# FLOCKVIZ: A VISUALIZATION TECHNIQUE TO FACILITATE MULTI-DIMENSIONAL ANALYTICS OF SPATIO-TEMPORAL CLUSTER DATA

MOHAMMAD ZAHID HOSSAIN

*A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfilment of the requirements of the degree of*

MASTER OF SCIENCE

Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada

## ABSTRACT

Visual analytics of large amounts of spatio-temporal data is challenging due to the overlap and clutter from movements of multiple objects. A common approach for analyzing such data is to consider how groups of items cluster and move together in space and time. However, most methods for showing Spatio-temporal Cluster (STC) properties, concentrate on a few dimensions of the cluster (e.g. the cluster movement direction or cluster density) and many other properties are not represented. Furthermore, while representing multiple attributes of clusters in a single view existing methods fail to preserve the original shape of the cluster or distort the actual spatial covering of the dataset. In this thesis, I propose a simple yet effective visualization, FlockViz, for showing multiple STC data dimensions in a single view by preserving the original cluster shape. To evaluate this method I develop a framework for categorizing the wide range of tasks involved in analyzing STCs. I conclude this work through a controlled user study comparing the performance of FlockViz with alternative visualization techniques that aid with cluster-based analytic tasks. Finally the exploration capability of FlockViz is demonstrated in some real life data sets such as fish movement, caribou movement, eagle migration, and hurricane movement. The results of the user studies and use cases confirm the advantage and novelty of the novel FlockViz design for visual analytic tasks.

# DECLARATION

The following images have been taken from different research papers and web sources. I have cited them in this thesis and got permission from the authors to use.

Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, Figure 7, Figure 9, Figure 67, Figure 69

*Winnipeg, April 2014*

Mohammad Zahid Hossain

# ACKNOWLEDGMENTS

First and foremost, I would like to thank Dr. Pourang Irani for his advice and for being a ceaseless motivator for last two years and four months. He has always encouraged me to work harder and keep an open mind. This thesis would not have been possible without his supervision and guidance.

I would like to thank my committee members, Dr. Ekram Hossain and Dr. Yang Wang, for their time, support and helpful comments.

I would like to express my gratuade to Dr. Irani, the Faculty of Graduate Studies, the Faculty of Science and the Department of Computer Science for providing scholarships to pursue my master's study.

I am very grateful to my fellow lab mates in the HCI lab who always support me in various ways. Dr. Amir Hossein Meghdadi, who is a post-doctoral fellow gets a special mention for putting so much time and effort into proof reading my thesis. I would also like to thank Khalad Hasan, Fereshteh Amini, Barrett Ens, and all other lab mates for their support, ideas and help. I am very grateful to Dr. Paul J. Blanchfield and his Student David Callaghan from Biological Sciences department and Dr. Micheline Manseau from NRI department of University of Manitoba for providing me different datasets to use in my thesis.

I would like to thank my loving, supportive and encouraging wife Dilara Alo for her faithful support throughout my master's study.

Last but not least, I would like to thank my family for their continuous support and confidence in me. Thanks to my brother and sisters for their inspiration and endless love. I am forever indebted to my parents. Their strong inspiration and unconditional support helped me to come to this level. I am what I am only due to their efforts.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES



# ACRONYMS

| | |
|---|---|
| STC | Spatio-temporal Cluster |
| CAM | Comparative Aggregation among Multiple Clusters |
| CAS | Comparative Aggregation within a Single Cluster |
| DAS | Distinct Aggregation within a Single Cluster |
| ANOVA | Analysis of Variance |
| BI | Basic Information from Data Files |
| DI | Derived Information using Data from Files |
| GI | Generated Information by Users |
| NRI | Natural Resource Institute |

# 1

# INTRODUCTION

Spatio-temporal data is highly common and is a critical information source for analyzing on objects' movement behaviour. When the volume of the data becomes large we need to summarize the data using various aggregation techniques. Clustering is one of the most common aggregation methods which helps analysts perform high level analytic and comparison tasks. To efficiently use clusters and to identify important patterns and trends in the aggregated data novel spatio-temporal cluster visualization techniques are needed. These should ideally show and allow the comparison of multiple information types in a single view. Current spatio-temporal visualization methods primarily focus on a single value representation and do not explore the benefit of displaying multiple data values at once. In this thesis I propose a novel spatio-temporal cluster visualization technique, FlockViz, to show multi-dimensional or multi-value properties of spatio-temporal cluster in a single view. I demonstrate its value in different application scenarios.

## 1.1 MOTIVATION

Spatio-temporal data sources such as sensors are obtained from multiple physical locations, that record data over time and space,

and possibly change of positions over time. One major challenge is the explosive growth of sensor deployments that monitor all forms of activity, including environmental conditions, road traffic, animal migration, health care procedures, security services and equipment stability due to the low-cost and low-power requirements of recent sensor technology. These sensors generate multi-dimensional data values such as vehicle type, speed, temperature and physical condition along with spatio-temporal information. These multi-dimensional data are important for doing high level analysis like city planning.


Figure 1: (a) A large traffic jam could be mitigated with proper visual analytic tools that cluster the movement data and show multi-dimensional attributes. (b) Animal migration generates large spatio-temporal datasets of interest to biologists, climate scientists and natural habitat planning. Image source [12, 56].

As an example, Figure 1 shows some real life scenarios where a huge traffic jam (a) is shown and the another (b) shows the migration of penguins in a geographic region. The traffic jam could be due to a lack of traffic planning in particular over a specific time period of the day. For better traffic planning we need sufficient information of vehicle movement at different road intersections. As the volume of

the vehicle movement data is large one solution could be to build clusters of vehicles for different regions and traffic intersections. Each cluster can represent aggregate measures (i.e, number of buses, number of cars, or number of vehicles during the day along with the purpose of travel, etc.) of different movement properties. Once we collect the statistics and present them visually we can control vehicle movement based on their multi-dimensional attributes and the importance to have them on the road during a particular time period. As another example, penguin migration data contains additional dimensions such as their origin, their genetic properties and their physical properties. These could be of interest to biological analysts. Therefore, instead of showing thousands of penguin movement an appropriate cluster visualization with a good representation of the multi-dimensional information could help analysts explore the dataset more efficiently.

Current analytical tools for activity monitoring lack the ability to show multi-dimensional properties over a visualization of clustered movement data. Advanced visualization techniques with filtering and animation capabilities also suffer from occlusion problems when many space-time trajectories are shown simultaneously. Finding ways to harness the potentially useful information that lies hidden in these large data repositories and turning these into knowledge and action is of major concern to all stakeholders of sensor network data, including governments, corporations, end-users and the general public.

## 1.2 PROBLEM STATEMENT

Summarizing and aggregating large datasets plays an important role in numerous applications, including identifying general traffic behaviour, analyzing and forecasting regional weather conditions and observing animal migration. To identify patterns and perform comparative analysis on groups of objects in such applications, analysts seek for clusters [33, 4, 19]. Such clusters aggregate values on several key dimensions and help analysts focus on the high level analysis. Current visualizations that show the aggregated values of the various properties of the clusters fail to show multiple properties simultaneously such as the cluster's shape, aggregate movement direction, and aggregate measures [33]. Efficient visualization of spatio-temporal cluster (STC) properties has yet to benefit from displaying multiple properties in a single view. However, keeping the original cluster shape is important for perceiving spatial position of consisted objects. As an example, in Figure 1(a) without showing the exact cluster shape or boundary for each road intersection we cannot identify the actual physical locations of offices, parking, and markets within that cluster that cause a large traffic jam. In [9, 14], authors showed that the level of details of cluster such as original shape, density map etc. are helpful to label cluster regions and detecting outliers. Most of the current methods also fail to meet this requirement.

Applying current solutions for multi-dimensional information visualization such as parallel [27] and star coordinates [30] for groups of objects or clusters is challenging and even more complicated

considering the spatial and temporal constraints of such data [58]. Similarly, existing STC visualization techniques [4, 5, 19] using directional arrows or bar charts fail to represent the original cluster shape and are limited to a few cluster properties (e.g. density, and aggregate measures of object type) at a time. These methods are also limited to a specific problem space such as traffic congestion since the visualizations were designed and tested with a few tasks within that problem domain. As a result these methods are unable to address the more generic forms of tasks such as comparative analysis of multiple properties of clusters.

## 1.3 RESEARCH QUESTIONS

To design an efficient visual platform for aggregated spatio-temporal datasets, major challenges include:

- showing multiple aggregate properties in a single view;
- gaining a broad understanding of what tasks are necessary on these clusters and how novel visualizations can handle these tasks;
- supporting analysis at multiple scales of data domain and across multiple platforms; and
- providing visualization methods that end-users can easily interact with.

I have focused my goals on the following research questions:

- What are the most common tasks and queries performed on spatio-temporal datasets? Can we build a basic framework of task categories for analyzing spatio-temporal clusters?
- What properties of spatio-temporal clusters are important for generic analytic tasks?
- How to represent spatio-temporal cluster information for high level analytics without distorting the cluster's shape?
- How many cluster properties can be efficiently visualized in a single view by preserving the original cluster shape? What is the limit on the number of variables or dimension that can be viewed simultaneously?
- How can interactive techniques assist in summarizing data sets and for exploring and examining novel visualizations of spatio-temporal content?

## 1.4 PROPOSED SOLUTIONS

To explore these research questions in this thesis, I design, implement and evaluate a novel visualization technique called FlockViz. FlockViz captures the multiple Spatio-Temporal Cluster properties including the:

- physical and actual boundary of the cluster;
- internal distribution of objects within a cluster;
- total number of objects in the cluster;

- temporal information of cluster formation;
- aggregate measures of different dimensional values;
- aggregate movement direction of the cluster; and
- quantitative value of any of the above cluster properties.

The main strategy of this visual design is to add varying components to the original cluster or flock by keeping its shape intact. This visualization aims at embedding these properties in a single view without sacrificing spatial and temporal information. An interactive exploration option in my system can further help analysts to highlight a specific property of a cluster, scale each part of the visualization separately, enable/disable any flock element, and apply rich filtering options to generate and visualize desired clusters. This technique can also be applied to a wide range of spatio-temporal datasets to solve various analytic tasks.

To assess FlockViz, I developed a set of analytic tasks that are common for doing cluster-based analysis. These include a combination of available data parameters (i.e. Space - S, Time - T, Attribute - A). By making any one or two parameters unknown at any given time, I generated 18 possible tasks under 3 high-level categories. I evaluated my visualization technique with these tasks using two types of datasets (consisting of low and high density) using case studies and experiments. In the experiment I compared FlockViz against Geotime [31] (a traditional STC visualization technique widely used in commercial applications) in terms of task completion time and error rate for doing those tasks. The reason for selecting Geotime as a

baseline tool, in addition to its widespread use, is for its unrestricted ability to visualize many generic datasets. The experimental results show that for most of the tasks, FlockViz is significantly faster and more accurate than the traditional way of visualizing STC properties. Finally I evaluated my proposed design by asking a number of domain experts in spatio-temporal analysis of caribou and fish movement data (these include Dr. Paul J. Blanchfield (a research scientists at the Department of Fisheries and Ocean), his MSc. student, David Callaghan and Dr. Micheline Manseau, an Associate professor from the Natural Resources Institute department, at the University of Manitoba). They participated in my case studies where they used FlockViz to analyze objects' group movement and visualize the cluster properties. We concluded from the results of the case studies that FlockViz has the capabilities to provide solution to various research and analytic questions related to spatio-temporal data which are not possible to answer using traditional visualization methods. They explored my proposed visualization method to solve some of the research questions they are currently interested in. They found this system very helpful and interesting to analyze clusters' properties.

## 1.5 CONTRIBUTIONS

The main contributions of this thesis are listed as follows:

- Developing a new visualization technique (FlockViz) for visualizing multiple properties of spatio-temporal clusters (STC) while preserving the shape of clusters

- Defining a taxonomy of tasks for analyzing STC properties
- Evaluation of FlockViz through case studies and empirical tests
- Exploring possible application scenarios for FlockViz.

# 2

# RELATED WORK

The topic of Spatio-Temporal Cluster (STC) attribute visualization is inspired by work on traditional spatio-temporal data visualization, clustering methods, multi-dimensional data visualization, and spatio-temporal data analysis. However, none of this existing work specifically focuses on effective visualization of multiple properties of STC in a single view. To understand the need of spatio-temporal clustering and to explore basic elements of STC's approach in the literature, I review the following topics.

## 2.1 SPATIO-TEMPORAL DATA VISUALIZATION

Visualization of spatio-temporal data is generally grouped under two camps: one in which the space dimension of the data can either be "fixed and stationary" and the other where the space attribute is "dynamic and changing" (i.e. movement data) over time. However, for particular data domains, there could also be additional attribute information attached to the temporal and spatial properties of the data [65, 52]. The primary method of visualizing movement traces is to use a 2D map. Researchers have identified methods for effectively representing time so that it can be nicely integrated with the 2D space information [45] as shown in the Figure 2(a). Turdukulov and

Kraak [61] indicate that there are four main types of representations in 2D: Single 2D map, multiple 2D maps and linked views, map animation, and 2D display of abstract spatial information.

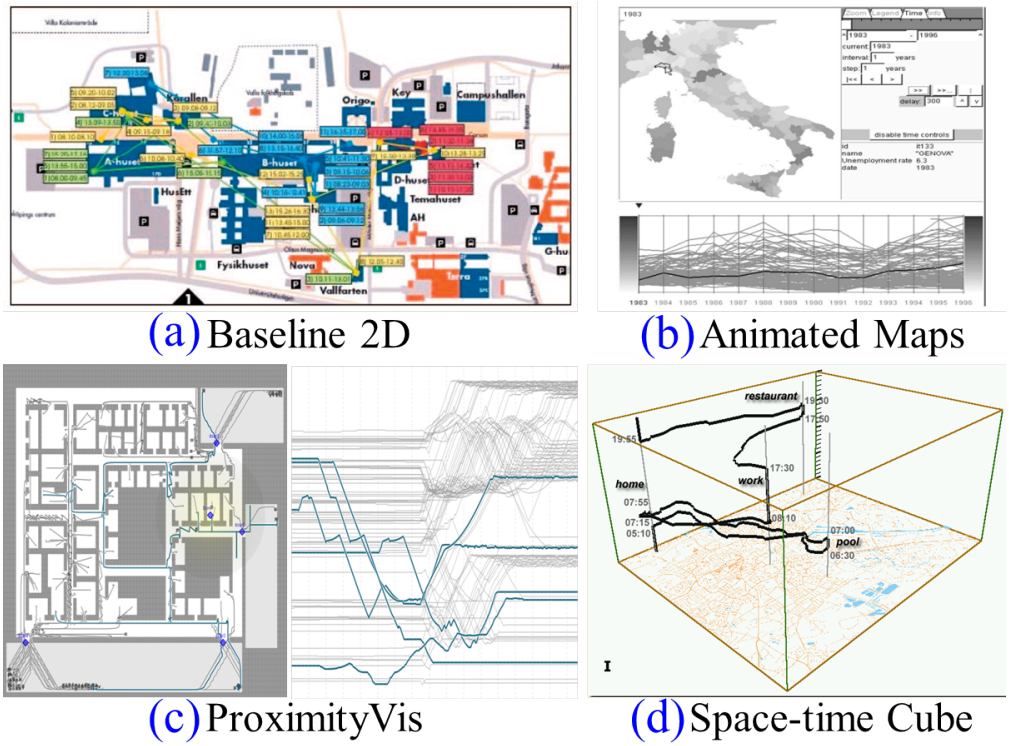



Figure 2: (a) In Baseline 2D [36] technique temporal information is embedded on top of each trajectory point. (b) In 2D map animation [25] the temporal information is linked to the bottom of the 2D map as a time slider and the trajectories in the map change with the animation of this widget. (c) In proximity-based 2D [15] view the position of objects are represented as a distance from a reference point (horizontal axis) and a line graph shows the change of that distance (vertical axis) in course of time (horizontal axis). (d) Space-time cube [24] method is used to show temporal information along z-axis while keeping the spatial position in 2D map.

To visualize the time as well as other dimensional attributes of spatio-temporal data (i.e. movement direction, objects' status etc.) tags, arrows and lines are usually used as visual cues on top of a single 2D map [62, 46, 68]. Different visual cues are used to repre-

sent additional attributes. For example, color can be used to show density of trajectories. A series of static 2D maps (also known as *small multiples*) can be used to show trajectories for different timestamps [60]. Ivanov *et al.* [28] describe a visualization system with a separate timeline view for temporal information which is linked to the 2D floor plan with embedded traces for showing movement. Similarly, in [4, 44, 57] linked views are used to visualize temporal data and spatial information in terms of a 2D map with trajectory overlays. Further extension of this approach results in animated 2D maps [26, 53, 25, 41] as shown in Figure 2(b). All of the above 2D approaches, keep the original spatial structure intact. Therefore, incorporation of extra attributes and time in the visualization makes them cluttered. There is, however, another group of 2D visualization methods proposed in the literature, which exploits abstract space representations. For instance, authors in [15] use the line graph metaphor to represent time as an abstract dimension. The result is a proximity-based visualization of movement traces in which the spatial relationships (e.g. the distance between objects) is preserved as shown in Figure 2(c).

Adding the third axis to represent time produces an alternative group of 3D visualizations which combine space and time in a single display. Hagerstraand [24] first proposed a visualization technique called the *space-time cube* (Figure 2(d)) where spatial and temporal information are shown by drawing trajectory points inside a cube. This idea has been expanded by other researchers in the field [34, 16]. A potential problem with this approach is occlusion if many trajectories are involved. To facilitate manipulation and perception of

information, the space-time cube has been augmented with interactive techniques [20] to rotate, zoom and translate data for better visualization. In [35] an advanced version of the space-time cube is presented where interactive time line, highlighting and linking of maps were incorporated. The enhanced version of this design that supports many of these features has been turned into a commercial software called GeoTime [31, 17]. Another drawback of the space-time cube approach, besides occlusion, is a distortion of both space and time due to projection, that makes it hard to perceive depth. Although 3D representation of movement data has been introduced, much research is being devoted to finding suitable forms of representing this complex dataset. However when the size of the dataset is out of control researchers seek methods to simplify the dataset and try to represent an aggregated form of visualization instead of thousand of trajectories. For such a purpose the concept of clustering traces plays a vital role. Figure 2 shows some examples of the state of the art 2D and 3D visualization techniques for spatio-temporal data.

## 2.2 SPATIO-TEMPORAL CLUSTERING ALGORITHMS

Clusters in general can be built based on different grouping models. The clustering method depends on the task at hand. There exist many clustering models but the most widely used models are hierarchical clustering, centroid based clustering, distribution based clustering and density based clustering.

- Hierarchical clustering [29] is divided into the top-down (or divisive) and bottom-up (or agglomerative) approach. The drawback of this clustering method is that it cannot provide a unique partition in datasets and gives a hierarchical structure, instead.
- K-means [39], k-medoids [32] within the group of centroid based clustering uses a predefined number of clusters.
- Distribution based clustering [69] is highly dependent on statistical measures and faces problem with over fitting data points.
- Density based clustering [18] is appropriate for movement data as it provides clusters based on spatial proximity.

Figure 3 shows an example dataset and corresponding clusters after applying the above clustering techniques. In all of these images each color represents a distinct cluster. It is clear from this example that the density-based approach clearly separates the clusters and hence provides space for cluster visualizations (with few misclassified data points). However, the method does not work well in the presence of measurement noise [7].

There are several types of clustering algorithms which support spatio-temporal data [18, 7]. Density based clustering is widely used for generating clusters of spatio-temporal data points. Among these methods DBSCAN [18] is designed to discover arbitrary-shaped clusters and distinguish noise points. DBSCAN accepts user-defined distance value, a radius value, and the number of minimum points within the radius. Starting from any arbitrary point, these criteria are checked at each step to identify neighboring points and once

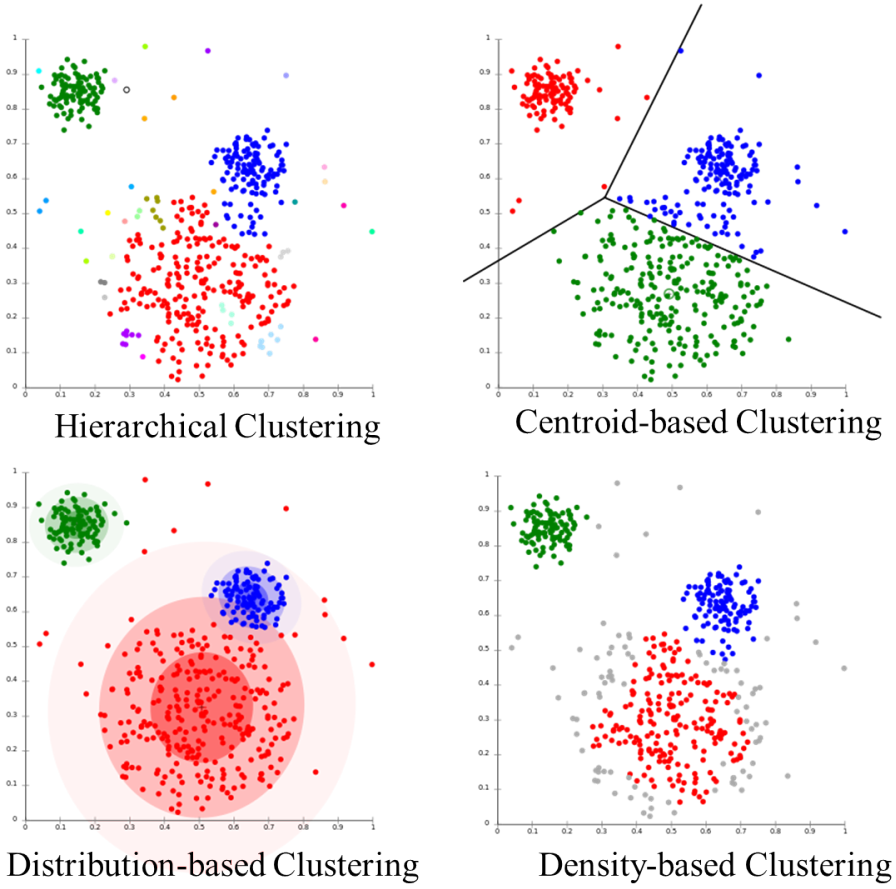



Figure 3: Examples of major clustering algorithms for a given set of data points.

the number of minimum points is reached the cluster is formed. To handle temporal constraints in the spatio-temporal data, a modified version of the DBSCAN algorithm called ST-DBSCAN [10] was proposed. In addition to the spatial distance, temporal constraints between points are applied in each step. In this thesis, I have used the ST-DBSCAN algorithm for generating the clusters. OPTICS [7] is a modified version of DBSCAN and it can generate clusters using each and every point of the dataset. The drawback of using this algorithm is that it may consider a point at a very large distance

from the centroid of the cluster as a potential member of the cluster. Due to the nature of this algorithm we may face overlaps while drawing the cluster boundary if any member of another cluster exists in the middle. Therefore I decided to use DBSCAN in my thesis for generating clusters of spatio-temporal data points.

## 2.3 TECHNIQUES FOR VISUALIZATION OF SPATIO-TEMPORAL CLUSTER PROPERTIES

Researchers have explored visualization of STC properties from different perspectives focusing on solving specific problems such as showing directional movement, aggregate measures, and overall cluster patterns. None of these methods provide general solutions for a wide range of datasets and analytic tasks. In particular, visualizations are mostly limited to one or two properties of the cluster. For example, Andreinko *et al.* [4, 5, 6], have proposed the use of directional arrow, circle size, and bar diagram to visualize cluster density and movement direction. The shapes are drawn based on the center point of the cluster thus providing information about the cluster location. The width of directional arrows represents the number of objects moving in that direction. Circle size and bar heights represent the number of points within a cluster. Similarly, Lodha and Verma [40] used a three dimensional square area with different colors and lengths to represent crime data density. The cluster areas are determined based on geographic boundaries. In another work, Andreinko *et al.* [2] showed how to trace aggregate history of trajectory data. To support their analysis, the authors used cluster

visualization where tessellation of the entire spatial map was done and color coding was used as shown in Figure 5(c) to represent any particular attribute's status for the cluster. Tominski *et al.* [59] presented a solution based on the space-time cube but the main objective of that technique was to show attribute values of each trajectory. The authors visualized those information by stacking 3D color-coded bands on a 2D map and ordering the bands based on the temporal information, the trajectories and their attributes where the temporal information is directly perceivable. Extra visual cues are also added to the bands to depict direction and other dimensions. Recently, Wallner *et al.* [64] proposed a polygonal shaped visualization with color coding to show a single property of clusters from game play data. In this thesis, I propose a visualization of multiple cluster properties in a single view allowing comparative analytic tasks which has not been addressed in previous works.

Prior research has considered methods for interacting with STC visualizations to analyze different properties at once. Filter-based interaction techniques [37, 22] are able to extract patterns in a large spatio-temporal dataset. These methods allow users to select a set of events within a region of interest and generate patterns by searching similar events in the whole dataset. Rinzivillo *et al.* [51] proposed progressive clustering to analyze movement data at each step of the filtering algorithm which allows to visually make sense of the data through clusters. The authors propose pie chart-based visualizations including the grid like location indicators to show properties of the clusters. In addition, using aggregate movement arrows conveys directional information about the clusters. This technique works well

when analyzing one attribute at a time or for performing analytic tasks. Utilizing human analysts is another approach to guide the system to find meaningful clusters [4]. A human analyst can classify the clusters by seeing group of events in the output and control the final cluster generation process by manually including or excluding each cluster. Training the clustering algorithms (such as employing genetic algorithm) using different datasets and following a particular pattern or similarity through users' interaction [3] is another approach previously proposed. These methods have a common visualization approach similar to Geotime [31] where analysts have to remember measures in each step as they go through the whole process. These methods work well for specific problem solving related to STC or pattern. But they cannot provide a visualization of the summarized results in a single view.

From the above review we see that most prior work focuses the on visualization of one single property of clusters. Figure 4 summarizes the important properties of spatio-temporal data and some of the corresponding state of the art visual representations. There are six important properties for spatio-temporal clusters that researchers have explored to solve different analytic tasks. These properties and the visual mappings previously used are listed as follows:

- Density (number of objects) of the cluster
  - Circle size (Figure 4(a)): In [5, 51] circle size has been used to represent number of objects present in each cluster. In this example the top right cluster has more objects as the circle size is bigger than the bottom right cluster.

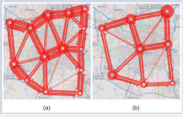
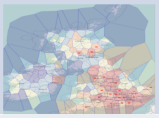
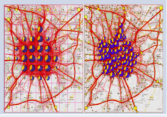
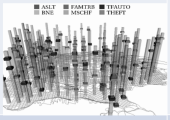
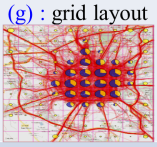
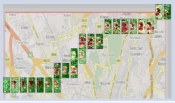
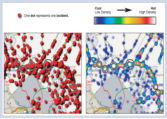
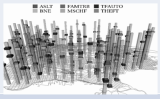
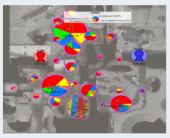
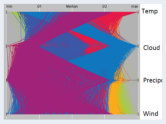
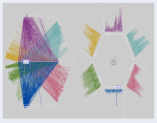
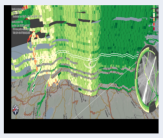
| Properties | Design1 | Design2 | Design3 | Design4 |
|---|---|---|---|---|
| Cluster density: # of objects in cluster | (a) : circle size  | (b) : arrow width  | (c) : color coding  | (d) : pie chart  |
| Spatial boundary: area/border of cluster | (e) : geographic  | (f) : Voronoi partition  | (g) : grid layout  | |
| Internal distribution: # of objects in regions within a cluster | (h) : mosaic diagram  | (i) : heat map  | (j) : tree map  | |
| Temporal information: Cluster formation time | (k) : bar length  | (l) : z-axis's values  | | |
| Direction: cluster movement direction | (m) : arrow  | (n) : color bars  | | |
| Aggregate measure: aggregate measure of dimensional values | (o) : pie segments  | (p) : parallel coordinate  | (q) : Time wheel and Multicomb  | (r) : Stacked 3D  |

Figure 4: State of the art visual mappings of different cluster properties of spatio-temporal data.

- Width of a particular shape (Figure 4(b)): Andreinko *et al.* [51] used width of arrows to show the number of objects. In this example, the variation in the size of the start and end points of the arrow represents the number of objects present in the cluster connected by that arrow.
- Color coding (Figure 4(c)): In [2] red to blue color coding was used to show number of objects where red represents high and blue represents low number of objects.

  - Pie chart (Figure 4(d)): Rinzivillo *et al.* [5] proposed to use different pie segments to represent the number of objects present in a particular cluster based on a given criteria.

- Spatial boundary of the cluster

  - Geographic border of a region (Figure 4(e)): In [40] authors used the original physical boundary of a city or region to show the position of a particular cluster. In this method at first the boundaries are made and other information of the cluster is drawn on top of it.
  - Voronoi partition (Figure 4(f)): Andreinko *et al.* [5] proposed to use centroids of all clusters to apply a Voronoi partitioning algorithm and the boundaries are generated based on those partitions.
  - Grid partition (Figure 4(g)): Tomniski *et al.* [51] used a grid layout to divide the whole map into equal sized regions to show a cluster's boundary.

    All these methods for showing cluster boundaries cannot provide information of the actual placement of its constituent objects. Here the original cluster shapes are not being preserved.

- Internal distribution of objects within a cluster

  - Mosaic diagram (Figure 4(h)): In [1] the color coding in each small segment within a cluster has been used to represent the density in that segment. This method generates

a mosaic like texture by merging all these segments of a cluster.

  - Heat map (Figure 4(i)): In prior work [67, 6, 54] a heat map was used to show the internal density of a cluster where color coding is used for each pixel of the map within a cluster boundary.
  - Tree map (Figure 4(j)): Andreinko *et al.* [6] used a specific set of discrete colors to represent the number of objects within a region or cluster.

- Temporal information of cluster formation
  - Bar length (Figure 4(k)): In [4] researchers showed mean time of each cluster using bar lengths.
  - Z-axis in 3D view (Figure 4(l)): Similar to the normal trajectory visualization, the third dimension or Z-axis [40] ] is also commonly used by researchers to show temporal information of the cluster.

- Aggregate movement direction
  - Arrow (Figure 4(m)): In [51] arrow width was used to show the number of objects moving towards the direction of the arrow and different color of the arrows was used to represent the intended purpose of those aggregate movements.
  - Color bars (Figure 4(n)): Andreinko *et al.* [1] presented a set of color bars at the centroid of the cluster where length of the bar represents the number of objects are moving and the color shows direction of the movement. In this

example the pink and green bars show the number of objects moving towards west and east.

- Aggregate measures of objects' dimensional values
  - Pie segments and color coded segments of a polygon (Figure 4(o)): In [64] authors proposed a pie chart to show a single dimension and the colored segments of that pie were used to represent the number of objects having a particular dimensional value. Similarly they divided a polygon into various segments based on the number of objects for each dimensional values.
  - Parallel coordinate plot (Figure 4(p)): This is a widely used method [27] to show multi-dimensional values using multiple parallel axes. In this example the right vertical axis represents different dimensions and the left vertical axis shows the change of those dimensions' values over the course of time. The top horizontal axis is used to show values of those dimensions.
  - Time Wheel and Multicomb (Figure 4(q)): Tomniski *et al.* [58] proposed a modified version of the parallel and star coordinate technique to show aggregate measures of dimensions and to fit the visualization into a specific shape. The figure shows that the parallel coordinate plot can be generated within a hexagonal region.
  - Stacked trajectories (color coded) in 3D (Figure 4(r)): Tomniski *et al.* [59] proposed this method to show the values of an attribute of objects over the course of time. In this

example the dark green color shows high value and the light green color shows low value of a particular attribute of the trajectory.

## 2.4 MULTI-DIMENSIONAL DATA VISUALIZATION

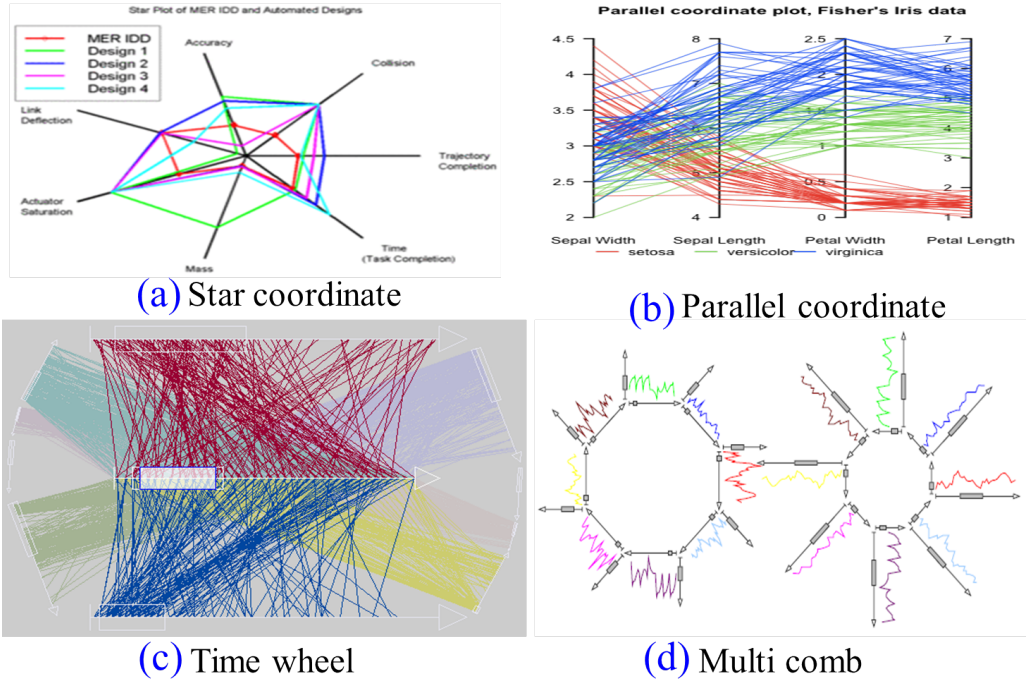



Figure 5: State of the art multi-dimensional data visualization techniques.

