

AN ABSTRACT OF THE THESIS OF

Mariam Guizani for the degree of Master of Science in Computer Science
presented on June 1, 2018.

Title: Coincident Nodes Multi-Edge Graph for Simultaneous Decision and
Objective Space Multi-Dimensional Visualization

Abstract approved: _____

Eugene Zhang

The importance of data visualization is becoming increasingly more substantial to the field of optimization and engineering design where a carefully designed visualization of the data on decision parameters (i.e Decision Space) and performance functions (i.e Objective Space) is critical to the success of the decision making process.

One of the main goals of data visualization is to unveil the different patterns, trends and relationships that the data encapsulates. However, this aforementioned goal becomes challenging when visualizing multidimensional decision and objective space data both qualitatively and quantitatively. In fact, in order to discover the patterns and inter-variable relationships that the data encapsulates, a holistic visualization approach, where all the variables are simultaneously represented, is

required. However, holistically mapping multidimensional data in a single 2D visual is a challenging task that could result in a cognitively overwhelming output. Consequently, we aim to reach a balance between the desired holistic view that facilitates pattern discovery, and a clear user friendly visualization.

In this thesis, we present a novel holistic visualization model for pattern discovery in multidimensional decision and objective space data structures and demonstrate its usage in a watershed conservation plan context. We use a coincident nodes and multi-edge network map visualization to represent users' decisions in terms of watershed conservation plan practices and goals without losing the geographical knowledge provided by a map. In reality the decision and objective space are highly related. This simultaneous combination of the geographical information, decision space and objective space yields to an efficient identification of existing patterns that are further validated using a set of predefined statistical methods.

©Copyright by Mariam Guizani
June 1, 2018
All Rights Reserved

Coincident Nodes Multi-Edge Graph for Simultaneous Decision and
Objective Space Multi-Dimensional Visualization

by

Mariam Guizani

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented June 1, 2018
Commencement June 2018

Master of Science thesis of Mariam Guizani presented on June 1, 2018.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Mariam Guizani, Author

ACKNOWLEDGEMENTS

First of all, I would like to thank my adviser Dr. Eugene Zhang for his guidance and help throughout my master's program and for introducing me to my research topic. I would also like to thanks Dr. Meghna Babbar-Sebens for her availability, help and insightful comments throughout the research phase of this program.

Second, I would like to thank the rest of my committee Dr. Yue Zhang and Dr. David Michael Kling for accepting to be part of my committee and giving me the opportunity to present my work.

Finally, I would like to express my profound gratitude to my parents for their love, support and continuous encouragement throughout my years of study in general and my master's program in particular.

TABLE OF CONTENTS

	<u>Page</u>
1 Introduction	1
2 Literature Review	3
3 Methodology	9
3.1 Predictive Data Centered Theory	10
3.2 Qualitative Color Coding	12
3.3 Aesthetic Criteria	13
4 Context and Case Study	16
4.1 Eagle Creek Watershed	17
4.2 Users Involved	18
4.3 Data Acquisition Process	18
5 Proposed Visualization Technique	22
5.1 Visualization Metrics	22
5.2 Proposed Multi-Variables Network Map Visualization	23
5.3 Analysis Methods	31
6 Results and Discussions	35
6.1 Patterns Prior to Human Guided Search	35
6.1.1 Variables Correlation Patterns	40
6.1.2 User Preferences Correlation Patterns	41
6.2 Pattern Post Human Guided Search	51
6.2.1 Variables Correlation Patterns	52
6.2.2 User Preferences Correlation Patterns	53
6.3 Inter-Session Comparison	59
6.4 Hot Spot Map and Network Map Visualization	62
7 Conclusion	65
7.1 Limitations and Future Work	66

TABLE OF CONTENTS (Continued)

	<u>Page</u>
Bibliography	67

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
2.1 Example from [18] of a decision and objective space representation using the three objective projection glyphs.	4
2.2 Example of the visualization of objective space variables[18].	5
4.1 Watershed example [33].	16
4.2 Eagle Creek outline [1].	17
4.3 Eagle Creek sub-basins [1].	17
4.4 Eagle Creek stream [1].	17
4.5 Map of spacial distribution of sub-basins of interest (SBint) in the watershed. Model A Surrogate represent the 7 surrogate users in this study and the top right map represents the water stream in the 6 local sub-basins of interest [26].	19
4.6 The sequence of the different sessions. HS stand for Human guided search session[26].	21
5.1 Nodes and edge example: (a) a node representation of two sub-basins: sub-basin 68 and sub-basin 69, (b) an example of a water stream edge representation connecting sub-basin 68 to sub-basins 69 and (c) the combination of the edge and the two nodes.	24
5.3 An example of the possible edge visualizations: (a) peak flow reduction, (b) sediment reduction (c) nitrate reduction, (d) cost reduction and (e) simultaneous visualization of all the objective space variables.	26
5.2 An example of visualizing all the decision space and x number of objective space variables. x being in the range of [1..4].	27
5.4 A side by side comparison between the watershed map and the network representation before adding the decision space and objective space variables. The numbers that are underneath each sub-basins represent the sub-basins labels.	28
5.5 An example of the simultaneous representation of the decision and objective space in the ECW.	29

LIST OF FIGURES (Continued)

<u>Figure</u>	<u>Page</u>
5.6 An example of visualizing one decision space variable (cover crops) and x number of objective space variables. x being in the range of [1..4]. . .	30
5.7 An example of all the different levels of decision space visualization. . .	32
5.8 An example of single variable decision space visualizations.	33
6.1 Decision and objective space coincident nodes multi-edge network map visualization of participant 1 to 4 in Modal A surrogate for the "I like it design" rating relative to a non-Interactive Genetic Algorithm based session.	36
6.2 Decision and objective space coincident nodes multi-edge network map visualization of participant 5 to 7 in Modal A surrogate for the "I like it design" rating relative to a non-Interactive Genetic Algorithm based session.	37
6.3 The objective space skeletons of all the individuals of group A surrogate taking into consideration only the "I like it design" rating generated by the non-Interactive Genetic Algorithm.	39
6.4 Average separation in the decision space between individual 2 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	42
6.5 Average separation in the decision space between individual 3 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	43
6.6 Average separation in the decision space between individual 4 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	44
6.7 Average separation in the decision space between individual 5 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	45
6.8 Average separation in the decision space between individual 6 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	46

LIST OF FIGURES (Continued)

<u>Figure</u>	<u>Page</u>
6.9 Average separation in the decision space between individual 7 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	47
6.10 Decision and objective space coincident nodes multi-edge network map visualization of participant 1 to 4 in Modal A surrogate for the "I like it design" rating relative to Interactive Genetic Algorithm sessions.	49
6.11 Decision and objective space coincident nodes multi-edge network map visualization of participant 5 to 7 in Modal A surrogate for the "I like it design" rating relative to Interactive Genetic Algorithm Sessions.	50
6.12 The objective space skeleton of all the individuals of group A surrogate taking into consideration the "I like it" designs rating generated by the Interactive Genetic Algorithm.	51
6.13 Average separation in the decision space between individual 2 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	53
6.14 Average separation in the decision space between individual 3 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	54
6.15 Average separation in the decision space between individual 4 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	55
6.16 Average separation in the decision space between individual 5 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	56
6.17 Average separation in the decision space between individual 6 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	57
6.18 Average separation in the decision space between individual 7 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.	58

LIST OF FIGURES (Continued)

<u>Figure</u>	<u>Page</u>
6.19 Correlation between interactive and non interactive sessions for each individual and each decision and objective space variable.	60
6.20 A side by side comparison between a single variable (cover crops) hot spot map representation and our holistic network map visualization.	62

Chapter 1: Introduction

In today's data driven society, data exploration and understanding are more prevalent than ever. With the rising awareness of the importance of data and its potential applications, data collection has grown extensively across different sectors of society. In addition, a lot of effort and resources have been put into understanding and unveiling meaningful insights from data sets.

One of the main approaches to presenting and exploring data sets is data visualization. Data visualization is a subfield of computer science that aims to uncover visual patterns about a particular data set. More specifically data visualization enables its users to make, explore and validate hypothesis from their data [11]. For all these reasons, a lot of attention has been surrounding data visualization across many disciplinary fields [13][27]. However, data visualization in general constitutes a challenging task and simultaneous multivariate data visualization in particular can be both tedious and confusing [6]. In fact, encoding multidimensional data in a single 2D representation while keeping the cognitive load in a manageable range can be an arduous task. In addition, data visualization is critical when it come to optimization and decision making. In fact carefully designed visualization of the data on decision parameters (i.e Decision Space) and performance functions (i.e Objective Space) is very important to the success of the decision making process. Moreover, representing data that sits into two different spaces (decision and objec-

tive space) linked by a clear semantic relationship, requires a lot of effort to be put into assigning the right graphical representation to each end every nominal variable [12]. Additionally, visualizing that multidimensional data both qualitatively and quantitatively, while keeping in mind the limitations of the human perception adds considerably to the complexity of the problem. A holistic, complete view of all the variables is usually more efficient and accurate than the aggregation of distinct visuals when it comes to uncovering meaningful patterns and insights but it can become cognitively challenging to understand [27]. For this reason, in order for the visualization to be successful, a balance between the desired holistic view and the usability and clarity of the representation must be reached.

Because of the identified challenges surrounding simultaneous multivariate decision and objective space visualization, we study different visualization methodologies as well as different theoretical foundations of information visualization to examine their guidelines and limitations in order to contribute with a novel visualization model that holistically visualizes a multivariate decision and objective space data set. The upcoming chapters provide a walk through the design choices and trade offs that were used in our novel visualization model as well as a demonstration and assessment of this visual technique in a watershed conservation plan context.

Chapter 2: Literature Review

Data visualization has been used in different fields in order to better understand and investigate a data set. It aims to efficiently communicate patterns, trends and relations present in the data set [36].

Different visualization like histograms, scatter plots, box plots and bar charts are effective in terms of outliers and gaps detection in low dimensionality scenarios [31]. Even though box plots provides statistical insight about a visualized data set, this method is only suitable for quantitative data and can only visualize a single attribute at a time. Similarly, even though histograms provide clear distribution visualization, they are limited to a single variable visualization. On the other hand, scatter plot could result in a cluttered visualization caused by the potential data overlap which is common when studying big data sets [23]. Additionally, scatter plots are limited to a maximum number of three attributes per visualization which make them unsuitable for multidimensional data set visualization [31]. Heat map is another visualization model that is limited in terms of the number of attributes displayed. In fact, heat maps are colored matrix representing a total of two variables at a time. Hot spot maps, However, are specifically targeted toward data that has geographical information as one of its dimensions. Even though hot spot maps are efficient for visualizing quantitative data with the added geographical context, this visualization model is only suitable for visualizing one variable at a

time.

Small multiples of any of these visualizations could cover all the dimensions of a multivariable data set, however, it would be challenging for the human brain to mentally link the visuals and discover meaningful and non misleading insight.

The three objective projection glyph [18] and the parallel coordinate plots [15] [37] are commonly used for decision and objective space representation.

