Methods for Multiclass Geospatial Data Visualization

by

Rui Zhang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2022 by the
Graduate Supervisory Committee:

Ross Maciejewski, Chair
Jorge Sefair
Chris Bryan
Sharon Hsiao

ARIZONA STATE UNIVERSITY

August 2022

ABSTRACT

Geographical visualizations are critical for multi-criteria analysis, optimization, and decision making, where the translation of spatial data into a visual form allows analysts to quickly see patterns, explore summaries and relate domain knowledge about underlying geographical phenomena. However, several critical challenges arise when visualizing large spatiotemporal datasets. While, the underlying geographical component of the data lends itself well to univariate visualization in the form of traditional cartographic representations (e.g., choropleth, isopleth, dasymetric maps), as the data becomes multivariate, cartographic representations become more complex, requiring new approaches for multiclass map visualization and exploration. In this thesis, novel visual analytics methods and frameworks are proposed to support multiclass map analysis. An interactive conservation portfolio development system that combines visualization, multicriteria analysis, optimization, and decision making is developed that showcases a novel visualization and interaction design to compare different purchasing profiles under various optimization constraints. Such multiclass map analysis is then extended using concepts from scalar field topology for hotspot analysis including the introduction of a novel visualization construct combining Merge Trees and Streamgraphs.

ACKNOWLEDGMENTS

I would like to express my most sincere gratitude to my advisor, Dr. Ross Maciejewski, for his consistent support and trust. Whenever I encountered difficulties, he always stood with me and helped me get through them. If I was ever confused about my research direction, Dr. Maciejewski was willing to give me an opportunity to explore and was very patient in teaching me. I could not imagine having a better advisor.

I would also like to thank Dr. Jorge Sefair, Dr. Chris Bryan, and Dr. Ihan Hsiao for being my committee members and for their valuable suggestions. Their insightful thoughts on the comprehensive exam and proposal helped me to dig into my research and critically think about my future work. I also thank all my collaborators, Dr. Jonas Lukasczyk, Dr. Yafeng Lu, and Dr. Hong Wang. I learned a lot from them. Thanks to Dr. Guoliang Xue for giving me the opportunity to study at ASU.

I would like to thank my friends, Michael Steptoe, Dr. Xin Ye, Dr. Zhaosong Huang, Rui Shi, Qun Zhao, and Dr. Gong Ze, for being with me. We learned together and played together. Because of you, my Ph.D. life became better.

In the end, I would like to thank my parents, Qiaoling Zhang and Haican Zhang. Thank you for raising me. Thank you my husband Yucen Tao and my cute son Luka Tao. Because of you, I feel loved and find new meaning in life.

This material is based upon work supported by the National Science Foundation under Grant No. 1350573 and the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-02-00. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER

1 INTRODUCTION 1

1.1 Motivation 1

1.2 Problem Statement 3

1.3 Aim of the Work 4

1.4 Outline and Individual Contributions 6

2 A VISUAL ANALYTICS FRAMEWORK FOR CONSERVATION PLANNING 8

2.1 Introduction 8

2.2 Related Work 11

2.3 Visual Analytics Framework 15

2.3.1 Data Storage and Downscaling 18

2.3.2 Multi-layer Map View 21

2.3.3 Attribute Selection View 22

2.3.4 Conservation Portfolio Optimization 24

2.3.5 Portfolio Comparison View 28

2.4 Case Studies 31

2.4.1 Multi-species Conservation Scenarios for the Judith Gap in Montana 31

2.4.2 Creating a Protected Area in Montana's Park and Sweet Grass Counties 36

2.5 Conclusions 38

CHAPTER Page

3 EXPLORING GEOGRAPHIC HOTSPOTS USING TOPOLOGICAL DATA ANALYSIS 40

3.1 Introduction 40

3.2 Related Work 43

3.3 Method 46

3.3.1 Kernel Density Estimation (KDE) 47

3.3.2 Superlevel Set Segmentation (SSS) 48

3.3.3 Merge Tree Segmentation (MTS) 49

3.3.4 Merge Tree Leaf Segmentation (MTLS) 50

3.3.5 Merge Tree Crown Segmentation (MTCS) 50

3.3.6 Morse Complex Segmentation (MCS) 52

3.3.7 Combining Segmentations (MTCS+MCS) 52

3.3.8 Topological Simplification 53

3.3.9 Visual Analytics Framework 55

3.4 Case Studies 57

3.4.1 Criminal Incident Reports from Tippecanoe County, IN 58

3.4.2 Gang Analysis 63

3.5 Conclusion 68

4 VISUALIZING MULTIPLE SCALAR FIELDS USING REPRESENTATIVE TOPOLOGICAL FEATURES 71

4.1 Introduction 72

4.2 Background and Related Work 73

4.2.1 Scalar Field Topology 73

4.2.2 Topology-Based Visualization of Multiple Scalar Fields 75

CHAPTER Page

4.2.3 Stacked Graphs 77

4.3 Method 79

4.3.1 Reference Fields 81

4.3.2 Merge Tree Segmentations 82

4.3.3 Visualization 82

4.3.4 Implementation 84

4.4 Usage Scenarios 85

4.4.1 Gang Activity in Chicago 87

4.4.2 Climate Ensemble 91

4.4.3 Asteroid Ocean Impacts 94

4.5 Conclusion 95

5 CONCLUSION 97

REFERENCES 102

LIST OF TABLES

Table Page

2.1 Variables and Data Sources Used 20

LIST OF FIGURES

Figure Page

2.1 A Visual Analytics Framework for Conservation Planning. (A) Map View: Visualizes Parcels Falling in the Analyst-defined Search Area. Colors in the Map Reflect the Median Ranking, Attributes, or Optimization Results. Users Can Create a Customized Portfolio with the Draw, Pan, and Zoom Controls in the Upper Left Corner. The Legend and the Differences of the Constraints and Goals Reached by the Portfolios Are Shown Separately in the Left and Right Bottom Corner. (B) Attribute Analysis View: Attribute Distributions Are Visualized by Line Charts and Parallel Coordinates Integrated with a Box Plot. (C) Optimization Configuration: The Median Ranking Slider Filters out Low-quality Parcels and Reduce the Optimization Algorithm Runtime. The Radio and Text Box Are Used to Input the Optimization Constraints and Objective Function. Users Click the 'optimization' Button to Run the Algorithm and 'save' the Current Portfolios on the Map to the 'user's Collection' for Comparisons. (D) Portfolio Comparisons: The Generated Portfolios Are Saved, and a Screenshot of the Selected Portfolios and Attributes Are Visualized. (E) Attribute Lists: Drop-down Menu for Selection. Event Sequence: 1) Select the Land Attributes to Explore. 2) Explore the Land Attributes to Define the Study Area. 3) Display the Distribution of the Attribute for the Parcels. 4) Filter Parcels by Brushing the Attributes Range and the Median Ranking Range. 5) Set the Constraints and Objective Function to Get the Optimal Result. 6) Adjust Portfolio Based on the Optimal Result. 7) Save the Current Portfolio for Future Comparisons. 16

2.2 Examples of Multi-layer Map Views. (A) Visualization of Distance to Metro Areas. The Pink Area Is the Metro Area and the Peripheral Region Is Colored from Brown to Yellow Based on the Distance Value of Each Patch. The Red/Blue Highlights Correspond to the User Selected Region of Interest. (B) Visualization of Fish Species Richness. The Region with Bluer Color Has Lower Species Richness for Fish. 21

2.3 Comparison of the Average Value for Each Attribute. The Selected Attribute with a Gray Arc Is the Pa Attribute of the Analyst's Portfolio in the Second Row. All the Comparison Result of Other Portfolios with This Selected Attribute Are given with Three Relations, "less Than", "larger Than", and "equal To". The Corresponding Signals in the Comparison View Are a minus Sign, a plus Sign, and an Equal Sign, Respectively. For Example, the Average Values of Pa for an Analyst's Selected Set of Parcels in the First Row Are Smaller than That of the Analyst's Selected Set of Parcels in the Second Row. Comparing Analyst's Portfolios of Plan1 and Plan2, Plan1 Can Buy 314 Million Square Meters Area Better than the 280 Million Square Meters in Plan2 with a Similar Cost. In Addition, the 7 Attributes out of 12 Attributes in Plan1 Perform Better than Those in Plan2. In General, Plan1 Is Much Better than Plan2. 29

2.4 Process to Define the Candidate Region: (A) PA Layer. (B) MA Layer Covering PA Layer. (C) HW Layer Covering MA and PA Layers. (D) HII Layer Covering PA Layer. (E) Cost Layer Covering PA Layer. (F) Selected Candidate Region with Median Ranking Covering PA Layer. . . 33

2.5 Generated Portfolios on the Map. (A) Portfolios Generated with $10m Budget Connecting Two Protected Areas. (B) Portfolios Generated with $15m Budget Connecting Two Protected Areas. (C) Portfolios Generated with $15m Budget Connecting Three Protected Areas. All the Parcels Are Prepossessed by Filtering Those with Less than 54 Mammal Species Richness. Portfolio Comparisons: (A,B) the Portfolios Generated Based on the Small Search Area with the Different Budgets Are Compared. (B,C) with the Same Budget, the Portfolios Generated Based on the Small Search Area and Larger Search Area Are Compared. 34

2.6 Case Study 2: (A) Initial Ranking given Selected Attributes in Montana's Park and Sweet Grass Counties. (B) Pre-processed Areas Using Median Ranking. (C) Optimal Results When Total Purchase Cost Is Minimized Subject to a Minimum Area of 300 Km^2; (D) Results after Manually Inducing Compactness in the Protected Area. (E) Comparison Between the Optimal Portfolio and the Connected and Compact Analyst-selected Area. 36

Figure Page

3.1 A Common Way to Characterize Hotspots in Point Events (A) Is to First Derive a Kernel Density Estimate (B), and Then to Extract All Areas That Exceed a Given Density Value Threshold (C-D). Here, Events Correspond to Reported Gang Activity in Chicago, IL, USA (Section 3.4.2), Where Each Gang Is Assigned a Random Color. As Demonstrated in the Bottom Row, the Chosen Threshold Has a Significant Impact on the Shape and Number of Extracted Hotspots (Colored Randomly). More Importantly, a Single Global Threshold Can Not Distinguish Between Hotspots at Different Scales, E.G., For the Threshold That Is Required to Detect All Small Hotspots Shown In (D), All Major Hotspots Shown in (C) Can No Longer Be Separated. 40

3.2 The Proposed Scalar Field Topology (SFT)-based Visual Analytics Interface Consists of Four Linked Views: (I.1) a Merge Tree, (I.2) a Persistence Diagram, (I.3) a Persistence Curve, and (I.4) a Geospatial Map. Given a Set of Geo-spatial Point Events—here, Crime Incident Reports in Tippecanoe County, IL, USA (Section 3.4.2)—the Framework Computes a 2D Kernel Density Estimate (Contours), and Then Analysts Can Interactively Extract Hotspots (Colored Elements) via Different SFT-based Characterizations (A.1-d.2). (A.1) and (A.2) Both Show Hotspots That Are Characterized via a Merge Tree Leaf Segmentation (MTLS), Where Hotspots of A.1 Are Colored Based on the Id of Their Corresponding Merge Tree Branch, and Hotspots of A.2 Are Colored Based on the Most Frequent (Majority) Crime Category in Each Hotspot. (B.1), (B.2), and (B.3) Show the Same Morse Complex Segmentations (MCS) Where Hotspots Are Colored by the Most Frequent, the Second Most Frequent, and the Third Most Frequent Crime Category, Respectively. (C) Shows a Combination of a Superlevel Set Segmentation (Sss) for Level 0.7 and a MCS, Where Hotspots Are Colored by the Majority Crime Type. (D.1) and (D.2) Show the Combination of a MCS with Two Different Merge Tree Crown Segmentations (MTCS), Where Hotspots Are Again Colored by the Majority Crime Type. 44

3.3 Hotspots of a Density Field Can Be Characterized via Several SFT-based Abstractions. Superlevel Set Segmentations (SSS) Identify Connected Areas That Exceed a given Level Threshold. The Merge Tree Segmentation (MTS) Associates Each Point of the Domain with a Branch of the Merge Tree. As a Special Case, the Merge Tree Leaf Segmentation (MTLS) Extracts Only Parts of the Domain That Are Associated with Leaf Arcs. Another Special Case Is the Merge Tree Crown Segmentation (MTCS) That Extracts Subtrees Based on a Relative Scalar Value Difference, Called Relevance. Finally, the Morse Complex Segmentation (MCS) Associates Each Point of the Domain with the Maximum That Is Reached by Following the Steepest Ascend. 51

3.4 Combining the Merge Tree Crown Segmentation (MTCS) and the Morse Complex Segmentation (MCS) Yields a Very Versatile, Robust, and Effective Hotspot Characterization. Hotspot Boundaries Are Determined by the MTCS, and the MCS Is Used to Differentiate Between Individual Hotspots. 52

3.5 Topological Simplification (Lukasczyk *et al.*, 2020; Edelsbrunner *et al.*, 2006) Makes It Possible to Remove Hotspots That Do Not Exceed a given Persistence Threshold (Red Dashed Line) by Symbolically Removing the Corresponding Maximum. Filtered Hotspots Are Assigned to the Region of Preserved Hotspots, Which Enables Analysts to Examine Hotspots in a Level-of-detail Approach. For Instance, Large Persistence Thresholds Preserve Only the Most Significant Hotspots, and by Continuously Decreasing the Persistence Threshold These Hotspots Are Incrementally Partioned into Smaller Hotspot Regions. For the Example Scalar Field Shown Here (A), the Maximum c Has Been Removed as Its Pair Does Not Exceed the Current Persistence Threshold of 1.5. Its Corresponding Region Is Therefore Assigned to the Region of Maximum d. This Is Also Indicated in the Merge Tree (B), Where We Render Filtered Branches as Gray Lines. To Choose Appropriate Persistence Thresholds, Topological Data Analysts Usually Examine Two Visualizations: The Persistence Diagram and the Persistence Curve (Edelsbrunner and Harer, 2010). The Persistence Diagram (C) Renders Each Persistence Pair as a Line Which Indicates for Which Level the Pair Is Born and Destroyed. The Persistence of Each Pair Is Encoded by the Height of Its Corresponding Line, and the Persistence Threshold Is Rendered as a Line Parallel to the Main Diagonal. Hence, Less Persistence Features Are Close to the Diagonal. The Persistence Curve (D) Shows the Total Number of Pairs (Y-axis) That Exceed a given Persistence Threshold (X-axis). 54

3.6 This Figure Shows a Compilation of Different Hotspot Characterizations for Gang Activity in Chicago, IL, USA (Section 3.4.2): (A) Merge Tree Leaf Segmentation, (B.1-B.2) Merge Tree Crown Segmentation, (C) Morse Complex Segmentation, and (D) Combination of a Merge Tree Crown Segmentation and a Morse Complex Segmentation. Each Characterization Is Shown as a Merge Tree-map Pair, Where Special Parameters of Certain Characterizations Are Stated in the Corresponding Subcaptions. In the Left Column, Hotspots Are Colored Based on the Id of Their Corresponding Merge Tree Elements, Whereas Hotspots in the Right Column Are Colored Based on the Most Active Gang Inside Each Hotspot (Except for (D), Which Is Always Colored by Branch Id). 62

4.1 The Proposed Stream Tree View (Center), Illustrating the Distribution of Gang Activity in Chicago, Illinois. Here, the Analyst-defined Reference Field r Corresponds to the Density Estimate of All Gang Activity in Chicago (Contours, Left). The Merge Tree of r (Black Edges, Center) Indicates for Which Density Value (Y-axis) Individual Gang Activity Hotspots Appear and Merge, I.E., Each Point of the Merge Tree Corresponds to a Unique Superlevel Set Component (Hotspot). Here, Each Region of the Map Is Colored by the Gang That Has the Largest Integral in the Corresponding Superlevel Set Component. With the Stream Tree, Analysts Can Determine at a Glance Which Gangs Are Active in the Individual Regions, and Which Regions Are Clearly Dominated by a Single Gang or Are Contested. For Example, the Hatched Region in the Lower Right Corner of the Map Corresponds to the Highlighted Edge of the Merge Tree. In This Area, Almost All Reported Gang Activity Is Attributed to the *Latin Kings* (Orange), Which Is Effectively Visualized with the Stream Tree and the Coloring of the Map. All Other Hotspots Are Clearly Dominated by the *Gangster Disciples* (Teal), Except for One Hotspot Where Slightly More Activity of the *Black P Stones* (Red) Gang Was Reported. To Further Reduce Visual Clutter and Emphasize Proportions, One Can Also Visualize Attribute Vectors with Donut Charts Drawn at the Critical Points of the Merge Tree (Right). 71

4.2 Illustration of the Proposed Visualization Pipeline Whose Input Is a Set of n Scalar Fields That Are Defined on the Same Domain $\mathcal{D}$ (A-C), I.E., A Vector Function $f : \mathcal{D} \to \mathbb{R}^n$. First, the Analyst Chooses a Reference Field r to Base Their Analysis on, Which Can Be a Specific Input Field or an Aggregation of the Input Fields (Section 4.3.1). Here, r Is the Sum of All Input Fields (D). Next, the Algorithm Computes the Merge Tree Segmentation $\phi : \mathcal{T} \to 2^{\mathcal{D}}$ of r That Maps Each Point of the Merge Tree $\mathcal{T}$ (Black Edges of E and F) to Its Corresponding Superlevel Set Component $\mathcal{S} \subseteq \mathcal{D}$ of r (Section 4.3.2). For Instance, the Red Point in (E) Maps to the Interior of the Dashed Region in (D). Given a Set of Sample Merge Tree Points $p \subseteq \mathcal{T}$ the Algorithm Then Computes the Component-wise Integral of f over the Area of the Corresponding Superlevel Set Components. This Yields, for Each Merge Tree Point $p \in P$, an Attribute Vector $a^{\phi(P)}$ Whose Entries Correspond to the Individual Integrals, I.E., $a_i^{\phi(P)} = \int_{\phi(P)} f_i(X)Dx$. In This Example, the Attribute Vector of the Red Point Records the Integrals of the Green, Yellow, and Red Scalar Field over the Area of the Dashed Region. The Magnitude and Ratio Between the Attribute Vector Entries Can Then Be Visualized with a Stream Graph (E) along the Merge Tree Edges (Section 4.3.3). Alternatively, It Is Possible to Visualize the Attribute Vectors with Donut or Pie Charts at the Critical Points of the Merge Tree (Left and Right Tree in (F)). 80

4.3 In This Gang Activity Analysis the Reference Field (Left) Corresponds to the Kernel Density Estimate of the *Gangster Disciples* (Green), Where Each Region Is Colored by the Second Most Reported Gang in That Region. As Shown in the Donut Tree (Right), the *Gangster Disciples* Are by Far the Most Reported Gang in All Their Hotspots. However, the Second Most Reported Gang Varies from Hotspot to Hotspot, Where the Donut Tree Clearly Shows Which Hotspots Are Contested by Which Gangs. For Instance, the Second Most Reported Gang in the Highlighted Hotspot (Thick Edge in the Tree and Hatched Area on the Map) Are the *Black Disciples* (Yellow). An Outlier Is the Purple Hotspot, Where the Second Most Reports Can Not Be Attributed to Any Specific Gang (*Not Available*) as There Are Many Original Input Reports in That Region That Have No Gang Label. . . . 86

4.4 This Analysis Focuses on Reports That Have No Gang Label, I.E., The Reference Field That Corresponds to the Kernel Density Estimate of the Reports with the Label *Not Available*. In the Map Hotspots Are Colored by the Most Reported Gang in That Region (Left). The Stream Tree with Fixed Scaling (Right) Shows Which Gangs Are Frequently Reported in Areas with Many Unlabeled Reports. Not Surprisingly, in Many of These Regions the *Gangster Disciples* Have the Most Reports, but the Stream Tree Clearly Indicates Regions in Which Other Gangs Are Rarely Reported (Such as in the Left Most Edge) or Other Gangs Are Almost as Frequently Reported (Such as the Third Edge from the Left). The Highlighted Hotspot, However, Stands out since in This Hotspot the Amount of Reports of the *Black P Stones* Is Almost Four times the Number of Reports of the *Gangster Disciples*; And Therefore Making Them the Most Likely Gang to Be Associated with Unlabeled Reports in That Region. This Example Indicates That Our Approach Can Also Be Used for Probabilistic Relabeling. 88

4.5 Analysis Results of the *Grand Ensemble* That Simulates Climate Change under Three Different RCP Scenarios: RCP 2.6, 4.5, and 8.5. Each Scenario Was Simulated 100 times, and the Three Scenarios Roughly Correspond to the Best, Intermediate, and Worst Case. First, We Compute for Each Scenario the Mean Surface Temperature (MST) for the Year 2100 by Aggregating the Corresponding Ensemble Members. The MST of the RCP 2.6 Scenario Is Shown in (A). Next, We Compute the Difference Fields Between the RCP 4.5 (B) and RCP 8.5 (C) Scenario with Respect to the RCP 2.6 MST Field. Then We Select the RCP 2.6 MST Field as a Reference Field, and the Two Difference Fields as the Integrands. The Corresponding Stream Tree (D) and Donut Tree (E) Show the Proportions Between the Temperature Increases Predicted by the RCP 4.5 and RCP 8.5 Scenarios in Superlevel Set Components (Segments) of the MST Field Predicted by the RCP 2.6 Scenario. Although the Stream Tree Makes It Easier to Follow Edges, It Is More Difficult to Judge the Proportions Between the Two Scenarios. The Donut Tree, on the Other Hand, Produces More Visual Clutter but Better Encodes the Ratios Between the Scenarios. . . 90

4.6 Results of the Asteroid Impact Example Where We Visualize the Proportions Between Water (Blue) and Asteroid (Brown) matter Inside Superlevel Set Components (Segments) of the Temperature Field. The Corresponding Stream Tree with Fixed Scaling (a) Clearly Indicates the Fraction Between Water and Asteroid Matter Inside the Individual Segments. The Stream Tree Makes It Possible to Quickly Identify Which Segments Consist Almost Entirely of Water, and Which Segments Contain Asteroid Matter. The Stream Tree Also Shows That as One Goes Further up in the Tree I.E., As the Corresponding Segments Contract Towards Their Respective Temperature Maxima the Proportion of Asteroid Matter Increases. This Indicates That Asteroid Fragments Are Close the Temperature Maxima, I.E., At the Core of These Segments. Utilizing the Stream Tree It Is Possible to Partition the Individual Segments into Two Groups: The Ones That Contain Some Asteroid Matter (B), and the Ones That Do Not Contain Asteroid Matter (C). 92

Chapter 1

INTRODUCTION

1.1 Motivation

Given the large-scale instrumentation of places and devices, multiclass geospatial data is widely available, and the analysis of such data requires novel tools and techniques to operationalize its use in real-world problem solving, such as emergency response planning, resource allocation, forestry, wildfires (Koutsias *et al.*, 2014), air traffic patterns (Lampe and Hauser, 2011), human activity (Hu *et al.*, 2014), animal movements (Sarkar *et al.*, 2015; Chirima and Owen-Smith, 2017), crime analysis (Nakaya and Yano, 2010; Levine, 2008), urban analysis (Zhou, 2015), health (Liadsky and Ceh, 2017), etc. To support multiclass geospatial data analysis, recent work has focused on how to display multiple attributes for the geospatial data in the analysis. For example, Turkay et al. (Turkay *et al.*, 2014) explores the geographic variation of multivariate data and developed the attribute signatures method to dynamically generated graphs to summarize the change of statistics over a sequence of geospatial data selection. (Ferreira *et al.*, 2015) proposed a 3D framework to help urban developers when planning new architectural structures. In their framework, the user can explore buildings and their environment through parallel coordinates and a table view to help urban planners identify environmental factors that could be critical for their building development.

However, the scale and dimensionality of available data still leads to challenges in data analysis and visualization. For exploring underlying relationships and patterns in multiclass geospatial data, recent works often apply aggregation techniques to reduce data sizes and provide an overview of data distributions. For example, Maciejewski et al. explores spatiotemporal changes in emergency department records using kernel density estimation (Maciejewski *et al.*, 2009). Scheepens et al. apply kernel density estimation to trajectory aggregation and use contour lines to help analysts predict the movement of ships (Scheepens *et al.*, 2014, 2012).

There is also a need to move beyond data exploration and develop tools that can support the decision-making process. (Konev *et al.*, 2014) proposed an automatic simulation-based approach for flood management. The decision trees are automatically generated and visualized by clustered timelines. (Rinner, 2007) developed a geographic visualization system to support multi-criteria decision making. An index rank for different land tracts is calculated, and users can explore attributes through a linked parallel coordinate plot. However, portfolio comparisons and interaction with the optimization results are limited in these systems. Given the lack of human involvement, it's not easy to generate a robust and optimal result satisfying users' preferences.

Given the challenges in data size and dimensionality as well as the need for robust decision making support, I have identified three main challenges related to multiclass map exploration: (1) Visualizing the multiclass geospatial data. (2) Adopting models/methods to explore the relationship and pattern of the data. (3) Involvement of a Human-in-the-loop for decision making. My proposed works explores multiclass geospatial data and proposes novel methods to support high-dimensional visualization. This work adopts topological methods to explore and detect geographic relationships and patterns, and I have designed and implemented visualization systems to support analysts in decision-making.

1.2 Problem Statement

The multiclass geospatial data contains a lot of information. To take advantage of the information, visualization of them is necessary and useful. The multiclass geospatial data is huge location-based information with multiple variables, which makes it difficult to explore. Exploring the multiclass geospatial data can be done in spatial scenes and non-spatial scenes. The geospatial data has location information. The common way to visualize them is to project them on the map with colors. However, there are some challenges to project the data on the map. First, the amount of data is usually huge. Loading such huge data on the map is difficult. The scale of the map also increases the difficulty of the projection. Second, the multiclass geospatial data contains multiple variables. No matter visualizing multiple variables in the spatial scenes or the non-spatial scenes, visualization of multiple variables is usually a hot topic. That is, visualizing huge geospatial data with multiple variables in real-time is still challenging. Beyond being visualized in the spatial scenes, the multiclass geospatial data can be visualized in the non-spatial scenes, such as tables, parallel coordinates, and many others. However, most of the work lacks the interaction between spatial visualization and non-spatial visualization.

Detecting and learning the underlying relationships and patterns of the multiclass geospatial data usually need the help of some methods. For example, how to measure the data with multiple variables. The data could have a ranking based on each variable. Given multiple different variables, it's not easy to evaluate them. For the point event data with multiple variables, how to get the hotspots of them? which variable contributes most to the hotspot? How does the variable change in different hotspots? All these questions need to solve with the application of other methods, such as some algorithms, models and etc. These methods need researchers to study and explore.

No matter how to explore the data and detect the data's pattern, the way has to satisfy analysts' requirements and preferences. Therefore, the involvement of humans is very important. Usually, analysts could have some basic interactions with the visualization view, such as hovering to highlight the selected data. However, adding analysts' preference into the visualization view and the generated result is barely studied in the recent work.

In summary, there is a lack of methods to visualize multiclass geospatial data, which involve visualizing huge geospatial data in real-time, representing multiple variables of geospatial data in the spatial and non-spatial scenes, interacting with humans, providing methods to detect the pattern and complex relationship of multiclass geospatial data.

1.3 Aim of the Work

The aim of the thesis is to develop novel methods to support multiclass geospatial data analysis. To explore multiclass geospatial data, I have developed a visual analytics framework for optimization and planning with multiclass geospatial data. The framework is developed to support conservation planning. Conservation agencies have limited resources (in terms of budget) for investing in new conservation areas and have differing priorities for conservation in terms of species, vegetation, human activities, etc. In general, decision makers need to look at these variables and explore the available units of land to purchase. All of these different priorities result in different optimization criteria that can be applied across the same spatial area, resulting in different land portfolio purchasing decisions. As the number of variables and size of the area of analysis increases, solving various spatial optimization formulations can be computationally infeasible. However, this problem lends itself well to a human-in-the-loop process. Here, human's can define key variables of interest, identify unimportant land parcels, and interact with the data to reduce the computational space. This system incorporates a multi-layer map view, a parallel coordinates attribute view, a control area for optimization modeling, and a small multiple portfolio visualization for decision comparison. To support automatic portfolio optimization, I have implemented a median ranking algorithm to allow parcel filtering by an aggregated indicator of all the attributes and an integer programming optimization algorithm to generate land purchase decisions given user-defined constraints and objective functions. The analytics procedure starts from search area selection on the map view, attribute exploration and parcel filtering, to automatic portfolio generation and user interactive modification. Multiple decisions can be generated and saved for comparison.

While the conservation planning portfolio developed novel views for multiclass map exploration, it became clear that new techniques were needed to effectively visualize differences between regions. To detect the complex relationship and patterns in multiclass geospatial data, I have explored the application of scalar field topology for multiclass map analysis. Point event data can be transformed into a scale field by an application of KDE (Kernel Density Estimation) method, and then analysts typically explore these fields looking for high-density areas or hotspots. The common way to extract hotspots is to extract the scalar fields with high density. However, this can obscure local peaks and unintentionally highlight noise and outliers. To overcome such issues, I have employed a scalar field topology (SFT)-based methodology for the interactive characterization and analysis of hotspots for density fields defined on a regular grid. This method makes it possible to filter hotspots by significance, to identify hotspot boundaries, and to understand their hierarchical and spatial relationship. I have instantiated these SFT methods in a visual analytics framework that includes a map view, a merge tree view, a persistence diagram view, and a persistence curve view. This enables the exploration of what SFT methods are relevant for multiclass map analysis.

Since SFT methods can explore hotspots using a level-of-detail approach, how to best profile multiclass regions based on the SFT methods will be interesting to explore. When hotspots contains multiple categories, such as different gang names, the goal is to explore how the various classes are distributed in the hotspots. I combine the merge three and stream graph into a novel visualization view, stream tree for multiple scalar fields that are defined over the same domain.

In summary, this thesis studies and tries to solve the challenges of multiclass geospatial data analysis. This thesis does the following:

- Develop novel methods to support the exploration of multiclass geospatial data.
- Apply novel methods to analyze the multiclass map and profile multiclass regions.
- Create tools to enable human-in-the-loop decision-making for multiclass map problems.

1.4 Outline and Individual Contributions

This thesis contains three chapters: Chapter 2 (one published paper), Chapter 3 (one published paper), and Chapter 4 (one submitted paper under review). I provide the title of the paper and the contribution of the first author. While the author of the thesis is the first author of the three papers. These papers are the results of several collaborations among researchers.

A Visual Analytics Framework for Conservation Planning

The first author **Rui Zhang** collected data, prepossessed data, designed and implemented the system, ran case studies, recorded the demo video, and wrote the paper with other authors.

Exploring Geographic Hotspots Using Topological Data Analysis

The first author **Rui Zhang** collected data, prepossessed data, set up the TTK environment guided by another author, designed and implemented the system, ran case studies, recorded the demo video, and wrote the paper with other authors.

Stream Trees: Visualizing Multiple Scalar Fields using Representative Topological Features

The first author **Rui Zhang** collected data, prepossessed data, set up the TTK environment guided by another author, designed and implemented the system, and wrote the paper with other authors.

