A Multi-Scale Correlative Approach for Crowd-Sourced Multi-Variate Spatiotemporal Data

Thomas Gorko
EPICS Healthcare
tgorko@gmail.com

Calvin Yau
Purdue University
yauc@purdue.edu

Abish Malik
Davista Technologies
amalik@davistatechnologies.com

Matt Harris
Purdue University
mharris@purdue.edu

Jun Xiang Tee
Google, Inc.
juntee@google.com

Ross Maciejewski
Arizona State University
rmacieje@asu.edu

Cheryl Qian
Purdue University
qianz@purdue.edu

Shehzad Afzal
Purdue University
safzal@purdue.edu

Bryan Pijanowski
Purdue University
bpijanow@purdue.edu

David Ebert
Purdue University
ebertd@purdue.edu

Abstract

With the increase in community-contributed data availability, citizens and analysts are interested in identifying patterns, trends and correlation within these datasets. Various levels of aggregation are often applied to interpret such large data schemes. Identifying the proper scales of aggregation is a non-trivial task in this exploratory data analysis process. In this paper, we present an integrated visual analytics environment that facilitates the exploration of multivariate categorical spatiotemporal data at multiple spatial scales of aggregation, focusing on citizen-contributed data. We propose a compact visual correlation representation by embedding various statistical measures across different spatial regions to enable users to explore correlations between multiple data categories across different spatial scales. The system provides several scale-sensitive spatial partitioning strategies to examine the sensitivity of correlations at varying spatial extents. To demonstrate the capabilities of our system, we provide several usage scenarios from various domains including citizen-contributed social media (soundscape ecology) data.

1. Introduction

The focus of most visual analytic techniques (e.g., [3], [13], [24]) has been on data that are comprised of scalar values. The utilization of qualitative categorical data, such as personal opinion, feelings, emotions, for correlative exploration remains a challenging task. These data are often more complex than scalar data, and can be spatiotemporal, multi-scale, and comprised of multiple set variables. Examples of such data include users' emotional states (e.g., happy and scared), and crime arrest records where an individual can be charged with multiple offense charges for an incident at a particular location/time. Public safety agencies are interested in understanding the relationships between different charges across different scales in order to better design effective mitigation measures. Other examples include data collected from the *Record the Earth* project [29], which is a world-wide effort to record sounds in the environment in order to understand underlying coupled natural-human system dynamics across multiple spatial and temporal scales [26] .

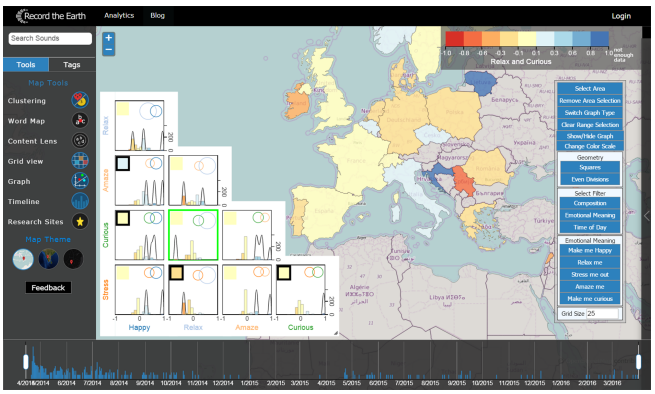


Figure 1: The overall visual analytics system includes a timeline selection slider, map tools, choropleth map divided by country or grid, and a correlation matrix interface to compare correlations between sound tag categories within the *Record the Earth* data set.

Researchers have utilized visual analytics [34] to assist casual experts[1] in interactively exploring their data in order to generate new insights (e.g., [25], [35]). These systems enable casual experts to form, explore, and validate hypotheses from their data. These systems are extremely important for social media and citizen-contributed data, especially when the goal is to enable the public to explore and analyze these

1. Casual experts: Experts in a domain but not necessarily data sciences. [27]

URI: http://hdl.handle.net/10125/50100
ISBN: 978-0-9981331-1-9

data, which encourages more citizen participation. Although these systems harness the strengths of both users' analytical skills and automated systems, solutions are still needed for two important challenges: understanding multivariate correlation, especially in multivalued categorical datasets, and solving cross-scale issues. Identifying the appropriate level of resolution at which to explore the data is still a fundamental challenge. For example, the data to be analyzed may be collected at a very fine scale/resolution, but the questions posed of the data may lie at a coarse resolution [20]. An even more difficult problem occurs when data are captured at a coarse scale, but the questions that are posed lie at a finer resolution. These challenges are exacerbated when the data are comprised of multiple attributes that may have different inter-correlations across multiple scales. However, such analyses over the different data dimensions and spatiotemporal scales can yield new insights into the underlying processes.

To solve these challenges and provide solutions for casual experts, we have developed a visual analytics approach that enables the exploratory analysis of spatiotemporal data across multiple categorical dimensions and geospatial scales. Our visual analytics system, shown in Figure 1, enables citizens and casual experts to interactively explore their data and develop insights into the workings of real world environments. Specifically, it provides novel views and methods for exploring and analyzing potential correlations between different qualitative data variables under uncertainty across multiple geospatial scales through an interactive correlation matrix view that incorporates Venn diagrams and mosaic plot visualizations [11]. Additionally, we incorporate visual indicators to encode the uncertainty that occurs from both the underlying data and the statistical methods. Our system was designed to enable citizens and soundscape ecologists to explore the *Record the Earth* data. However, we note that the techniques presented in this paper are versatile and can be adapted to any spatiotemporal data that are comprised of multi-variate nominal and ordinal attributes, and is especially valuable to citizen-contributed data and other social media data. The main contributions of this paper include the following:

- A visual analytics environment for exploring multi-scale correlations among spatiotemporal categorical data with polytomous variables.
- An interactive approach for exploring global/local correlations among multiple variables under uncertainty.
- A visual analytics process for performing scale sensitive geospatial correlative exploration.