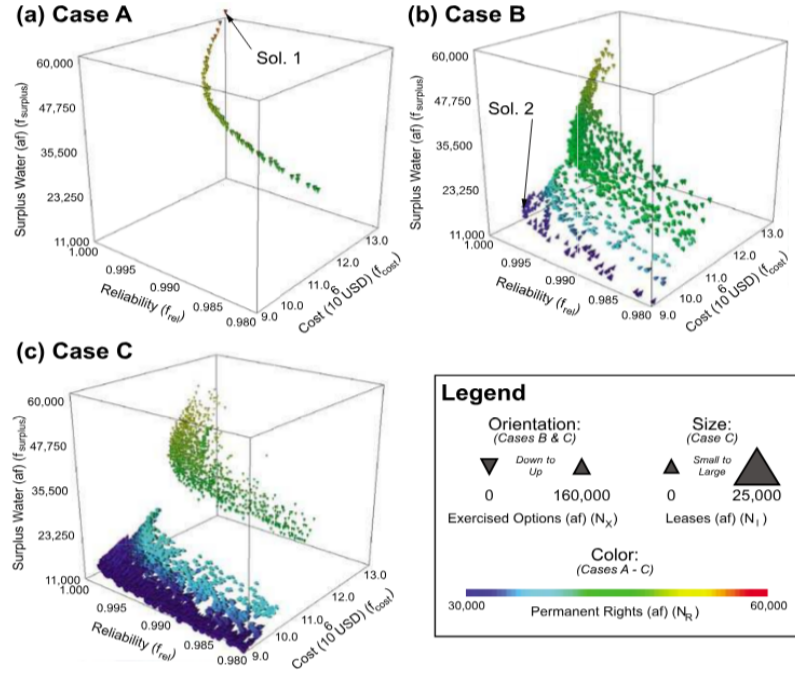


Figure 2.1: Example from [18] of a decision and objective space representation using the three objective projection glyphs.

[Figure 2.1] and [Figure 2.2] have been used in [18] for a visual analysis and comparison of four water supply strategies in the Lower Rio Grande Valley in Texas.

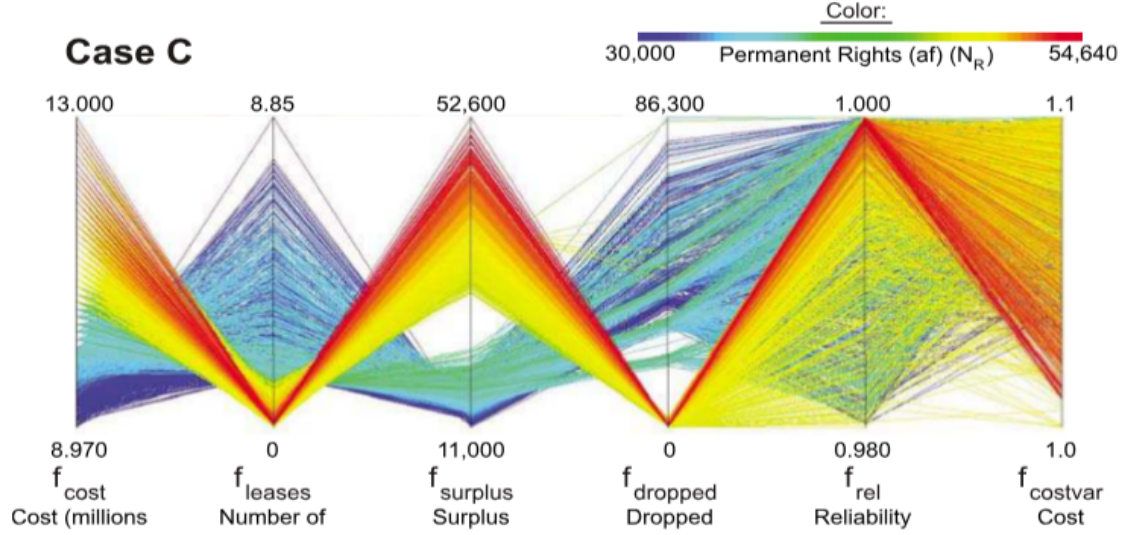


Figure 2.2: Example of the visualization of objective space variables[18].

These two visualization methods also used in [14] and [40] have the advantage of being multivariable visualizations that perform better than other visualization methods like box plots and scatter plots when it comes to displaying multiple variables at a time. Projection glyphs visualizations [Figure 2.1] have the advantage of representing up to three variables in both the decision and objective space. However, although the representation in the objective space provides easily distinguishable results, augmenting the cardinality of the decision space can diminish readability. In fact, when considering the legend in [Figure 2.1] the decision attributes represented by the size and direction of the arrows are hardly visible when using the "big picture" without zooming in. Moreover, the absolute maximum of a three variables representation in both the decision and objective space presents a limitation especially when a study is taking more than three objective variables into consideration which is very common in practice. One way that has been used

to face that limitation is to pair the projection glyphs visualization with a parallel coordinate visualization [18]. However, mentally linking two different visualizations could be both challenging and misleading. The parallel coordinate visualization is a multi-objective variable representation that is particularly suitable for the detection of conflicts and correlation between objective variables. However, this previously mentioned visualization method is limited to the visualization of one decision variable at a time [18] which makes it unsuitable in terms of simultaneously visualizing the decision and objective space.

When representing structured data with all its inherent relations [13] graph drawing visualizations are a suitable solution. Since simultaneously visualizing the decision and the objective space could be considered as a structured data representation problem, it would be more suitable to consider graph related representations.

2.0.0.1 Graph Drawing and Flow Maps

Graph drawing or network visualization diagrams are interdisciplinary information visualization methods that have been used in different applications such as social network analysis, bioinformatics, linguistics, economics, chemistry and computer network diagrams. Network diagrams are two dimensional representations composed of a set of nodes and edges displayed using a specific layout [39]. Different design options of the same network diagram fulfill different aesthetic and usability criteria.

There are different ways of visualizing graphs. The tree layout [35] [29] is a

layout where each pair of nodes is connected by only one edge in a hierarchical ordering. Radial map [9] is a way of representing a tree graph while expanding outwards on the periphery of concentric circles. Balloon tree layout [13], represents the children of each node on the circular surrounding of that node. These aforementioned layouts fall into the simple graph category where each pair of nodes is linked by means of only one edge. On the other hand, multiple edge graphs or multigraphs allow the presence of loops around a pair of nodes. The presence of multiple edges between nodes allows information gain in terms of visualization especially when representing complex relationships that cannot be contained in a simple graph.

Flow map visualization models are mainly used in cartography [38]. They are obtained by the superposition of maps and flow charts. The main goal of a flow map is to represent a certain movement between regions [16] [17]. It is common for a flow map visualization to assimilate the importance of a movement with the thickness of the edge between two geographical regions.

Even though multiple edge graph drawing gives some flexibility in terms of the number of variables visualized, that multi-variable visualization is restricted to qualitative visualization. Moreover, using edges for visualizing certain variables implies that those variables represent a semantic connection between two nodes.

When aiming for a simultaneous qualitative and quantitative visualization of both the decision and objective space of a data set, we should think about easily distinguishable shapes [12] and also a way of visualizing the quantity of each variable.

In the next sections of this thesis, we propose a novel coincident nodes, multi-edge network visualization for both decision and objective spaces.

Since our proposed method shares some of its characteristics with graph drawing and flow maps, we follow some of the graph drawing and flow maps aesthetic criteria that have been proven to be efficient for usability purposes and relevant for our specific representation model.

Chapter 3: Methodology

Visualizing multidimensional data with two distinct qualitative spaces, imposes significant challenges. On the other hand, separately visualizing variables in different graphs and trying to mentally link the different variables together can constitute a big cognitive load for the user and can end up being misleading. Simultaneously visualizing both qualitative and quantitative multidimensional data, can easily result in a cluttered representation, thus decreasing its usability. Even though [Section 2.0.0.1] presents interesting representations that could encapsulate the connection between two different spaces of variables, these visualizations are not the best candidates for a multidimensional data representation.

Because of the absence of a general framework for information visualization, it is hard to predict in advance the outcome as well as the success and validity of a certain visualization. In order to compensate this theoretical lack, the authors of [27] have drawn some frameworks from associated fields to provide guidelines for information visualization. From assimilating the understanding of a visual representation to the understanding of a language authors of [27] develop three information visualization theories based on the analogy with the linguistic model and communication theory.

Considering the specific visualization goal that we want to achieve, we choose to focus on the first visualization theory presented in [27]. Predictive data centered

theory is going to help guide the construction of our visualization model.

3.1 Predictive Data Centered Theory

Predictive Data Centered Theory [27] focuses on the data itself in order to facilitate its exploration and understanding. This theory is based on the abstraction of characteristic features in order to enable pattern discovery. It is important to take into consideration the type, structure, properties of the data in order to have a preliminary idea about the type of pattern to target.

A data set could be thought of as a set of values within a particular context that is characterized by [27]:

- Attributes which constitute a list of the observed properties.
- Referrer is a data element that encapsulates an aspect of the context that the data is reflecting. In our case the decision and objective spaces could be referrers.

The goal of visualizing a data set is to discover some meaningful patterns that lead to a better understanding of the phenomena it is presenting. Authors of [10] define a pattern as an "expression E in some language L describing facts in a subset F_E of a set of facts F ".

Since we are visualizing numerical attribute values that are ordered and both the decision and objective space that have a semantic ordering between them, the

patterns that we are looking to observe are mainly "increase, decrease, peaks and low points" [27].

Using characteristics that best align with the aforementioned theory, we introduce our Coincident Nodes Multi-Edges for Simultaneous Decision and Objective Space Visualization in its general form.

Let the graph $G = (V, E)$ be composed of two types of sets, each set represents a different space where:

- $V = \bigcup_{i=1}^n v_i$ is the set of nodes representative of the first referrer: the decision space.
- $E = \bigcup_{i=1}^m e_i$ is the set of edges representative of the second referrer: the objective space.

The decision space corresponds to the set of variables that are modifiable and the objective space represents the set of variables that result from the decision space.

Since the decision space variables of each node influence all the objective space variables, the linkage between each pair of nodes simultaneously represents all the objective space variables.

This representation can visualize a maximum of two spaces each space containing n and m variables respectively. The number of variables in the different spaces could be but is not necessarily the same.

The differentiation between the two spaces or referrers is based on the shape (e.g nodes, edges) that characterizes each space. Additionally, the link between the

two spaces is representative of the actual link present in the data set. Moreover, our proposed visualization aims to represent both quantitative and qualitative data.

The quantitative data is represented by the diameter of the node and the width of the edge. This representation is meant to help discover pattern of "increase, decrease, peaks and low points" [27] as described in section 3.1.

On the other hand, the qualitative data is represented using the color coding guidelines presented in the following section [Section 3.2].

3.2 Qualitative Color Coding

Even though color choices can be highly subjective, the color palette selection in data visualization is not solely based on aesthetic criteria [32][7]. In color theory, the use of color to encode information depends on the type of data itself. There are three types of data:

- Sequential data: Data that goes from a minimum low value to a maximum high value. This type of data is better visualized using a gradient scale with the darkest hue corresponding to the maximum value and the lightest hue corresponding to the minimal value [32][30][7].
- Divergent data: Data values that sit on opposite sides of a spectrum with a neutral value in the middle. This type of data requires the usage of two different colors that decrease in intensity toward the middle of the scale [32][7][30].

- Qualitative data: Qualitative data is nominal [5] rather than numerical. This type of data requires the usage of as many distinct hues as the distinct variables present.

In our visualization case, we are using colors to encode qualitative data. However, multi-dimensional data can pose certain difficulties when selecting a color palette [5]. In fact, when mapping colors to a qualitative set of data, the colors have to be distinguishable and contrasted. In addition to that, the usage of color coding should not exceed seven colors [5] in order for the human brain to be able to discriminate information easily [20]. Kelly the author of [19] came up with a list of the twenty two most contrasted colors that help distinguish the different variables they represent. We based the color selection of this visualization on Kelly’s list of color as well as our own experimentation with different color palettes.

Additionally, in order to obtain an easily readable visualization, we follow some aesthetic guidelines that are presented in the following section [Section 3.3]

3.3 Aesthetic Criteria

Even though there is no consensus on the effectiveness of curved edges over straight edges in graph drawing [17], some research [28] show that curved edges with a minimal number of bends improves the usability of a visualization. In addition, we notice a wide use of curved lines in flow map representation [Section 2.0.0.1].

