 September 2010

Keywords:
Web-mapping
Geovisualization
Client–server system
Open standards
Open source
Flora collections
California Herbaria

ABSTRACT

The potential for physical flora collections to support scientific research is being enhanced by rapid development of digital databases that represent characteristics of the physical specimens held in those collections and make this information available remotely. One example is the unified database of California flora observations from the Consortium of California Herbaria that was developed to support the exploration of plant diversity patterns, distribution ranges of species, and vegetation associations for specimens held in physical collections. Many of the records in the herbaria database, and in complementary databases elsewhere, are geo-referenced; but, current web tools for accessing the data do not take advantage of that georeferencing. In this paper, we report on development and implementation of a web-based client–server map interface to facilitate open mapping and exploration of the dataset. Three research objectives were addressed: (1) develop a method for efficient web-map client–server interaction involving large volumes of spatiotemporal point data, (2) develop a symbology and symbol scaling method for representing those spatial–temporal data in the client, and (3) develop an interface for client–server interactions and data exploration. With a focus on cartographically-sound visualization and user-friendly interaction, we introduce HerbariaViz, a web mapping application that provides space–time–species data query responses efficiently. Following a discussion of relevant literature, we present open-source methods for aggregating point data spatially and temporally, outline our approach to sound cartographic representations of those data, and detail the design of a client interface for making requests and mapping responses. A focus group session involving domain experts was performed to provide user evaluation of the application. In our discussion, we present potential avenues of future work, including: facilitating query response comparisons, handling incomplete and inaccurate data, and generalizing the method presented here.

1. Introduction

Geovisualization research in the past 15 years has directed substantial effort toward improving dynamic, interactive maps, especially those intended for internet distribution (MacEachren et al., 2008). The number of these applications has grown exponentially in recent years with the advent of application programming interfaces (APIs), such as the Google Maps APIs and the OpenLayers open-source API. Complementary research in exploratory geovisualization has resulted in substantial efforts to understand the use and improve the usability of web-based interactive map applications (Bhowmick et al., 2008; Lienert et al., 2009). While interest in generating both exploratory geovisualization applications and web-based interactive maps grows, it is important to apply principles from research on map use and design to yield results that are both useable and useful.

Advances in both exploratory geovisualization and in web-mapping more generally are occurring in parallel with dramatic increases in the availability of data that contain geographic referencing. These increases result from a combination of advances in remote sensing, integration of GPS into a wide array of devices, and advances in computational methods to extract place references from text sources. Commercial distributors, such as Oracle and MySQL, and open-source distributors, such as PostgreSQL, all now support spatial data storage, analysis, and retrieval natively or through extensions.

Web-based map applications can play an important role in facilitating the analysis and communication of spatially-enabled datasets. But, many current systems allow only for relatively static display of large datasets. Systems often have no capability to deal with spatio-temporal data. The next generation of web-map tools will require unique solutions to navigating, querying, displaying, and interpreting millions of data points in both time and space. Additionally, existing cartographic design principles must be adapted to cope with on-the-fly generation of custom maps associated with

* Corresponding author. Tel: +1 401 789 6224; fax: +1 401 789 1932.
E-mail address: mthomasauer@gmail.com (T. Auer).

[1] Applied Science Associates, Inc., 55 Village Square Dr., South Kingstown, RI 02882, USA.

[2] ESRI, 380 New York Street, Redlands, CA 92373, USA.

1574-9541/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2010.09.001

large volumes of data and to provide intuitive filtering pathways for discovering and comparing spatiotemporal patterns. A goal in the work presented here is to adapt and apply map design principles that focus on minimizing the cognitive and perceptual burden of exploring large, complex data sets.

Complementary to a focus on effective design, web-mapping technologies are likely to have the most impact if they adopt open standards in order to enable interoperability of services, allowing components to be easily mixed, matched, and upgraded over time without the need to re-engineer the entire application. In 1999, an effort was begun by the Open Geospatial Consortium (OGC) to establish Web Map Services standards. Version 1.0.0 of the standard was released in 2000 (OGC, 2000) with a draft specification of the Web Feature Services following in 2002 (OGC, 2002). While the OGC Web Map Service (WMS) allows 'a client to overlay map images for display served from multiple Web Map Services on the Internet,' the Web Feature Service (WFS), 'allows a client to retrieve and update geospatial data encoded in Geography Markup Language (GML) from multiple Web Feature Services' (OGC, 2005, p. 10). The latter provides more flexibility with geospatial web-mapping products than merely overlaying maps as images. The impact of retrieving and updating geographic information on demand via a client interface opens the door for web-mapping capabilities to be much more flexible for both the designer and the user of any web-mapping application. An open standards approach is particularly effective when coupled with open source software that allows developers to customize technology to fit application needs and to share development resources easily.

Many initial web-mapping applications have been directed at producing regional atlases (Cobb and Olivero, 1997; Richard, 2000) or supporting public health (Boulos, 2004; Croner, 2003; Kobayashi et al., 2009; Lu, 2004; MacEachren et al., 2008). Recently, other efforts have applied web-mapping methods to environmental and biodiversity data projects (Best et al., 2007; McGuire et al., 2008). Demand exists within these latter communities for methods to support web-mapping for biodiversity data management and knowledge discovery (Guralnick et al., 2007). The focus of the work reported here is on development and implementation of client–server web-mapping technologies to support mapping and analysis of a biodiversity-relevant dataset of California flora made available by the Consortium of California Herbaria (Moe et al., 2009).

1.1. Goals and objectives

The overall goal for the research presented here is to develop flexible, user-centered client–server web-mapping solutions that support information retrieval and knowledge discovery with large, spatiotemporal biodiversity datasets. In this paper, we focus on three specific objectives within this goal: (1) to develop a method that uses spatial and temporal data aggregation to support efficient web-map client–server work with large volumes of spatiotemporal point data, (2) to develop a symbology and symbol scaling method for representing those data in the client map, and (3) to combine the developments above into a flexible, extensible user interface for client–server interactions that facilitates exploration of the Consortium of California Herbaria data set. The third objective, more specifically, focuses on developing interactive display capabilities that allow researchers and others to explore the geographic distribution in species represented by the collection and its relationships to ecoregions (or other relevant data) along with the temporal components of the data that represent trends in collection patterns over time as well as seasonality. In addition to developing and implementing strategies to meet each of the objectives above, we obtained structured input from a group of domain experts using a focus group method as an initial assessment of interface usability and utility. The remainder of this section describes the dataset that we use and the approaches taken in addressing these objectives.

1.2. Dataset

The work presented here addresses two challenges related to the growing repositories of geographically and temporally indexed data. First, we address scaling web-mapping methods, most of which were developed for merging overlays, generating push-pin type maps of relatively small data sets, or for representation of pre-aggregated data in choropleth maps (those that are represented by color fill typically representing low to high data values with light to dark colors). Second, we address cartographic design for interactive maps to support exploratory geovisualization. Most past research and development directed at interactive maps has focused on interacting with relatively small geographic data sets in desktop applications. Specifically, we focus on client–server web-mapping methods for interactively investigating the spatial history of a large, rich dataset of plant collections provided by the Consortium of California Herbaria (CCH). Currently, the CCH "is a gateway to information from California vascular plant specimens that are housed in herbaria throughout the state" (Data provided by the participants of the Consortium of California Herbaria (ucjeps.berkeley.edu/consortium/)). As of December, 2009, users are able to generate push-pin web maps of species with mouse-over access to selected record information (accessed through the Jepson Interchange: http://ucjeps.berkeley.edu/consortium/). Map-based analysis beyond specimen location plotting requires download of data and substantial effort to import data into a desktop geographic information system (GIS) for analysis.

We seek to enhance the potential to use spatially-enabled species distribution information, whether from the CCH or other sources, by creating a user-friendly and powerful web application for querying and displaying attributes of spatiotemporal flora data. In the research presented here, we have created a prototype interactive web-mapping application to demonstrate the potential for direct, map-based exploration of spatio-temporal herbaria data without the need to first download and process the data. Our prototype relies upon a copy of the CCH data that is limited to the data containing georeferencing; the copy was obtained in mid-2008. This subset of the complete dataset contains over 377,000 spatially-referenced plant samples collected in California between 1860 and 2007.

The challenge of accessing a large volume of records is addressed by creating query pathways that take advantage of the hierarchical nature of taxonomic data, allowing the user to drill down through the natural categorical breaks of the phylogenetic hierarchy and increasing levels of attribute detail. In the following, we provide background context for our approach to addressing this challenge.

1.3. Background

In this section, we present background and literature on methods for handling the web delivery and representation of large, point-based spatiotemporal datasets and provide the context within which we developed the novel techniques presented here. Related to our first objective, we first provide brief background on flexible spatial data aggregation strategies. To contextualize our second objective, we discuss existing cartographic methods for representing aggregated point data. Third, we outline key interface design and web-mapping environments for online delivery that we leveraged and/or modeled methods upon as they relate to our third objective. Finally, we provide background on the use of the focus group method for assessing geovisualization tools.

1.3.1. Custom regional aggregation

The generation of spatiotemporal data is currently outpacing the development of methods to analyze, synthesize and communicate those data (Chen et al., 2008). In relation to the specific challenge of facilitating the exploration and analysis of massive spatiotemporal data collections, Andrienko and Andrienko (2007) identify, "a need to combine visualization with computational analysis methods, database queries, data transformations, and other computer-based operations". The interactive, client–server, web-mapping methods described here address this need.

The concept of aggregation is central to our first objective of developing a method for efficient web-map client–server interaction involving large volumes of spatiotemporal point data. Geographic aggregation involves the collection of data at one level of detail into regional or areal features at a coarser level of detail. As described by Andrienko and Andrienko (2005, 2007) and implemented by Kraak and van de Vlag (2007), aggregation can be used in geovisualization as a way of exploring large datasets, making their management easier, and facilitating visual exploration of the data. Taking advantage of aggregation for managing large databases, Fredrikson et al. (1999) work around large volumes of data in drill-down map interfaces by aggregating at broad scales, providing more information as the user zooms in, subscribing to the "overview first, details on demand" mantra popular in both info- and geo- visualization (Chen et al., 2008; Plaisant et al., 1995; Shneiderman, 1996).

Data form has a significant impact on aggregation methods and output. In GIS analysis, data are often aggregated to a grid, counting the number of points that fall within each cell, to build a surface (Andrienko and Andrienko, 2007). While this method is general, our target here is support for spatiotemporal point data, to support exploration of the CCH flora dataset. In that data set, each instance is a field sample or observation of a plant species at a given location at a given time, made by an observer or group of observers. For typical uses of these data and similar flora specimen data, we contend that using more domain-relevant, vector-based polygon collection units (e.g. ecoregions, management units, conservation regions) can make the resulting aggregations more meaningful for many applications.

Aggregation can be based on space, time, and/or attribute. Andrienko and Andrienko (2007) suggest implementing support for temporal aggregation by existing divisions of time as a way of managing temporal exploration when there is a large temporal extent and high temporal resolution. Similarly, Edsall et al. (2000) demonstrate that the aggregation of objects by available time intervals can facilitate temporal pattern and periodicity discovery. For data such as California flora, which can display meaningful temporal patterns at multiple scales (e.g., across years, or by season), supporting data aggregated to months is particularly relevant.

The data of interest here have not only high temporal resolution, but also high spatial resolution. Spatial aggregation provides a solution to the visual overload that would be created by displaying thousands of data points at the same time (e.g., Scrophulariaceae Mimulus, with 6503 specimen locations around the state on a push-pin map would produce a completely illegible set of over-plotted symbols). Using spatial aggregation, query results of user-selected attributes can be aggregated spatially by their presence within one of a set of contiguous, pre-defined sub-regions of a polygon layer, such as simple 5 km grid cells or thematically relevant units such as ecoregions and census tracts.