Current research efforts to visualize property values irrespective of the data domain have used parallel [27] and star coordinate [30] views which promise to capture multiple dimensions in a single view. In the star coordinate technique different areas of a star can represent different dimension and a point along that line indicates the value of that dimension of a particular item. As shown in Figure 5(a), here different colors represent different design of a model and the

values for different dimensions (i.e. accuracy, task completion time of those designs) are placed along the lines of the star. In the parallel coordinate method without using the star shape a number of vertical lines are used for different dimensions and a point along that line shows the value of that dimension. In Figure 5(b) each color represents different types (i.e. type of flower) and the vertical lines represent various dimensions (sepal length, sepal width, petal length, petal width) of each flower. These two techniques are widely used for showing multi-dimensional properties of data. But due to their visual shape constraints, they cannot be easily embedded in the spatial dimension. For example, if we have any arbitrary shaped cluster it will not be possible to draw a star or parallel coordinate plot on top of that cluster shape. Tomniski *et al.* [58] have proposed a modified version of the parallel and star coordinate, called Time Wheel and MultiComb respectively which can be embedded in the spatial dimension by restricting the visualizations to a particular shape as shown in Figure 5(c). Similar to Parallel and Star coordinates in these designs each color represents different objects and the corresponding line graph shows values of a particular dimension of that object. Additionally in these designs the change of values in course of time can also be tracked through the sliders as shown in Figure 5(d). This modified design can hold particular shapes such as hexagons, pyramids, etc. Only clusters having that particular shape can use these methods to show multi-dimensional properties. However, as observed, these designs cover a significant portion of the display's real estate, making it difficult to embed additional cluster properties on top of it.

## 2.5 "FLOCK" VISUALIZATION

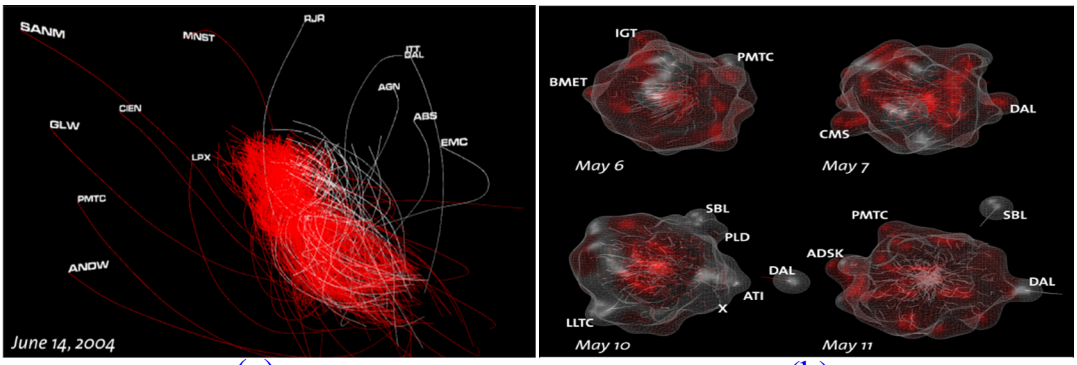



Figure 6: (a) Individual boids separating from the main flock. (b) Boid shape generation and evolution.

To represent the behaviour of a group of objects and to visualize their dimensional values, researchers adapted natural behaviour, such as the movement of a flock of birds, a school of fish, or a herd of land animals [42, 50]. The objectives of these prior work were to show complex relationships (i.e. aggregate movement pattern, proximity based on group behavior) between data points. Later, Proctor and Winter [47] extended this concept to visualize time-varying datasets. Their proposed method can generate dynamic visual patterns based on long-term or short-term temporal similarities within datasets. The visualization consists of a three-dimensional blob shape with embedded boids to represent global and directional trends for each flock region as shown in the Figure 6. In this figure the price trend of different products in a stock market is shown and the behaviour of price changes of a group of products are represented by the boid shapes. Figure 6(a) shows prices of products that are increasing (red

color) or decreasing (white color) from a previous trend. In Figure 6(b) the whole outer surface reflects the potential values of each boid or product through color to present the overall state of the flock or prices of a group of products. This design of the flock is helpful when we want to simulate any behavioural system such as dynamic actions or movements of people or objects. Although these works use the term "flock" they do not represent spatio-temporal clusters. As such, FlockViz is completely different in terms of objective and visual representation.

# 3

# FLOCKVIZ DESIGN

The structure of flocks in nature motivated the design of my cluster visualization (the common pattern of birds that move together). Variations can be seen in diverse animals groups, such as ducks, fish (school), birds (flock), and cows (herd) as shown in Figure 7 (a) to (d), respectively. It is very interesting to see that each flock has a different shape and the flock present various attributes in a single view. From these images we can predict the percentage of objects present in a particular part of the flock or the direction in which the objects are moving. However, it is not clear how to get information about when the flock was formed, the time taken to generate the flock and different dimensional values such has how many male, female or old objects are present in that flock. Based on my initial motivation, I designed a visualization scheme with enhanced capability which can represent all these additional information as well.

## 3.1 PROPERTIES OF CLUSTERS - REQUIREMENTS FOR A STC VISUALIZATION

An effective visualization needs to unveil and represent hidden information. In terms of clusters, we need to consider additional properties or dimensions such as how many objects formed that

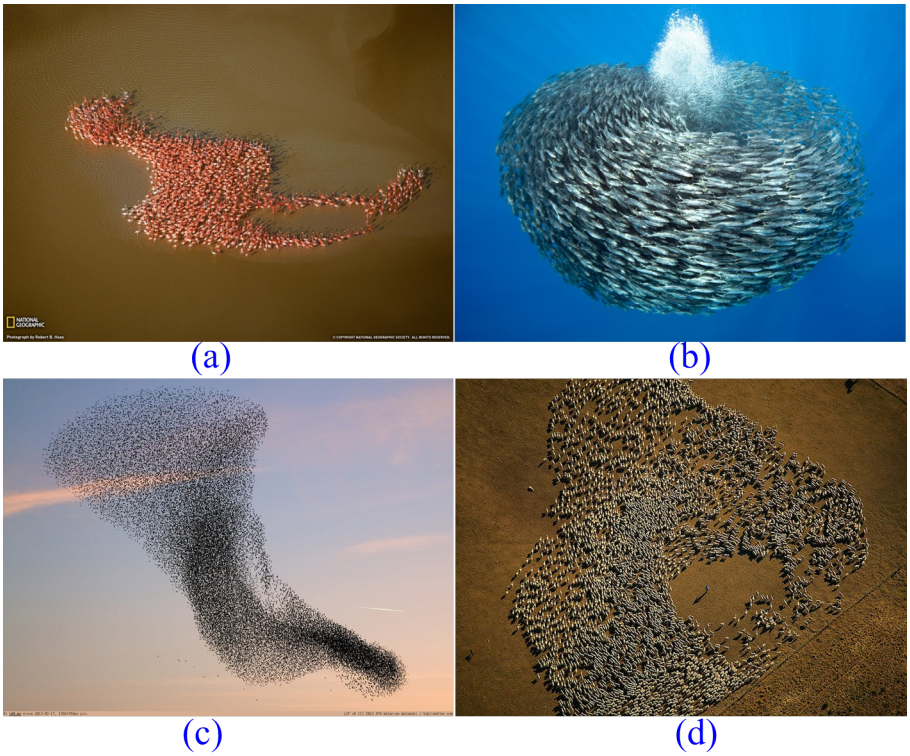



Figure 7: Some examples of biological pattern like flock commonly seen in nature. Image (a) shows movement of ducks in a group, (b) shows the school of fish in an ocean, (c) shows a flock of bird in the sky and (d) shows a herd of cows in a farm. Image source [23, 70, 63, 8].

cluster, different statistics based on various object properties, temporal formation, etc. Presenting this information can make the task of the analyst more effective [14, 9, 21]. From the literature (also see related work section) I have identified major properties of cluster that should be included in an ideal visualization of a sptio-temporal cluster:

- Cluster region or area: cluster area is the common physical boundary of all the objects present in a cluster. In general it should be the most outer convex polygon drawn using all the points within a cluster.

- Cluster density: cluster density is the measure of the number of objects per unit area. For many analyses such as finding objects' placement within a cluster this information is crucial.

- Cluster size: irrespective of density, the overall count of objects gives a good analytic measure to include or exclude the cluster for further analysis.

- Temporal formation of cluster: temporal information about a cluster is another important component of a visualization. The existence and duration of a cluster for any particular period of time can lead to important analytic conclusions.

- Aggregate movement direction of a cluster: objects' movement direction before and after cluster formation helps to realize how the clusters evolve over the course of time.

- Aggregate measures of a cluster's multi-dimensional attributes: another interesting element about spatio-temporal data is the presence of multiple variables within a cluster. For example in the animal migration data, gender, age and size are some important dimensions on which an analyst may reach different conclusions.

The above list is a set of requirements that will guide the design of my novel spatio-temporal cluster visualization technique. The primary challenge consists of designing and implementing a visualization that can make visible as many of these variables at once. Making such multi-dimensional information visible in a single visualization can be useful for complex queries. For example a query

asking the following: "find the most recent densely populated cluster of animals which has the largest number of males?". For such a query, we require information of spatial, temporal, density and multi-dimensional attributes of cluster in the visualization at the same time. This is an example of a scenario where current cluster visualization methods fail to provide such information efficiently. Instead, analysts can adopt FlockViz for easily answer such queries. Database query can also find the solution but cannot identify the cluster position visually.

## 3.2 DETAIL DESIGN OF FLOCKVIZ

As described in the related work section, prior concentrated on visualizing cluster properties such as a) shape, b) internal distribution, c) temporal formation, d) density, e) movement direction and f) various aggregate measures of objects' dimensional values. I incorporated these attributes on different parts of the flock that considers the cluster as a single body. Throughout my design, I preserved the shape of the cluster intact which is one of the main goals of my thesis. Another important objective of this design is to show all the information in a single view at the same time. Therefore, I used different parts to show different information so that one part of the flock cannot occlude the other. This is crucial as it helps to analyze multiple properties simultaneously without switching the visualization mode.

I used the ST-DBSCAN [10] algorithm to extract the sets of points for a cluster or flock. This is the most recent clustering algorithm for spatio-temporal data in the DBSCAN family of algorithms. Differ-

ent clustering algorithms will generate different clusters and hence different visualizations. However, I do not measure or evaluate the performance of clustering algorithms. Instead, with FlockViz, my main focus is to visualize clusters so that different properties can be represented through a single visualization without compromising their spatial structures. Before describing different elements of FlockViz, I will first briefly describe the ST-DBSCAN [10] algorithm that finds the sets of points for inclusion in each cluster.

DBSCAN (Density based spatial clustering), is designed to generate any arbitrary shaped cluster in a database and to distinguish noise points or outliers. It works with two parameters *Eps* and *MinPts* where *Eps* is the user defined radius of the cluster and *MinPts* is the minimum number of points that should be present within that radius. To understand DBSCAN clustering technique at first we need to understand the following definitions.

- Eps-Neighborhood: Two points $p$ and $q$ are Eps-neighbors to each other if the distance between them is smaller than *Eps*: $dist(p, q) <= Eps$.
- Core Object: A point $p$ is considered as a core object if it has at least *MinPts* number of Eps-neighbors.
- Directly density-reachable: Two points $p$ and $q$ are directly density reachable if they are Eps-neighbor to each other and both are core objects.
- Density reachable: Two objects $p$ and $q$ are density reachable with respect to *Eps* and *MinPts*, if there is a chain of objects

*p1,p2,.....pn, p1=p, pn=q* and *pi+1* is directly density reachable from *pi* with respect to *Eps* and *MinPts, for 1<=i<=n*

- Density connected: Two objects *p* and *q* are density connected with respect to *Eps* and *MinPts* if there is an object *o* such that both *p* and *q* are density reachable from *o*.

- Border object: An object *p* is a border object if it is not a core object but density reachable from another core object.

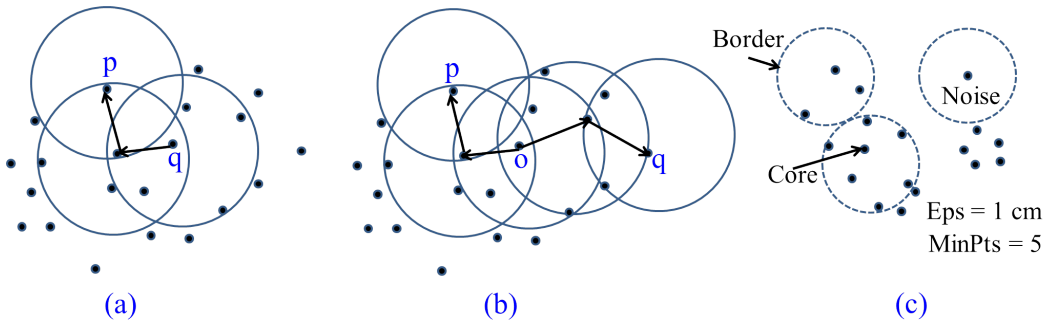



Figure 8: Basic concepts and terms used in ST-DBSCAN algorithm: (a) *p* density-reachable from *q*, (b) *p* and *q* density-connected to each other by *o* and (c) border object, core object and noise.

Figure 8 shows some examples that will help to understand these definitions. The density based cluster definition according to DB-SCAN algorithm is defined as follows:

**Density-based cluster:** A density based cluster *C* is a non-empty subset of database *D* which satisfies the following two conditions:

- Maximality: For all *p,q* if *q* is in *C* and *p* is density reachable from *q* with respect to *Eps* and *MinPts* then *p* is also in *C*

- Connectivity: For all *p,q* in *C*; *p* must be density connected to *q* with respect to *Eps* and *MinPts*.

```
Algorithm ST_DBSCAN (D, Eps1, Eps2, MinPts, Δε)
   // Inputs:
       // D={o1, o2, ..., on}  Set of objects
       // Eps1 : Maximum geographical coordinate (spatial) distance value.
       // Eps2 : Maximum non-spatial distance value.
       // MinPts : Minimum number of points within Eps1 and Eps2 distance.
       // Δε : Threshold value to be included in a cluster.
   // Output:
       // C={C1, C2, … Ck} Set of clusters

 Cluster_Label = 0

 For i=1 to n                                          //(i)
   If oi is not in a cluster Then                      //(ii)
      X=Retrieve_Neighbors(oi , Eps1, Eps2)            //(iii)

      If |X| < MinPts Then
            Mark oi as noise                           //(iv)
      Else                              //construct a new cluster (v)
         Cluster_Label = Cluster_Label + 1

         For j=1 to |X|
           Mark all objects in X with current Cluster_Label
         End For

         Push(all objects in X)                        //(vi)

         While not IsEpmty()
            CurrentObj = Pop()
            Y= Retrieve_Neighbors(CurrentObj, Eps1, Eps2)

            If |Y| >= MinPts Then
               ForAll objects o in Y                   //(vii)
                  If (o is not marked as noise or it is not in a cluster) and
                     |Cluster_Avg() - o.Value| <= Δε Then
                     Mark o with current Cluster_Label
                     Push(o)
                  End If
               End For
            End If
         End While
      End If
   End If
 End For
End Algorithm
```

Figure 9: ST-DBSCAN algorithm [10]. Image source: [10].

DBSCAN starts with an arbitrary point in the database and finds all the neighbor points that are within *Eps* distance. If the total number of neighbors reaches the *MinPts*, the start point is considered as a core object and a new cluster is formed. All the neighbor points are assigned to this new cluster. It iteratively collects all the neighbor points of all core points and finalize the cluster until all the points in the database have been processed. ST-DBSCAN follows the same concept for cluster generation but here instead of two parameters four parameters are used which are *Eps1*, *Eps2*, *MinPts* and *DeltaE*. *Eps1* is the distance parameter for spatial attribute and *Eps2* is the distance parameter for non-spatial or temporal attribute. This second

attribute also define the similarity among points within a cluster. Here two distance measures are checked every time for finding neighbouring points. The last parameter *DeltaE* is used to prevent discovery of the combined cluster due to very little difference in the non-spatial values of neighbour locations. A pseudo code of this algorithm from the original paper is given in the Figure 9.

Once I get the points for each cluster after applying a clustering algorithm the next step is to visualize that cluster considering the properties outlined in the previous section. In the following subsections I will describe the detail design of each element of my visualization which represents different properties of the cluster.

### 3.2.1 *Flock shape - Physical boundary of the cluster*

*Flock shape design description*

This part of the visualization covers one of the most important properties of clusters which is the physical boundary or cluster shape. Cluster shape is important because it gives the actual geographical space covered by the objects in the cluster. In many analyses, the cluster shape can reveal interesting properties of the group of items. For example, how a particular crowd moves in a public place, whether people in a meeting sat circularly or in another way, in what structure animals follow each other in a cluster, etc. Thus keeping the cluster shape as intact with the true properties of its elements is an important property. Currently most cluster visualization methods use the symbolic shapes such as an arbitrary circle or rectangle to

represent a cluster which fails to represent the geographical region covered by that cluster.

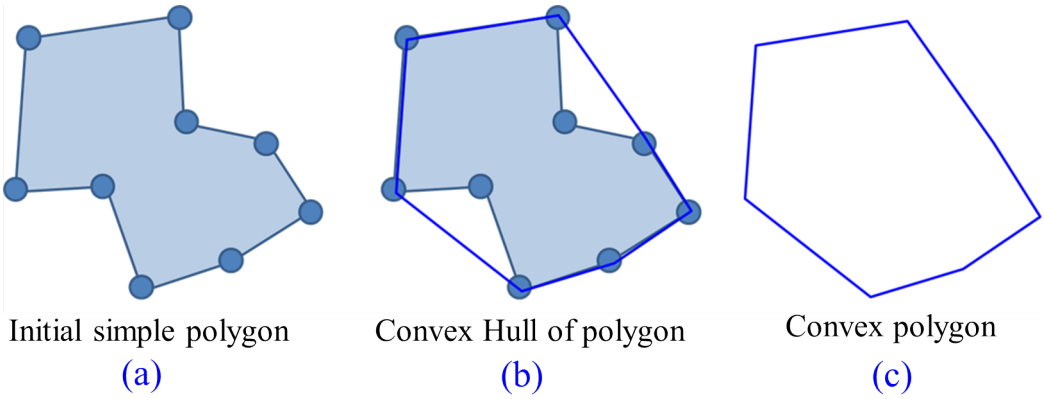



Figure 10: Cluster shape generation using a convex polygon. This polygon takes the smallest area to fit all the points inside it and ensures that we will always have to turn either right or left to traverse the whole polygon.

To design this part, at first we need to define what should be the shape of the cluster and how that should be generated given a set of points. There exist different algorithms for generating a polygon for a set of points such as concave polygon, convex polygon, rectangle of most outer points. There could also be a continuous curved shape drawn using the points instead of polygons. In this thesis, I used the most outer convex polygon with minimum number of points to represent the *flock shape*. Let's assume we have a set of trajectory points (Figure 10(a)) as our cluster after applying the ST-DBSCAN algorithm. Now there could be several designs on how we will draw the polygon. As shown in Figure 10(a) connecting all the points can give a basic cluster shape with a concave polygon. However, this shape is complex and hard to describe. Also, providing additional information on top of it will be difficult due to the complexity of the

shape. We will see this scenario in later designs of the other flock parts. Considering these problems I chose to use the convex hull polygon using the minimum number of points as it gives a simpler shape for the cluster. Outliers can affect the convex hull by making a large empty space which later make the visualization overlapped. I discussed this scenarios in the discussion section. But DBSCAN tries to minimize those noise points. Figure 10(b,c) shows the final cluster shape.

*Flock shape - alternate designs, pros and cons*

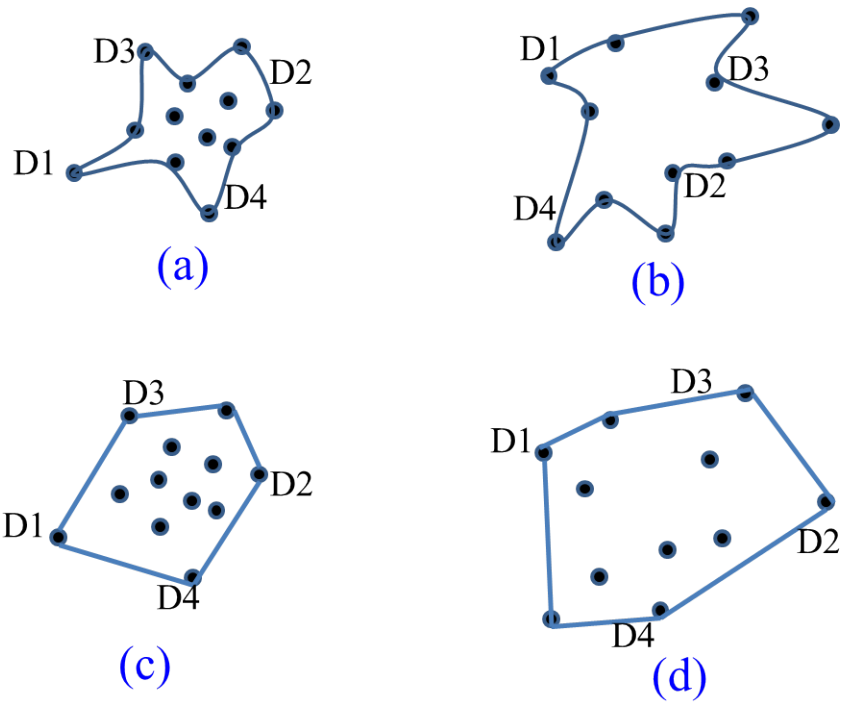



Figure 11: (a, b) Example of an alternate *flock shape* design using curved boundary. (c, d) Convex polygons show cluster shape generation for those clusters.

There are different possible alternate designs for *flock shape*. It can be a concave polygon as shown in Figure 10(a) or a curved area

(Figure 11(a, b)). These designs cannot provide benefits for further incorporation of additional information (as another flock part) on top of them. For example, if we want to place dimensional values as *flock shield* segments (described in more details in section 3.2.5) in the outer surface of these shapes, the position of those segments in different clusters will not be consistent. If we have four dimensions of each cluster such as D1, D2, D3 and D4 then the curved surface will place the visual cues of those dimensions in the *flock shield* regions inconsistently as shown in Figure 11(a, b). As I need to measure the distance for next dimension's placement curved surface is not appropriate for that. Here consistency refers the similarity between two clusters to place dimensions. However, my proposed convex polygon will keep the placement of dimensions consistent in different clusters (Figure 11( c, d)). In this example the position of the dimensions in cluster *a* is very different than in cluster *b*. This increases mental demand of users while analyzing different clusters at the same time. The curved surface is the main reason for this inconsistency. But according to my design cluster *c* and *d* have almost identical placement of dimensions. My proposed design might also have this inconsistency if the polygon have many complex edges. But in most of the cases it will give better result compared to the curved surface.

### 3.2.2 *Flock body - Internal distribution of the cluster*

*Flock body design description*

Internal distribution of objects present in the cluster is an important property which mainly shows the density map of the internal structure of the cluster. In my proposed design I discretized the *flock body* into sub-regions and I used color coding (green to red, where green represents few objects and red represents more) to fill those regions. In my design, there is an input parameter for the users where they can choose the number of discretization to build this density map. Therefore, a very high level as well as a granular density mapping is possible in this design.

Let's assume that we get a set of points after applying ST-DBSCAN clustering algorithm as shown in Figure 12(a). To preserve the original shape of the cluster, I build a convex polygon using the points which determine the boundary of the cluster as marked by ABCDEF in Figure 12(b) according to the previous section. Next, in order to discretize the *flock body*, I find the centroid of the convex polygon marked as O in Figure 12(c) and draw lines to each corner of the polygon. After doing this for this example dataset, I get six initial regions within the cluster. Then, based on the number of divisions from the centroid to the cluster border (which is an input parameter for FlockViz and in this example I assign it to 3), I equally divide the initial regions (Figure 12(d)). This leaves us with discretized sub-regions and based on the number of trajectory points present in each sub-region I calculate the regional density. Using the density values

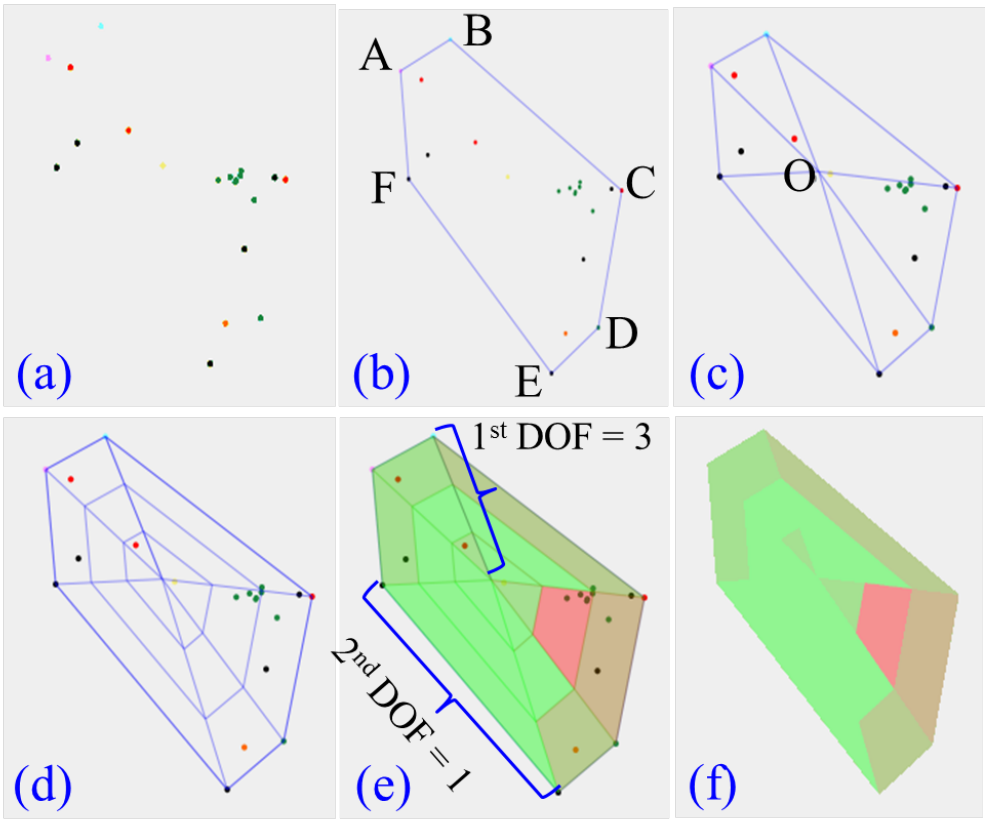



Figure 12: Steps for building an internal distribution map of the cluster.

with normalization (making the region of having highest number of objects to 1 and dividing other regions by that number ), I apply green to red color coding to get the final *flock body* (Figure 12(e, f). In my tool, this area is scalable so users can view the cluster structure at the desired scale. Users can hover over a particular sub-region to show detailed information associated with that sub-cluster in a pop-up window. These features can also be enabled or disabled to give priority to other elements of FlockViz. One additional option for the *flock body* is that it can have two degrees of freedom for discretization. The first degree is the number of levels from the centroid to the border and the second one is the number of divisions along

the border lines as shown in the Figure 12(e). The second level is set to 1 in this example. In FlockViz design, we can increase the values of these degree of fredoms. If we use a large values for these parameters we will get a continuous internal distribution map for the *flock body*. Therefore, without loosing a very granular level of distribution this method provides a rich internal structure.

*Flock body - alternate designs, pros and cons*

To represent the internal density of a cluster, prior work has applied other methods such as using a heat map, a mosaic digram, a tree map [44, 57, 53, 4]. All these methods can be used in the internal part of the flock. Among these methods, the heat map is widely used and popular. However, the main problem with this technique is that it gives a continuous density map of the *flock body* for each pixel rather than a sub-region within the body. We need to consider many regions from continuous space. Therefore, using a heat map is very difficult when we want to show different measures of a particular region within the *flock body* and do analytic comparisons based on that.

### 3.2.3 *Flock heart - Total number of objects in the cluster*

*Flock heart design description*

The most common information for any cluster analysis is the number of points in the cluster. Researchers have visualized this information through different designs such as the width of an arrow, bar length,

circle size, etc [4, 5]. Amongst all, the use of circle size has been more common and hence I also decided to use this mapping for this purpose. It is also very easy to place this visual cue in the center position of the flock without affecting the other parts. However, to keep the cluster shape intact and avoid clutter, the centroid of the *flock body* is chosen as an appropriate place. I call this location in the flock as the "*flock heart*" because it shows the most significant information that is overall size of the cluster. The size of this circle depends on the number of objects present in the cluster. I also give users the flexibility to scale, disable, or change the color to see the circle clearly. In Figure 13, the *flock heart* is depicted where the larger circle for cluster *c* shows that there are almost two times more objects present in this cluster compared to cluster *a*. Although the area of cluster *b* is the largest, that cluster does not have that highest number of objects. In this design, the circle colors are transparent black to avoid conflict with the colors of the *flock body*. The opaque and single color design makes this unique based on its position in the heart. However, in my design users can easily find the exact number of objects present in the cluster by hovering the mouse over the *flock heart*.

*Flock heart - alternate designs, pros and cons*

There could be other possible designs of *flock heart* based on its placement and icon representation. For example, we could place it at the top or bottom of the flock. In another design we could only use the color of the circle to represent the number of objects by keeping the size of the circle constant. As I used color coding to fill

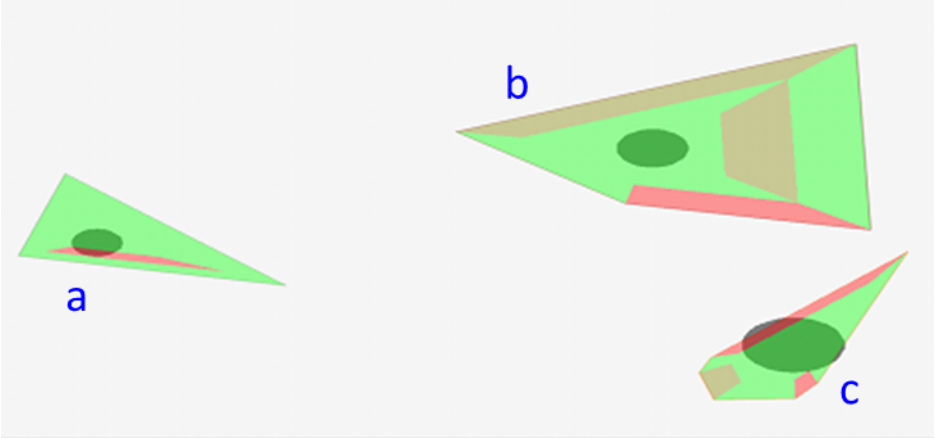



Figure 13: *Flock heart* design as drawn circles at the centroid of the clusters.

the *flock body* it would create confusion and increase the cognitive load of users if we use color for *flock heart* too. In my design the reason for placing it in the centroid is that it does not conflict with the other parts as I have those visual cues for different properties in the surrounding regions of the *flock body*. Other alternate designs of the *flock heart* could be using square shape, bar height etc. However circle is one of the most common and widely used method [5, 51] to represent this kind of measure which also gives a quick visual result for comparison based analysis.

### 3.2.4 *Flock head - Temporal information of cluster formation*

*Flock head design description*

The *flock head* represents temporal information of the cluster. I call it "*flock head*" as it shows information at the top and by just looking at this, users can identify solutions to many cluster-based questions. In Figure 14, *flock head* is shown as circles to represent temporal

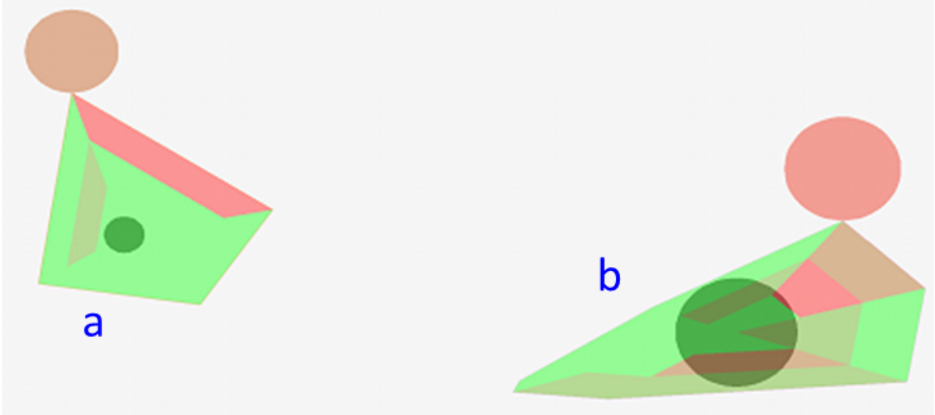



Figure 14: *Flock head* for representing temporal information. The size of the circles drawn at the top of each cluster represent the duration of cluster formation and the color coding (green to red) of those circles is used to show recency of cluster formation.

information of the clusters. They are placed at the top point of the flock. The size of the circle represents the total time needed to form that cluster and color coding (green to red) is used to represent temporal recency of the cluster. The color red represents the most recently formed cluster and the color green represent the earliest with colors in between representing the recency from red to green order. In this particular example the cluster *b* took almost the same time to form and was generated more recently than the cluster *a*. In this design users have the flexibility to see the actual time by hovering over the *flock head* at any time. They can also scale it or change the color coding during analysis.