Authors of [17] have conducted a user study that proves that curved lines in flow maps minimizes the error rate compared to straight lines. According to these

previous research findings and to the specificities of our visualization model that represents multiple edges between two nodes, straight edges are not a suitable option. Thus, we choose to use symmetrically curved edges while minimizing the number of bends.

Moreover, findings in graph drawing, show that minimizing the number of edge crossing facilitates graph readability by reducing the error rate [17].

Even though in flow maps arrow heads are the most common design option for indicating direction, graph drawing has different ways of indicating direction such as tapered flow width or using convention that eliminate the necessity of the direction indication. In our case, knowing that we are displaying multiple variables at a time and knowing that the edge width is going to be used to represent quantitative data, we choose to follow a top down direction convention [17].

To summarize, different design principles have been used in our model in order to increase readability and minimize visual clutter. Some of these methods were inspired from efficient aesthetic criteria used in graph drawing and flow map diagrams [17] and others were proper to the characteristics of our multi-variable visualization method:

- Symmetric curved edge.
- Higher curve for longer edges.
- Radial representation of edges around nodes.
- Minimal edge on edge and edge on node crossing.

- Minimal edge overlapping.
- No arrow heads for direction.
- Coincident nodes with different opacity.

Chapter 4: Context and Case Study

A watershed [Figure 4.1] is an area of land between different water flowing streams that separates rivers, sub-basins, aquifer or even the ocean [33].



Figure 4.1: Watershed example [33].

The data we are working with is the results of an interactive optimization method embedded in the web-based WESTORE tool [4]. The WESTORE tool [4] is based on interactive genetic algorithms [3] which are evolved optimization methods that adapt their search according to the users' feedback. Because of the fact that watershed sub-basins stakeholders are in the center of the WESTORE [4] tool, this process takes into account both the quantifiable goals and the unquantifiable subjective preferences.

4.1 Eagle Creek Watershed

The proposed visualization model is demonstrated for the Eagle Creek Watershed (ECW) [Figure 4.2] [2] [24]. Located in central Indiana, USA , this watershed is one of the water supplies of the city of Indianapolis. It is mainly composed of agricultural lands used for growing corn and soybean [26]. The major concerns of this watershed and similar watersheds located in the Midwest region are related to the water quality and the frequent flood episodes [Figure 4.4] in the last few years [26]. Now, multiple efforts have been put into increasing the implementation of best management practices and numerous stakeholders are interested in studying conservation plan practices and their impact. The Eagle Creek Watershed, is divided into 130 [Figure 4.3] sub-basins. Out of the total number of sub-basins, ECW contains 108 rural sub-basins where conservation plans can be applied and 22 remaining urban sub-basins that are not considered for conservation practices.



Figure 4.2: Eagle Creek outline [1].

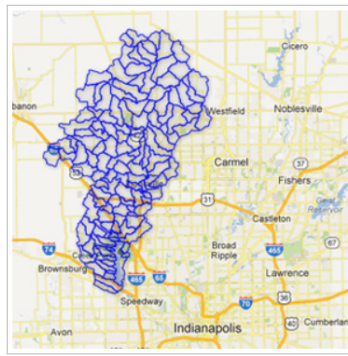


Figure 4.3: Eagle Creek sub-basins [1].

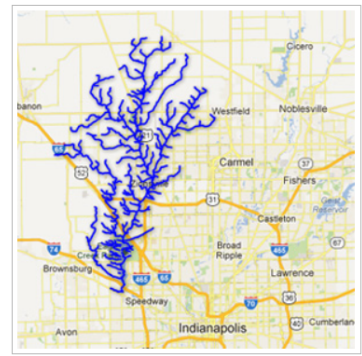


Figure 4.4: Eagle Creek stream [1].

4.2 Users Involved

A previous study [26] recorded the WESTORE interaction of 20 participants. Out of these 20 participants and for data consistency reasons we used the data of 7 surrogate volunteers (four females and three males) from both Indiana University and Oregon State University. These students have been trained to act like sub-basins stakeholders and thus constitute a representative sample of the desired population. Moreover, each user has been assigned a different sub-basins group from the seven sub-basins groups in Figure 4.5. Six out of the seven sub-basins groups were a randomly chosen set of neighboring sub-basins that represent different local regions of the watershed [26]. The seventh group, on the other hand, included the entire watershed. Only after being assigned their sub-basins of interest, do the users start evaluating watershed conservation plan based on both quantitative goals and subjective preferences relative to their sub-basins of interest.

4.3 Data Acquisition Process

The users from whom we are acquiring the data for our study belong to the Modal A Surrogate users. Modal A represents the original calibrated SWAT (Soil and Water Assessment Tool) model [21] used for the Eagle Creek Watershed [24]. After being familiarized with the WESTORE tool, the participants proceed to the experiment. Once the user starts a new experiment with specific conservation practices and watershed goals, each one of the conservation plan related variables

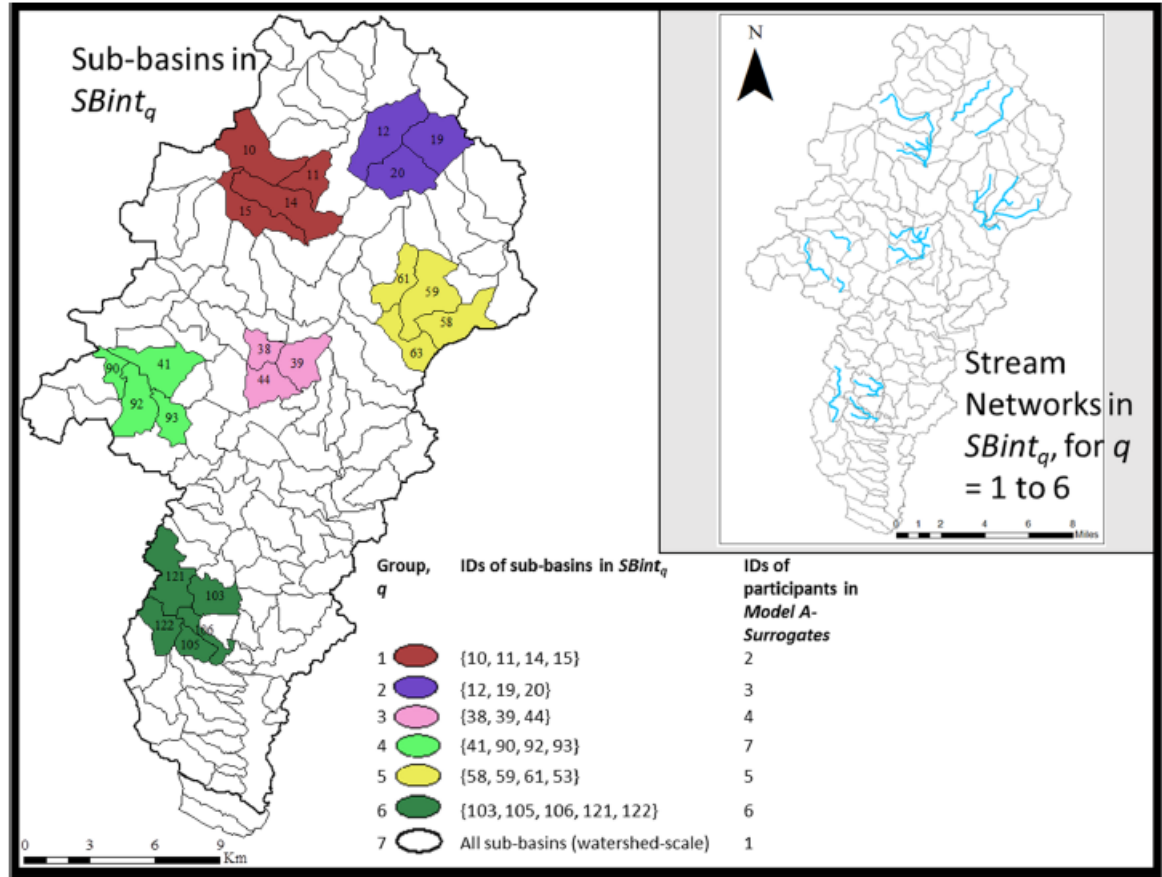


Figure 4.5: Map of spacial distribution of sub-basins of interest (SBint) in the watershed. Model A Surrogate represent the 7 surrogate users in this study and the top right map represents the water stream in the 6 local sub-basins of interest [26].

would correspond to a decision variable for each and every sub-basin and each watershed goal variable would be mapped to an objective function for each and every sub-basin [26].

In this particular study, each sub-basin's decision space is composed of two variables relatives to the chosen conservation plan practices:

- Cover crops: Represented by a binary value corresponding to either the presence or the absence of cover crops in every eligible sub-basin. Value 1 encodes a positive decision toward cover crop implementation and value 0 represents a negative decision toward the implementation of cover crops.
- Filter strips: Represented by a real number that corresponds to the width of the filter strip applied on each eligible sub-basin.

As for the objective space resulting from each sub-basin's conservation plan, the WESTORE tool optimizes a total of four performance variables:

- Peak flow reduction: Reduction of the maximum water flow resulting from the conservation plan applied to each eligible sub-basin.
- Cost reduction: Relative to reducing the cost of the decision implementation in each eligible sub-basins.
- Nitrate reduction: Corresponds to the resulting nitrate reduction in each eligible sub-basin after implementing a particular conservation plan.
- Sediment reduction: Corresponds to the resulting sediment reduction in each eligible sub-basin after implementing a particular conservation plan.

In each experiment, the user goes through interactive sessions. In each session the participant gets a set of twenty design alternatives with different decision and objective values. The participant would then rate the design on a scale of one to three ("I don't like it", "Neutral", "I like it") [26].

There are two different types of sessions for the user to interact with [26]:

- Introspection sessions: These sessions originate either from a previous non-interactive search or from the most highly rated designs resulting from the previously completed human guided search sessions.
- Human guided search sessions: These sessions are the result of an interactive genetic algorithm [4] where previous users ratings are used as one of the objective functions to generate the next population of designs.

The data we are using is retrieved from two sets of introspection session and two sets of human guided search sessions [Figure 4.6].

The first introspection session is the same for all the users. In fact in this session the users interact with the same twenty initial designs. However, the second introspection session depend on the most highly rated designs by each user in the prior human guided search sessions.

Once the experiment is over, the WESTORE tool saves all the raw data relative to the highly rated designs by each user in all the sessions. The data collected during this experiment, is what we are going to use in this thesis in order to demonstrate the results of our novel visualization methodology.



Figure 4.6: The sequence of the different sessions. HS stand for Human guided search session[26].

Chapter 5: Proposed Visualization Technique

5.1 Visualization Metrics

The metrics used to quantify the decision spaces alternatives depend on the type of the decision variables:

- Binary variables are quantified by the probability of implementation of a particular decision in each sub-basin (e.g cover crop attribute).

$$Prob_{s,R_i,k} = \frac{\sum_{l=1}^{l_{ik}} ccimp_{s,R_i,l,k}}{L_{R_i,k}} \quad (5.1)$$

where:

- $ccimp_{s,R_i,l,k}$ [25] is the binary decision in the sub-basin s of the design l . This decision is relative to user k that attributed a rating R_i to the design l .
- $L_{R_i,k}$ [25] represent the total number of designs that got assigned the rating R_i by user k
- Real variables (e.g filter strip attribute) are quantified by calculating the mode. The mode is the most repeated value in a sub-basin. This method

has been selected among other methods in order to identify the most liked or disliked values relative to a particular practice.

$$Mode_{s,R_i,k} = \operatorname{argmax}(\operatorname{count}(FSW_{s,R_i,l,k})) \quad (5.2)$$

where:

- $FSW_{s,R_i,l,k}$ [25] is a unique value of the filter strip width to be implemented in the sub-basin s of the design l . Moreover R_i represents the rating that user k assigned to the design alternative l .