2. Related Work

Researchers in the visual analytics field have proposed numerous approaches for exploring multi-dimensional and spatiotemporal datasets. Using interactive maps to explore spatiotemporal dataset at multiple spatial scales is a common approach. Rich interactions facilitate effective level-of-detail investigations of patterns hidden in the large-scale dataset. In this section, we discuss previous works that relate to the multi-scale geographical visualization and multi-dimensional visualization.

2.1. Scale and Geography

In order to gain insights in geographical datasets across multiple spatial scales, researchers have proposed several solutions that visualize the statistic results at different levels of aggregation [21]. Goodwin et al. [13] point out that an overemphasis on global statistics obfuscates local correlations that may deliver valuable insights for visualization purposes and propose a framework that visualizes the relationships between multiple variables concurrently based on various scales. Ferreira et al. [9] propose Birdvis, a visualization system that analyzes spatiotemporal bird distribution models. The system facilitates the analysis of model parameter inter-dependencies, and provides coordinated views for determining local correlations and patterns in the models. Goodwin et al. [12] explore multivariate data visualization over various scales in the context of geodemographic classifications. In contrast to the previous work, our system visualizes the correlations of multiple scales based on several spatial partition strategies and also allows the users to interactively navigate through different spatial scales. Moreover, we visualize the correlations at multiple scales using a compact visual design that seamlessly integrates both global and local scales into a single visualization.

Thom et al. [33] introduce a quadtree-based approach that performs spatial partition adaptively based on the density of data points. The partitions effectively indicate the population density in the geospatial regions. Guo et al. [14] introduce another quadtree-based algorithm that solves the overcrowding problem of geographical map glyphs due to prolonged, continuous scale change. Our application employs a similar data-driven quadtree-based visualization.

2.2. Multi-dimensional data visualization

Multi-dimensional data visualization is an important research topic in visual analytics. Some related work utilizes matrix based visualizations and compares two variables in each cell of the matrix. Im et al. [17] introduce the Generalized Plot Matrix (GPLOM) to visualize the pair-wise relationships of variables. Similar work has been proposed by Emerson et al. [7]. Zhang et al. [39] examine various correlation representation methods. Our application also uses a half matrix based correlation visualization technique that utilizes several visual indicators to encode global and local correlations.

Several previous works also focus on visualizing multiple categorical data variables. Friendly [10] gives a comprehensive overview of common techniques used to visualize categorical data. Horrigan [16] explores visualization and analysis of multi-dimensional categorical and ordinal data. Shneiderman et al. [30] develop GRIDL, a two-dimensional visualization tool based on categorical and hierarchical axes. Stoffel et al. [32] propose a novel technique to visualize

two proportions in categorical data, addressing limitations of choropleth maps. Kolatch and Weinstein [19] develop CatTrees, an enhancement of Treemaps that creates a hierarchy of categorical data. Kerbs [18] proposes BLUE, an interactive visualization system that establishes meaningful decision trees out of database entries that are made up of categorical attributes. Our system also displays multiple categorical variables, and allows analysts to compare the correlations between polytomous categorical variables using a matrix view.

Other previous works focus on exploration and visualization of temporal categorical data. Fails et al. [8] introduce PatternFinder to visualize temporal patterns of multivariate and categorical data sets that identifies the patterns as user-defined event sequences with in-between time spans. Wongsuphasawat and Shneiderman [36] propose M&M (Match & Mismatch), a temporal categorical similarity measure that matches events between a target and records aligned by sentinel events. They also propose Similan for visualizing temporal categorical records to allow parameter customization. Chang et al. [5] introduce WireVis to visualize correlations between bank accounts and transaction keywords over time. Alsakran et al. [1] demonstrate that entropy-related measures improve the effectiveness of visualization, since the measures can help interpret categorical data and reduce visual clutter. Ma and Hellerstein [23] address the difficulty in visualizing categorical values with no semantic order, and develop an algorithm that orders the values by identifying relationships between values of distinct categorical attributes. We also display temporal categorical data and allow for the selection of a time range to enable the comparison of categorical attributes at different time ranges.

2.3. Set visualization

Finally, we note that researchers have also explored different set visualization techniques. Lex et al. [22] propose UpSet to visualize and analyze set intersections, and propose several visual encoding to solve multiple scales and scalability issues. Delaney et al. [6] explores the visualization of a large number of sets using Euler diagrams. Alsallakh et al. [2] introduce Radial Sets to compare overlapping sets. Our system utilizes Venn diagrams and mosaic plots to show the relationships between multiple pairs of sets at a time.

3. Exploring multi-variate spatiotemporal correlations across multiple scales

Our system has been designed to enable social scientists, sound ecologists, and citizens to explore the correlations and relationships between the different categories of their data at different spatial scales. Our approach focuses on enabling users to explore both local and global geospatial trends and correlations among different data variables while factoring in the uncertainty that is inherent in the applied statistical processes (e.g., due to sparse data, statistical significance tests).

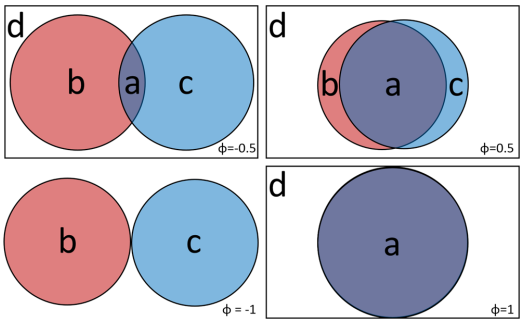


Figure 2: Four Venn diagrams with varying phi-coefficients.

Users can visually and interactively explore the spatiotemporal relationships between different attributes in polytomous categorical data. For example, variable X (Emotion) with state x_1 (Stress) and variable Y (Culture) with state y_1 (Talking) are positively correlated in New York, but negatively correlated in Texas (Figure 9). Correlation is usually represented by a quantitative scale (e.g., between -1 and $+1$) that indicates the strength of the relationship between two variables. In our work, since each variable can have only two possible outcomes - "yes" or "no", we utilize the phi-coefficient [38] to measure the correlations between a pair of data attributes that ranges from -1 to 1. The phi-coefficient is calculated using Equation 1:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \quad (1)$$

Here, a, b, c, and d are the number of points that fall into the corresponding category as shown by the Venn diagram in Figure 2. We take d into account to factor in for the conditional probability. For example, if each attribute is distributed independently with a constant red number of points, when the red number of points of the entire data space grows, less of an intersection would be expected.