1.3.2. Symbolizing aggregated data

The second objective of the research reported here is to develop and implement strategies to symbolize aggregated data appropriately to support interactive exploration of space, time, and species. Commonly used methods for symbolizing positional data (e.g., field samples) aggregated to contiguous polygons (census tracts, ecoregions, etc.) include choropleth mapping (maps in which numerical data for contiguous polygons are represented by color fill that typically represents low to high data values with light to dark colors) and proportional point symbols (maps in which size of discrete symbols is used to depict the magnitudes of the data values) (Slocum et al., 2008). There are four reasons for symbolizing aggregated data as point symbols (that no longer represent the original positional data, but instead represent the aggregated regions). First, since the original phenomenon is composed of points, representing the aggregated data with point symbols maintains the conceptual link between the two. Second, point symbols can be more practically used to represent multiple subcomponents of the data at the same time as they represent the whole (such as temporal divisions in the data, a topic discussed later). Third, it is practical with point symbols to separately symbolize potentially relevant covariates and/or contextual variables represented by continuous polygons (e.g., ecoregions). This allows users to visually relate the phenomenon depicted by the points symbols (the primary focus) and the one depicted by the polygons (the related variable); when both are depicted as polygons, paired maps are typically used, making comparison harder. Fourth, Brewer and Campbell (1998) support proportional symbol mapping for count data aggregated to polygons because it does not suffer from the variation of enumeration size common with choropleth mapping, while it preserves spatial structure across the entirety of the map (thus point symbols do not exaggerate the importance of polygons that happen to be geographically large even though they may have a small count).

The use of proportional symbols presents three challenges for mapping. First, a data-to-display mapping algorithm (the method used to translate data values into symbol sizes or temporal divisions) must be selected. Second, a standard scaling size, which determines how large or small the symbols for the largest and smallest data values will be, must be selected. Third, proportional mapping methods for use with large, spatiotemporal datasets must cope with extreme ranges of values present in the data. The remainder of this subsection will address these three issues.

Data-to-Display mapping algorithms translate values in the dataset into symbol sizes on the map. Research on and discussion of such mapping methods for scaling proportional symbols has a long cartographic history, with a starting point generally considered to be an empirical study by Flannery (1971). Building from that and subsequent research, authors of current cartographic textbooks discuss the relative advantages and disadvantages of mathematical, perceptual, and range-graded scaling methods (Kraak and Ormeling, 2003; Slocum et al., 2008). Although early research (and past textbooks) advocated symbol scaling methods that attempted to account for human vision (that, on average was found to result in underestimation of differences in symbol size), current texts advocate for range-graded scaling, where raw data are grouped into classes and each class is represented with a different sized symbol. They consider range-grading advantageous, since map readers are better able to discriminate between symbol sizes and more easily match map symbols to the legend. However, web-mapping makes this advantage less useful for two reasons. First, selecting pre-defined, or automatically-generated, range classes is difficult in the face of undetermined and highly variable data ranges generated by user queries. Second, interactive maps no longer rely on the abilities of users to match map and legend symbols, as interactivity allows users to retrieve precise data values from map symbols directly by mousing-over them. Despite these detractors to the use of range-grading, it has specific advantages related to symbol readability, extreme data values, and data distributions.

A common issue with scaling map symbols is ensuring that while the smallest symbol is visibly readable, the largest symbol does not overwhelm the map. This is one problem that the range-grading solution above addresses indirectly. With range grading, the range in data value from the smallest value being depicted to the largest is reduced to that representing the range from the median of the lowest category to the median of the highest category. If data have a very large range, even range grading will not eliminate the problem of impractical size for the smallest or largest symbols, if the symbols are scaled in proportion to their data value or that of the range's median. This issue is commonly addressed by setting an arbitrary maximum size, scaling all other symbols based on that size, but grouping all values that would result in a symbol smaller that a specified minimum into a common "low" category.

Large spatiotemporal datasets often contain a broad range of values, including extremes. Presenting a solution similar to the one above, Kraak and Ormeling (2003) discuss the issue of handling data extremes, suggesting the use of a threshold below which all values would be represented differently. This type of solution is especially relevant in the context of dynamic web-mapping, where a user may generate multiple maps with different value ranges, and implementing a single scaling scheme, consistently applied across multiple data generations, might result in a confusion of relative scale by the user. A more developed solution would be to map values above or below a threshold (e.g., a box plot) differently (thus range-grading a portion of the values), while mathematically scaling values that do not fall above or below the threshold. This scheme could then be applied consistently across multiple dynamic generations of the dataset.

1.3.3. Interface design and query structures

Our third objective is to combine the above developments into a flexible, extensible user interface for client–server interactions that facilitates exploration of the California flora data set (or other similar data sets). An approach to this goal needs to support multiple levels of aggregation for geography, time, and taxonomy. First, using drill-down (or multi-scale) interface navigation, data in both geographically aggregated and un-aggregated forms can be displayed, depending on map scale. Second, using temporal filtering tools, the entire time span or subsets of time can be specified, depending on the user's demands. Third, the phylogenetic hierarchical structure of flora can be leveraged to guide query design flow in the interface.

Fredrikson et al. (1999) acknowledge the difficulty that exists in creating interactive aggregated representations from spatial, temporal and categorical data for use in data exploration interfaces. However, they recognize the utility of such interactive methods for handling the representation of large data sets, especially those that have rich spatial, temporal, or attribute breadth. Using a drill-down interface, the authors take advantage of aggregating data by spatial, temporal, and attribute dimensions individually, progressively displaying more data as a user "drills down" through levels of aggregation.

Kumar et al. (1997) discuss the importance of "iterative refinement" or "progressive querying" in the process of navigating a hierarchy of data choices. This idea builds on the notion of first providing an overview of the data in question and then, on-demand, the details of that data (Shneiderman, 1996). Designing a query structure around a phylogenetic hierarchy conforms to this concept, starting with the broadest attribute category aggregations first (i.e., plant family) and progressively refining to more detail (i.e., plant genus, species or subspecies), based on user demands. While using a hierarchy such as that found in taxonomy achieves the goal of "progressive querying", hierarchies used in query structures should be flexible and allow user-definition for optimal success.

The concept of "drilling-down" or "overview first, details on demand," also applies to temporal and spatial data. Temporal data can be treated in this manner, by providing an interface element that provides control of time (Andrienko et al., 2000; Harrower and Fabrikant, 2008), through a "time slider" or "temporal navigator" that allows a user to select ranges of time easily. The application of this concept to spatial data results in an interactive, multi-scale map that generalizes target data at a broad scale, revealing detail upon zooming in to a finer scale, as Fredrikson et al. (1999) use.

A method for guiding the design of an application interface such as the one presented in this paper is to develop application scenarios that detail potential uses for the application (Zhang et al., 2007). Having developed prototypical user scenarios early in the development of our application, we extend these in the next section to provide example user perspectives that helped guide interface design and that can be used to help evaluate the success of the completed application.

1.3.4. Usability and utility assessment methods

Given the intended use of the interface by scientists and other experts, assessment using qualitative methods and expert participants is more appropriate than a traditional controlled laboratory study based on accuracy and response time for narrowly defined tasks. Thus, an initial assessment using the focus group method was carried out. Focus groups are used frequently to provide input on usability and utility of computer technology in a range of domains (e.g., Tremblay et al., 2010). For the assessment of geovisualization tools, specifically, they have been used to provide a systematic means for soliciting opinions of typical users, in this case herbaria experts, on the utility and usability of an application (Kessler, 2000; Monmonier and Gluck, 1994; Weaver et al., 2007). A discussion facilitator leads the group through a sequence of questions targeting a small number of topics and attempts to keep the discussion purposefully informal and on topic with probes to prompt elaboration when needed. The approach can provide a good understanding of user preferences, likes, dislikes and reasoning behind these reactions, and overall trends and patterns in participant impressions of the tool (Monmonier and Gluck, 1994). Limitations to focus group effectiveness include less control over discussion than one-on-one interviews (e.g. repetitive statements, off-topic comments), difficulty in analyzing data, and challenges to recruitment and scheduling of participants.

The remainder of this paper will present our application context (the dataset and user scenarios), methods, and results for addressing the three objectives presented in this section. First, the application context is presented with a focus on the data and their pre-processing and on application scenarios that guided interface design. Second, we present the approach and implementation of our solutions for aggregation, symbol scaling, and interface design, including subsections on client–server architecture, query structure, symbology, and interface design. Third, we reflect on insights gained from a focus group session with domain experts to evaluate the application interface, linking back to the application scenarios that were presented as user task exercises in the focus group session. Fourth, we make suggestions for future work. Finally, a summary of our efforts will be given.

2. Application context

Here, we present the materials used in the process we applied to develop a web-based application that utilized aggregation of spatiotemporal point data for exploratory analysis of a large dataset. This includes description of the original dataset and the process for preparing it for use in the web-map interface, as well as presentation of the user scenarios that guided development and design of the interface. Our database management approach adopts open standards for handling geographic data and our implementation leverages open source methods. The combination is intended to make it practical for others to build on our approach.

2.1. Data processing

The design of both the database and query structure was driven by the underlying hierarchical nature of the species taxonomy present in the CCH data. This suggested a drill-down approach to attribute querying as the most effective method for navigating the contents of this large, branched dataset, acting as the guiding influence in the concurrent development of both the interface and database construction. In order to implement a logical, stepped query approach, additional attributes were added to the original dataset. As a result, preparing the database for the client–server web-map involved four overlapping stages: 1) formatting and cleaning the spatial and temporal data, 2) separating and adding additional levels of taxonomic detail, 3) importing the dataset into a PostgreSQL database for access via GeoServer, and 4) defining the query structure.

The raw data were received in mid-2008 from the CCH as a comma separated text file containing 888,227 individual records. As of October 2009, the CCH database contained 1,083,390 records, with approximately half containing coordinate location. Of the records received in 2008, 488,730 were missing latitude/longitude information and were removed from the set used to demonstrate the prototype described here, leaving 387,881 records with spatial reference. The spatially-referenced point locations had varied precision and accuracies, as a result of various geolocation methods, resulting in a large number of the samples having high location uncertainty Aggregation to a broader spatial scale mitigates much of the uncertainty associated with fine-scale, individual point location. The dataset also contained some points lying outside the California border. These were removed, leaving the final set of 377,977 sample point locations. Table 1 details the attribute information fields present in the data table.

Text fields in the data file provide location information, which was often based on descriptions of the sample locations or township/range/section information. When included, the specific tool or method of providing latitude/longitude from this location information was listed as a separate text field, typically one of the following: Maptech, Biogeomancer, Terrain Navigator, or TRS2LL (township/range/section to latitude/longitude converter). Of the points containing datum information, approximately half specified the NAD 27 datum and half specified the WGS 84 datum. While the data file carried a field used to indicate the datum that latitude and longitude were based on, a large number of records had no datum specified. To compensate, records indicating that they used the NAD 27 datum or records that were older than 1984 without datum information were transformed from the NAD27 datum to the WGS 84 datum. Records from or after 1984 or records that used the WGS 84 datum were left untransformed. Transformation reduced linear error to approximately 10–15 m, compared to the 200–300 m of error that would have been introduced if the records using the NAD27 datum had been left untransformed. Final client–server interactions were based on the WGS 84 datum.

Once spatial cleaning was accomplished, temporal information had to be transformed to enable easy sorting and querying on the client. Three different date fields were present for each sample: one containing the original date information in more than 10 different formats (e.g. 2-Jun-76, 04-23-1937, May 26 2000, Sep 1884, etc.), and two fields containing start and end Julian date codes, which were created by previous researchers. For use in client–server communication Julian dates were transformed to epoch milliseconds (both Julian and epoch formats have similar behavior but different temporal origins), using a linear equation to compute the transformation.

Table 1
California flora dataset fields.