The metric used to quantify the objective space variables is the average reduction:

$$Avr Reduc_{s,R_i,k} = \frac{\sum_{l=1}^{l_{ik}} Reduc_{s,R_i,l,k}}{L_{R_i,k}} \quad (5.3)$$

where $Reduc_{s,R_i,l,k}$ [25] is the resulting reduction value in the sub-basin s of the l^{th} design. In addition, R_i represent the rating given by the user k to the design alternative l .

5.2 Proposed Multi-Variables Network Map Visualization

Additionally to the aesthetic criteria presented in [Section 3.3], our visualization method demonstrated for the watershed conservation plan is characterized by fixed node positions. In fact, the position of the nodes corresponds to the map placement

of the sub-basins thus keeping the additional referrer relative to the geographical information present.

Knowing that commonly used watershed visualizations fail in representing all the conservation plan variables, our goal with the proposed network map visualization is to simultaneously represent both the decision space and the objective space without losing the geographical information provided by a map.

In this watershed conservation plan network structure, each node represents a sub-basin and each edge represent a water stream connecting a pair of sub-basins.

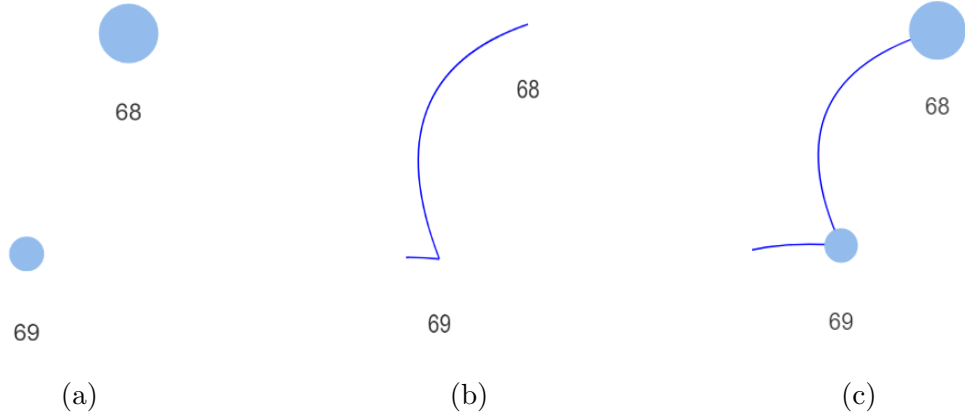


Figure 5.1: Nodes and edge example: (a) a node representation of two sub-basins: sub-basin 68 and sub-basin 69, (b) an example of a water stream edge representation connecting sub-basin 68 to sub-basins 69 and (c) the combination of the edge and the two nodes.

In order to describe the user’s decision in each sub-basin, we superpose two nodes with different colors [19] and opacity:

- A blue opaque node represents the cover crop decision and its size is based

on the probability of cover crop implementation $Prob_{s,R_i,k}$ [Eq 5.1] [25]

- A green transparent node represents the filter strips decision and its diameter is based on the normalized mode $normalized(Mode_{s,R_i,k})$ of all the design alternatives in a particular sub-basin that have been assigned the same R_i rating from participant k . For each one of the 108 implementable sub-basins, the mode $Mode_{g,R_i,k}$ [Eq 5.2] [25] is computed. The normalized mode is obtained by using the following scaling equation:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (5.4)$$

Where x_{min} and x_{max} are respectively the minimal and maximal $Mode_{s,R_i,k}$ from the set of computed decision modes relative to user k for all the implementable sub-basins with the same rating R_i .

The resulting objective space is represented by four edges with distinct colors [19], one for each objective variable:

- A blue edge represents the peak flow reduction.
- A red edge represents the cost reduction.
- A black edge represents the sediment reduction.
- An orange edges represents the nitrate reduction.

The color choice is purely based on maximizing the contrast [19] between the edges. Also, the curvature between the edges is a design decision chosen in order to conform to an aesthetic criteria that has been proven to improve readability [17].

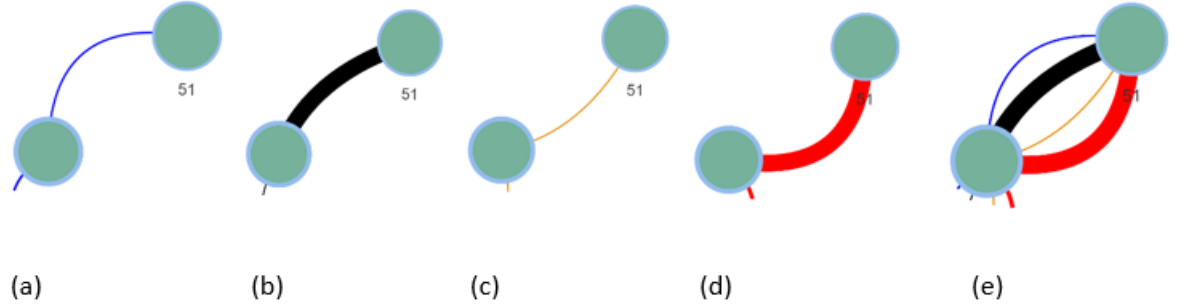
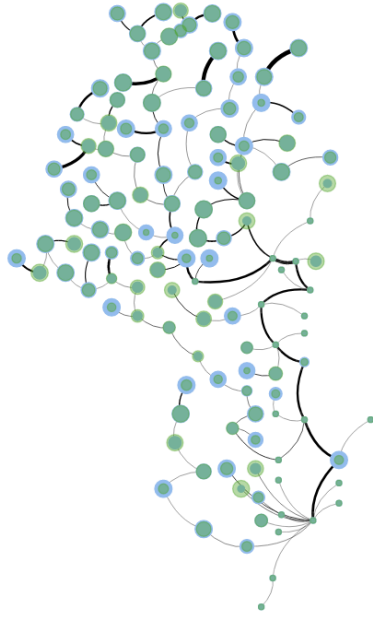


Figure 5.3: An example of the possible edge visualizations: (a) peak flow reduction, (b) sediment reduction (c) nitrate reduction, (d) cost reduction and (e) simultaneous visualization of all the objective space variables.

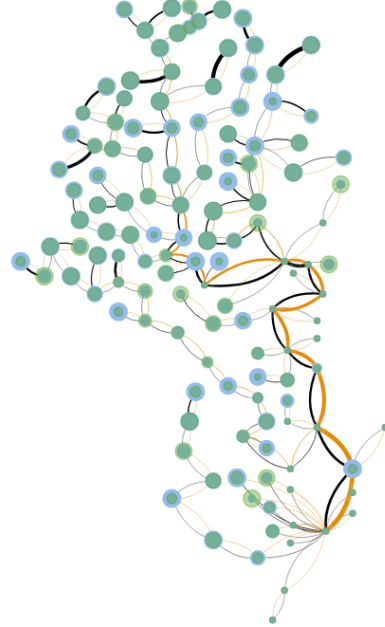
For each one of the objective space variables the width of the corresponding edge is based on the normalized average of all the design alternatives reduction values that have been assigned the same rating R_i from the same user k . The average reduction value $AvrReduc_{s,R_i,k}$ [25] of each one of the resulting objective function (peak flow, cost, sediment and nitrate) is based on [Eq 5.3].

The normalization of an average reduction value at a particular sub-basin is computed using [Eq 5.4].

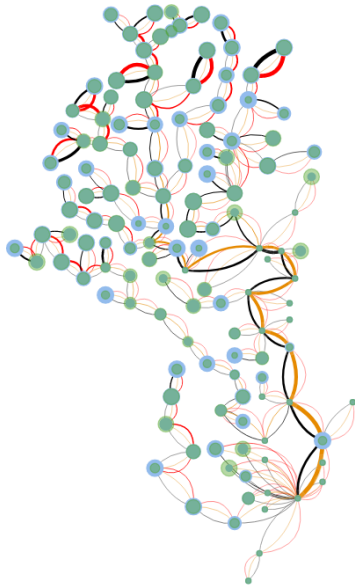
Our proposed network visualization covers the totality of the variables from both the decision and objective space without losing the geographical proximity



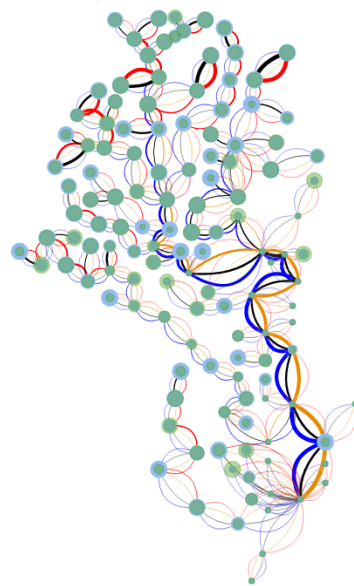
(a) Two decision variables and one objective variable



(b) Two decision variables and two objective variables



(c) Two decision variables and three objective variables



(d) Two decision variables and four objective variables

Figure 5.2: An example of visualizing all the decision space and x number of objective space variables. x being in the range of $[1..4]$.

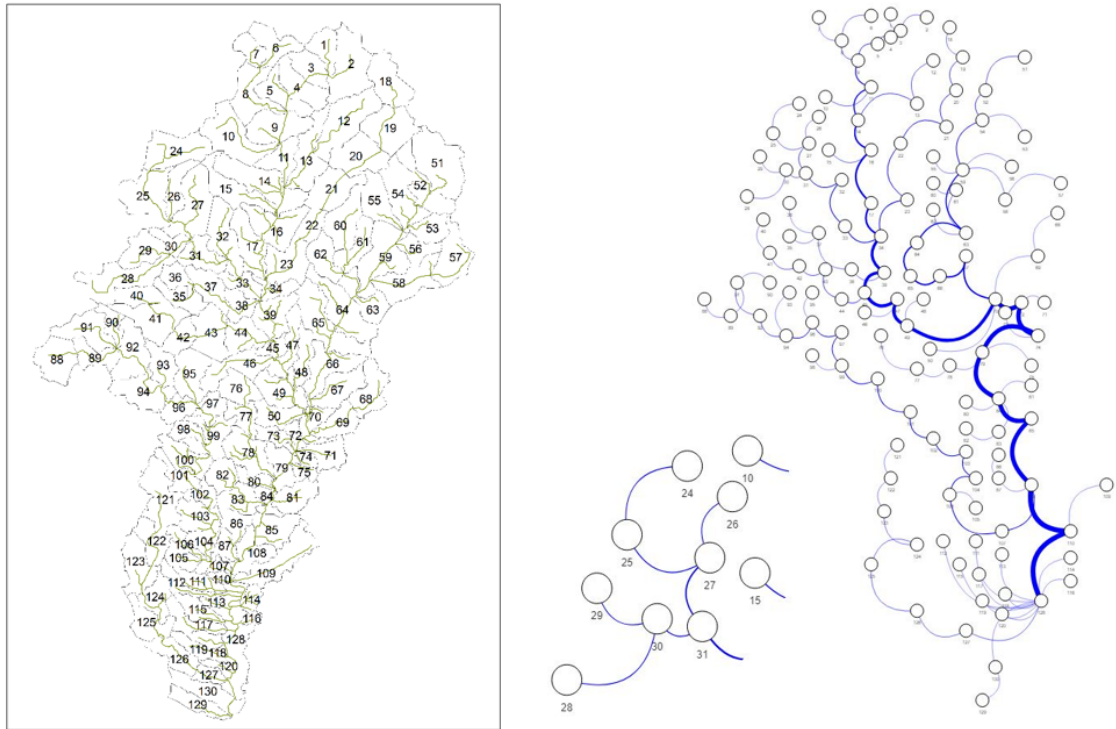


Figure 5.4: A side by side comparison between the watershed map and the network representation before adding the decision space and objective space variables. The numbers that are underneath each sub-basins represent the sub-basins labels.