We note that a high phi-coefficient value does not necessarily indicate that there is a significant correlation between the two attributes, as it does not take into account the total number of points. Accordingly, we apply the chi-squared test ($\chi^2 = N\phi^2$) [37], where N is the total number of points. The correlation is significant when the obtained χ^2 value is higher than the value for the desired significance with 1 degree of freedom.

3.1. Domain Requirements

In this section, we discuss the domain related tasks and characterize the main challenges faced by users in their use of spatiotemporal polytomous categorical data. Our discussions have been motivated by conversations with experts from the Purdue University Department of Forestry and Natural Resources, a U.S. research institution, who study human-environment interactions in order to explore the impacts of humans on the environment, and vice versa. One of the main areas of focus was soundscape ecology (i.e.,

the study of landscape dynamics using sound). The goal of soundscape ecology is to use the information encoded in sound to understand the broader impacts of land use, climate change, and habitat disturbance across different landscapes and over time. To gather the domain requirements and to understand challenges associated with ecological and soundscape data, we conducted several requirements elicitation interviews with researchers and scientists from varied backgrounds, including a seabird biologist, a soundscape ecologist, a computational musicologist, an environmental education specialist, and a ecological statistics researcher. The focus of these discussions was in regard to their use of spatiotemporal categorical data.

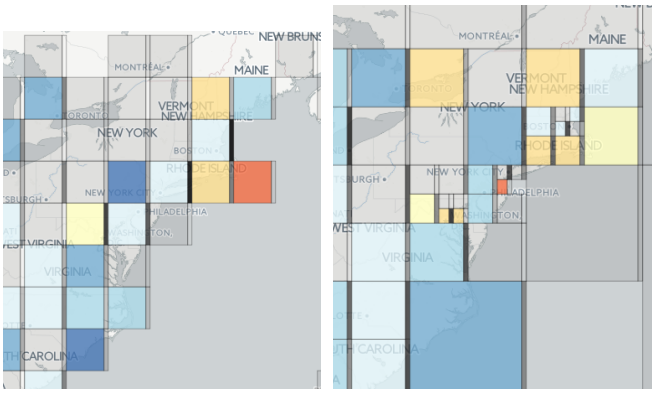
3.1.1. Design Goals. During interview sessions with domain experts, we noted a frequently occurring theme which highlighted the need to explore trends and patterns with spatiotemporal data. There was also an emphasis on the need to understand relationships between categorical data and their associated values over the various geospatial scales. Time was also brought up during the sessions. The domain experts explained their desire to understand patterns in datasets that span over many years or decades. Accordingly, we designed our system with the following goals:

- G1 **Explore correlations between multiple data categories**: The system should enable exploration of correlations between different categorical data variables and their possible values.
- G2 **Visualize correlations across multiple geospatial scales**: The system should provide the ability to visually explore correlations between user-selected categories in geospace across multiple scales of analysis.
- G3 **Visualize the set relations between the different categories**: The system should enable users to visually explore relationships and the degrees of overlap between different categories.
- G4 **Provide significance testing results for multiple correlation computations**: The system should provide users with the ability to visualize statistical significance of correlation results.
- G5 **Summarize and provide access to raw data**: The system should provide visual summaries and access to the original data for further exploration.

3.2. Visual Analytics Environment

Figure 1 shows a snapshot of our visual analytics system. The system enables users to explore and analyze their multi-variate data at multiple spatial scales. The users can select, filter, and highlight across different views to examine different perspectives of the dataset.

Our system is comprised of several interactive linked views. The main view of our system is the map view that enables users to geospatially visualize the correlations between different data categories (Section 3.2.1). The system also utilizes an interactive correlation matrix view that



(a) Uniform rectangular grids (b) Quad-Tree

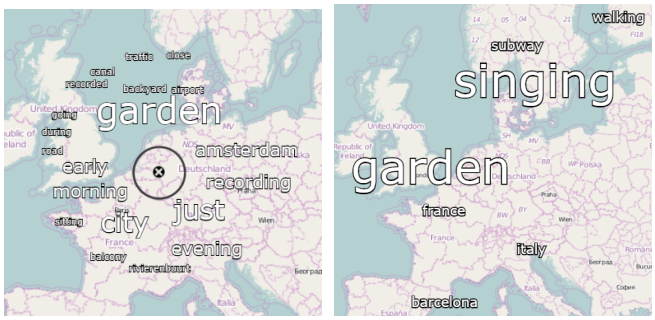
Figure 3: Choropleth map. A color scale shows the relationship between "frogs" and "insects" using a (a) square grid or (b) quad-tree method.

provides a compact visual design to enable the exploration of correlations between data categories (Section 3.2.2). We also provide an interactive time series view that visualizes the evolution of the data volume over time using a bar chart (Figure 1). This view enables users to perform temporal filtering by specifying a temporal range on the bar chart. Finally, the system provides several different filtering options to enable users to examine different data perspectives. Users can choose to filter by specific keywords using a search bar to explore the semantic knowledge hidden in the dataset. They can also perform geospatial filtering by drawing a polygonal region or specifying certain spatial units for filtering. The system also provides users with the ability to filter their data by the different associated data categories and their corresponding values.

3.2.1. Map View. The map view enables users to visualize the correlations between any two user-selected categories geospatially. Users can select the categories using the interactive correlation matrix view (Section 3.2.2). The system fragments geospace using a user-selected geospatial aggregation technique and computes the correlation value between the two data categories for each sub-region.