Chapter 2

A VISUAL ANALYTICS FRAMEWORK FOR CONSERVATION PLANNING

2.1 Introduction

Biodiversity is declining at rapid rates due to human-driven habitat loss and landscape deterioration (Stokstad, 2010). Human activities have resulted in species extinctions at 10 – 100 times normal 'background' extinction levels (Sala *et al.*, 2000; Pimm *et al.*, 1995). This rapid biodiversity decline threatens the provision of key ecosystem services such as food, clean water, and crop pollination, resulting in negative consequences for economies and human health (Mace *et al.*, 2012). Therefore, protecting remaining natural areas is fundamental to preserve biodiversity and to mitigate the negative consequences of ongoing environmental change (Johnson *et al.*, 2017). While 15% of the Earth is in some kind of protection (Belle *et al.*, 2018), this is still insufficient due to substantial gaps in land coverage and increasing threats (Rodrigues *et al.*, 2004). Recent studies suggest the need of a drastic increase in protected lands by 2050 to maintain current rates of resource extraction (Watson and Venter, 2017; Dinerstein *et al.*, 2019). This ambitious goal contrasts with the limited resources available to local institutions to design and implement networks of protected areas (Bicknell *et al.*, 2017).

Conservation biologists apply systematic conservation planning approaches to design protected area networks that are cost-effective while meeting conservation goals. This systematic process is composed of six steps that include: (1) biodiversity data collection and analysis, (2) identification of conservation goals, (3) analysis of current conservation areas, (4) identification of a set of additional areas, (5) implementation of proposed conservation actions, and (6) preservation of required conservation values (Margules and Pressey, 2000). Out of these, we focus on Step (4), the identification of additional conservation areas—one of the most challenging steps in the conservation planning process. These conservation areas are selected with multiple conservation goals in mind, such as maximizing biodiversity representation while attenuating future threats, and remaining within a limited budget (Luck *et al.*, 2012). This is a complex selection process that, if performed using inadequate quantitative tools, may result in landscape or seascape portfolios that are not optimal in terms of their budget and priorities. Therefore, the success of systematic conservation planning rests, in part, in the development of appropriate data-driven methodologies for designing protected area networks at the regional level (Williams *et al.*, 2005).

Web-based geographic information systems (WB-GIS) provide an ideal setting to translate the result of complex spatial mathematical models used in systematic conservation planning into simple qualitative visual scenarios (Dragicevic, 2004). These WB-GIS can summarize multiple layers of information, allowing planners to analyze various future hypothetical scenarios (Rao *et al.*, 2007). While multiple mathematical models are available to prioritize areas for conservation (Sarkar *et al.*, 2006; Moilanen *et al.*, 2009), designing a network of protected areas requires the quantification, visualization, and adjustment of multiple hypothetical scenarios almost simultaneously (Tress and Tress, 2003; Pettit *et al.*, 2011). In this context, typical questions faced by conservation analysts include, what areas should be selected as part of a network of protected areas to have all species of conservation concern under protection while minimizing the acquiring costs? If we decrease the budget by 10%, which areas should be protected? What happens if, instead, we increase the budget by 5%? Therefore, there is a need for WB-GIS applications for conservation planning that combine cost optimization with efficient visualization tools that can provide alternative future scenarios in real time (Portman, 2014).

In this thesis, we present an interactive conservation portfolio development system that combines visualization, multicriteria analysis, optimization, and decision making that enables conservation planners and scientists to explore different land purchasing portfolios under a variety of constraints in real time. Our system incorporates a multi-layer map view, a parallel coordinates attribute view, a control area for optimization modeling, and a multiple portfolio visualization for solution comparison. To support automatic portfolio optimization, we implemented a median ranking algorithm to allow parcel filtering by an aggregated indicator of all the attributes and an integer programming model to generate land purchase recommendations given user-defined constraints and objective function. The visual analytics system is designed to support the efficient selection of conservation areas by enabling portfolio generation and interactive modification. Multiple land portfolios can be generated and saved for comparison. Our system complements the existing body of tools by providing new visual, analytical, and mathematical features, while also allowing loading of a (shape compatible) conservation plan obtained with any other tool for further visual analysis.

From the software systems perspective, we propose a novel combination of visualization components, where our design has focused on featuring credibility, saliency, and legitimacy (White *et al.*, 2010). The software tool itself serves as a boundary object to enable decision making. Our design ranges from providing detail-on-demand for the data source to enable analysts to determine credibility of data layers, interactive selection of optimization criteria, and provenance analysis. Specifically, for supporting provenance analysis and comparison, we propose new visualization designs to capture different portfolios and provide comparison between them. Along with novel integration of techniques and the proposal of a visualization design, we have also designed a pre-processing scheme to match data across different levels of granularity. Our down-sampling technique allows data comparisons at a high-resolution level, and supports land purchases which can only occur at the parcel level. The human-machine combination is also innovative, where our framework is designed to present an optimal solution within the problem formulation; however, the problem formulation needed to be computationally efficient for real-time exploration. By providing an optimal conservation portfolio as a first pass, we allow users to refine their choices in a human-machine teaming process. Our system enables conservation planners to develop consecutive portfolios in real-time and adjust the outputs of the multiple criteria optimization selections. From an optimization perspective, our proposed approach uses the analyst's preferences to drastically reduce the problem size. By supporting interaction with the optimization results, planners can utilize species specific knowledge to enforce different landscape features (i.e., connectivity, compactness, corridor width) which can be challenging for automatic optimization either due to the difficulty in representing the corresponding constraints or the computational complexity of the resulting model. Although we propose a simple (and fast) optimization model to support land acquisition decisions, the proposed system can accommodate the results from other spatial models. As a result, our system can be seen as a visualization framework that supports user interaction with an optimal solution. To our knowledge, currently there is no tool available with the proposed features to support conservation decisions.

2.2 Related Work

Our system is designed to support decision making through multicriteria analysis and solution comparison. In this section, we summarize previous work in visual analytics for decision making and multicriteria optimization for conservation planning.

Recent visualization work has focused on how to best display multiple attributes for analysis, a key component of multicriteria analysis. (Turkay *et al.*, 2014) explored the geographic variation of multivariate data and developed attribute signatures consisting of dynamically generated graphs to summarize the change of statistics over a sequence of geospatial data selection. (Pajer *et al.*, 2017) proposed WeightLifter, a technique that allows the exploration of weight space with up to ten criteria and helps to explore the sensitivity of candidate solutions to the change of weights. (Sorger *et al.*, 2016) proposed a visual analytics system, LiteVis, to support lighting design. LiteVis integrates spatial scene visualization, non-spatial model parameterization and multi-objective ranking to help build and compare lighting designs. (Weng *et al.*, 2018b) designed a visual analytics system, ReACH, which helps analysts identify their ideal home given multiple purchasing constraints. (Ferreira *et al.*, 2015) proposed a 3D framework to help urban developers when planning new architectural structures. In this framework, the user can explore buildings and their environment through parallel coordinates and a table view to help urban planners identify environmental factors that could be critical for their building development.

Common amongst many of these systems are the use of parallel coordinates plots, and a wide variety of extensions to parallel coordinate plots. (Lind *et al.*, 2009) proposed a many-to-many relational parallel coordinates plot, and (Johansson *et al.*, 2005) proposed a multi-relational 3D parallel coordinates plot for displaying complex patterns. (Rosenbaum *et al.*, 2012) proposed a progressive parallel coordinates plot which reduces the amount of data shown while retaining the underlying patterns. (Xie *et al.*, 2017) visualized the probability distribution of each data attribute in a parallel coordinates plot before and after filtering as annotated line charts on top of their parallel coordinates. Although the systems extending the parallel coordinated plots provided a means to easily explore multivariate data, other works have focused more on supporting analysis and decision making through the integration of interactive models. (Afzal *et al.*, 2011) developed a decision support environment to evaluate disease control strategies by predicting the course of an outbreak and analyzing the response measures. The severity of the epidemic is visualized by different color intensities on the map, and a custom split timeline is used to show the solution path. (Konev *et al.*, 2014) proposed an automatic simulation-based approach for flood management. The decision trees are automatically generated and visualized by clustered timelines. (Rinner, 2007) developed a geographic visualization system to support multi-criteria decision making. An index rank for each tract is calculated, and users can explore attributes through a linked parallel coordinate plot. Similar to the work of (Rinner, 2007), (Cassol *et al.*, 2017) proposed a framework to explore the optimal evacuation plan for crowd egress based on multiple factors, which were taken as input by the proposed metric to calculate the optimal plan. In both systems, interactive optimization methods are not fully considered. These systems support multi-criteria analysis through interactions with a parallel coordinate plot and quality indices (similar to our use of median ranking). However, portfolio comparisons and interaction with the optimization results are limited.

Other major issues underlying such decision support systems are the mechanisms used to compare across candidate solutions. The work by (Gleicher, 2018) summarized the basic designs of comparison into three categories, juxtaposition—i.e., which places the compared items are in different screen spaces, superposition—i.e, which places the compared items fit into the same screen space, and explicit encoding—i.e., visualization of the relationship between the compared items. (Kehrer *et al.*, 2013) and (Munzner *et al.*, 2003) utilized juxtaposition design for their comparisons of bar charts, lists, and trees. (Dasgupta *et al.*, 2015) combined juxtaposition and superposition for climate model comparison. (Law *et al.*, 2018) developed Duet, a visual analytics system for pairwise comparison integrating all three categories. Duet uses visualizations and textual descriptions to explain the recommended object groups which are similar to, or different from, the user-specified object with a focus on the similarity and difference. (Weng *et al.*, 2018a) proposed a spatial ranking visualization technique to explore and analyze ranking datasets and annotate the cause of the ranking with spatial context, which involves the three design categories of comparison.

From the optimization perspective, multi-criteria analysis and modeling have been integrated in a number of visual analytics systems for domains ranging from epidemiology to emergency response. However, little work in the visual analytics community has focused on conservation planning. Conservation planning requires the integration of optimization algorithms for conservation portfolios given the myriad of possible parcel configurations available. These conservation portfolios must allocate resources efficiently while considering current and future threats and their influence on the biodiversity assets. The problem of designing natural reserves has received considerable attention since the 1980s (Kirkpatrick, 1983; Cocks and Baird, 1989), mostly through the use of exact optimization models (Ando *et al.*, 1998; Church *et al.*, 1996; Polasky *et al.*, 2001; Sefair *et al.*, 2017; Acevedo *et al.*, 2015) and heuristic approaches (Pressey *et al.*, 1997; Arthur *et al.*, 1997; Margules *et al.*, 1988). The use of Operations Research techniques in this area have become more prevalent in recent years, including deterministic and stochastic approaches (see (Moilanen *et al.*, 2009) for a comprehensive review). Moreover, these methodological efforts have evolved into free software designed to support conservation planning decision-making processes (e.g., Zonation and MARXAN) (Lehtomäki and Moilanen, 2013; Ball *et al.*, 2009). Although available tools cover several pressing issues in conservation problems, some contemporary challenges are still unsolved. Some of the existing approaches focus on cost-minimization subject to ecological outcomes, ignoring the more realistic *dual* problems of maximizing such outcomes subject to a given budget. Approaches that optimize the ecological performance of the conservation portfolio approximate the quality of candidate patches by species representation (i.e., whether a species is present in a patch) and other single static patch attributes (Toregas and Revelle, 1973; Underhill, 1994; Williams and ReVelle, 1996; Camm *et al.*, 2002; Moilanen *et al.*, 2009), ignoring the multiobjective nature of the conservation decisions.

Other works focus on desirable geographical properties of protected areas such as landscape connectivity (Önal and Briers, 2006; Dilkina and Gomes, 2010; Dissanayake *et al.*, 2012; Jafari and Hearne, 2013) and compactness (Önal and Briers, 2002; Nalle *et al.*, 2002; Dissanayake *et al.*, 2012; Jafari and Hearne, 2013) but ignore the subjacent ecological processes. The majority of the works studying connected and compact reserves are mixed-integer programming models that are difficult to solve for realistic-size instances and that provide a single solution (i.e., a single connected and compact set of parcels to purchase). Without the visual support, analysts cannot easily modify an existing solution to incorporate expert knowledge and other attributes not included in the optimization model. Although optimization models in conservation planning are difficult to solve (Margules and Pressey, 2000), they are a fundamental part in the conservation decision process. However, ignoring other equally important components such as the interaction with experts for the inclusion of non-quantifiable or other aspects that are hard to express as constraints or objectives may reduce their applicability in real life conservation decisions.

2.3 Visual Analytics Framework

This section describes the design process and components of the proposed framework. The design of the proposed system is the result of a collaborative effort with a variety of stakeholders including donors, ecologists, and conservation plan-

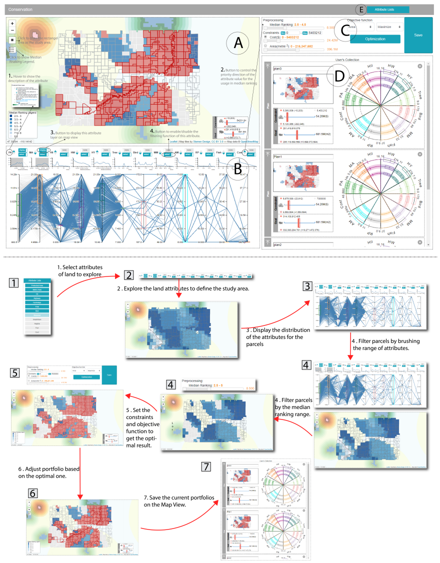

A
B
C
D
E
1
1. Select attributes of land to explore
2
2 . Explore the land attributes to define the study area.
3
3 . Display the distribution of the attributes for the parcels
4
4 . Filter parcels by brushing the range of attributes.
4
4 . Filter parcels by the median ranking range.
5
5 . Set the constraints and objective function to get the optimal result.
6 . Adjust portfolio based on the optimal one.
6
7. Save the current portfolios on the Map View.
7

Figure 2.1: A Visual Analytics Framework for Conservation Planning. (A) Map View: Visualizes Parcels Falling in the Analyst-defined Search Area. Colors in the Map Reflect the Median Ranking, Attributes, or Optimization Results. Users Can Create a Customized Portfolio with the Draw, Pan, and Zoom Controls in the Upper Left Corner. The Legend and the Differences of the Constraints and Goals Reached by the Portfolios Are Shown Separately in the Left and Right Bottom Corner. (B) Attribute Analysis View: Attribute Distributions Are Visualized by Line Charts and Parallel Coordinates Integrated with a Box Plot. (C) Optimization Configuration: The Median Ranking Slider Filters out Low-quality Parcels and Reduce the Optimization Algorithm Run-time. The Radio and Text Box Are Used to Input the Optimization Constraints and Objective Function. Users Click the 'optimization' Button to Run the Algorithm and 'save' the Current Portfolios on the Map to the 'user's Collection' for Comparisons. (D) Portfolio Comparisons: The Generated Portfolios Are Saved, and a Screenshot of the Selected Portfolios and Attributes Are Visualized. (E) Attribute Lists: Drop-down Menu for Selection. Event Sequence: 1) Select the Land Attributes to Explore. 2) Explore the Land Attributes to Define the Study Area. 3) Display the Distribution of the Attribute for the Parcels. 4) Filter Parcels by Brushing the Attributes Range and the Median Ranking Range. 5) Set the Constraints and Objective Function to Get the Optimal Result. 6) Adjust Portfolio Based on the Optimal Result. 7) Save the Current Portfolio for Future Comparisons.

ners. Through discussions and planning with domain experts, we identified key data needs, tasks, and design requirements. The proposed framework avoids the manual processing of the attributes of each candidate parcel to determine its relative convenience with respect to other parcels in the area of interest. It also consolidates the data processing, visualization, and optimization processes into a single intuitive tool.

The interaction with potential users resulted in the following functionalities of our framework.

- Data Storage and Downscaling: Stores map data for conservation planning, including biological, physical or socio-economic attributes. Currently includes 12 attributes suggested by conservation planners, and is scalable to further attributes. The data set is categorized into land use, physical-geospatial, and biodiversity layers. Input data is downscaled to the parcel scale to facilitate the calculation of the quality of land portfolios.

- Multi-layer Map View: Allows the investigation of parcel attribute values and the visualization of one of more attributes over a common area of analysis.

- Attribute Selection View: Filters parcels whose attributes fall within a certain range of interest. Provides the distributions of attribute values in any selected search area and allows the user to turn on/off each attribute layer on the map, define the ranking order of each attribute (e.g., higher values are preferred), and filter parcels based on a ranking aggregation metric calculated using selected attributes.

- Conservation Portfolio Optimization: Allows the specification of requirements for the land purchase portfolio, such as area of interest on the map, desired criteria for candidate parcels, objective to optimize, constraints, and maximum budget. Embeds a multicriteria optimization functionality to automatically provide land purchase recommendations and allow the user to visually interact with a solution to induce other desirable performance metrics (e.g., landscape connectivity and compactness).

- Porfolio Comparison View: Provides comparison tools to help portfolio managers explore their criteria of interest, compare land purchase portfolios, and work together to realize their final solution space.

We build upon previous works on multicriteria analysis and visualization, integrating geographic visualization and optimization to recommend land portfolios. Our system is designed to support the comparison of candidate land portfolios generated between the optimization recommendation and the analyst adjustments. Similar to previous work, our system uses a color code to visually inform the analyst on the quality of patches and land portfolios, linking attribute analysis and filtering to a parallel coordinates plot. Instead of displaying a sequence of portfolios to illustrate the impact of parameter changes, our system provides a unique visualization method to help comparing the attributes and their differences between various candidate portfolios. Our target users are conservation planning decision-makers in the broad sense. This could be an analyst assessing the ecological benefits of land patches, an ecologist surveying alternatives to expand current reserves, or NGOs and government agencies deciding which patches of land to restore or purchase. Figure 2.1 shows a snapshot of our system and its features, which is freely available at (Zhang *et al.*, 2021a). We have deployed this system to conservation planners, and our use cases demonstrate the effectiveness of optimizing their decision process given limited resources. A step-by-step demonstration video is available at (Zhang *et al.*, 2021a). The final product functionalities are explained in detail in Sections 2.3.1–2.3.5.

2.3.1 Data Storage and Downscaling

Typical data for conservation planning is characterized by biological, physical or socio-economic attributes. Our framework includes 12 common attributes and is scalable to additional data. We use the state of Montana as an example to describe the properties of a typical dataset for this system and the data downscaling steps. Table 2.1 describes the used datasets and their attributes that were chosen by conservation experts.

Our system supports a wide variety of shapefiles, geotiffs, open street map layers, among other types, including conservation portfolios built in other tools (e.g., MARXAN) as long as they are compatible with the shapefiles in our system. We note that our system is flexible to any geographical data, where users only need to select a base layer for analysis. Typically, the parcel layer would be used for this purpose, as this is the level at which land can be purchased. Once the base spatial unit is chosen, attributes are aggregated or dis-aggregated through a downscaling step to conform to the level of spatial granularity under analysis. For each data category, we use different processing rules to derive the corresponding attribute(s).

Other than COST, which is directly provided in the parcel shapefile dataset, we downscale the remaining datasets to calculate the parcel-level attributes. We calculate the distances to the existing protected areas, metro area, highway and hydrology areas, and aggregate the HII and other biodiversity attributes. Some conservation attributes measure the distance from a parcel to a feature of interest. In our dataset, examples include PA, MA, HW, and HY, which require us to calculate the distance from the parcel to the areas described in the attribute datasets. We first discretize each dataset into 30×30 m^2 *patches* (a request from our conservation planning partners), which will be later used to calculate the attributes of the larger-sized *parcels*. Parcels can be different in shape and size, and there are a variety of geographic aggregation methods that can be employed to calculate their attributes out of the patch attributes (Unwin, 1996). Then, we calculate the distance from the center of a patch to its nearest feature of interest. From there, we can aggregate all patches that fall within a parcel using min, max, average, or other aggregation functions. In our system, we use the average value of all patches within a parcel. Other attributes focus

Table 2.1: Variables and Data Sources Used

Category	Attribute Name (Abbr.)	Explanation	Data Source
Land Use Layer	Distance to protected area (PA)	The average distance of a parcel to its nearest protected area(s)	The shapefile of protected area by state from the USGS (USGS, 2018b)
	Distance to a metro area (MA)	The average distance of a parcel to its nearest metro area(s)	The shapefile of the 129 incorporated cities and towns in Montanan from Montana.gov (State of Montana, 2018)
	Distance to highway (HW)	The average distance of a parcel to its nearest highway(s)	The shapefile of primary and secondary roads by state from Census.gov (USCB, 2017)
	Human influence index (HII)	HII values range from 0 to 64 and measure the direct human influence on terrestrial ecosystems (Sanderson *et al.*, 2003). The average of the HII values within a parcel.	The shapefile of HII by North America from the socioeconomix data and applications center of NASA (SEDAC, 2018)
	Cost per square meter (COST)	total cost per square meter	The shapefile of parcels from Montana (Loveland Tech., 2018).
Physical Geospatial Layer	Distance to hydrology area (HY)	The minimum distance from the center of the parcel to the nearest hydrology area	The shapefile of hydrology area by state from the USGS (USGS, 2018a).
Biodiversity Layers	Richness of trees (TREE)	The total species richness of trees in the parcel.	The TIF file of richness of trees by state from BiodiversityMapping.org (Jenkins, 2017).
	Richness of birds (BIRD)	The total species richness of birds in a parcel.	The TIF file of richness of birds by state from BiodiversityMapping.org (Jenkins, 2017).
	Richness of fishes (FISH)	The total species richness of fishes in a parcel.	The shapefile of richness of fishes by state from BiodiversityMapping.org (Jenkins, 2017).
	Richness of amphibians (AM)	The total species richness of amphibians in a parcel.	The shapefile of richness of amphibians by state from BiodiversityMapping.org (Jenkins, 2017).
	Richness of mammals (MM)	The total species richness of mammals in a parcel.	The TIF file of richness of mammals by state from BiodiversityMapping.org (Jenkins, 2017).
	Richness of reptiles (RP)	The total species richness of reptiles in a parcel.	The TIF file of richness of reptiles by state from BiodiversityMapping.org (Jenkins, 2017).

on measurements and estimates from sensors, reports, and other sources. Examples of these attributes include TREE, BIRD, FISH, and other attributes in the biodiversity layer. We overlay the parcels onto the datasets and perform an aggregation operation to estimate the parcel attributes.

2.3.2 Multi-layer Map View

In order to support the multicriteria analysis during the decision-making process, we incorporated a multi-layer map view to visualize each attribute and their combinations over space. For distance-based attributes (PA, MA, HW, and HY), a sequential color scheme is used. The darker color means a patch is closer to the feature of interest. As an illustration, Figure 2.2.A shows the visualization of the distance to the metro area (MA) attribute. The pink region is the metro area, and the peripheral region around the metro area is colored based on the distance. The red and blue highlighted regions in Figure 2.2.A are the parcels in the user selected region of interest.

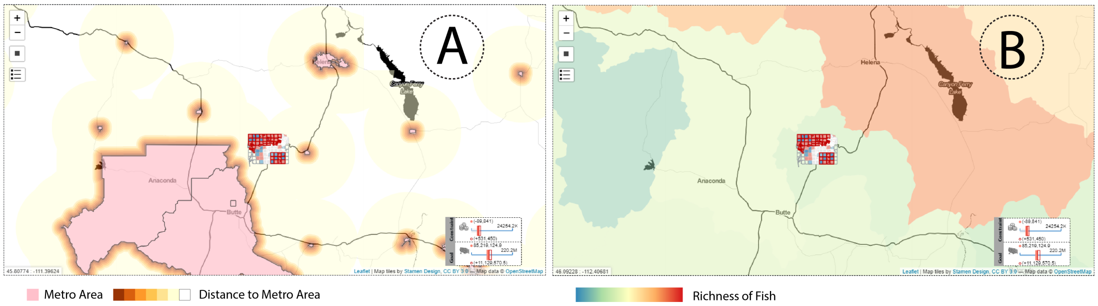


Figure 2.2: Examples of Multi-layer Map Views. (A) Visualization of Distance to Metro Areas. The Pink Area Is the Metro Area and the Peripheral Region Is Colored from Brown to Yellow Based on the Distance Value of Each Patch. The Red/Blue Highlights Correspond to the User Selected Region of Interest. (B) Visualization of Fish Species Richness. The Region with Bluer Color Has Lower Species Richness for Fish.

For region-based attributes (HII, COST, TREE, FISH, BIRD, AM, MM, and RP), the original datasets are overlaid on the map and colored based on their attribute values using diverging color schemes. The color scheme is designed to match the NASA analysis (SEDAC, 2018) and BiodiversityMapping.org (Jenkins, 2017). Figure 2.2.B shows the visualization of FISH. The region with redder color has higher species richness, while the region with bluer color has lower species richness for this variable. In the map view, the user can define their conservation area by drawing a rectangle on the map. Once the area is selected, the optimization algorithm suggests which parcels within this area to buy (the red/blue area seen in Figure 2.2). The parcels are colored based on an aggregation of the parcel attributes through a ranking function (see Section 2.3.4), filtering updates, the optimization algorithm's solution, and other user modifications. Results in the selection are influenced by the Attribute Selection View.

2.3.3 Attribute Selection View

The attribute selection view integrates parallel coordinates, line charts, and attribute controllers (see Figure 2.1.B). The user can explore value distributions of attributes in the search area, turn on/off each attribute layer on the map, define the ranking order of each attribute in the median ranking, and filter attributes for median ranking and based on the attribute value.

On each attribute controller, the user can click the top switch button (see Figure 2.1.B.3) to turn on/off the corresponding attribute layer, and mouse over the attribute name (see Figure 2.1.B.1) to see the explanation of the attribute and the color legend or the layer. The bottom switch button (see Figure 2.1.B.4) is used to enable/disable the filtering function of this attribute. The user can still explore the value distribution of an attribute when its filtering function is disabled, but interactions on disabled attributes won't impact the map view or the optimization model. The triangles pointing up and down (see Figure 2.1.B.2) are used to decide the priority direction of the attribute value when used in the median ranking aggregation (e.g., whether near or far proximity is desirable). For example, if the user wants to buy parcels near a protected area, then the priority direction is non-decreasing. That is, the user prioritizes low PA values by turning on the up triangle for the PA attribute. By turning on their down triangles, the user prioritizes high values of TREE, BIRD, FISH and AM attributes.

The line charts and parallel coordinates display the value distribution of each attribute and support parcel filtering by attribute value. Such filtering is only active when the attribute's filtering function is enabled. To explore attribute correlations and observe patterns of the data, the user can drag the axes of the parallel coordinates to change the order of the attributes. On each axis of the parallel coordinates, we add a box plot to help reveal the statistical distribution of the data. We use a categorical color scheme for the box plots to represent different attributes, and the attribute uses the same color in the portfolio comparison view, which we describe in detail in Section 2.3.5.

When the number of parcels increases, it is difficult to observe the distribution on the parallel coordinates due to visual clutters. Therefore, each attribute is also associated with a line chart where the x-axis represents the attribute value and the y-axis represents the frequency of the attribute value. The line chart is adjacent to each axis in the parallel coordinate plot and is used to show the value distribution of both the original data and the filtered data. To filter parcels, brush interaction is supported on the axes of the parallel coordinates as well as on the x-axis of the line chart. Parcels removed from the filtering will be grayed out on the parallel coordinates, while brushed parcels are highlighted in blue. On the top line chart, the black line shows the value distribution of all parcels in the search area, and, once filtered, a blue line is used to display the value distribution of the filtered parcels, and the original line will become gray.

In our system, all the interactions are coordinated with the map view. Once attributes are selected, the parcels in the user selected area will be colored based on their median ranking order. The legend for the median ranking results is in the left bottom of the map. The result of the median ranking depends on which parcels are selected and which filters have been applied to the data. The attributes of the parcels in the selected area are then used to generate a potential conservation portfolio.

2.3.4 Conservation Portfolio Optimization

Once the region and attributes are defined, our system employs a mathematical programming model to identify an optimal portfolio of patches for conservation. We define P as the set of candidate parcels eligible for purchase and A as the set of attributes of interest. We assume that all attributes are (or can be converted to) numerical values, and that all attributes are available for each parcel. We denote the value of attribute $j \in A$ for parcel $i \in P$ by a_{ij}.

Depending on the discretization of the area of analysis (chosen by the user), the number of candidate parcels may be very large. To reduce the computational effort in our system, we implement two pre-processing techniques. Both aim to reduce the set of candidate parcels by ignoring some that are not of interest for the decision-maker. The first technique is based on user-defined attribute filters. In this case, the user explicitly sets thresholds for a subset of the attributes, and the system discards those parcels with attributes violating the thresholds. Mathematically, we denote the set of attributes with threshold values as $\bar{A} \subseteq A$, and the corresponding lower and upper threshold values by $\bar{a}_j$ and $\underline{a}_j$ for attribute $j \in \bar{A}$, respectively. Using these values, the set of eligible parcels can be calculated as $\bar{P} = \{i \in P : \underline{a}_j \leq a_{ij} \leq \bar{a}_j, \forall j \in \bar{A}\}$. The $\bar{a}$- and $\underline{a}$-parameters are calculated via user interactions with the map and the attributes' value distribution.

Depending on the magnitude and meaning of an attribute, it may not be intuitive for the user to specify the $\bar{a}$- and $\underline{a}$-parameters. We also determine the set of eligible patches using a ranking-based procedure that describes the *relative* performance of a parcel with respect to other parcels. Parcels are sorted in non-decreasing order based on each attribute and then ranked such that $r_{ij} < r_{kj}$ if $a_{ij} < a_{kj}$, where $r_{ij} \in \{1, \ldots, |P|\}$ is the rank of candidate parcel $i \in P$ on attribute $j \in A$. In other words, the smaller the value of an attribute the higher the ranking of the parcel on that attribute (i.e., closer to 1). In the case where attributes with larger values are preferred (e.g., distance to human settlements), then the attribute values are sorted in non-increasing order and ranked such that $r_{ij} < r_{kj}$ if $a_{ij} > a_{kj}$. If two parcels have the same value on a particular attribute, then their ranking on that attribute is the same, i.e., $r_{ij} = r_{kj}$ if $a_{ij} = a_{kj}$. The ranking describes the parcel's relative performance on each attribute. We aggregate such rankings into a single number $\tilde{r}_i$ using the median value of the rankings across attributes. In other words, $\tilde{r}_i = \text{median}(r_{i,1}, \ldots, r_{i,|A|})$, $\forall i \in P$. We add the aggregated rank to the set of attributes for each parcel, allowing the user to specify more intuitive filters on the $\tilde{r}$-values. For instance, the user can choose to discard parcels that are not among the top k parcels–according to the median ranking–by setting the corresponding $\bar{a}$-parameter to k. We use median ranking aggregation because, among other properties, it eliminates the effect of extreme r-values and it can be computed efficiently (Sculley, 2007). Our system is flexible to accommodate any other ranking aggregation model. For a review on ranking aggregation, readers are referred to (Sculley, 2007), (Lin, 2010), and (Ailon *et al.*, 2008) and the references therein.