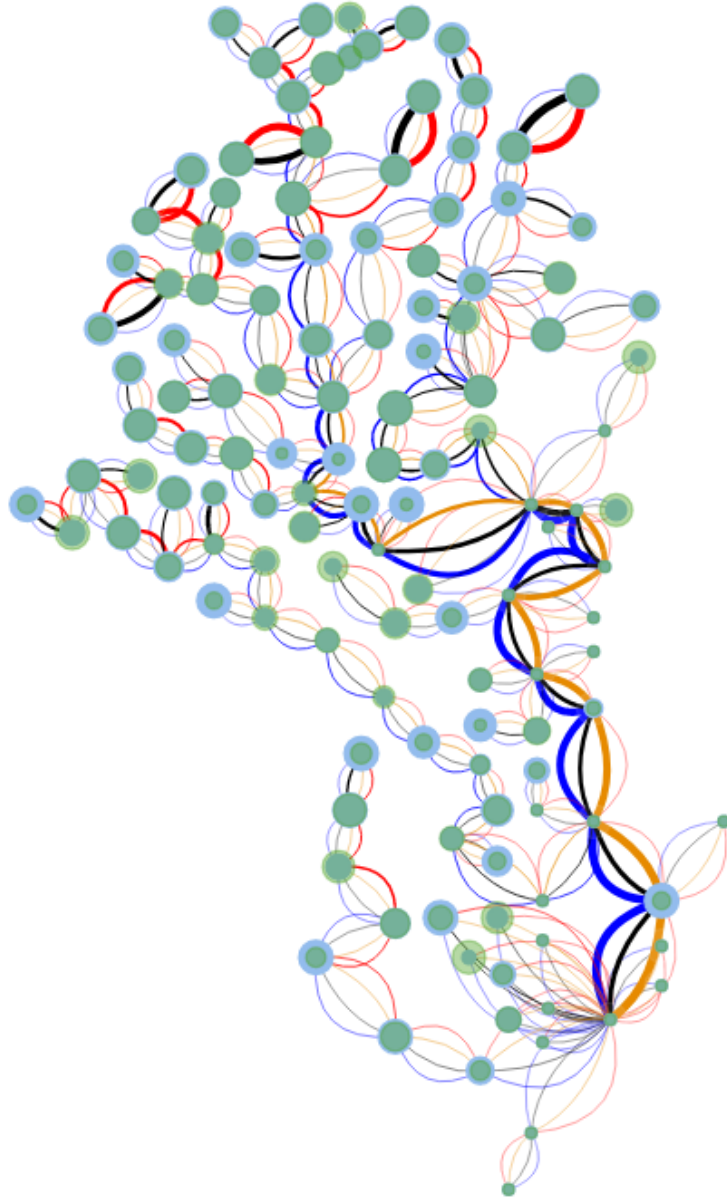
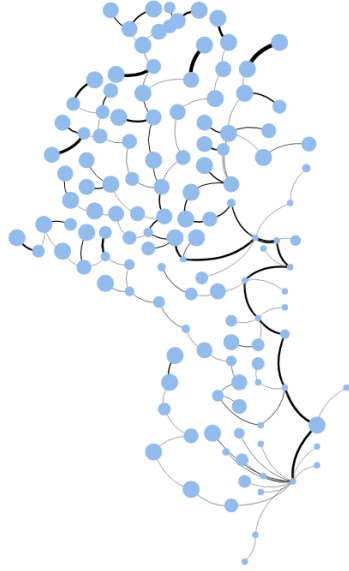
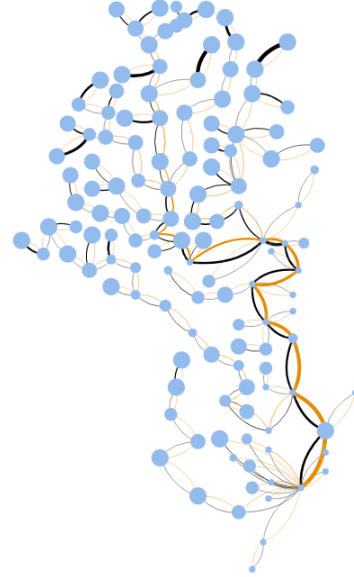


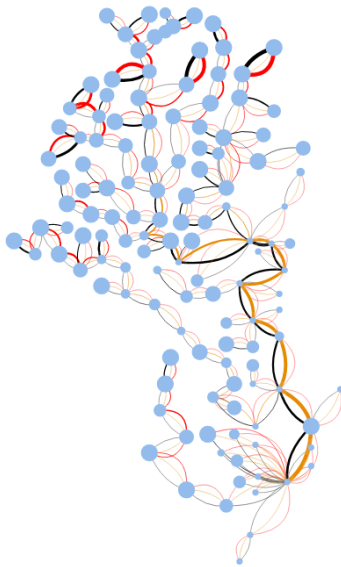
Figure 5.5: An example of the simultaneous representation of the decision and objective space in the ECW.



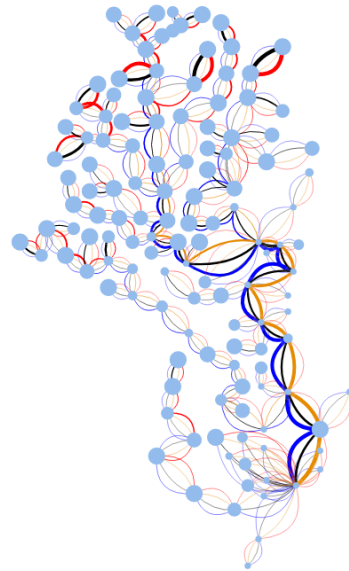
(a) One decision variable and one objective variable.



(b) One decision variable and two objective variables.



(c) One decision variable and three objective variables.



(d) One decision variable and four objective variables.

Figure 5.6: An example of visualizing one decision space variable (cover crops) and x number of objective space variables. x being in the range of $[1..4]$.

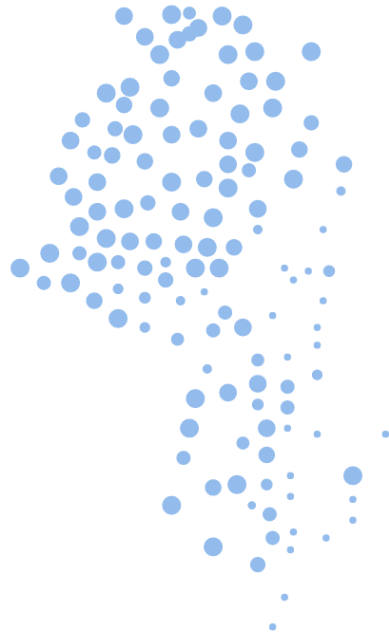
between the sub-basins. In fact, the positioning of the nodes is consistent with the actual positions of the sub-basins on the map which minimizes the overlap and adds a spatial context to the visualization [Figure 5.4][Figure 5.5]. On a broader sense, the simultaneous combination of geographical information, decision space and objective space [Figure 5.5] yields to the discovery and identification of existing patterns presented in more details in the result chapter.

In addition, the proposed decision and objective space network visualization implementation provides various levels of the visualization of the watershed as shown in [Figure 5.2][Figure 5.6] [Figure 5.7][Figure 5.8]. This visualization offers different granularity levels. The visualization can be customized from a global "big picture" visualization [Figure 5.5] of all the decision and objective space variables to a more targeted version focusing on one or more chosen variables. Some of the possible combinations are shown in [Figure 5.2][Figure 5.6] [Figure 5.7][Figure 5.8] but the user has the possibility to choose any other combinations.

5.3 Analysis Methods

The strength of the relationship between the different variables is evaluated using the correlation matrix [11]. The correlation metric is a way of analyzing how different variables are related to each other.

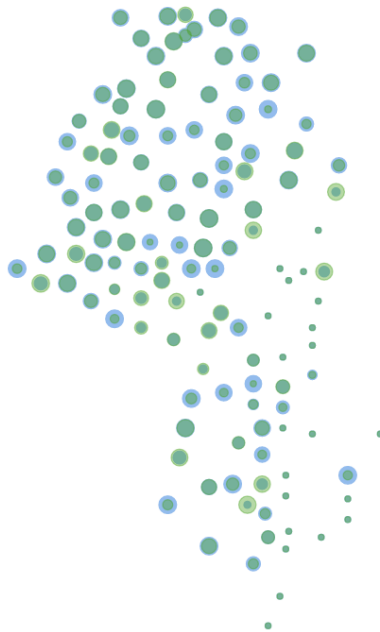
In order to compute the correlation matrix between m different variables, we first evaluate the variance - covariance matrix $\Sigma_{m \times m}$ where the covariance between



(a) Cover Crops.

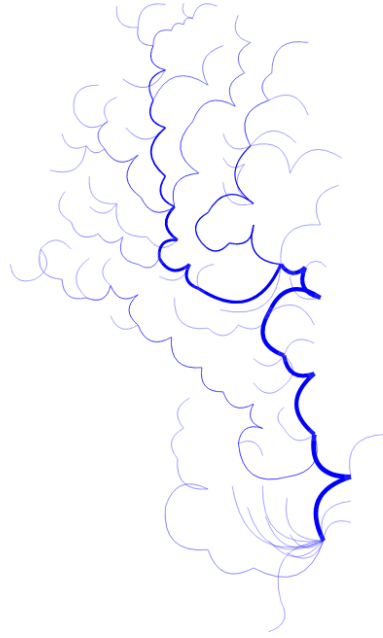


(b) Filter Strips.

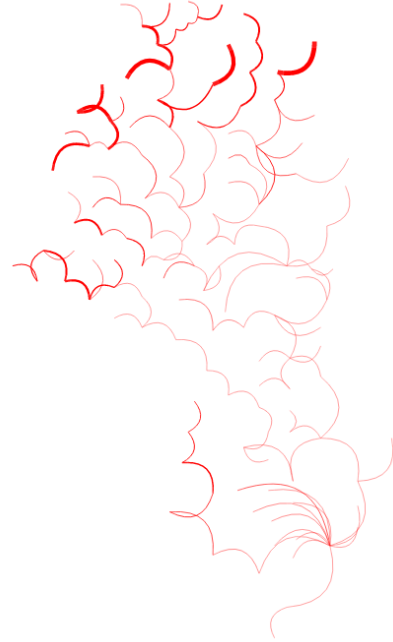


(c) Cover Crops and Filter Strips.

Figure 5.7: An example of all the different levels of decision space visualization.



(a) Peak Flow Reduction.



(b) Cost Reduction.



(c) Sediment Reduction.



(d) Nitrate Reduction.

Figure 5.8: An example of single variable decision space visualizations.

a variable x and a variable y is defined by [Eq 5.5]:

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (5.5)$$

where :

- x_i is the observed variable relative to a particular sub-basin.
- \bar{x} is the average of the observed variable for all the sub-basins.
- n is the total number of sub-basins.

The correlation matrix is then defined using [Eq 5.6]:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (5.6)$$

Where σ_{xy} is the covariance between a variable x and a variable y and σ_x and σ_y represent respectively the standard deviation of variable x and variable y .

Chapter 6: Results and Discussions

In this section we present the visual and analytical results for our novel coincident nodes multi-edge network map visualization demonstrated for the watershed conservation plan practices. The analysis part in general and the correlation analysis in particular is used as an established statistical method [27] to validate the visual patterns observed.