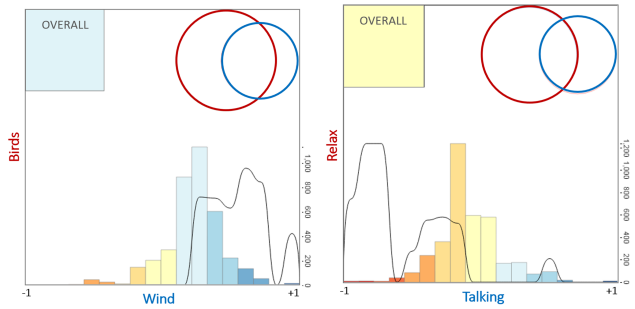
In order to facilitate the exploration of the data at multiple spatial scales, the system provides several spatial aggregation strategies, including:

- Uniform rectangular grids: In this scheme, we fragment geospace into evenly divided regular rectangular grids (Figure 3a).
- Density-based quad-trees: The geospace is evenly and recursively divided into quarters in order to ensure that the data volume inside each spaital grid is below a pre-defined threshold (Figure 3b).
- Man-made spatial regions: The geospace is divided based on man-made administrative partitions (e.g., census blocks, voting districts, countries (Figure 1)).



(a) Content Lens (b) Tag Map

Figure 4: Language tools. The content lens (a) shows most common words within in a small region while the tag map (b) shows the most common words across all regions within the window extent.



(a) Positive Correlation (b) Negative Correlation

Figure 5: Correlation histogram and Venn diagrams. Chart (a) indicates a mostly positive correlation between "birds" and "wind" while (b) indicates a mostly negative correlation between "relax" and "talking".

- Contour areas based on the data density: The geospace is divided into areas of high, medium, and low density.

Dividing the map into uniform rectangular grids enables users to visually explore the geospatial correlations. Both the uniform rectangular grids and density-based quad-tree visualizations are adaptive to the scope of the data within the current view and geospatial scale. As users zoom in on the map, the size of each sub-region in the uniform rectangular grid mode automatically changes to a finer scale allowing users to explore more local patterns. Similarly, zooming out enables users to explore the global correlation patterns. The system also provides users with the ability to interactively change the size of the sub-regions; thereby, dynamically changing the number of data points in each sub-region and hence the corresponding results. The density-based quad-tree approach is also sensitive to the data contained in the scope of the map. In this mode, the geospatial regions are divided into quarters to ensure that the data are more evenly distributed throughout each sub-region. While the quad tree aproach typically shows the same overall trends, it does a better job at showing the data at the resolution of data we have. There is a minimum size for the regions to ensure they remain readable.

However, a potential drawback of utilizing the quad-tree approach is that regions with sparse data can be large and more prominent. To solve this challenge, we provide a vertical bar for each sub-region on its right side to encode either the data volume or uncertainty (i.e., statistical significance).

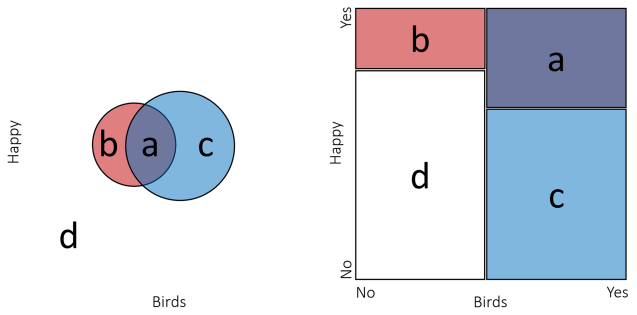
The contour area lines has the advantage of being generated from the data, and as a result gives a better picture of how the data is distibuted. These areas are generated based on the total number of points so if there are fewer points in the data set, fewer points are needed for a region to be considered high density.

The map view also employs several other visualization techniques to help users explore the semantic information in different geographical regions. These include glyph based visualizations to show the distribution of the incidents. We also utilize the content lens [4] technique that is an interactive lens that changes with the mouse movement to show the most representative keywords extracted from the textual data for the points inside the lens using a word cloud layout. (Figure 4a). Finally, we provide a tag map visualization [31] that utilizes a context-aware approach and shows the representative keywords of the data points in a global view. (Figure 4b).

3.2.2. Correlation Matrix View. Since the map view only shows the correlation between two variables at a time, we introduce a correlation half matrix view that visualizes the correlations of multiple categorical variables to enable the exploration of global and local trends. The matrix view provides a compact visual design that couples a global matrix layout along with several types of charts/glyphs that are embedded in the cell of the matrix. The matrix consists of an n by n grid where n is the number of selected attributes. The matrix view allows the users to flexibly add or remove attributes in the matrix so that they can focus only on a subset of variables of their interest. Inside each cell of the matrix, we utilize several visual designs to indicate both the global and local correlations across multiple spatial scales. We now discuss the different visual components embedded in each cell (Figure 5) of the matrix view.

Global correlation: Rectangular representation. The rectangle on the top left of each cell shows the global correlation of the corresponding two variables. The color of the rectangle encodes the phi-coefficient value based on the aforementioned divergent color scheme. The rectangle is drawn with a thick black colored border if the phi-coefficient value is statistically significant (Figure 5).

Global relationship: Venn diagram. In the top right of the cells shown in Figure 5, we utilize an area-proportional Venn diagram to indicate the relationship of the corresponding two variables. The size of the two circles in the Venn diagram encodes the volume of the corresponding data



(a) Venn Diagram (b) Mosaic Plot

Figure 6: Example of Venn diagram (a) and mosaic plot (b) from *Record the Earth* showing relationship between the tags "birds" and "happy". Points with "birds" are more likely to have "happy" than without.

variables . Note that the color of the circles is consistent with the color of the labels for the corresponding attribute. The Venn diagram visually indicates the quantitative value of the union, intersection, and difference between the two variables. Our Venn diagrams are area proportional; that is, the ratio between the area of the circles and their intersection is the same as the ratio between the corresponding data volume. We choose the Venn diagram in our work because it is a popular visual representation for logical relationships of statistical variables.