To find an optimal set of parcels for conservation, we use an integer programming model with variables x_i, where $x_i = 1$ if parcel i is recommended for purchase, and

$x_i = 0$ otherwise, $\forall i \in \bar{P}$. The model constraints represent conditions that a *portfolio* of parcels must satisfy, as opposed to the individual parcel conditions described in the pre-processing analysis. These include land purchase budget, minimum population area to protect, among others. We use linear constraints reflecting that the *aggregated* value of an attribute for the selected parcels must be less than (or greater than) or equal to a threshold value. We denote by $A^{\leq}$ and $A^{\geq}$, the set of attributes with a less than or equal to and greater than or equal to constraints, respectively. We use b_j as the threshold value for attribute $j \in A^{\leq} \cup A^{\geq}$. Note that not all attributes need to be included in such constraints, which means that $A^{\leq} \subseteq A$ and $A^{\geq} \subseteq A$. We pay special attention to the cost and area of each parcel, which we denote by c_i and α_i, $\forall i \in \bar{P}$, respectively. Our optimization problem maximizes the total purchase area (2.1a), subject to attribute constraints (2.1b)–(2.1c), and variable-type constraints (2.1d). The optimal purchased area will be a subset of the available area given that the purchasing cost will be part of $A^{\leq}$, with a corresponding b-parameter equal to the budget available for land purchases.

$$\max \quad \sum_{i \in \bar{P}} \alpha_i x_i \tag{2.1a}$$

$$\text{s.t.} \quad \sum_{i \in \bar{P}} a_{ij} x_i \leq b_j, \quad \forall j \in A^{\leq} \tag{2.1b}$$

$$\sum_{i \in \bar{P}} a_{ij} x_i \geq b_j, \quad \forall j \in A^{\geq} \tag{2.1c}$$

$$x_i \in \{0, 1\}, \quad \forall i \in \bar{P} \tag{2.1d}$$

An alternative model minimizes the total purchase cost, subject to constraints (2.1b)–(2.1d). In this case, the area will be part of $A^{\geq}$, with a corresponding b-parameter equal to a minimum required area to conserve. Mathematically, this problem can be written as min $\sum_{i \in \bar{P}} c_i x_i$, subject to (2.1b)–(2.1d). Although some of the constraints in our models may indirectly induce some landscape attributes (e.g., landscape connectivity by selecting the distance to existing protected areas as an attribute), the conservation portfolio produced by our models may not satisfy some those landscape requirements. This is because of the complexity and computing demand of enforcing such constraints for any arbitrary sized area selected by the user. Instead, our system allows the user to interactively modify an existing solution (through clicks on the map) to induce these landscape features. This allows the exploration of solutions that are infeasible for the optimization model, but that provide a good compromise between the ecological values gained and the extra cost required. The user is allowed to add or remove attribute constraints, as well as select the objective function to optimize (maximize the protected area or minimize the purchasing cost). Our models produce an optimal purchasing plan that satisfies all the selected attribute constraints at the same time. Using the optimal values of the decision variables, denoted by x_i^*, we define an optimal conservation portfolio as $P^* = \{i \in \bar{P} : x_i^* = 1\}$. These optimal portfolios are displayed for further user analysis.

The analyst interacts with the optimization model through the configuration view (see Figure 2.1.C). The analyst can filter parcels based on their median ranking and sets the constraints and objective function of the optimization model. The analyst can also “save” the current portfolio to the comparison view for further exploration and comparison. The median ranking slider shows the rankings of all selected parcels, and the analyst can drag the two ends of the slider to remove low-ranked or high-ranked parcels. The filtering tool changes the parcels used in the automatic optimization algorithm. The sliders under “constraints” are used to set the constraints for the optimization model. Currently, our system is able to answer the questions: *What is the largest total area that can be protected given a fixed budget and other ecological* *constraints?* and *What is the least-expensive set of parcels to protect with an area of at least b* km^2 *while satisfying other ecological constraints?* Therefore, the analyst needs to select one variable between "cost" and "area" to be the constraints, and leave the other to be the objective function. Our mathematical algorithm and system can support multiple constraints. For both "cost" and "area", the maximum value of the slider updates to represent the sum of the cost and area of user-selected parcels. Dragging the ends of the slider can change the value range we set for the constraint. To set the objective function, the analyst can choose either to maximize or minimize the variable. When the configuration is done, the analyst can click on the "Optimization" button to run the algorithm. For an easy comparison of multiple optimal portfolios under different right-hand-sides of the constraints, the constraint value from the previous run of the optimization algorithm is recorded in the slider.

2.3.5 Portfolio Comparison View

Multiple land purchasing portfolios may satisfy the planners' requirements under different attribute priorities. The analyst can make different modifications on top of the same suggested portfolio or change the selected parcels. Figure 2.3 shows our portfolio comparison view which uses a multiple portfolio visualization to display all saved portfolios. Each portfolio visualization has three visual components, the map screenshot, the optimization setting, and the attribute pie. The map screenshot represents the exact status of the map view when the portfolio is saved, and it records the details of the parcel selection in the portfolio. The optimization setting uses the same design as the lower right legend on the map to present the constraint and objective function for the optimization algorithm. The attribute pie is a glyph designed to visualize the attribute distribution of selected parcels under the setting of each portfolio and allow the analyst to compare their customized portfolio to that suggested

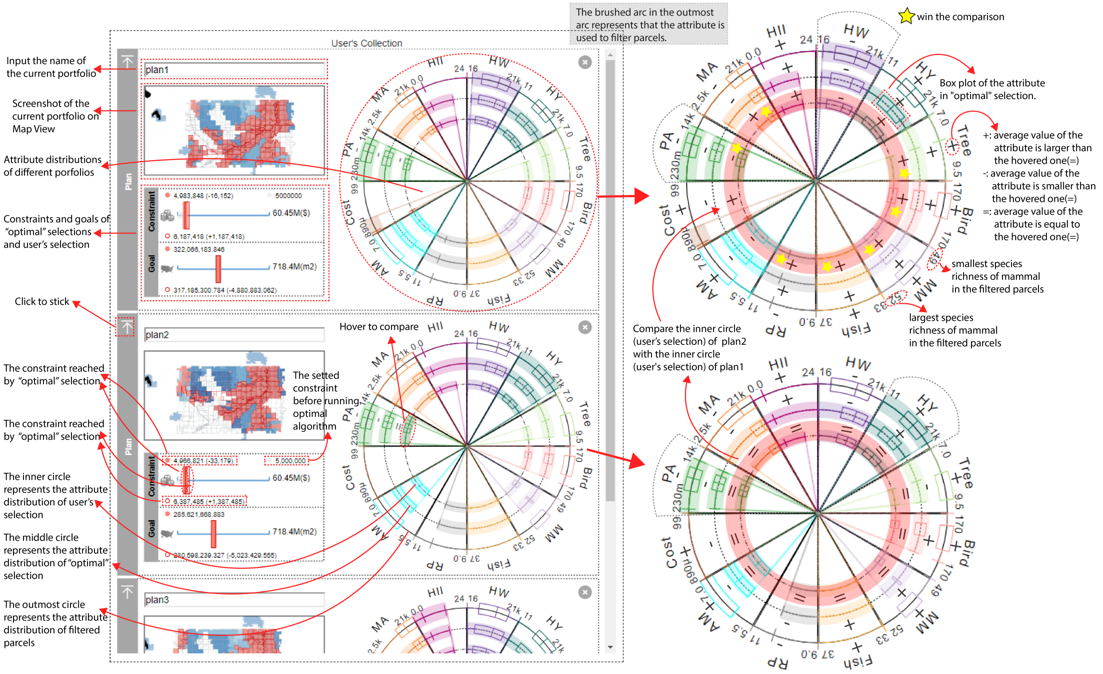

User's Collection
Input the name of the current portfolio
Screenshot of the current portfolio on Map View
Attribute distributions of different porfolios
Constraints and goals of "optimal" selections and user's selection
Click to stick
The constraint reached by "optimal" selection
The constraint reached by "optimal" selection
The inner circle represents the attribute distribution of user's selection
The middle circle represents the attribute distribution of "optimal" selection
The outmost circle represents the attribute distribution of filtered parcels
Hover to compare
The setted constraint before running optimal algorithm
The brushed arc in the outmost arc represents that the attribute is used to filter parcels.
win the comparison
Box plot of the attribute in "optimal" selection.
+: average value of the attribute is larger than the hovered one(=)
-: average value of the attribute is smaller than the hovered one(=)
=: average value of the attribute is equal to the hovered one(=)
smallest species richness of mammal in the filtered parcels
largest species richness of mammal in the filtered parcels
Compare the inner circle (user's selection) of plan2 with the inner circle (user's selection) of plan1
Plan
Constraint
Goal
plan1
plan2
plan3
4,983,848 (-16,152)
5000000
60.45M($)
6,187,418 (+1,187,418)
322,066,183.846
718.4M(m2)
317,185,300.784 (-4,880,883.062)
4,966,821 (-33,179)
5,000,000
6,387,485 (+1,387,485)
285,621,668.883
280,598,239.327 (-5,023,429.555)
HII
HW
HY
Tree
Bird
MM
Fish
RP
AM
Cost
PA
MA
24 16
21k 11
21k 7.0
9.5 170
170 49
52 33
37 9.0
11 5.5
7.0 8900
99 230m
14k 2.5k
21k 0.0

Figure 2.3: Comparison of the Average Value for Each Attribute. The Selected Attribute with a Gray Arc Is the Pa Attribute of the Analyst's Portfolio in the Second Row. All the Comparison Result of Other Portfolios with This Selected Attribute Are given with Three Relations, "less Than", "larger Than", and "equal To". The Corresponding Signals in the Comparison View Are a minus Sign, a plus Sign, and an Equal Sign, Respectively. For Example, the Average Values of Pa for an Analyst's Selected Set of Parcels in the First Row Are Smaller than That of the Analyst's Selected Set of Parcels in the Second Row. Comparing Analyst's Portfolios of Plan1 and Plan2, Plan1 Can Buy 314 Million Square Meters Area Better than the 280 Million Square Meters in Plan2 with a Similar Cost. In Addition, the 7 Attributes out of 12 Attributes in Plan1 Perform Better than Those in Plan2. In General, Plan1 Is Much Better than Plan2.

by the optimization model. The pie shows all attributes with evenly split sectors and each attribute is assigned one color. This is because even if not all the attributes are used to filter parcels, their value distribution may need to be considered in the final decision-making process.

To compare the influence of an attribute value in the portfolio, three circles with different radius are used. The outermost circle represents all the parcels in the search area, the middle circle represents the parcels suggested by the optimization model, and the inner circle represents the analyst selected parcels, which are those finalized in the portfolio. The three circles are arranged into the same value scale, which ranges from the minimum to the maximum of all parcels in the search area (the range of the outermost circle). For the middle and the innermost circle, the value range is also emphasized with a brushed color arc. Within this color arc, a box plot visualizes the attribute value's statistical distribution. We use a brushed color arc on the outermost circle to indicate the attribute value range that was used by the analyst to filter the attribute. If there is no such colored arc, it means the analyst did not filter on this attribute. The box plot on the outermost circle shows the quartiles of the attribute value with these filtered parcels. By using these glyphs, analysts can compare the attribute distribution of filtered parcels, suggested parcels and the user selected parcels to explore how the choice of parcels affects the attribute distribution. Analysts can also directly compare different portfolios, allowing multiple analysts to provide input and serving as a mechanism for both provenance and analysis. The map screenshots of the portfolios provide an overview of the differences between search areas and parcel selection. A black vertical line across all saved portfolio appears when the analyst mouses over the optimal setting view so that the analyst can easily compare the value of the constraint and objective function (cost and area) for these portfolios. The analyst can also compare the attribute value distribution of different portfolios by mousing over one arc of an attribute to turn on the comparison signs of this attribute for all portfolios. In this case, the average attribute values are compared both between the parcels represented of the three circles within one portfolio and also between the parcel selections represented by the circles of other portfolios. The reference circle arc is colored gray. If the average value equals the reference value, it shows an $=$ sign. When the average value of the parcels represented by the circle is larger than the reference value, a $+$ sign will appear, and when the value is smaller, a $-$ sign will appear.

2.4 Case Studies

In this section, we illustrate the use of our tool for the selection of conservation areas in Montana, USA. The state of Montana has a long wildlife conservation tradition dating back to 1895 when the Game and Fish Commission was established (Brownell, 1987). The evolution of wildlife legislation in this state reflects a serious commitment to the protection of wildlife; yet, less than 3.7% of its total area is designated as a wilderness protection area. Furthermore, most of the currently designated protected areas are composed of isolated mountain ranges clustered in a limited number of counties. Therefore, there is a need to complement existing protected areas by establishing new protection zones in counties that have limited designated wilderness areas and establishing corridors that facilitate movement and gene flow among wildlife populations living in isolated conservation areas (Hodgson *et al.*, 2009).

2.4.1 Multi-species Conservation Scenarios for the Judith Gap in Montana

In this case study, a conservation planner (the “analyst” hereafter) selects a set of areas to acquire (or restore) near the Judith Gap in Wheatland County. This gap represents a region of unprotected land between protected areas in the Little Belt Mountains in the west and the Big Snowy Mountains in the east. The analyst’s overall goal is to identify the largest possible total area to purchase subject to a budget constraint, while at the same time maximizing the number of terrestrial vertebrate species under protection within the corridor. There is evidence showing that in many instances the negative effects of human populations on protected areas decreases with distance to population centers (Mcdonald *et al.*, 2009). In this case, our system allows the analyst to visually explore a variety of attributes related to human use. Using the distance to metro area (MA) layer, as shown in Fig. 2.4b, it is possible to assess the spatial relationship between existing protected areas and urbanized centers. The figure shows how protected areas are generally distant from major urban centers. The highway layer (HW), as shown in Fig. 2.4c, illustrates how major roads may influence accessibility to protected areas. The figure shows how highway 91 is located between the two major conservation areas that the analyst seeks to connect. Alternatively, the analyst can visualize a human influence index (HII), as shown in Fig. 2.4d, which summarizes in a scale from 0–64 the overall influence of humans on terrestrial ecosystems. This view shows that areas near the metro and highway areas usually have high human influence index.

Meeting cost constrains is a central goal of conservation planning because resources for conservation are always limited (Naidoo *et al.*, 2006). Fig. 2.4e shows the spatial distribution of costs and its relationship with existing protected areas or other attributes. The cost layer shows that the average cost to purchase land near the Big Snowy Mountains is higher than that near the Little Belt Mountains. After the exploratory spatial analysis of existing protected areas, human influence, and cost, the analyst can define a candidate region between the Big Snowy Mountains and Little Belt Mountains conservation areas in Fig. 2.4f using the drawing tool.

Protecting sites that are closer to existing conservation areas (both east and west) will encourage connectivity. Therefore, the analyst selected PA as an attribute for the ranking calculation and the optimization model, as well as terrestrial vertebrate species richness which includes mammals, birds, reptiles, and amphibians because the overall goal is to promote movement and gene flow of wildlife species among existing conservation areas. Mammal conservation is a regional conservation priority, thus, using the brushed axis the analyst imposed a constraint to include sites that have a total richness of mammals index of at least 54 species. The model's goal is to maximize the area under protection while adding a budget as a constraint as it is

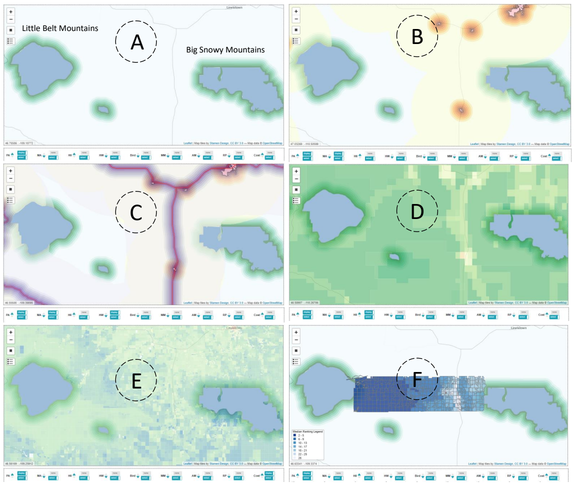


Figure 2.4: Process to Define the Candidate Region: (A) PA Layer. (B) MA Layer Covering PA Layer. (C) HW Layer Covering MA and PA Layers. (D) HII Layer Covering PA Layer. (E) Cost Layer Covering PA Layer. (F) Selected Candidate Region with Median Ranking Covering PA Layer.

a common practice in conservation planning (Cabeza and Moilanen, 2001; Williams *et al.*, 2005). Acquiring the whole candidate region would cost \$46M, which is higher than conservation budgets in many instances. Therefore, the analyst sets a target total cost of \$0 — \$10,000,000 to test if this budget range allows to meet the conservation goal of acquiring land to connect the conservation areas. The prescribed solution is shown in Figure 2.5a. The figure shows that the current budget allows purchasing a limited number of isolated patches that will contribute little to the overall goal of promoting connectivity. The budget is increased to \$15,000,000, obtaining the area in Figure 2.5b, which better promotes connectivity between the two existing protected areas. This budget level also allows connecting the southern portion of the Little Belt Mountains.

As is common in conservation planning, the prescribed optimal set requires manual refinement by the analyst to incorporate expert opinion on attributes that are not necessarily accounted for in the optimization model. For example, a land parcel may be already zoned for other uses, it may be prone to fire or flood disturbance, or it may

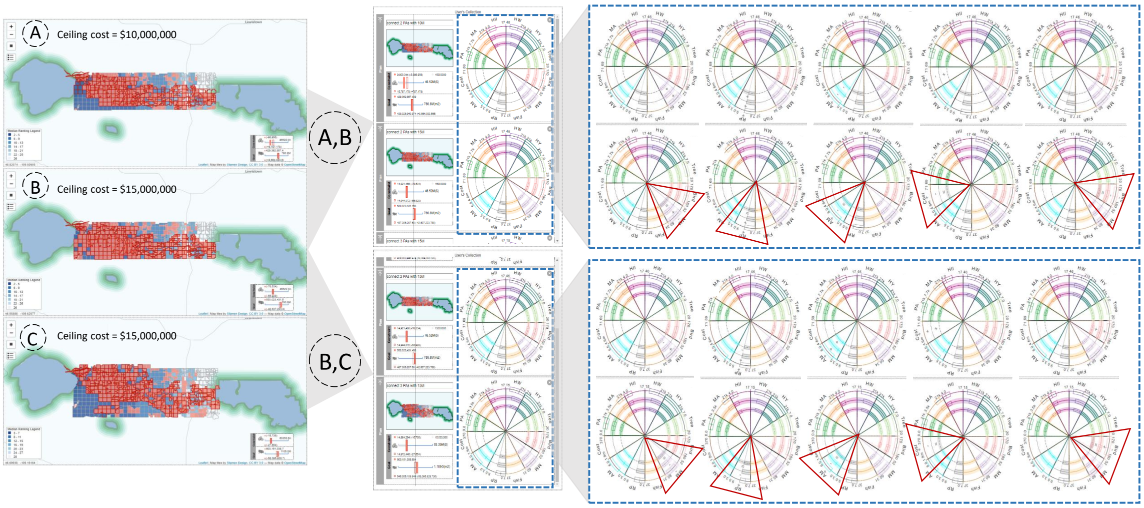

A
Ceiling cost = $10,000,000
B
Ceiling cost = $15,000,000
C
Ceiling cost = $15,000,000
A,B
B,C

Figure 2.5: Generated Portfolios on the Map. (A) Portfolios Generated with \$10m Budget Connecting Two Protected Areas. (B) Portfolios Generated with \$15m Budget Connecting Two Protected Areas. (C) Portfolios Generated with \$15m Budget Connecting Three Protected Areas. All the Parcels Are Prepossessed by Filtering Those with Less than 54 Mammal Species Richness. Portfolio Comparisons: (A,B) the Portfolios Generated Based on the Small Search Area with the Different Budgets Are Compared. (B,C) with the Same Budget, the Portfolios Generated Based on the Small Search Area and Larger Search Area Are Compared.

be spatially isolated and therefore not desirable as a conservation unit. This manual refinement is a key component of the conservation planning process that is lacking in many computational applications and is intuitively incorporated in this tool given its spatial nature. In these examples, the analyst replaced isolated regions with little contribution wildlife movement with areas in the west that ensure connection to the Big Snowy Mountains. Because the map reports the total selected area and cost after any analyst action (e.g., selection or removal of a parcel), the analyst was able to select an area within the given budget while using the manual refinement tool. The customized portfolio in Fig. 2.5c results in a set of areas to protect of $\sim$ 545 km^2 and a cost of \$14,972,446. Although this solution is not optimal (i.e., the optimal solution recommends the purchase of $\sim$ 603 km^2 within the same budget), it reflects the complementary insight of the mathematical model and expert judgment based on attributes such as landscape connectivity that are not included in the mathematical model.

We compare the three portfolios in the left part of Fig. 2.5 using the spatial and non-spatial information. The first two examples have the same search area and different budgets. The customized portfolio in the second example consists of a larger area (within its budget) than that in the first example, which exceeds its budget as shown in Fig. 2.5ab. The screenshot of the map shows how the parcels of each portfolio distribute. Besides the difference of constraints and goals reached by the portfolios, the change on the distribution of each attribute is visualized on the arcs in Fig. 2.5ab. We hover the inner arc to get the comparison result of the average attribute value among different portfolios. To show the hovered result of each attribute, we list five attributes we concern in the right part of Fig. 2.5ab. We observe that the customized portfolio in the second example has a higher average richness of amphibian species and the lower average cost. Moreover, it consists of a larger area within its budget to connect the two protected areas. The analyst decides that the second portfolio is better than the first one. With the same budget, we generate the third portfolio based on a larger search area. Fig. 2.5bc shows the comparison result of the second and the third portfolios. The third portfolio consists of a larger area within the same budget to connect three protected areas. In addition, it has a higher average richness of mammal species and reptile species, and lower cost. Based on our analysis from Fig. 2.5, the analyst selects the third portfolio as the final choice. The comparison result of the customized and suggested portfolios in the third example, which is represented near the inner and middle arc, gives more evidence to support the analyst's decision. In Fig. 2.5bc, the average richness of mammal, reptile and amphibian species is higher in the customized portfolio.

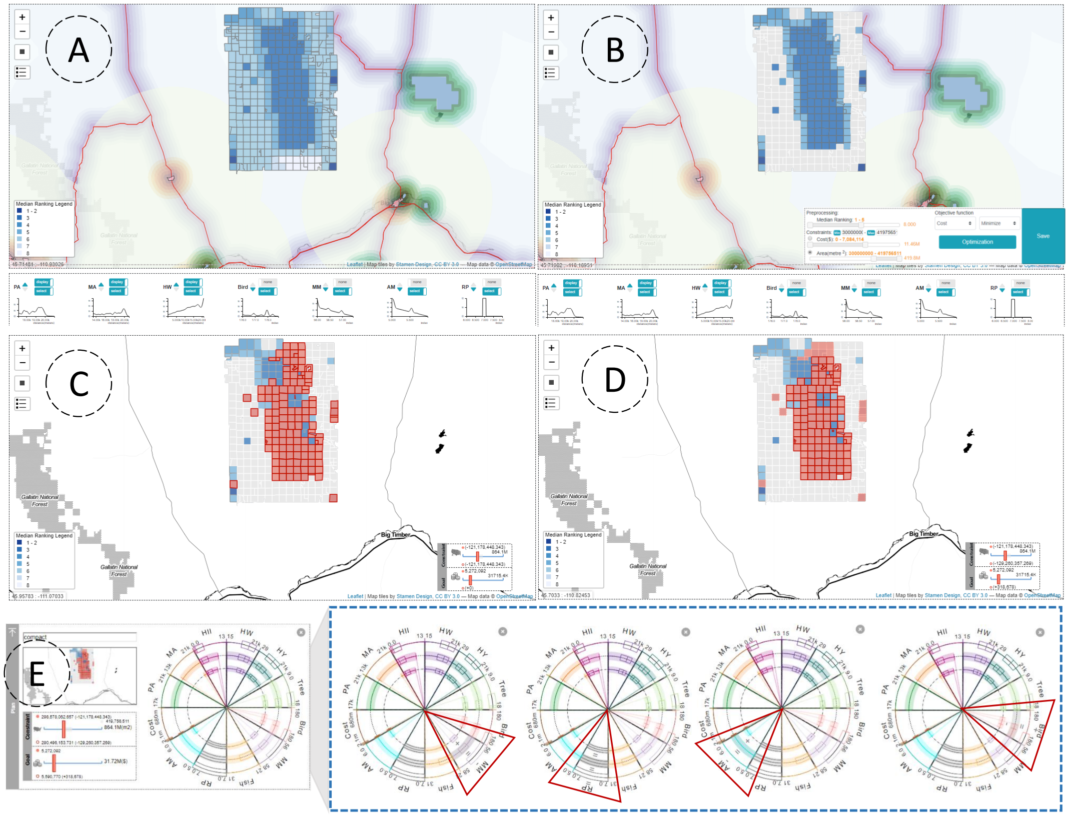


Figure 2.6: Case Study 2: (A) Initial Ranking given Selected Attributes in Montana's Park and Sweet Grass Counties. (B) Pre-processed Areas Using Median Ranking. (C) Optimal Results When Total Purchase Cost Is Minimized Subject to a Minimum Area of 300 Km^2; (D) Results after Manually Inducing Compactness in the Protected Area. (E) Comparison Between the Optimal Portfolio and the Connected and Compact Analyst-selected Area.

2.4.2 *Creating a Protected Area in Montana's Park and Sweet Grass Counties*

In this section, we illustrate the creation of a protected area at the boundary of Montana's Park and Sweet Grass counties, between highways 89, 90, 191, and 371. The region of interest consists of federal land and other unprotected areas and is within 100 mi from urban areas such as Bozeman, Livingston, Big Timber, and White Sulphur Springs, as well as other unincorporated communities. While in the first case study we were interested in designing a conservation area distant from areas of human influence, in this case study we have an opposite goal. Recent studies argue for a positive role of nature parks and protected areas close to human population centers (More *et al.*, 1988). Proximity to natural areas has been associated with improved mental health (Sturm and Cohen, 2014) and positive attitudes towards nature (Lin *et al.*, 2014). Therefore, in this case the analyst is interested in creating conservation areas that promote the protection of biodiversity, while at the same time being accessible by the community. In addition to the mammals, reptiles, amphibians, and bird richness layers, the analyst includes the distance to metropolitan areas and the distance to highway as attributes in non-decreasing order using the Attribute Analysis View. In this way, areas closer to highways and metro areas are given a higher preference.

Using the selected attributes, Fig. 2.6a shows the initial ranking of areas within the region of interest. This ranking combines both biological and geographical features. Because the cost of purchasing the whole region of interest is prohibitively high ($\sim$ \$32M), the analyst decided to exclude from the analysis such areas whose median ranking is larger than 5 for the selected attributes. In other words, discards those areas that are not ranked in the top five in at least half of the selected attributes. This was done using the pre-processing slider in the Optimization Configuration panel, which reduced the area from $\sim$864 km^2 to $\sim$420 km^2, with an updated total cost of $\sim$ \$11.5M (see Fig. 2.6b). The optimization model's goal is to minimize the total purchasing cost subject to a minimum protected area of 300 km^2. The results of this baseline scenario are shown in Fig. 2.6c. The size of the optimal area is $\sim$ 298 km^2 with a total cost of $\sim$ \$5.28M. This area is neither connected nor compact, having some isolated parcels and gaps inside the main cluster of selected areas. To improve the geographical properties of the selected area, the analyst manually induced these properties using the point-and-click feature of our system, ultimately producing the area shown in Fig. 2.6d. In this case, the size of the analyst-selected area is $\sim$ 290 km^2 with a total cost of $\sim$ \$5.6M. Regarding the ecological features, the attribute comparison in Fig. 2.6e shows that the analyst-selected landscape has a higher average richness for mammals, reptiles, and amphibians, but not birds. In this case, the increase in some species coverage as well as the connectivity and compactness properties of the resulting landscape are achieved at the expense of a higher land purchase cost with respect to the baseline scenario ($\sim$ \$320K).

2.5 Conclusions

In this chapter, we propose a visual analytics framework to help conservation planners and scientists to explore, compare, and modify conservation portfolios under a variety of constraints. To explore the candidate parcels, the system proposes the multi-layer map view and the parallel coordinates-based attribute analysis view. The suggested portfolios and the user-defined portfolios are generated based on an optimization model and users' domain knowledge. The comparison between these portfolios is supported by the portfolio comparison view. Using our system, analysts can incorporate their decision preferences and add selection attributes that are not easily incorporated as constraints or objectives, or that delay the construction of a portfolio given the resulting model complexity. The optimization model is fast for moderately sized landscapes and allows the construction of what-if scenarios almost in real time.

The framework has been validated by conservation experts through two case studies, which demonstrate how the framework can help analysts to generate conservation portfolios for different goals under a variety of constraints. Moreover, the system has been received design feedback from multiple conservation experts including two co-authors and four external partners. Although the feedback received was generally positive, some limitations have been identified for future work. Specifically, analysts appreciated the option to compare portfolios; however, more automation for supporting detailed comparison could improve the analysis process. The analysts also noted that while the framework is flexible to the underlying optimization approach, an API that would allow users to directly integrate their own optimization routines could greatly enhance their workflow. A possible avenue is to explore alternative multi-objective approaches to explore the trade-off between objectives in the portfolio optimization (see, e.g., (Miettinen, 2012) and (Sawaragi *et al.*, 1985) for alternatives). Further work will focus on the automatic comparison of candidate portfolios and add customized algorithms to induce other spatial properties to the framework in case the user decides to use them (e.g., connectivity and compactness). An interesting conjecture is whether adding human interaction with the optimization helps with the run-time issues when spatial properties are enforced. Further studies exploring the tradeoffs between human input and ability to explore reasonable solutions is an interesting future direction. As of now, the analyst can manually load a candidate conservation portfolio for further analysis using a shapefile or a file specifying whether a parcel is selected. We will add modifications in this aspect to facilitate the upload and compatibility check of a candidate portfolio, as this will allow our system to complement the analysis of other existing tools like MARXAN and Zonation. Although the framework focuses on conservation planning decisions, it can be extended to other spatial problems, including electoral districting, location of urban parks, and land-use planning. Such applications will require the proper data inputs and specification of the related optimization problems.

This work develops novel methods to support the exploration multiclass geospatial data. While the detection of the patterns and the relationship for the multiclass geospatial data needs methods to do. In the following chapter, I proposed the topological method to explore the multicalss map analysis.