Field	Example
Record Number	CDA100002
Taxonomy	*Acer glabrum* var. *torreyi*
Observer(s)	G.F. Hrusa, B.Smith, T.D. Wilfred
Collection number prefix	PNF
Collection number	15174
Collection number suffix	b
Early Julian Date	2449193
Late Julian Date	2449193
Text Date	July 16 1996
County	Pumas
Elevation (varied units)	1900 m
Comments	Sierra Nevada Flora – Plumas National Forest | Base of partially shaded talus slope on USFS 23N11A. Large colony, in site apparently moist, at least seasonally. Shrub, 3 m tall. W/*Prunus emarginata*.
Lat	39.86667
Long	−120,.633
Datum	WGS84
Geolocation method	Biogeomancer
Township and range	20N06E24

To enable a guided, drill-down attribute query, it was necessary to add additional levels of taxonomic detail to the "cleaned" set of records. The taxonomy specification in the source data was formatted as a single text field with combined Genus, species and sub-species (where applicable) information for each sample. To support hierarchical query, the Genus and species + sub-species information was split apart into two separate fields, and an additional Family field was added, representing the top level of the hierarchy. The Family name that corresponded to each Genus in the dataset was added to each record using semi-automated methods. The necessary information to assign each record to a family was obtained through the U.S. Department of Agriculture State Plants Checklist for California (http://plants.usda.gov/dl_state.html). This listing of all known plants in California was parsed for the Family and Genus information, and a list of all unique Family/Genus pairs was created. From this array, a Microsoft Excel VLOOKUP command was used to fill in the missing Family information by matching the Genus in our dataset to the Family/Genus USDA pairs. The result was 3 separate levels of taxonomy for each sample: Family–Genus–Species (+ sub-species) (Table 2).

2.2. Application scenarios

As discussed above, we developed application scenarios to guide application and interface design. Two application scenarios are presented below. These application scenarios provide example user situations and their resulting demands on the interface of the application. More specifically, the scenarios illustrate how users can browse the California Flora dataset, execute queries with space–time–species constraints, and prompt hypothesis generation about space–time patterns. In addition to their use in guiding application and interface design by helping to identify key tasks that the interface must support to be successful, these scenarios were incorporated as user tasks carried out by expert participants as input to the focus group assessment. Insights derived from participant performance of these tasks are discussed in the assessment section of the paper below.

In the first scenario, a botanist at a small university would like to visually analyze the seasonal distribution of a single flowering plant species in California. To start, she wants to confine her search to a single species, Orange Bush Monkey-Flower (*Mimulus aurantiacus*), using only samples from the most recent twenty years. With this information mapped, she wants to study, specifically, where the species is distributed throughout the state, making visual connections between the samples and where they are found in California's ecoregions. Digging deeper, she wants to understand how the species varies seasonally and how that seasonal variation differs across the state of California. To adequately support this goal, she needs the ability to flexibly select her target species and time range, generating a map that displays the result using a symbology that allows her to explore seasonal variation geographically.

Table 2
Original taxonomic information ("Before") is expanded to include Family, Genus, and species as separate hierarchical fields ("After").

Before	*Abies concolor* var. *lowiana* (Genus + species)		
After	Pinaceae (Family)	*Abies* (Genus)	*concolor* var. *lowiana* (species)

In the second scenario, a field crew leader for an invasive plant species control project is planning a field season for mapping existing and new areas of invasion and for directing control efforts. He needs to find locations where Purple Loosetrife (*Lythrum salicaria*) has already been found and where a field crew might find more in an area that is at the edge of the plant's existing range in California. First, he wants to start at a broad scale, to identify regions in California where the plant has been collected and during what time of year the plant has been collected frequently. Second, he will need to study the distribution of samples at the landscape scale, to see exactly where the plant has been collected, so that he can direct his field crew to potential locations for more detailed mapping of new areas of invasion, to direct control efforts, and to monitor for further expansion. In supporting this goal, the interface needs to provide drill-down capabilities, giving him the ability to generate, first, broad-scale maps to identify range boundaries, and second, fine-scale maps to identify specific locations for directing action.

3. Implementation

This section presents our implementation of the methods used to accomplish the primary research objectives identified above. Specifically we detail methods implemented to perform aggregation of spatiotemporal point data, symbolize that aggregated data, and design and build a client interface for data exploration. We first discuss the specifics behind aggregating the point data in the database and structuring the database to facilitate hierarchical querying and temporal focusing in the client. Second, we discuss solutions to the symbol scaling problem and how those were implemented in the client. Then, we present an overview of our client and its use.

3.1. Client–server structure

PostgreSQL, PostGIS, and the GeoServer open-source database and server-side software were employed to efficiently process query requests and serve results to the client, which was built using Adobe Flex and compiled for a web browser into an Adobe Flash application. The completed, cleaned Microsoft Excel data table was imported to PostgreSQL. This open-source relational database provides powerful search, query, update, and management tools and can support databases several orders of magnitude larger than the 377,977 records managed here. Access to the database by the Flex application is in turn provided by GeoServer, which acts as a gate-keeper web service to the actual PostgreSQL data table. Based on the user-specified query, a WFS request is sent from the Flex-based client to GeoServer using GML filter tags. GeoServer then translates this request into an SQL statement that queries the PostgreSQL database. The results of the SQL query are then gathered and returned to GeoServer, which converts the resulting data to GML. These results are then sent back to the Flex-based client to be parsed and displayed as graphics.

Following a straightforward PostgreSQL import of the cleaned dataset, a geometry type field was created from the latitude and longitude values for use by the PostGIS extension of PostgreSQL. Creating a geometry type field, instead of storing the geographic data in simpler numeric type fields, allows for easier spatial analysis within the database and retrieval through GeoServer.

A set of polygons for aggregation were selected from the four-tier (province, region, subregion, and district) system of ecologically-defined units from The Jepson Manual (Jepson and Hickman, 1993). Within California there are only 3 provinces and 10 regions, which would have made for overly generalized aggregations of the data. However, the 35 sub-regions (referred to by the data provider and hereon as “sub-ecoregions – see below) made a suitable level of aggregation as it provided distinct regionalization within the state, but did not occlude an overview of the entire region with too many map symbols. We were unable to find an enumeration of districts, but it would have likely produced too long of a list of regions to feasibly aggregate and display.

An ArcGIS polygon shapefile of Jepson sub-ecoregions was acquired from the Information Center for the Environment at the University of California, Davis. It was loaded into the database using the PostGIS utility (shp2pgsql) and an SQL script tagged each point spatially according to which region/polygon it fell within. An arbitrary numeric code corresponding to each sub-ecoregion, based upon the results of the previous script, was added to the PostgreSQL table for each point in the dataset and served as the basis for the spatial aggregation and display of the query results. This numeric code allows for an efficient direct lookup instead of the intensive spatial lookup. Database performance was further enhanced by indexing sub-ecoregion and point geometry values, improving query efficiency. Prior to indexing, database queries took as long as three to 4 min to complete, compared to just a few seconds after indexing, greatly improving usability.

Aggregation of the data to Jepson sub-ecoregions, regions pertinent to this dataset, was done for three reasons: (1) to demonstrate a key capability of the system – to aggregate field-collected point data to user-defined contiguous regions using regions of interest to ecologists who use the herbaria collection; (2) to reduce large volumes of web client data transfer, making it possible to generate maps on-the-fly in response to user-supplied choices; and (3) to generate useful overviews of a dataset that is too large to make the display of all raw data at once practical, from both a data transfer and visual display perspective. Regarding the third point, plotting raw data is inconvenient considering commonly available user internet bandwidth and the sheer number of sightings in the database that could be delivered in response to a given query. While it is technically possible to display all the data, the volume here would fill about 1/3 of the pixels of a 1280 × 720 display if only 1 pixel was used to depict each point and many locations are closer together than a one-pixel width at display size; thus abstraction is needed. In addition, aggregating to sub-ecoregion expedites server–client delivery such that drawing in the client can happen in a reasonable amount of time (<15 s) with access to decent bandwidth (DSL or better).

3.2. Query structure

An underlying goal for our web-map application is to support flexible, on-the-fly map generation based on user-specified constraints. Each of the four primary data categories were possible entry points for user driven filtering: taxonomy (family, genus, species, and sub-species), time (sample collection date), space (latitude/longitude, elevation, county, Jepson sub-ecoregion), and observer. While querying within each of these fields provides a unique and valuable method of focusing the data, our prototypical application scenarios indicated that typical users, such as botanists, are likely to focus first on a specific plant Genus or species, secondly constraining the data displayed by time, place, or observer. This assumption, combined with the hierarchical structure of the taxonomic data suggested that the query process should start with higher-level (Family) plant information and then provide the ability to filter additional attributes once a subset of plants has been selected. Hierarchies such as this are one of many possible standard hierarchies that could be used to guide query structure. In our case, alternative family classifications do exist. An ultimately flexible query structure design would allow users to define the hierarchy itself.

The inherent phylogenetic hierarchy present in the California flora dataset presented a challenge for interfacing with such a large database, as plant taxonomy categories (family, genus, species, etc.) are not logically queried independently (one would not want to search for all records based on species name alone, but instead genus and species, after family). However, this hierarchy also presented an opportunity to design a natural progressive query that started with plant families and offered more detail as users selected into the hierarchy. The concept of a hierarchical menu, itself, was adapted from Bhowmick et al. (2008), who, based on user studies, judged it to be a successful interface organization.

As few users know the family of a given plant species, requiring the user to start a query by selecting a family name is not ideal. A flexible structure would use a more general entry point, allowing the user to select a familiar and general category, such as life form (tree, shrub, etc.) or common name group (chestnuts, orchids, etc.), with subdivision of the query proceeding based on user selection. For example, the Jepson Interchange (http://ucjeps.berkeley.edu/interchange.html) allows users to search by scientific or common name, list by county or bioregion, and browse by thematic qualities (e.g., native, endemic, or new). While a structure of this kind supports the notion of a user-defined query structure, datasets are rarely laterally-flexible enough to provide a number of linked categories from which to choose. To help guide the user in their initial selection of a family, we have placed a button linking to the Jepson Interchange near the family selection dropdown menu so that they may go off-site and use the Interchange to find a specific, scientific family name, returning to the client to find it in the dropdown menu. Providing further taxonomic context, buttons are placed to the right of both the family and genus dropdown menus, linking to Wikipedia entries for selected items in those menus.

Within the client, the queries for the subdivisions of the taxonomic data are arranged in a logical, top–down order, beginning with the broadest division, Family, then progressing through Genus and species. At each level of query, once a selection is made, the request is submitted to the GeoVISTA servers at Penn State and the next level of results are returned to the Adobe Flex-based client for further querying. The total number of samples matching the query at any stage is reported in the "Current Feature Count" and informs the user of the effects of each query choice on the total set of matching features. At any step in the Family/Genus/Species selection, the results can be plotted on the map pane by clicking on "add". These results are plotted on region centroids for the Jepson sub-ecoregions, with the data having been previously aggregated to those sub-ecoregions as described above. Within the map pane itself, the user can query the names of individual Jepson sub-ecoregions by mousing-over the symbol of interest. If a user progresses through the hierarchical query, selecting a single species, results can be further filtered by the names of individuals who have collected that particular species.

Temporal filtering can be done at any point in the selection process. There have been a wide array of novel temporal interface tools proposed to support query and analysis of time series data (e.g., André et al., 2007; Hochheiser and Shneiderman, 2004; Javed and Elmqvist, 2010). However, since the temporal controls in HerbariaViz are intended primarily for simple filtering that supports subsequent exploration of the spatial patterns in data retrieved, we opted for a minimalist temporal control. The Herbaria dataset contains plant sample data collected between 1860 and 2006, and the results can be focused using the two time sliders in the interface. To give a simple visual overview of the temporal distribution for the selected attributes, the interface incorporates a time-series sparkline (Fig. 1); a sparkline is a small schematic graph that provides a quick but unobtrusive overview of a data distribution. For detailed a discussion of sparklines as an information visualization tool providing quick data summaries in small spaces, see Tufte (2006). A sparkline is drawn and updated at each query step, providing an easy way to visualize and interact with the time domain to focus on specific collection histories. While dynamic update of the map as a user moves the temporal control slider would be ideal, the client–server methods used here to allow the system to work on the web do not support this level of interactivity for a data set this large.