6.1 Patterns Prior to Human Guided Search

The "I like it design" user preferences relative to the introspection 1 session are summarized in [Figure 6.1] [Figure 6.2].

Simultaneously visualizing the decision and objective space, enables us to explore the quantitative and qualitative data in its entirety and thus uncover useful insight about the data set.

From a visual perspective [Figure 6.1] [Figure 6.2], we notice some similarities in the decision space preferences and almost an identical preference in terms of the objective space variables at the watershed level. Moreover, we notice an interesting salient pattern in the main watershed stream where the reductions in peak flow, sediment and nitrate reach their maximum. In addition, we notice another repeated pattern in the upper region of the network map where the cost and

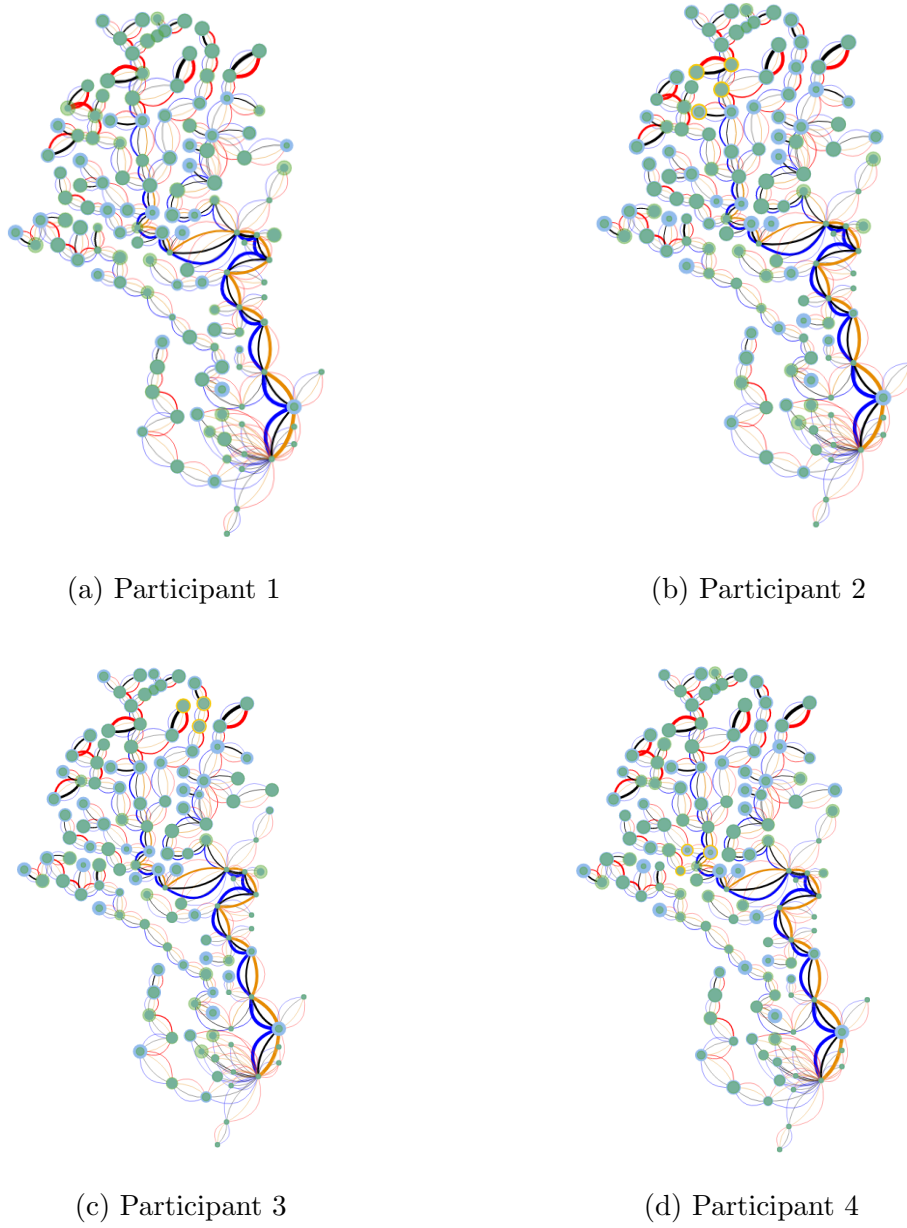
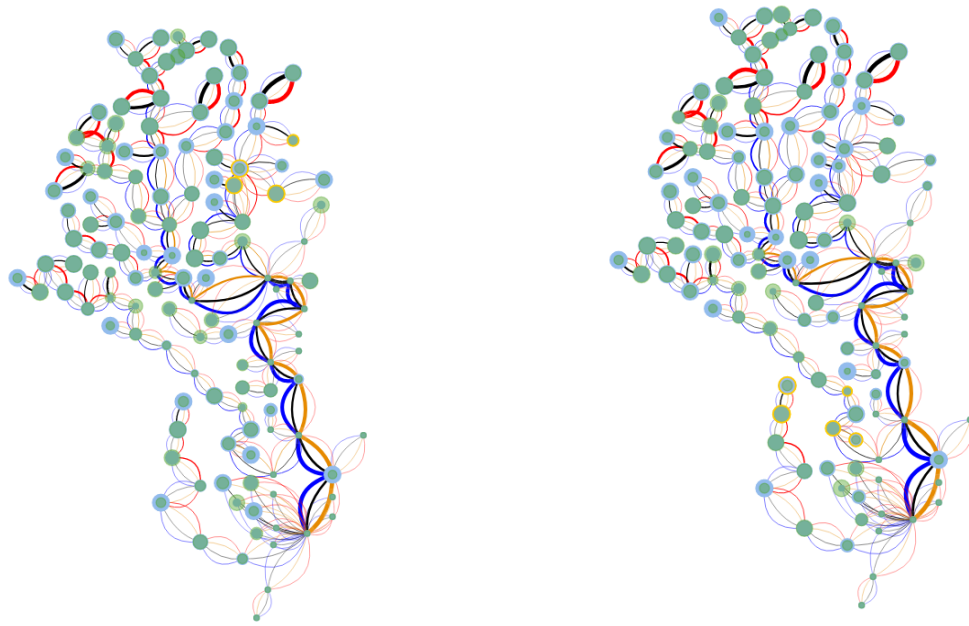
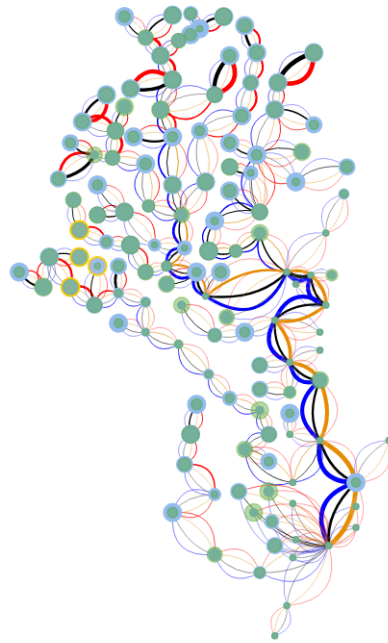


Figure 6.1: Decision and objective space coincident nodes multi-edge network map visualization of participant 1 to 4 in Modal A surrogate for the "I like it design" rating relative to a non-Interactive Genetic Algorithm based session.



Participant 5

(a) Participant 6



(b) Participant 7

Figure 6.2: Decision and objective space coincident nodes multi-edge network map visualization of participant 5 to 7 in Modal A surrogate for the "I like it design" rating relative to a non-Interactive Genetic Algorithm based session.

sediment reduction reach their maximum.

The main watershed stream is representative of the sub-basins where the peak flow reduction [Eq 5.3] is greater or equal to 50%. The yellow boundary [Figure 6.1] [Figure 6.2] around certain sub-basins represents the sub-basins of interest of each particular participant.

Solely from a visual assessment [Figure 6.1] [Figure 6.2], we can already see how, even though, the decision space presents some differences between users in the whole watershed scale, the objective space holds the same patterns and presents extremely similar objective variable values [Figure 6.3].

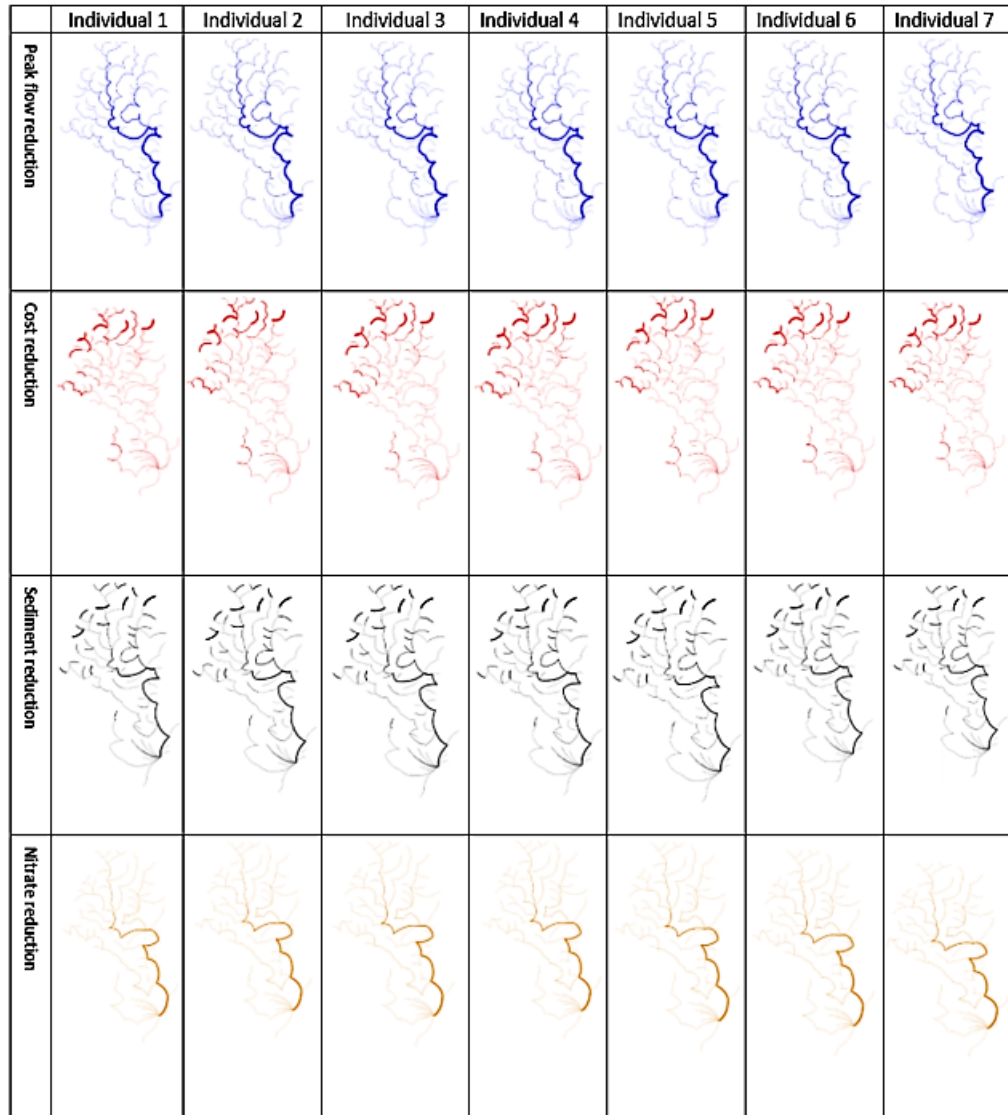


Figure 6.3: The objective space skeletons of all the individuals of group A surrogate taking into consideration only the "I like it design" rating generated by the non-Interactive Genetic Algorithm.