Chapter 3

EXPLORING GEOGRAPHIC HOTSPOTS USING TOPOLOGICAL DATA ANALYSIS

3.1 Introduction

The spatial analysis of point events has been widely examined (Bailey and Gatrell, 1995; O'Sullivan and Unwin, 2010; Shiode, 2011; Chen *et al.*, 2018; Bizimana and Nduwayezu, 2020), and a variety of methods have been developed for detecting and visualizing hotspots (e.g. (Kulldorff, 1997; Krige, 1951; Borruso, 2008; Nakaya and Yano, 2010; Hu *et al.*, 2014; Zhang *et al.*, 2020)). A popular method for analyzing point data is to transform the data into a density function across a regular grid using functional approximations, for instance via Kernel Density Estimation (KDE) (Silver-

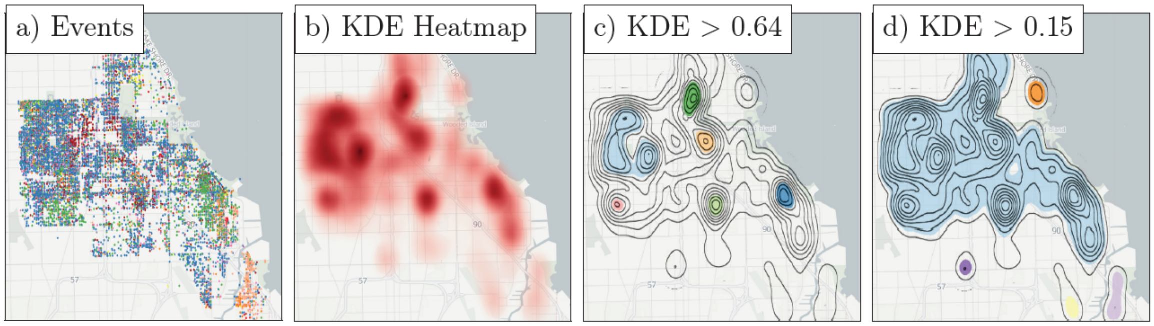


Figure 3.1: A Common Way to Characterize Hotspots in Point Events (A) Is to First Derive a Kernel Density Estimate (B), and Then to Extract All Areas That Exceed a Given Density Value Threshold (C-D). Here, Events Correspond to Reported Gang Activity in Chicago, IL, USA (Section 3.4.2), Where Each Gang Is Assigned a Random Color. As Demonstrated in the Bottom Row, the Chosen Threshold Has a Significant Impact on the Shape and Number of Extracted Hotspots (Colored Randomly). More Importantly, a Single Global Threshold Can Not Distinguish Between Hotspots at Different Scales, E.G., For the Threshold That Is Required to Detect All Small Hotspots Shown In (D), All Major Hotspots Shown in (C) Can No Longer Be Separated.

man, 1986). The results of the analysis produce an informative and visually appealing surface that communicates the intensity of point data. Because of their utility, density functions have been used to study a wide variety of phenomenon including: wildfires (Koutsias *et al.*, 2014), air traffic patterns (Lampe and Hauser, 2011), human activity (Hu *et al.*, 2014), animal movements (Sarkar *et al.*, 2015; Chirima and Owen-Smith, 2017), crime (Nakaya and Yano, 2010; Levine, 2008), urban analysis (Zhou, 2015), health (Liadsky and Ceh, 2017), etc. The most common approach is to identify hotspots as spatial regions that exceed a specified event density threshold (Bizimana and Nduwayezu, 2020; Hu *et al.*, 2014; Chainey *et al.*, 2002; Lukasczyk *et al.*, 2015; Chainey and Ratcliffe, 2013; Johansson *et al.*, 2015). While such approaches have proven valuable, the general approach of characterizing hotspots on density alone can obscure local peaks and unintentionally highlight noise and outliers. For example, Figure 3.1 shows a standard hotspot analysis, where point event data (Figure 3.1.a) is transformed into a probability estimate and rendered as a heatmap (Figure 3.1.b). To identify hotspots, an analyst-defined threshold is chosen and regions with values greater than the chosen threshold are extracted. Here, we observe some of the classical issues with thresholding. In Figure 3.1.c, selecting a reasonably high threshold results in several distinct geographic hotspots, such as the dark green hotspot and the light blue hotspot; however, regional hotspots that may be local maximas but are below the selected threshold may be missed. If the threshold value is lowered to try and capture these features, Figure 3.1.d, the missing local hotspots, such as the orange and purple regions, now appear. However, the unique peaks in the first pair now merge into a single large hotspot (light blue), obscuring local variations.

Given that a density estimate is an ordinary scalar field, current geographic hotspot exploration methods can be augmented based on the theory and application of scalar field topology (SFT). SFT provides various data abstractions that can be used for robust, hierarchical feature characterization, such as the persistence diagram (Edelsbrunner *et al.*, 2002), the merge tree (Carr *et al.*, 2003; Gueunet *et al.*, 2017), and the Morse complex (De Floriani *et al.*, 2015; Robins *et al.*, 2011; Gyulassy *et al.*, 2018). In the past, these abstractions have been successfully applied in a number of visualization and analysis tasks (Heine *et al.*, 2016); including astrophysics (Sousbie, 2011; Shivashankar *et al.*, 2016), biological imaging (Carr *et al.*, 2004; Bock *et al.*, 2018; Anderson *et al.*, 2018), chemistry (Bhatia *et al.*, 2018; Guenther *et al.*, 2014; Olejniczak *et al.*, 2019), fluid dynamics (Laney *et al.*, 2006; Kasten *et al.*, 2011; Bremer *et al.*, 2016), material sciences (Gyulassy *et al.*, 2015; Lukasczyk *et al.*, 2017a; Soler *et al.*, 2019), and turbulent combustion (Bremer *et al.*, 2011; Gyulassy *et al.*, 2014). However, the application of SFT theory for geographical hotspot analysis is not yet widely utilized.

To facilitate the application of SFT in this context, we demonstrate how the concepts of SFT can be used to interactively extract hotspots in density estimates derived from geo-spatial point events. We demonstrate that SFT-based hotspot characterizations make it possible to filter hotspots by significance, to identify hotspot boundaries, and to understand their hierarchical and spatial relationship. These advantages make SFT-based characterizations far superior to the common approach of simple thresholding (Figure 3.1). We also describe a novel characterization that combines Morse complex cells with superlevel set components and crown components. This hybrid characterization has the advantage that connected regions can be further subdivided into individual hotspots with natural boundaries that align with the density gradient. Finally, we instantiate our SFT-methodology in a visual analytics framework (Figure 3.2) that supports analysts in comparing different hotspot characterizations. We demonstrate the effectiveness of the framework via a detailed analysis of two crime datasets and discuss how this framework improves upon current geovisual analytics systems. The contributions of our paper are:

- The application of SFT in the area of hotspot characterization of geo-spatial point events;
- A robust and effective hotspot characterization that is based on a combination of merge tree crown components and Morse complex segmentations;
- A novel hotspot characterization that combines Morse complex cells with superlevel set components and crown components that allows connected hotspot components to be further subdivided into individual hotspots with natural boundaries that align with the density gradient; and
- A visual analytics framework that enables the effective exploration and comparison of SFT-based hotspot characterizations.

3.2 Related Work

Point events appear in many application domains, such as law enforcement (Johansson *et al.*, 2015; Malik *et al.*, 2014), zoology (Travaini *et al.*, 2007; Sarkar *et al.*, 2015), epidemiology (Lukasczyk *et al.*, 2015; Bizimana and Nduwayezu, 2020), food service (Zhang *et al.*, 2020), internet applications (Guan *et al.*, 2014) and environmental science (Bröring *et al.*, 2015). Researchers and practitioners are often interested in regions with a relatively high number of observed events. Such areas are commonly referred to as hotspots, but there is no universal definition of what a hotspot actually is. For instance, spatial scan statistics (Kulldorff, 1997) identify hotspots as subsets of the point events. Other techniques define hotspots based on a continuous representation of the point events, which can be derived via kernel density estimation (Silverman, 1986), Kriging (Krige, 1951; Chilès and Desassis, 2018; Oliver and

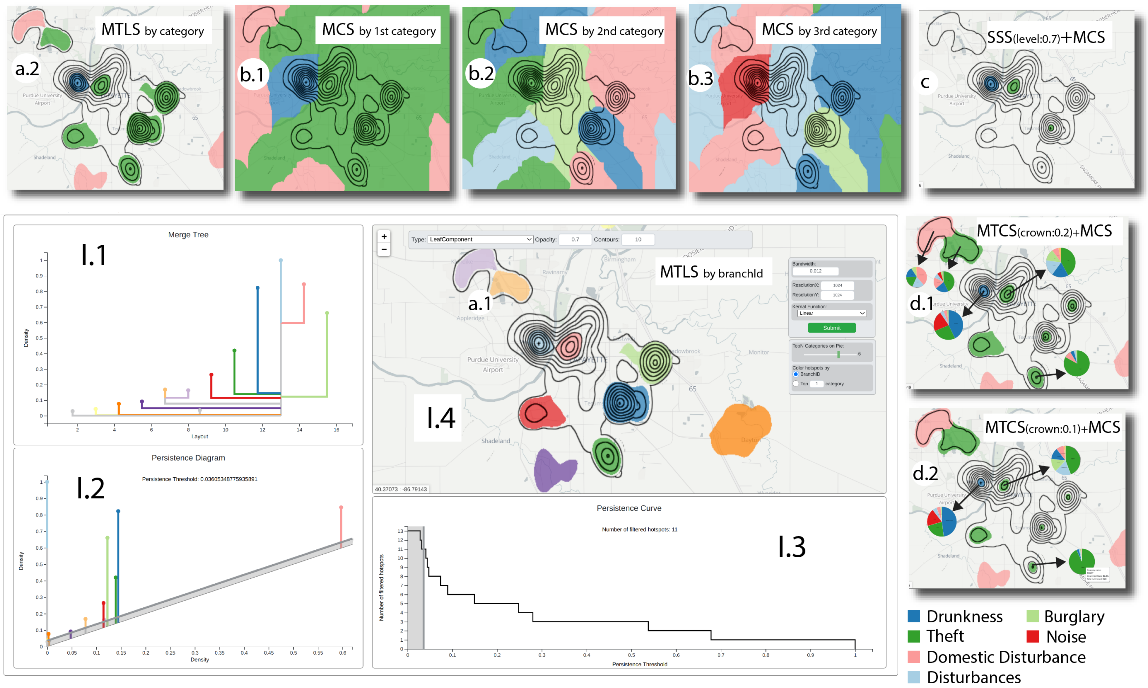


Figure 3.2: The Proposed Scalar Field Topology (SFT)-based Visual Analytics Interface Consists of Four Linked Views: (I.1) a Merge Tree, (I.2) a Persistence Diagram, (I.3) a Persistence Curve, and (I.4) a Geospatial Map. Given a Set of Geo-spatial Point Events—here, Crime Incident Reports in Tippecanoe County, IL, USA (Section 3.4.2)—the Framework Computes a 2D Kernel Density Estimate (Contours), and Then Analysts Can Interactively Extract Hotspots (Colored Elements) via Different SFT-based Characterizations (A.1-d.2). (A.1) and (A.2) Both Show Hotspots That Are Characterized via a Merge Tree Leaf Segmentation (MTLS), Where Hotspots of A.1 Are Colored Based on the Id of Their Corresponding Merge Tree Branch, and Hotspots of A.2 Are Colored Based on the Most Frequent (Majority) Crime Category in Each Hotspot. (B.1), (B.2), and (B.3) Show the Same Morse Complex Segmentations (MCS) Where Hotspots Are Colored by the Most Frequent, the Second Most Frequent, and the Third Most Frequent Crime Category, Respectively. (C) Shows a Combination of a Superlevel Set Segmentation (Sss) for Level 0.7 and a MCS, Where Hotspots Are Colored by the Majority Crime Type. (D.1) and (D.2) Show the Combination of a MCS with Two Different Merge Tree Crown Segmentations (MTCS), Where Hotspots Are Again Colored by the Majority Crime Type.

Webster, 1990), or other regression-based techniques. The resulting scalar fields are then visualized via heatmaps and iso-contours, where hotspots can be identified as local peaks (Chainey *et al.*, 2002; Lukasczyk *et al.*, 2015; Johansson *et al.*, 2015).

In this thesis, we focus on hotspot characterization of density estimates, where a hotspot is an area on a map that has a high number of events. All examples in this thesis utilize kernel density estimation for creating the density field, and there are numerous sample applications of KDE for geographic visual analytics. For example, Maciejewski et al. explore spatiotemporal changes in emergency department records using kernel density estimation (Maciejewski *et al.*, 2009). Scheepens et al. apply kernel density estimation to trajectory aggregation and use contour lines to help analysts predict the movement of ships (Scheepens *et al.*, 2014, 2012). Razip el al. present a mobile toolkit to help citizen and law enforcers assess risk levels in urban areas where risk was visualized as a density estimation (Razip *et al.*, 2014), and de Queiroz Neto et al. used Marching Squares to quickly generate high resolution hotspots (de Queiroz Neto *et al.*, 2016) and characterized tasks for hotspot analysis (de Queiroz Neto *et al.*, 2020). Hu et al. (Hu *et al.*, 2014) analyze human activity based on the mobility hotspots calculated by kernel density estimation. Zhang el al. (Zhang *et al.*, 2020) apply the kernel density estimation to restaurant POI data for food culture analysis. Nakaya and Yano (Nakaya and Yano, 2010) propose space-time kernel density estimation to analyze crime hotspots.

A common approach among such systems is to identify hotspots as the individual connected areas that exceed a given event density threshold (Chainey *et al.*, 2002; Lukasczyk *et al.*, 2015; Maciejewski *et al.*, 2009). This threshold, however, controls the shape and the number of hotspots, and even slight threshold variations can have a significant impact on analysis results (Lukasczyk *et al.*, 2017b). Appropriate thresholds are also often not known a priori, and need to be adjusted interactively (Chainey *et al.*, 2002; Malik *et al.*, 2014). Here, scalar field topology (SFT) (Edelsbrunner and Harer, 2010) can provide a family of generic, robust, and efficient feature characterizations, adding another suite of tools for geographical analysis. SFT theory has already been successfully applied in different application domains (Heine *et al.*, 2016); however, while SFT essentially provides a Swiss army knife for feature characterization, it is not clear which specific characterization should be used to describe hotspots. As noted by Heine et al. (Heine *et al.*, 2016), this is often application dependent and requires an SFT expert for fine-tuning. In this work, we examine the pros and cons of several SFT-based feature characterizations for the extraction of hotspots present inside density estimates, we provide a novel hotspot characterization that combines Morse-Smale complex cells with superlevel set and crown components, and develop a visual analytics framework that enables analysts that do not have a background in SFT to conveniently and efficiently explore and compare different characterizations.

3.3 Method

In this thesis, we discuss how SFT-based feature characterizations can be used to identify hotspots. We note that any density field on a regular grid can be explored using our proposed framework. Our examples all utilize kernel density estimation for approximating a density field from geographically reference crime points. For completeness, we first describe how to compute kernel density estimates to obtain a density map of geographically referenced point data. In the context of density maps, we then introduce the core concepts of SFT and we demonstrate that various feature definitions derived from SFT (superlevel set components, crown components, and Morse Complex cells) can be used to characterize hotspots. We showcase the advantages and disadvantages of each characterization and their ability to highlight different aspects of the inherent data. We also describe a novel characterization that combines Morse Complex cells with superlevel set components and crown components. This hybrid characterization has the advantage that connected regions can be further subdivided into individual hotspots with natural boundaries that align with the density gradient, and provides a novel means of visualizing hotspot boundaries and spatial extent. Finally, we discuss our visual analytics interface, which provides multiclass analytical capabilities for aggregated groups, and allows analysts to compare the SFT-profiles of hotspots in terms of the prevalence of event types.

3.3.1 Kernel Density Estimation (KDE)

In this work, we consider scalar fields that correspond to radial symmetrical kernel density estimates (Silverman, 1986; Gramacki, 2018). For a list of 2D point events $\bar{X} = [x_1, \ldots, x_n]$ with $x_i \in \mathbb{R}^2$, such an estimator $f_h : \mathbb{R}^2 \to \mathbb{R}$ is defined as

$$f_h(x) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{\|x - x_i\|_2}{h}\right), \tag{3.1}$$

where $K(u)$ is a univariate kernel function, and h is the kernel bandwidth. This estimator is usually evaluated on each point of a regular grid, where the kernel computes the contribution of each event based on its distance to the current grid point. The kernel bandwidth adjusts this contribution radius and has the most significant impact on the resulting estimate. Choosing appropriate bandwidths is itself a challenging research area (Heidenreich *et al.*, 2013). For the sake of simplicity, in this thesis we will focus on estimates derived with a linear kernel

$$K_L(u) = \begin{cases} 1 - |u| & \text{if } |u| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

and an initial bandwidth suggested by Silverman's *Rule of Thumb* (Silverman, 1986):

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} \sigma n^{-\frac{1}{5}}. \tag{3.3}$$

Here, σ is the standard deviation of the n samples. However, if necessary, the kernel and the bandwidth can be interactively adjusted in our visual analytics framework. Contour lines of an example kernel density estimate are shown in the top row of Figure 3.3. Note that any density field can be explored with the proposed framework, as well as ratio/risk maps, to normalize against a population of interest, as long as the data is captured on a regular grid.

3.3.2 Superlevel Set Segmentation (SSS)

One of the most common ways to identify areas of interest in a scalar field is to extract all regions that exceed a given scalar threshold. In SFT-terms, this threshold is referred to as a level, and every individual connected area that exceeds the level is called a superlevel set component (Figure 3.3, second and third row). However, even slightly changing the level can drastically change the geometry and the number of extracted components (Lukasczyk *et al.*, 2015; Bremer *et al.*, 2016; Lukasczyk *et al.*, 2017b). Superlevel set segmentations are also not a true multi-resolution feature characterization since components for different levels are not explicitly set in context with each other, i.e., components (hotspots) are shown with a single color, independent of the fact that they exhibit a nesting structure. Another limitation of these segmentations is the fact that if the density field exhibits peaks with a small event density, then, in order to extract these small density peaks, the threshold level has to be sufficiently small. For such small thresholds, hotspots of higher event density can no longer be separated. For instance, in the example of Figure 3.3, there does not exist a global level that would extract all four hotspots as individual superlevel set components, i.e., to extract the orange hotspot around maximum A, the level has to be smaller than 3, but for such level values, the red and the green hotspot already merged. SFT provides several feature characterizations that overcome these limitations.

3.3.3 Merge Tree Segmentation (MTS)

Much SFT theory is derived from observing the behavior of superlevel set components during a continuously level sweep. Consider a level that is greater than the largest density value. For this level no superlevel set components exist. However, if we now continuously decrease the level then new superlevel set components appear at local maxima (discs in Figure 3.3), and already existing components merge at so-called saddles (diamonds in Figure 3.3). These points are also called the critical points of the function, and the topological evolution of the components they induce is recorded in a topological abstraction called the merge tree (Carr *et al.*, 2003), whose leaf nodes correspond to maxima, intermediate nodes correspond to saddles, and edges indicate which components merge at which saddle (Figure 3.3, right). Every time branches merge at a saddle, then we define that the branch starting at the largest maximum continues and all other branches terminate at the saddle. This convention is called the *Elder Rule* (Edelsbrunner and Harer, 2010). Based on this rule, every maximum can be uniquely paired with a saddle (except the global maximum which is paired by convention with the global minimum). SFT actually requires that all critical points must have unique values in order to never encounter a tie when components merge at saddles. Although density scalar fields usually do not satisfy this criterion, it is always possible to trivially enforce this criterion symbolically with a method called *Simulation of Simplicity* (Edelsbrunner and Mücke, 1990). The resulting pairs are called persistence pairs, where the persistence (the significance) of a pair corresponds to the absolute scalar value difference of its two points. Performing this level sweep over the entire scalar range yields the merge tree segmentation (MTS), which maps every point of the domain to a merge tree edge and vice versa (Figure 3.3, fourth row).

The merge tree provides valuable insight about the structure of the scalar field. For example, the nodes of the tree indicate important level values for which superlevel set components will appear and merge. Furthermore, the number of edges that include a given level value correspond to the number of extracted superlevel set components at this level (Figure 3.3, second and third row). However, the primary advantage of the merge tree segmentation is that it partitions the domain into a hierarchical structure, which we will utilize in the following characterizations.

3.3.4 Merge Tree Leaf Segmentation (MTLS)

Given an MTS, it is trivial to extract all regions that correspond to the leaf arcs of the merge tree. The resulting domain segmentation is called the merge tree leaf segmentation (MTLS), which highlights the parts of the domain that are uniquely associated with a single maximum (Figure 3.3, fifth row). An advantage of the MTLS over the SSS is that it can identify hotspots independent of a level value, and the resulting hotspots can have different density ranges.

3.3.5 Merge Tree Crown Segmentation (MTCS)

Instead of using a global density threshold (as used for SSSs), it is also possible to extract hotspots based on local thresholds (Bremer *et al.*, 2016). We can extract subtrees of the merge tree that are rooted at maxima and span a specific scalar range. Each such subtree is called a crown, and the corresponding scalar range is called the crown height. In density estimates, crown components correspond to regions whose points have at most a certain density difference relative to their associated maximum, i.e., the hotspot center (Figure 3.3, sixth row).

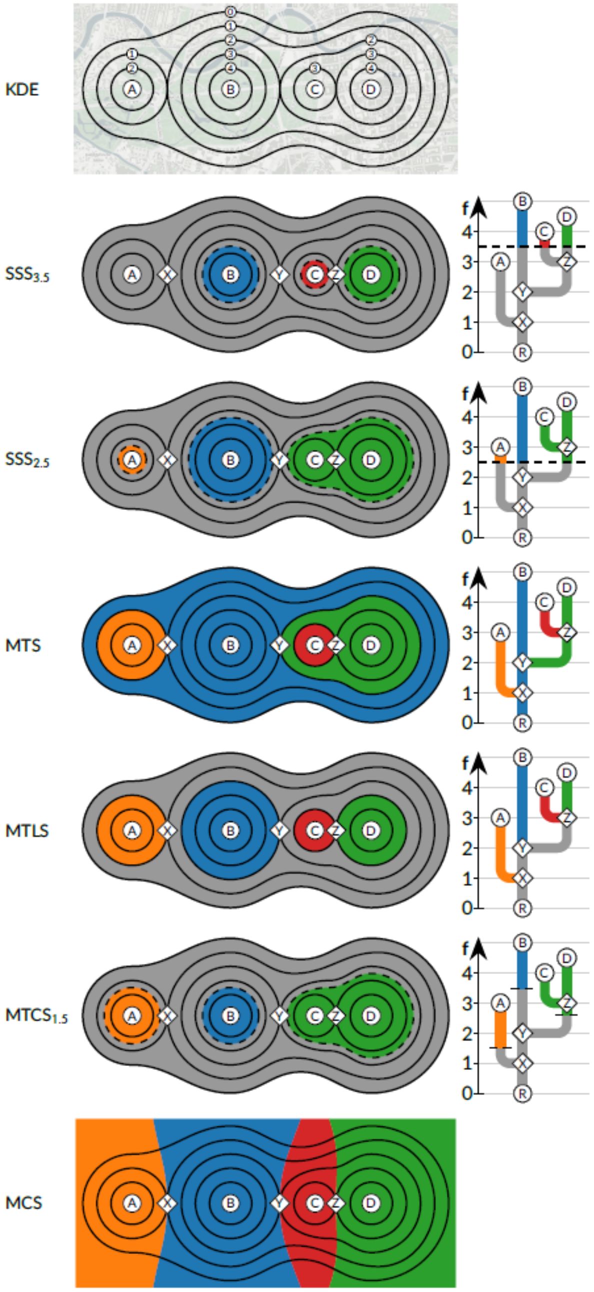


Figure 3.3: Hotspots of a Density Field Can Be Characterized via Several SFT-based Abstractions. Superlevel Set Segmentations (SSS) Identify Connected Areas That Exceed a given Level Threshold. The Merge Tree Segmentation (MTS) Associates Each Point of the Domain with a Branch of the Merge Tree. As a Special Case, the Merge Tree Leaf Segmentation (MTLS) Extracts Only Parts of the Domain That Are Associated with Leaf Arcs. Another Special Case Is the Merge Tree Crown Segmentation (MTCS) That Extracts Subtrees Based on a Relative Scalar Value Difference, Called Relevance. Finally, the Morse Complex Segmentation (MCS) Associates Each Point of the Domain with the Maximum That Is Reached by Following the Steepest Ascend.

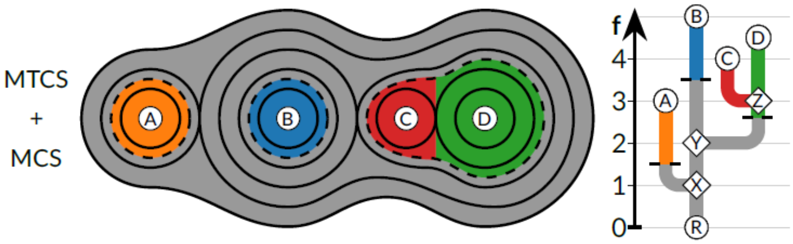


Figure 3.4: Combining the Merge Tree Crown Segmentation (MTCS) and the Morse Complex Segmentation (MCS) Yields a Very Versatile, Robust, and Effective Hotspot Characterization. Hotspot Boundaries Are Determined by the MTCS, and the MCS Is Used to Differentiate Between Individual Hotspots.

3.3.6 Morse Complex Segmentation (MCS)

The final SFT-based feature characterization we consider is based on the Morse complex (Milnor *et al.*, 1963; Forman, 2002). In contrast to the merge tree, the Morse complex is derived based on the gradient field of the density estimate. By always following the steepest ascent, every point of the domain can be uniquely mapped to a single maximum (this is again a result from the *Simulation of Simplicity* (Edelsbrunner and Mücke, 1990) procedure). Hence, the domain can be partitioned into parts whose points would all end up in the same maximum. In SFT terms, the resulting partition is called the ascending manifold, but we refer to this as the Morse complex segmentation (MCS) for simplicity, (Figure 3.3, seventh row). In the context of hotspot analysis, the boundaries between regions of the MCS can be interpreted as the most conservative hotspot boundaries, providing a novel means of visualizing hotspot boundaries.

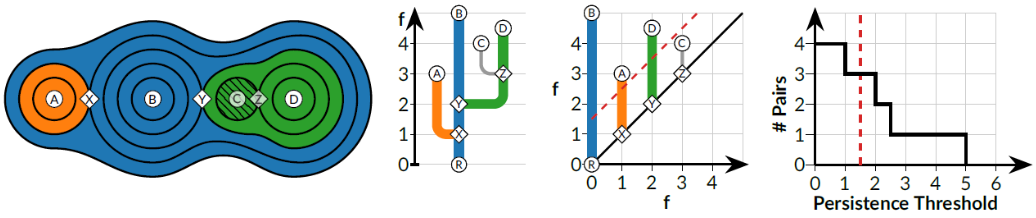
3.3.7 Combining Segmentations (MTCS+MCS)

It is possible to combine the previously described segmentations to derive advanced feature characterizations. In our experiments, we propose a novel combination of the merge tree crown segmentation and the Morse complex segmentation. As shown in

Figure 3.4, this characterization uses the MTCS to extract the areas with a significantly higher event density relative to their local neighborhood, and then these areas are further subdivided into individual hotspots based on the MCS. Note, it is not possible to extract the boundaries of the MTCS+MCS segmentation with a single superlevel set segmentation. However, by incorporating the Morse complex segmentation, the extracted crown components can be further partitioned into individual hotspots, which can then be explored with respect to other features of the geographically referenced points (e.g., crime type). This has the advantage that connected regions can be further subdivided into individual hotspots with natural boundaries that align with the density gradient and provides a robust hotspot extraction when compared to the previous SFT segmentations.

3.3.8 Topological Simplification

As previously stated, SFT-based feature characterizations primarily depend on the maxima of the scalar field. However, scalar fields usually exhibit noise in the form of undesired maxima. For instance, a density field might contain several small peaks that are the result of sampling artifacts or under-smoothing. To remove undesired maxima—and therefore their corresponding features—one can use a procedure known as topological simplification (Lukasczyk *et al.*, 2020; Edelsbrunner *et al.*, 2006). This procedure filters features by symbolically removing their corresponding maxima from the scalar field, which can be imagined as if the corresponding hill would be flattened to a plateau. After this procedure, all SFT-based feature characterizations presented in the previous sections can be computed on the simplified field, and the resulting abstractions will omit the undesired features (Figure 3.5a,b). The merge tree can be used to intuitively understand this simplification procedure as undesired branches (features) are absorbed into the preserved branches they are attached to. This makes



a) Simplified KDE b) Simplified MT c) Persistence Diagram d) Persistence Curve

Figure 3.5: Topological Simplification (Lukasczyk *et al.*, 2020; Edelsbrunner *et al.*, 2006) Makes It Possible to Remove Hotspots That Do Not Exceed a given Persistence Threshold (Red Dashed Line) by Symbolically Removing the Corresponding Maximum. Filtered Hotspots Are Assigned to the Region of Preserved Hotspots, Which Enables Analysts to Examine Hotspots in a Level-of-detail Approach. For Instance, Large Persistence Thresholds Preserve Only the Most Significant Hotspots, and by Continuously Decreasing the Persistence Threshold These Hotspots Are Incrementally Partioned into Smaller Hotspot Regions. For the Example Scalar Field Shown Here (A), the Maximum c Has Been Removed as Its Pair Does Not Exceed the Current Persistence Threshold of 1.5. Its Corresponding Region Is Therefore Assigned to the Region of Maximum d. This Is Also Indicated in the Merge Tree (B), Where We Render Filtered Branches as Gray Lines. To Choose Appropriate Persistence Thresholds, Topological Data Analysts Usually Examine Two Visualizations: The Persistence Diagram and the Persistence Curve (Edelsbrunner and Harer, 2010). The Persistence Diagram (C) Renders Each Persistence Pair as a Line Which Indicates for Which Level the Pair Is Born and Destroyed. The Persistence of Each Pair Is Encoded by the Height of Its Corresponding Line, and the Persistence Threshold Is Rendered as a Line Parallel to the Main Diagonal. Hence, Less Persistence Features Are Close to the Diagonal. The Persistence Curve (D) Shows the Total Number of Pairs (Y-axis) That Exceed a given Persistence Threshold (X-axis).

it possible to explore features in a level-of-detail approach as analysts can first examine the most significant features, and then iteratively break these features apart, enabling the exploration of localized geographic neighborhoods.

In this work, we use persistence to rank and filter features. Persistence is a natural choice for a significance measure since the importance of a hotspot can be intuitively measured by its relative event density (which is exactly captured by persistence). However, topological simplification supports any analyst-defined importance measure, such as the surface area or the location of a hotspot. To examine suitable persistence thresholds, analysts can consult the persistence diagram and the persistence curve (Figure 3.5c,d). Our visual analytics framework enables analysts to adjust the persistence threshold at interactive framerates.

3.3.9 Visual Analytics Framework

In order to demonstrate how SFT constructs can support geographic hotspot analysis, we have developed a visual analytics framework (Figure 3.2) for exploring categorical geospatial point events (e.g. criminal incident reports where a type of crime, such as theft, is reported at a given location). This visual analytics framework is built upon a c++ backend that computes all topological procedures via the Topology ToolKit (Tierny *et al.*, 2017; Masood *et al.*, 2019) and ParaView (Ahrens *et al.*, 2005), and a JavaScript frontend for the interactive visualization of the computed density estimates and characterizations. Key components of the interface (Figure 3.2) include the merge tree (I.1), the persistence diagram (I.2), the persistence curve (I.3), and a geospatial map (I.4). The addition of these topological visualization structures have not been explored in previous geovisual analytics systems. By providing linked controls between the topological visualizations and the geographic views, we provide analysts with access to a new suite of mathematical tools for hotspot extraction.

The most important parameter for any SFT-based characterization is the persistence threshold. As previously stated, large persistence thresholds only preserve the most significant hotspots (in terms of density range), whereas small thresholds also preserve small scale hotspots. By decreasing the persistence threshold, hotspots are partitioned into smaller, less persistent hotspots. The persistence curve (Figure 3.2.I.3) shows the persistence distribution of hotspots, and analysts can interactively adjust the persistence threshold by dragging the vertical filtering line. As analysts drag the line, the persistence diagram (I.2), merge tree (I.1), and map view (I.4) interactively update.