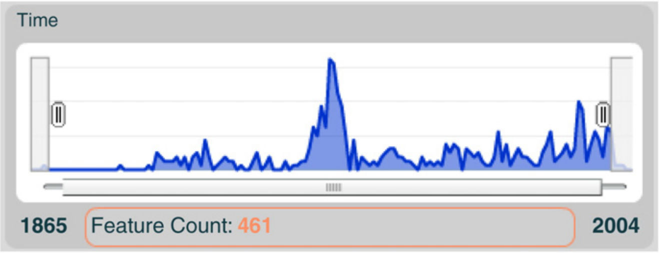


Fig. 1. The time-series graph, with sparkline, drawn at each query step.

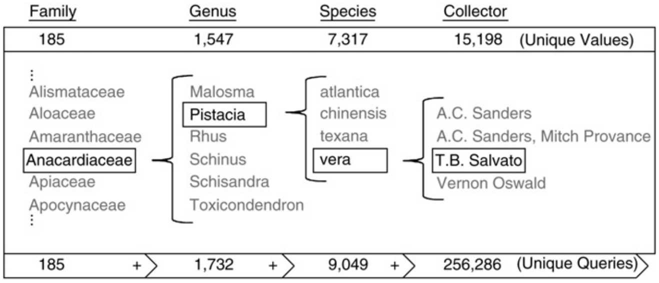


Fig. 2. All possible query combinations, including unique values for each field and each selection hierarchy.

The total possible query combinations using the taxonomy and observer attributes are summarized in Fig. 2. The top of the figure shows the number of unique values for each individual attribute, while the bottom shows the cumulative number of unique results at each step of the query hierarchy. It is worth noting that these cumulative figures do not take into account the additional ability to temporally filter the results at each step, which would multiply the existing number of unique queries by 146 factorial (1.17 e254), as the dataset currently spans 146 years.

3.3. Symbology

As discussed above, using graduated point symbols to represent the aggregated count data has multiple advantages, including enabling users to easily see the distribution of a species in relation to the ecoregion map being used as a background. Using graduated point symbols and developing an interactive web mapping environment that supports user-generated maps based on visually-specified queries to the database requires developing and implementing symbol scaling procedures that produce interpretable and informative maps while also being flexible enough to deal with large geographic and temporal differences in count data for specific user selections, particularly when the data distribution for the selection to be mapped is highly skewed (a typical situation). Our solutions for implementing these facets follow.

3.3.1. Symbol choice and design

The most common form of graduated symbol for representing numerical data on maps is a graduated circle in which the area of the circle is scaled in proportion to the data value (or when values are grouped into classes, in proportion to the class median). We implemented this method as one of two options, thus enabling users to get a full impression of the geographic distribution of any species of interest for the time specified.

The second, user-selectable, symbol option implemented is the coxcomb (Fig. 3). Coxcombs enhance user opportunity to identify temporal periodicity present in the dataset. Specifically, the coxcomb, or polar area chart, uses slices to represent data count totals for ordered sub-divisions, in this case months of the year. As plants have many natural cycles (appearance, disappearance, blooming, fruiting, etc.) tied to seasons, the use of months as the temporal divisions in the coxcomb makes it possible to easily identify temporal periodicity

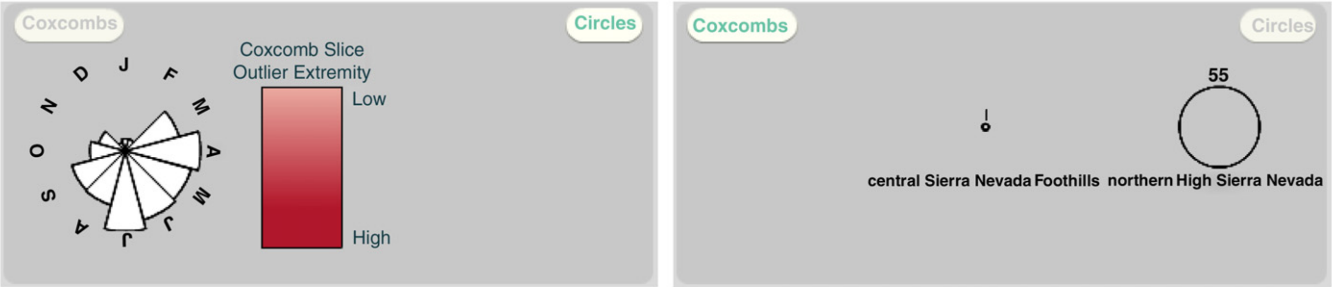


Fig. 3. On the left, the legend for Coxcomb symbolization, showing monthly totals for all aggregation regions. On the right, the simpler legend for graduated circle symbolization, showing the largest and smallest circles on the map.

found in the data. Coxcombs, however, are more complex than graduated circles and require more consideration when choosing data-to-display algorithms for mapping.

The underlying code for HerbariaViz incorporate a modular design that allows the addition of different symbolization methods; the graduated circles included as an option in the current interface are a proof-of-concept that this modularity works. Since we developed the system as open source software, other developers have the option to add their own symbolization methods to fit their needs.

3.3.2. Symbol scaling

Our discussion of the literature identified the need to define a data-to-display algorithm, a method for determining the size of a standard value that others will be scaled in relation to, and a way of handling outliers and skewed distributions in user-generated data ranges. A single solution was to group data values into classes defined using box plots (Tukey, 1977) and then represent the class each value falls in rather than the specific value. The resulting map is less precise than a map of the raw data, but will typically be much more understandable; and users who want to know specific data values can retrieve the values easily using a mouse-over of any symbol. The method is implemented as follows.

When a user makes a data selection and generates a map, the client calculates a box plot for that selection. The box plot is then used to determine how either the coxcomb slices or the circles representing each sub-ecoregion are symbolized. Rules were set to scale and color each symbol differently depending on where the symbol's value falls on the box plot.

Two visual variables are used for each symbol: size and color. The size for all values between the upper (right) and lower (left) whiskers (all non-outliers) are determined directly by the data value. Color is used to represent whether data values are between the whiskers or are outliers beyond them (either above or below). Values above the upper whisker (outliers that are beyond the upper quartile plus 1.5 times the interquartile range) are represented by a symbol size equal to that for the upper whisker, but are colored red. Values below the lower whisker are represented by a symbol size equal to that for the lower whisker, but are colored blue. By using the box plot in this way, the data-to-display algorithm is defined, as is the method for scaling the symbols based on the user's data range selection. Additionally, by coloring based on box plot, outliers and skewed distributions that can be found in many data selections are symbolized in a way that makes it easy for the user to understand the distribution of data values they selected. When a user places their mouse over a Coxcomb slice for a particular month and region, its value is brushed in a box plot legend to assist in helping the user match symbol values to the legend.

3.4. Interface design

The most novel aspect of our application is its ability to give a user easy, quick access over the web to a large collection (>377,000 records) of spatiotemporally referenced plant collection data. This success is achieved through the use of relevant spatial aggregation, the inherent hierarchical structure in the data that allows easy navigation, and the application of a variety of interactive geovisualization methods originally developed for desktop mapping that are transitioned to the web-based maps used here. This section describes the interface flow, starting with phylogenetic hierarchy selection, through temporal focusing, and finishing with map generation.

Design of the interface for this tool revolved around navigating the hierarchical structure for querying and symbolizing queries efficiently. The final interface is published as an embedded Adobe Flash Small Web Format (SWF) file. Programming was done using ActionScript 3.0 and MXML in Adobe Flex Builder, leveraging the highly graphical nature of that software and its ability to easily interact with a web-server / database configuration. Open-source code can be found for free in the GeoVISTA Resource Library (http://www.geovista.psu.edu/).

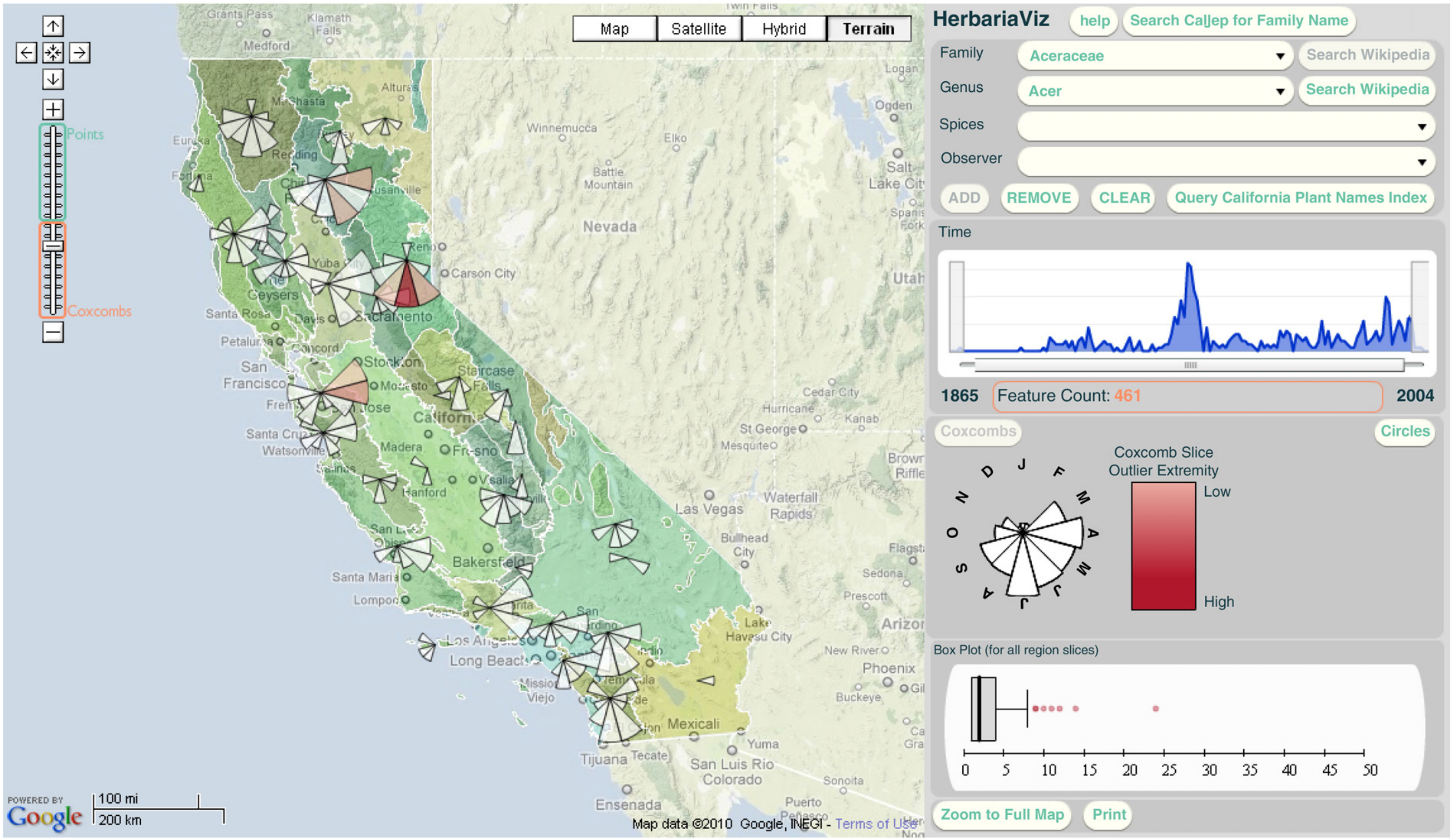
A method for navigating the phylogeny of the dataset was required to allow simple attribute query. There is limited attention in the literature towards design of query interfaces for hierarchically organized data. As previously introduced, our solution was to first ask the user to make a taxonomic selection, starting with at least a family, optionally progressing to the finer levels of genus and species. If the query progressed to species, the observers of that species are also available for filtering. However, this query structure is not optimally flexible, as a user may want to start with observers, instead of the taxonomy. Future implementations should work to make query structures more customizable by the user, even before they interface with the map and visualization.

Once the hierarchy is navigated to the genus level, it is possible to focus the data query temporally. A small, navigable time series graph is drawn for selections of genus, species, and observer, providing the user a temporal reference of data density. This information helps users to avoid blindly selecting periods lacking data. Instead, they can make an informed temporal focusing choice that will produce relevant maps with limited trial and error. Interactive slider handles allow a user to select the temporal query window (through starting and ending dates).