Our system also employs mosaic plots to visualize the relationships between data variables [15]. Mosaic plots are used to show the correlations between two attributes by showing their relative frequencies and overlaps. Mosaic plots are area proportional representations of all the data points . Figure 6b shows an example of the mosaic plot and compares it with its corresponding Venn diagram visualization. Mosaic plots enable users to discern if an attribute occurs more or less often than another attribute [11].

Local correlations: Bar chart. We integrate the bar chart visualization in the matrix cell that is positioned at the bottom of the cell to reveal the correlations related to different local regions. The bar chart visualizes the correlation distribution across the local geospatial regions in the current scope of the map view. The x-axis represents the correlation value from purely negative (-1) to purely positive $(+1)$, while the y-axis represents the summed data volume inside all the regions of the corresponding correlations. The y-axis is normalized based on the maximum value across the entire matrix. Figure 5a shows an example where most of the regions are positively correlated for the sounds that are tagged with attributes "birds" and "wind", while Figure 5b provides an example of mainly negative correlations for the sounds tagged with attributes "relax me" and "talking". The correlation bar charts allow users to select a range on the x-axis to filter the geospatial regions with the selected correlation values. This action also filters the other cells of the matrix view to reflect the selection.

Uncertainty visualization: Line chart. Finally we add a line chart layer on top of the bar charts to show the statistical significance of the local correlations shown by the bar charts in each cell of the matrix. The value of the line chart for the corresponding correlation value encodes the percentage of points in regions whose correlation values obtained are statistically significant with $p < .05$. The underlying correlation bar charts enable analysts to visually explore the correlation distribution of the data over geospace. Correspondingly, the uncertainty line chart visualization provide analysts with the results of the statistical significance tests for each correlation value (Figure 5).

4. Case Studies

In this section, we utilize the data from the *Record the Earth* [29] project to demonstrate the capabilities of our system. We also briefly explain a use case with criminal arrest records from Tippecanoe County, Indiana.

4.1. Soundscapes: *Record the Earth* Case Study

Here, we explore the relationships between different data categories in this dataset [29]. The data utilized in this subsection comes from the sounds and the associated meta-data uploaded by people through a publicly available mobile application, crowdsourcing both sounds and options/emotions about the sound. Before uploading a sound, users must go through a list of categories and tag the associated sounds that are relevant for the recording being uploaded. The fields in this dataset include the date, time, and location of the sound, user-generated free-form text description, and categories that describe the sound, including the emotion (e.g., happy, relaxed, stressed), sounds that animals make (e.g., birds, mammals), geophysical sounds (e.g., rain, water, thunder), manmade sounds (e.g., vehicles, airplanes, trains), and cultural sounds (e.g., talking, instruments, singing). In this discussion, we will refer to the values for these categories (e.g., happy, talking, birds, etc.) as tags. We now provide a scenario where a soundscape ecologist is utilizing our system to explore the relationships between different data categories over multiple geospatial scales.

4.1.1. Global Trends. The analyst begins by utilizing the matrix view with the intent to investigate whether sounds with tags "stress", "birds", and "vehicles" have any correlations over a region of interest (Figure 7). She observes an overall statistically significant positive correlation between "birds" and "vehicles". This indicates that the sounds tagged with birds usually also contain vehicular sounds. She also notices a negative correlation between "birds" and "stress". While the level of confidence in these correlation values is lower, she notes that this conforms to what she would expect as she hypothesizes that bird sounds should generally tend to have a positive influence on peoples' behaviors. Finally, she notes a positive correlation between "stress" and "vehicles" with high statistical significance (Figure 7b). She observes

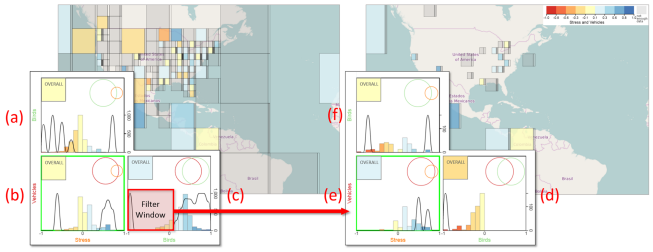
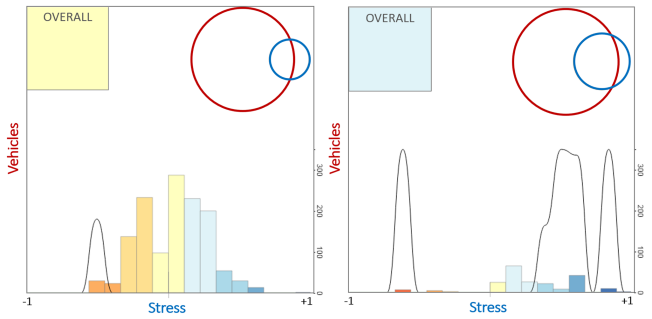


Figure 7: Choropleth map and matrix interface comparing multiple tag categories in the *Record the Earth* data set. On the left the "vehicles" vs. "stress" window is selected, the map shows a distribution of positive and negatively correlated regions across the United States. A filter is applied highlighting only regions with negatively correlated "vehicles" and "birds". The right side shows updated regions on the map and changes in the matrix reflect the updated map distribution. The correlation between "vehicles" and "stress" is now much higher.



(a) United States (b) Europe

Figure 8: Correlation matrix comparing "stress" and "vehicles" using regional selection in (a) the United States and (b) Europe. The histogram for Europe has been scaled to match.

that vehicular and traffic sounds have a tendency to cause stress in people.

The analyst now chooses to visualize the correlation between the tags "stress" and "vehicles" over geospace by clicking on the corresponding view in the matrix view (Figure 7b). Given the uneven geospatial distribution of the uploaded sounds, she selects the quad-tree spatial division technique (Section 3.2.1) to visualize the results on the map. She observes that the sounds with "vehicles" exhibit a relatively higher correlation with "stress" in Europe, whereas the correlations observed in the United States are more uniform (i.e., different regions exhibit both positive and negative correlations). To investigate this further, she applies a geospatial filter to select only the sounds in the United States using the region selection tool (Section 3.2). The resulting correlation bar graph (Figure 8a) confirms that "stress" and "vehicles" is more uniformly distributed. The corresponding Venn diagram (Figure 8a) also shows that more than half of the sounds with "stress" also have "vehicles" in the United States. Upon further investigation, the analyst observes that the regions of positive correlation in the U.S. gravitate more toward highly populated cities (e.g., New York City, San Francisco), whereas the negatively correlated regions are located near less populated areas (e.g., Wyoming).