The map view (I.4) is used to display the hotspots and serves as the primary view for exploring spatial relationships between hotspots. In our framework, hotspots can be colored in two ways, by branchID or by the ith largest category within the hotspot (where i is user defined). The hotspots on the map correspond to the branches of the merge tree, and colors are applied consistently to the branch edges in the Merge Tree (I.1) and Persistence Diagram (I.2), and hotspots in the map view (I.4). By hovering over a branch in the merge tree or an edge of the persistence diagram, the corresponding hotspot will be highlighted on the map.

On the right side of the map view, there is a configuration toolbar where analysts can adjust the kernel function, the resolution of the sampling grid, and the kernel bandwidth. The resulting density estimate is then visualized via contours, and other menu options allow analysts to control the color opacity of hotspots and to toggle between color modes via radio buttons.

In order to also support the analysis of the features within a hotspot (e.g., the gangs active in a region or the type of crimes occurring), our framework utilizes geospatial probing (Butkiewicz *et al.*, 2008). For each detected hotspot, the analyst can select/deselect the corresponding branch in the merge tree (I.1) or the corresponding edge in the persistence diagram (I.2) to display/hide a distribution pie-chart that displays the proportions of categories within the geographic hotspot. In order to avoid numerous small pie segments, the number of segments can be adjusted, such that if the analyst chooses three segments, the pie-chart will show the distribution of the three largest categories within a region, as well as a fourth category that aggregates all the remaining categories. The pie-chart is displayed as a glyph on the map view, and the size of the pie-chart is proportional to the total number of events within the hotspot region. For detailed information, our framework also supports mouse over events on the pie-chart to reveal the exact proportion values for a selected pie slice. Furthermore, the pie glyphs will also interactively adjust as the hotspot characterization is modified. The novelty here is that we can now enable a multiclass analysis of aggregated groups which allows analysts to quickly compare distributional profiles between hotspots.

Analysts can use the configuration menus to interactively switch between the previously introduced SFT-based feature characterization to extract hotspots. However, many of the other SFT-based feature characterizations require parameter tuning, i.e., superlevel set components require a level and crown components require a crown height. Our framework allows analysts to directly input parameters or interact with the merge tree to adjust these parameters. For instance, the height of crown components can be adjusted by dragging any of the height markers (black lines of a branch in Figure 3.6.b.2). We also borrow interaction mechanics from the *Flexible Isosurface* interface proposed by Carr et al. (Carr *et al.*, 2010), as we enable feature characterization via merge tree selections. All interactions with the merge tree interactively update the map view.

3.4 Case Studies

Density fields are frequently used in crime mapping and analysis (Chainey, 2013; Ratcliffe, 2010; Chainey *et al.*, 2008; Ratcliffe, 2002). Given this popularity, SFT can enhance analyses of point-based crime data. In order to demonstrate how SFT can be used to enhance geographical hotspot analysis, we present a detailed analysis of two crime datasets, where the first dataset contains general crime incident reports in Tippecanoe County, IN, and the second datatset consists of incident records of gang violence in Chicago, IL.

3.4.1 Criminal Incident Reports from Tippecanoe County, IN

In this case study, we explore differences in the geographic distributions of criminal incident reports in Tippecanoe County, Indiana, home of Purdue University. Our dataset consists of 4,961 geographically referenced criminal incident reports, each of which is attributed to one of eight crime types in 2014, Figure 3.2. The KDE is computed using a linear kernel function with a bandwidth of 1000 meters, and was then normalized to $[0, 1]$. Here, we consider crime hotspots to be defined as areas on a map that have high crime intensity compared to other areas. In our examples, we explore aggregates of all crimes, and then compare regional differences between crimes that make up hotspots.

Next, we demonstrate how an STF-based analysis approach is able to quickly capture known features within the dataset. In Figure 3.2, we visualize the density distribution of all criminal incident reports in Tippecanoe county along with landmarks as a reference for the discussion. Based on historical knowledge of the data from discussions with local law enforcement agencies, several peaks in this dataset are anomalies and should be considered noise. Thus, to remove noise, i.e., insignificant peaks in the KDE, we ignore all peaks that do not exceed a persistence density threshold of 0.04. This threshold is indicated by the persistence curve (Figure 3.2.a.1), since it is located right after the steepest falloff, which usually indicates the point at which all small scale features have been removed. This threshold preserves 11 leaf nodes (from 40) in the merge tree, which indicates the existence of 11 unique, relevant hotspots.

Merge Tree Leaf Segmentation (MTLS)

In Figure 3.2.a.1, the hotspots identified via a MTLS are colored by the branch ID. We can immediately identify two hotspots (light blue and pink) separated by the river. The western hotspot is located near the campus downtown region and the traditional bar district for students. The eastern hotspot is also a bar district; however, the distance from campus typically requires driving to this area and tends to have a different clientele.

Morse Complex Segmentation (MCS)

In order to further understand the hotspot separation, we link the hotspots to their majority crime type (Figure 3.2.a.2), and immediately observe the different distributions of crime between the two regions. The bottom-right corner of Figure 3.2 provides the color map for Figure 3.2, mapping topological components to the corresponding crime hotspot. In Figure 3.2.a.2, we can observe that the largest number of recorded crimes in the hotspots in Tippecanoe County are *drunkness* (blue), *theft* (green) and *domestic disturbance* (pink) incidents. We can display the Morse Complex Segmentation (MCS) colored by the majority crime category to further partition the map into regions. Figure 3.2.b.1 is colored by the majority crime type for each region, and we can see a clear separation of the University region (blue). This is a well-known phenomenon for local law enforcement; However, depending on the choice of threshold for hotspots, the separation of these two peaks is not always immediately obvious.

Here we note that while the color corresponds to the majority crime, our interface is also able to display the second (Figure 3.2.b.2), third (Figure 3.2.b.3), fourth, etc. largest proportion for a given region to enable analysts to quickly explore regional differences, and this is coupled with a tooltip that will display the distribution of all crimes in a region as a pie chart to further explore regional differences. In Figure 3.2.b.2 and Figure 3.2.b.3, we can observe that there are *noise* (red), *burglary* (light green) and *disturbances* (light blue) making up large portions of the crime types, along with the previously identified *drunkness* (blue), *theft* (green) and *domestic disturbance* (pink) incidents. Again, the campus region appears to have a different crime profile than the surrounding areas as shown in the pie chart tooltip.

Combining SSS and MCS

Next, we can use the Superlevel Set Segmentation (SSS) to identify the densest crime areas, and then color the resulting hotspots using MCS to partition the dense area into individual hotspots. Figure 3.2.c shows the three areas that exceed a density threshold of 0.7. These areas are dominated by *drunkness* (blue) and *theft* (green). The corresponding branch of the blue hotspot is the highest in the merge tree, which means the blue hotspot is the densest area, and its color indicates that it is dominated by *drunkness* (blue) arrests.

Combining MTCS and MCS

While the MCS extracts the densest regions, we can apply MTCS to detect all local peaks, and even group them by increasing the crown height (i.e, the density threshold). Figure 3.2.d.1 shows the detected hotspots for crown height 0.2. We can obverse that most hotspots have *theft* (green) as the majority crime, some local hotspots have *domestic disturbance* (pink) as the majority, and that the most dense hotspot has *drunkness* (blue) as the majority. To explore the blue hotspot in detail, analysts can click on the hotspot to examine the crime distribution via a pie-chart. The chart will show as many categories as specified by the analyst, whereas the remaining categories are summarized in a miscellaneous slice. Here, the pie-charts show the distribution of the top six crime types. There is a total of 303 crimes that were reported in the extracted hotspot. The *drunkness* (blue) occupies 42.9%, *theft* (green), *noise* (red), *disturbance* (light blue), *domestic disturbance* (pink), and *burglary* (light green) occupy 25.7%, 21.8%, 4.0%, 3.6% and 1.7%, respectively. We hypothesize that the blue hotspot majority is *drunkness* since it is located near the local university, and this area is home to many restaurants and bars.

The combination of the crown and Morse segmentation is also able to detect local hotspots, such as the hotspot in the top left corner of Figure 3.2.d.1 and d.2. The Morse complex can separate this hotspot into two regions, which actually features different majority crime types: *domestic disturbance* (pink) and *theft* (green). Note, this hotspot is hard to detect with the classical approach of thresholding the density estimate since its crime density is relatively low compared to the downtown and campus area. Moreover, the classical approach is not able to partition the hotspot into separate areas, and by revealing these separate patterns, different patrol profiles could be developed for the areas.

Again, to explore the hotspots shown in Figure 3.2.d.1 in detail, we can decrease the crown height to separate them into individual peaks (Figure 3.2.d.2). Here, we observe that the majority crime type of all hotspots does not change compared with Figure 3.2.d.1, while the order of other crime types changes for some hotspots, as indicated by the pie-charts. *Drunkness* (blue) in the middle green hotspot drops from the second-largest crime type in Figure 3.2.d.1 to the fourth-largest crime type from Figure 3.2.d.2. 83% of all reported incidents in the local hotspot at the bottom of the map in Figure 3.2.d.1 correspond to *theft* (green), while for its core (Figure 3.2.d.2) the proportion increases to 93%. This is near a local shopping area, which is well-known to have issues related to theft.

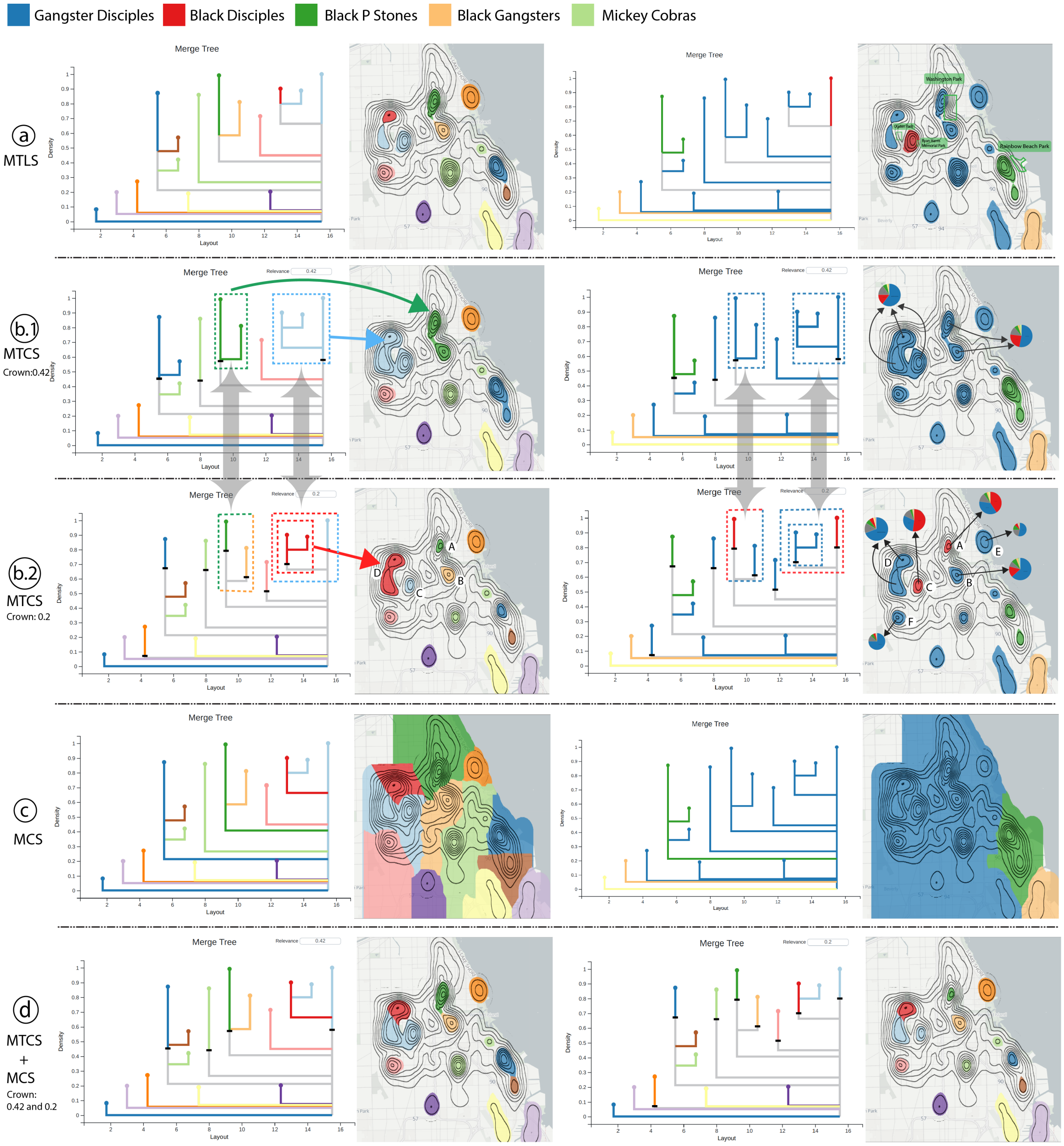


Figure 3.6: This Figure Shows a Compilation of Different Hotspot Characterizations for Gang Activity in Chicago, IL, USA (Section 3.4.2): (A) Merge Tree Leaf Segmentation, (B.1-B.2) Merge Tree Crown Segmentation, (C) Morse Complex Segmentation, and (D) Combination of a Merge Tree Crown Segmentation and a Morse Complex Segmentation. Each Characterization Is Shown as a Merge Tree-map Pair, Where Special Parameters of Certain Characterizations Are Stated in the Corresponding Subcaptions. In the Left Column, Hotspots Are Colored Based on the Id of Their Corresponding Merge Tree Elements, Whereas Hotspots in the Right Column Are Colored Based on the Most Active Gang Inside Each Hotspot (Except for (D), Which Is Always Colored by Branch Id).

3.4.2 Gang Analysis

In this case study, we explore how SFT can support the exploration of gang activities. Our dataset consists of 38,268 geographically referenced criminal incident reports, each of which have been attributed to one of sixty-four gangs in Chicago, IL from 2011 to 2014. We describe our findings in the context of high-profile news stories and discuss how SFT-based geographic analysis can support the analytic process. The KDE is computed using a linear kernel function with a bandwidth of 800 meters. We normalize the density value range to [0, 1]. Figure 3.6 illustrates how various SFT methods are applied, and we discuss how MTLS (Section 3.3.4), MTCS (Section 3.3.5), and MCS (Section 3.3.6) can be applied for geographic hotspot analysis.

Merge Tree Leaf Segmentation (MTLS)

The goal of our SFT-based visual analytics interface is to provide alternative methods for hotspot extraction. First, we discuss the Merge Tree Leaf Segmentation (MTLS - Section 3.3.4), which is capable of extracting any peak independent of a density threshold. As shown in Figure 3.6.a, every leaf arc of the merge tree corresponds to a local peak in the kernel density estimate, where the boundary of each hotspot corresponds to the largest contour of its corresponding peak that does not intersect with a contour of another peak. The leaves and the corresponding hotspots in Figure 3.6.a (left) are colored by branch id to highlight separate hotspots.

In Figure 3.6.a (right) all hotspots and their corresponding leaf arcs are colored by the gang with the most reported incidents in the extracted area. In the following, we will call such a gang the *majority gang* of that hotspot. However, we note that the percent majority needs to be further investigated via the pie chart tool tip as we apply only the concept of a simple majority.

From the hotspots colored by the majority gang in Figure 3.6.a (right), we can observe that there are four main gangs in the global and local hotspots. The *Gangster Disciples* (blue) are the majority gang in most of the hotspots in Chicago. The *Black Disciples* (red) are active between Ogden Park and Ryan Harris Memorial Park. The *Black P Stones* (green) are the majority gang in the bottom right corner on the map, near Rainbow Beach Park. The Black Gangsters (orange) are only active in one hotspot on the east side.

Merge Tree Crown Segmentation (MTCS)

While the MTLS method can automatically extract hotspots and their boundaries, the Merge Tree Crown Segmentation (MTCS - Section 3.3.5) can additionally constrain the extracted hotspots based on a local density threshold, i.e., the so-called crown height. For example, the MTCS can be used to define hotspots as regions surrounding peaks spanning at most a relative density difference. This makes it possible to join neighboring peaks—i.e., connected branches in the merge tree—while still preserving hotspots at different scales (as opposed to SSSs; Figure 3.1).

Figure 3.6.b.1 and b.2 show two MTCSs for different local density thresholds (crown heights). In Figure 3.6.b.1 (left), colors correspond to branch IDs, where subbranches are colored based on the largest crown branch they are connected to. For instance, the merge tree highlights a dark green crown consisting of two branches, and a light blue crown consisting of three branches; their corresponding regions are shown on the map. In Figure 3.6.b.1 (right), we color crowns based on the majority gang, and we can observe that the majority of the hotspots are dominated by a single gang, the *Gangster Disciples* (blue). We can further investigate the gang distributions by selecting individual hotspots. For example, analysts can click on a merge tree branch to add a pie chart to the map that shows the gang distribution in that hotspot. The number of segments in the pie chart is user-defined such that if the analyst chooses to see only the top three gangs in a region, the pie chart will have four slices, three slices corresponding to the proportions of the top three gangs, and a fourth slice corresponding to the proportion of all remaining gangs. In Figure 3.6.b.1 (right), the probing reveals that the two close hotspots are dominated by the same gang, Gangster Disciples (blue), while their content(gang distribution) is different.

To further explore areas of potential gang rivalries, we can decrease the crown height of the MTCS to separate hotspots, i.e., explore them in a level-of-detail approach (Figure 3.6.b.2). As shown in the left column, the hotspots of Figure 3.6.b.1 split into multiple, separate hotspots for the crown height used in Figure 3.6.b.2. Specifically, the dark green and the light blue crown split each into two crowns. Note, that the color of the main branches are preserved, and subbranches are assigned a new color. Next, by coloring the branches by the majority gang and probing the branches Figure 3.6.b.2 (right), we can begin to observe different local regional stability with respect to gang distribution. Again, we observe that the majority of the hotspots have the majority of gang related arrests associated with the *Gangster Disciples* (blue - hotspots B, D, E, and F); however, there were local regional differences that were obscured in Figure 3.6.b.1. Now we can observe areas in red representing the *Black Disciples* (red - hotspots A, C) that are in close proximity to the blue hotspots. Here we can hypothesize that the regions with the most even proportions in the pie chart are more prone to violence. From the pie chart, we can observe that hotspot A has a large mix of different gangs. Hotspot A is adjacent to Washington Park, and, based on local news reports, we can see that this area is noted to have one of the highest murder rates in Chicago (Mos, 2022; Was, 2017). Similar to hotspot A, hotspot C—near Ogden Park— has the *Black Disciples* (red) as the majority gang. Activity of the *Gangster Disciples* (blue) was also frequently reported in these areas, as indicated by big slices in the pie-charts. As such, hotspot C would appear to be a critical hotspot for potential gang violence, and an exploration of local news reports reveals several high profile gang murders in this area. For instance, according to local news reports, in 2012, an 18-year-old rapper named *JoJo* was killed near hotspot C by a member of the Black Disciples, and Jennifer Hudson's family was also murdered by a gang member near here in 2008. As such, the combination of SFT methods and probing is able to quickly support the analysis process, and future work may consider novel representations of the merge tree to help highlight distributional patterns.

We also observe that the local hotspots detected with MTLS are still preserved when applying a MTCS, see hotspot E. This region is interesting as it is near Harsh Park and has significantly less gang incidents than the downtown area. Yet, the park—relative to the area surrounding it—still exhibits an elevated amount of incident reports. If only regions with a high threshold (large number of gang incidents) are explored, local hotspots, such as this one, might be missed. From local news reports, this area was the home of a number of gang violence incidents that made national news. For example, in 2013, a 15-year old black girl, Hadiya Pendleton, was murdered by two gang members in Harsh Park. These gang members belonged to two rival factions. This serves as a meaningful example as to why local hotspots analysis is also critical.

Morse Complex Segmentation (MCS)

While we have shown how our framework is able to support the exploration of global and local hotspots based on an analyst-defined crown height (MCTS), analysts may also want to look at more general zones or territories in the data. This may support resource allocation, or serve as a mechanism for identifying gang territory boundaries. To support region partitioning, we utilize the Morse Complex Segmentation (MCS -

Section 3.3.6).

Figure 3.6.c shows the application of MCS to partition the map into different regions corresponding to hotspot territories. In Figure 3.6.c (left), the colors correspond to the branch ID, and in Figure 3.6.c (right), the colors correspond to the majority gang. Figure 3.6.c (right) lets the analyst observe the general gang territories and quickly identify potential boundaries relating to a gang's territorial geography, which has been identified as a critical component of gang rivalry (Radil *et al.*, 2010). Future work will explore how to map the identified boundaries to local geographic features (roads, waterways, etc.) to further enhance regional territory extraction.

Combining MTCS and MCS

Along with applying MTCS and MCS, we also explore the impact of combining such segmentations (MTCS+MCS - Section 3.3.4). Here, the combination of the MTCS and the MCS provides a different approach to separate crowns based on the boundaries of the Morse complex. In the context of gang activity, the boundary of the Morse complex corresponds to gang territory boundaries. So crowns that consist of multiple peaks can be separated according to the gang territory. Note, this is different to subpartioning crowns based on different heights (3.4.2), since crown boundaries correspond to contours of the KDE, while Morse boundaries follow the gradient of the KDE.

Figure 3.6.d (left) shows the combination of the MCS and the MTCS for crown height 0.42. Here, we can see that we are not able to separate the green hotspot of the pure MTCS shown in Figure 3.6.b.1 (left) into two subhotspots (green and light orange), although they are connected. The same can be observed for the crowns with height 0.2 shown in Figure 3.6.b.2. Hence, the combination of both segmentations may make it possible to further subdivide extracted hotspot clusters along gang territory boundaries, providing further insight into how the features of the point distributions vary within local regions.

3.5 Conclusion

To summarize, our system provides different characterizations and customized preferences for the comprehensive analysis of hotspots. We explored the advantages and disadvantages of several scalar field topology (SFT)-based feature characterizations such as merge tree segmentations and Morse complexes for the exploration of geo-spatial event hotspots inherent in kernel density estimates. As demonstrated in two case studies (3.4), these provide novel methods for extracting and characterizing local geographic features which may be missed when simply extracting connected areas that exceed a fixed, user-specified density threshold, i.e., superlevel set components. However, there are limitations to the application of SFT for hotspot analysis.

A limitation of kernel density estimates is their strong dependence on the kernel bandwidth. In our experiments, we only considered the bandwidth suggested by Silverman's Rule of Thumb (3.3), but in future work we plan to investigate if SFT can also be used for bandwidth optimization. Along with bandwidth selection, it is also necessary to select the correct SFT-based characterization based on the estimate at hand, and all characterizations depend on at least one parameter: the persistence threshold. Our visual analytics framework enables analysts to compare/adjust characterizations and explore hotspots in a level-of-detail approach by adjusting the persistence threshold.

Another limitation comes from the central topological abstraction used in the proposed SFT-based characterizations, the merge tree. If the topological complexity of the underlying density estimate increases, then so does the number of nodes and edges of the tree. At some point, the tree becomes too complex and can no longer be used as an interaction device (Carr *et al.*, 2010). In future work, we will investigate if this issue can be overcome by, for example, partitioning the kernel density estimate or by displaying only relevant parts of the tree. Furthermore, we want to explore how to enhance the merge tree to capture more information about the underlying feature distributions that are shown in the Pie Chart.

Overall, the primary advantage of the discussed SFT-based characterizations is their capability to describe hotspots at multiple resolutions, which makes it possible to rank, filter, and decompose hotspots based on their significance. In this thesis, we have demonstrated that a novel combination of crown components and Morse complex cells yields a very robust, versatile, multi-scale hotspot characterization. Furthermore, this framework presents an opportunity for crime analysts and law enforcement to work together collaboratively to map particular crime types for a given region. The tool is flexible enough to allow analysts to solicit input from patrol officers and change persistence thresholds and explore various hotspot segmentation strategies. Analysts can also use the MTCS+MCS segmentation to help officers understand hotspots of particular type of crime. Plotted over time, with similar KDE specifications, side-by-side comparisons of the hotspots produced from this segmentation can help law enforcement understand how the distribution of crime types changes over time within the study area of interest.

We believe that the ability to segment regions and patterns has an immediate benefit for directing targeted law enforcement that can best respond to the typical types of crimes in a neighborhood. For example, looking at the maps of the communities of public intoxication on the west and east sides of the river in Figure 3.2, we can observe that the hotspots cover known parking structures and undergraduate dorms in the dataset. By doing random foot patrols optimized by these locations, officers can potentially prevent drunk driving (which is a more serious offense). Further analysis and characterization of hotspots can lead to customized patrol routing and be linked to optimization packages for improving police coverage and response. As such, the use of SFT for characterizing geospatial hotspots can provide analysts with novel capabilities for hotspot identification and extraction. Beyond the crime case studies presented in this thesis, we believe that the application of SFT to density analyses has myriad other applications, as does the use of KDE alone. From this perspective, the framework outlined in this thesis presents the opportunities for deeper insights into multiple types of phenomenon characterized by point level data (e.g. wildfires, air traffic patterns, animal movements and food environments) and their derived density estimates. Given that one of the most common density estimation techniques is KDE, future research will also investigate how SFT could be used to characterize the impact of the kernel function, bandwidth, and resolution of the KDE on the derived hotspots. Incorporating this kind of analysis into a visual analytics tool would enable analysts to compare and choose between different parameters. For instance, as one increases the bandwidth the KDE loses details and becomes more blurred (small detailed hotspots disappear while other hotspots merge). Providing statistics about the number and properties of hotspots for different bandwidths highlights salient bandwidth values that will either yield too coarse or too detailed hotspot partitions.

In this chapter, the application of scalar field topology for multiclass map analysis is discussed. While the methods to profile the multiclass regions in the map analysis can be rarely explored. In the next chapter, I combine the merge three and stream graph into a novel visualization view, stream tree for multiple scalar fields that are defined over the same domain.

Chapter 4

VISUALIZING MULTIPLE SCALAR FIELDS USING REPRESENTATIVE TOPOLOGICAL FEATURES

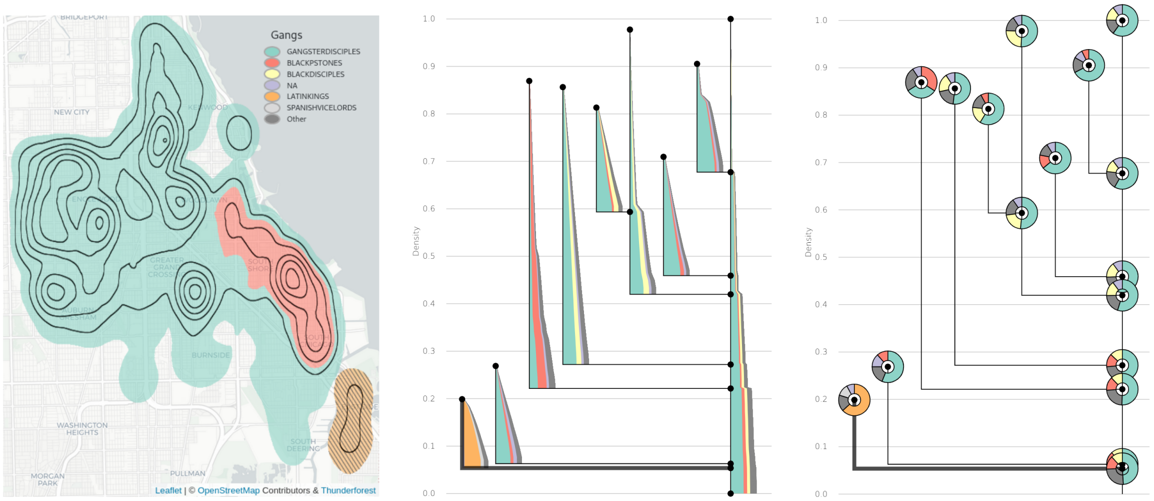


Figure 4.1: The Proposed Stream Tree View (Center), Illustrating the Distribution of Gang Activity in Chicago, Illinois. Here, the Analyst-defined Reference Field r Corresponds to the Density Estimate of All Gang Activity in Chicago (Contours, Left). The Merge Tree of r (Black Edges, Center) Indicates for Which Density Value (Y-axis) Individual Gang Activity Hotspots Appear and Merge, I.E., Each Point of the Merge Tree Corresponds to a Unique Superlevel Set Component (Hotspot). Here, Each Region of the Map Is Colored by the Gang That Has the Largest Integral in the Corresponding Superlevel Set Component. With the Stream Tree, Analysts Can Determine at a Glance Which Gangs Are Active in the Individual Regions, and Which Regions Are Clearly Dominated by a Single Gang or Are Contested. For Example, the Hatched Region in the Lower Right Corner of the Map Corresponds to the Highlighted Edge of the Merge Tree. In This Area, Almost All Reported Gang Activity Is Attributed to the *Latin Kings* (Orange), Which Is Effectively Visualized with the Stream Tree and the Coloring of the Map. All Other Hotspots Are Clearly Dominated by the *Gangster Disciples* (Teal), Except for One Hotspot Where Slightly More Activity of the *Black P Stones* (Red) Gang Was Reported. To Further Reduce Visual Clutter and Emphasize Proportions, One Can Also Visualize Attribute Vectors with Donut Charts Drawn at the Critical Points of the Merge Tree (Right).

4.1 Introduction

To date, limited approaches have been proposed for visualizing multiple scalar fields that are defined over the same domain; such as a set of kernel density estimates (Figure 4.1). Common visualization approaches, such as small multiples or overdrawing, do not scale with the number of fields. As an alternative to these approaches, we introduce *stream trees*, a compact feature-centric visualization for datasets that can be represented as a vector function $f : \mathcal{D} \rightarrow \mathbb{R}^n$ which maps a point of a simply connected manifold $\mathcal{D}$ to n real values. The core idea of our approach is to derive a spatial segmentation of $\mathcal{D}$ such that every segment represents a feature of interest, and then we integrate f inside each feature to derive an attribute vector for that feature. This attribute vector records the proportions between the individual field integrals inside the feature. For example, if the fields represent the amount of reported gang activity, then the attribute vector can be used to visualize the ratio between the activity of individual gangs in an analyst-defined region. To derive these regions i.e., meaningful domain segmentations we introduce the notion of an analyst-defined reference field $r : \mathcal{D} \rightarrow \mathbb{R}$, which can, for example, be a specific input field or an aggregation of all input fields. In many applications, features of interest can be characterized as superlevel set components of r, i.e., connected regions of the domain that exceed a given level threshold l. The merge tree is a topological abstraction whose edges record at which levels superlevel set components appear and merge (black edges of Figure 4.1 center), and we partition the domain according to the merge tree segmentation of the reference field. Note, each point on the merge tree corresponds to a unique superlevel set component, so we record at each point the attribute vector of the corresponding component. These attribute vectors can then be visualized along the edges of the merge tree with a stream graph, which yields a novel, compact visual representation of f with respect to topological features of r: the stream tree (Figure 4.1 center). We also discuss two alternative visualizations of the attribute vectors: the donut and pie tree (Figure 4.1 right). We demonstrate the benefit of using these visualizations on datasets from geographically referenced crime data and scientific visualization.