The remainder of the web-map interface development followed traditional interface design guidelines. Progressing from hierarchy navigation, through temporal selection, and finally to map generation, the user has flexibility in making changes and stopping at various levels in the attribute-time hierarchy. Map generation following querying and filtering creates a map depicting totals for each polygon to which data are aggregated (sub-ecoregions here). Maps can be cleared quickly for a new query. Final query results are displayed on the centroids of aggregating polygons using proportional circles or coxcombs that provide additional information. For a full overview of the interface, see Fig. 4 or visit (http://www.geovista.psu.edu/herbaria/v3/index.html).

3.5. Data limitations

Two particular aspects associated with the dataset that we used present limitations to knowledge that can be derived from using our application. First, skew as a result of biased collection and museum



Family: Aceraceae

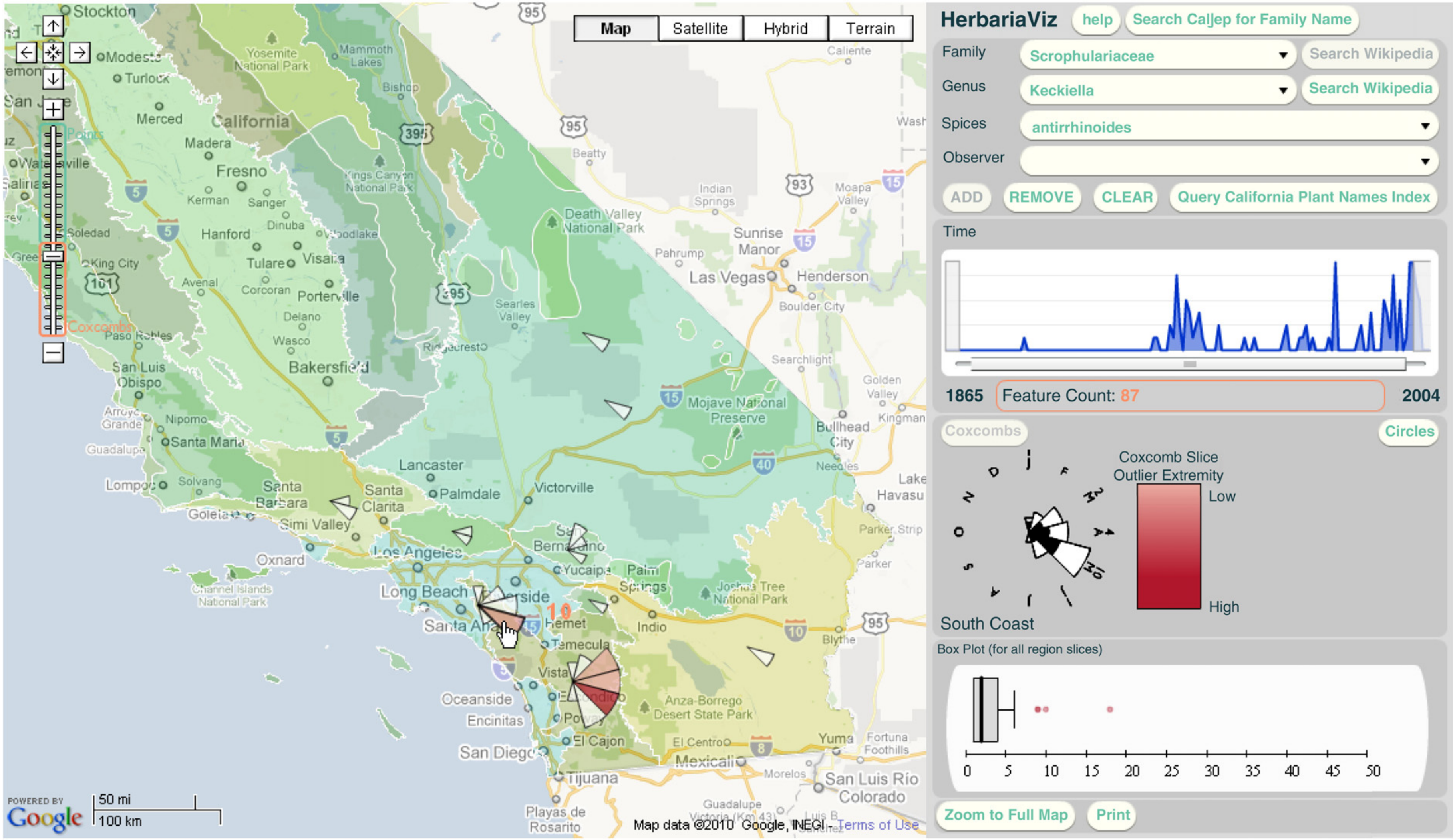
Genus: Acer

Spices:

Observer:

1865 Feature Count: 461 2004

Fig. 4. A screenshot of the interface, showing mouse-over highlighting.



HerbariaViz

Family: Scrophulariaceae

Genus: Keckiella

Spices: antirrhinoides

Observer:

Time: 1865 – 2004

Feature Count: 87

South Coast

Box Plot (for all region slices)

Fig. 5. A screenshot of Snapdragon Penstemon, showing a peak of collection in late spring.

management add further noise to the distribution signal of specimen records in the database. Second, discrepancies in taxonomic classification may impact how general the results of a query may be. In this section we discuss these two limitations, present an example based on our application, and discuss how our interface may help users cope with some of these limitations.

Biodiversity datasets formed from non-random collection methods are subject to sampling biases (Hijmans et al., 2000). Particular biases related to plants include specimens: from easily accessed areas, that are easy to study and identify, collected during dry seasons and those found near roads, rivers, towns, and biological stations (Funk and Richardson, 2002; Graham et al., 2004). Collectors tend to be opportunistic, sampling specimens that are either the focus of their efforts (rare species) or those that are easy to identify (Williams et al., 2002), resulting in a disproportionate representation of both rare and common species in museum collections, such as the ones that the dataset used here draws from. While reconciling such sampling biases for use in modeling species ranges and occurrence is a difficult and necessary task (Loiselle et al., 2008), the maps generated in the application presented in this paper may help researchers to more readily identify these kinds of bias. Fig. 5 shows an example of Snapdragon Penstemon (Keckiella antirrhinoides), a shrub that is largely noticed during the brief period that it flowers in the latter half of the spring season. As a result, it has a high occurrence of collection when it is known to flower, compared to the rest of the year when it is not flowering, despite the fact that it is still present.

Our application may potentially help a user identify some of these skews. By aggregating the data temporally, seasonal patterns in collection may be revealed, and a knowledgeable user may be able to take advantage of the display we offer to infer that such a temporal distribution does not conform with the true, natural distribution of a plant species (a perennial only being collected during summer months). Further, the absence of physically large species (such as trees) in areas that a user knows a plant to exist could potentially alert the user to the fact that a species may be under collected in a particular region. Finally, allowing drill-down spatial representation, with individual record mapping may allow a user to notice that records are clustered near roads, in parks, or on public land, as seen in Fig. 6.

However, in the end, the ability of the user to ultimately identify skew in the data will depend on the depth of their knowledge as it relates to the underlying distributions.

A second concern is that of taxonomic uncertainty, which presents challenges to all research involving biology (Isaac et al., 2004). Addressing research on taxonomic standards and the problems they pose to using biological data is a large and complex topic that is largely beyond the scope of this paper. An effort by the Taxonomic Data Working Group (TDWG) to put forth a Taxon Concept Schema (Kennedy et al., 2005) for addressing problems with varying taxonomic standards and classifications represents a good starting point for seeking to understand concepts related to the topic.

Relevant to our work, consider the application scenario example using Orange Bush Monkey-Flower (*Mimulus aurantiacus*). First, the family placement of this species has not been agreed upon, with records falling under either Phrymaceae and Scrophulariaceae. Second, infraspecfic taxa are often not agreed upon. In the *M. aurantiacus* example, botanists disagree as to whether it belongs in either the Mimulus or Diplacus genus. This means that a query may locate only a portion of the total records for a given species. The following example, showing nine variants of *Mimulus aurantiacus*, demonstrates this issue in Fig. 7. Compare the example of *M. aurantiacus*, to that of *Rosa minutifolia*, or Ensenada Rose, in Fig. 8, which shows an extremely limited range, having only been collected in a very small corner of the state where it has been discovered.

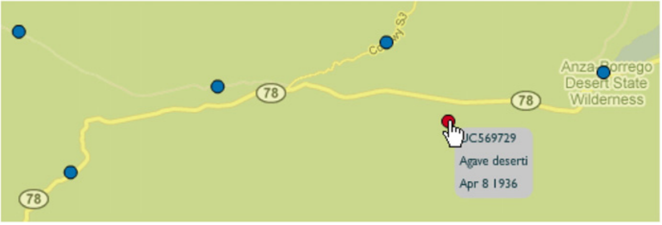
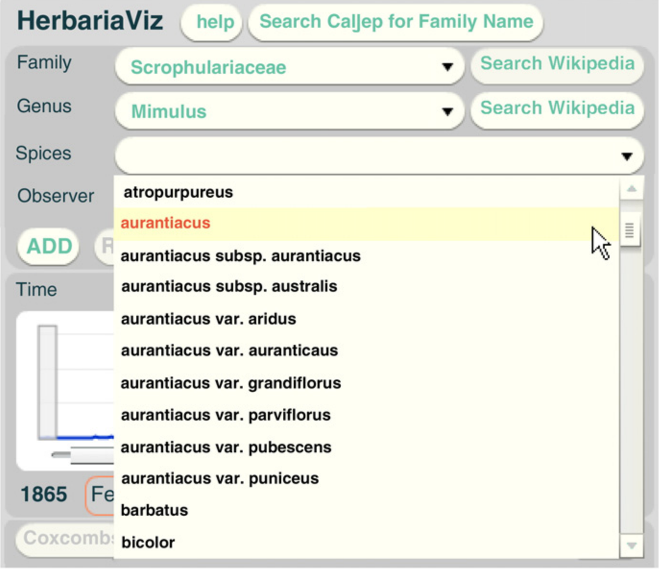


Fig. 6. A cluster of Agavaceae collections, with all but one being recorded near a prevalent road.