Next, the analyst applies the geospatial filter to focus on the European regions. The resulting bar graph obtained (Figure 8b) shows that "stress" and "vehicles" are positively associated with an overall $p < .05$, and that they have a positive association in many of the regions. The corresponding Venn diagram also shows that almost all sounds tagged with "stress" are also tagged with "vehicles". She notes that the reason for these observations might be because of the higher population density in urban areas of Europe, along with the fact that most of the sounds recorded through this project in Europe originate in larger cities. While Europe does have many fewer sounds in the database than the US, the distribution of sounds there is significantly different.

Next, the analyst returns to the global geospatial view of the whole world by removing the applied geospatial filters. She uses the matrix view to filter out the regions where "birds" and "vehicles" are positively correlated (Figure 7c) by selecting the negative range inside the window "birds" and "vehicles". The results from this action are shown in Figure 7d. She observes that the tags "stress" and "vehicles" are now more positively correlated (Figure 7e). She notes that since the positive effect of "birds" has now been reduced by applying this filter, the relationship between "vehicles" and "stress" gets amplified. However, she finds that the correlation between "birds" and "stress" does not change by much (by comparing Figure 7f and 7a). This indicates that removing the sounds that have "vehicles" and "birds" as positively correlated does not have much effect on the relationship between "birds" and "stress". That is, the correlation between "birds" and "stress" in the filtered regions is only minimally influenced by "vehicles". She decides to further investigate this phenomena and listens to the sound files using our system. After her analysis, she concludes that in these sounds birds are much more noticeable than vehicles which tend to be heard from a distance. As a result, the analyst can get insights into the underlying correlations between these different variables over geospace.

4.1.2. Regional Trends. In this example, our analyst is interested in investigating the relationships between the tags "stress" and "talking" in the *Record the Earth* data. She is specifically interested in exploring the correlations between the western and eastern regions of the United States. In order to explore the data, she utilizes the region selection tool of our system to select the two regions of interest. She fragments the geospace using uniform grids (Section 3.2.1) for this analysis and obtains the results shown in Figure 9. She notes that the eastern half of the country has more regions that have positive correlation. She also finds that certain regions in the East Coast (e.g, Maine) tend to be more positively correlated for the selected tags. She observes

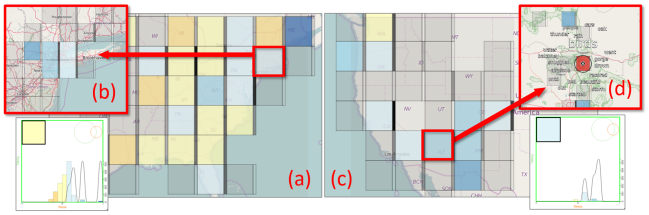


Figure 9: Comparison of "stress" and "talking" in United States. The eastern half of the country is selected in (a) and the western half of the country is selected in (c). (b) shows a close up of view of New York City and (d) shows a close up in Arizona usiing the content lens.

that the region surrounding New York city has a slightly positive correlation between stress and talking. She decides to investigate further and zooms into this region. The system adapts the scale to the new zoom level automatically (Figure 9b). This micro scale analysis reveals that although the coarse scale analysis indicates a positive correlation, certain regions at the fine scale have a strong negative correlation with respect to the selected tags of "talking" and "stress".

The analyst notices similar trends in the western part of the country (Figure 9c). She finds that the regions in Texas have a negative correlation between stress and talking. She also notes the regions covering the states of Arizona and New Mexico (at a global scale) have an overall positive correlation. However, when she zooms in further to this region, she finds a particular region where the sounds are negatively correlated for the tags stress and talking (Figure 9d). She uses the content lens feature of our system (Section 3.2.1) to investigate the textual content of the sound uploads and finds the keywords "thunder", "beautiful", "rain", and "airplane". She decides to listen to a few sounds from this region using our system and determines that these sounds have probably been uploaded by vacationers who do not find talking to be a stressful activity. The system therefore enables her to explore both micro and macroscopic trends using the data.

4.2. Crime Data Case Study

We now further demonstrate the system's capabilities through examining the criminal arrest records from Tippecanoe County, Indiana. We present the discoveries from our analyst exploring the connections between different criminal charges. The analyst first filtered and utilized criminal incidents with two or more associated charges, then studied the correlations between different charges and geospatial relationships. The analysts discovered a strong positive correlation between both theft and vandalism, and theft and fraud, signaling these charges are often associated. One theory is the possibility of thieves fraudulently selling stolen goods to other people. The analyst observed that the correlations between trespassing and theft tend to center around 0 in more densely populated regions near the city center, but are more positive around rural areas indicating theft charges may be more likely to occur in conjunction with trespassing charges. Finally, our analyst looked into the associations between operating while intoxicated (OWI) and driving while suspended offenses through geospatially aggregating the data using the quad-tree approach, and discovered that high-density population areas and the more rural areas in the south of Lafayette have a stronger positive association between OWI and driving while suspended. These are several examples of how our system enables analysts to gain new insights using different geospatial aggregations.