These visual results enable us to make a preliminary hypothesis about the po-

tential agreement between the individuals of group A surrogate for the "I like it design" rating using a non-Interactive Genetic Algorithm. This potential agreement is going to be further emphasized in the remaining of this section.

6.1.1 Variables Correlation Patterns

The correlation metric is known to be the natural choice when measuring similarity [22]. In fact it gives us an idea about the covariation in a particular data set. Using the correlation metric [Eq 5.6] at the whole watershed level between all the decision and objective space variables of each participant one at a time, we notice a low positive correlation of 0.55 between cover crops and filter strips for all individuals. Moreover, Peak flow and nitrate reduction appear to be very highly correlated. This high correlation of 0.972 between peak flow and nitrate reduction implies that the increase of one variables is followed by the increase of the other variable. The later observation is noticeable visually in [Figure 5.8d][Figure 5.8a]. In fact, the nitrate and peak flow skeleton follow the same visual patterns, having minimal values at the whole watershed level except in the main stream area where both reductions reach their maximum. In addition, even though, the peak flow and sediment reduction seem to be highly correlated in the main watershed stream [Figure 5.8a] [Figure 5.8c], they don't present a high positive correlation on the whole watershed level. This analytical finding is further emphasized visually by the fact that contrary to the peak flow reduction, the sediment reduction reaches also its maximum values in the upper region of the map [Figure 5.8c].

At the main watershed stream level, the correlation analysis for all the participant one at a time, confirms the visual hypothesis [Figure 6.1] [Figure 6.2]. In fact, the peak flow, sediment and nitrate reduction [Figure 5.8a] [Figure 5.8c] [Figure 5.8d] are very highly positively correlated with a value of 0.91 for all individuals. This pattern is justifiable by the fact that the water flow is usually proportional to the presence of sediment and nitrate. Moreover, we notice a slight increase in the correlation between the cover crops and filter strips reaching at its highest 0.71 for individual 2 and at its lowest 0.57 for individual 7.

6.1.2 User Preferences Correlation Patterns

Objective Space: The correlation analysis between all the individuals and considering the whole watershed, emphasizes the visual agreement hypothesis relative to the conservation plan goals between participants [Figures 6.3]. In fact, these high correlation values fluctuate between a minimum of 0.989 and a maximum of 1 for the objective space.

Decision Space: The correlation analysis between participants at the level of the whole watershed, shows that the cover crop correlation values are in the range [0.856,0.984] which indicates that when the cover crop implementation increases for a particular individual, it follows that same trend for the other individuals.

Moreover, the filter strips is the only variable where the inter-participant correlation reaches a minimum as low as 0.678. This value is relative to individual 7,

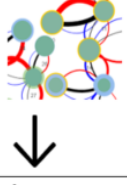
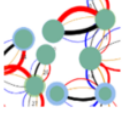
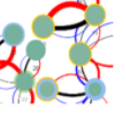
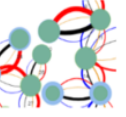
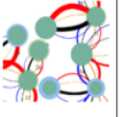
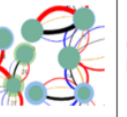
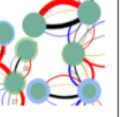
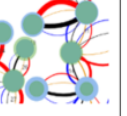
Average Separation in Decision Spaces of Individual 2 and i^{th} Individuals	Individual 1: sub basins {10, 11, 14, 15}	Individual 2: sub basins {10, 11, 14, 15}	Individual 3: sub basins {10, 11, 14, 15}	Individual 4: sub basins {10, 11, 14, 15}	Individual 5: sub basins {10, 11, 14, 15}	Individual 6: sub basins {10, 11, 14, 15}	Individual 7: sub basins {10, 11, 14, 15}
							
Average Separation in cover crops decisions	0.004545	0	0.024621	0.017045	0.045455	0.087121	0.087121
Average Separation in filter strips decisions	0	0	0.001249	0	0	0	0.050228

Figure 6.4: Average separation in the decision space between individual 2 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

who seems to show the least amount of positive correlation with the other users in terms of filter strip width. On the other hand, the set of filter strip inter-individuals correlation reaches a maximum of 0.961 with the majority of the values ranging in the interval $[0.813, 0.898]$.

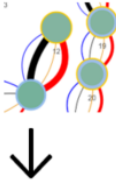
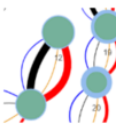
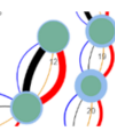
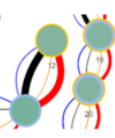
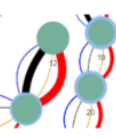
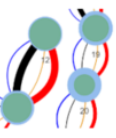
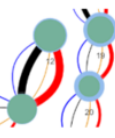
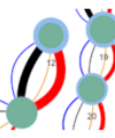
Average Separation in Decision Spaces of Individual 3 and i^{th} Individual	Individual 1: sub basins {12,19,20}	Individual 2: sub basins {12,19,20}	Individual 3: sub basins {12,19,20}	Individual 4: sub basins {12,19,20}	Individual 5: sub basins {12,19,20}	Individual 6: sub basins {12,19,20}	Individual 7: sub basins {12,19,20}
							
Average Separation in cover crops decisions	0.033333	0	0	0	0	0.055556	0.055556
Average Separation in filter strips decisions	0.08187	0.08187	0	0.044156	0.08187	0.08187	0.132905

Figure 6.5: Average separation in the decision space between individual 3 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

Whether it is from a visual or analytical perspective, we notice some differences in the decision preferences at the level of the whole watershed scale. Moreover, in order for a conservation plan to be successful, we need to be able to detect and take into consideration the potential disagreement between stakeholders in their respective sub-basins of interest. Since each participant has only a particular group of sub-basins of interest, we choose to further investigate the average separation between the decision space variables relative to individual j sub-basins of interest and the values of the decision space variables that the rest of the individuals have

chosen for the sub-basins of interest of individual j .

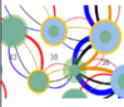
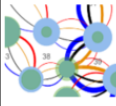
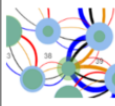
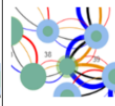
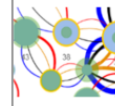
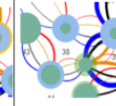
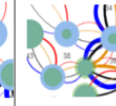
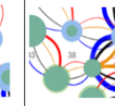
Average Separation in Decision Spaces of Individual 4 and i^{th} Individual	Individual 1: sub basins {38,39,44}	Individual 2: sub basins {38,39,44}	Individual 3: sub basins {38,39,44}	Individual 4: sub basins {38,39,44}	Individual 5: sub basins {38,39,44}	Individual 6: sub basins {38,39,44}	Individual 7: sub basins {38,39,44}
 ↓							
Average Separation in cover crops decisions	0.116667	0.143939	0.222222	0	0.138889	0.194444	0.25
Average Separation in filter strips decisions	0	0	0.12775	0	0	0	0.127314

Figure 6.6: Average separation in the decision space between individual 4 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

[Figure 6.4 to 6.9] represents the separation in decision space preferences between the values chosen by a particular individual j for his sub-basins of interest and the values chosen by the rest of the participants for the same group of sub-basins.

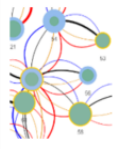
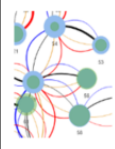
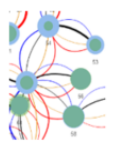
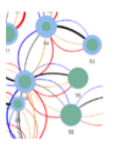
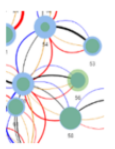
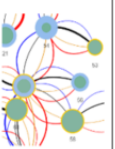
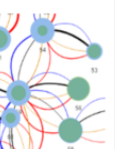
Average Separation in Decision Spaces of Individual 5 and i^{th} Individual	Individual 1: sub basins {53, 58, 59, 61}	Individual 2: sub basins {53, 58, 59, 61}	Individual 3: sub basins {53, 58, 59, 61}	Individual 4: sub basins {53, 58, 59, 61}	Individual 5: sub basins {53, 58, 59, 61}	Individual 6: sub basins {53, 58, 59, 61}	Individual 7: sub basins {53, 58, 59, 61}
							
Average Separation in cover crops decisions	0.108333	0.083333	0.166667	0.083333	0	0.041667	0.125
Average Separation in filter strips decisions	0.033624	0.033624	0.25623	0.114886	0	0.250098	0.444396

Figure 6.7: Average separation in the decision space between individual 5 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

In this analysis, we omit the separation relative to individual 1 sub-basins of interest since it consists of the whole watershed.

The least amount of filter strip separation is relative to individual 2 sub-basins of interest [Figure 6.4] with four out of six separation values being zero and with an overall separation percentage of 0.857%. As for cover crops, the highest amount of agreement is found in the sub-basin group of individual 3 [Figure 6.5]. In fact, the values of cover crop separation are in the range $[0, 0.056]$ with the average overall percentage being equal to 2.41%.

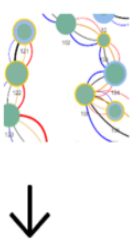
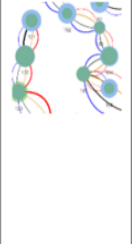
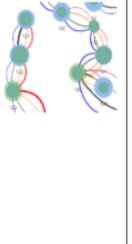
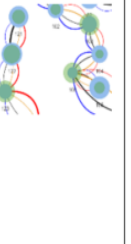

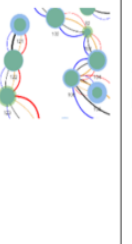
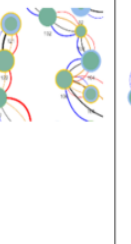
Average Separation in Decision Spaces of Individual 6 and i^{th} Individual	Individual 1: sub basins {103, 105, 106, 121, 122}	Individual 2: sub basins {103, 105, 106, 121, 122}	Individual 3: sub basins {103, 105, 106, 121, 122}	Individual 4: sub basins {103, 105, 106, 121, 122}	Individual 5: sub basins {103, 105, 106, 121, 122}	Individual 6: sub basins {103, 105, 106, 121, 122}	Individual 7: sub basins {103, 105, 106, 121, 122}
							
Average Separation in cover crops decisions	0.073333	0.133333	0.216667	0.15	0.166667	0	0.1
Average Separation in filter strips decisions	0.04409	0.060005	0.188601	0.05657	0.04409	0	0.138004

Figure 6.8: Average separation in the decision space between individual 6 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

In addition, considering all the separation values [Figure 6.4 to 6.9], we notice that the minimum separation value is 0% for both the variables of the decision space. On the other hand, the maximum separation value reaches 25% for the cover crop decision variable and 44% for the filter strip decision space variable. The average separation analysis, shows that the the cover crop separation average is 9.23% and the filter strip separation average reaches 8.53%.

The separation analysis centered around the sub-basins of interests of each individual, proves that the differences in the subjective decision space preference using

a non interactive genetic algorithm session is less than 10% and thus emphasizes that the stakeholders are rather converging toward an agreement.