*Flock head - alternate designs, pros and cons*

Like the previous alternate designs of the *flock heart*, it is possible to use different shapes such as square, triangle or bar to place in the *flock head*. I decided to use circle to be consistent with the other

designs (i.e, *flock heart*) so that users do not have to give too much effort to understand various visual cues. Another reason is that a circle is commonly used to represent "head" of any symbolic physical entity. Another alternate design for representing temporal information could be using an analogue clock within the heart of the flock. There will be two parts of the clock for the start and end time of the cluster formation. It will be the normalized time considering all the clusters rather than the actual cluster formation time. So, by seeing the angle between two legs of clock we can easily tell how long it took to form the cluster and the end leg will indicate how recent the cluster is. In Figure 15 temporal information of the clusters is shown by drawing red legs of clock within the *flock heart*. Here the same clusters in Figure 14 are shown where for both of the clusters the angles or duration of cluster formation is almost the same but the end time of cluster *b* is more recent than the cluster *a* according to the end leg's position. This alternate design for showing temporal information of clusters is very helpful to place a particular cluster along the time line of the dataset. However as this design uses the real clock concept it is easy for the users to set their minds in that context to extract temporal information.

### 3.2.5 *Flock shield - Aggregate measures of different dimensional values*

*Flock shield design description*

Simplification of datasets by aggregation is one of the main benefits of generating clusters. As this is the main goal of analysts for using

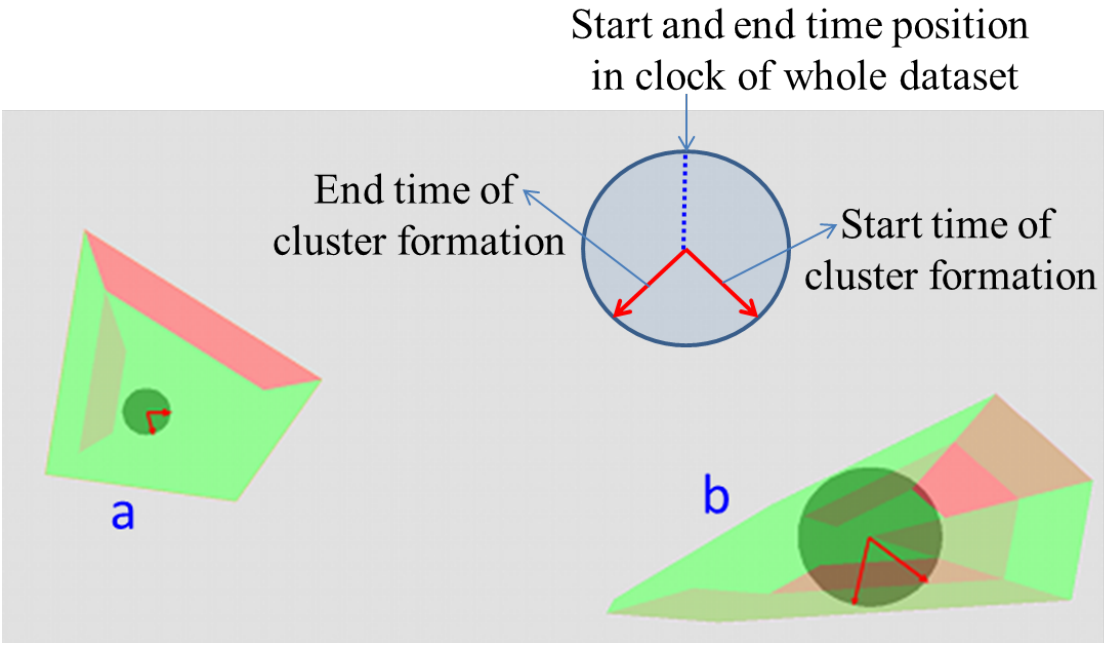



Figure 15: Alternate design for temporal information of the cluster using clock legs.

clusters, they usually seek different aggregate measures of different dimensional values of objects within that cluster visualization. Most of the complex spatio-temporal datasets have different dimensions which are attributes such as age, gender, color, physical condition, activity type, status etc. Each of these dimensions could have multiple values such as child, old, young values for the age dimension. When we generate clusters, we also need to store different statistics like number of objects, sum of weight etc. for each of these dimensional values. To help analysts work with these aggregate measures, an ideal cluster visualization must be able to show these values. I propose the *flock shield* to capture and show these measures without affecting the cluster's actual physical structure.

*Flock shield* is an additional surface area surrounding the whole *flock body* (for example, the area indicated by ABCDEF in Figure 16(a)).

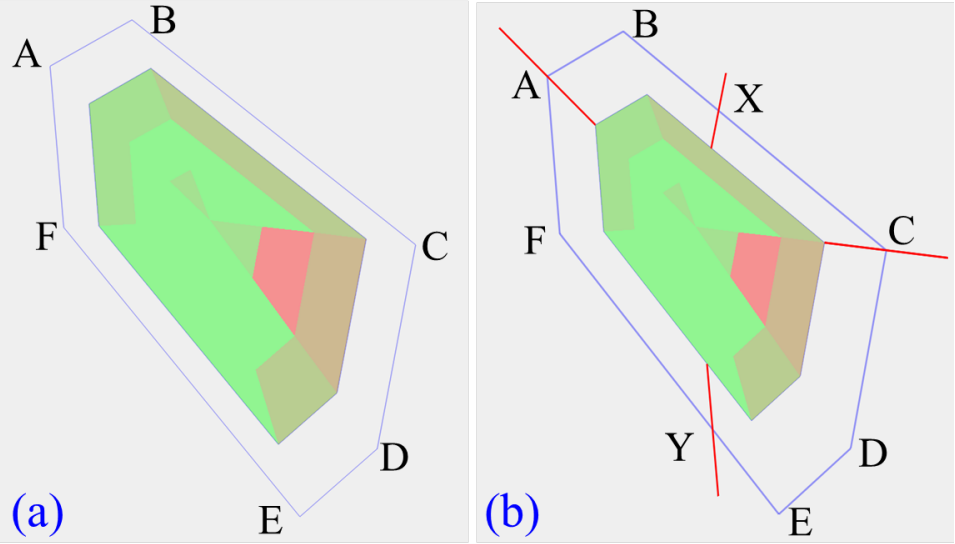



Figure 16: First two steps for building *flock shield*.

The outer polygonal area satisfies the primary requirements to keep the original cluster shape and structure intact. I called it "*flock shield*" based on its position and as it represents various attributes of the cluster which is the main strength of the cluster visualization. The distance or width of the *flock shield* from flock boundary can be variable and users can set that at any time. It can be dynamically divided into sub parts based on the number of dimensions for which we want to show the aggregate measures. To construct the *flock shield* divisions at first I find the leftmost point (A) and the rightmost point (C) dividing the shield into the upper segment ABC and the lower segment AFEDC (Figure 16(a)). I choose this design to make the visualization rotation-invariant. There are several design alternatives for further dividing these two segments. For example, more dimensions can be placed in the larger segment than in the smaller segment. To take advantage of users' memorability, I decided to have the same

number of divisions in each segment. To avoid even-odd conflict for n number of dimensions, the first Integer {(n-1)/2}+1 dimensions will be in the upper segment and the rest in the lower segment of the shield. I preserve 1% of the total segment length for each value of the dimensions so that if any value is zero, I still can show that section and fill it in with black. Now, if I have four dimensions in the dataset then each segment (ABC and AFEDC) of the *flock shield* will be divided into two equal length parts indicated by the red lines in Figure 16(b). The resultant four segments of the whole *flock shield* can be used to represent those dimensions. For example, ABX can be used to represent gender (i.e. male and female), XC is for age (i.e. old, child, and young), AFY is for location (i.e. park, hill, lake area, and near highway) and YEDC is for the activity (i.e. sleeping, running, idle, and breeding) dimension.

To assign dimensions to the divided shield regions, I divide the shield part ABX based on the normalized values of aggregate measures of the gender dimension. I calculate the number of male and female objects in this cluster and divide this shield part based on those normalized values. In this example the number of male objects (ABG) is greater than female objects (GX) (Figure 17(a)). Similarly, other shield parts can be divided based on the aggregate measures for the corresponding dimensional values.

This part of the *flock shield* design helps to make within-cluster comparison of the values related to a particular dimension. Another important visualization of the *flock shield* is the use of color coding which is again green to red to fill the segments of the *flock shield*. Here the more red in color for a particular shield part compared to the

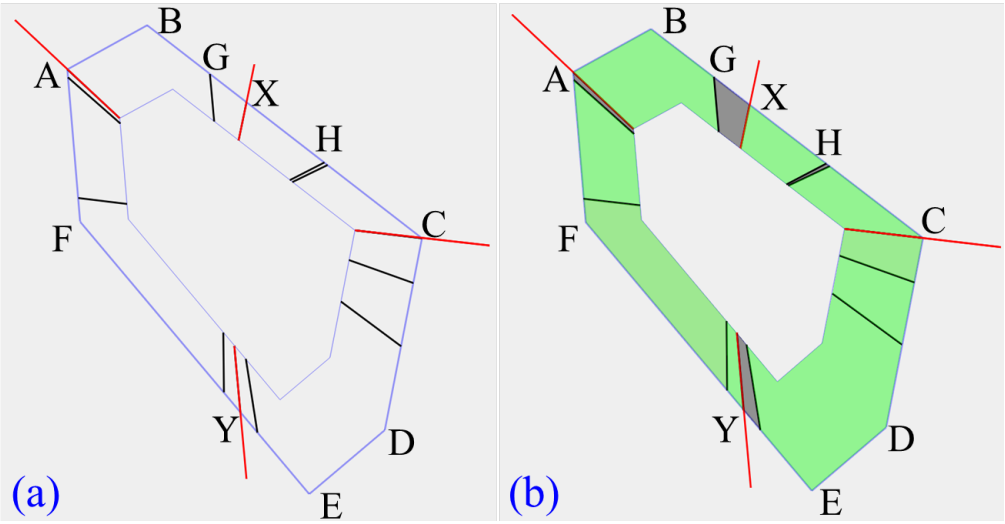



Figure 17: Steps for assigning dimensional values into *flock shields*.

other clusters means the more number of objects are present in the cluster corresponding to that dimensional value. For example, in the segment XC (which represents the age dimension) in Figure 17(b), the number of objects of "old" value represented by XH is green. Now, if in another cluster *k*, the same part of the age segment is red, then we can say that there are more old-aged objects in the cluster *k* compared to this cluster. I carefully designed the *flock shield* to reduce cognitive load of the users. I considered peoples' flexibility of reading in a sequence and kept the position of the shield parts consistent in my design. My tool also provides interactive features to scale the *flock shield* as well as to enable or disable it. Users can highlight any particular dimension's value at the bottom of each cluster and easily compare it among clusters as shown in detail in section 3.2.7.

*Flock shield - alternate designs, pros and cons*

I propose an additional design of *flock shield* which can be used for representing a large number of dimensions. Users can choose how many layers of shield they want to use and how many dimensions they want to see in each layer. In that case there will be layers of *flock shields* and if we have eight dimensions to visualize one option will be to show the first four in the first layer of *flock shield* and the rest in the second layer of the *flock shield*. There are some alternate designs of *flock shield* where we could use the same outer region of the *flock body* and various kinds of visual cues instead of using the segments of the *flock shield*. In the first option we can have multiple circles with different sizes and colors. The circle size can represent within-cluster values and color can be used to show between-cluster values. Secondly, we can use vertical bars with different length for within-cluster values and color of those bars for between cluster values. Also, the width of the bars or edge curliness can also be used for between cluster value representation.

Another alternate design of *flock shield* could be variable pie charts drawn in each shield segment. In this design we can fit a larger number of dimensions than the original *flock shield* design for each layer. Here the percentage of the pie or pie length can be used for representing within cluster values and fillness to the center or pie width can be used for between cluster values. This design was well appreciated in my post evaluation design session and I did some analysis based on this modified design in case study section. Figure 18 shows this alternate design of *flock shield*. Similar to the

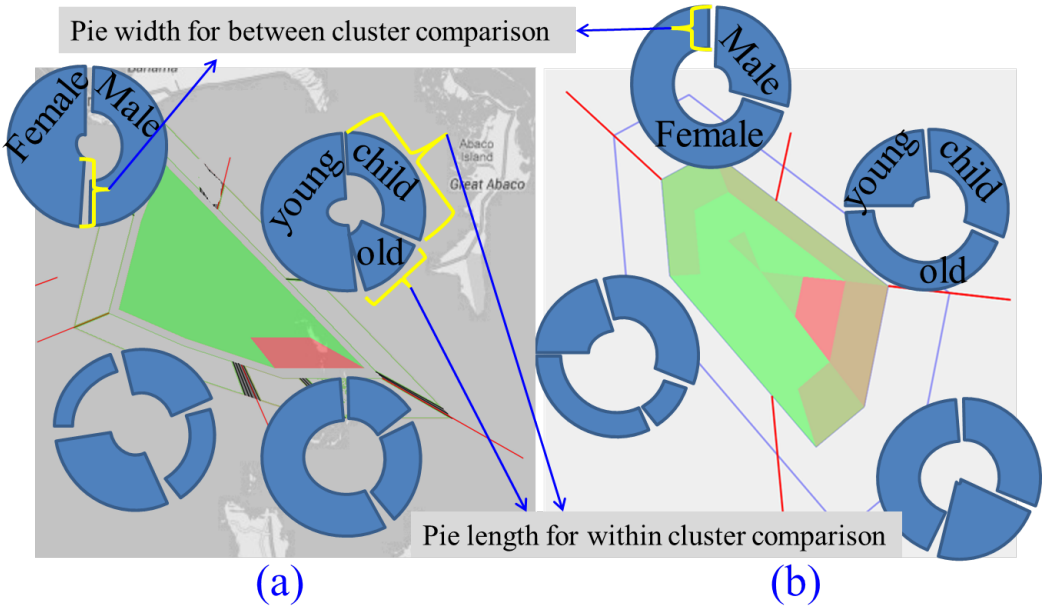



Figure 18: Alternate design of *flock shield* using pie segments.

previous example, there are four dimensions in this dataset and the values for each dimension are shown in different *flock shield* regions. Every dimensional value starts from the start position of a conventional clock. In this data set, for gender dimension we have male, female; for age group dimension we have child, old and young values and so on. Since the pie segment length shows within cluster values, in the cluster *a* there are equal number of male and female objects. But in the cluster *b* there are more female object than male objects. Here fillness of pie to the center shows between cluster comparative values. Therefore, in the cluster *a* there are less male objects than the *b* as the corresponding pie segment in the cluster *b* is more filled towards the center. The advantage of all these alternate designs is that they can be incorporated in the outer part of the original flock without affecting the cluster shape. This pie segment

based *flock shield* design can represent a large number of dimensions as in this design we have more space in the shield region compared to the polygonal *flock shield* design.

### 3.2.6 *Flock wings - Aggregate movement direction of the cluster*

*Flock wings desing description*

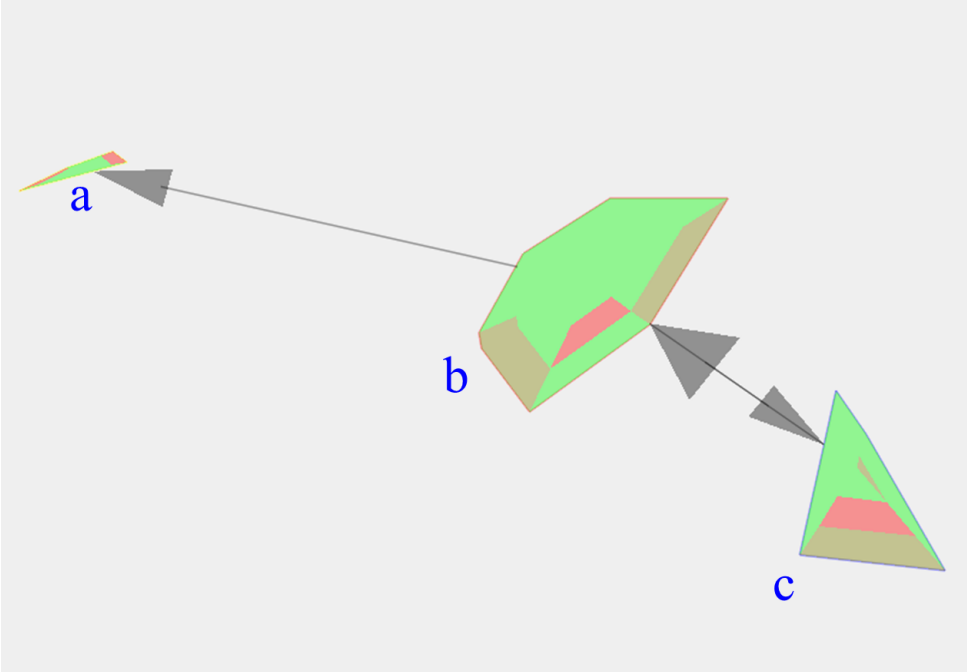



Figure 19: *Flock wings* as directional arrows to show aggregate direction of cluster movement. Here the width of the arrow head shows number of objects moving toward that direction from one cluster to the other.

Movement of objects from one cluster to another is an important information for predicting activity and migration patterns. My proposed visualization method ( "*flock wing*") uses arrows to represent this information. In FlockViz, each flock can have a) out-directional wings, b) in-directional wings, or c) no wings. The width of the

arrow head represents how many objects moved from one cluster to another. For example, in Figure 19, from the cluster *b*, equal number of objects are moving to the cluster *a* and *c*. A larger number of objects are moving from the cluster *c* towards the cluster *b* as the width of the arrow head is greater compared to others. Also, the cluster *a* has no outward wings which means that the data has no trace of where the objects from this cluster are moving to after entering this region or the objects no longer moved.


Figure 20: Alternate design of *flock wings* which uses only arrow head to represent directional movement.

*Flock wings - alternate designs, pros and cons*

Arrow is one of the most common visual cues to show directional information. In my proposed *flock wings* design, we can use only arrow heads for in and out directional movement of objects at the border of the clusters as shown in Figure 20. Here we do not need

to draw lines which helps to have a clutter-free map when there are many directional arrows.

### 3.2.7 *Flock tail - Highlighted measure of any particular cluster properties*

*Flock tail design description*

In many cluster analysis tasks, we need to compare several attributes among all the clusters and find the cluster that has a maximum or minimum value with respect to that attribute. To answer this type of question an ideal cluster visualization must have an option to explicitly show only that required measure in a particular part of the visualization. In my proposed method, I used the *flock tail* to show the highlighted measure. Hovering over a particular dimensional value or any other measure will trigger the corresponding measure to be displayed in the *flock tail* of all clusters. I used a small triangle to represent this visual cue. The length of the tail will vary for each cluster based on the value of the highlighted measure. This visual cue is situated at the bottom of the *flock body*. Users can enable or disable this part at any time. There is a pop-up information window added at the hovering point to help users identify which dimension is currently activated for highlighting as shown in Figure 21 marked by x. Here the pop-up window shows that user asked to highlight how many storms in each cluster have a wind speed of between 65 to 95 mile per hours. The *flock tails* give a quick and clear comparison of those measures among all clusters. In this example the cluster *b* had the most number of storms that had a wind speed 65 to 95 mph.

I use a constant blue color to fill this tail part. This flock element is useful for comparing a very broad range of views by just seeing the tail length of all clusters.

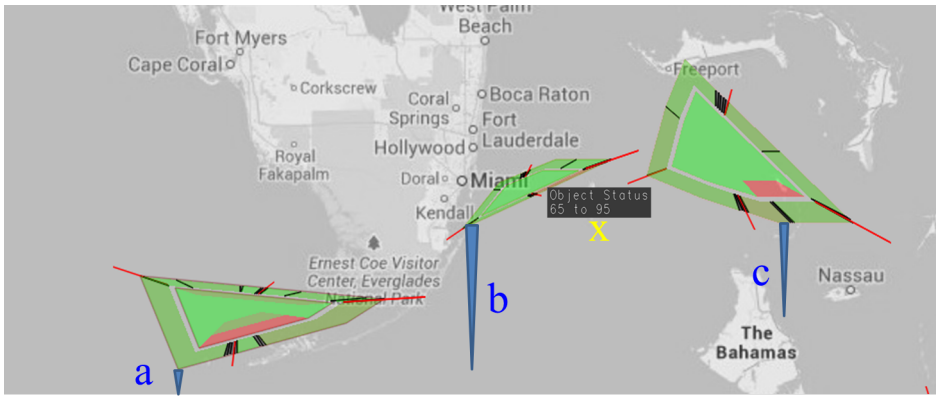



Figure 21: *Flock tails* to show highlighted measure using the height of the blue triangles.

*Flock tail - alternate designs, pros and cons*

There could be other possible designs of *flock tail* based on its placement and icon representation. For example, the nearest bottom corner of the centroid can be used to visualize this part. Besides different shapes such a circle, rectangle bar etc. can also be used instead of thin triangle.

# 4

# EVALUATION OF FLOCKVIZ THROUGH EXPERIMENT

FlockViz is designed to solve a wide range of analytic tasks with various dataset types. To evaluate the performance of FlockViz, I used five data sets in different domains such as people's movement, eagle migration, storm movement, fish movement and caribou movement data. To evaluate the effectiveness of FlockViz, I generated a framework of task categories related to identification and visual analytics of spatio-temporal clusters. I ran a user study and measured the performance of FlockViz in terms of task completion time, error rate and users' preference against a traditional approach for cluster analysis. Finally, I present some case studies using the datasets that are currently being used by the Natural Resources Institute (NRI) and the Biological Sciences departments at the University of Manitoba. Using these case studies, I demonstrated how FlockViz assists in solving their research questions and provides data analytic value. The following sections describe the datasets and my proposed task categories as well as the analyses of the results.

## 4.1 DATASETS

One of the main motivations of my proposed visualization is to be universally applicable for a wide range of spatio-temporal datasets. To test FlockViz (and by keeping this challenge in mind), I used data sets from five different domains including storm movement, people's movement in cities, caribou movement in parks, fish movement in a lake, and also bird migration data. I classified the datasets which were used in the user study into low (around 1000 movement points), and high density (around 3000 movement points) based on the number of movement points. Each data set has eight data fields such as object identification, latitude, longitude, timestamp, object category/type, object status, high level location name or terrain, and object's activity type. The first four fields are common for all datasets and are the core data fields to visualize any spatio-temporal data [4]. The last four data fields are considered as objects' dimensions. In order to keep the number of dimensions consistent across all datasets, I restricted them to four dimensions. However, FlockViz can support more than four dimensions by splitting the shield region into that number and adding additional layers of shields as discussed in the design section (3.2.5). Also, in the alternate design of *flock shield* we can place more pies surrounding the *flock body* to handle a large number of dimensions.

### 4.1.1 *Hurricane movement data*

This dataset contains movement data of hurricanes between June and December 2005 in the Atlantic Ocean [66]. A total of 900 trajectory points for 29 hurricane movement paths exist in this dataset. In my study, this dataset is considered as having low level density based on the total number of trajectory points. Four high level properties (e.g. storm type, speed of wind, terrain, movement direction) of each hurricane is also captured and mapped to the general dimensions as shown in Figure 22. I used this dataset for the main user study. The reason for using this dataset in the user study is that it's a real dataset and has low density.

### 4.1.2 *People movement data*

This is a simulated data of three peoples' movement in Berlin, Germany and recorded time between 09/02/13 and 12/02/13. The movements are related to various trips such as going to the office, or to the university using various types of transportation like bus, auto mobile, etc. A total of 1,707 movement points is recorded. The dimensional values (e.g. movement purpose, movement medium, traffic position, movement pattern) of peoples are shown in Figure 22. This dataset was used for the practice session of the user study. As I simulated this data I did not use it for the main user study.

| Data set | General properties | Object type | Object status | Location | Activity |
|---|---|---|---|---|---|
| Hurricane movement | Mapping | Storm type | Speed of wind | Terrain | Movement direction |
| | No of Instances | 7 | 4 | 4 | 5 |
| | Instances | tropical dprs<br>tropical storm<br>Hurricane-1<br>Hurricane-2<br>Hurricane-3<br>Hurricane-4<br>Hurricane-5 | 10-30 mph<br>35-60 mph<br>65- 95 mph<br>100-150 mph | near land<br>main land<br>ocean<br>forrest | within ocean<br>ocean to land<br>within land<br>land to ocean<br>along shore |
| People movement | Mapping | Movement purpose | Movement medium | Traffic position | Movement pattern |
| | No of Instances | 5 | 4 | 4 | 4 |
| | Instances | office<br>market<br>home<br>university<br>recreation | walking<br>in bus<br>in car<br>in bike | signal<br>crossing<br>on road<br>parking | pick up<br>drop off<br>both<br>idle |
| Golden Eagle migration | Mapping | Gender | Age | Terrain | Activity |
| | No of Instances | 2 | 3 | 4 | 4 |
| | Instances | male<br>female | old<br>child<br>young | park<br>hill<br>lake area<br>high way | sleeping<br>running<br>idle<br>breeding |
| Caribou movement | Mapping | Gender | Age | Terrain | Activity |
| | No of Instances | 2 | 3 | 4 | 4 |
| | Instances | male<br>female | old<br>child<br>young | lake<br>grass land<br>urban region<br>road | sleeping<br>running<br>idle<br>breeding |
| Fish movement | Mapping | Gender | Size | Weight | Species |
| | No of Instances | 2 | 3 | 4 | 4 |
| | Instances | male<br>female | small<br>medium<br>large<br>extra large | light<br>medium<br>heavy<br>extra heavy | Lake trout<br>Northern pike<br>Burbot |

Figure 22: Description of objects' dimensional values.

### 4.1.3 *Golden Eagle migration data*

This data set contains Golden Eagles' movement behaviour during a migration period in western North America and is available online for public use [43]. There are 2,708 trajectory points for 43 Golden eagles in this dataset which were recorded between July-1997 to August-2000. This dataset is considered as the high density. In the original dataset, there are no dimensional values of Eagles. I ran-

domly added these dimensional values to make the analysis and visualization more interesting. Four high level properties (i.e. gender, age, terrain, activity) of this data set are listed in Figure 22. This dataset was also used in the experiment. The reason for using this dataset in the user study is that it's a real dataset and have high density.

### 4.1.4 *Caribou movement data*

This dataset has been collected from the NRI department at the University of Manitoba. Dr. Micheline Manseau and her research group are currently using this dataset to solve various research questions. The data set is divided into two subsets containing data for the time periods 1992-1996 and 2005-2010. In total there are 0.2 million movement records of 66 Caribous. The movements were recorded by tracking the animals in some parks in Saskatchewan. The dimensional values (gender, age, terrain, activity) of Caribou are shown in Figure 22. I used this dataset for case studies. I explored different research questions (Figure 24) that Dr. Micheline Manseau is currently working on and provided a guide to solve those questions using my proposed FlockViz technique.

### 4.1.5 *Fish movement data*

The researchers in the Biological Sciences department at the University of Manitoba are working on fish movement data in the Alexie

lake of the Northwest Territories, Canada. This dataset is also very large in size. The researchers collected about 20 positions of each fish for every hour. It has 0.5 million position records of 40 different fish over one and a half month period (from 14th September 2012 to 26th October 2012). The dimensional values ( gender, size, weight, species) of fish are shown in Figure 22. Dr. Paul Blanchfield and his student David Callaghan from the University of Manitoba are currently working with this dataset to study group movements of fish, their re-production process, their activities in a particular lake region etc. I explored some of their queries in the case study section using FlockViz and my proposed visual solution helped them to identify different statistics of fish movement.

## 4.2 TASK CATEGORIES

Researchers have studied Spatio-Temporal Clustering (STC) from different perspectives. Most of these research works [64, 3] are focused on solving specific problems rather than evaluating the visualization technique. For example, Guo *et al.* [22] evaluated their technique by solving problems in traffic analysis domain. Rao *et al.* [48] proposed a wide range of tasks related to STCs but did not categorize them into a general framework. I fill these gaps by proposing a framework for task categories covering questions that can be asked in the STC analysis.

While designing task categories for the STC analysis, I applied two different approaches and later generalized the categories by merging the two into one. In the first approach I categorize the tasks into the

following categories based on the common information of interest to users:

- Basic Information (BI) from data files: This includes spatial (i.e. latitude, and longitude), temporal, and attributes (i.e. dimensions of objects).
- Derived Information (DI) from BI: These are attributes we can derive from BI such as speed, and direction.
- Generated Information (GI) by users during exploration: Informations concerning users' drawn regions, path, users' defined time segment, etc.

Since I do not perform post-processing on data, I assume DI and GI are generated and available in advance as attribute values in BI. Hence, in my task category design, I have a spacial (S), temporal (T) and attributes (A) as pieces of information where A can be anything related to objects' properties and can be from DI and GI. Considering one or two BI unknown at a time we can get total 6 combinations.

Once I have these basic pieces of information to work with clusters, the next step is to define how the aggregation should be designed. I collected and classified several common aggregation functions (i.e. count, sum, average, minimum, and maximum). There could be one or more cluster at a time for which users will perform analytic tasks based on these aggregations. Hence I categorized these aggregation functions into two high level groups:

- Distinct Aggregation (DA - sum, count, and average) which is focused on Single Cluster (DAS).

- Comparative Aggregation (CA - min, and max) which could be for both Single (CAS) or Multiple clusters (CAM).

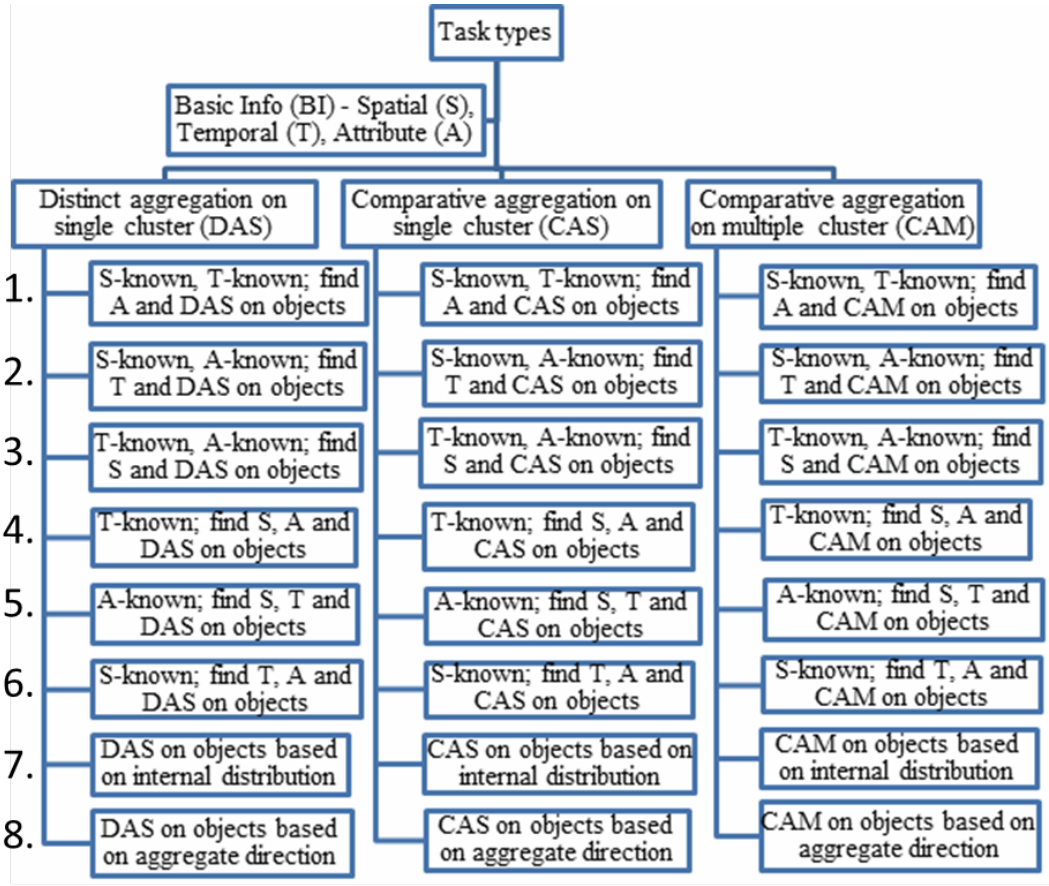



Figure 23: Framework of task categories for spatio-temporal cluster analysis.

Therefore, I have three high level aggregation tasks that I can perform along with the previous 6 combinations. Using this structure, I can define 18 distinct tasks for each dimension of objects. Based on the dataset this can be multiplied by the number of objects' properties. For example, let's take a case where "S", and "T" are known and "A" is unknown and we want to find a CAS on "activity" dimension. If we consider the Hurricane dataset, one possible task fulfilling this criteria will be to find a cluster near Bermuda (S -

known) which happened during August-2005 (T-known) and in this particular cluster (CAS) what type of hurricane movement (A - unknown) happened most?

| Dataset | Research questions ( NRI and Biological Sciences departments of the University of Manitoba are currently studying) | Classification according to framework | Explanation of classification |
|---|---|---|---|
| Caribou | Did the movement of individual animals change between two time periods? | CAM-4 | T-known, asking in general the comparison of pattern of S and A between clusters |
| | Did the extent (annual home range) of their movement changed between those 2 time periods? | CAM-3 | Same as the previous but only status of S is asking |
| | Did some unique characteristics of their movement change e.g. are the movements more clustered now, are the movements more clustered now only at certain times of year? | CAM-1 to CAM-7 | T may or may not be known, asking for general change in cluster which could be any informaiton |
| | How much fidelity do animals present to calving sites (15 May - 15 June), late winter foraging areas (15 March - 30 April), or rutting areas (15 Sept - 30 Oct) between years? | Mix of DAS-1,7,8 and CAM-1,7,8 | S-known and T-known, asking for mainly cluster shape, direction and distribution for individual cluster on different years |
| | Do animals tend to avoid roads and cut block in the same way between different time periods and does that differ between different times of year? | Mix of DAS-1,7,8 and CAM-1,7,8 | T-known, S-known and it is asking for cluster behaviour of movement comparison |
| | Do the movement characteristics of animals differ between times of year or for animals of smaller and larger home ranges? For animals locate in more fragmented areas? | CAM-1 to CAM-7 | T may or may not be known, asking for general change in cluster which could be any informaiton |
| Fish | When and how long fish gather in a particular areas of the lake? Compare | DAS, CAM-6 | S-known and asking for temporal information |
| | Is there any change in fish movement behavour between day and night time? | CAM-4,7 | T-known, asking for comparison of different clusters |
| | Is there any scenario of cluster formation where fish go back and forth often or just stay in that area? | DAS, CAM - 8 | Asking for directional movement of a single and surrounding clusters |
| | What is the difference between male and female fish movement behavour? | CAM-5,7 | A-known, asking for comparison of different clusters on various properties |
| | What are the different statistics of each cluster of fish in particular region? Is there male of female dominant? Any variation in terms of size of fish present there? | DAS, CAS-6,7 | S-known, find different statistics on a single cluster |

Figure 24: Real analytic and research questions related to spatio-temporal clusters and their classification into task categories using my proposed framework.