5. Domain Expert Feedback

Our system was assessed by six domain experts in the Department of Forestry and Natural Resources at a U.S. research center. The system was presented to them with the *Record the Earth* data. For each user we conducted a brief training session involving demonstrations and explanations of the different functions and assistance as they explored data of their interest, after which we asked for their feedback. While our analysts found the system useful and appreciated the visual analytics support, one analyst expressed a slight concern on the interface being too complex for her less analytical-focused need. We wish to later perform a formal qualitative study with citizens to better evaluate the system usability. Aside from that, the researchers noted that most of their analysis was usually performed using spreadsheets or statistical packages such as R [28], so they found our interactive visual analytics system to be a welcome addition. They stressed the need to conduct their analyses over multiple scales of aggregation. During one interview, a seabird ecologist noted that his analysis on seagull populations usually spanned over multiple geospatial regions and required him to explore the data across multiple geospatial scales (e.g., cities, countries, continents) in order to correlate the different seagull data categories (e.g., seagull health, demographics) with human demographics data (e.g., population, income, pollution).

This feature in particular, that the system enabled them to visualize correlations between variables in geospace (Section 3.2.1) at various scales, was viewed as a strong benefit by the researchers. One researcher commented: "In biology we have several scales of processes. We study species, communities, ecosystems, and biomes, and there are different processes between these levels of organization. We are reductionists and need to look for trends at level of communities, and at the level of species. Tools like this can help us visualize trends using data taken at the level of the species and look at bigger processes happening in our data."

Another benefit called out by the researchers was the different levels of aggregation that the system provides, specifically the multiple methods to spatially aggregate the data (e.g., using a shapefile, quad-tree, regular grids) in order to perform their correlation exploration was discussed:

"This part of the analysis process is kind of a game, to play with the data to explore and search for trends. In this sense the part of the tool that allows you to easily break down space into sub-samples without going back to

the database to manually break down the space, is really helpful."

The analysts noted that our method of using the geospatial bars to encode either the geospatial density distribution or statistical significance was effective, and they could visually associate them with the corresponding correlation values. They commented that statistical significance testing was also important in their analysis and were pleased that our system provided them with a visual way to explore the same.

The analysts had positive feedback for the matrix view (Section 3.2.2). Analysts said that although the matrix view was quite information dense and a bit overwhelming at first, they quickly got used to the different visual encodings. They found the ability to visualize the correlations between the multiple categories in the unified matrix view to be important. Analysts especially liked that we had used the same color scheme for the correlation histogram visualization as that of the map visualization. They also found the correlation view to be highly interactive and it allowed them to filter the geospace by the correlation values. With respect to the uncertainty visualization, the analysts found the line graph visualization on top of the correlation histogram visualization to be confusing at first; however, after further explanation, they were able to relate the correlation values to the results of the statistical significance testing. They also stated that scaling the uncertainty line graph differently than the correlation histogram (as opposed to scaling them to the corresponding histogram bar height) (Section 3.2.2) was beneficial as it highlighted the statistically significant regions irrespective of the number of data points contained within them. We note, however, that although we have received largely positive feedback on our visual design to encode the uncertainty in the matrix correlation view, we believe that further evaluations are needed to fully understand their efficacy in communicating the correlation values and their corresponding uncertainty. We leave this as future work.

The analysts also provided positive feedback regarding the Venn diagram visualizations and found them to be intuitive. However, they found the mosaic plot visualization to be cumbersome and preferred the Venn diagram visualization. Although a few analysts commented that the mosaic plot visualization was helpful in allowing them to examine and compare the "yes-yes" relationships between the different variables, they largely preferred the Venn diagram visualizations. We believe that this is because people are more accustomed to Venn diagrams. However, the benefits of the mosaic plot over the Venn diagram visualization remain to be tested, and accordingly we leave this as future work. The analysts also had positive feedback for the timeline view and especially liked that they could temporally filter their data using this view. Finally, we note that all the analysts wanted to apply our system onto their other ecological datasets and suggested that we provide a data input interface that would enable them to upload their data into our system. We leave this as future work.

6. Conclusion and Future Work

This paper presents a visual analytics approach for categorical spatiotemporal data that can be comprised of multiple set based variables (e.g., polytomous variables). Our system enables users to explore data for potential correlations between the different data categories and their values over time and across multiple geospatial scales of analysis. We provide several linked views, including an interactive map view, data selection tools, and a correlation half matrix view that utilizes visual design elements to encode the global and local correlation values. We also utilize uncertainty visualizations directly in the geospatial view and in aggregate within each correlation window within the matrix view. We have provided several case studies that demonstrate how our system can be used to help find relationships in polytomous categorical datasets such as the soundscape ecology project *Record the Earth* data. Expert feedback has pointed out benefits of our interactive data exploration system such as finding global or regional pairwise correlations in multivariate categorical data and a simple method for selecting and aggregating data across macro and micro spatial scales. We note that our system is extendable to other polytomous categorical spatiotemporal datasets.

Future work includes extending our system to enable the exploration of multivariate correlations. Our current system is based on the users selecting only two categories at one time in order to explore their relationships in geospace. We also plan on extending our work to enable the exploration of correlations over multiple temporal scales. These include the exploration of spatiotemporal data over multiple temporal aggregations (e.g., by day, week, month, year) and different temporal regions for the same temporal aggregation level. This will enable users to explore the changes withing these different relationships between variables over time. We plan on performing a qualitative study on the effectiveness and the usability of the system, especially of citizens. Finally, we plan on investigating automated methods that provide guidance to users for selecting appropriate geospatial and temporal scales and variables to examine, and implementing in-system training tools to teach new users how to use the system.

Acknowledgments

The authors would like to thank the researchers at the Center for Global Soundscapes for their help, as well as the reviewers for enhancing the structure and the content of this paper. This work was funded by the NSF and U.S. Department of Homeland Security VACCINE Center's under Award Number 2009-ST-061-CI005.

References

[1] J. Alsakran, X. Huang, Y. Zhao, J. Yang, and K. Fast, "Using entropy-related measures in categorical data visualization," *2014 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 81–88, 2014.

[2] B. Alsallakh, W. Aigner, S. Miksch, and H. Hauser, "Radial sets: Interactive visual analysis of large overlapping sets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2496–2505, 2013.

[3] N. Andrienko and G. Andrienko, *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media, 2006.

[4] E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose, "Toolglass and magic lenses: the see-through interface," *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pp. 73–80, 1993.