4.2 Background and Related Work

In this section, we review related work on topology and stacked graphs.

4.2.1 Scalar Field Topology

Scalar Field Topology (SFT) (Edelsbrunner and Harer, 2010) provides various topological abstractions which have been shown to effectively characterize features in many application domains; including astrophysics (Sousbie, 2011; Shivashankar *et al.*, 2016), bioimaging (Carr *et al.*, 2004; Bock *et al.*, 2018; Anderson *et al.*, 2018), chemistry (Bhatia *et al.*, 2018; Guenther *et al.*, 2014; Olejniczak *et al.*, 2019), fluid dynamics (Laney *et al.*, 2006; Kasten *et al.*, 2011; Bremer *et al.*, 2016), geostatistics (Zhang *et al.*, 2021b; Lukasczyk *et al.*, 2015), material sciences (Gyulassy *et al.*, 2007, 2015; Lukasczyk *et al.*, 2017a; Soler *et al.*, 2019), and turbulent combustion (Bremer *et al.*, 2011; Gyulassy *et al.*, 2014).

One of the most prominent topological abstractions is the merge tree (Carr *et al.*, 2003) that records either the evolution of superlevel set components (in which case it is alternatively called the split tree) or the evolution of sublevel set components (in which case it is also called the join tree). In this thesis, we focus on superlevel set components i.e., split trees but our method can symmetrically be applied to sublevel set components, i.e., join trees. To understand the construction of merge trees, imagine the scalar field shown in Figure 4.2d as a 2D landscape where the elevation of every point corresponds to the scalar value at that point. Now imagine that the entire landscape is under water for a water level that exceeds the highest elevation. If the water level now continuously decreases, then isolated islands will appear at the maxima of the landscape. As the water level further decreases these islands will grow and eventually merge together at so-called saddle points. The merge tree tracks the evolution of these islands (in mathematical terms superlevel set components) during this level sweep. Every time the level reaches a maximum, then the merge tree creates a new leaf arc, and every time two islands (superlevel set components) merge then their corresponding arcs merge in the tree. Following the so-called Elder rule (Edelsbrunner and Harer, 2010), we say that when two arcs merge at a saddle, then the arc that corresponds to the lower maximum terminates, and the arc of the larger maximum continues. This way, every maximum is uniquely paired with a saddle, except for the global maximum, which, by convention, is paired with the global minimum. Moreover, during the computation each point of domain is associated with a point on the merge tree, and vice versa. This mapping is referred to as the merge tree segmentation.

We can also measure the significance of a merge tree edge via the absolute scalar value difference of the corresponding paired points. This importance measure is called the persistence (Edelsbrunner *et al.*, 2002) of the arc. Furthermore, one can simplify the merge tree and even the original scalar field to remove arcs below a user-specified persistence threshold (Lukasczyk *et al.*, 2020). This is very useful since in many applications arcs with a low persistence value often correspond to noise, while arcs with a high persistence value correspond to features of interest.

The proposed approach is implemented in the Topology ToolKit (TTK) (Tierny *et al.*, 2017; Masood *et al.*, 2019) and utilizes existing features such the topological simplification (Lukasczyk *et al.*, 2020) and the merge tree computation (Gueunet *et al.*, 2017). Many topological algorithms including the ones implemented in TTK require that the input scalar function is injective (e.g., to prevent ambiguity). Although this constraint is usually not met in practice, one can always enforce injectivity by performing a symbolic perturbation of the input scalar field with a pre-processing procedure inspired by *Simulation of Simplicity* (Edelsbrunner and Mücke, 1990). Our implementation handles this implicitly through the existing TTK infrastructure.

In this thesis, we build upon the concept of the merge tree by augmenting edges with visualizations that provide insight into multiple fields. Traditionally, a merge tree is an abstraction of a single scalar field, and to examine multiple fields via merge trees analysts have to compare small multiples, where the interplay between different fields can be difficult to interpret. We seek to derive a new visualization by aggregating multiple fields with respect to a merge tree segmentation of an analyst-defined reference field. This way we can visualize aggregated data of multiple fields on one merge tree, which provides a compact visual representation that summarizes multiple fields with respect to the merge tree of the reference field.

4.2.2 Topology-Based Visualization of Multiple Scalar Fields

Due to its abstracting nature and utility for scalar field analysis and visualization, researchers have sought to generalize topology-based visualization techniques to treat multiple scalar fields.

Edelsbrunner and Harer introduce the *Jacobi set* of k real Morse functions on a d-dimensional domain ($d \geq k$) as the set of critical points of one function restricted to the preimages of the remaining functions (Edelsbrunner and Harer, 2004). For $k = 2$ generic functions, the Jacobi set is a 1-manifold. In the piecewise linear setting, they present an algorithm for computing Jacobi sets as a set of grid edges, decomposed into simple cycles. Jacobi sets are furthermore useful to determine local or global similarity of functions (Edelsbrunner *et al.*, 2004). Nagaraj and Natarajan (N and Natarajan, 2010) use this to introduce simplification of the Jacobi set of two Morse functions on a 2-manifold.

Schneider et al. (Schneider *et al.*, 2008) describe *largest contour segmentation*, obtained from computing two field's contour trees (Schneider *et al.*, 2008), and use spatial overlap to compute and visualize similarity among contours of different fields. Schneider et al. later extend this approach to multifields (Schneider *et al.*, 2013), using an approach based on mutual information.

Generalizing *Reeb graphs*, the *Reeb space* of k functions on a common d-manifold is described by Edelsbrunner et al. (Edelsbrunner *et al.*, 2008) as the quotient space of the connected components of function preimages. For two scalar fields, Carr et al. (Carr *et al.*, 2015) define *fiber surfaces* as a bivariate equivalent to isosurfaces; each fiber in a fiber surface corresponds to a preimage in the Reeb space and thus generalizes univariate contours. Tierny and Carr (Tierny and Carr, 2017) present *Jacobi Fiber Surfaces* as an efficient computational approach. Sakurai et al. (Sakurai *et al.*, 2020) investigate how fiber surfaces can be used for exploratory visualization.

Singh et al. (Singh *et al.*, 2007) presented a structure called *Mapper* for computing a descriptive simplicial complex for scalar multifields, which can be viewed as an approximation of the Reeb space, and visualize the resulting graph structure using force-directed layout. Carr and Duke(Carr and Duke, 2013) proposed a similar construction called *joint contour nets* to analyze multiple scalar fields, computed by discretizing each function's range into equally-sized intervals and constructing a graph from connected components of the finite number of preimages of the discretized function. They give an algorithm to compute joint contour nets for piecewise linear data and present them as node-link diagrams, also using force-directed graph drawing. Combining Reeb spaces and Jacobi sets, Chattopadhyay et al. (Chattopadhyay *et al.*, 2014) propose *Jacobi structures*, i.e. projections of Jacobi sets of a multifield onto its Reeb space, to obtain more expressive visualization. They further use this approach to define a scaling-invariant topological simplification scheme for multifields on domains with simple topology (Chattopadhyay *et al.*, 2016).

Huettenberger et al. (Huettenberger *et al.*, 2013) employ the notion of Pareto optimality to visualize k scalar fields on a common domain to identify Pareto extrema; these extend the notion of local extrema to multifields; however, an analogue to saddle points is not given. They describe the functions' *Pareto sets* as the union of all non-regular points. The authors gave an algorithm to compute Pareto sets for piecewise linear functions, and used a strategy similar to Nagaraj and Natarajan [NN11] to simplify Pareto sets and remove noise (Huettenberger *et al.*, 2014).

Focusing on the case of ensemble visualization, Lohfink et al. (Lohfink *et al.*, 2020) use tree alignment to align contour trees of multiple functions. This alignment can be used for tree layout and interactive visualization. They later extend this to the case where multiple time steps of a scalar field are visualized (Lohfink *et al.*, 2021).

4.2.3 Stacked Graphs

The proposed stream tree visualization utilizes stacked graphs as a means of displaying information about multiple fields with respect to a merge tree. Stacked graphs and their variations have found wide spread use in the visualization community, most commonly in time series representations where multiple time series are stacked on top of each other using filled shapes to represent magnitude. Several variations of the stacked graph exist. Byron and Wattenberg proposed the stream graph to visualize an individual's listening pattern of particular artists. In the stream graph, the x-axis represents time. Each strip of the stream graph represents an artist. The thickness of the strip represents the number times an artist's song was listened to. Similarly, the New York Times also visualized the box office revenue (Cox and Byron, 2008) using a stream graph. The y-axis represents time. Each layer of the stream graph represents the movie. The thickness of the layer represents the revenue of the movie.

Havre et al.(Havre *et al.*, 2000) extended this concept to ThemeRiver, which represents the changes of the theme for documents over time. Each current is one theme in the document. The width of the current represents the number of documents where the theme exists. The current is drawn from left to right in the horizontal direction. The currents are displayed symmetrically around the horizontal axis of the river. The pattern and changes of currents can be easily explored by ThemeRiver. Work by Cui et al. (Cui *et al.*, 2011) and Xu et al. (Xu *et al.*, 2013) add threads representing keywords into each layer of the topic in the stacked graph. The combination of thread and layer supports the exploration of concurrent variables in one view while observing their correlation with each other. Sun et al.(Sun *et al.*, 2014) apply a similar method to explore the relationship of topic leaders.

Recent work by Bartolomeo and Hu (Di Bartolomeo and Hu, 2016) developed methods for improving the ordering algorithm of the stream graph to fit more general data. Their method focused on solving the distortion issue, and the proposed wiggle value of the ordering is used to evaluate the performance of the generated stream graph. To improve the readability of the stream graph, Bu et.al (Bu *et al.*, 2020) focused on the sine illusion effect. The distance between two adjacent layers of the stream graph should be the vertical distance rather than the orthogonal distance. While the sine illusion effect makes human focus on the orthogonal distance. The paper improved the ordering algorithm and baseline selection algorithm to minimize the illusion effect of the stream graph.

Our work utilizes the stacked graph design to augment the merge tree. Here, each edge of the merge tree can be related to underlying features across multiple scalar fields. By visualizing the integral of these features, we are able to provide a quick overview of the distribution of these fields with respect to a reference field.

4.3 Method

The input of the proposed approach is a vector function $f : \mathcal{D} \to \mathbb{R}^n$ that maps each point of a simply connected domain $\mathcal{D}$ to n positive real values, i.e., n scalar functions defined over the same domain. The proposed approach aims to derive a meaningful segmentation of the domain, and then integrate the n scalar fields individually over each segment. This approach can be seen as a data reduction technique that reduces the values of f inside a segment $\mathcal{S} \subseteq \mathcal{D}$ into a single attribute vector $A^{\mathcal{S}} \in \mathbb{R}^n_+$, where the i-th entry of the attribute vector records the integral of the i-th function over the segment, i.e.,

$$A_i^{\mathcal{S}} = \int_{\mathcal{S}} f_i(x)dx. \tag{4.1}$$

Thus, the attribute vector represents the reduced information of the segment. Since we are ultimately interested in visualizing the proportions between the attribute vector entries, we require that the attribute vector only contains positive values for any possible segment of the domain. This can be enforced by first applying a transformation to f in Eq. 4.1.

In short, the proposed approach reduces any vector function no matter how many components into an analyst-defined number of attribute vectors, where the number and value of the vectors depend on the number and shape of the domain segments, respectively. This approach therefore relies heavily on the chosen domain segmentation. In the following, we will first demonstrate how to derive meaningful domain segmentations with respect to so-called reference fields (Section 4.3.1). Specifically, we derive a merge tree segmentation of the reference field and then compute the corre-

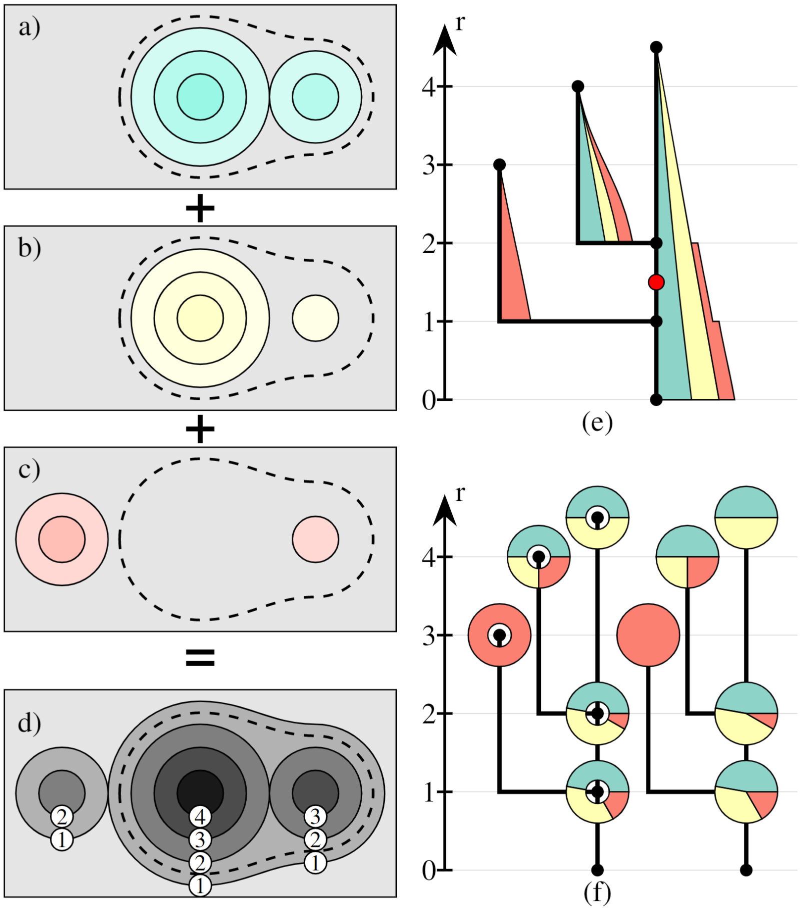


Figure 4.2: Illustration of the Proposed Visualization Pipeline Whose Input Is a Set of n Scalar Fields That Are Defined on the Same Domain $\mathcal{D}$ (A-C), I.E., A Vector Function $f : \mathcal{D} \to \mathbb{R}^n$. First, the Analyst Chooses a Reference Field r to Base Their Analysis on, Which Can Be a Specific Input Field or an Aggregation of the Input Fields (Section 4.3.1). Here, r Is the Sum of All Input Fields (D). Next, the Algorithm Computes the Merge Tree Segmentation $\phi : \mathcal{T} \to 2^{\mathcal{D}}$ of r That Maps Each Point of the Merge Tree $\mathcal{T}$ (Black Edges of E and F) to Its Corresponding Superlevel Set Component $\mathcal{S} \subseteq \mathcal{D}$ of r (Section 4.3.2). For Instance, the Red Point in (E) Maps to the Interior of the Dashed Region in (D). Given a Set of Sample Merge Tree Points $p \subseteq \mathcal{T}$ the Algorithm Then Computes the Component-wise Integral of f over the Area of the Corresponding Superlevel Set Components. This Yields, for Each Merge Tree Point $p \in P$, an Attribute Vector $a^{\phi(P)}$ Whose Entries Correspond to the Individual Integrals, I.E., $a_i^{\phi(P)} = \int_{\phi(P)} f_i(X)Dx$. In This Example, the Attribute Vector of the Red Point Records the Integrals of the Green, Yellow, and Red Scalar Field over the Area of the Dashed Region. The Magnitude and Ratio Between the Attribute Vector Entries Can Then Be Visualized with a Stream Graph (E) along the Merge Tree Edges (Section 4.3.3). Alternatively, It Is Possible to Visualize the Attribute Vectors with Donut or Pie Charts at the Critical Points of the Merge Tree (Left and Right Tree in (F)).

sponding attribute vectors (Section 4.3.2). Finally, we visualize the attribute vectors by augmenting the merge tree with stream graphs, donut charts, or pie charts (Section 4.3.3). This procedure is illustrated in Figure 4.2.

4.3.1 Reference Fields

As a first step, analysts have to define a so-called reference field r. As the name suggests, the following analysis will be in reference to topological features of this field. Specifically, r should be chosen such that superlevel set components of r correspond to features of interest in the application problem at hand. For instance, the first reference field used in the gang activity example (Section 4.4.1) corresponds to the total density estimate of all reported gang activity. In law enforcement, it is common to inspect the total KDE to get a first overview of generic gang hotspots, i.e., areas with elevated gang activity without differentiating between individual gangs. The total KDE is therefore an excellent candidate for a reference field to further examine the distribution between the different gangs in these hotspots. In the next step of the analysis we show that one can use the KDE of a specific gang as a reference field to examine the activity of other gangs inside their hotspots. Similarly, one can choose individual members of a scalar field ensemble as a reference field to examine how much other members agree or disagree with the chosen member (Section 4.4.2). It is even possible to chose a reference field that is not even in the same dimension/scale of the other fields; as demonstrated in the asteroid impact example in which we explore the distribution between water and asteroid matter in superlevel set components of the temperature field (Section 4.4.3). In the end, it is up to the analysts to determine suitable reference fields.

4.3.2 Merge Tree Segmentations

In the next step of the approach we derive a merge tree segmentation $\phi : \mathcal{T} \to 2^{\mathcal{D}}$ of a given reference field r, where $2^{\mathcal{D}}$ is the powerset of $\mathcal{D}$. The concept behind the merge tree computation is summarized in Section 4.2.1, but in short the merge tree segmentation ϕ maps a merge tree point $p \in \mathcal{T}$ to a superlevel set component $\mathcal{S} \subseteq \mathcal{D}$ of r. We compute ϕ with the implementation provided in the Topology ToolKit (Gueunet *et al.*, 2017). Optionally, one can simplify r by persistence (Lukasczyk *et al.*, 2020) to suppress noise prior to computing the segmentation.

Given a set of sample points $P \subseteq \mathcal{T}$ on the merge tree and a vector function f, we can then compute for each point $p \in P$ the attribute vector $A^{\phi(p)}$ of the corresponding domain segment $\phi(p)$ with Eq. 4.1.

4.3.3 Visualization

We propose three approaches to visualize the attribute vectors recorded on the merge tree points: stream trees, donut trees, and pie trees. All visualizations are augmentations of the merge tree, which we render with a branch-centric layout where new edges always branch off to the left. Thus, merge tree edges are rendered as vertical lines that are connected at their lowest point to their corresponding saddle via a horizontal line (black edges in Figure 4.2e). We extended the existing layout algorithm of TTK to handle this additional layout constraint.

Stream Trees

To render a stream tree we first reserve some fixed space on the right side of every merge tree edge. Then we determine a global stacking order of the attribute vector components by sorting the components by their largest value in descending order.

Next, we render the attribute vectors with a stream graph by horizontally stacking their components, where we only color the n largest components per stream graph; the remaining components are aggregated in a misc category that is consistently colored in dark gray. This procedure yields a stream graph for each merge tree edge, and since edges always branch off to the left the visualization can not exhibit any edge crossings.

To determine the actual width of the individual stream graphs we support three different scaling approaches: global, local, and fixed scaling. *Global scaling* normalizes all stream graphs according to the largest magnitude of all attribute vectors to fit the graphs in the reserved space (Figure 4.2e). Global scaling has the advantage that one can compare ratios as well as the values of the attribute vectors across all stream graphs. However, attribute vectors with small values become almost unreadable, especially at the tip of the edges where the integration area of the corresponding superlevel set components is just too small in comparison to the ones further down in the tree. With *local scaling*, each edge individually determines the largest magnitude of its attribute vectors, and then uses it to normalize its stream graph (Figure 4.1 center). Local scaling has the advantage that even small attribute vectors are visualized along the edges at the fully available space, while still encoding the evolution of the values and magnitudes along an individual edge. Although the ratios between the individual values of different stream graphs are still comparable, their absolute values which are encoded by width are no longer in the same scale. Local scaling also has the same problem as global scaling that the stream graph becomes more narrow as one approaches the top of the edges. With *fixed scaling*, each attribute vector is normalized according to their own magnitude, i.e., the width of every stream graph always corresponds to the available space (Figure 4.4). Fixed scaling no longer encodes the values of the attribute vectors, but emphasizes the proportions between the values.

Donut and Pie Trees

Alternatively, attribute vectors can be rendered via donut or pie charts at the critical points of the merge tree, thus each attribute vector component is represented by an arc (Figure 4.2f). In contrast to stream trees, donut and pie charts make it easier for analysts to judge the proportions between the components. We noticed that pie trees are less visually cluttered, but the holes of the donut charts make it possible to still show the branching of the merge tree.

Moreover, a donut or pie chart does not have to visualize the attribute vector at that specific point. For example, a donut chart shown at a maximum node can summarize all attribute vectors along its edge until the next saddle point. Recall, each merge tree edge corresponds to a single superlevel set component, which grows as we go further down the tree until it finally merges with another component at a saddle point. Thus, the donut tree shown at a maximum can summarize all attribute vectors that are associated with this single component just right before it will merge with another component, and these attribute vectors are exactly those vectors that are located on the edge between the maximum and the saddle. So, for instance, the donut chart shown at a maximum could visualize the component-wise maximum of all these vectors, or their mean. In this case, the donut chart is representing an entire edge, not just a single point.

4.3.4 Implementation

All analysis steps i.e., the computation of the merge tree segmentation and the attribute vectors are implemented in TTK (Tierny *et al.*, 2017; Masood *et al.*, 2019). The merge tree and the attribute vectors can then be imported in our web-based visualization tool, which uses D3 (Bostock *et al.*, 2011) to render the stream, donut, and pie trees (Figure 4.1 center). If the domain of the reference field corresponds to a geo-spatial map then we utilize Leaflet (Crickard III, 2014) and WebGL (Parisi, 2012) to render the reference field on top of an interactive map (Figure 4.1 left). For all other domains, such as 3D volume data (Figure 4.6), we use ParaView (Ahrens *et al.*, 2005) to render the reference field.

Our web-based front end also supports linking and brushing. Analysts can click on any part of the merge tree and see the corresponding superlevel set component as a hatched area in the map view. By default, we render the current reference field via contour lines, and color each point of the field according to the largest component of its corresponding superlevel set component. For instance, the map view of Figure 4.1 indicates that the teal component (the *Gangster Disciples*) is the largest component in almost any part of the map, except for the red and orange areas. Our tool also enables analysts to color the map based on the second largest component, and so forth.

4.4 Usage Scenarios

In this section, we present three usage scenarios that demonstrate how the Stream Tree visualization can support the analysis of multiple scalar fields from various domains. First, we analyze geographically referenced criminal activities and compare different hotspot profiles of topology. Then, we demonstrate the flexibility of the Stream Tree by analyzing a climate change ensemble dataset. Finally, we explore how the Stream Tree can be used to support the identification of unique 3D structures using an asteroid impact simulation.

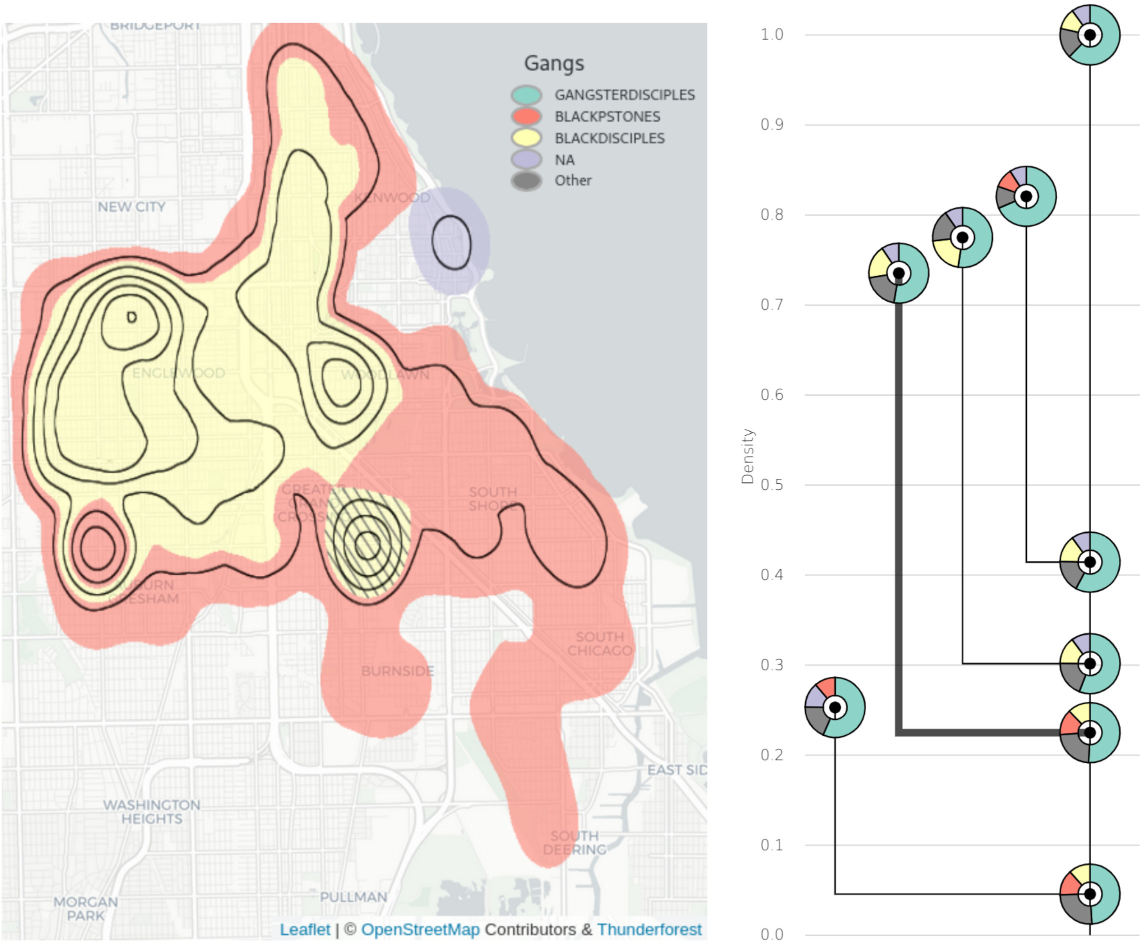


Figure 4.3: In This Gang Activity Analysis the Reference Field (Left) Corresponds to the Kernel Density Estimate of the *Gangster Disciples* (Green), Where Each Region Is Colored by the Second Most Reported Gang in That Region. As Shown in the Donut Tree (Right), the *Gangster Disciples* Are by Far the Most Reported Gang in All Their Hotspots. However, the Second Most Reported Gang Varies from Hotspot to Hotspot, Where the Donut Tree Clearly Shows Which Hotspots Are Contested by Which Gangs. For Instance, the Second Most Reported Gang in the Highlighted Hotspot (Thick Edge in the Tree and Hatched Area on the Map) Are the *Black Disciples* (Yellow). An Outlier Is the Purple Hotspot, Where the Second Most Reports Can Not Be Attributed to Any Specific Gang (*Not Available*) as There Are Many Original Input Reports in That Region That Have No Gang Label.

4.4.1 Gang Activity in Chicago

In this usage scenario we explore georeferenced reports of gang activity in Chicago, IL (USA) between 2011 and 2014. Here, we want to explore the distribution of gang activity with respect to global and local gang hotspots. First, we transform the point events to a set of scalar fields. For each gang, we compute a kernel density estimate (KDE) of their corresponding events in a preprocessing step. Let E denote the set of all n events, and E_i all events that are associated with gang i. Then each field f_i of the vector function f is defined as

$$f_i(x) = \frac{1}{nh^2} \sum_{e \in E_i} K\left(\frac{\|x - p(e)\|_2}{h}\right), \tag{4.2}$$

where the function p returns the latitude-longitude position of an event, K is the kernel function, and h is the spatial bandwidth. For each gang we use the same Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \tag{4.3}$$

and spatial bandwidth

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} \sigma n^{-\frac{1}{5}}, \tag{4.4}$$

which has been chosen according to Silverman's *Rule of Thumb* (Silverman, 1986) with σ being the standard deviation of all n events. This results in a set of multiple scalar fields, each of which provides the local topology of a single gang activity. Furthermore, these fields can be aggregated together to obtain a global topology of gang activity in the region. Analysis can be done at both the local and global level.

Next, we demonstrate results for two different reference fields. First, we examine the case where the analyst-defined reference field corresponds to a specific KDE of a single gang. Here we explore the following question: ”How active are other gangs in hotspots of one particular gang?”. This can serve to highlight regions with potential gang disputes.

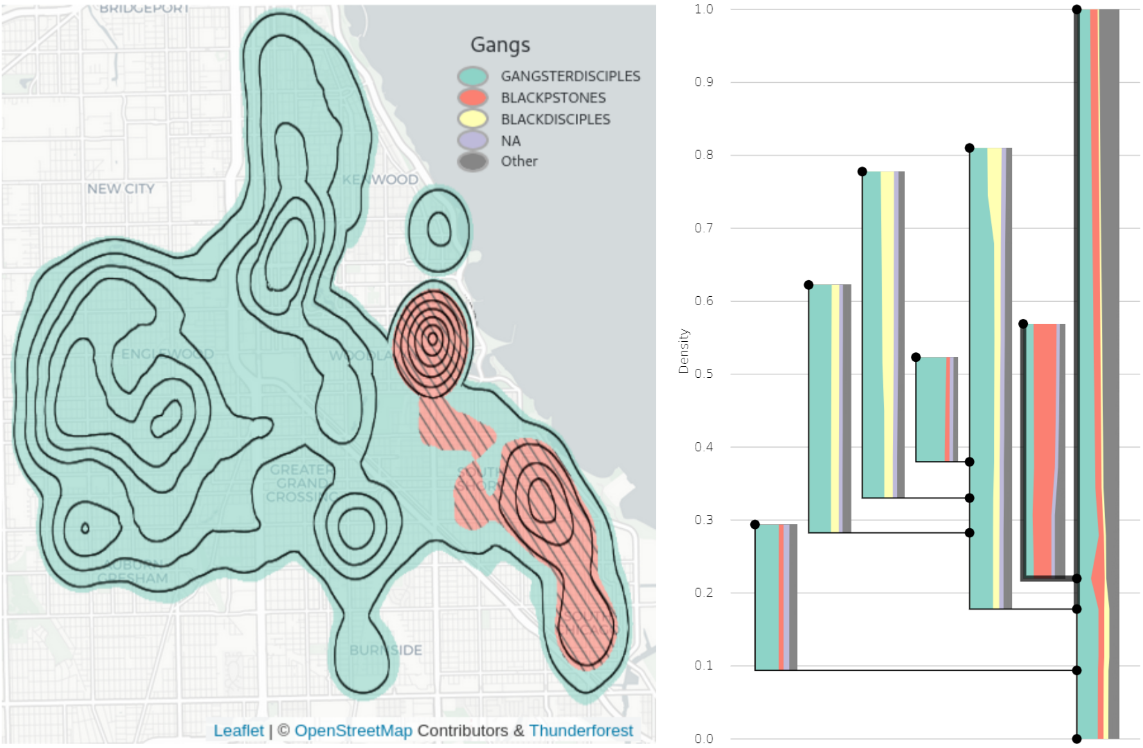


Figure 4.4: This Analysis Focuses on Reports That Have No Gang Label, I.E., The Reference Field That Corresponds to the Kernel Density Estimate of the Reports with the Label *Not Available*. In the Map Hotspots Are Colored by the Most Reported Gang in That Region (Left). The Stream Tree with Fixed Scaling (Right) Shows Which Gangs Are Frequently Reported in Areas with Many Unlabeled Reports. Not Surprisingly, in Many of These Regions the *Gangster Disciples* Have the Most Reports, but the Stream Tree Clearly Indicates Regions in Which Other Gangs Are Rarely Reported (Such as in the Left Most Edge) or Other Gangs Are Almost as Frequently Reported (Such as the Third Edge from the Left). The Highlighted Hotspot, However, Stands out since in This Hotspot the Amount of Reports of the *Black P Stones* Is Almost Four times the Number of Reports of the *Gangster Disciples*; And Therefore Making Them the Most Likely Gang to Be Associated with Unlabeled Reports in That Region. This Example Indicates That Our Approach Can Also Be Used for Probabilistic Relabeling.