My second approach to formulate the task categories is to identify what are the most desired properties of the clusters for visualization.

Then we can identify tasks for each property. The most important properties in the literature are: cluster density, internal distribution, movement direction, aggregate measures on object dimensions, and formation time.

Most of the analytic tasks in the first approach cover these properties. In an effort to merge the two approaches, I found two other task sets (i.e. aggregate direction, and internal distribution) which cannot be addressed by previous approaches. By multiplying these two with the three high level aggregation tasks we get 6 additional tasks. The result is a total of 24 different questions for any clustering scenario possible with this framework as shown in the Figure 23. In this figure first six task categories under each high level task (DAS, CAS and CAM) directly come from the first approach and the last two tasks basically cover the gaps between first and second approach.

To validate my proposed task categories, I considered to categorize the research and analytic questions used in the NRI and Biological Sciences departments at the University of Manitoba. Most of these questions can be classified using my proposed task category framework. One important observation while analyzing these real life analytic questions is that we cannot directly classify them into a specific task category because most of the questions are very high level and cover more than one category at a time. However, each question can be classified into a set of tasks from my proposed framework. In Figure 24 I present all these research and analytic questions and corresponding classification of my proposed task category.

## 4.3 EXPERIMENTAL DESIGN

To evaluate FlockViz and show the importance of cluster visualization, I conducted a user study. In this study I used the traditional STC visualization similar to Geotime [31] as a baseline. In the traditional visualization (TraditionalViz), only the cluster area and the consisted objects were shown and hovering over a cluster triggers a display of all the information for the cluster properties (Figure 25(a)). In FlockViz, visual cues represent different properties of the cluster (Figure 25(b)).

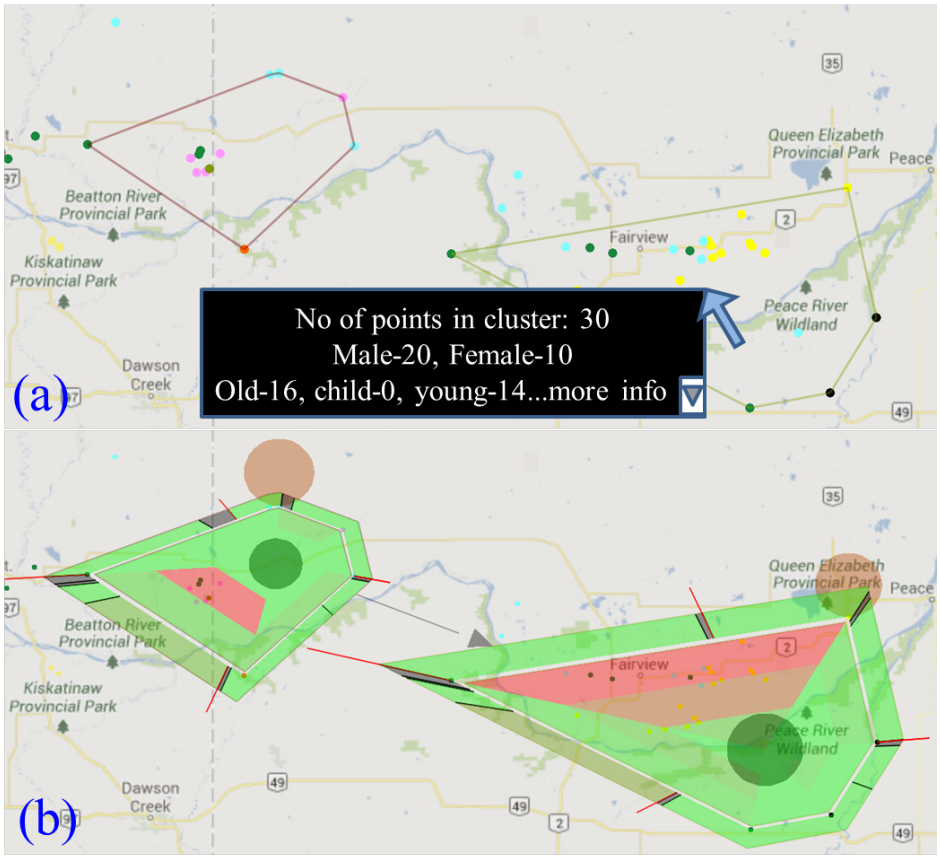



Figure 25: Experimental setup for two different visualization techniques of spatio-temporal cluster.

I used the Hurricane and Eagle datasets described in section 4.1 with two levels of density (low, and high) for the actual experiment and the People dataset for training participants. There were two high level tasks (CAS, CAM) as discussed in section 4.2 in this experiment. The tasks of DAS category was not used in the experiment as it is related to finding an exact numeric value of a particular cluster property. However if we want to include DAS then we will have to provide a pop-up window in FockViz similar to Figure 25(a) and in that case both techniques would have the same options to answer these questions. Therefore, I excluded this type of question from the experiment. Finally in the experimental design the independent variables were techniques ( FlockViz, TraditionalViz (Figure 25)), dataset density (low,high), and task category (CAS,CAM) and the dependent variables were task completion time, and number of errors.

The experiment was performed on a Dell OptiPlex 9010 Desktop Computer (Intel Core i7 3.4 GHz CPU, 16 GB RAM, an AMD Radeon HD 7570 graphics card) running Windows 7 Professional, connected to a Dell 23 inch monitor. Participants interacted with the visualization using a mouse. For FlockViz only control panels marked by (c) in Figure 26 were visible to interact with different flock parts. For TraditionalViz, the control panel marked by (b) was only visible to show or scale objects within clusters. Each question was displayed at the top of the visual panel marked as (d) in Figure 26 and there were four multiple choice answers at the top-left side of the window (a). I conducted my experiment with 16 participants (ages between 15 to 35 years) majoring in computer science.


Figure 26: Different parts of the user interface in the experimental environment.

For designing questions in the experiment I divided CAS and CAM into density, aggregate measures of attributes, temporal information and internal distribution based questions which cover the first 21 questions in my proposed task category framework. I did not include any question related to direction because using traditional visualization it is almost impossible to find aggregate direction from the movement of all trajectories. I also did not include questions related to density and temporal information in the CAS category because for a single cluster those questions will be similar to DAS

| Task category | Task on | Sample question |
|---|---|---|
| **CAM** | Density | Which cluster has largest number of objects? |
| **CAM** | Internal distribution | Which cluster has objects well distributed over all cluste area? |
| **CAM** | Temporal information | Which cluster took longest duration? |
| **CAM** | Aggregate measuresl | Which cluster has largest number of objects of X type? |
| **CAS** | Internal distribution | In cluster X where most of the objects were located? |
| **CAS** | Aggregate measures | In cluster X which type of objects are greater than all? |

Table 1: Sample questions for experiment

type. Therefore, in total I had 6 different questions under two high level tasks (4 questions for CAM and 2 questions for CAS category) and I generated variations of these questions for two datasets and two techniques. In total I asked 24 ($6 \times 2 \times 2$) questions per trial and 12 (for a single dataset) questions for a practice session. Examples of these questions are given in Table 1.

## 4.4 EXPERIMENTAL RESULTS

I used an Analysis of Variance (ANOVA) test at the significance level of $\alpha = 0.05$ using the Bonferroni adjustment to carry out all the statistical analysis for this study. There are no transformations (e.g. logarithmic transformations) applied to the data.

The within-subject variables were visualization technique (i.e. FlockViz vs TraditionalViz), data density (i.e. low vs high), and

high/low task categories as described above. The dependent variables were task completion time or duration, and number of errors. The latter was measured in terms of the error rate. All the independent variables were counter balanced to avoid learning effects and effects on the performance under each condition.

### 4.4.1 *Task Completion Time*

The mean duration of time to completion of tasks for each technique is shown in Figure 27(a). There is an overall main effect of visualization technique on the task duration ($F_{1,15}$ = 8.91, P = 0.02) with FlockViz performing significantly faster (P = 0.001). Therefore, this result of Visualization technique vs Task completion time shows:

- In overall, with FlockViz users complete the tasks significantly faster than TraditionalViz.

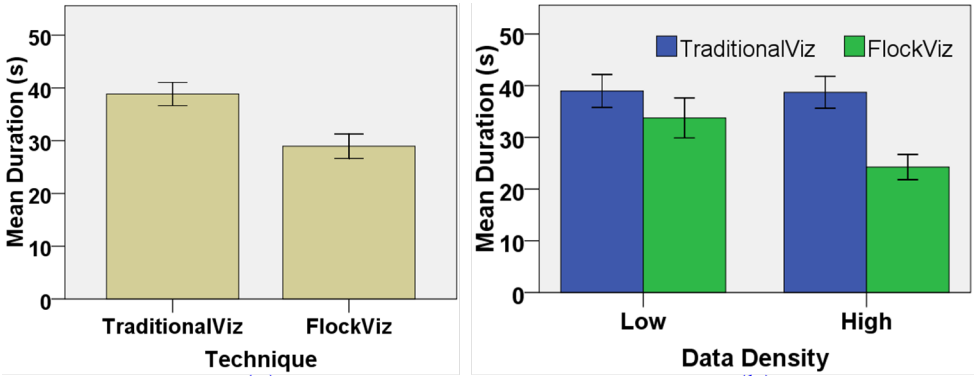



Figure 27: (a) Visualization technique vs Task completion time. (b) Data density vs Task completion time.

There are no main effects of dataset density for the task duration. However, there is a significant interaction effect of data density ×

visualization technique ($F_{1,15}$ = 8.716, P = 0.021). It is interesting to note that FlockViz showed to be significantly faster with highly density datasets ($F_{1,15}$ = 11.896, P = 0.001) whereas with the low density datasets, the visualization techniques showed no significant differences for the task duration time (Figure 27(b)). This makes sense considering the fact that cluster visualizations would not be as useful and effective if there are not many objects and trajectories to form clusters. Therefore, we can conclude that:

- Data density has no effect on task completion.
- FlockViz complete the tasks on high density data significantly faster than TraditionalViz.

Further analysis of the results to investigate the effect of task categories reveal the main effect of high level task categories on the task duration ($F_{1,15}$ = 22.972, P = 0.002) with CAS takes a significantly lower amount of time to finish ($F_{1,15}$ = 13.908, P < 0.001).There ware no overall main interaction effect of visualization technique $\times$ high level task categories (Figure 28(a)). Therefore, I further analyze the tasks through the low level task categories to understand the possible reason. But this result of High level task category vs Task completion time shows:

- CAS type tasks take significantly lower amount of time to complete than CAM type tasks.
- In overall, visualization techniques $\times$ high level task categories have on effect on task completion.

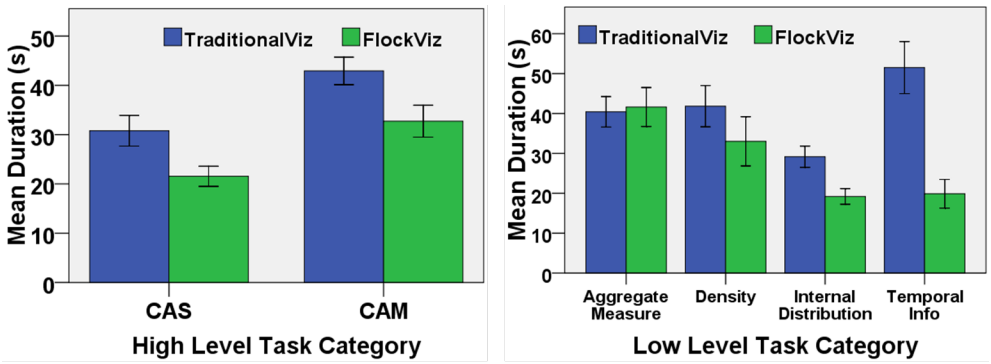



Figure 28: (a) High level task category vs Task completion time. (b) Low level task category vs Task completion time.

The results of the task duration in the breakdown of high level task categories to the low level task categories is depicted in Figure 28(b). There is an overall main effect of low level task categories on task duration ($F_{3,45}$ = 15.767, P < 0.001) with "Internal Distribution" tasks being significantly faster than other tasks in the low level task categories ($F_{3,45}$ = 8.286, P<0.001). There is also an interaction effect of visualization technique $\times$ low task categories ($F_{3,45}$ = 5.93, P = 0.004). In particular, it is interesting to see that FlockViz performed significantly faster than TraditionalViz in "Internal Distribution" tasks ($F_{3,45}$ = 3.866, P = 0.051) and "Temporal Info" tasks ($F_{3,45}$ = 20.197, P < 0.001). However, the task related to "Aggregate Measure" nullified the main interaction effect of high level task category. Therefore, we can conclude that:

- overall, low level task categories have significant effect on task completion time.
- "Internal Distribution" tasks take significantly lower amount of time than other low level task categories.

- FlockViz performed significantly faster than TraditionalViz on completion of "Internal Distribution" tasks.

To further investigate the effect of "Aggregate Measure" related tasks, I separately analyzed the result under CAS and CAM high level category. The results show that under CAS category, TraditionalViz significantly performed better than FlockViz but under the CAM category there is no significant effect. The reason for this result is the pop-up information of TraditionalViz was enough and faster to answer CAS type questions for a limited number of options (there were maximum 4 options for each question) in the question - answer panel. As we do not need to go back and forth between multiple clusters for answering these questions TraditionalViz performed better than FlockViz. However, if the number of possible options increases there will be a chance that TaditionalViz will not perform better than FlockViz in this type of question.

### 4.4.2 *Error Rate*

Figure 29(a) shows error rate averages for each visualization technique. Overall, low error rates of less than 15% were measured across all conditions. I observed an overall main effect of visualization technique on error rate ($F_{1,15} = 19.305$, $P = 0.003$) with FlockViz performing significantly better (i.e. Lowest error rate) in comparing to TraditionalViz. I did not find any main interaction effects of visualization technique $\times$ dataset density nor across any of the task

categories (Figure 29(b), Figure 30). The detail experiment results are given in the appendix section. This result on error rate shows:

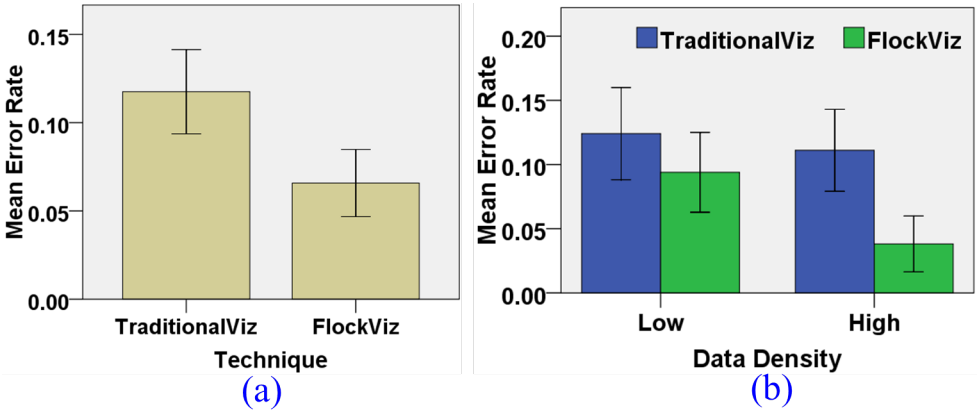



Figure 29: (a) Visualization technique vs Error rate. (b) Data density vs Error rate

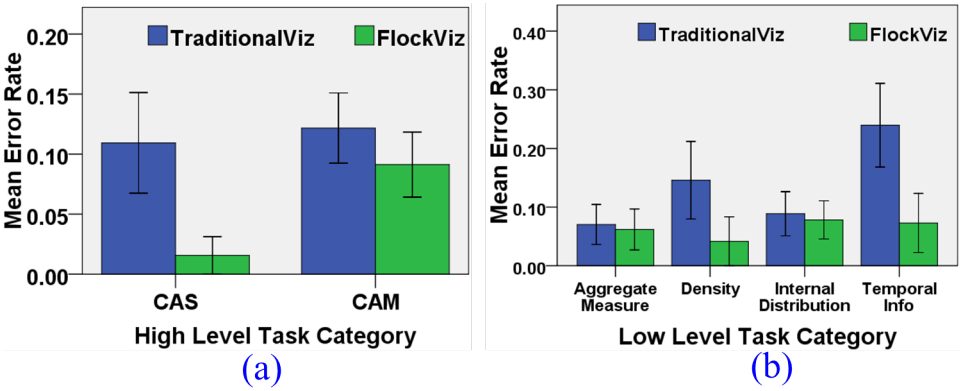



Figure 30: (a) High level task category vs Error rate. (b) Low level task category vs Error rate

- overall, visualization technique has significant effect on error rate where FlockViz has significantly lower error rate than TraditionalViz.

### 4.4.3 *Subjective Feedback*

At the end of the experiment I asked participants to rank each technique for each question type based on their preference where the level of preferences are 1 (Not preferred) to 5 (Preferred). In Figure 31 the average rating of each question ranked by the participants are shown where questions 2 to 7 are same as the questions in Table 1. The first question was on overall preference and the last question was on overall difficulty of using these two visualization techniques throughout the experiment. For difficulty, 1 represents very difficult and 5 represents very easy.

Overall, these results empirically demonstrate the effectiveness of using FLockViz for Spatio-Temporal Cluster analytic tasks. I also demonstrate the additional value this technique provides as I describe case studies in the next chapter.

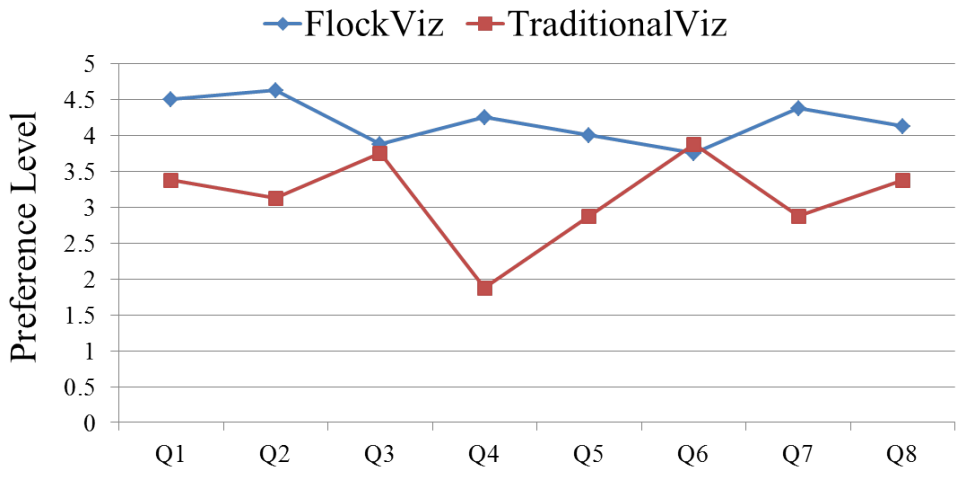



Figure 31: Average preference rating by users of each distinct question type

# 5

# CASE STUDIES TO EVALUATE THE PROBLEM SOLVING CAPABILITIES OF FLOCKVIZ

FlockViz is designed to assist with analytic tasks on spatio-temporal clusters (STC). The ability to show multi-dimensional attributes with FlockViz in a single view offer a significant advantage during analysis. Since FlockViz uses several components to visualize the different parts of the flock, users can apply rich filtering options to enable or disable different data attributes simultaneously. These unique features of FlockViz ensure to have a high level of aggregated information for each cluster and keeps the original trajectories' positions intact. In this chapter, I demonstrate how these FlockViz features can help in solving research questions for various data domains. The datasets are described in section 4.1. I have also included comments from domain experts at the end of the first two case studies.

## 5.1 CASE STUDY 1: UNDERSTANDING HOW FISH BEHAVE DURING SPAWNING SEASON

Preservation of natural habitat in lakes is an important part of our overall ecology. A significant element is the healthy development of fish in this ecosystem. Having such knowledge assists with controlling the fish population at various periods. Keeping these objectives

in mind, biologists track movements of fish in different lakes to conduct experiments. Since they are interested in analyzing movement behaviour, they need to visualize the data to get meaningful information. However, when the dataset becomes very large in size, a traditional approach for trajectory visualization is not sufficient to answer their queries. In this case study I will show how FlockViz can help to answer some of these questions. The data set for this case study was described in section 4.1.5 and a general trajectory visualization of this dataset is shown in Figure 32. The figure shows fish movement during one week (18th September, 2012 to 25th September, 2012). Each color represents a particular fish. There are only 40 fish in this dataset, but the movement data is enormous as the sampling rate is 20 position records per fish per hour. This visualization is practically unusable. The research questions obtained from researchers in the department of Biological Sciences at the University of Manitoba concerning this dataset are listed in section 4.2. The use of FlockViz for analysis of this data is shown as follows.

The first research question is "What are the times, duration and location of fish gatherings? Is there any pattern for those gatherings such as overpresence of male fish? How do these compare between two regions." We show how FlockViz can be used to answer these questions through the following example.

Example: In applying the FlockViz approach to the fish dataset [section 4.1.5] we first identify the criteria to filter the clusters. We choose a the minimum number of points we wish to see in a cluster (i.e. 10 as an example and a maximum distance between two points to be 15 meters. We do not select any parameter for temporal informa-

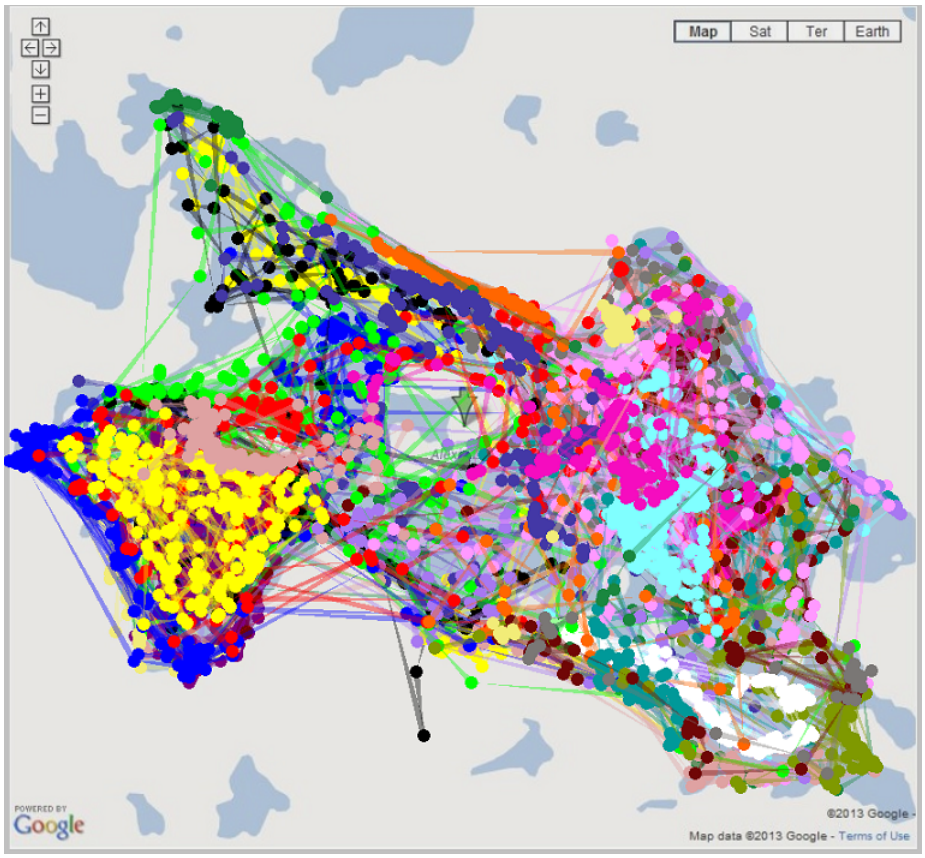



Figure 32: General trajectory visualization of fish dataset in lake Alexie of NT, Canada. Each color represents a particular fish. Triangular edges from one point to another show their movement direction.

tion as there is no constraint on time for generating these gatherings. The initial output of FlockViz is shown in the Figure 33. If we look carefully at this output we will find some important properties of fish gatherings in that lake. As FlockViz preserves the original cluster shape, we see that most of the clusters cover a limited geographical area and those events mostly happen near the lake shore. However, the density map of the *flock body* shows that almost all clusters have some dense (red) internal regions implying that in each cluster, there are fish that stay very close to each other. Therefore, these visual

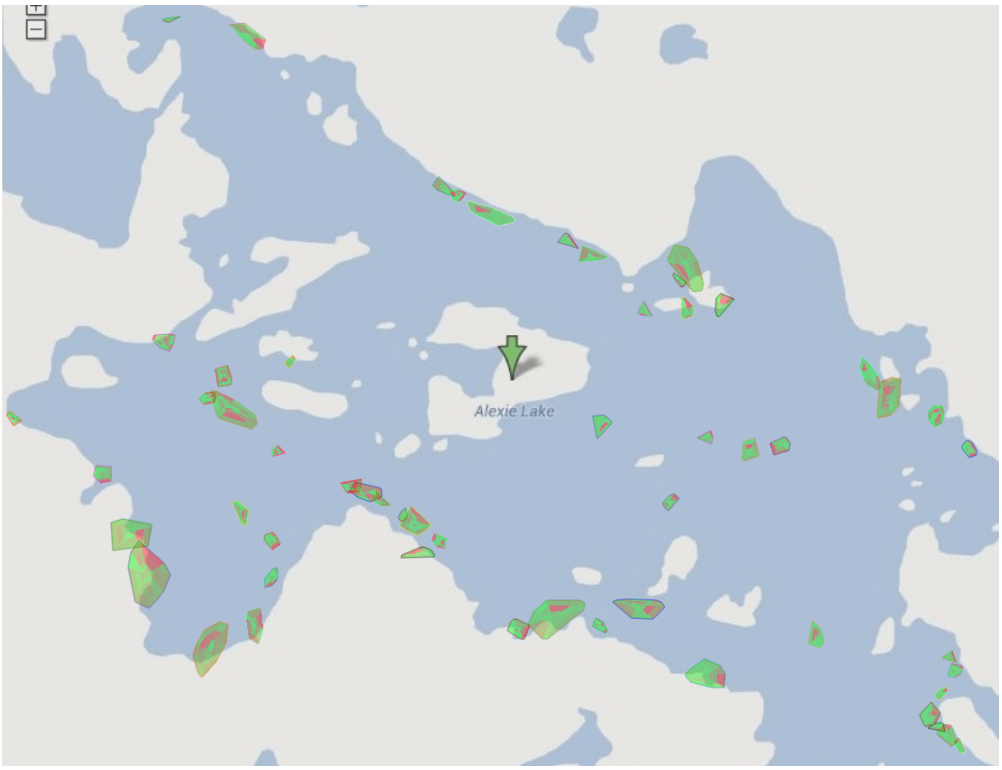



Figure 33: Output of FlockViz as a set of clusters after applying a clustering algorithm.

outputs can explain if any particular area of the lake has any cluster of fish and how the cluster is structured.

To find the temporal information, we can enable the alternate design (analogue clock) of showing temporal information in the *flock heart*. The output is shown in Figure 34. Using the clock design (small leg and long leg to represent the start and end times, respectively) we can tell the relative duration and recency of each cluster formation where the reference time for the start point is 18th September 2012 and the reference time for the endpoint is 25th September 2012 ( as per the dataset). In this design, both the start and end reference points are at the 12 o'clock position. This example output of FlockViz shows that most of the clusters near the lake shore regions start


Figure 34: Analogue clock in each *flock heart* shows the temporal duration and recency of the cluster.

gathering earlier and keep the gathering for a long time. Specially in the bottom-left corner of the lake the cluster spans throughout the whole timespan. This output can also be used to compare between two different regions of the lake. For example, in the middle and top-middle of the lake the cluster formation times are not consistent. These clusters start generating lately and take a small amount of time to form. During these time periods very few clusters are formed in those regions. But in the bottom-left and bottom-right regions of the lake there are some gatherings of fish during the whole time period. We can explore more findings by changing the input parameters of cluster generation and by filtering different properties of the fish such as using particular gender, species etc.


Figure 35: Gathering of fish during day time [06:01 to 18:00].

The second question about this dataset is: "Is there any change in fish's group movement behaviour between day and night time?". This question can have a very broad range of answers because any change in the properties of the cluster for each time segment can be incorporated in the result. To answer this question at first I set the basic clustering parameters same as the previous question. Then I apply temporal filtering where only day time (considering 06:01 to 18:00) for all days are considered. After setting all these parameters we get the generated clusters as shown in the Figure 35. The same thing is done for the night time (18:01 to 06:00) as shown in the Figure 36. These two outputs give a comparative picture of fish movement during the day and night. The figures show that during the day, fish make groups in different regions of the lake and the

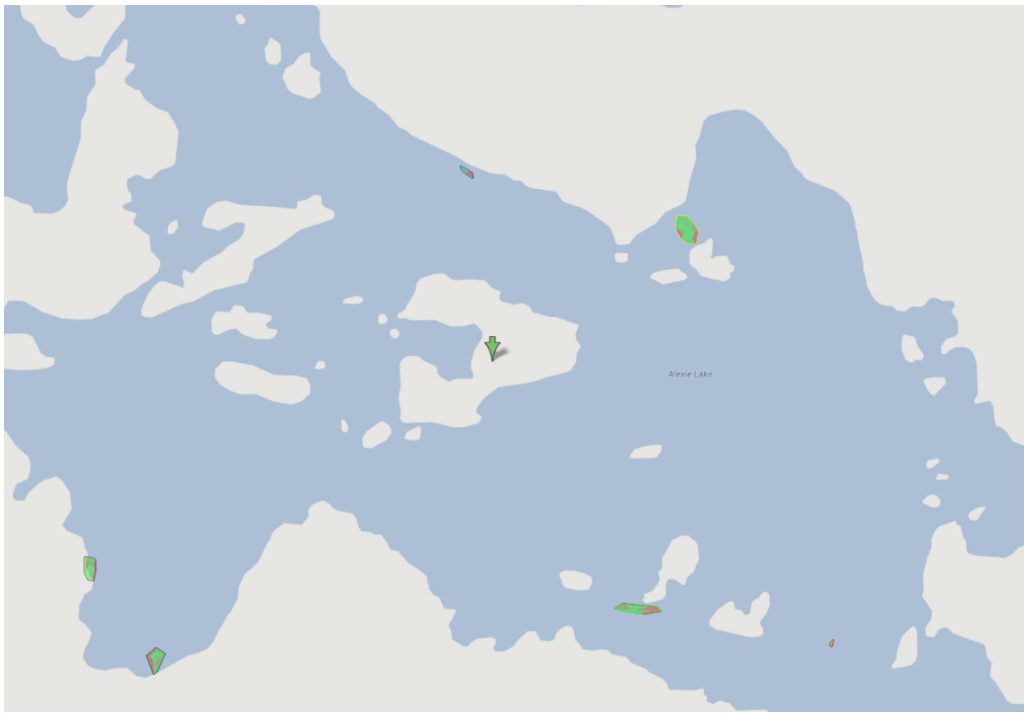



Figure 36: Gathering of fish during night time [18:01 to 06:00].

grouping is scattered throughout the lake but during the night fish gather in some particular areas of the lake. One reason for this difference can be that most of the fish prefer to stay alone at night.