Family Scrophulariaceae

Genus Mimulus

Spices

Observer

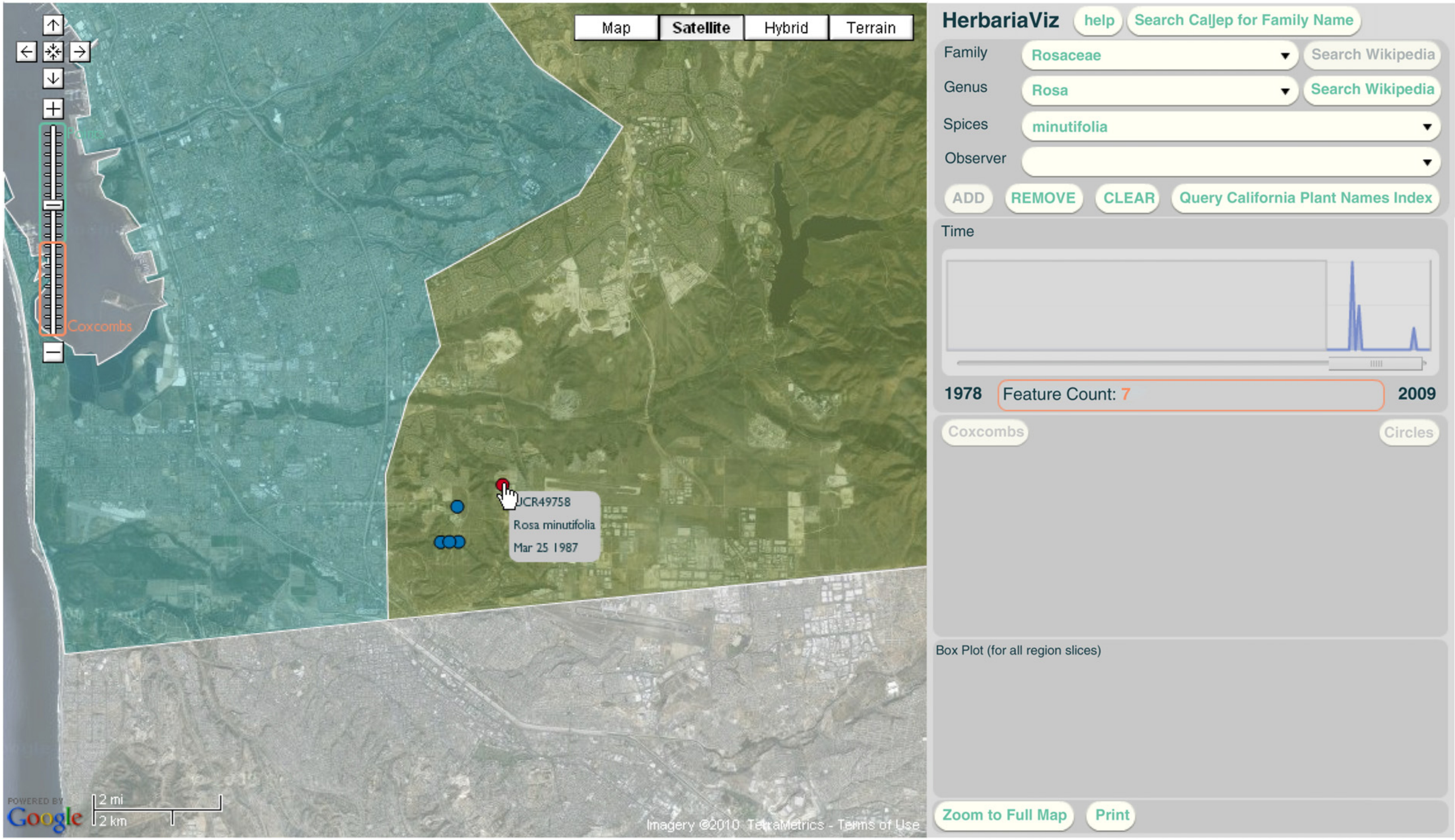
atropurpureus
aurantiacus
aurantiacus subsp. aurantiacus
aurantiacus subsp. australis
aurantiacus var. aridus
aurantiacus var. auranticaus
aurantiacus var. grandiflorus
aurantiacus var. parviflorus
aurantiacus var. pubescens
aurantiacus var. puniceus
barbatus
bicolor

Fig. 7. The nine variants of *Mimulus aurantiacus*, including both subspecies and variant entries (some of the nominate race).

A potential solution to this problem relates to our earlier discussion of flexible, user-defined hierarchies. An ideally adaptable query structure would allow a user to custom group subspecific taxa and generate a map based on that grouping, as opposed to requiring the user to select individual hierarchy instances that may be subject to disagreements in classification. Existing research relating to methods for visualizing relationships between taxonomic entities and across classifications (Graham and Kennedy, 2007a,b) has developed methods demonstrated to be useful in helping ecologists study and understand how taxa are differently classified, leveraging concept annotation to link specimen data categorized under different classifications (Graham and Kennedy, 2007b). However, to our knowledge these methods have not been implemented with map interfaces and are potentially too complex for the purpose of simply selecting a species for spatial and temporal exploration. Incorporating such methods without spatiotemporal tools is beyond the scope of the present research. But, subsequent research is clearly needed to investigate strategies for integrating advances in how the complexities of taxonomic classification are communicated with geovisualization methods that leverage the spatial and temporal components of data.

4. Usability and utility assessment

A focus group was conducted with three researchers from Penn State University who have research experience in botany and plant life ecology ranging from seven years to over twenty. As domain experts, these participants provide valuable insight informed by knowledge and problem solving strategies representative of target users of HerbariaViz. While the focus group was small, the group assembled represented a high level of expertise and were very



HerbariaViz

Family: Rosaceae

Genus: Rosa

Spices: minutifolia

Observer:

Feature Count: 7

UCR49758
Rosa minutifolia
Mar 25 1987

Fig. 8. A screenshot of Ensenada Rose, showing a cluster of collections in a very small area near the Mexican border.

engaged in the process of reviewing HerbariaViz functionality and providing input on its usability and utility.

The session was conducted over 90 min. It began with a twenty-minute demonstration of HerbariaViz, followed by completion of two seven-minute tasks (modeled on the scenarios described above) that were completed by each participant at an individual personal computer. After participants completed the tasks using HerbariaViz, a structured group discussion was carried out. This discussion was moderated by a PhD student with training in software use and usability who was not involved in the design or development of HerbariaViz. The discussion targeted input on four primary issues: taxa query interface, map design and interaction; timeline design and interaction; overall utility.

Data from the focus group include informal observations made while participants were completing the two tasks plus systematic processing and organization of input from the structured discussion. Data were collected through audio recording, transcribing the audio into text, and then organizing the text statements by participant. Statements were treated individually and coded (using codes generated prior to coding) with both topic and positive/negative impression, with some statements receiving multiple topic codes. Our discussion of the results will cover, first, responses regarding overall utility, then input on feature usability, followed by the suggested extensions to the application, and, finally, participant success with the task-based exercises derived from the application scenarios presented earlier in this paper.

4.1. Overall utility

Each of the participants commented on the utility of HerbariaViz for a number of tasks during the demo, exercise and discussion portions of the focus group. During the discussion session, P1 indicated (and the other participants concurred) that HerbariaViz would support their activities better than tools currently being used. See Table 3 for a summary of positive and negative comments on each topic, showing a generally positive response with the exception of comments on query.

In relation to HerbariaViz utility, participants indicated multiple possible applications that they felt the tools could support, with an emphasis on the ability to understand spatial patterns as they vary over time. During the focus group discussion, P3 noted specifically that the tool could be used to explore phenology by region and the variations over time.

> P3: "...from north to south. You can tell that cheat grass lasts longer in the south, has a longer growing season in the north and if you just had points you would say, ok it's there. And you wouldn't be able to see that seasonality. So this could also be very useful for phenology were you might be able to see if you took at a narrow sliver in time and just moved it forward to see if that seasonality is changing..."

P1 noted the utility of the tool for both examining spatial–temporal patterns of invasive species and special status species.

Table 3
Coded focus group statements by topic and impression.

	Subtotal	Positive	Negative
Coxcomb	5	4	1
Map	7	6	1
Outliers	1	–	1
Query	4	–	4
Timeline	3	2	1
Usability	11	7	4
Total	30	19	11

> P1: "I would think like we saw with the purple loosestrife that was interesting to look at the distribution over time and see how it came in the 70's and it had a peak...But for tracking both special status species and invasive species that would be a really great tool because it covers that big time period. You can see when the invasive species came into the state, you know, given that somebody went out and collected it."

P1 expanded on this idea to identify potential users as botanists and environmental, including the impacts of climate change scientists maintaining natural heritage programs and policy analysts examining the effect of exotic species management programs. P3 agreed, commenting that those working for the US federal government could make use of the tool provided the government convention for using abbreviations was supported.

4.2. Feature utility

Following the brief demonstration, participants indicated that they found the HerbariaViz interface easy to use. Positive comments on the usability of the map identified the ease with which seasonality could be interpreted based on the aggregation of data by month and region presented through the coxcomb plots and the supplemental point presentation of individual observations at larger map scales. The placement of the coxcomb plots at the center of each sub-ecoregion (to represent the aggregate total for that region) was not clear to one participant initially. This participant initially expected each coxcomb to represent a specimen collection site, thus that the location of the coxcombs represented specific samples. This aspect of the map design became clear to the user upon further inspection of the map.

> P1: "Ok. Alright. I did not realize that until now. It just puts them right in the middle of the [sub-]ecoregion. It does not tell you anything about the distribution in that ecosystem."P3: "Yeah, you have to drill-down to the points view for that."

The scaling of radial lengths for coxcomb wedge symbols and the hue and value color encoding of these wedge symbols did require some clarification. In each case, the participants had the correct interpretation of these encodings but were not sure until the moderator confirmed the interpretation to be correct. The proportional scaling of coxcomb wedges for small frequency count ranges on the map (i.e., ranges from 1 to 4) versus the radial length scaling for the aggregate view on the legend did produce some confusion. Exploring a larger set of observations eliminated this confusion by providing greater variation in frequency counts on the map. Both genus and species filters and time interval constraints affect the number of data points represented on the map.

The timeline generate only one suggestion for improvement. P2 suggested that adding a means to directly enter dates would be useful.

> P2: "I would actually suggest instead just the timeline that you would actually have boxes where you could type in a specific range of time or just single date instead of having to do the slider bar."

Quoted earlier, P3's comment referencing the exploration of phenology by dragging the slider successively through time intervals suggests the need for both forms of interaction with the time element of the application.

The majority of usability problems focused on the evolving nature of taxonomic groups used to name flora in the CCH data. Participants suggested support for synonymy and for cross-referencing multiple versions of the Jepson manual to enable querying species that are known by different Latin names over time and in different communities of practice. Doing so would help overcome some of the problems presented by changes in species classification that has occurred since the date of the specimens first collected over 150 years ago. Tools for accessing external references intended to help users relate common names and scientific names were also found to be difficult to use.

> P2: "[Tools for accessing Jepson manuals or taxonomic grouping should be] dynamic where it could be updated."

4.2.1. Suggested extensions

Participants generally presented an enthusiastic attitude towards the tool, suggesting a number of ways the application would be useful in its current form. However, they also identified the need for some extensions and offered constructive suggestions. The primary extensions envisioned for the tool were: (a) include the full set of species listings in the Jepson manual, not just for samples represented in the collection, and (b) support lookup functions across taxonomic groups. The participants also commented that HerbariaViz would be a useful tool in their own work in the mid-Atlantic region of the US if populated with other regional flora datasets.

4.3. Application scenario task exercises

The focus group session included a section where participants completed two brief exercises, on their own, at a personal computer. These exercises were based on summarized versions of the two application scenarios presented earlier in the paper. Participants had no difficulty completing the tasks successfully within a reasonable amount of time. Interaction with the second of two tasks (the example involving Purple Loosestrife) lead to positive discussion of the application by the first participant, as highlighted above in Section 4.1. Results showed that participants were able to both read the coxcomb symbology properly and identify the sub-ecoregions a coxcomb symbolized.

One critical incident observed during the exercise portion of the session revealed the need for greater support and refinement of the scientific name lookup functionality (in usability engineering, a critical incident is considered to be events that have an important effect on the final outcome). When asked to add records for Orange Bush Monkey-Flower, one participant knew the genus but needed to confirm the family. This participant followed the link labeled in HerbariaViz as the tool to access the CalJep website with a utility for family lookup by common name. A search by "Orange Bush Monkey Flower "did not produce a result. The user located links on the site for taxa by genus that produced an answer, but this manual search of an external website seemed unnecessarily inefficient. Last, the result produced was the contemporary naming convention – *Mimulus aurantiacus* as a member of the family Phrymaceae. However, the herbaria list observations for Orange Bush Monkey-Flower under the family Scrophulariaceae. The other two participants abandoned the above strategy of following links from within HerbariaViz, as presented during the demonstration, and choose independently to perform an external web search to obtain the required family, genus, and species.

This incident independently identified the issue associated with the ability of users to successfully utilize the fixed taxonomic hierarchy query structure, an issue we had suspected existed. In addition, query was the only topic in the focus group session to receive more negative statements than positive (4 negative, 0 positive). It is obvious that when combined with the problem of inconsistent species classification, a fixed taxonomic query structure requiring a user to first begin with a family name is not an entirely successful design. Regardless, its use did not prevent our focus group participants from finding utility in the application. As discussed earlier in this paper, a more flexible, user-defined query structure would likely be more suitable and can be identified as a necessary area of further research. Documentation of poor user experience with this particular query structure and generally negative comments about it should not only help us improve design of HerbariaViz, but may help others seeking to design query interfaces for taxonomic.