In Figure 4.3, the right part is the donut tree extracted from gang data in Chicago, IL. The donut on each merge tree segmentation visualizes the top three contributions to the gang crime activity in the corresponding areas. The rest of the contributions are aggregated and represented as the gray portion of the donut view. We can observe that all the segments (hotspots) are dominated by Gangster Disciples (green). In these hotspots, we would like to know the activity of other gangs. To quickly get a glance at the most competing gang with the Gangster Disciples in these hotspots, the map view will color each hotspot with the second dominant gang. The left part of Figure 4.3 is the map view, where colors for each hotspot are the second dominant integral color in the donut tree branch, enabling the exploration of other gang activity relative to the reference gang's hotspots. We can see the Black P Stones (red) and Black Disciples (yellow) are the second most active gangs in these territories.

While using one gang distribution as the reference field can help explore how other activity is distributed in the reference gang's hotspot, our approach can also be used to explore unlabeled data. In the gang dataset, a large portion of the records remain unattributed to gangs. By setting the reference field to correspond to the ”Not Available” label, we can explore those hotspots to identify what other fields overlap, Figure 4.4. Here, the map hotspots are colored by the most reported gang in that region. We use a stream tree with fixed scaling to show which gangs are frequently reported in the reference field hotspot. As expected, the majority of the regions are primarily composed of teal; however, there are two regions where the dominant contributor is red. We believe that this approach could be used to create a probabilistic relabeling of the data as unlabeled events are likely drawn from the underlying related scalar fields.

Finally, we look at an analyst-defined reference field that corresponds to the sum of all KDEs. In Figure 4.1, we use the aggregate density field of all gangs as our

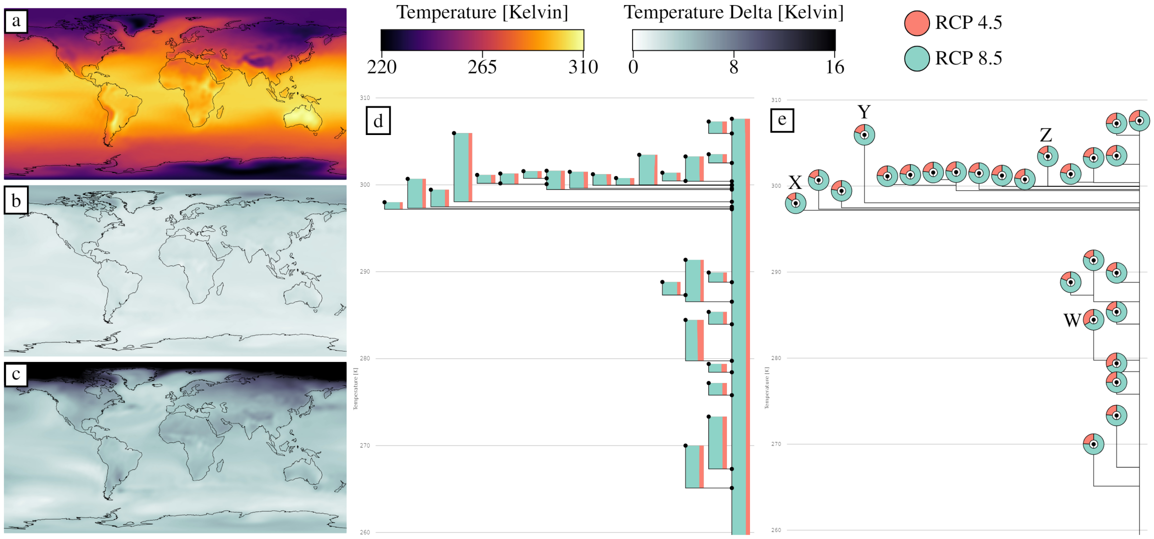


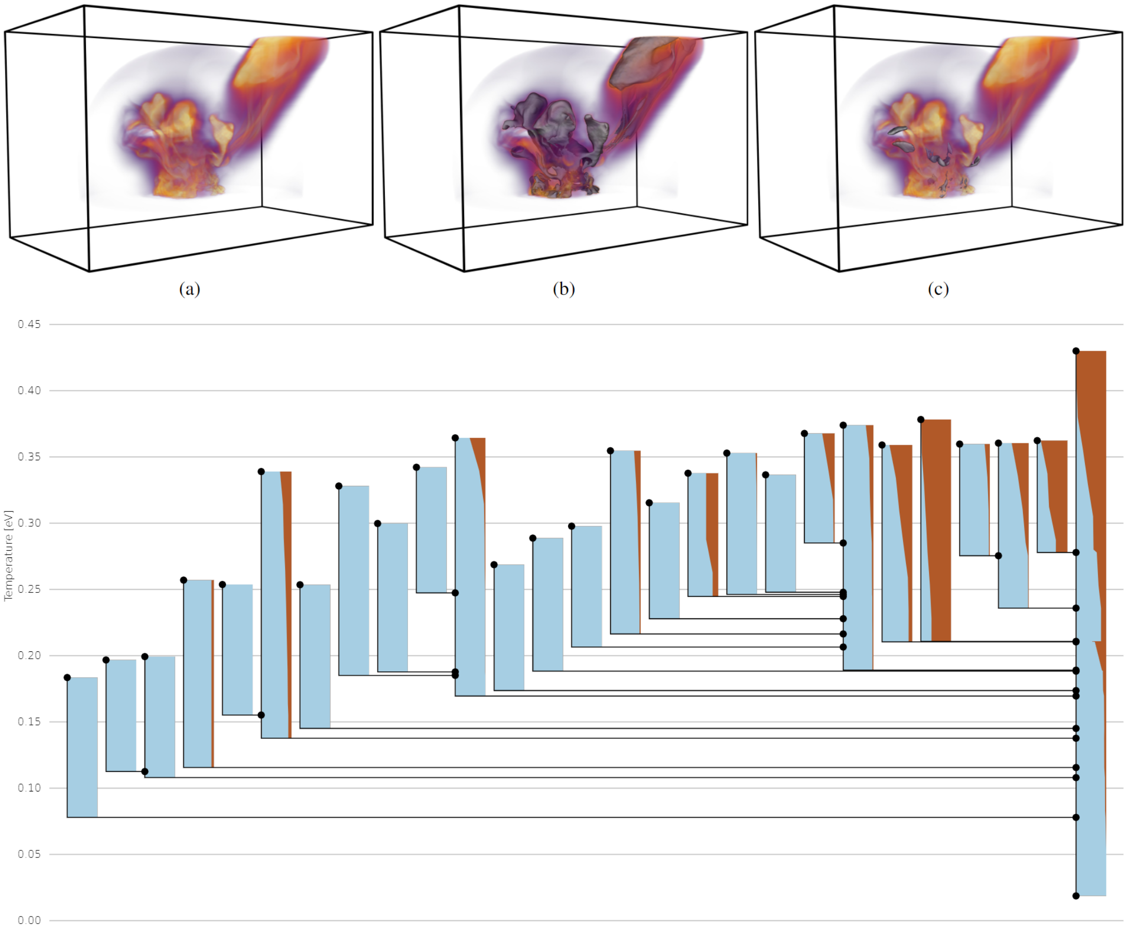
Figure 4.5: Analysis Results of the *Grand Ensemble* That Simulates Climate Change under Three Different RCP Scenarios: RCP 2.6, 4.5, and 8.5. Each Scenario Was Simulated 100 times, and the Three Scenarios Roughly Correspond to the Best, Intermediate, and Worst Case. First, We Compute for Each Scenario the Mean Surface Temperature (MST) for the Year 2100 by Aggregating the Corresponding Ensemble Members. The MST of the RCP 2.6 Scenario Is Shown in (A). Next, We Compute the Difference Fields Between the RCP 4.5 (B) and RCP 8.5 (C) Scenario with Respect to the RCP 2.6 MST Field. Then We Select the RCP 2.6 MST Field as a Reference Field, and the Two Difference Fields as the Integrands. The Corresponding Stream Tree (D) and Donut Tree (E) Show the Proportions Between the Temperature Increases Predicted by the RCP 4.5 and RCP 8.5 Scenarios in Superlevel Set Components (Segments) of the MST Field Predicted by the RCP 2.6 Scenario. Although the Stream Tree Makes It Easier to Follow Edges, It Is More Difficult to Judge the Proportions Between the Two Scenarios. The Donut Tree, on the Other Hand, Produces More Visual Clutter but Better Encodes the Ratios Between the Scenarios.

reference topology. As such, the merge tree provides the topological structure where branches are regional peaks of gang activity. Each point on the merge tree corresponds to a unique superlevel set component (commonly referred to as a hotspot when exploring georeferenced events). The merge tree is augmented by the stream graph (Figure 4.1 (Middle)), and the dominant components of the merge tree branches are immediately visible. Here, three distinct primary regions emerge, where the teal (Gangster Disciples), red (Black P Stones), and orange (Latin Kings) are the majority scalar field contributor. Here, we shade the aggregate field on the map based on the majority contributing field from the merge tree branch (Figure 4.1 (Left)). We can also observe varying levels of yellow (Black Disciples) across several branches, primarily in teal regions. According to Wikipedia (Wikipedia contributors, 2022), the Gangster Disciples and Black Disciples are bitter rivals, indicating that regions with the blue-yellow make up might be prone to violence, or these regions may even be territories that are in dispute. From the donut tree view (Figure 4.1 (Right)), we can quickly observe branches that have a higher blue/yellow distribution, and by selecting a branch, we can interactively filter hotspots on the map to further explore regions.

A traditional approach for geographic analysis would be to either overlay all the events on the map, leading to overplotting and occlusion while also making it extremely challenging to estimate the ratio of events in local areas. Another option is to use a series of small multiples to show the distribution of individual events, which requires the analyst to mentally integrate the images. What the stream tree and donut tree provides is a quick summary of how features from multiple scalar fields are distributed with respect to a reference field (the base topology). This enables a broad overview of the data distributions and can quickly guide analysts to locations where interesting interactions may be occurring.

4.4.2 Climate Ensemble

In this usage scenario we demonstrate how stream and donut trees can be used to explore climate ensembles. Here, we analyze the *Grand Ensemble* provided by the Max Planck Institute for Meteorology (Maher *et al.*, 2019). The *Grand Ensemble* includes simulations that model the evolution of climate change from the year 2005 until 2100 with respect to three different representative concentration pathway (RCP)


(a) (b) (c)

(d)

Figure 4.6: Results of the Asteroid Impact Example Where We Visualize the Proportions Between Water (Blue) and Asteroid (Brown) matter Inside Superlevel Set Components (Segments) of the Temperature Field. The Corresponding Stream Tree with Fixed Scaling (a) Clearly Indicates the Fraction Between Water and Asteroid Matter Inside the Individual Segments. The Stream Tree Makes It Possible to Quickly Identify Which Segments Consist Almost Entirely of Water, and Which Segments Contain Asteroid Matter. The Stream Tree Also Shows That as One Goes Further up in the Tree I.E., As the Corresponding Segments Contract Towards Their Respective Temperature Maxima the Proportion of Asteroid Matter Increases. This Indicates That Asteroid Fragments Are Close the Temperature Maxima, I.E., At the Core of These Segments. Utilizing the Stream Tree It Is Possible to Partition the Individual Segments into Two Groups: The Ones That Contain Some Asteroid Matter (B), and the Ones That Do Not Contain Asteroid Matter (C).

scenarios. RCPs model trajectories of greenhouse gas concentrations, which vary based on different assumptions such as being compliant to the Paris climate agreement or the availability of fossil fuels. These three different RCP scenarios are labeled RCP 2.6, 4.5, and 8.5, which can roughly be categorized as best, intermediate, and worst case scenarios, respectively. Each scenario includes multiple random variables, and to incorporate the effects of this internal variability the *Grand Ensemble* contains 100 simulations of each scenario. Each simulation contains a scalar field defined on the globe that corresponds to the surface temperature.

Our analysis focuses on the temperature differences of the three scenarios in the last simulated year: 2100. To this end, we compute for the last timestep the mean surface temperature (MST) field by aggregating all 100 simulations for each scenario. We also simplify the fields by a low persistence threshold to suppress noise. Next, we compute the absolute difference between the MST of RCP 2.6 and the MST fields of RCP 4.5 and 8.5, which are shown in Figure 4.5b and c. Then we select the MST field of RCP 2.6 as our reference field, and the two difference fields as the integrands in Eq. 4.1. The corresponding steam and donut trees are shown in Figure 4.5d and e. These trees show how much the RCP 4.5 and 8.5 scenarios differ in respect to hot regions of RCP 2.6. The trees indicate that overall RCP 8.5 predicts almost three times the temperature increase of RCP 4.5; a global trend that has already been shown in prior analysis of the RCP scenarios (Stocker, 2014; Maher *et al.*, 2019). The stream and donut trees, however, provide more detail as they partition the globe into segments that can be explored separately. The vertical location of the merge tree edges also sets these individual segments into context. Here, all edges above 295K correspond to segments around the equator, while the other edges correspond to segments closer to the poles. In the donut tree it is easy to see that the temperature increase in most segments around the equator exhibit similar proportions as the global trend. However, there exists three outliers (X, Y, and Z) where the donut charts indicate that in the corresponding segments the RCP 8.5 scenario predicts four to five times more temperature increase than the RCP 4.5 scenario. These outliers correspond to segments of east Africa and Argentina, which indicates that these areas will be most affected if the RCP 8.5 scenario comes to pass. Segments close to the poles also exhibit an outlier. The donut chart with label (W) corresponds to the area of the Caspian Sea, and although RCP 8.5 still predicts a more significant temperature increase than RCP 4.5, the increase is relatively small compared to the other segments.

4.4.3 Asteroid Ocean Impacts

In this usage scenario we examine one timestep of an asteroid ocean impact simulation ensemble that was made publically available for the 2018 scientific visualization contest (IEEE VIS, 2018; Patchett *et al.*, 2016). This ensamble contains simulation runs that model the impact of an asteroid in the ocean, where the ensemble members differ in the size of the asteroid, the impact angle, and the height of a potential air burst before impact. Here, we examine the impact scenario with no airburst, an impact angle of 45 degrees, and an asteroid diameter of 250 meters. Figure 4.6a shows a volume rendering of the temperature field at cycle time 28649; shortly after the impact. In addition to the temperature field (electron volt), the dataset also provides three scalar fields that record per cell the total mass density (grams per cubic meter), as well as the volume fraction of water and asteroid matter.

We aim to derive a visualization that encodes the matter composition of volume segments that exhibit extreme temperatures. To this end, we select the temperature field as our reference field, and in a preprocessing step we compute the absolute mass of water and asteroid per cell by multiplying the respective fraction fields with the mass density field. The two resulting mass fields are then used as the integrands in Eq. 4.1. The corresponding stream tree with fixed scaling is shown in Figure 4.6d. The stream tree clearly indicates the ratio between asteroid and water matter inside individual high temperature segments, as well as the evolution of these proportions as the temperature in these segments increases. All segments exhibit the same trend that if they contain some asteroid matter than the proportion of asteroid matter increases as the temperature in the corresponding segment increases. Moreover, it enables analysts to partition the segments based on the proportion of asteroid matter. Figure 4.6b shows in gray all segments that exhibit along the corresponding edge an asteroid-to-matter-ratio of at least 4-to-1. The remaining segments i.e., the segments that consist predominantly of water are shown in Figure 4.6c.

4.5 Conclusion

In this thesis, we presented stream trees, a novel visualization technique for multiple scalar fields that are defined over the same domain. By first determining a reference field whose superlevel set components correspond to features of interest, our approach numerically integrates the input fields over the area of these features, and then displays the individual integrals and their proportions by augmenting the merge tree of the reference field with stream graphs or donut charts. We demonstrated in three usage scenarios that this visualization effectively summarizes information of the input scalar fields, showcasing that our technique is flexible to problems from traditional information visualization and scientific visualization domains. Overall, stream trees serve as a compact view for exploring the relationship between feature distributions. This enables analysts to reason about drivers behind various topological constructs. Furthermore, this approach can serve as an alternative to overplotting and small multiples, providing a single view to help describe mutiple scalar field relationships.

We also see several avenues for future work. In future work, one could explore if the described concepts can be extended to other topological abstractions that yield domain segmentations, such as the contour tree or the Morse-Smale Complex. In addition to numerical integration, one could also explore other aggregation techniques, such as summarizing features at various contour bands. We also believe that this technique could serve as the basis for a variety of rich interactions to support quick filtering and exploration, and future work will explore a more robust interaction space, expanding upon our current branch selection mechanisms.

Chapter 5

CONCLUSION

With the large-scale instrumentation of geographical locations and digital devices, a massive amount of multiclass geospatial data is generated. In order to facilitate domain experts in analyzing such extensive geospatial data, I have to answer a series of challenging questions: How to assist analysis in exploring this data? How to aid insights discovery from this data? How to recommend domain experts make actionable decisions? Due to the size and dimensionality of the multiclass geospatial data, there are no trivial answers to these questions. In this thesis, I proposed and implemented various methods, strategies, and frameworks for multiclass geospatial data visualization.

During the exploration of multiclass geospatial data, I collaborated with domain experts to understand their key analysis requirements and explored various visualizations to reveal the underlying patterns effectively. To accomplish these requirements, firstly, I proposed and implemented a visual analytics framework for conservation planning. The framework is designed to help conservation experts to generate land purchase portfolios for the protection of species. The parcel (the minimum unit of land purchasing) contains multiple attributes, such as the distance to the existing protected areas, the richness of some special species, land cost, and others. The conservation experts have to analyze millions of parcels in the US, then generate the land purchasing portfolio with different priorities of concerns. To assist the conservation experts, I visualized the land attributes on the map separately by map layers. The land (geospatial data) contains multiple spatial elements, including lines, points, and polygons. Besides the massive size of geospatial data for analysis, these spatial elements are often associated with multiple attributes. In order to support the online exploratory of such geospatial data, I preprocessed the data into images and stored these images as different map layers for further analysis. In order to support the online exploratory of such geospatial data, I preprocessed the data into images and stored these images as different map layers for further analysis. Since different users may have distinct perspectives and preferences, the generated land purchasing portfolios have to be customized. Therefore, a visual analytics system that integrates human-in-the-loop is necessary to fulfill solicitations by domain experts. My proposed framework includes effective visualizations for multi-attributes geospatial data, multiple coordinated views for parcel analysis and portfolio generation, and intuitive brushing and linking interactions among multiple views. For instance, the filtering on parallel coordinates enables users to sift land based on the range of particular attributes. Even though users can narrow the investigation land to a reasonable area, the number of parcels for calculation is still too large to be analyzed. In order to speed up the process, a build-in optimization algorithm is provided to generate optimal solutions with user-specified constraints and objective functions. Furthermore, a comparison view is created to compare multiple solutions and assist users in finding the best land purchasing solution to protect the diversity of species. This work combines geospatial visualizations, multicriteria analysis, and automatic optimization that facilitate conservation planners and scientists exploring different land purchasing portfolios under a variety of constraints in real-time.

Besides visualizing each attribute as separate map layers, I further explored the patterns and relationships of hotspots for multiple attributes. For instance, sophisticated hotspot pattern analysis can reveal the high-risk areas for different crimes, discover the geographical association among high-risk areas, and highlight local high-risk areas. To satisfy these analysis requirements, I adopted the scalar field topology(SFT) method for hotspot identification. The standard method identifies hotspots as spatial regions that exceed a specified event occurrence density threshold. However, the method may obscure local peaks and highlight noise and outliers. The method also lacks knowledge about the relationship between hotspots and has difficulty determining hotspot boundaries. To mitigate these issues, I applied the SFT-based method that exploits feature characterizations to identify hotspots. I first calculate the kernel density estimation (KDE) for the scalar field and then apply a topology method to the scalar field to detect different types of hotspots. The proposed hotspot characterization methods include Superlevel Set Segmentation (SSS), Merge Tree Leaf Segmentation (MTLS), Merge Tree Crown Segmentation (MTCS), and Morse Complex Segmentation (MCS). Similar to the standard method of detecting hotspots, SSS can identify the hotspot as the specific area that exceeds a particular threshold of density. However, slightly changing the threshold can drastically change the geometry (boundary) and the number of extracted hotspots. As an improvement, MTLS can identify hotspots independent of a single threshold value. Furthermore, MTCS can extract hotspots based on local thresholds, which could detect the center of each hotspot. In addition, MCS can reveal the boundaries of hotspots. To aid the hotspot analysis using different SFT-based methods, I developed a visual analytics interface composed of four views: a geospatial map, a merge tree, a persistence diagram, and a persistence curve. The hotspots can be displayed on the map, and users can adjust the configurations on the other three views to compare hotspots generated in different settings and explore the geospatial relationships between hotspots. Overall, the preliminary advantage of the discussed SFT-based characterizations is their capability to represent hotspots at multiple resolutions, which makes it possible to rank, filter, and decompose hotspots based on their significance.

Since the SFT-based method explores the relationship and patterns of multiclass spatial data, a further step is to profile multiple attributes of the spatial data. For event datasets with multiple categories, not only the total event density is of interest, but also the ratio between the different event types in a given region. In my work, the combination of Merge Tree and Stream Graph explores the different categories in the hotspots. Expanding on my previous work exploring hotspots in a level-of-detail approach, I focus on the multiple categories for the hotspots. Oftentimes, a hotspot can be composed of multiple categories, but the contributions of each category are not straightforward. For example, I explored gang crime data and detected the high-risk areas to answer the following questions related to hotspot analysis: Which gang is the most active in the high-risk area? Are there any competing gangs such that the conflicts lead to violence? How is the gang active in different hotspots? I combined the stream graph with a merge tree to visualize the distributions of individual gangs and their ratios among multiple gangs for each hotspot. The stream tree also displays the comparison of gangs in different hotspots. As compensation for displaying the ratio of gangs, Pie charts and Donut charts are provided as alternatives to the stream graph on the merge tree. The stream tree provides a novel visualization to explain the relationship of the hotspots and explore the feature distributions.

In general, this thesis provides fruitful methods and strategies to visualize geospatial data, and these proposed methods overcome the challenges of exploring, learning, and using multiclass geospatial data. In addition, involving humans in the analytics process is necessary to satisfy users' different requirements. With the effective demonstration of geographical patterns, users can make more informed decisions. However, there are some limitations to the proposed methods. For example, comparing generated portfolios in the conservation planning system is not automatic. Users have to specify their objectives in the system since the built-in optimization algorithm is not customized. I adopted the topology-based method to discover the pattern and relationship of multi-attribute spatial data. While the foundation of the topology-based method is kernel density estimation (KDE), my SFT-based method takes the simple kernel function and bandwidth. Further studies are needed to understand the influence on hotspots using different kernel functions and bandwidths.

REFERENCES

"Washington park il murder/homicide rate 2011-2017 — macrotrends", URL: https://www.macrotrends.net/cities/us/il/washington-park/murder-homicide-rate-statistics, accessed on 06/08/2022 (2017).

"Most dangerous neighborhoods in chicago, il", URL: https://www.areavibes.com/chicago-il/most-dangerous-neighborhoods/, accessed on 06/08/2022 (2022).

Acevedo, M. A., J. A. Sefair, J. C. Smith, B. Reichert and R. J. Fletcher Jr., "Conservation under uncertainty: Optimal network protection strategies for worst-case disturbance events", Journal of Applied Ecology **52**, 6, 1588–1597 (2015).

Afzal, S., R. Maciejewski and D. S. Ebert, "Visual analytics decision support environment for epidemic modeling and response evaluation", in "Visual Analytics Science and Technology", pp. 191–200 (IEEE, 2011).

Ahrens, J., B. Geveci and C. Law, "ParaView: An End-User Tool for Large-Data Visualization", The Visualization Handbook pp. 717–731 (2005).

Ailon, N., M. Charikar and A. Newman, "Aggregating inconsistent information: ranking and clustering", Journal of the ACM (JACM) **55**, 5, 23 (2008).

Anderson, K., J. Anderson, S. Palande and B. Wang, "Topological Data Analysis of Functional MRI Connectivity in Time and Space Domains", in "MICCAI Workshop on Connectomics in NeuroImaging", (2018).

Ando, A., J. Camm, S. Polasky and A. Solow, "Species distributions, land values, and efficient conservation", Science **279**, 5359, 2126–2128 (1998).

Arthur, J. L., M. Hachey, K. Sahr, M. Huso and A. Kiester, "Finding all optimal solutions to the reserve site selection problem: formulation and computational analysis", Environmental and Ecological Statistics **4**, 2, 153–165 (1997).

Bailey, T. C. and A. C. Gatrell, *Interactive spatial data analysis* (Longman Scientific & Technical ; J. Wiley, Harlow Essex, England; New York, NY, 1995).

Ball, I. R., H. P. Possingham and M. Watts, "Marxan and relatives: software for spatial conservation prioritisation", Spatial conservation prioritisation: Quantitative methods and computational tools pp. 185–195 (2009).

Belle, E., N. Kingston, N. Burgess, T. Sandwith, N. Ali and K. MacKinnon, "Protected planet report 2018", UNEP-WCMC, IUCN and NGS: Cambridge UK, Gland Switzerland and Washington, DC, USA. URL https://livereport.protectedplanet.net/pdf/Protected_Planet_Report_2018.pdf (2018).

Bhatia, H., A. G. Gyulassy, V. Lordi, J. E. Pask, V. Pascucci and P.-T. Bremer., "TopoMS: Comprehensive Topological Exploration for Molecular and Condensed-Matter Systems", J. of Computational Chemistry (2018).

Bicknell, J. E., M. B. Collins, R. S. Pickles, N. P. McCann, C. R. Bernard, D. J. Fernandes, M. G. Miller, S. M. James, A. U. Williams, M. J. Struebig *et al.*, "Designing protected area networks that translate international conservation commitments into national action", Biological Conservation **214**, 168–175 (2017).

Bizimana, J. P. and G. Nduwayezu, "Spatio-temporal patterns of malaria incidence in rwanda", Transactions in GIS (2020).

Bock, A., H. Doraiswamy, A. Summers and C. T. Silva, "TopoAngler: Interactive Topology-Based Extraction of Fishes", IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS) (2018).

Borruso, G., "Network density estimation: a gis approach for analysing point patterns in a network space", Transactions in GIS **12**, 3, 377–402 (2008).

Bostock, M., V. Ogievetsky and J. Heer, "D^3 Data-Driven documents", IEEE transactions on visualization and computer graphics **17**, 12, 2301–2309, `https://d3js.org/` (2011).

Bremer, P.-T., A. Gruber, J. Bennett, A. Gyulassy, H. Kolla, J. Chen, and R. Grout, "Identifying turbulent structures through topological segmentation", Communications in Applied Mathematics and Computational Science **11**, 37–53 (2016).

Bremer, P.-T., G. Weber, J. Tierny, V. Pascucci, M. Day and J. Bell, "Interactive Exploration and Analysis of Large-Scale Simulations Using Topology-Based Data Segmentation", IEEE Transactions on Visualization and Computer Graphics (2011).

Bröring, A., A. Remke, C. Stasch, C. Autermann, M. Rieke and J. Möllers, "envirocar: A citizen science platform for analyzing and mapping crowd-sourced car sensor data", Transactions in GIS **19**, 3, 362–376 (2015).

Brownell, J. L., *The genesis of wildlife conservation in Montana*, Ph.D. thesis, Montana State University-Bozeman, College of Letters and Science (1987).

Bu, C., Q. Zhang, Q. Wang, J. Zhang, M. Sedlmair, O. Deussen and Y. Wang, "Sinestream: Improving the readability of streamgraphs by minimizing sine illusion effects", IEEE Transactions on Visualization and Computer Graphics **27**, 2, 1634–1643 (2020).

Butkiewicz, T., W. Dou, Z. Wartell, W. Ribarsky and R. Chang, "Multi-focused geospatial analysis using probes", IEEE Transactions on Visualization and Computer Graphics **14**, 6, 1165–1172 (2008).

Cabeza, M. and A. Moilanen, "Design of reserve networks and the persistence of biodiversity", Trends in Ecology and Evolution **16**, 5, 242–248 (2001).

Camm, J. D., S. K. Norman, S. Polasky and A. R. Solow, "Nature reserve site selection to maximize expected species covered", Operations Research **50**, 6, 946–955 (2002).

Carr, H. and D. Duke, “Joint contour nets: Computation and properties”, in “2013 IEEE Pacific Visualization Symposium (PacificVis)”, pp. 161–168 (2013).

Carr, H., Z. Geng, J. Tierny, A. Chattopadhyay and A. Knoll, “Fiber surfaces: Generalizing isosurfaces to bivariate data”, Computer Graphics Forum **34**, 3, 241–250 (2015).

Carr, H., J. Snoeyink and U. Axen, “Computing Contour Trees in All Dimensions”, Computational Geometry **24**, 2, 75 – 94 (2003).

Carr, H., J. Snoeyink and M. Van De Panne, “Flexible Isosurfaces: Simplifying and Displaying Scalar Topology using the Contour Tree”, Computational Geometry **43**, 1, 42–58 (2010).

Carr, H. A., J. Snoeyink and M. van de Panne, “Simplifying Flexible Isosurfaces Using Local Geometric Measures”, in “IEEE VIS”, (2004).

Cassol, V. J., E. S. Testa, C. R. Jung, M. Usman, P. Faloutsos, G. Berseth, M. Kapadia, N. I. Badler and S. R. Musse, “Evaluating and optimizing evacuation plans for crowd egress”, IEEE Computer Graphics and Applications **37**, 4, 60–71 (2017).

Chainey, S. and J. Ratcliffe, *GIS and crime mapping* (John Wiley & Sons, 2013).

Chainey, S., S. Reid and N. Stuart, *When is a Hotspot a Hotspot? A Procedure for Creating Statistically Robust Hotspot Maps of Crime* (Taylor & Francis, London, England, 2002).

Chainey, S., L. Tompson and S. Uhlig, “The utility of hotspot mapping for predicting spatial patterns of crime”, Security journal **21**, 1, 4–28 (2008).

Chainey, S. P., “Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime”, Bulletin of the Geographical Society of Liege **60**, 7–19 (2013).

Chattopadhyay, A., H. Carr, D. Duke and Z. Geng, “Extracting Jacobi Structures in Reeb Spaces”, in “EuroVis - Short Papers”, edited by N. Elmqvist, M. Hlawitschka and J. Kennedy (The Eurographics Association, 2014).

Chattopadhyay, A., H. Carr, D. Duke, Z. Geng and O. Saeki, “Multivariate topology simplification”, Computational Geometry **58**, 1–24 (2016).