Now we can use the capabilities of FlockViz to show multiple information in a single view and explore group movement behaviour of fish during the day and night time. I will show one of those different capabilities of FlockViz to analyze a few cluster properties. Assume that we want to know the statistics of different species of fish gatherings in combination with their gender information during the day and night. To explore this movement behaviour, we can use a combination of *flock shield* and *flock tail* where *flock shield* can show gender information and *flock tail* can be used to highlight particular species and vice versa. Here, we use the alternate design of *flock shield*

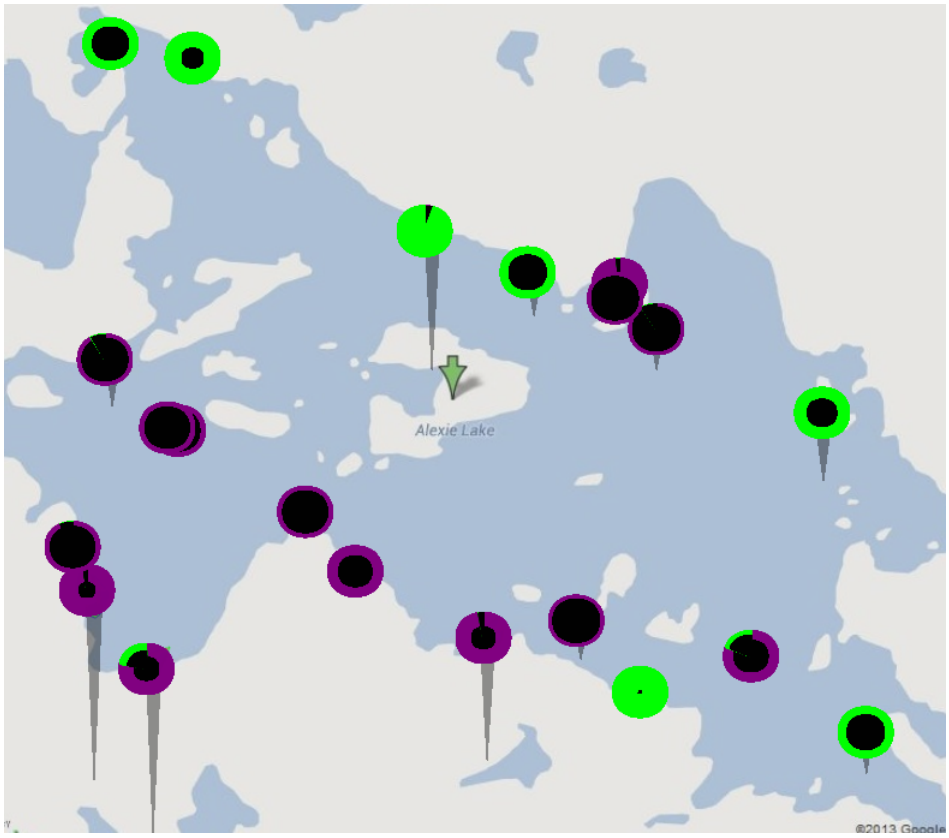



Figure 37: Statistics of "Lake Trout" fish movement along with their gender information during the day [06:01 to 18:00].

as pies where the segments of each pie represents the number of fish of a particular dimensional value. Assume we want to investigate the movement of a particular species (Lake Trout). Figure 37 shows clusters with *flock shields* (represented as pies) and *flock tails* during day time. Figure 38 shows the same visualization for night time. Purple color in each pie is used to represent the number of male fish and green color is used for the number of female fish. The tail length shows how many "Lake Trout" were in that cluster. The Figure 37 shows that most of the male fish gather at the bottom-left region of the lake whereas the female fish gather at the top-left region during the day. But during the night time the gatherings are reduced and

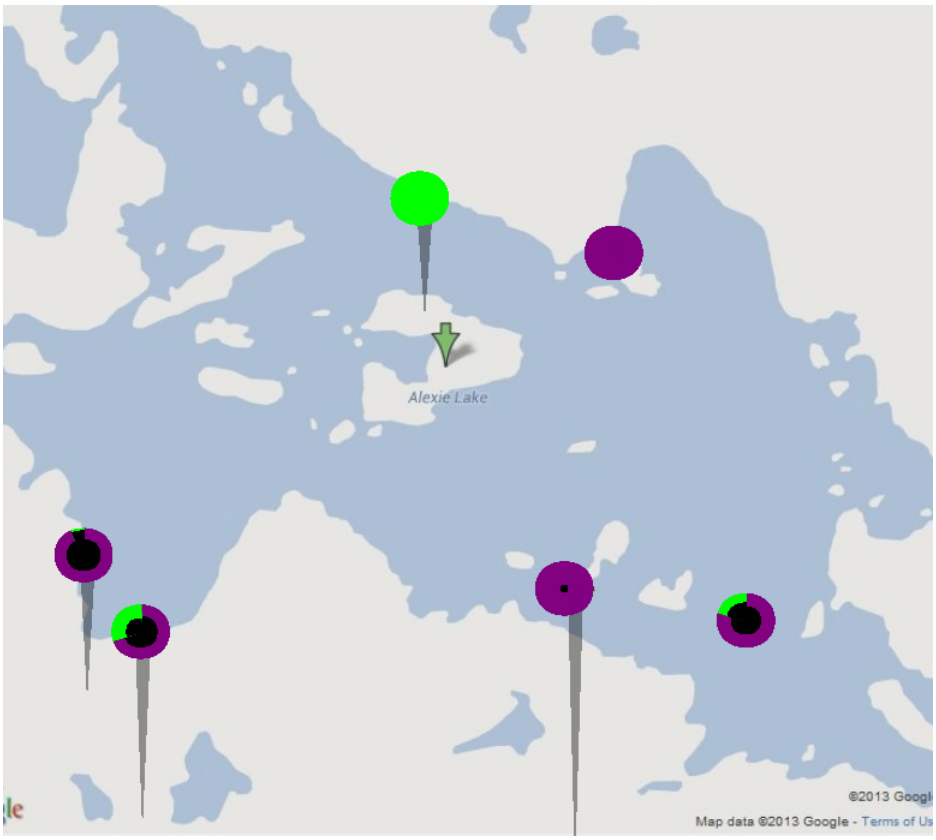



Figure 38: Statistics of "Lake Trout" fish movement along with their gender information during the night[18:01 to 06:00].

female fish leave the top-left region of the lake and get scattered in different regions of the lake without forming any cluster (Figure 38). In general, most of the male fish stay together at the bottom of the lake and female fish at the top-middle of the lake during both the day and night time. "Lake Trout" keep the gatherings consistent both during the day and night time as there is not that much difference between the lengths of *flock tails* in these two images. Other properties of fish such as length, weight based day-night movement can also be explored using the same approach.

The next research question is: "Is there any scenario of cluster formation where fish move back and forth between clusters or are


Figure 39: Visualization of aggregate movement of fish between clusters using *flock wings*.

they sedentary, i.e. just stay in one area?" To answer this question at first I generate clusters with normal input parameters like the day time clustering scenario of the previous question. Then I can activate the *flock wings* of FlockViz which will show if any fish is moving out of the cluster to another (an out directional arrow) or the fish are staying there (no out directional arrow) or the fish are moving in another separate destination (no arrow). Figure 39 is the output that shows all these possible scenarios. According to the *flock wing* design, the width of the arrow head shows how many fish are moving in that direction. In the Figure 39 we see that in the top-middle and bottom-left corner regions of the lake there are many bi-directional arrows that means fish are moving from one group to another in those regions frequently. However, as the arrow

heads' sizes are almost the same, we can say that in those regions the same number of fish are moving back and forth between clusters. Therefore, the clusters are not stable in those regions. On the other hand on the top-left corner and middle-left corner of the lake there are uni-directional arrows which mean fish travelling to that cluster stay longer. One reason of this scenario could be they feel safe in those regions which made them spend more time there. Another explanation has to do with wind direction, which also dictates their spawning behaviours. Other clusters on the bottom-right side of the lake has no arrows which means the fish are not coming here directly from another group. They have travelled a lot to reach here and when they leave this region, they also do that individually, not in groups.

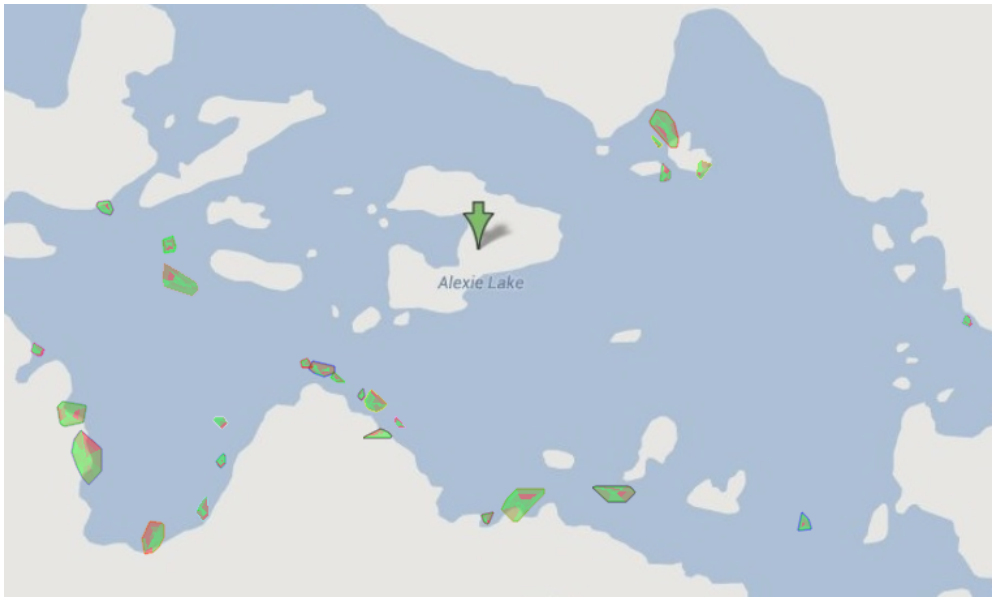



Figure 40: Clusters of male fish.

The fourth question is "What is the difference between male and female fish movement behaviour?" This is very similar to the sec-

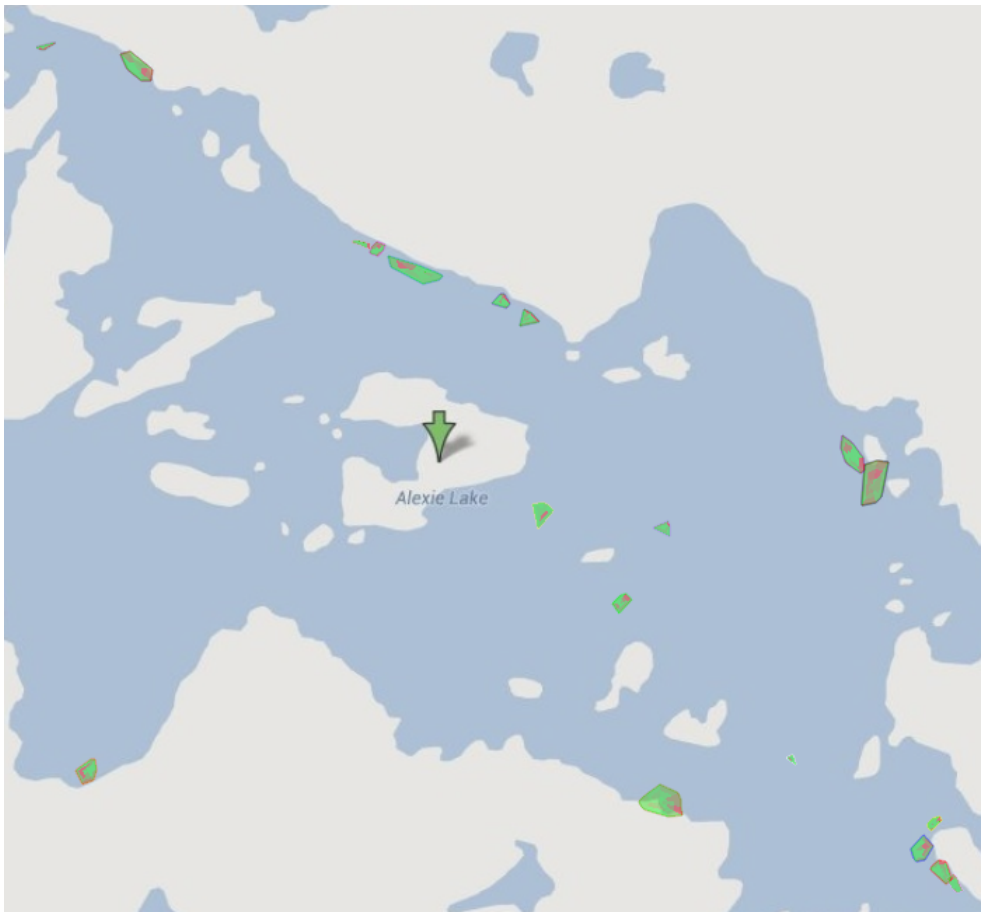



Figure 41: Clusters of female fish.

ond question but instead of filtering for time we filter for the sex dimension for generating clusters. By setting the same parameters and applying male and female filter we get the clusters as shown in Figure 40 and Figure 41 respectively. It is interesting to see that male fish mostly gather at the bottom-left side of the lake where the female fish stay together at the top and bottom-right regions of the lake. Besides the internal density map of the cluster of male fish is more uniform than the cluster of female fish as there are many dense (red) regions within a cluster of male fish. This implies that male fish get scattered within the group whereas the female fish tend to make

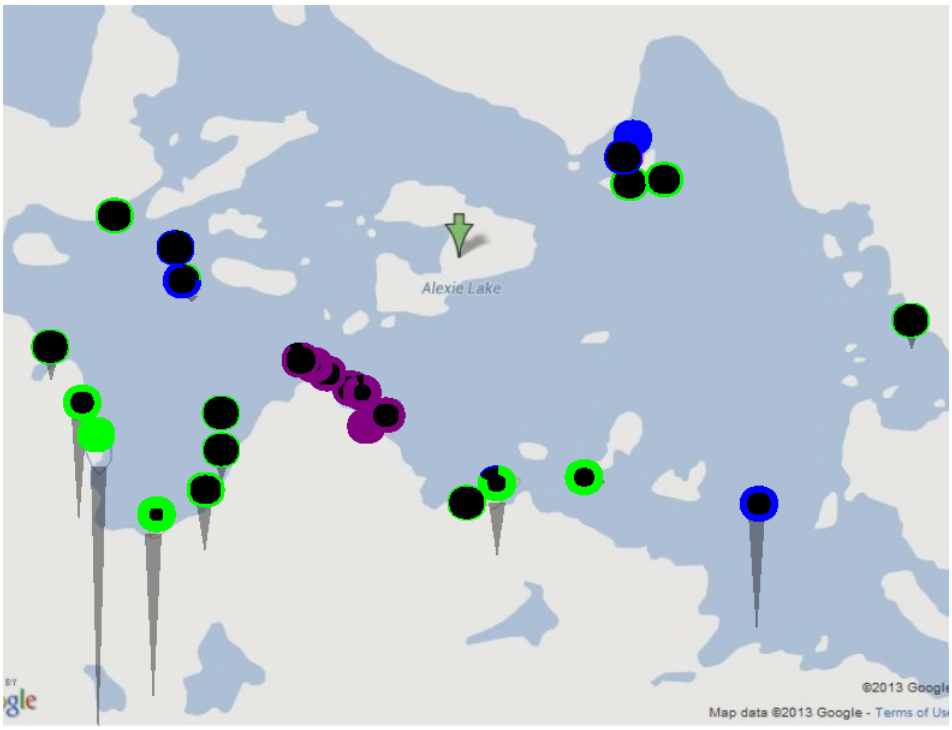



Figure 42: Cluster of male fish where *flock shields* as pies show different species and *flock tails* show the number of "Heavy" fish.

internal groupings strong and choose some particular region within a gathered area.

Now similar to question 2 we can explore various options of FlockViz to identify differences between male and female fish movements. Assume we want to analyze the "Heavy" weight fish's movement along with the statistics of the number of different species. In order to do that we use a *flock shield* region as pies for representing different species and *flock tail* length to represent the number of "Heavy" fish in each cluster as shown in Figure 42 and Figure 43. Here purple, green and blue color in the pie segments represent Burbot, Lake Trout and Northern Pike species, respectively. From the clusters of male fish (Figure 42) we see that male Lake Trout mostly stay at the bottom-left region of the lake where male Northern Pike fish gather

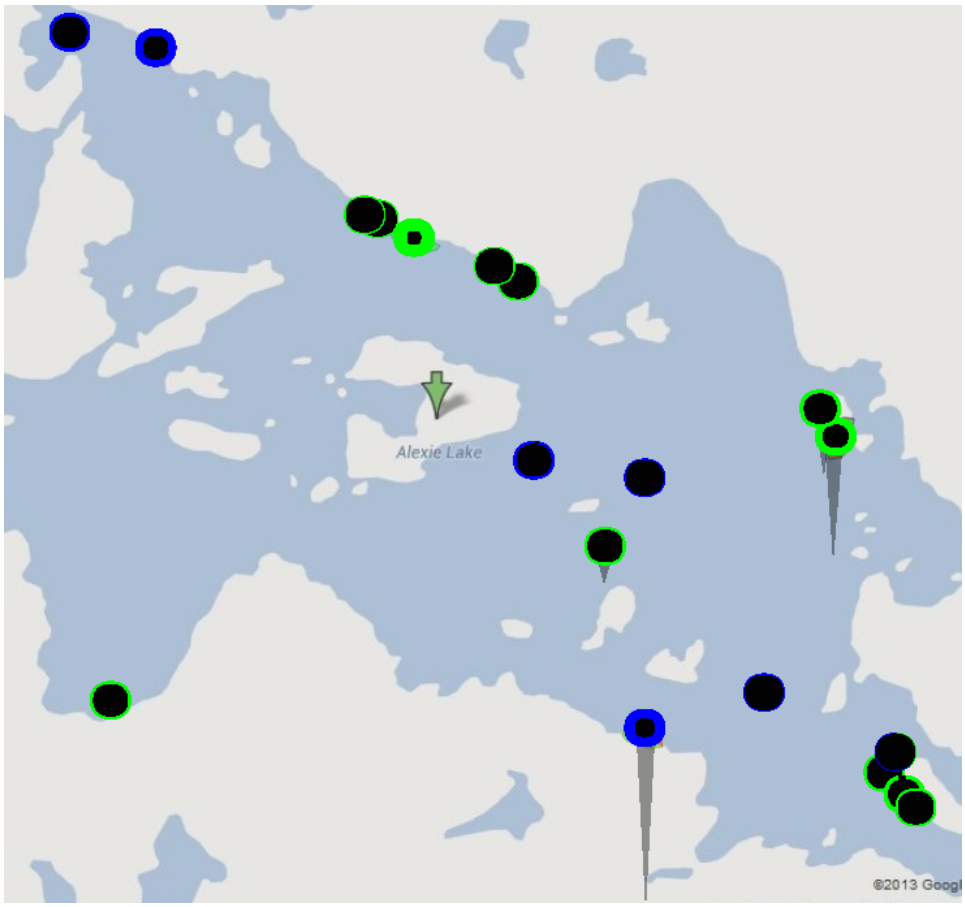



Figure 43: Cluster of female fish where *flock shields* as pies show different species and *flock tails* show the number of "Heavy" fish.

at different places. But in the cluster of female fish (Figure 43), Lake Trout are mostly found at the top and right side of the lake. Since *flock tail* represents the number of "Heavy" fish, we can say that the heavy male fish gather at the bottom of the lake but there are few female heavy fish who love to be in the group on the right side of the lake.

The last question is "What are the different statistics of each cluster of fish in a particular region? Is there male or female dominancy? Is there any variation in terms of size of fish present there?" This question has a very broad range of possible answers. In order to generate

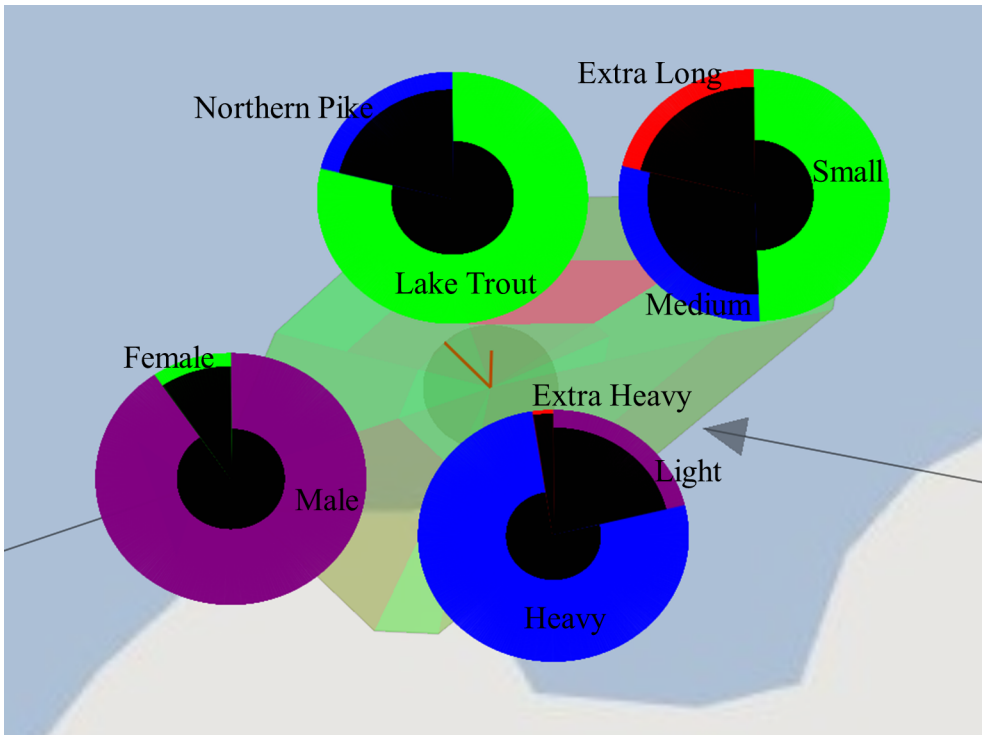



Figure 44: Visualization of different statistics of a cluster using FlockViz.

clusters for this question we follow the same basic parameter settings without applying any filter. The output of clusters will be the same as shown in Figure 33. Now we notice that this question is asking to analyze different statistics of a single cluster. Let us consider a particular cluster (at the bottom-middle region of the lake) of that whole set of clusters in Figure 33. We zoom into that cluster and enable all the features of FlockViz which are shown in Figure 44. Here we see that in a single view, various information of a cluster can be shown using the full capabilities of FlockViz. According to my FlockViz design we can identify different interesting properties of this cluster. For example, male fish are dominant in this region (the length of the purple segment of the pie is greater in the pie which represents a gender dimension) but there are other clusters in

this lake where we can find more male fish as the pie segment for the male is not filled up to the center. Similarly there are no Burbot fish here and the number of Northern Pike is less than the number of Lake Trout fish. The fish that are small in length are dominant in this cluster but most of them are heavy in weight. Other dimensional values can also be explored using the same exploration technique of *flock shields*. Now according to the circle size of the heart of this flock we see how many fish gathered in this region. We can compare this measure with other clusters by comparing circle sizes. The clock drawn in the heart shows the start (small leg) and end (long leg) time of cluster formation. From this view we can say that the cluster took a long time to form except a small gap at the end. The overall shape of the flock shows how the fish are distributed within this region ( where red represents a large number of fish and green represents a small number of fish). Therefore, at the top-middle part of this cluster many fish can be found compare to other regions. Finally the arrow shows that the fish are coming from a cluster in the right side and going from this region to a cluster on the left side. We can also use *flock tail* like the previous questions to highlight particular property of this cluster. Thus various statistics of a particular cluster in a particular lake region can be visualized in a single view using FlockViz technique.

The department of Biological Sciences at the University of Manitoba is doing research on the behaviour of fish movement. I collected this dataset from Dr. Paul Blanchfield and his Masters student David Callaghan of that department. They have used my proposed visualization system to explore all the above research questions. David

spent several hours working with different features of FlockViz to find the answers to these research questions. He was very pleased with this visualization technique and provided me a written comment about the exploration capability of FlockViz. The comments are quoted as follows:

"The cluster analysis program (FlockViz) is great for identifying areas of interest during important events such as animal reproduction. This is especially useful when studying fish, since they live underwater and are not easily located visually. By tracking the fish movements then analyzing and plotting the cluster data, we can quickly and easily determine where fish aggregations (clusters) occur and what attributes define them. This sort of spatio-temporal analysis will be incredibly useful in trying to determine fish reproductive strategies of males and females as well as how attributes such as length and weight influence these strategies to ensure reproductive success."

## 5.2 CASE STUDY 2: FINDING ACTIVITY CHANGES IN MOVEMENT OF CARIBOU

Animal migration is a very common biological phenomenon. This can dictate how the climate is affecting given geographical areas. As a result animals move to a new place so that they can survive and do their usual activities. In order to ensure a friendly environment in the new place for the animals, governing bodies in various countries take the necessary initiatives. To help make effective decisions researchers conduct a study about where the animals travel mostly, animals' activity during and after migration, where they usually group, what

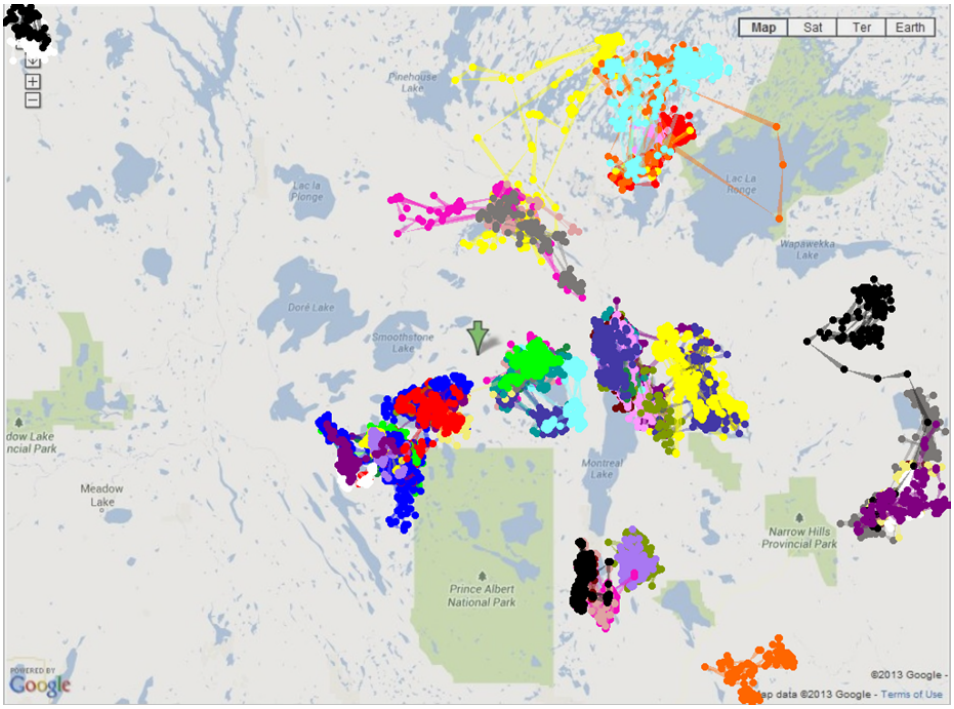



Figure 45: Caribou dataset in different park areas of Saskatchewan. Here each color represents a particular Caribou and the triangular arrow shows their movement direction.

type of animals form groups and what are their activities in a group etc. The NRI (Natural Resources Institute) department of the University of Manitoba is currently investigating Caribou movement in various parks of the Saskatchewan province. The main objective of their research is to identify the change in behaviour of Caribou movement and tracking growth of their home range. These findings will help the government to make effective decisions before making any structure in the regions where Caribou migrate, gather and do their activities. Besides different statistics of Caribou movement such as gender, age, regional movement are also important to predict their habitat so that we can restrict public movement in those habitats. I mentioned the Caribou dataset and the research questions related to

this in the 4.1.4 and 4.2 sections. In this case study I will show how FlockViz can visualize the information to solve those questions. The dataset used for this case study is shown in the Figure 45.

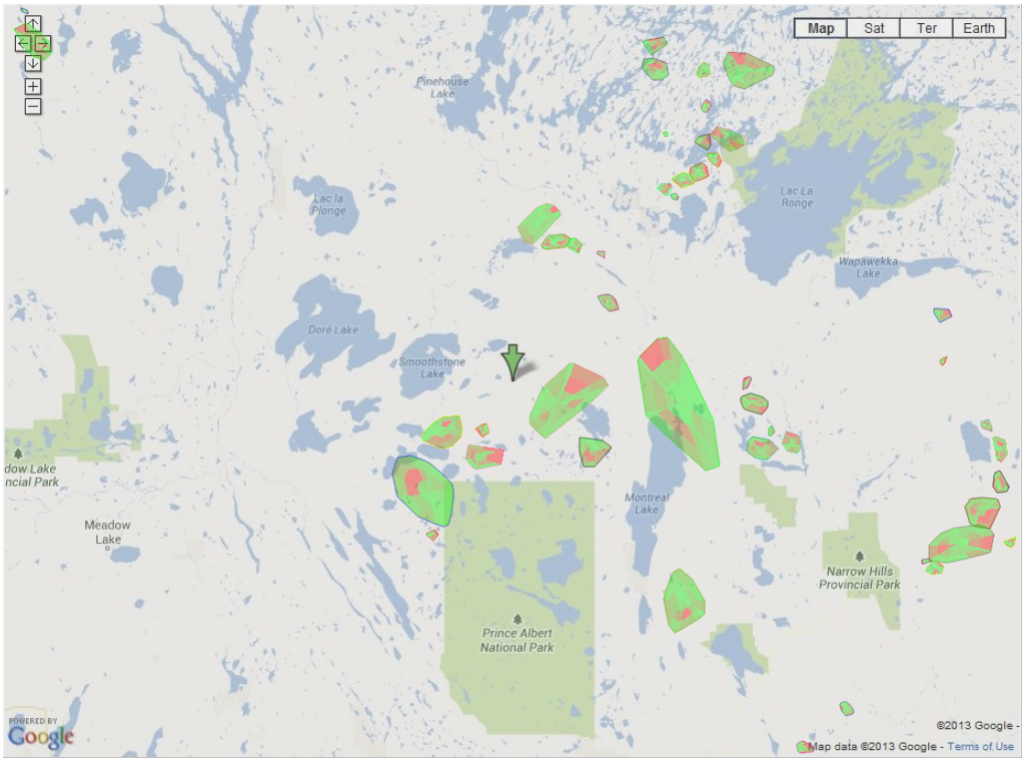



Figure 46: Group of Caribous during 1992 to 1996.

The first research question is "Did the movement of individual animals or group of animals change between 2 time periods?" As FlockViz is designed for cluster analysis we will investigate only movement of a group of animals here for this question. Here we need to compare Caribou movement between two user defined time periods. In the Caribou dataset we have two different time segments (1992 to 1996 and 2005 to 2010). Lets choose the first time period between 1992 to 1996 and the second time period between 2006 to 2010. Then we set the basic parameters of minimum number of points to 10 and maximum distance between two points to 1.8 kilometers.

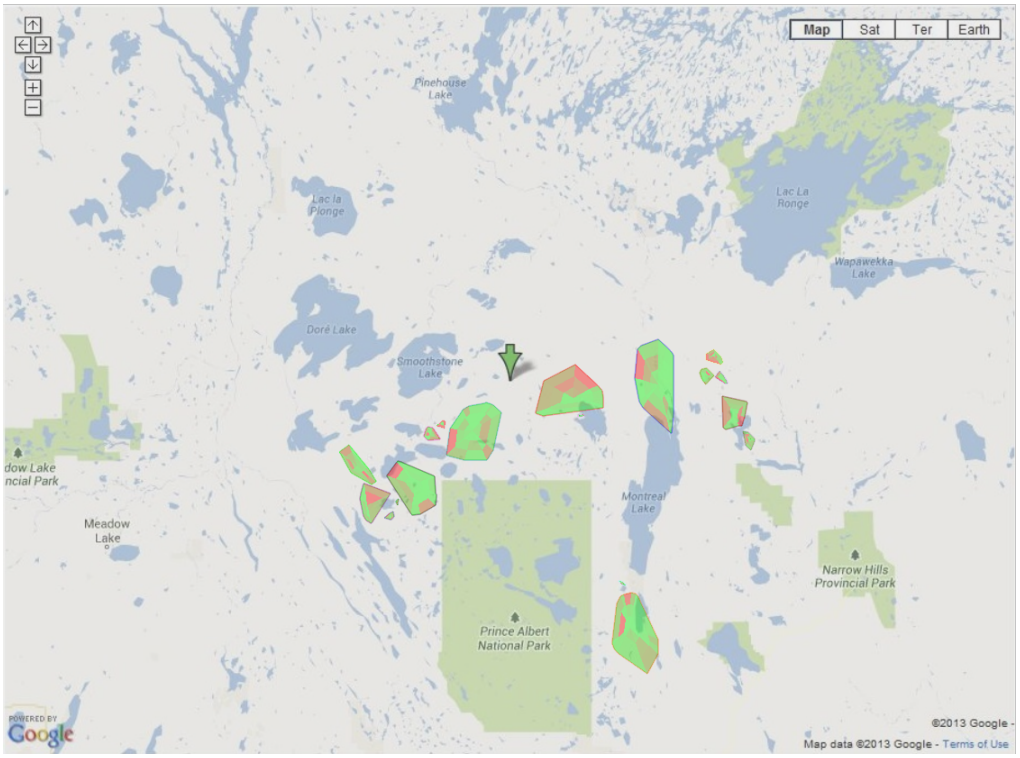



Figure 47: Group of Caribous during 2006 to 2010.

Once we do the clustering based on these parameter settings and generate clusters for two different time periods the output will be as shown in Figure 46 and Figure 47. From this initial output we see that there are many differences in the behaviour of Caribou grouping between these two time periods. During 1992 to 1996 (Figure 46) Caribous made group in various regions especially in the area near the lake La Ronge (the big lake at the top-right side). But during 2006 to 2010 (Figure 47) their grouping behaviour has changed and they are more concentrated to form group near the large park and human locality during this time period

Now if we want to see these clusters' movement based on their size we can use *flock wings* and *flock heart* visual cues of FlockViz. In FlockViz we can enable these two features and the results for the two

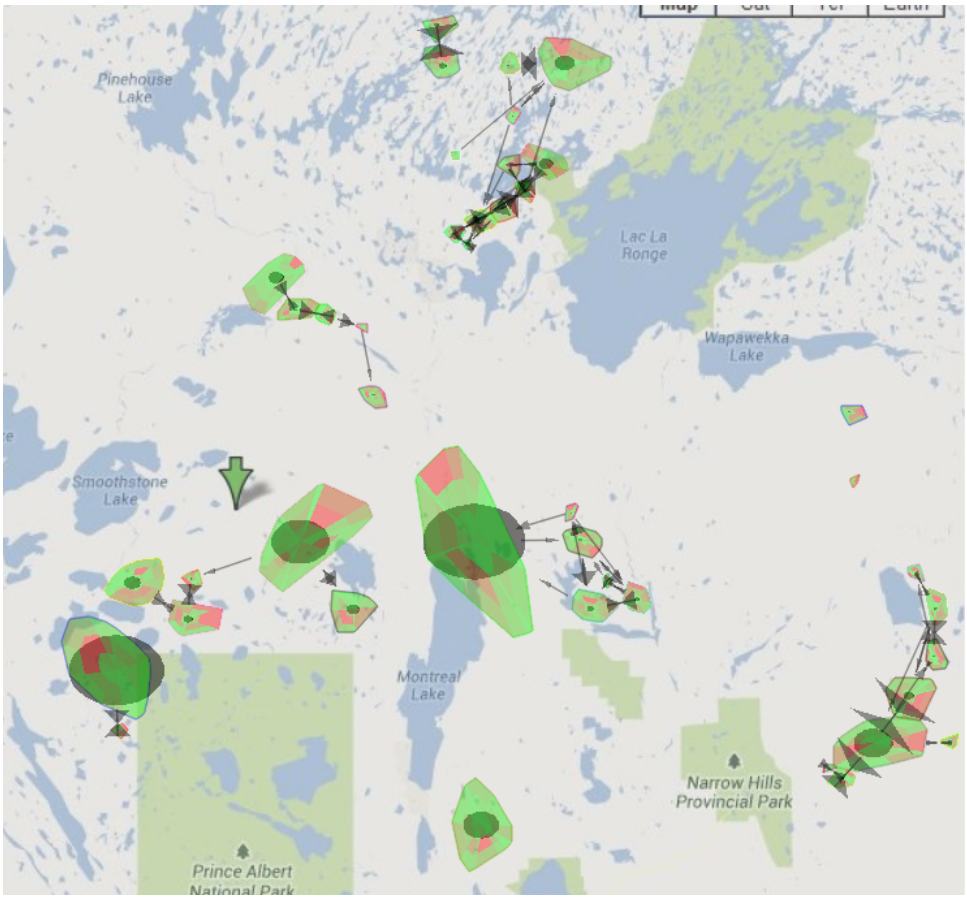



Figure 48: Group size and movement activity of Caribou during 1992 to 1996.

time periods are shown in Figure 48 (1992 to 1996) and Figure 49 (2006 to 2010). We see that in both cases a large number of groupings are formed just outside the Prince Albert National park. However, during the first time period there were more frequent movement between the different groups as shown by many bi-directional arrows. But in the later time period the movement is much reduced. One reason could be the new generation of Caribou do not prefer to change group much or they do less activity compared to the previous generation.