The query interface implemented here forces users to know the taxonomy of a species to initiate a query that builds the map of that species. This approach has obvious limits; if a scientist knows the species they want, it would be most efficient to query for that species directly. Making the spatial aggregation and online mapping methods introduced here useful for a wider range of users will require implementation of a more comprehensive and flexible approach to query that allows users to enter the system in multiple ways; initial ideas for extension include adding search using common names, using free form text entry with relevance ranking of matching and partially matching results, using faceted query, and using visual interfaces that display database metadata using various information visualization strategies.

5. Future work

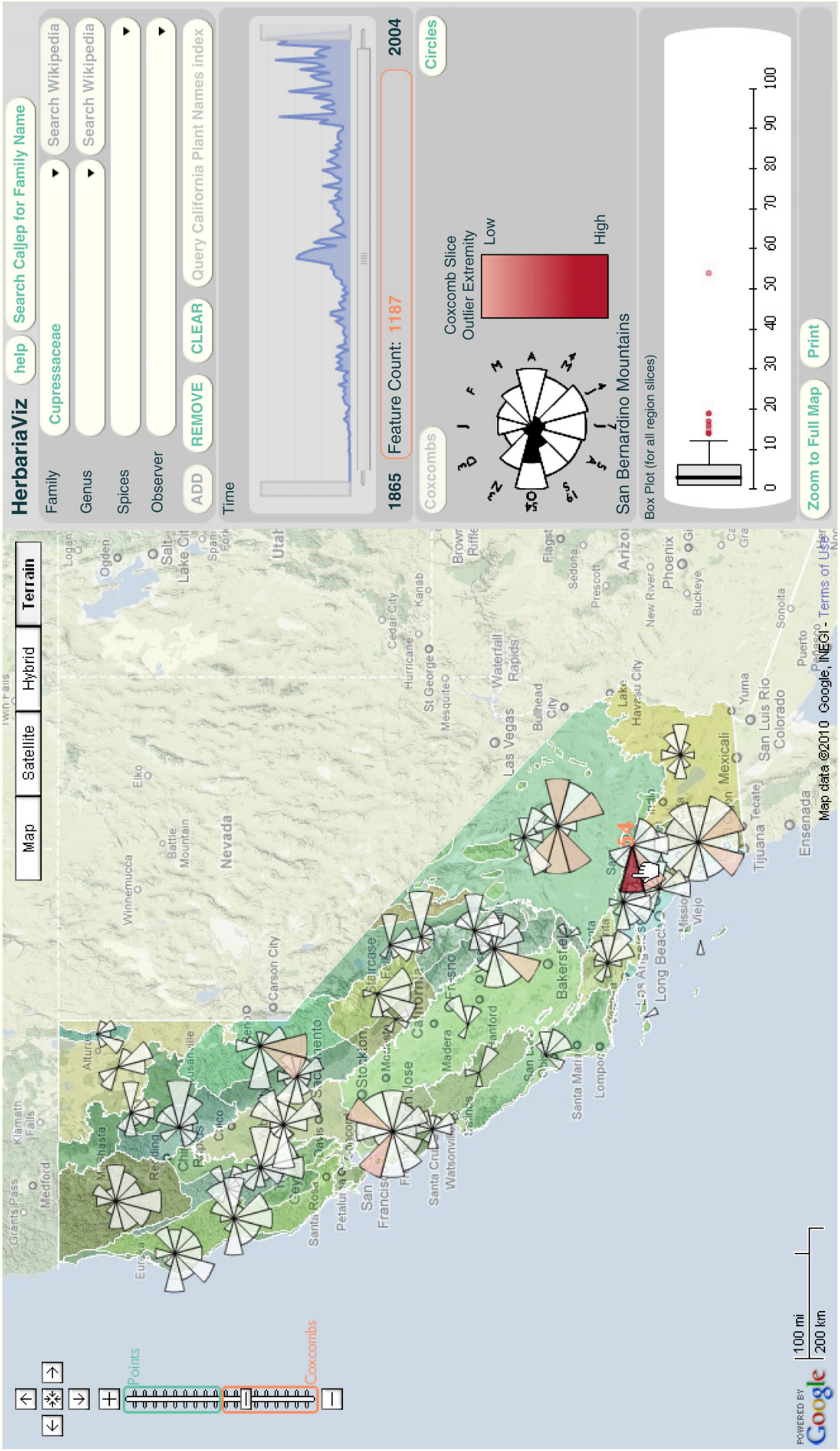
In developing this method to handle large datasets of space–time points, a number of potential avenues of research were revealed that were not within the scope of our initial efforts. These potential routes of investigation include cartographically and statistically sound methods for allowing multi-view comparison of user-generated query results, strategies for coping with inaccurate and/or incomplete data, a generalized method for allowing users to generate customized aggregations of their own datasets for use with an adaptable interface, and incorporating additional levels of data detail that accommodate inquiries related to invasive species. These will be discussed below.

5.1. Comparison symbology

Often, with an application that generates maps such as these, a user will want to compare multiple selections. The current application (and most other web-map applications) allow only map one selection at a time. Therefore, solutions are needed to facilitate comparison analysis. These solutions require developing symbol scaling methods that work across multiple dynamic query generations that have the potential to produce substantially different counts. The real challenge is to create a scaling method that can adapt to multiple selections with different data value ranges in order to make comparison possible. We have identified no published guidelines for addressing this problem. Any solution needs to generate comparable maps while also helping the user keep track of any changes in scaling that might be necessary for the specific maps generated to be informative.

More specifically, establishing one scaling method for the entirety of the dataset is impractical and unfeasible, considering the broad range of potential values found in map selections. The core problem here is that the data ranges for any specific selection may be very different from the range of the whole data set and from other selections. For example, comparing only records of *Cupressus macrocarpa* (Monterey Cypress) (Fig. 9) with all records for the entire family Cupressaceae, as in Fig. 10, reveals the significant disparity in value distribution and hints at the associated difficulty in creating a scaling method to handle both.

In these circumstances, scaling each selection independently would produce individual maps that represented their data well but, like graphs with different scales on their axes, would be difficult to compare (a symbol of a specific size would represent a different data value on each map). On the other hand, applying a constant symbol scaling function might produce maps with symbols that are extremely small or large if the data range was very different between



HerbariaViz

Family: Cupressaceae

Genus

Spices

Observer

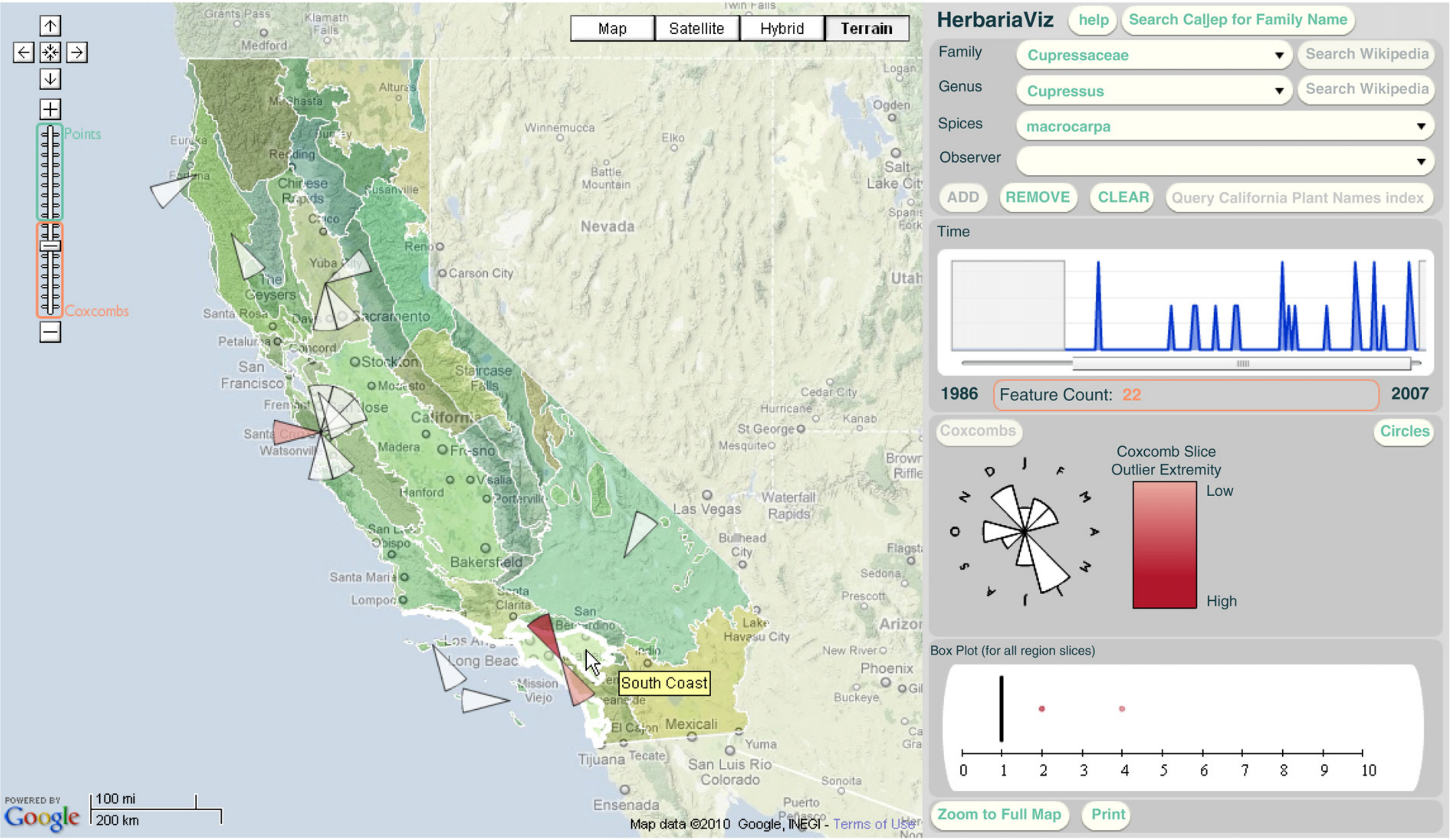
Time: 1865 – 2004

Feature Count: 1187

San Bernardino Mountains

Box Plot (for all region slices)

Fig. 9. The distribution for the family Cupressaceae with 1,187 records.



HerbariaViz

Family: Cupressaceae

Genus: Cupressus

Spices: macrocarpa

Observer:

Time: 1986 – 2007

Feature Count: 22

Fig. 10. The distribution for *Cupressus macrocarpa* (Monterey Cypress), a member of the Cupressaceae family, with 22 records.

selections; and, in an interactive system, there is no way to predict what selection a user might make for subsequent maps.

For creating paper map series using graduated symbols, Monmonier (1977) suggests establishing a regression method by imposing a shared minimum and maximum circle size for two (or more) maps designed for comparison. His assumption is that the map-maker knows what all the value ranges are going to be before creating the series. In a dynamic web-based scenario, as noted above, this is not possible. Monmonier's work does suggests that a dynamic, automatic update of the symbol scaling based on the most recent selection, rescaling the initially created map(s), might be successful in imposing minimum and maximum circle size shared by the different selections. However, this does not address the fact that the maps viewed prior to the new selections would have had an entirely different scaling, and comparison between the user's memory of the maps with the old scaling and future maps with the new scaling may be altered. The impacts of this are unknown. Regardless, potential solutions need to address: scalability, user memory, and computational cartographic practices for ensuring that scaling methods still produce readable and interpretable maps.

5.2. Data inaccuracy and inconsistency

Inherent variation in record formats and levels of accuracy in large, heterogeneously-formed datasets, presents challenges for creating database-driven web-maps that attempt to utilize space, time, and attribute information consistently. With the CCH dataset, record and accuracy variation was a result of compiling observer information from many different individuals and institutions over a relatively long period of time (~150 years). Two issues relevant to data processing and client–server interactions were inconsistencies and errors in the observer information and in the collection date.

Each database sample is associated with one or more observers. A list of observers, generated in the application by making a taxonomic selection, is sorted alphabetically, according to the first letter of the name that comes first in the record. However, a wide variety of naming formats exists. For example, the observer John Quincy Smith may be listed as: John Q. Smith, J. Quincy Smith, J.Q. Smith, J. Smith, Smith J., etc. With no easy method of recognizing and coding each of these naming conventions as the same individual, the query result often lists the same individual multiple times in different formats or orders. There is currently no way to view all samples of a given species associated with a single individual (including those samples that had multiple observers). Given the difficulty of linking multiple name formats to a single individual in the database, providing users with the means of selecting multiple observers for map generation would partially circumvent the issue. However, the issue of multiple individuals sharing a similar name would remain unsolved.