Chen, Y., Z. Huang, T. Pei and Y. Liu, “Hispatialcluster: A novel high-performance software tool for clustering massive spatial points”, Transactions in GIS **22**, 5, 1275–1298 (2018).

Chilès, J.-P. and N. Desassis, “Fifty Years of Kriging”, in “Handbook of Mathematical Geosciences”, pp. 589–612 (Springer, Cham, 2018).

Chirima, G. J. and N. Owen-Smith, “Comparison of kernel density and local convex hull methods for assessing distribution ranges of large mammalian herbivores”, Transactions in GIS **21**, 2, 359–375 (2017).

Church, R. L., D. M. Stoms and F. W. Davis, “Reserve selection as a maximal covering location”, Biological conservation **76**, 103–112 (1996).

Cocks, K. and I. A. Baird, “Using mathematical programming to address the multiple reserve selection problem: an example from the eyre peninsula, south australia”, Biological Conservation **49**, 2, 113–130 (1989).

Cox, A. and L. Byron, “The ebb and flow of box office sales, 1986-2007”, The New York Times (2008).

Crickard III, P., *Leaflet. js essentials* (Packt Publishing Ltd, 2014), `https://leafletjs.com/`.

Cui, W., S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu and X. Tong, “Textflow: Towards better understanding of evolving topics in text”, IEEE transactions on visualization and computer graphics **17**, 12, 2412–2421 (2011).

Dasgupta, A., J. Poco, Y. Wei, R. Cook, E. Bertini and C. T. Silva, “Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison”, IEEE Transactions on Visualization and Computer Graphics **21**, 9, 996–1014 (2015).

De Floriani, L., U. Fugacci, F. Iuricich and P. Magillo, “Morse complexes for shape segmentation and homological analysis: discrete models and algorithms”, Comp. Grap. For. (2015).

de Queiroz Neto, J. F., E. M. dos Santos and C. A. Vidal, “Mskde-using marching squares to quickly make high quality crime hotspot maps”, in “2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)”, pp. 305–312 (IEEE, 2016).

de Queiroz Neto, J. F., E. Santos, C. A. Vidal and D. S. Ebert, “A visual analytics approach to facilitate crime hotspot analysis”, Computer Graphics Forum **39**, 3, 139–151 (2020).

Di Bartolomeo, M. and Y. Hu, “There is more to streamgraphs than movies: Better aesthetics via ordering and lassoing”, in “Computer Graphics Forum”, vol. 35, pp. 341–350 (Wiley Online Library, 2016).

Dilkina, B. and C. P. Gomes, “Solving connected subgraph problems in wildlife conservation”, in “International Conference on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming”, pp. 102–116 (Springer, 2010).

Dinerstein, E., C. Vynne, E. Sala, A. Joshi, S. Fernando, T. Lovejoy, J. Mayorga, D. Olson, G. Asner, J. Baillie, N. Burgess, K. Burkart, R. Noss, Y. Zhang, A. Baccini, T. Birch, N. Hahn, L. Joppa and E. Wikramanayake, “A global deal for nature: Guiding principles, milestones, and targets”, Science Advances **5**, 4, eaaw2869 (2019).

Dissanayake, S. T., H. Önal, J. D. Westervelt and H. E. Balbach, "Incorporating species relocation in reserve design models: An example from ft. benning ga", Ecological modelling **224**, 1, 65–75 (2012).

Dragicevic, S., "The potential of web-based gis", Journal of Geographical Systems **6**, 2, 79–81 (2004).

Edelsbrunner, H. and J. Harer, *Jacobi Sets*, pp. 37–57, London Mathematical Society Lecture Note Series (Cambridge University Press, 2004).

Edelsbrunner, H. and J. Harer, *Computational Topology: An Introduction* (American Mathematical Soc., 2010).

Edelsbrunner, H., J. Harer, V. Natarajan and V. Pascucci, "Local and global comparison of continuous functions", in "IEEE Visualization 2004", pp. 275–280 (2004).

Edelsbrunner, H., J. Harer and A. K. Patel, "Reeb spaces of piecewise linear mappings", in "Proceedings of the Twenty-Fourth Annual Symposium on Computational Geometry", SCG '08, p. 242–250 (Association for Computing Machinery, 2008).

Edelsbrunner, H., D. Letscher and A. Zomorodian, "Topological Persistence and Simplification", Discrete & Computational Geometry **28**, 4, 511–533 (2002).

Edelsbrunner, H., D. Morozov and V. Pascucci, "Persistence-Sensitive Simplification Functions on 2-Manifolds", in "Symp. on Comp. Geom.", (2006).

Edelsbrunner, H. and E. P. Mücke, "Simulation of Simplicity: A Technique to Cope with Degenerate Cases in Geometric Algorithms", ACM Transactions on Graphics (tog) **9**, 1, 66–104 (1990).

Ferreira, N., M. Lage, H. Doraiswamy, H. Vo, L. Wilson, H. Werner, M. Park and C. Silva, "Urbane: A 3d framework to support data driven decision making in urban development", in "Visual Analytics Science and Technology", pp. 97–104 (IEEE, 2015).

Forman, R., "A User's Guide to Discrete Morse Theory", Sém. Lothar. Combin **48**, 35pp (2002).

Gleicher, M., "Considerations for visualizing comparison", IEEE Transactions on Visualization and Computer Graphics **24**, 1, 413–423 (2018).

Gramacki, A., *Nonparametric Kernel Density Estimation and its Computational Aspects* (Springer, 2018).

Guan, X., B. Cheng, A. Song and H. Wu, "Modeling users' behavior for testing the performance of a web map tile service", Transactions in GIS **18**, 109–125 (2014).

Guenther, D., R. Alvarez-Boto, J. Contreras-Garcia, J.-P. Piquemal and J. Tierny, "Characterizing Molecular Interactions in Chemical Systems", IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS) (2014).

Gueunet, C., P. Fortin, J. Jomier and J. Tierny, “Task-Based Augmented Merge Trees with Fibonacci Heaps”, in “IEEE Symposium on Large Data Analysis and Visualization 2017”, (2017).

Gyulassy, A., P. Bremer, R. Grout, H. Kolla, J. Chen and V. Pascucci, “Stability of Dissipation Elements: A case study in combustion”, Computer Graphics Forum (2014).

Gyulassy, A., P. Bremer and V. Pascucci, “Shared-Memory Parallel Computation of Morse-Smale Complexes with Improved Accuracy”, IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS) (2018).

Gyulassy, A., M. A. Duchaineau, V. Natarajan, V. Pascucci, E. Bringa, A. Higginbotham and B. Hamann, “Topologically Clean Distance Fields”, IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS) (2007).

Gyulassy, A., A. Knoll, K. Lau, B. Wang, P. Bremer, M. Papka, L. A. Curtiss and V. Pascucci, “Interstitial and Interlayer Ion Diffusion Geometry Extraction in Graphitic Nanosphere Battery Materials”, IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS) (2015).

Havre, S., B. Hetzler and L. Nowell, “Themeriver: Visualizing theme changes over time”, in “IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings”, pp. 115–123 (IEEE, 2000).

Heidenreich, N.-B., A. Schindler and S. Sperlich, “Bandwidth Selection for Kernel Density Estimation: A Review of Fully Automatic Selectors”, AStA Advances in Statistical Analysis **97**, 4, 403–433 (2013).

Heine, C., H. Leitte, M. Hlawitschka, F. Iuricich, L. De Floriani, G. Scheuermann, H. Hagen and C. Garth, “A Survey of Topology-Based Methods in Visualization”, in “Computer Graphics Forum”, vol. 35, pp. 643–667 (Wiley Online Library, 2016).

Hodgson, J. A., C. D. Thomas, B. A. Wintle and A. Moilanen, “Climate change, connectivity and conservation decision making: back to basics”, Journal of Applied Ecology **46**, 5, 964–969 (2009).

Hu, Y., H. J. Miller and X. Li, “Detecting and analyzing mobility hotspots using surface networks”, Transactions in GIS **18**, 6, 911–935 (2014).

Huettenberger, L., C. Heine, H. Carr, G. Scheuermann and C. Garth, “Towards multifield scalar topology based on pareto optimality”, Computer Graphics Forum **32**, 3pt3, 341–350 (2013).

Huettenberger, L., C. Heine and C. Garth, “Decomposition and simplification of multivariate data using pareto sets”, IEEE Transactions on Visualization and Computer Graphics **20**, 12, 2684–2693 (2014).

IEEE VIS, “Scientific Visualization Contest 2018”, `http://sciviscontest2018.org/` (2018).

Jafari, N. and J. Hearne, “A new method to solve the fully connected reserve network design problem”, European Journal of Operational Research **231**, 1, 202–209 (2013).

Jenkins, C. N., “GIS layers of biodiversity data”, URL: `https://biodiversitymapping.org/wordpress/index.php/download/`, instituto de Pesquisas Ecológicas, Accessed on 10/08/2018 (2017).

Johansson, E., C. Gåhlin and A. Borg, “Crime Hotspots: An Evaluation of the KDE Spatial Mapping Technique”, in “2015 European Intelligence and Security Informatics Conference”, pp. 69–74 (IEEE, 2015).

Johansson, J., M. Cooper and M. Jern, “3-dimensional display for clustered multi-relational parallel coordinates”, in “Information Visualisation”, pp. 188–193 (IEEE, 2005).

Johnson, C. N., A. Balmford, B. W. Brook, J. C. Buettel, M. Galetti, L. Guangchun and J. M. Wilmshurst, “Biodiversity losses and conservation responses in the anthropocene”, Science **356**, 6335, 270–275 (2017).

Kasten, J., J. Reininghaus, I. Hotz and H. Hege, “Two-Dimensional Time-Dependent Vortex Regions based on the Acceleration Magnitude”, IEEE Transactions on Visualization and Computer Graphics (2011).

Kehrer, J., H. Piringer, W. Berger and M. E. Gröller, “A model for structure-based comparison of many categories in small-multiple displays”, IEEE Transactions on Visualization and Computer Graphics **19**, 12, 2287–2296 (2013).

Kirkpatrick, J., “An iterative method for establishing priorities for the selection of nature reserves: an example from tasmania”, Biological Conservation **25**, 2, 127–134 (1983).

Konev, A., J. Waser, B. Sadransky, D. Cornel, R. Perdigao, Z. Horváth and M. Groller, “Run watchers: Automatic simulation-based decision support in flood management”, IEEE Transactions on Visualization and Computer Graphics **20**, 12, 1873–1882 (2014).

Koutsias, N., P. Balatsos and K. Kalabokidis, “Fire occurrence zones: kernel density estimation of historical wildfire ignitions at the national level, greece”, Journal of Maps **10**, 4, 630–639 (2014).

Krige, D. G., *A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand*, Ph.D. thesis, University of the Witwatersrand (1951).

Kulldorff, M., “A Spatial Scan Statistic”, Communications in Statistics-Theory and methods **26**, 6, 1481–1496 (1997).

Lampe, O. and H. Hauser, “Interactive visualization of streaming data with kernel density estimation”, in “Pacific Visualization Symposium (PacificVis), 2011 IEEE”, pp. 171–178 (2011).

Laney, D., A. Mascarenhas, P. Miller, V. Pascucci *et al.*, "Understanding the Structure of the Turbulent Mixing Layer in Hydrodynamic Instabilities", IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS) (2006).

Law, P.-M., R. C. Basole and Y. Wu, "Duet: Helping data analysis novices conduct pairwise comparisons by minimal specification", IEEE Transactions on Visualization and Computer Graphics (2018).

Lehtomäki, J. and A. Moilanen, "Methods and workflow for spatial conservation prioritization using zonation", Environmental Modelling & Software **47**, 128–137 (2013).

Levine, N., "Crimestat: A spatial statistical program for the analysis of crime incidents", in "Encyclopedia of GIS", pp. 187–193 (Springer US, 2008).

Liadsky, D. and B. Ceh, "The interaction between individual, social and environmental factors and their influence on dietary intake among adults in toronto", Transactions in GIS **21**, 6, 1260–1279 (2017).

Lin, B. B., R. A. Fuller, R. Bush, K. J. Gaston and D. F. Shanahan, "Opportunity or orientation? who uses urban parks and why", PLoS one **9**, 1, e87422 (2014).

Lin, S., "Rank aggregation methods", Wiley Interdisciplinary Reviews: Computational Statistics **2**, 5, 555–570 (2010).

Lind, M., J. Johansson and M. Cooper, "Many-to-many relational parallel coordinates displays", in "Information Visualisation", pp. 25–31 (IEEE, 2009).

Lohfink, A.-P., F. Gartzky, F. Wetzels, L. Vollmer and C. Garth, "Time-varying fuzzy contour trees", in "2021 IEEE Visualization Conference (VIS)", (IEEE, 2021).

Lohfink, A.-P., F. Wetzels, J. Lukasczyk, G. H. Weber and C. Garth, "Fuzzy contour trees: Alignment and joint layout of multiple contour trees", Computer Graphics Forum **39**, 3, 343–355 (2020).

Loveland Tech., "LOVELAND: mapping every parcel on the planet", URL: `https://landgrid.com/reports/parcels`, (Accessed on 10/08/2018) (2018).

Luck, G. W., K. M. Chan and C. J. Klien, "Identifying spatial priorities for protecting ecosystem services", F1000Research **1** (2012).

Lukasczyk, J., G. Aldrich, M. Steptoe, G. Favelier, C. Gueunet, J. Tierny, R. Maciejewski, B. Hamann and H. Leitte, "Viscous Fingering: A Topological Visual Analytic Approach", Applied Mechanics and Materials (2017a).

Lukasczyk, J., C. Garth, R. Maciejewski and J. Tierny, "Localized Topological Simplification of Scalar Data", IEEE Transactions on Visualization and Computer Graphics (2020).

Lukasczyk, J., R. Maciejewski, C. Garth and H. Hagen, “Understanding Hotspots: A Topological Visual Analytics Approach”, in “Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems”, GIS ’15, pp. 36:1–36:10 (ACM, 2015).

Lukasczyk, J., G. Weber, R. Maciejewski, C. Garth and H. Leitte, “Nested Tracking Graphs”, in “Computer Graphics Forum”, vol. 36, pp. 12–22 (2017b).

Mace, G. M., K. Norris and A. H. Fitter, “Biodiversity and ecosystem services: a multilayered relationship”, Trends in Ecology and Evolution **27**, 1, 19–26 (2012).

Maciejewski, R., S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis and D. S. Ebert, “A Visual Analytics Approach to Understanding Spatiotemporal Hotspots”, IEEE Transactions on Visualization and Computer Graphics **16**, 2, 205–220 (2009).

Maher, N., S. Milinski, L. Suarez-Gutierrez, M. Botzet, M. Dobrynin, L. Kornblueh, J. Kröger, Y. Takano, R. Ghosh, C. Hedemann *et al.*, “The Max Planck Institute Grand Ensemble: enabling the exploration of climate system variability”, Journal of Advances in Modeling Earth Systems **11**, 7, 2050–2069 (2019).

Malik, A., 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 **20**, 12, 1863–1872 (2014).

Margules, C. R., A. Nicholls and R. Pressey, “Selecting networks of reserves to maximise biological diversity”, Biological conservation **43**, 1, 63–76 (1988).

Margules, C. R. and R. L. Pressey, “Systematic conservation planning”, Nature **405**, 6783, 243 (2000).

Masood, T. B., J. Budin, M. Falk, G. Favelier, C. Garth, C. Gueunet, P. Guillou, L. Hofmann, P. Hristov, A. Kamakshidasan *et al.*, “An Overview of the Topology ToolKit”, in “TopoInVis 2019-Topological Methods in Data Analysis and Visualization”, (2019).

Mcdonald, R. I., R. T. Forman, P. Kareiva, R. Neugarten, D. Salzer and J. Fisher, “Urban effects, distance, and protected areas in an urbanizing world”, Landscape and Urban Planning **93**, 1, 63–75 (2009).

Miettinen, K., *Nonlinear multiobjective optimization*, vol. 12 (Springer Science and Business Media, 2012).

Milnor, J. W., M. Spivak, R. Wells and R. Wells, *Morse Theory* (Princeton University Press, 1963).

Moilanen, A., K. A. Wilson and H. Possingham, *Spatial conservation prioritization: quantitative methods and computational tools* (Oxford University Press, 2009).

More, T. A., T. Stevens and P. G. Allen, “Valuation of urban parks”, Landscape and urban planning **15**, 1-2, 139–152 (1988).

Munzner, T., F. Guimbretiére, S. Tasiran, L. Zhang and Y. Zhou, “Treejuxtaposer: scalable tree comparison using focus+ context with guaranteed visibility”, ACM Transactions on Graphics **22**, 3, 453–462 (2003).

N, S. and V. Natarajan, “Simplification of jacobi sets”, in “Mathematics and Visualization”, pp. 91–102 (Springer Berlin Heidelberg, 2010).

Naidoo, R., A. Balmford, P. J. Ferraro, S. Polasky, T. H. Ricketts and M. Rouget, “Integrating economic costs into conservation planning”, Trends in Ecology and Evolution **21**, 12, 681–687 (2006).

Nakaya, T. and K. Yano, “Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics”, Transactions in GIS **14**, 3, 223–239 (2010).

Nalle, D. J., J. L. Arthur and J. Sessions, “Designing compact and contiguous reserve networks with a hybrid heuristic algorithm”, Forest Science **48**, 1, 59–68 (2002).

Olejniczak, M., A. S. P. Gomes and J. Tierny, “A Topological Data Analysis Perspective on Non-Covalent Interactions in Relativistic Calculations”, International Journal of Quantum Chemistry (2019).

Oliver, M. A. and R. Webster, “Kriging: a method of interpolation for geographical information systems”, International Journal of Geographical Information System **4**, 3, 313–332 (1990).

Önal, H. and R. A. Briers, “Incorporating spatial criteria in optimum reserve network selection”, Proceedings of the Royal Society of London B: Biological Sciences **269**, 1508, 2437–2441 (2002).

Önal, H. and R. A. Briers, “Optimal selection of a connected reserve network”, Operations Research **54**, 2, 379–388 (2006).

O’Sullivan, D. and D. Unwin, *Geographic Information Analysis* (Wiley, Hoboken, N.J, 2010), 2 edition edn.

Pajer, S., M. Streit, T. Torsney-Weir, F. Spechtenhauser, T. Möller and H. Piringer, “Weightlifter: Visual weight space exploration for multi-criteria decision making”, IEEE Transactions on Visualization and Computer Graphics **23**, 1, 611–620 (2017).

Parisi, T., *WebGL: up and running* (” O’Reilly Media, Inc.”, 2012).

Patchett, J., G. Gisler, B. Nouanesengsy, D. H. Rogers, G. Abram, F. Samsel, K. Tsai and T. Turton, “Visualization and Analysis of Threats from Asteroid Ocean Impacts”, Los Alamos National Laboratory (2016).

Pettit, C. J., C. M. Raymond, B. A. Bryan and H. Lewis, “Identifying strengths and weaknesses of landscape visualisation for effective communication of future alternatives”, Landscape and Urban Planning **100**, 3, 231–241 (2011).

Pimm, S. L., G. J. Russell, J. L. Gittleman and T. M. Brooks, “The future of biodiversity”, Science **269**, 5222, 347–350 (1995).

Polasky, S., J. D. Camm and B. Garber-Yonts, “Selecting biological reserves cost-effectively: an application to terrestrial vertebrate conservation in oregon”, Land Economics **77**, 1, 68–78 (2001).

Portman, M. E., “Visualization for planning and management of oceans and coasts”, Ocean and Coastal management **98**, 176–185 (2014).

Pressey, R., H. Possingham and J. Day, “Effectiveness of alternative heuristic algorithms for identifying indicative minimum requirements for conservation reserves”, Biological Conservation **80**, 2, 207–219 (1997).

Radil, S. M., C. Flint and G. E. Tita, “Spatializing social networks: Using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in los angeles”, Annals of the Association of American Geographers **100**, 2, 307–326 (2010).

Rao, M., G. Fan, J. Thomas, G. Cherian, V. Chudiwale and M. Awawdeh, “A web-based gis decision support system for managing and planning usda’s conservation reserve program (crp)”, Environmental Modelling and Software **22**, 9, 1270–1280 (2007).

Ratcliffe, J., “Crime mapping: spatial and temporal challenges”, in “Handbook of quantitative criminology”, pp. 5–24 (Springer, 2010).

Ratcliffe, J. H., “Aoristic signatures and the spatio-temporal analysis of high volume crime patterns”, Journal of quantitative criminology **18**, 1, 23–43 (2002).

Razip, A., A. Malik, S. Afzal, M. Potrawski, R. Maciejewski, Y. Jang, N. Elmqvist and D. Ebert, “A mobile visual analytics approach for law enforcement situation awareness”, in “IEEE Pacific Visualization Symposium (PacificVis)”, pp. 169–176 (2014).

Rinner, C., “A geographic visualization approach to multi-criteria evaluation of urban quality of life”, International Journal of Geographical Information Science **21**, 8, 907–919 (2007).

Robins, V., P. J. Wood and A. P. Sheppard, “Theory and Algorithms for Constructing Discrete Morse Complexes from Grayscale Digital Images”, IEEE Trans. Pattern Anal. Mach. Intell. (2011).

Rodrigues, A. S., H. R. Akcakaya, S. J. Andelman, M. I. Bakarr, L. Boitani, T. M. Brooks, J. S. Chanson, L. D. Fishpool, G. A. Da Fonseca, K. J. Gaston *et al.*, “Global gap analysis: priority regions for expanding the global protected-area network”, BioScience **54**, 12, 1092–1100 (2004).

Rosenbaum, R., J. Zhi and B. Hamann, “Progressive parallel coordinates”, in “Pacific Visualization Symposium”, pp. 25–32 (IEEE, 2012).

Sakurai, D., K. Ono, H. Carr, J. Nonaka and T. Kawanabe, “Flexible fiber surfaces: A reeb-free approach”, in “Mathematics and Visualization”, pp. 187–201 (Springer International Publishing, 2020).

Sala, O. E., F. S. Chapin, J. J. Armesto, E. Berlow, J. Bloomfield, R. Dirzo, E. Huber-Sanwald, L. F. Huenneke, R. B. Jackson, A. Kinzig *et al.*, “Global biodiversity scenarios for the year 2100”, Science **287**, 5459, 1770–1774 (2000).

Sanderson, E., M. Jaiteh, M. Levy, K. Redford, A. Wannebo and G. Woolmer, “The human footprint and the last of the wild”, BioScience **52**, 10, 891–904 (2003).

Sarkar, D., C. A. Chapman, L. Griffin and R. Sengupta, “Analyzing animal movement characteristics from location data”, Transactions in GIS **19**, 4, 516–534 (2015).

Sarkar, S., R. L. Pressey, D. P. Faith, C. R. Margules, T. Fuller, D. M. Stoms, A. Moffett, K. A. Wilson, K. J. Williams, P. H. Williams *et al.*, “Biodiversity conservation planning tools: present status and challenges for the future”, Annual Review of Environment and Resources **31** (2006).

Sawaragi, Y., H. NAKAYAMA and T. TANINO, *Theory of multiobjective optimization* (Elsevier, 1985).

Scheepens, R., H. van de Wetering and J. J. van Wijk, “Contour based visualization of vessel movement predictions”, International Journal of Geographical Information Science **28**, 5, 891–909 (2014).

Scheepens, R., N. Willems, H. van de Wetering and J. van Wijk, “Interactive density maps for moving objects”, IEEE Computer Graphics and Applications **32**, 1, 56–66 (2012).

Schneider, D., C. Heine, H. Carr and G. Scheuermann, “Interactive comparison of multifield scalar data based on largest contours”, Computer Aided Geometric Design **30**, 6, 521–528, foundations of Topological Analysis (2013).

Schneider, D., A. Wiebel, H. Carr, M. Hlawitschka and G. Scheuermann, “Interactive comparison of scalar fields based on largest contours with applications to flow visualization”, IEEE Transactions on Visualization and Computer Graphics **14**, 6, 1475–1482 (2008).

Sculley, D., “Rank aggregation for similar items”, in “Proceedings of the 2007 SIAM international conference on data mining”, pp. 587–592 (SIAM, 2007).

SEDAC, “Last of the wild (geographic), v2 (1995–2004)”, URL: `http://sedac.ciesin.columbia.edu/data/set/wildareas-v2-last-of-the-wild-geographic/data-download`, socioeconomic Data and Applications Center, Accessed on 10/08/2018 (2018).

Sefair, J. A., J. C. Smith, M. A. Acevedo and R. J. Fletcher Jr., "A defender-attacker model and algorithm for maximizing weighted expected hitting time with application to conservation planning", IISE Transactions **49**, 12, 1112–1128 (2017).

Shiode, S., "Street-level spatial scan statistic and stac for analysing street crime concentrations", Transactions in GIS **15**, 3, 365–383 (2011).

Shivashankar, N., P. Pranav, V. Natarajan, R. van de Weygaert, E. P. Bos and S. Rieder, "Felix: A Topology Based Framework for Visual Exploration of Cosmic Filaments", IEEE Transactions on Visualization and Computer Graphics (2016).

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, vol. 26 (CRC press, 1986).

Singh, G., F. Memoli and G. Carlsson, "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition", in "Eurographics Symposium on Point-Based Graphics", edited by M. Botsch, R. Pajarola, B. Chen and M. Zwicker (The Eurographics Association, 2007).

Soler, M., M. Petitfrere, G. Darche, M. Plainchault, B. Conche and J. Tierny, "Ranking Viscous Finger Simulations to an Acquired Ground Truth with Topology-Aware Matchings", in "IEEE Symposium on Large Data Analysis and Visualization", (2019).

Sorger, J., T. Ortner, C. Luksch, M. Schwärzler, E. Gröller and H. Piringer, "Litevis: integrated visualization for simulation-based decision support in lighting design", IEEE Transactions on Visualization and Computer Graphics **22**, 1, 290–299 (2016).

Sousbie, T., "The Persistent Cosmic Web and its Filamentary Structure: Theory and Implementations", Royal Astronomical Society (2011).

State of Montana, "Geographic information clearinghouse: Administrative boundaries", URL: `http://geoinfo.msl.mt.gov/Home/msdi/administrative_boundaries`, accessed on 10/08/2018 (2018).

Stocker, T., *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change* (Cambridge university press, 2014).

Stokstad, E., "Despite progress, biodiversity declines", Science **329**, 5997, 1272–1273 (2010).

Sturm, R. and D. Cohen, "Proximity to urban parks and mental health", The journal of mental health policy and economics **17**, 1, 19 (2014).

Sun, G., Y. Wu, S. Liu, T.-Q. Peng, J. J. Zhu and R. Liang, "Evoriver: Visual analysis of topic coopetition on social media", IEEE transactions on visualization and computer graphics **20**, 12, 1753–1762 (2014).

Tierny, J. and H. Carr, "Jacobi fiber surfaces for bivariate reeb space computation", IEEE Transactions on Visualization and Computer Graphics **23**, 1, 960–969 (2017).

Tierny, J., G. Favelier, J. A. Levine, C. Gueunet and M. Michaux, “The Topology ToolKit”, IEEE Transactions on Visualization and Computer Graphics **24**, 1, 832–842 (2017).

Toregas, C. and C. Revelle, “Binary logic solutions to a class of location problem”, Geographical Analysis **5**, 2, 145–155 (1973).

Travaini, A., J. Bustamante, A. Rodríguez, S. Zapata, D. Procopio, J. Pedrana and R. Martínez Peck, “An Integrated Framework to Map Animal Distributions in Large and Remote Regions”, Diversity and Distributions **13**, 3, 289–298 (2007).

Tress, B. and G. Tress, “Scenario visualisation for participatory landscape planning—a study from denmark”, Landscape and urban Planning **64**, 3, 161–178 (2003).

Turkay, C., 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 **20**, 12, 2033–2042 (2014).

Underhill, L., “Optimal and suboptimal reserve selection algorithms”, Biological Conservation **70**, 1, 85–87 (1994).

Unwin, D. J., “Gis, spatial analysis and spatial statistics”, Progress in Human Geography **20**, 4, 540–551 (1996).

USCB, “TIGER/Line® shapefiles: Roads”, URL: `https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2017&layergroup=Roads`, accessed on 5/24/2020 (2017).

USGS, “The national map”, URL: `http://prd-tnm.s3-website-us-west-2.amazonaws.com/?prefix=StagedProducts/Hydrography/NHD/State/HighResolution/Shape/`, united States Geological Service, Accessed on 10/08/2018 (2018a).

USGS, “Protected areas database of the united states (PAD-US) data download”, URL: `https://gapanalysis.usgs.gov/padus/data/download/`, united States Geological Service, Accessed on 10/08/2018 (2018b).

Watson, J. E. and O. Venter, “Ecology: a global plan for nature conservation”, Nature **550**, 7674, 48–49 (2017).

Weng, D., R. Chen, Z. Deng, F. Wu, J. Chen and Y. Wu, “Srvis: Towards better spatial integration in ranking visualization”, IEEE Transactions on Visualization and Computer Graphics (2018a).

Weng, D., H. Zhu, J. Bao, Y. Zheng and Y. Wu, “Homefinder revisited: finding ideal homes with reachability-centric multi-criteria decision making”, in “Proceedings of the CHI Conference on Human Factors in Computing Systems”, p. 247 (ACM, 2018b).

White, D., A. Wutich, K. Larson, P. Gober, T. Lant and C. Senneville, "Credibility, salience, and legitimacy of boundary objects: water managers' assessment of a simulation model in an immersive decision theater", Science and Public Policy **37**, 3, 219–232 (2010).

Wikipedia contributors, "Gangster disciples — Wikipedia, the free encyclopedia", URL `https://en.wikipedia.org/w/index.php?title=Gangster_Disciples&oldid=1079092796`, [Online; accessed 31-March-2022] (2022).

Williams, J. and C. ReVelle, "A 0–1 programming approach to delineating protected reserves", Environment and Planning B: Planning and Design **23**, 5, 607–624 (1996).

Williams, J. C., C. S. ReVelle and S. A. Levin, "Spatial attributes and reserve design models: a review", Environmental Modeling and Assessment **10**, 3, 163–181 (2005).

Xie, C., W. Zhong and K. Mueller, "A visual analytics approach for categorical joint distribution reconstruction from marginal projections", IEEE Transactions on Visualization and Computer Graphics **23**, 1, 51–60 (2017).

Xu, P., Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. Zhu and H. Qu, "Visual analysis of topic competition on social media", IEEE transactions on visualization and computer graphics **19**, 12, 2012–2021 (2013).

Zhang, H., X. Zhou, G. Tang, L. Xiong and K. Dong, "Mining spatial patterns of food culture in china using restaurant poi data", Transactions in GIS (2020).

Zhang, R., Y. Lu, K. Adams, J. A. Sefair, H. Mellin, M. A. Acevedo and R. Maciejewski, "A visual analytics framework for conservation planning optimization", Environmental Modelling and Software **145**, 105178 (2021a).

Zhang, R., J. Lukasczyk, F. Wang, D. Ebert, P. Shakarian, E. A. Mack and R. Maciejewski, "Exploring geographic hotspots using topological data analysis", Transactions in GIS (2021b).

Zhou, Q., "Comparative study of approaches to delineating built-up areas using road network data", Transactions in GIS **19**, 6, 848–876 (2015).