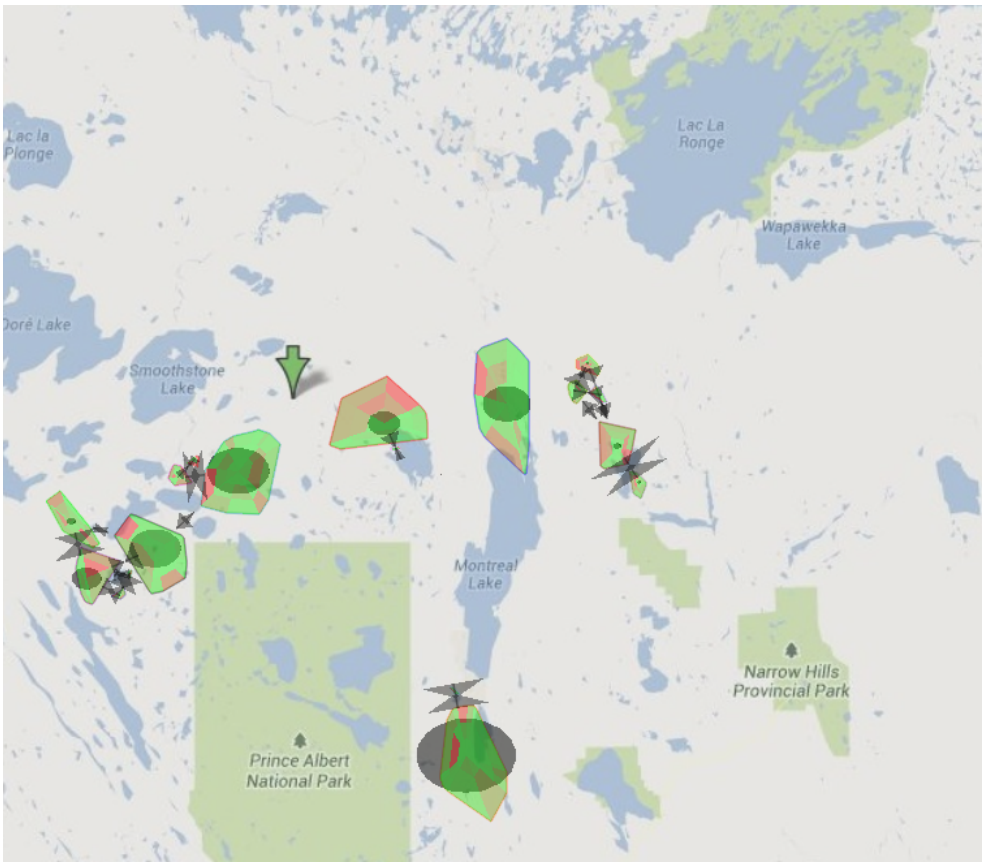



Figure 49: Group size and movement activity of Caribou during 2006 to 2010.

The second question is "Did the extent (annual home range) of their movement change between those 2 time periods?" This question is very similar to the first one and we answered part of this question using Figure 46 and Figure 47. One additional analysis could be changing the basic parameters of cluster formation to see if the pattern gets changed or not. This is dependent on user requirements for defining home range and corresponding values of the basic parameters to generate those home ranges.

The next question is "Did some unique characteristics of their movement change e.g. is the movement more clustered now, are the movements more clustered only at certain times of the year?" To

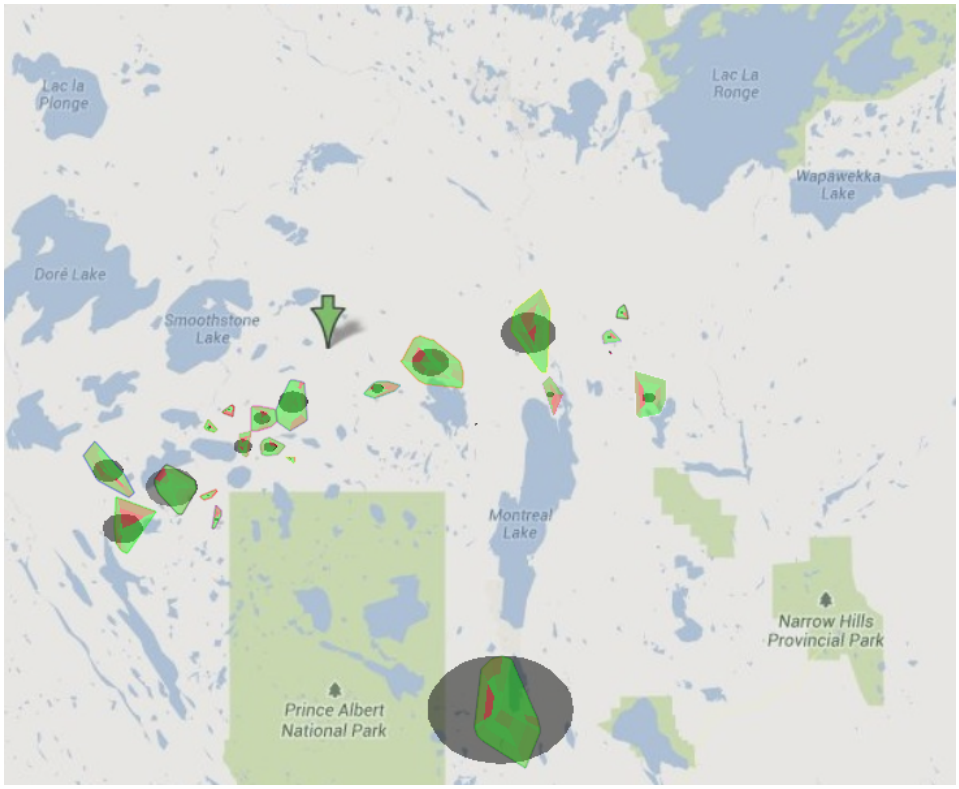



Figure 50: Clusters of Caribou from January to April.

answer this question we need to focus on the size and area of the cluster and compare these measures between different time segments. From Figure 48 and Figure 49 we notice that during 1992 to 1996 there were huge variations in cluster area and size compared to the time period between 2006 to 2010. During 2006 to 2010 most of the clusters had average size and area. Therefore, in recent years the Caribou are more structured in cluster formation than before. To answer the last part of the question we further divide the temporal segment into a different group of months. For example, we divide the temporal segment of 2006 to 2010 in January to April, May to August and September to December. The results of Caribou groupings for these three time periods are given in Figure 50, Figure 51 and Figure 52.

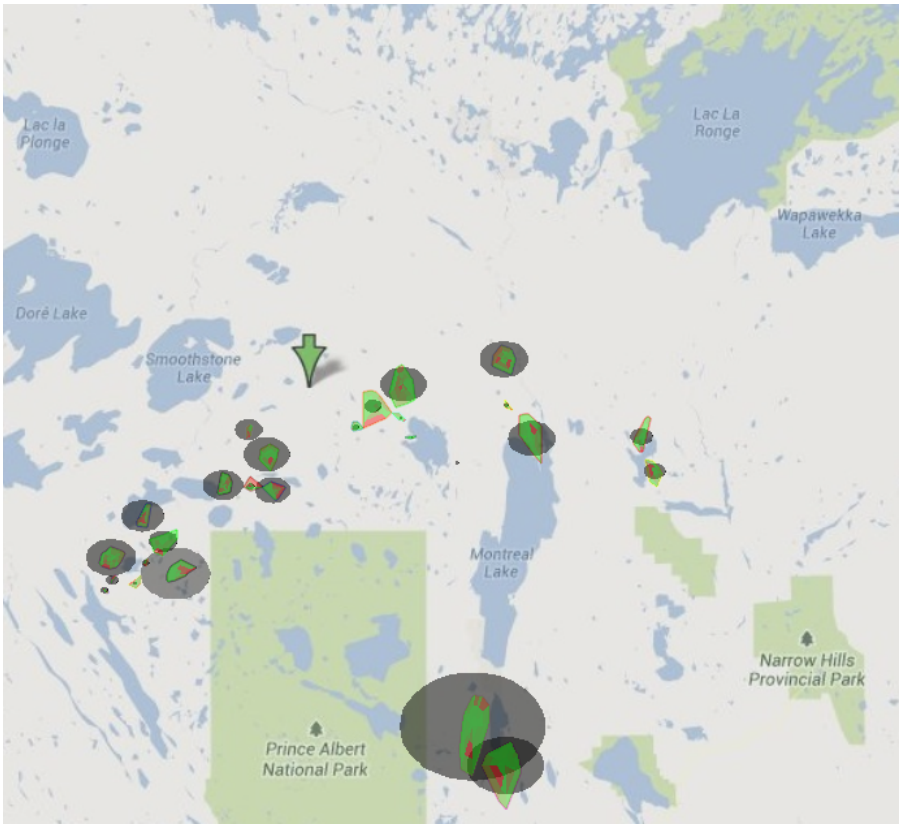



Figure 51: Clusters of Caribou from May to August.

The circle size of *flock heart* shows the number of Caribou present in those clusters. We see that at the end of the year (September to December) Caribous are likely to form smaller groups than the other period of the year. But at the beginning of the year almost an equal structure of gatherings happen. However, at the middle of the year the animals are more clustered in terms of their population compared to the other time periods.

The fourth question is "How much fidelity do animals present to calving sites (15 May - 15 June), late winter foraging areas (15 March - 30 April), or rutting areas (15 September - 30 October) between the years?" In order to answer this question we need to know the areas within the map where those activities happen. Then we can

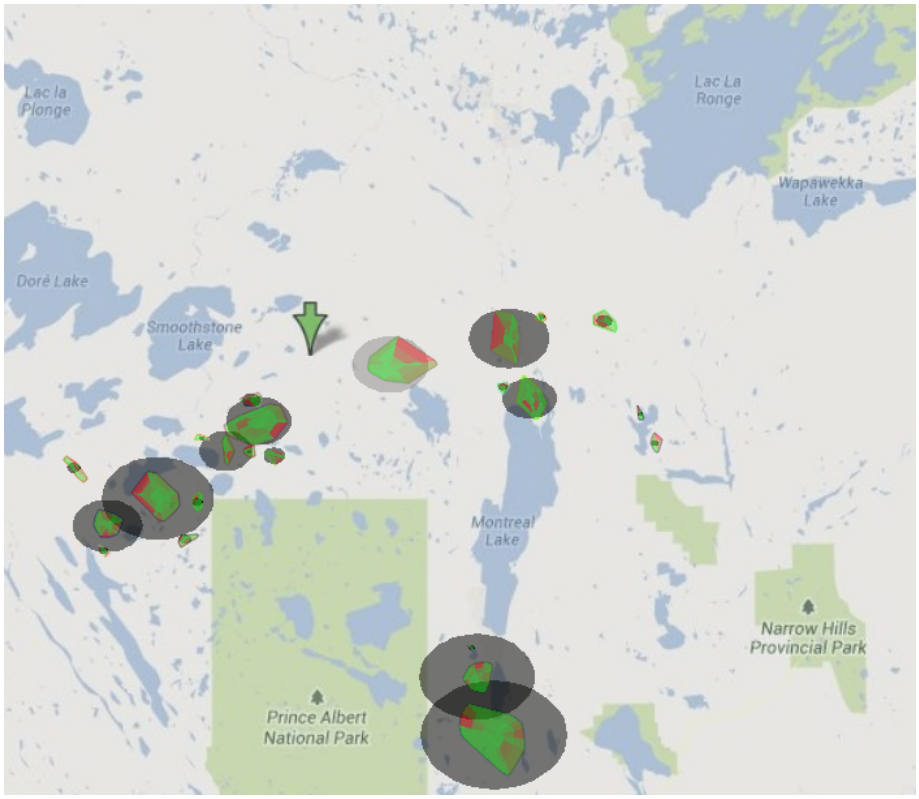



Figure 52: Clusters of Caribou from September to December.

build clusters and see if the animals gather in those areas during the specified time period. In this case study I will show the cluster of Caribou in the late winter foraging areas. These areas are the adjacent regions of each lake in this map. We build clusters using the same basic parameters like the previous questions but set the time segment between 15th March to 30 April. The resulting clusters using FlockViz are shown in Figure 53. We see that many clusters are formed near different water sources compared to the other clusters' location we saw before. Now to investigate Caribou movement between these groupings we activate the *flock wings* as shown in the Figure 54. Here we see that within small lake regions very frequent movement happens. To further analyze the fact we show the *flock shield* segment

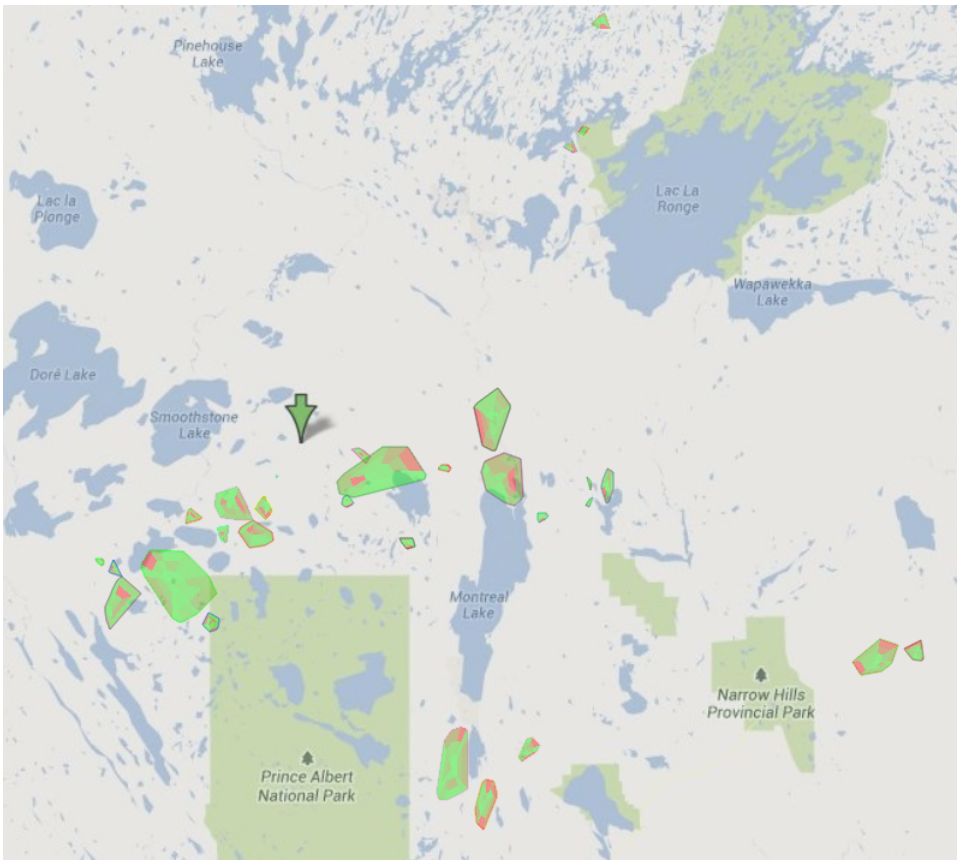



Figure 53: Clustes of caribou during 15 March to 30 April of different years.

( as pie) of age dimension (Figure 55). Here purple color is used to represent the number of old Caribou. Similarly green and blue color is used to represent the number of young and child Caribou. It is also interesting to see that child Caribous mostly gather in small water regions but the young and old Caribous are equally distributed in several lakes. The reason of this distribution is that child Caribous present much fidelity in late foraging areas compared to others and in the small water source areas most of the movement happens.

The fifth question is "Do animals tend to avoid roads and cut block in the same way between different time periods and does that differ between different times of the year?' This question can be answered


Figure 54: Clusters' movement direction during late winter time.

by tracking the movement of a particular Caribou in particular road and cut block areas. However FlockViz still can answer this question by showing a group of animal movement in those areas. Due to the lack of information about roads and cut blocks I could not identify and incorporate those areas in my software tool. In future I will gather these information and incorporate with my tool to answer this type of question.

The last question is "Do the movement characteristics of animals differ between times of year or for animals of smaller and larger home ranges? For animals are located in more fragmented areas?" This question can have a very broad range of answers. In the previous

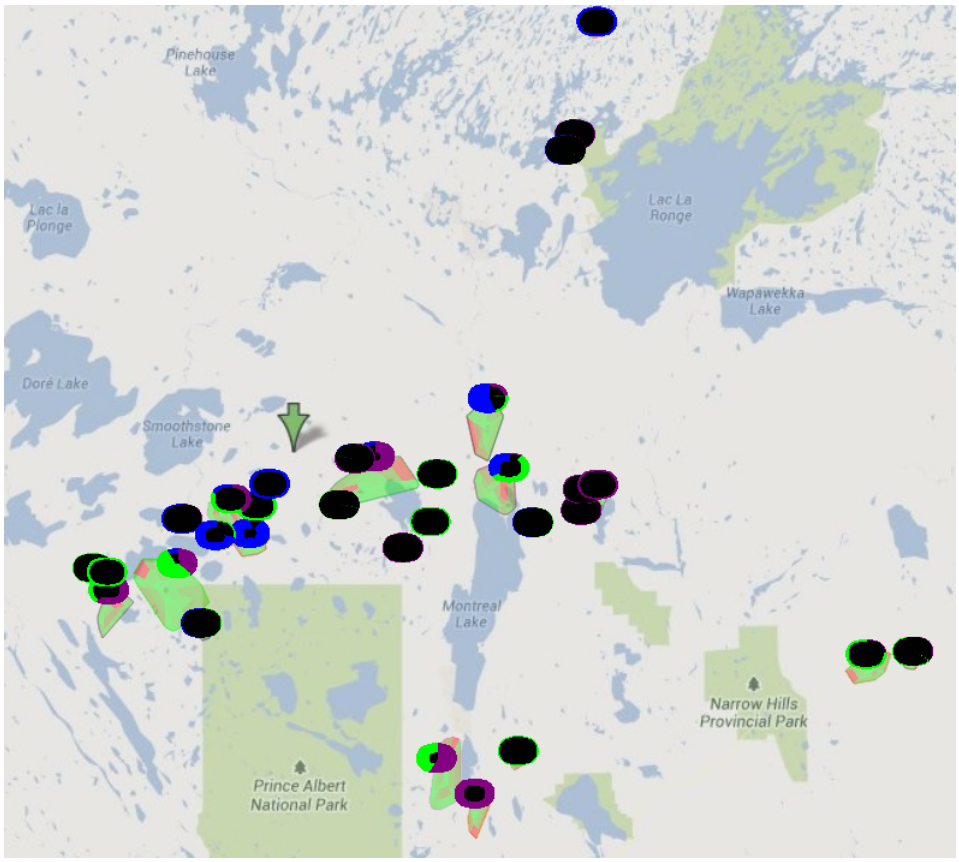



Figure 55: Age group wise statistics of clusters during late winter time.

questions we have explored many information based on time. Now we want to concentrate on two fragmented areas of the map and we will compare the characteristics of clusters between those areas. At first we build the clusters using basic parameters (minimum no of points as 15 and distance between points as 2.5 kilometers) and we do not apply any temporal constraint. The result is shown in the Figure 56. Now to compare between two fragmented clusters lets consider the cluster *a* and *b* as marked in the Figure 56. To answer this question we can enable all the parts of my FlockViz design and compare different statistics between these two clusters. The detail view of FlockViz for these two clusters is shown in Figure 57.

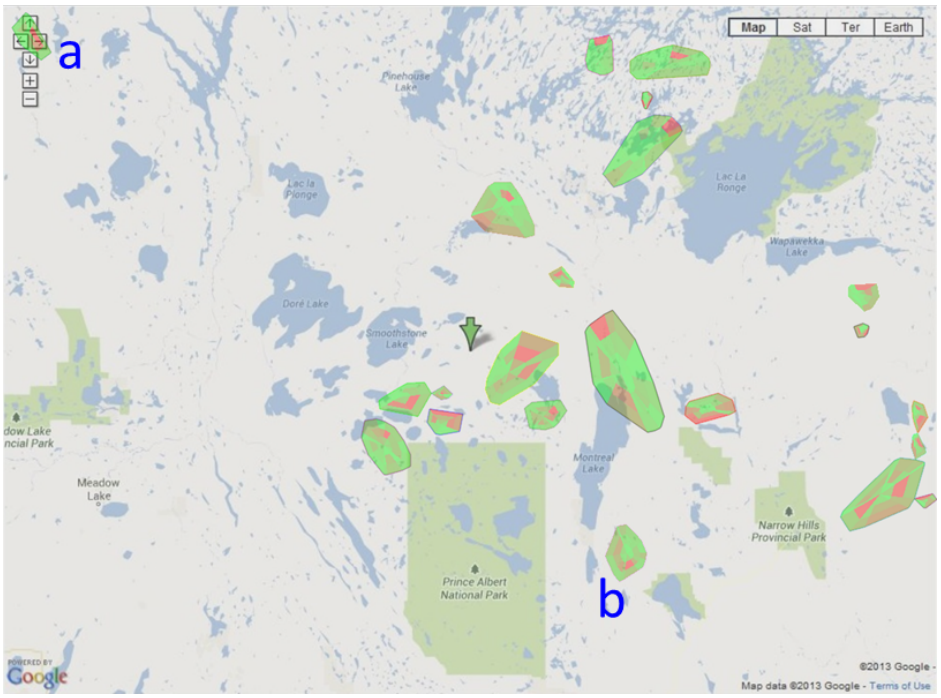



Figure 56: Visualization of FlockViz after generating clusters.

According to my FlockViz design we can say that the cluster *a* has both male and female Caribou but the cluster *b* has only female caribou. However, the number of female caribou in the cluster *b* is larger as the pie is more filled towards the center compared to the cluster *a*. Similarly there are only young Caribuous in the cluster *b* and most of them are heavy in weight. But the cluster *a* is old dominant and their weights are not within a particular range, it varies. Other dimensional values can also be compared in this manner. Another difference between these two clusters is the overall size of the cluster *b* is bigger than the cluster *a* according to the *flock heart* drawn in the middle. The clock in that heart shows that both of them took the same amount of time to form those clusters. Finally the density map of the cluster *b* is more uniform (red regions are well distributed) compared to the cluster *a*. Therefore, in the cluster

*b* Caribous are found almost equally all over the places. But in the cluster *a* most of the Caribous stay in the middle-right side of the cluster.

The department of NRI at the University of Manitoba is doing research on the behaviour of Caribou movement. I showed the capabilities of my FlockViz design to answer the above queries to Dr. Micheline Manseau of NRI department. She was very pleased with this visualization technique and found my proposed methods helpful to explore their research questions.

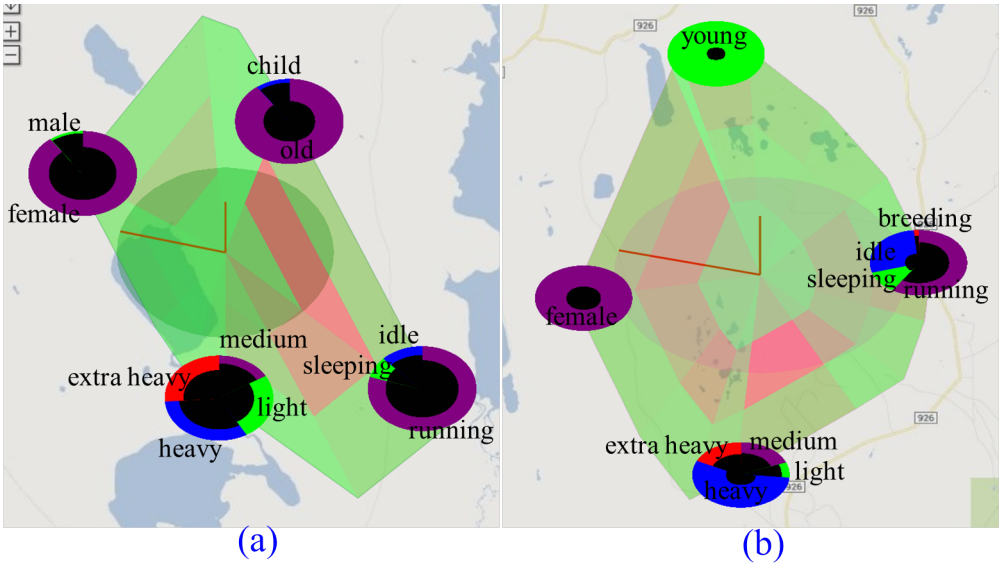



Figure 57: Comparison of different properties of clusters by showing all the features of FlockViz in a single view.

## 5.3 CASE STUDY 3: FINDING DISTRIBUTION OF STORMS IN CRITICAL ZONES

In the next two case studies I will explore one or two research questions for each dataset. Here some of the old design of *flock shield*

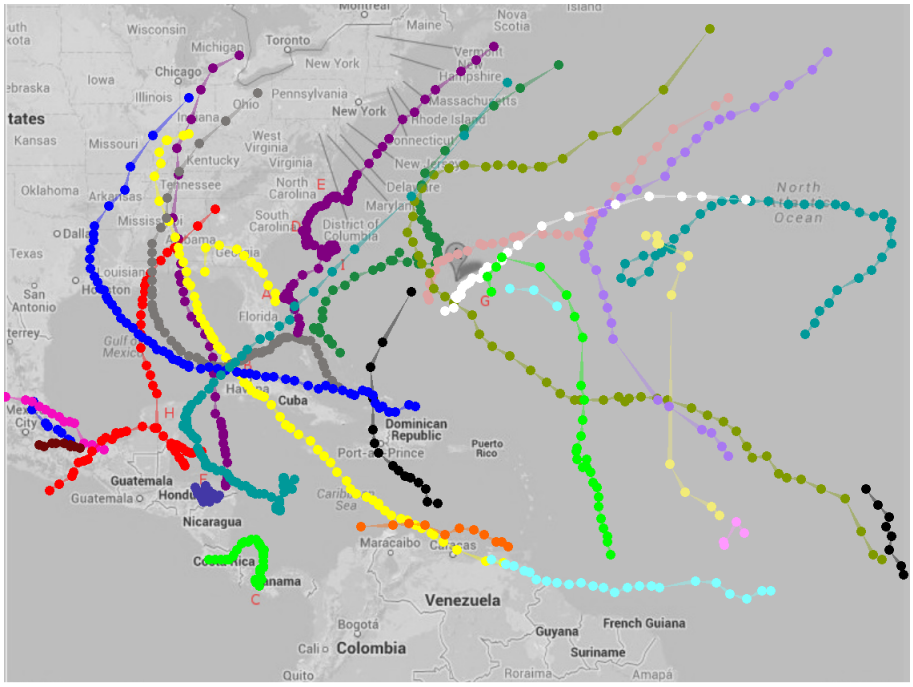



Figure 58: Movement of storms in pacific ocean during 2005.

and *flock head* will be used to answer different questions. In the first case study we will see how to generate a visual map of critical zones which were previously affected by storms. I will investigate the dataset of storms in the North Pacific Ocean region which were collected in 2005. This dataset was described in section 4.1.1. The movement trajectories of this dataset are shown in Figure 58. One of the interesting tasks associated with this dataset is to find a visual map of the critical lands near the ocean which were attacked by several hurricanes. It is also important to see how different hurricane types were distributed in these regions. In order to do this, I visualize clusters of hurricanes in these areas and do comparative analysis of aggregate measures for the hurricane category, hurricane location, most dominant wind speed, and hurricanes' movement pattern.

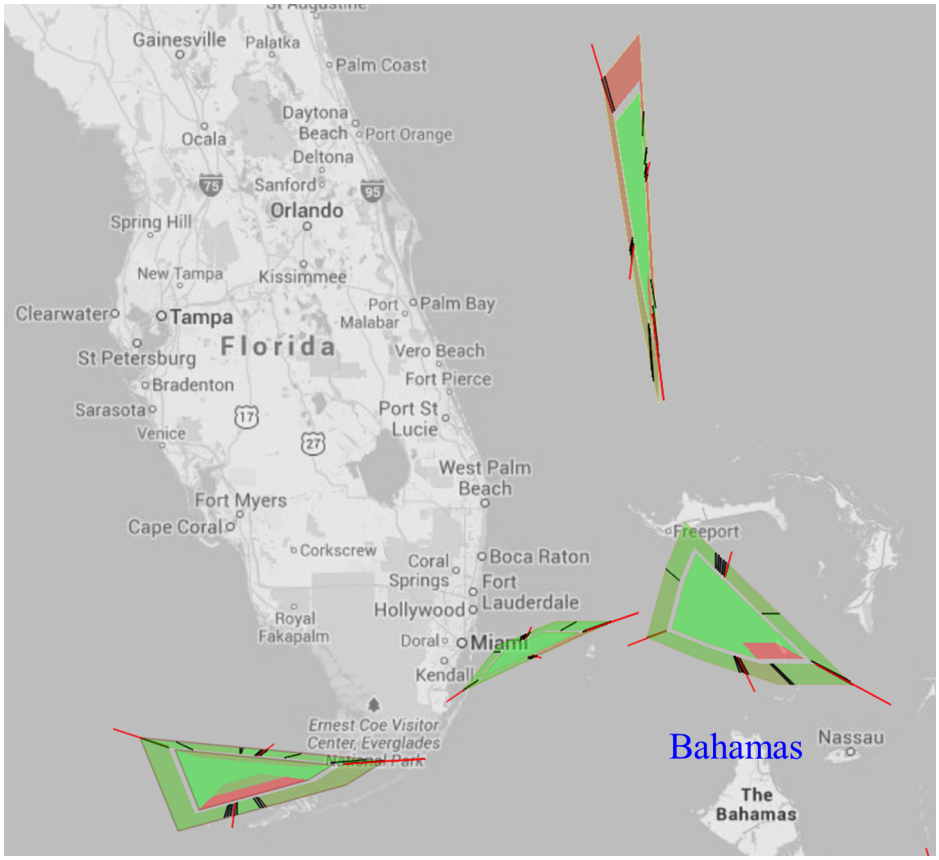



Figure 59: Clusters of storms in critical zones.

To define the clusters, from my FlockViz tool, I choose a minimum of 3 trajectory points and the distance between points to be 50 kilometers. I do not choose any temporal constraint since I want to see all storms which happened in 2005. Next, I choose "near land" location from location dimension values as it represents storms in critical zones and from activity dimension I choose "ocean to land" movement of storms. This will result in a display of the visual map with the clusters in the critical zones (Figure 59 shows a subset of all zones). Now if we want to see the distribution of hurricane types for a particular critical zone we zoom into that cluster. For example, the Figure 60 shows clusters of Hurricane in Bahamas critical zone.

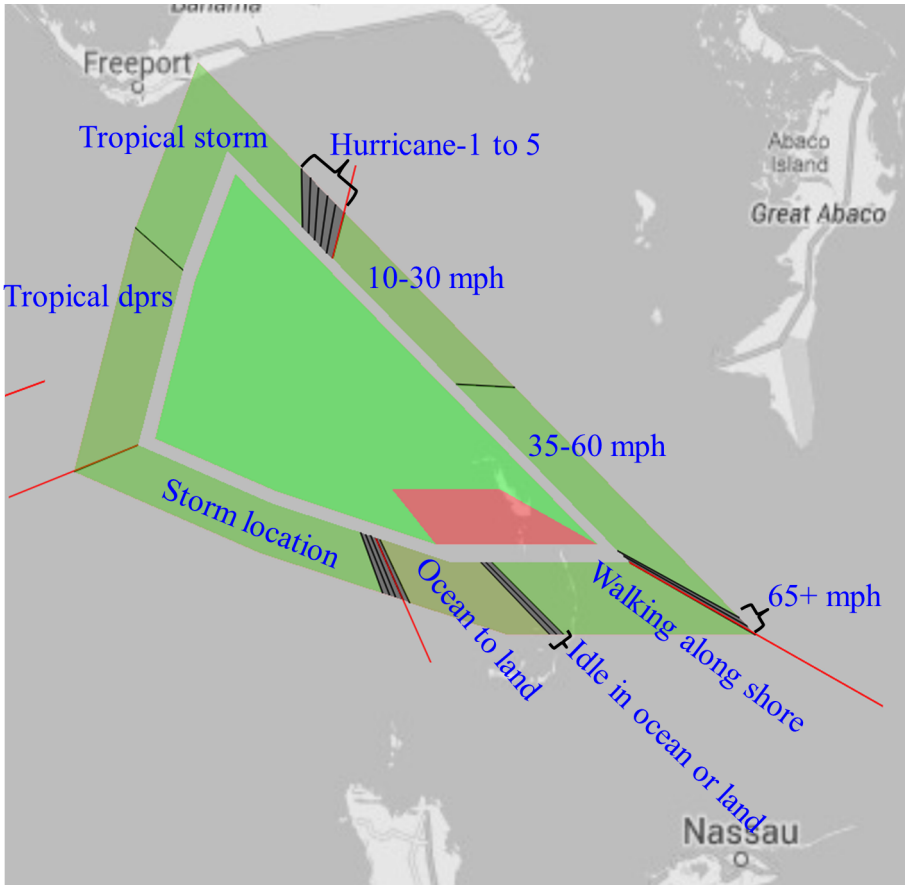



Figure 60: Analyzing the distribution of storm's properties in a particular cluster near Bahamas critical zone.