As noted earlier, a wide variety of date formats and accuracies are associated with each sample. While consistent Julian dates exist for most records in the dataset, dates which were missing day or month information (e.g. March 1995 or 1896) and those that covered a range of dates (e.g. July 20, 1906 – August 3, 1906) were problematic. To simplify temporal filtering and symbol drawing, each sample was associated with a single, finite moment in time, using only the first of the two Julian dates each sample carried (Table 1, tagged as "Early Julian Date"). For many of the samples, the two Julian dates were the same; the sample was collected on a single day. However, if a range of dates were provided, the sample was attributed to only the first day in the range. Similarly, the sample was coded as being collected on the first of the month if the day was missing, and on January 1st of the given year if both the day and month were missing. Given that the collections are aggregated monthly in HerbariaViz, missing day information is not a significant issue, though samples that are lacking month accuracy may cause collections in January to be slightly over-represented.

5.3. Generalizing the method

Generalizing the method presented here for on-the-fly aggregation of any space–time point dataset is an important component of future work. Such a method would allow a user to select or upload a spatially and temporally referenced point dataset, select or upload a polygon layer for aggregation, and then choose relevant attributes upon which to query, filter and focus data in the visualization stage of the process. Achieving a generalization of this method requires innovative development at all levels, including database and server functionality, client–server interaction, and smart interface tools that adapt to the users need.

6. Conclusions

In this paper, we presented the development of a practical method for handling large spatiotemporal point datasets for exploratory geovisualization. Specifically, we implemented this method with a prototype application focused on query and map-based display of California flora data. We provided solutions via client–server web-mapping using existing open source database and server technology for handling geographic information linked to a mapping client that was built using Adobe Flex (the client-side tools we produced are being distributed under an open-source license).

In successfully creating this application, we addressed the three goals of this paper: (1) to develop a method for efficient web-map client–server interaction involving large volumes of spatiotemporal point data, (2) to develop a symbology and symbol scaling method for representing that data in the client, (3) to develop an interface for client–server interactions and data exploration. Ongoing research is directed at solving the problem of scaling data for multiple user-generated query results in a multi-view comparison scenario, working around incomplete and inaccurate data, and generalizing the method.

Acknowledgements

We extend our appreciation to Dick Moe at the Jepson Herbarium, UC Berkeley for making the data used here available to us and for his insightful comments about the system on an early draft of material presented here. This work is supported in part by the U.S. National Science Foundation (grant EIA-0306845) and the U.S. Department of Homeland Security University Programs.

References

André, P., Wilson, M., Russell, A., Smith, D., Owens, 2007. Continuum: designing timelines for hierarchies, relationships and scale. In: Proceedings of the 20th annual ACM symposium on User interface software and technology: Oct. 7–10, 2007, Newport, RI. ACM. 101–110.

Andrienko, G., Andrienko, N., 2005. Visual exploration of the spatial distribution of temporal behaviors. Proceedings of the Ninth International Conference on Information Visualisation (IV'05), pp. 799–806.

Andrienko, N., Andrienko, G., 2007. Designing visual analytics methods for massive collections of movement data. Cartographica: The International Journal for Geographic Information and Geovisualization 42, 117–138.

Andrienko, N., Andrienko, G., Gatalsky, P., 2000. Supporting visual exploration of object movement. Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 217–220.

Best, B.D., Halpin, P.N., Fujioka, E., Read, A.J., Qian, S.S., Hazen, L.J., Schick, R.S., 2007. Geospatial web services within a scientific workflow: predicting marine mammal habitats in a dynamic environment. Ecological Informatics 2, 210–223.

Bhowmick, T., Griffin, A.L., MacEachren, A.M., Kluhsman, B.C., Lengerich, E.J., 2008. Informing geospatial toolset design: understanding the process of cancer data exploration and analysis. Health & Place.

Boulos, M.K., 2004. Web GIS in practice: an interactive geographical interface to English Primary Care Trust performance ratings for 2003 and 2004. International Journal of Health Geographics 3 (1).

Brewer, I., Campbell, A.J., 1998. Beyond graduated circles: varied point symbols for representing quantitative data on maps. Cartographic Perspectives 29 Winter 1998.

Chen, J., MacEachren, A.M., Guo, D., 2008. Supporting the process of exploring and interpreting space–time multivariate patterns: the visual inquiry toolkit. Cartography and Geographic Information Science 35, 33–50.

Cobb, D., Olivero, A., 1997. The Massachusetts electronic atlas: an interactive web site for access to maps and geographic data for the commonwealth of Massachusetts. The Journal of Academic Librarianship 23, 231–235.

Croner, C.M., 2003. Public health, GIS and the Internet. Annual Review of Public Health 24, 57–82.

Edsall, R.M., Harrower, M., et al., 2000. Tools for visualizing properties of spatial and temporal periodicity in geographic data. Computers and Geosciences 26, 109–118.

Flannery, J., 1971. The relative effectiveness of some common graudated point symbols in the presentation of quantitative data. Canadian Cartographer 8, 96–109.

Fredrikson, A., North, C., Plaisant, C., Shneiderman, B., 1999. Temporal, geographical and categorical aggregations viewed through coordinated displays: a case study with highway incident data. Proceedings of the 1999 Workshop on New Paradigms in Information Visualization and Manipulation in conjunction with the Eighth ACM International Conference on Information and Knowledge Management, pp. 26–34.

Funk, V.A., Richardson, K.S., 2002. Systematic data in biodiversity studies: use it or lose it. Systematic Biology 51, 303–316.

Graham, M., Kennedy, J., 2007a. Exploring Multiple Trees through DAG Representations. IEEE Transactions on Visualization and Computer Graphics 13 (6), 1294–1301.

Graham, M., Kennedy, J., 2007b. Visual Exploration of Alternative Taxonomies through Concepts. Ecological Informatics 2 (3), 248–261.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19, 497–503.

Guralnick, R.P., Hill, A.W., Lane, M., 2007. Towards a collaborative, global infrastructure for biodiversity assessment. Ecology Letters 10, 663.

Harrower, M., Fabrikant, S., 2008. The role of map animation in geographic visualization. Geographic visualization. Wiley and Sons, Chichester UK, pp. 49–65.

Hijmans, R.J., Garrett, K.A., Huaman, Z., Zhang, D.P., Schreuder, M., Bonierbale, M., 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. Conservation Biology 14, 1755–1765.

Hochheiser, H., Shneiderman, B., 2004. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. Information Visualization 3, 1–18.

Isaac, N.J.B., Mallet, J., Mace, G.M., 2004. Taxonomic inflation: its influence on macroecology and conservation. Trends in Ecology & Evolution 19, 464–469.

Javed, W., Elmqvist, N., 2010. Stack Zooming for Multi-Focus Interaction in Time-Series Data Visualization. In: Proceedings of the IEEE Pacific Visualization Symposium: March 2-5, 2010. Taipei, Taiwan.

Jepson, W.L., Hickman, J.C., 1993. The Jepson manual: higher plants of California. University of California Press, Berkeley and Los Angeles, California.

Kennedy, J., Kukla, R., Paterson, T., 2005. Scientific Names Are Ambiguous as Identifiers for Biological Taxa: Their Context and Definition Are Required for Accurate Data Integration. Lecture Notes in Computer Science 3615, 80–95.

Kessler, F.C., 2000. Focus Groups as a Means of Qualitatively Assessing the U-Boat Narrative. Cartographica 37, 33–60.

Kobayashi, S., Fujioka, T., Tanaka, Y., Inoue, M., Niho, Y., Miyoshi, A., 2009. A Geographical Information System Using the Google Map API for Guidance to Referral Hospitals. Journal of Medical Systems.

Kraak, M.J., Ormeling, F., 2003. Cartography: visualization of geospatial data. Pearson Education.

Kraak, M.J., van de Vlag, D.E., 2007. Understanding spatiotemporal patterns: visual ordering of space and time. Cartographica: The International Journal for Geographic Information and Geovisualization 42, 153–161.

Kumar, H.P., Plaisant, P., Shneiderman, B., 1997. Browsing hierarchical data with multilevel dynamic queries and pruning. International Journal of Human-Computers Studies 46, 103–124.

Lienert, C., Weingartner, R., et al., 2009. Real-time cartography in operational hydrology. Cartography & Geographic Information Systems 36, 45–58.

Loiselle, B.A., Jørgensen, P.M., Consiglio, T., Jimenez, I., Blake, J.G., Lohmann, L.G., Montiel, O.M., 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes. Journal of Biogeography 35, 105–116.

Lu, X., 2004. Web-GIS-based SARS epidemic situation visualization. Fourth International Conference on Virtual Reality and Its Applications in Industry, City, SPIE, pp. 445–452.

MacEachren, A.M., Crawford, S., Akella, M., Lengerich, G., 2008. Design and implementation of a model, web-based, GIS-enabled cancer atlas. The Cartographic Journal 45, 246–260.

McGuire, M., Gangopadhyay, A., Komlodi, A., Swan, C., 2008. A user-centered design for a spatial data warehouse for data exploration in environmental research. Ecological Informatics 3, 273–285.

Moe, R., Markos, S., Vanderplank, S., 2009. The consortium of California herbaria: a community approach to maintaining specimen information. CNPS 2009 Conservation Conference: Strategies and Solutions. California Native Plant Society, Sacramento, CA.

Monmonier, M., 1977. Regression-based scaling to facilitate the cross-correlation of graduated circle maps. The Cartographic Journal 14, 89–98.

Monmonier, M., Gluck, M., 1994. Focus groups for design improvement in dynamic cartography. Cartography and Geographic Information Science 21, 37–47.

Open GIS Consoritum, 2005. Web Feature Service Implementation Specification, Version 1.1.0. Open Geospatial Consortium, Inc., Wayland, MA.

Open GIS Consortium, 2000. OpenGIS Web Map Server Interface Implementation Specification, Revision 1.0.0. Open Geospatial Consortium, Inc., Wayland, MA.

Open GIS Consortium, 2002. Web Feature Service Implementation Specification, Version 1.0.0. Open Geospatial Consortium, Inc., Wayland, MA.

Plaisant, C., Carr, D., Shneiderman, B., 1995. Image-browser taxonomy and guidelines for designers. IEEE Software 12, 21–32.

Richard, D., 2000. Development of an Internet atlas of Switzerland. Computers and Geosciences 26, 45–50.

Shneiderman, B., 1996. The eyes have it: a task by data type taxonomy for information visualizations. Proceedings of the 1996 IEEE Symposium on Visual Languages. IEEE Computer Society Press, Boulder, Colorado.

Slocum, T.A., McMaster, R., Kessler, F.C., Howard, H.H., 2008. Thematic cartography and geovisualization. Prentice Hall, Upper Saddle River, NJ.

Tremblay, M.C., Hevner, A.R., Berndt, D.J., 2010. The use of focus groups in design science research. In: Hevner, A., Chatterjee, S. (Eds.), Design Research in Information Systems. Springer, New York, pp. 121–143. 22.

Tufte, E.R., 2006. Beautiful evidence. Graphics Press, Cheshire, CT.

Tukey, J.W., 1977. Exploratory data analysis. Addison-Wesley, Reading, MA.

Weaver, C., Fyfe, D., Robinson, A., Holdsworth, D., Peuquet, D., MacEachren, A.M., 2007. Visual analysis of historic hotel visitation patterns. Information Visualization 6, 89–103.

Williams, P.H., Margules, C.R., Hibert, D.W., 2002. Data requirements and data sources for biodiversity priority area selection. Journal of Biosciences 4, 327–338.

Zhang, J., Pennington, D.D., Liu, X., 2007. GBD-explorer: extending open source java GIS for exploring ecoregion-based biodiversity data. Ecological Informatics 2, 94–102.