[5] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto, "Wirevis: Visualization of categorical, time-varying data from financial transactions," *Visual Analytics Science and Technology, 2007*, pp. 155–162, 2007.

[6] A. Delaney, E. Kow, P. Chapman, and J. Nicholson, "Generating and navigating large euler diagrams," in *Proceedings of the fourth International workshop on Euler diagrams*, 2014.

[7] J. W. Emerson, W. A. Green, B. Schloerke, J. Crowley, D. Cook, H. Hofmann, and H. Wickham, "The generalized pairs plot," *Journal of Computational and Graphical Statistics*, vol. 22, no. 1, pp. 79–91, 2013.

[8] J. A. Fails, A. Karlson, L. Shahamat, and B. Shneiderman, "A visual interface for multivariate temporal data: Finding patterns of events across multiple histories," *Visual Analytics Science And Technology, 2006*, pp. 167–174, 2006.

[9] N. Ferreira, L. Lins, D. Fink, S. Kelling, C. Wood, J. Freire, and C. Silva, "Birdvis: visualizing and understanding bird populations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2374–2383, 2011.

[10] M. Friendly, "Graphical methods for categorical data," *SAS User Group International Conference Proceedings*, vol. 17, pp. 190–200, 1992.

[11] ——, "Mosaic displays for multi-way contingency tables," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 190–200, 1994.

[12] S. Goodwin, J. Dykes, and A. Slingsby, "Visualizing the effects of scale and geography in multivariate comparison," in *2014 IEEE Conference Visual Analytics Science and Technology (VAST)*. IEEE, 2014, pp. 251–252.

[13] S. Goodwin, J. Dykes, A. Slingsby, and C. Turkay, "Visualizing multiple variables across scale and geography," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 599–608, 2016.

[14] H. Guo, F. Chen, and J. Peng, "Geographic elements selection algorithm based on quadtree in variable-scale visualization," *Journal of Multimedia*, vol. 8, no. 2, pp. 137–144, 2013.

[15] H. Hofmann, A. P. Siebes, and A. F. Wilhelm, "Visualizing association rules with interactive mosaic plots," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000, pp. 227–235.

[16] S. Horrigan, "Visualizing relationships among categorical variables," 2008.

[17] J. F. Im, M. J. McGuffin, and R. Leung, "Gplom: the generalized plot matrix for visualizing multidimensional multivariate data," *Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2606–2614, 2013.

[18] R. W. Kerbs, "Blue: An interactive visualization system for categorical data technical note."

[19] E. Kolatch and B. Weinstein, "Cattrees: Dynamic visualization of categorical data using treemaps," *Project report*, 2001.

[20] N. S.-N. Lam and D. A. Quattrochi, "On the issues of scale, resolution, and fractal analysis in the mapping sciences*," *The Professional Geographer*, vol. 44, no. 1, pp. 88–98, 1992.

[21] S. A. Levin, "The problem of pattern and scale in ecology: the robert h. macarthur award lecture," *Ecology*, vol. 73, no. 6, pp. 1943–1967, 1992.

[22] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "Upset: visualization of intersecting sets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1983–1992, 2014.

[23] S. Ma and J. Hellerstein, "Ordering categorical data to improve visualization," *INFOVIS-99*, 1999.

[24] A. Malik, R. Maciejewski, S. Towers, S. McCullough, and D. S. Ebert, "Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1863–1872, 2014.

[25] W. Meulemans, N. H. Riche, B. Speckmann, B. Alper, and T. Dwyer, "Kelpfusion: a hybrid set visualization technique," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 11, pp. 1846 – 1858, 2013.

[26] B. C. Pijanowski, A. Farina, S. H. Gage, S. L. Dumyahn, and B. L. Krause, "What is soundscape ecology? an introduction and overview of an emerging new science," *Landscape ecology*, vol. 26, no. 9, pp. 1213–1232, 2011.

[27] Z. Pousman, J. T. Stasko, and M. Mateas, "Casual information visualization: Depictions of data in everyday life," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1145–1152, 2007.

[28] "The r project for statistical computing," https://www.r-project.org/, accessed: 2016-03-30.

[29] "Record the earth," https://www.recordtheearth.org.

[30] B. Shneiderman, D. Feldman, A. Rose, and X. F. Grau, "Visualizing digital library search results with categorical and hierarchical axes," *Proceedings of the fifth ACM conference on Digital libraries*, pp. 57–66, 2000.

[31] A. Slingsby, J. Dykes, J. Wood, and K. Clarke, "Interactive tag maps and tag clouds for the multiscale exploration of large spatio-temporal datasets," in *Information Visualization, 2007. IV'07. 11th International Conference*. IEEE, 2007, pp. 497–504.

[32] F. Stoffel, H. Janetzko, and F. Mansmann, "Proportions in categorical and geographic data: visualizing the results of political elections," *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 457–464, 2012.

[33] D. Thom, H. Bosch, and T. Ertl, "Inverse document density: A smooth measure for location-dependent term irregularities," in *COLING*, 2012, pp. 2603–2618.

[34] J. J. Thomas and K. A. Cook, Eds., *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[35] C. Turkay, A. Slingsby, H. Hauser, J. Wood, and J. Dykes, "Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2033–2042, 2014.

[36] K. Wongsuphasawat and B. Shneiderman, "Finding comparable temporal categorical records: A similarity measure with an interactive visualization," *Visual Analytics Science and Technology, 2009*, pp. 27–34, 2009.

[37] F. Yates, "Contingency tables involving small numbers and the χ 2 test," *Supplement to the Journal of the Royal Statistical Society*, vol. 1, no. 2, pp. 217–235, 1934.

[38] G. U. Yule, "On the methods of measuring association between two attributes," *Journal of the Royal Statistical Society*, vol. 75, no. 6, pp. 579 – 652, 1912.

[39] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller, "Visual correlation analysis of numerical and categorical data on the correlation map," *Visualization and Computer Graphics*, vol. 21, no. 2, pp. 289–303, 2015.