According to my first design of *flock shield*, the first top left segment of the shield represents an object category or storm type. Other segments represent wind speed, storm location, and storm activity. Referring to Figure 60, by maintaining the order in which I read dimension values (i.e. left to right or top to bottom) as shown in the Figure 22, we can see that there were no storms of type Hurricane-1 to 5 (indicated by last five small black divisions) but had tropical depression and tropical storm (indicated by the first two divisions). Second segment shows that only wind speed ranges of 10 to 30 mph

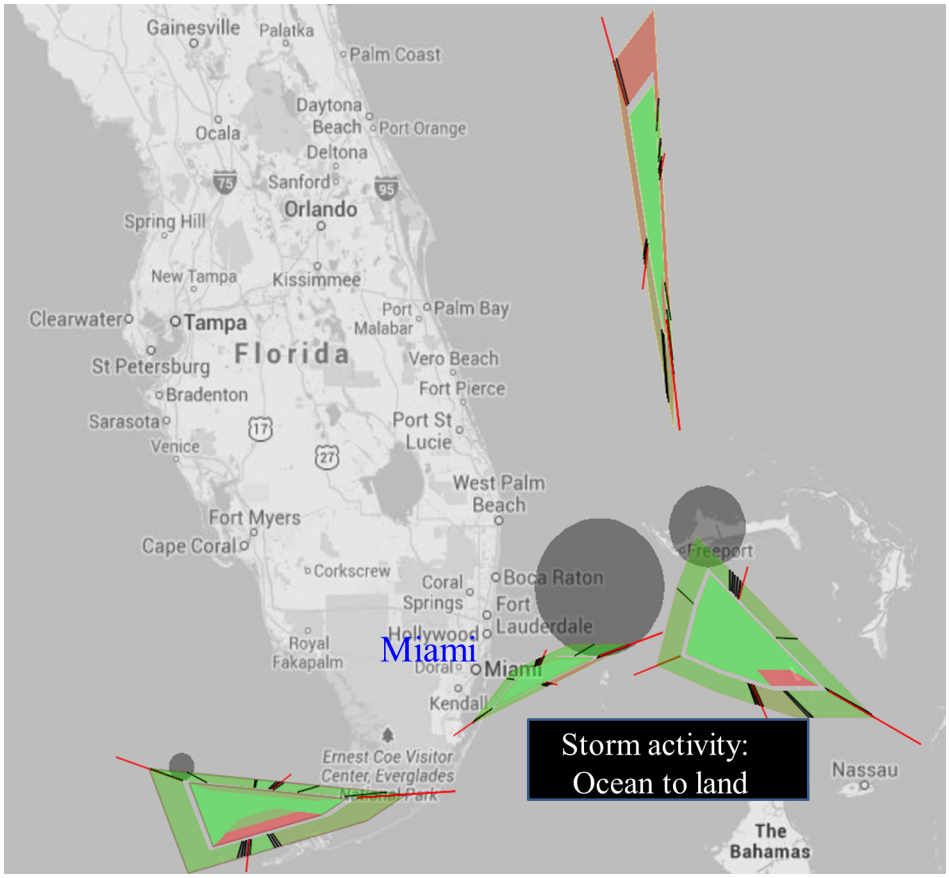



Figure 61: Comparison of storm activity (ocean to land movement) among critical zones using *flock head*.

and 35 to 60 mph happened there. Finally, the last segment shows only "ocean to land" and "walking along shore" activities happened.

In order to compare these statistics across different critical regions, I use interactive features of FlockViz which help to highlight particular aggregate measure at the top of each cluster. For example to find which critical region mostly had "ocean to land" directional movement, I can hover over the corresponding *flock shield* part. The resulting visual cue representing the aggregate measures is shown using the *flock head* (Figure 61). The black circle sizes show the number of storms containing the activity in each critical region. From

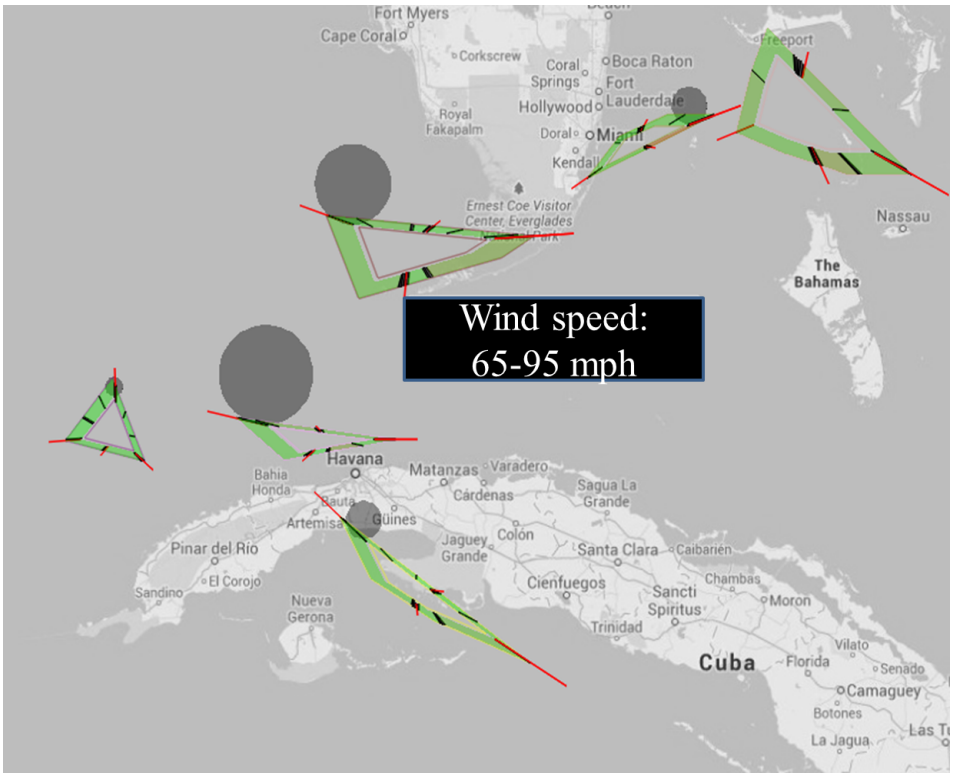



Figure 62: Comparison of storm speed among critical zones.

this view, we can say that all the storms near the Miami critical region (i.e. the cluster with the largest circle in the *flock head*) had the most number of "ocean to land" movement in comparison to others. Similarly, Figure 62 shows a broad angle view of wind speed of 65 to 95 mph among all critical zones.

## 5.4 CASE STUDY 4: FINDING EAGLES' WINTER ACTIVITIES

The golden eagle migration observation and tracking (Figure 63) is of importance to researchers in the field [43, 49]. This dataset was described in section 4.1.3. Some of the interesting analysis tasks include finding how eagles spend their winter, what areas they cover

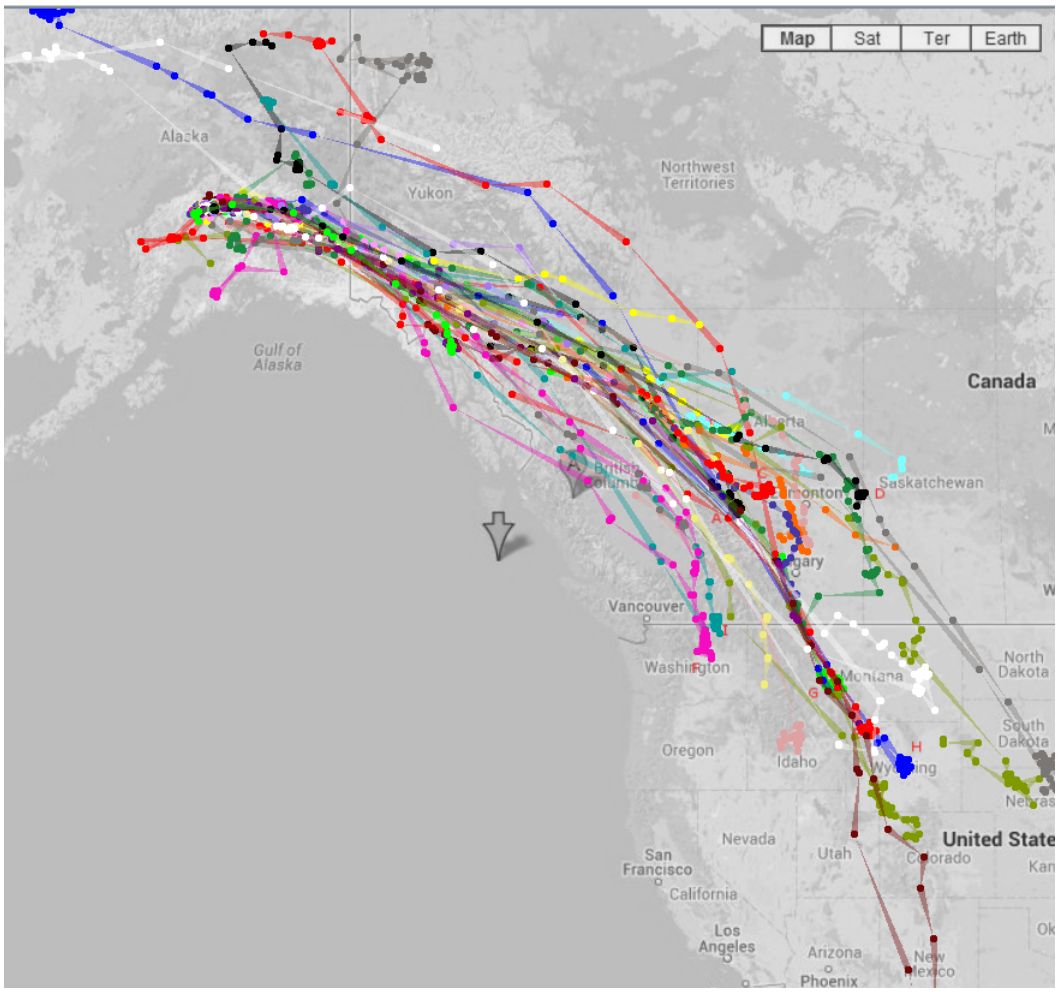



Figure 63: Golden eagle migration data.

mostly during winter, their duration of stay, and gender/age based statistics on their migration.

To do this case study we assume winter migration duration fall between August and February in the regions shown in Figure 63. The FlockViz technique gives an option to build clusters in different discrete time segments. In order to find winter time clusters I choose four time segments for each year (1997 to 2000) between August and February. For this dataset, I select a minimum of 5 trajectory points and 50 kilometers distance as clustering parameters to generate the

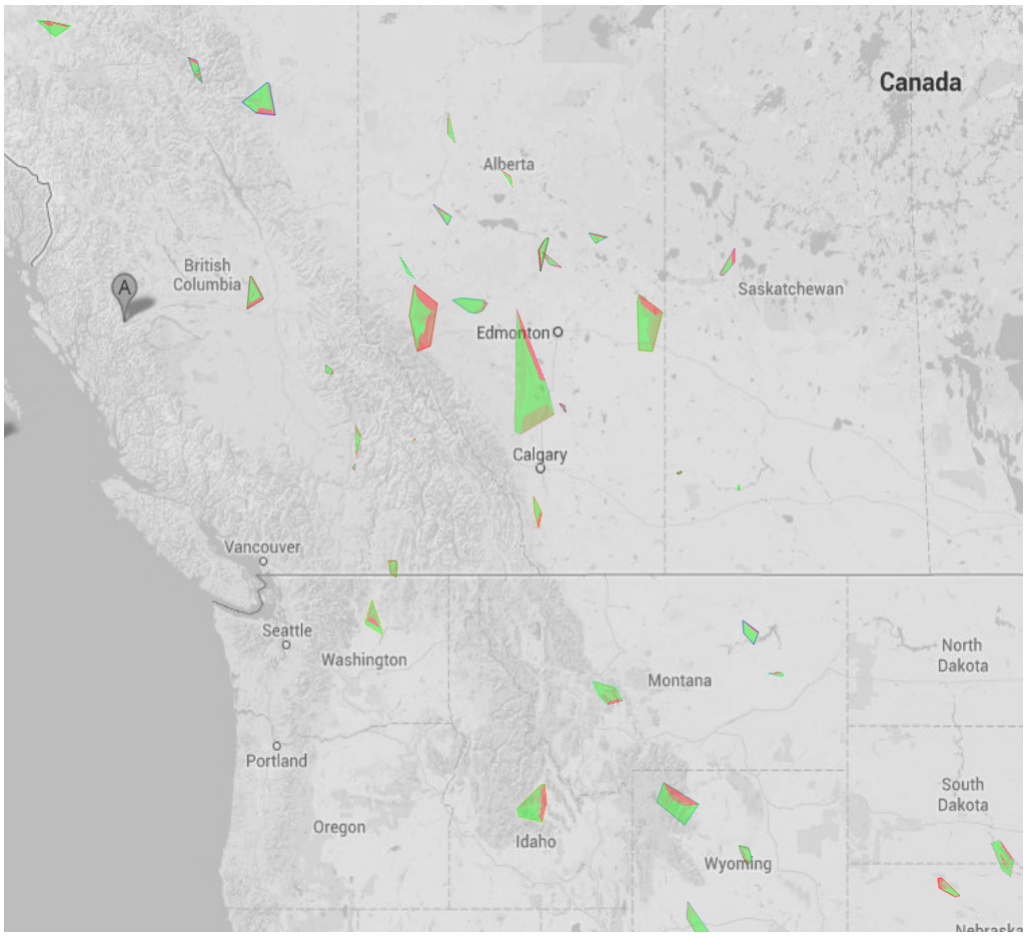



Figure 64: Clusters of eagles during winter migration.

clusters (Figure 64). We can see that during the winter migration period, almost all the eagles start moving from Alaska migrating to mostly regions in Alberta and some part of the US. Very few eagles reach Mexico. Another interesting finding is that the eagles mostly gather on the lakes than in hills or forest regions according to the terrain map of Figure 64. I assume the reason for this winter activity is eagles tend to go close to human locales to get to warmer places.

Finally, if we want to investigate how clusters of eagles are positioned in each month during the winter migration and compare it between two different years we can use the color pattern of *flock heads*

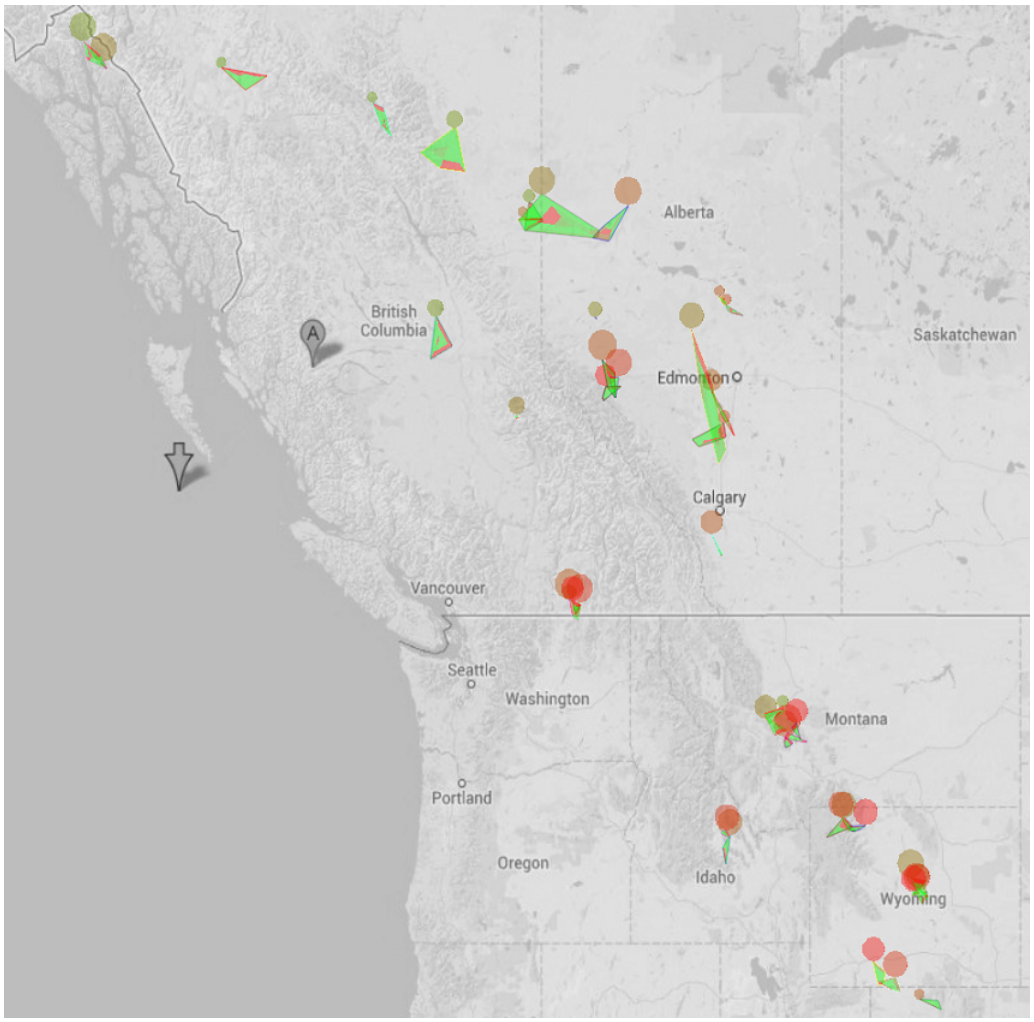



Figure 65: Clusters of eagles at different months during 1997.

(which is an alternate design for showing temporal information) and easily compare a broad range of views. I choose seven different time segments from August to February in 1997. Then setting the clustering parameters like before, I get clusters as shown in Figure 65 with temporal information embedded in the *flock heads*. Similarly, Figure 66 shows the clusters for the year 1999. As per my *flock head* design for temporal information, the circle color shows how recent the cluster is and circle size shows how long it takes to form the cluster. The green color of the circles will represent the start of the

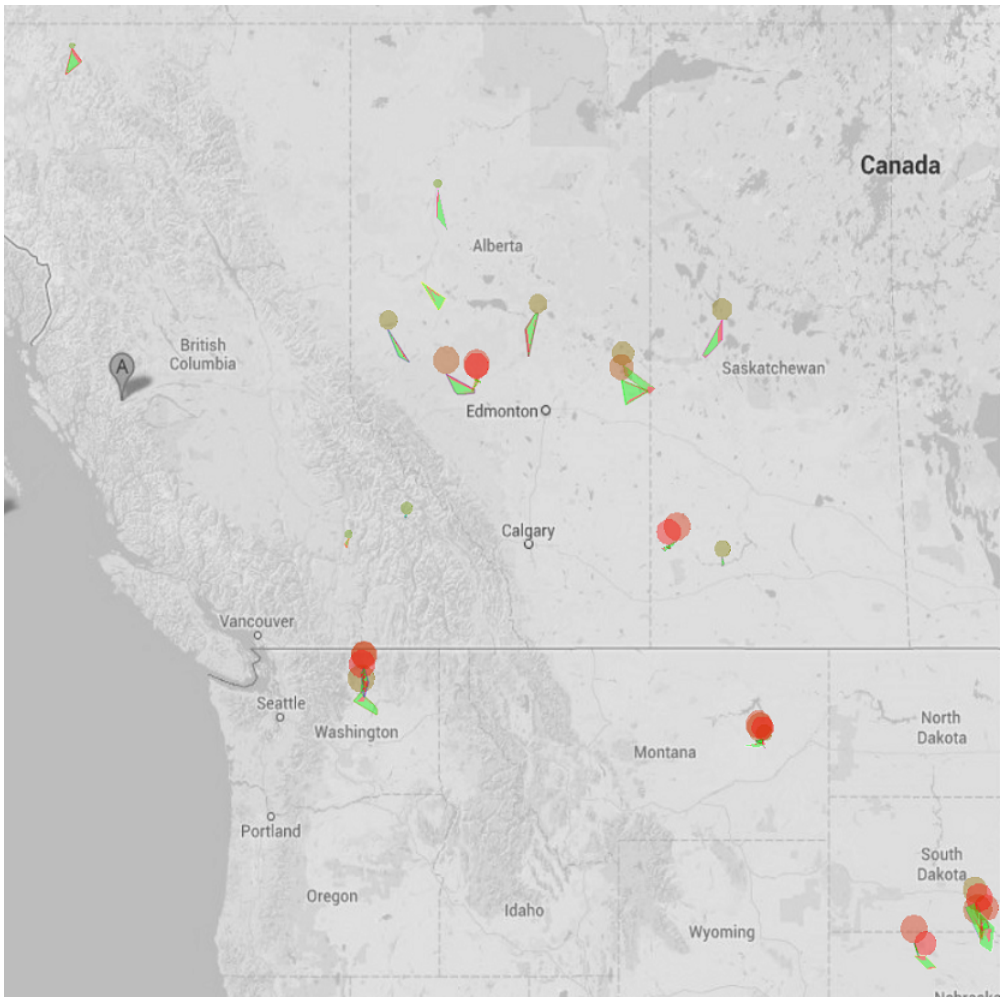



Figure 66: Clusters of eagles at different months during 1999

selected time segment and the red color will represent the end of the time segment. The gradient colors between these two will fall in the middle of the whole time period. By observing Figure 65 and Figure 66, we can say that eagles start their journey from Alaska in August, then during October to November they stay in Alberta (Canada) and reach the US between January to February. For both years the cluster regions are almost the same which indicates that eagles migrate using well-learned path every year.

I showed several case studies using various kinds of data set in this chapter. We see that FlockViz has a wide range of capabilities to analyze various research questions. However, the multiple information representation of FlockViz in a single view makes it unique from the traditional visualization approach. FlockViz not only increases the capabilities of analysis at multiple levels at the same time but also give a good quality of output to present the meaningful results. The comments of the expert users from the NRI and Biological Sciences department of the University of Manitoba confirm its value for analytic evaluation.

# 6

# APPLICATION SCENARIOS

Data visualization is one of the most important steps of analysis for making effective decisions in planning, monitoring, designing models, etc. Analyzing different characteristics of a group of objects by showing multiple information in a single view is very helpful for such kinds of applications. FlockViz can provide this single view representation making the analysis easier. In the following sections, I will show how FlockViz can be used to assist the analysis in some real life applications.

## 6.1 DESIGNING VEHICLE CONTROL SYSTEM AT A ROAD INTERSECTION

One of the major challenges to design an infrastructure of a town is building proper traffic system so that the traffic jam is minimized. It is hard to change the design once a traffic control system is implemented. However, if we need to change the traffic control system, we should make sure that the new system generates very few traffic congestions compared to the previous one. While designing a traffic control system we need some statistics of current traffic environment. These information may include a) the status of nearby roads at different times of a day, b) how many vehicles can travel through this

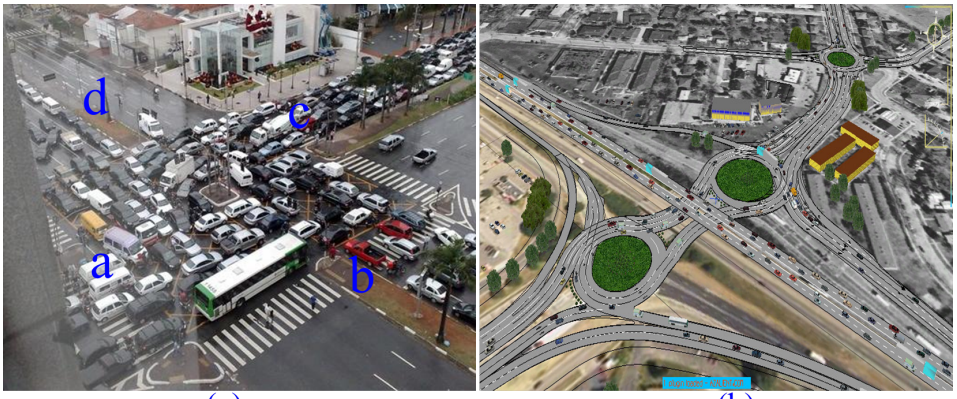



Figure 67: (a) Typical traffic jam condition due to poor traffic planning. (b) Simulated design of a road intersection to resolve traffic jam. Image source [13, 55]

road or intersection, c) what types of vehicle are most likely to use this road and d) what are the possible destinations of those vehicles. Based on these information they can design the wings of a road, waiting time in a particular signal, restrictions on vehicle movement to particular direction, etc. Since vehicle movement data are generated every second through GPS tracking systems, the volume of this data is huge. Analyzing such information using clustering is one of the important approaches that is helpful to city planners. Figure 67(a) shows a traffic jam and we assume that it is due to poor design and planning of vehicle controls in that road intersection. Figure 67(b) shows a simulated design of another road intersection to solve such kind of problems. Next, I will discuss how FlockViz can assist to design for proper traffic movement at a road intersection.

To design a road intersection, FlockViz can provide information such as hourly or temporal segment wise statistics of vehicles in that road intersection, the size of a queue in each road connecting to that intersection, various properties of those roads' usage (e.g.

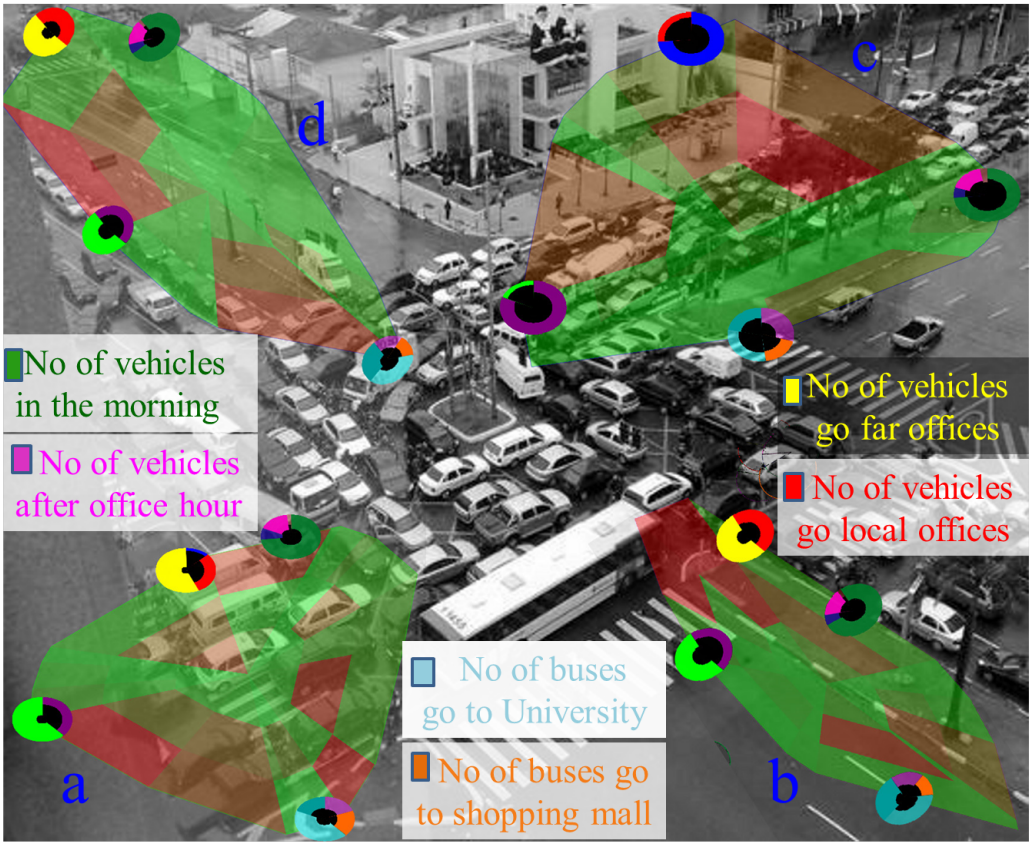



Figure 68: FlockViz embedded to a road intersection to show different statistics of vehicle movement.

number of vehicles going to offices, schools, or shopping centers etc.). Let's assume we have a four way road intersection (for example Figure 67(a)). After generating clusters of vehicles for each connected road (*a*, *b*, *c*, and *d* as marked in Figure 67(a)) we can represent those clusters as FlockViz. Let's place those flocks at the end of each road just before the crossing as shown in Figure 68. The *flock shields* represented by colored pie segments give a comparative aggregated measures of various dimensional values among those connected roads. Some important vehicle statistics are highlighted and tagged in this figure such as the number of vehicles movement through

this intersection before and after work hours. This one snapshot of FlockViz can suggest several traffic designs. For example, according to the *flock shield* design of my proposed FlockViz, all the roads have almost an equal number of vehicles passing after office hours (pink pie segment). Therefore, all the signals of this road intersection must have similar priority at that time. However, during the morning roads *a* and *b* must get highest priority because the number of vehicles at that time (dark green pie segment) is larger (pie segment is filled more towards its center) than the other roads. Now if we look at the yellow segment of pie (number of vehicles which are not using offices near this intersection), we see that the vehicles in road *c* have destinations very close to this intersection (no yellow pie segment). But the destination for the vehicles in other roads is very far from this intersection (significant amount of yellow pie segment). Therefore, this visualization can suggest to build three long overpasses for the roads *a*, *b* and *d* so that a large number of vehicles can avoid this signal and reach their destination quickly. Thus we can reduce the traffic jam significantly by implementing this design. Similarly other information can be expressed using this FlockViz design and traffic signal priority can be implemented effectively. It is important for buses to reach the University on time during peak hours. For example, we can control the waiting time for vehicles at the signal of each road during peak hours. If the Figure 68 shows clusters of vehicles during peak hours we see that the number of buses going to the University (light blue pie segment) is different in each cluster. Here road *a* has the largest number of buses going to the University during this time period and then road *b*, *d*, and *c* respectively. Thus

we can control the waiting time in the signals during peak hours according to this visual output of FlockViz.

## 6.2 DESIGNING AN ALERT SYSTEM IN A SURVEILLANCE AREA

Figure 69: Image captured by a surveillance camera in a shopping mall. Image source [38]

Video surveillance is an important feature to track any unusual events and monitor the status of a particular region of interest. As FlockViz has very high dimensional value representation capability by preserving the original cluster area intact, we can use this method to design an alert system based on different statistics of peoples' movement. Let's assume in a shopping mall [Figure 69] the surveillance system is installed and through the image of the

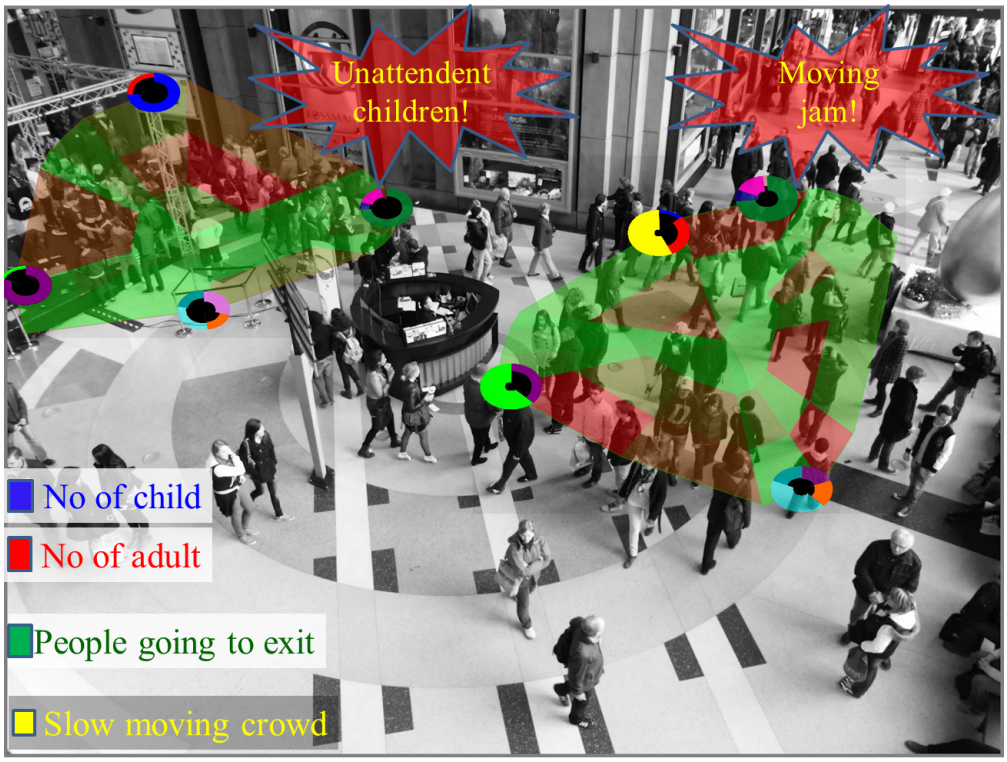



Figure 70: FlockViz assisting surveillance system to identify unusual gathering of peoples in a particular region of interest

cameras or using tracking systems we can detect the movement of people. With a machine learning algorithm we can identify the biometrics of passerbys and detect how many of them were present in a particular area. We can also detect whether they are entering or leaving a particular region of interest or not. If we have all these data available for different regions of interest we can use FlockViz to show the statistics of peoples' movement in those regions. It is also possible to dynamically change the FlockViz based on peoples' movement. Figure 69 shows an image captured through a surveillance camera of a shopping mall and Figure 70 shows how FlockViz can visualize different statistics of peoples' movement for two particular regions of interest. In Figure 70 the density map of the *flock body* shows

how people are distributed within that region of interest and *flock shields* (pies) represent gender, age and movement direction based statistics. Some interesting statistics are highlighted and tagged in this figure such as where we will find a large number of children or in which region more people are moving slowly. Therefore, FlockViz can help surveillance data analysts to identify if any particular area within that region gets over crowded or any particular type of people are gathering there in large amounts. Now based on the statistics of the number of people with a specific biometric trait the system analysts can generate alert messages. For example, Figure 70 shows that the left region of interest has more children (blue pie segment) than adults (red pie segment) which may cause accidents due to the absence of most of the parents. Similarly the right region of interest detected a slow moving crowd (yellow pie segment) and it may cause a serious jam in that area within an hour. Surveillance data analysts can identify all these unusual events using FlockViz and send alert messages to security persons. After getting these alert messages security persons can take the necessary actions by going to that region of interest and help to avoid possible damages.

## 6.3 ANIMAL MIGRATION AND THEIR SEASONAL ACTIVITIES MONITORING

Animal migration is one of the most significant biological phenomenon. Researchers are actively working with migration data to identify animal's movement behaviour, changes in climate and environment patterns. The result of their research helps government

to take necessary steps to ensure a healthy environment for those animals. Thus the balance in nature will be preserved well if they can provide sufficient information to the government. In order to do this, researchers are deploying tracking devices inside the animals to collect data about their movement. But analyzing individual animal's movement is not sufficient to get an overall idea of their migration pattern. However if we plot a visualization for all the animals' movement it becomes difficult to get meaningful information due to overlapping and cluttered view. To resolve these issues one possible solution is to build an application which will generate clusters and analyze information for each group. FlockViz can help to build such an application as it can give multiple properties of the cluster in a single view by preserving the actual cluster area. It can also show the temporal information to form those gatherings, their internal distribution within the cluster, how many male or female animals gathered there, how many young animals travelled and many other dimensional values. I discussed different scenarios of animal movement in the case study sections and demonstrated how FlockViz was used to solve various research questions.

# 7

# DISCUSSION, CONCLUSION AND FUTURE WORK

## 7.1 DISCUSSION

In this thesis I presented a visualization technique (FlockViz) and evaluated it's performance and applicability through empirical evaluation and case studies. In each step of this thesis, from the FlockViz design to the evaluation process I identified some important design choices that should be considered for Spatio-temporal cluster visualization. In the following sections I will discuss some of these important findings and the limitations of my proposed design. I will also discuss some of the results of the experiments and clarify some design choices that I made in the thesis.

### 7.1.1 *Experimental results and conditions*

FlockViz was designed for high-level analysis and comparison among spatio-temporal clusters (STCs). FlockViz is not intended to display and compare exact numerical values of cluster properties. However, we can easily add this option to display detailed information through pop-up windows. In the experiments, I intentionally disabled this option in FlcokViz. It made FlockViz competitive and challenging to answer the questions with respect to TraditionalViz. Therefore,

I did not include questions that require exact numerical values. I also excluded some task categories related to movement direction from the experiments because TraditionalViz is not able to answer those questions. I discussed the reason in detail in the task categories section (4.2).

Most of the experimental results showed significant favour for FlockViz. But I found that for high level task categories (Comparative Aggregation among Multiple clusters (CAM)) there was no significant differences in performance between FlockViz and TraditionalViz. The reason is that among the four low level tasks under the CAM category the tasks related to aggregate measures nullified the overall significance effect. To understand the performance of this task category I investigated the questions further. One of the questions under this category was "Which cluster took the longest duration to form?". Although this question required a comparative aggregation among multiple clusters, I had a maximum four clusters in the options panel for answering this question. Users were asked to compare the temporal information among those clusters. As TraditionalViz was not difficult to explore a few number of clusters, this question was not enough to judge the performance for comparison among a large number of clusters. I assume that by introducing more options for this type of question we could see the significant benefit of using FlockViz compared to TraditionalViz. Moreover the additional time taken by users to learn the placement of aggregate measures in different flock parts was also the reason for this experimental result. But I allowed the users the same amount of time to be familiar with the visualization techniques.

### 7.1.2 *Trained users and their learnability effect on performance*

FlockViz is inherently more complex than traditional visualization and users need more training to understand different flock elements compared to the TraditionalViz. In the experiment I provided equal training time for the users for both techniques. I assume that the users who are familiar with FlockViz design can perform significantly faster in all task categories than the TraditionalViz. I observed this fact while working with domain experts in my case studies. I noticed that at first they were slow in explaining analytic results of FlockViz. But their performance increased rapidly after understanding the FlockViz design well. Therefore, we need to provide sufficient training to users on a relatively complex visualization before giving providing the analytic task. A longitudinal study could be one way to investigate users performance if they work with the data and visualization techniques for a long time. It is also interesting to see how this learnability affects the performance over time. In my case studies, I studied this effect by observing the performance of domain experts over the course of time. For example at the beginning users frequently asked to enable the pop-up window (which was used to show the measures of different properties of a cluster (Figure 25(a)) for each cluster like the TraditionalViz to explain the data. However, gradually their performance improved and they could explain the data very well just by looking at the FlockViz instead of using the pop-up window.

### 7.1.3 *Limitation of objects' dimensional value representation in FlockViz*

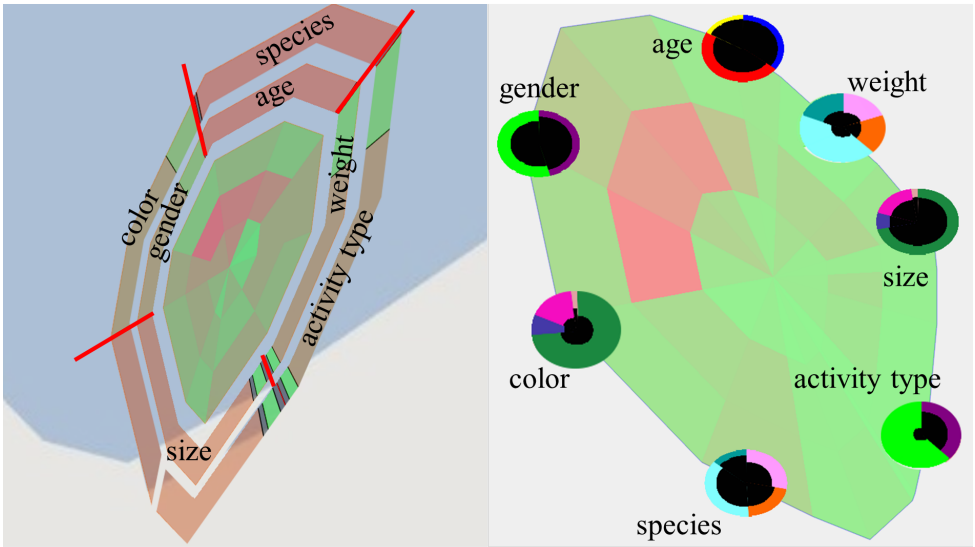



Figure 71: Enhanced *flock shields* to show a large number of dimensions' values. The left image shows the enhanced original *flock shield* design by adding multiple layers and the right image shows the alternate *flock shield* design where a large number of pie segments can be added in *flock shield* regions.

In the design section of FlockViz, we see that it can show a large number of dimensional values of a cluster in the *flock shield* regions. Figure 71 shows how FlockViz can visualize the statistics of 7 dimensions. Therefore, FlockViz can represent more than four dimensions by adding more pie segments within the shield or by including layers surrounding the main shield in the rectangular area design. The number of dimensions that FlockViz can represent (without any overlapping and cluttered view) might be limited. Investigating such limitations is outside of the scope of this thesis and is left for future works. However, such limitations depend on the number of values

for each dimension and how the overall cluster is distributed in space. If there are many clusters adjacent to a particular cluster then we will be able to show a very few number of dimensions in *flock shield*. All these factors of FlockViz design are very important when we work with a large number of dimensions in a single view.

### 7.1.4 *Overlapping scenarios between clusters in FlockViz design*

One of the main goals of cluster generation and visualization is to reduce overlapped and cluttered view. The cluster generation process (ST-DBSCAN algorithm) of FlockViz helps to overcome this problem significantly compared to other clustering methods as I discussed in related work section (2.2). However, there are few cases where FlockViz can have overlapping scenarios. According to my *flock shape* design the boundary of a cluster is defined by convex polygon. If there exists a large empty region (i.e, no objects present or the region is color coded as green according to my design) within that polygon then there might be a chance that some objects from other cluster occupy this empty space. Thus an overlapping may occurs between clusters. Figure 72 shows an example of such overlapping scenarios in FlockViz design. Here cluster *b* has an empty region marked by *x* which is occupied by some objects of cluster *a*. This kind of overlapping scenarios happen rarely and it does not affect the visualization of FlockViz. However, I identified this as one of the limitations of my FlockViz design.

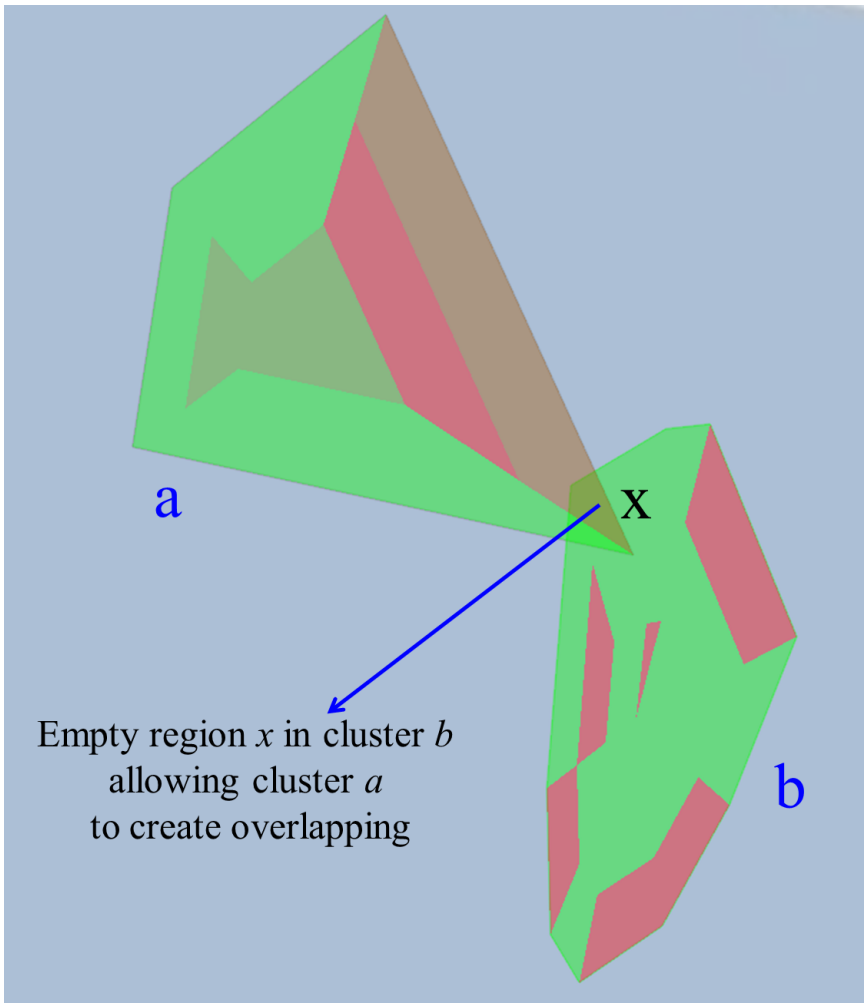



Figure 72: Example of overlapping situation in FlockViz design.

### 7.1.5 *Requirements of appropriate STC visualization and analysis*

Solving problems through cluster analysis depends on many factors. When we want to analyze different characteristics of the cluster at the same time and in a single view, then the design requirements get more complicated. While analyzing different datasets I found some factors which we should give importance for doing a successful spatio-temporal cluster analysis. First of all there must have a large number of unique objects in the dataset to build an informative set

of clusters. As some of the important users' requirements are to generate clusters of distinct objects, we should ensure that we have enough distinct objects in our dataset. Another important factor of cluster visualization is the clustering algorithm. It greatly affects the cluster generation and formation process. DBSCAN is a widely used clustering algorithm for spatio-temporal datasets. I used this in my thesis due to its wide acceptance in this domain. It might be possible that for a particular dataset other algorithms generate good results. Therefore, we should apply different clustering algorithms before analyzing a dataset. Another important criteria for spatio-temporal cluster analysis is basic parameter setting. This also affects the cluster generation and visualization process. For answering a particular analytic question we may look for clusters having a certain shape such as circular or in a certain location. To get those results at first we should test different parameters to get a sense of how the clustering is working for the particular dataset. Thus appropriate parameter setting can help us finding effective and accurate solutions.

## 7.2 CONCLUSION

Spatio-temporal movement is a common occurrence. With recent technological improvements in tracking technologies, a large amount of movement data is generated every day. To facilitate advanced spatio-temporal data analysis, such raw data needs to be converted into understandable visual information. Among many existing approaches, generation of clusters and appropriate visualization of clusters properties is one of the widely used high-level methods.

Future visual analytics systems will experience the need of such a method as a core component for the analytic pipeline. In this thesis, I explored various properties of spatio-temporal clusters and identified the need of it's information presentation capability. In this thesis, my primary contribution is a novel visualization, FlockViz, of spatio-temporal clusters and their multi-dimensional attributes.

In this thesis, I proposed FlockViz, a novel shape preserved spatio-temporal cluster (STC) visualization technique for simultaneous visualization of multiple properties of a cluster in a single view. FlockViz has several features that can show different properties of a single cluster both individually and also with respect to other clusters. To evaluate my proposed method, I conducted an experiment and explored several case studies using a wide range of spatio-temporal datasets. I also developed a framework of task categories which can be used to classify any kind of high-level analytic tasks related to spatio-temporal clusters. This task category framework is also useful to test any new visualization design of spatio-temporal clusters. Through a set of experiments, I demonstrated that FlockViz performs significantly better in terms of task completion time and error rate compared to the traditional approaches of analyzing STC. I also explored several research questions related to real life datasets and provided a guideline to solve those questions using FlockViz. These case studies show that FlockViz is capable of answering a wide range of analytical questions that are the subject of current research.. The evaluation section of my thesis indicates that FlockViz increases users' performance and quality of experience in performing analytic tasks. Finally, my thesis suggests that there is a need for efficient STC

visualization for analyzing spatio-temporal data and FlockViz gives a model to solve these problems by overcoming current limitations. Using such kind of visualization can help analysts find answers to queries relatively quickly and effectively in large spatio-temporal datasets.

In this thesis I have learned how to efficiently design spatio-temporal clusters for solving analytic tasks involving multi-dimensional attributes. My proposed visualization technique, FlockViz was able to solve these tasks efficiently which we noticed in the evaluation and case studies section. Finally some take home messages from my thesis are:

- Spatio-temporal cluster visualization of multi-dimensional attributes is very important in performing various high-level analytic tasks. While traditional visualization methods are not efficient in such cases, FlockViz can effectively visualize multiple attributes of clusters.

- While designing STC visualization, we should preserve original cluster shape and have the facility to show multiple attributes in a single view.

- Spatio-temporal cluster visualization methods should address a wide range of task categories and datasets from different domain to prove its universal usage.

- Various applications such as traffic planning, animal migration pattern finding, etc. can benefit from multi attribute STC visualization (implemented in FlockViz). Case studies in this

thesis use domain experts in those fields to evaluate my STC visualization method.

## 7.3 FUTURE WORK

The performance of my proposed method, FlockViz in the user experiment and case studies suggest that this visual design is extensible in various directions. I discuss these next.

### 7.3.1 *Conducting user studies with domain experts*

In this thesis I recruited students to participate in experiments to empirical test the performance of FlockViz. For a specific domain, user studies with participation of domain experts could gives us better results, specially subjective feedback based on the experience of domain experts would add a significant value to the evaluation process. It is also important to conduct a longitudinal study to analyze users' usage behaviour for any visualization system. By doing this study the users will be more familiar with the experimental dataset and it will be easy for them to answer complex analytic question. In future I will discuss with the analysts of different data domain such as a transit system, animal migration, surveillance system, etc. to set up a longitudinal study. Then I will ask domain experts to perform various complex analytic tasks in a longitudinal study to analyze the effect of long term use of FlockViz on the results.

### 7.3.2 *Testing alternate designs of FlockViz*

During the user study I used the original design of each flock part. In the FlockViz design section I mentioned some of the alternate designs of each part of the flock. In the future I will include alternative designs for each flock part to overcome some of the current limitations of FlockViz. I will test the performance of each alternate design to solve various analytic tasks. I will also conduct an experiment to rank those alternate designs based on users' performance and preference. I will try to come up with a mapping of task categories to those designs. This mapping will guide analysts to efficiently use FlockViz for finding the solution of a particular task.

### 7.3.3 *Determining the limitations of FlockViz for representing multiple data dimensions*

Most of the multi-dimensional data visualization methods are limited in terms of the number of properties that can be visualized simultaneously. FlockViz might also have the limitation on the number of dimensions that can be easily shown in the *flock shield* without affecting users' performance. In future I will run a user study to identify this limitation in various possible conditions such as for solving different types of tasks. From the result of those studies I will propose several alternate designs of *flock shield* to overcome those limitations. This research will help to guide users on handling a large number of dimensions using appropriate FlockViz designs.

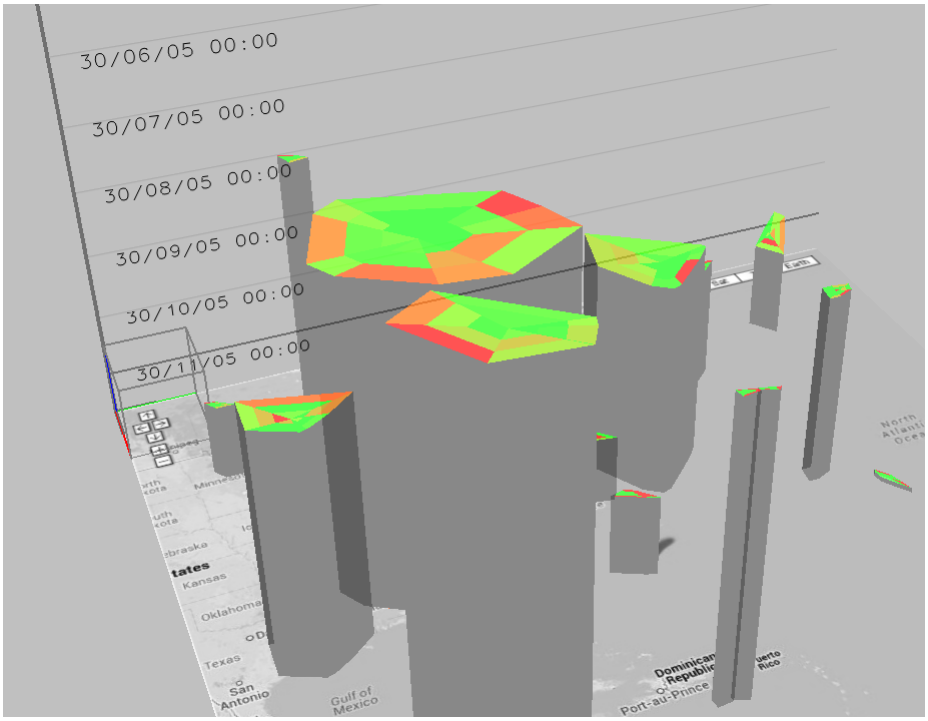



Figure 73: Current trial version of 3D FlockViz where z-axis provide temporal information.

### 7.3.4 *Extending FlockViz to 3D*

One of the main future works of my research will be to design, test and implement the full 3D version of my proposed FlockViz. In this thesis I have explored the 2D version of this design in detail. In order to test the possible 3D shape, I implemented only the *flock body* having very basic features of original FlockViz. Figure 73 shows my current 3D version of FlockViz design. In a 3D design the third dimension can be utilized to represent additional properties of the cluster. For example, in the Figure 73 I used the third dimension to show temporal information. However, it will be challenging to convey appropriate depth and color information in 3D FlockViz. Because in 3D there will be a chance that one cluster will occlude

the other and we will get a blended color due to this occlusion. I will do an extensive research on this problem and provide efficient solutions for to overcome these challenges for the 3D FlockViz.

# A

# RESULTS FROM EXPERIMENTS

## A.1 EXPERIMENT RESULTS

*Task Completion Time*

| **Main and Interaction effect** | | |
|---|---|---|
| *Visualization Techniques* | $F_{1,15} = 8.91$ | $p = 0.02$ |
| *Data Density* | $F_{1,15} = 2.93$ | $p = 0.089$ |
| *High Level Task Categories* | $F_{1,15} = 22.972$ | $p = 0.002$ |
| *Low Level Task Categories* | $F_{3,45} = 15.767$ | $p < 0.001$ |
| *Visualization Techniques×Data Density* | $F_{1,15} = 8.716$ | $p = 0.021$ |
| *Visualization Techniques×High Level Task Categories* | $F_{1,15} = 13.908$ | $p < 0.001$ |
| *Visualization Techniques×Low Level Task Categories* | $F_{3,45} = 5.93$ | $p = 0.004$ |

*Error rate*

| **Main and Interaction effect** | | |
|---|---|---|
| *Visualization Techniques* | $F_{1,15} = 19.305$ | $p = 0.003$ |
| *Data Density* | $F_{1,15} = 1.809$ | $p = 0.181$ |
| *High Level Task Categories* | $F_{1,15} = 1.988$ | $p = 0.160$ |
| *Low Level Task Categories* | $F_{3,45} = 9.65$ | $p < 0.001$ |
| *Visualization Techniques×Data Density* | $F_{1,15} = 0.388$ | $p = 0.553$ |
| *Visualization Techniques×High Level Task Categories* | $F_{1,15} = 1.883$ | $p = 0.212$ |
| *Visualization Techniques×Low Level Task Categories* | $F_{3,45} = 9.26$ | $p < 0.001$ |

# BIBLIOGRAPHY


[1] Gennady Andrienko and Natalia Andrienko. Interactive visual clustering of large collections of trajectories. In *VAST'08, Visual Analytics Science and Technology*, pages 51–58. IEEE, 2008.

[2] Gennady Andrienko, Natalia Andrienko, Martin Mladenov, Michael Mock, and Christian Politz. Discovering bits of place histories from people's activity traces. In *IEEE Symposium on VAST 2010*, pages 59–66, 2010.

[3] Gennady Andrienko, Natalia Andrienko, Salvatore Rinzivillo, Mirco Nanni, Dino Pedreschi, and Fosca Giannotti. Interactive visual clustering of large collections of trajectories. In *IEEE Symposium on VAST 2009*, pages 3–10, 2009.

[4] Gennady Andrienko, Natalia Andrienko, and Stefan Wrobel. Visual analytics tools for analysis of movement data. *SIGKDD Explorations Newsletter*, 9.2:38–46, 2007.

[5] N aatalia Andrienko and Gennady Andrienko. Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics*, 17:205–219, February 2011.

[6] Natalia Andrienko and Gennady Andrienko. Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, 12:3–24, 2013.

[7] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, pages 49–60, New York, NY, USA, 1999. ACM.

[8] Yann Arthus-Bertrand. Stunning images of herds from above. `http://www.environmentalgraffiti.com/featured/herds-from-above/9450`.

[9] Michael Balzer and Oliver Deussen. Level-of-detail visualization of clustered graph layouts. In *6th International Asia-Pacific Symposium*, pages 133–140. IEEE, 2007.

[10] Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.*, 60(1):208–221, January 2007.

[11] Robert Bringhurst. *The Elements of Typographic Style*. Hartley & Marks, 2002.

[12] Hong Bui. Traffic jam - government's responsibility or citizens' awareness? `http://sociologyiu09.wordpress.com/2010/01/13/traffic-jam-%E2%80%93-government%E2%80%99s-responsibility-or-citizens%E2%80%99-awareness/`.

[13] cab4fun. Traffic jams from hell. `http://www.cab4fun.com/traffic-jams-from-hell/`.

[14] Keke Chen and Ling Liu. Clustermap: Labeling clusters in large datasets via visualization. In *Thirteenth ACM international conference on Information and knowledge management,*, pages 285–293. ACM, 2004.

[15] Tarik Crnovrsanin, Chris Muelder, Carlos Correa, and Kwan-Liu Ma. Proximity-based visualization of movement trace data. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 11–18, Atlantic City, New Jersey, USA, October 2009.

[16] Urska Demsar and Kirsi Virrantaus. Space-time density of trajectories: Exploring spatio-temporal patterns in movement data. *Int. J. Geogr. Inf. Sci.*, 24(10):1527–1542, October 2010.

[17] Ryan Eccles, Thomas Kapler, Robert Harper, and William Wright. Stories in geotime. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, VAST '07, pages 19–26, Washington, DC, USA, 2007. IEEE Computer Society.

[18] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 96:226–231, 1996.

[19] Anna Fredrikson, Chris North, Catherine Plaisant, and Ben Shneidrman. Temporal, geographical and categorical aggregations viewed through coordinated displays: a case study with highway incident data. In *NPIVM '99 Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, pages 26–34, New York, NY, USA, 1999. ACM.

[20] Peter Gatalsky, Natalia Andrienko, and Gennady Andrienko. Interactive analysis of event data using space-time cube. In *Proceedings of the Information Visualisation, Eighth International Conference*, IV '04, pages 145–152, Washington, DC, USA, 2004. IEEE Computer Society.

[21] Diansheng Guo, Jin Chen, Alan M. MacEachren, and Ke Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, November 2006.

[22] Hanqi Guo, Zuchao Wang, Bowen Yu, Huijing Zhao, and Xiaoru Yuan. Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. In *Proceedings of the 2011 IEEE Pacific Visualization Symposium*, PACIFICVIS '11, pages 163–170, Washington, DC, USA, 2011. IEEE Computer Society.

[23] Robert Haas. Duck-like image of a flock of pink flamingos. `http://thexodirectory.com/2011/08/duck-like-image-of-a-flock-of-pink-flamingos/`.

[24] Torsten Hagerstrand. What about people in regional science? *Papers in regional science*, 24(1):7–24, 1970.

[25] Mark Harrower. A look at the history and future of animated maps. *Cartographica: The international journal for geographic information and geovisualization*, 39(3):33–42, 2004.

[26] William L. Hibbard, Brian E. Paul, David A. Santek, Charles R. Dyer, André L. Battaiola, and Marie-Françoise Voidrot-Martinez. Interactive visualization of earth and space science computations. *Computer*, 27(7):65–72, July 1994.

[27] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: A tool for visualization multi-dimensional geometry. In *Human Machine Interactive Systems*, pages 199–233, USA, 1991. Springer.

[28] Yuri Ivanov, Christopher Wren, Alexander Sorokin, and Ishwinder Kaur. Visualizing the history of living spaces. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1153–1160, November 2007.

[29] Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[30] Eser Kandogan. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. *IEEE Information Visualization Symposium*, 650, 2000.

[31] Thomas Kapler and William Wright. Geotime information visualization. In *INFOVIS '04 Proceedings of the IEEE Symposium on Information Visualization*, pages 25–32, Washington, DC, USA, 2004. IEEE Computer Society.

[32] Leonard kaufman and Peter Rousseeuw. Clustering by means of medoids. 1987.

[33] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Gorg, Jorn Kohlhammer, and Guy Melancon. Visual analytics: Definiton, process, and challenges. pages 154–175, 2008.

[34] M.-J. Kraak. The space-time cube revisited from a geovisualization perspective. In *21st International Cartographic Conference*, pages 1988–1996, 2003.

[35] M.-J. Kraak. Timelines, temporal resolution, temporal zoom and time geography. In *22nd International Cartographic Conference*, Spain, 2005.

[36] Per Ola Kristensson, Nils Dahlbäck, Daniel Anundi, Marius Björnstad, Hanna Gillberg, Jonas Haraldsson, Ingrid Mårtensson, Mathias Nordvall, and Josefine Ståhl. An evaluation of space time cube representation of spatiotemporal patterns. *IEEE Transactions on Visualization and Computer Graphics*, 15(4):696–702, July 2009.

[37] Robert Kruger, Dennis Thom, Michael Worner, Harald Bosch, and Thomas Ertl. Trajectorylenses-a set based filtering and exploration technique for long term trajectory data. *Computer Graphics Forum*, 32:451–660, June 2013.

[38] Lau. Paint the sky with stars. `http://lau-paint-the-sky.blogspot.ca/2012/04/shopping-saturday-with-mimi-and-kiki.html`.

[39] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, September 2006.

[40] Suresh K. Lodha and Arvind K. Verma. Spatio-temporal visualization of urban crimes on a gis grid. In *Proceedings of the 8th ACM International Symposium on Advances in Geographic Information Systems*, GIS '00, pages 174–179, New York, NY, USA, 2000. ACM.

[41] Alan M. MacEachren. How maps work: Representation, visualization, and design. 2004.

[42] Andrew Vande Moere. Time-varying data visualization using information flocking boids. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 97–104, Washington, DC, USA, 2004. IEEE Computer Society.

[43] Movebank. Eagle research. `https://www.movebank.org/`.

[44] Maxwell Guimaraes de Oliveira and Claudio de Souza Baptista. Geostat-a system for visualization, analysis and clustering of distributed spatiotemporal data. In *XIII GEOINFO*, pages 108–119, Campos do Jordao, Brazil, November 2012.

[45] Donna J. Peuquet. Making space for time: Issues in space-time data representation. *Geoinformatica*, 5(1):11–32, March 2001.

[46] Doantam Phan, Ling Xiao, Ron Yeh, Pat Hanrahan, and Terry Winograd. Flow map layout. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, INFOVIS '05, pages 29–, Washington, DC, USA, 2005. IEEE Computer Society.

[47] Glenn Proctor and Chris Winter. Information flocking: Data visualisation in virtual worlds using emergent behaviours. In *Virtual Worlds*, pages 168–176, Berlin Heidelberg, 1998. Springer.

[48] K Purushottama Rao, Uppe Nanaji, and Y Swapna. Spatiotemporal data mining: Issues, tasks and applications. *International Journal of Computer Science and Engineering Survey*, 3, 2012.

[49] Raptor. Research projects. `http://www.raptorview.org`.

[50] Craig W. Reynolds. Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH Computer Graphics*, 21:25–34, July 1987.

[51] Salvatore Rinzivillo, Dino Pedreschi, Mirco Nanni, Fosca Giannotti, Natalia Andrienko, and Gennady Andrienko. Visually driven analysis of movement data by progressive clustering. *Information Visualization*, 7(3):225–239, June 2008.

[52] Aidan Slingsby, Jo Wood, and Jason Dykes. Treemap cartography for showing spatial and temporal traffic patterns. *Journal of Maps*, 6(1):135–146, 2010.

[53] Terry A. Slocum, Stephen C. Yoder, Fritz C. Kessler, and Robert S. Sluter. Maptime: Software for exploring spatiotemporal data associated with point locations. *Cartographica: The international journal for geographic information and geovisualization*, 37(1):15–32, 2000.

[54] O. Spakov and D. Miniotas. Visualization of eye gaze data using heat maps. *Electronics and electrical engineering*, 2:55–58, 2007.

[55] Strand.com. Transportation. `http://www.strand.com/services/transportation/roundabouts/`.

[56] Tara and Karina. Tara and karina go out. `http://www.taraandkarinagoout.com/wp-content/uploads/2010/09/brown-bluff-tay-head-and-one-emperor-penguin-036.jpg`.

[57] Alice Thudt, Dominikus Baur, and Sheelagh Carpendale. Visits: A spatiotemporal visualization of location hitories. In *EuroVis 2013, Eurographics Conference on Visualization*. IEEE, 2013.

[58] Christian Tomniski, James Abello, and Heidrun Schumann. Axes-based visualizations with radial layouts. In *Sysmposium on Applied computing*. ACM, 2004.

[59] Christian Tomniski, Heidrun Schumann, Gennady Andrienko, and Natalia Andrienko. Stacking-based visualization of trajectory attribute data. volume 18, pages 2565–2574. IEEE, 2012.

[60] Edward R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Cheshire, CT, USA, 1997.

[61] Ulanbek Turdukulov and Menno-Jan Kraak. Visualization of events in time-series of remote sensing data. In *ICC 2005: Proceedings of the 22nd international cartographic conference: mapping approaches into a changing world*, pages 9–16, 2005.

[62] Irina Rev Vasiliev. Mapping time. *Cartographica: The international journal for geographic information and geovisualization*, 34(2):1–51, 1997.

[63] Walerian Walawski. Did you know...what that "flock" of birds is actually called? `http://www.pacificnorthwestbirds.com/tag/owls/`.

[64] Günter Wallner and Simone Kriglstein. A spatiotemporal visualization approach for the analysis of gameplay data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1115–1124, New York, NY, USA, 2012. ACM.

[65] Colin Ware, Roland Arsenault, Matthew Plumlee, and David Wiley. Visualizing the underwater behavior of humpback whales. *IEEE Comput. Graph. Appl.*, 26(4):14–18, July 2006.

[66] Wikipedia. 2005 atlantic hurricane season. `https://en.wikipedia.org/wiki/2005_Atlantic_hurricane_season`, 2005.

[67] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63, 2009.

[68] Niels Willems, Roeland Scheepens, Huub van de Wetering, and Jarke J.van Wijk. Visualization of vessel traffic. In *Situation Awareness with Systems of Systems*, pages 73–87. Springer, 2013.

[69] Xiaowei Xu, Martin Ester, Hans-Peter Kriegel, and Jörg Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *Proceedings of the Fourteenth International Conference on Data Engineering*, ICDE '98, pages 324–331, Washington, DC, USA, 1998. IEEE Computer Society.

[70] Ed Yong. How the science of swarms can help us fight cancer and predict the future. `http://www.wired.com/wiredscience/2013/03/powers-of-swarms/all/`.

## COLOPHON

This thesis was typeset with the `pdflatex` LaTeX $2_\varepsilon$ interpreter using Hermann Zapf's *Palatino* type face for text and math and *Euler* for chapter numbers. The listings were set in *Bera Mono*.

The typographic style of the thesis was based on André Miede's wonderful `classicthesis` LaTeX style available from CTAN. My modifications were limited to those required to satisfy the constraints imposed by my university, mainly 12pt font on letter-size paper with extra leading. Miede's original style was inspired by Robert Bringhurst's classic *The Elements of Typographic Style* [11].

*Final Version* as of April 24, 2014 at 12:16.