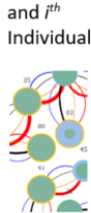
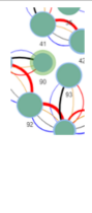


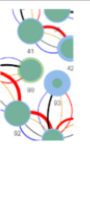
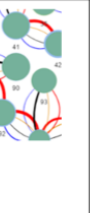
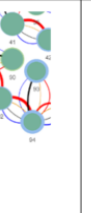

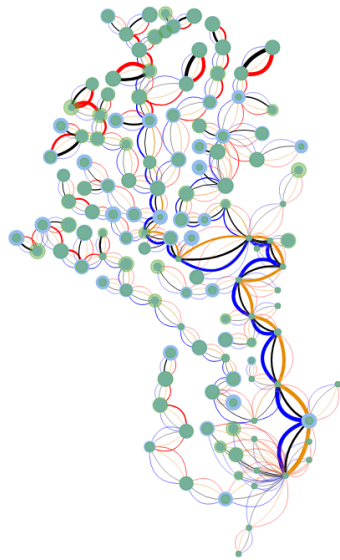
Average Separation in Decision Spaces of Individual 7 and i^{th} Individual	Individual 1: sub basins {41, 90, 92, 93}	Individual 2: sub basins {41, 90, 92, 93}	Individual 3: sub basins {41, 90, 92, 93}	Individual 4: sub basins {41, 90, 92, 93}	Individual 5: sub basins {41, 90, 92, 93}	Individual 6: sub basins {41, 90, 92, 93}	Individual 7: sub basins {41, 90, 92, 93}
							
Average Separation in cover crops decisions	0.1	0.090909	0.0625	0.0625	0.041667	0.041667	0
Average Separation in filter strips decisions	0.202906	0	0.008423	0	0.207346	0.175801	0

Figure 6.9: Average separation in the decision space between individual 7 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

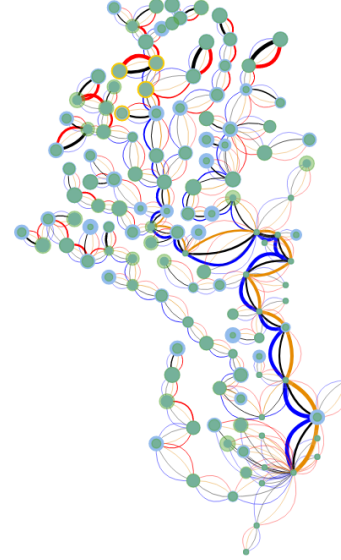
Analytically, the population studied has a relatively low average separation for the decision space variables and a very high positive correlation for the objective space variables. This highlights the general visual agreement between the participants and also emphasizes the noticeable salient similarities in the objective space.

Both the visual and analytical findings converge toward the same results. Furthermore, the inter-participants decision space similarities and the conformity ob-

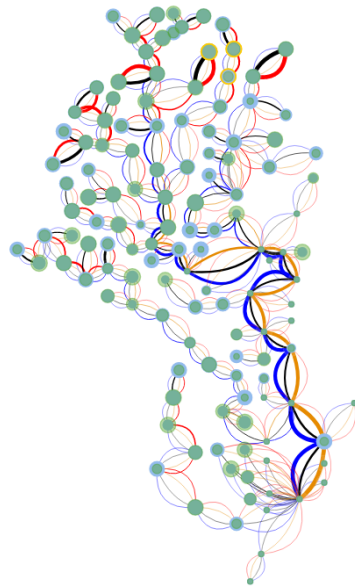
jective space wise, gives important insight to the decision makers and thus greatly increases the likelihood of a successfully chosen conservation plan.



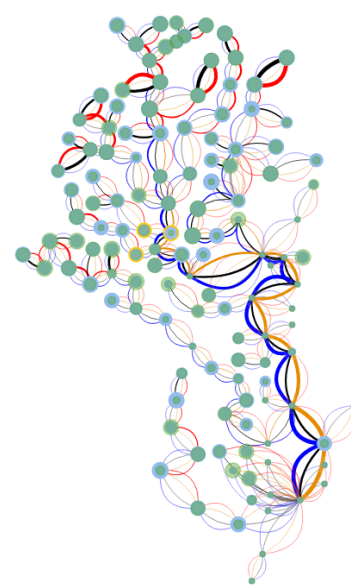
(a) Participant 1



(b) Participant 2

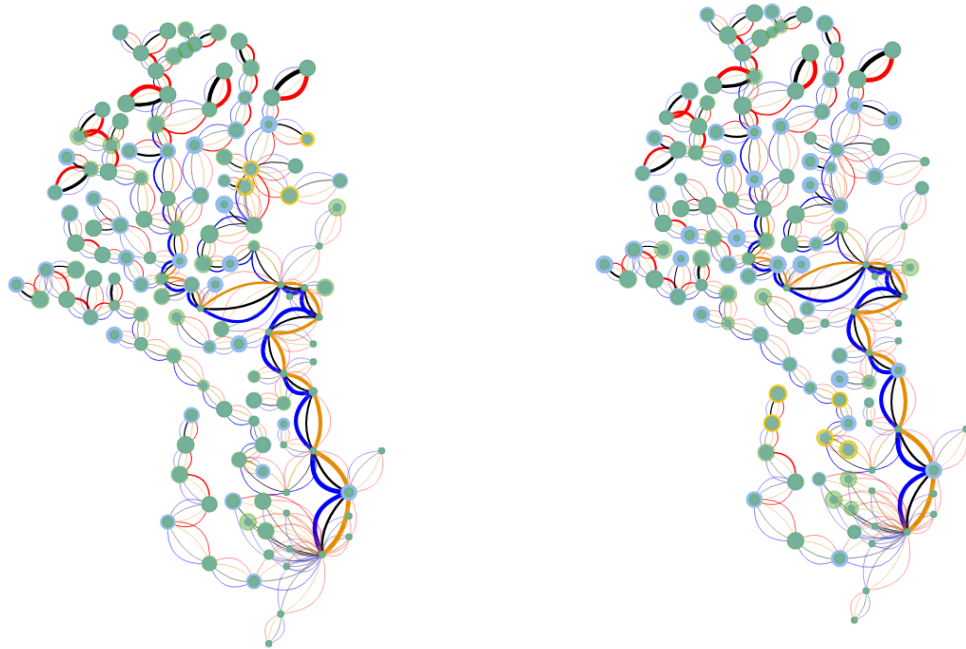


(c) Participant 3



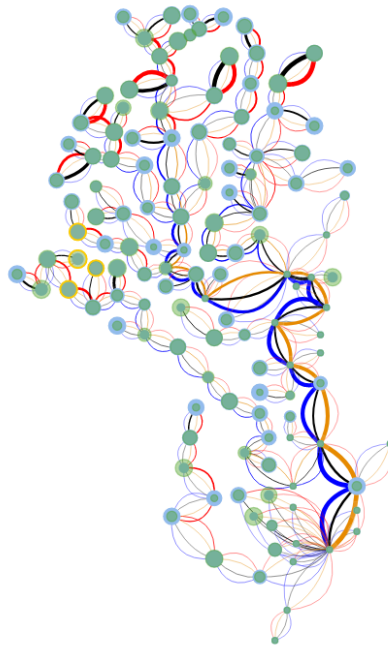
(d) Participant 4

Figure 6.10: Decision and objective space coincident nodes multi-edge network map visualization of participant 1 to 4 in Modal A surrogate for the "I like it design" rating relative to Interactive Genetic Algorithm sessions.



Participant 5

(a) Participant 6



(b) Participant 7

Figure 6.11: Decision and objective space coincident nodes multi-edge network map visualization of participant 5 to 7 in Modal A surrogate for the "I like it design" rating relative to Interactive Genetic Algorithm Sessions.

6.2 Pattern Post Human Guided Search

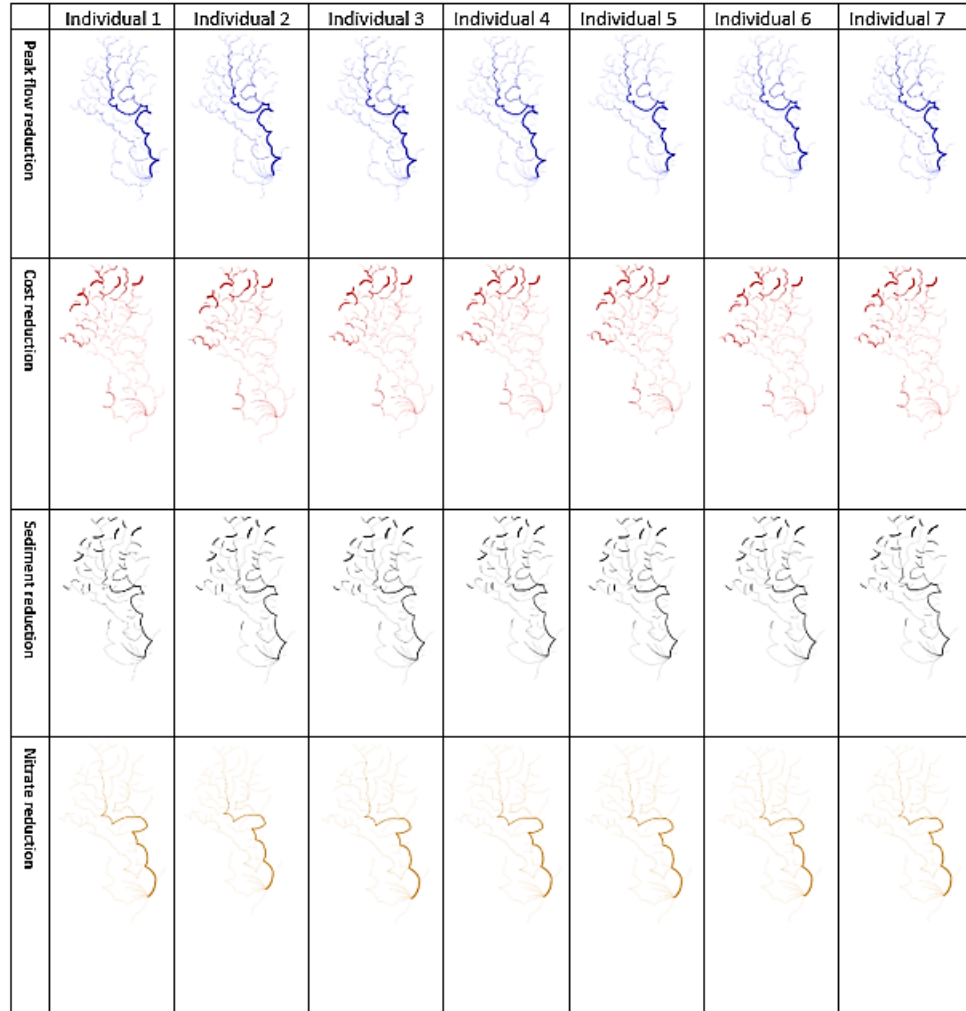


Figure 6.12: The objective space skeleton of all the individuals of group A surrogate taking into consideration the "I like it" designs rating generated by the Interactive Genetic Algorithm.

After analyzing the results relative to the non interactive session, we study the results of the sessions resulting from the Interactive Genetic Algorithm. In this

session we study the same sample population of group A surrogate for the same "I like it design" rating. Figure 6.10 and Figure 6.11 show the same patterns that appeared previously in the non-interactive session. In fact, considering the same river channel threshold presented in [section 6.1], we clearly notice the presence of the same patterns relative to the high positive correlation between peak flow, sediment and nitrate. Moreover, looking at the "big picture" with both the decision and objective space variables, we observe how the different users present the same layout with patterns that are resemblant to the introspection 1 patterns.

Looking into the different network map visualizations [Figure 6.10 and Figure 6.11], the same visual insights as the insights relative to [Section 6.1] can be made. Actually, even though there seems to be some difference between the participant in the decision preferences, the objective space [Figure 6.12] is clearly holding the same patterns between different users as well as between the different types of sessions.

6.2.1 Variables Correlation Patterns

Similar to the session based on the non Interaction Genetic Algorithm, the correlation between the cover crop and filter strips is represented by a rather low positive value. The highest positive correlation is the one relative to the peak flow and nitrate reduction.

Going from a non Interactive Genetic Algorithm to an Interactive one, no change in the main river channel is noticed. In fact, the peak flow, sediment and

nitrate reduction are still very highly positively correlated [Figure 6.12].

6.2.2 User Preferences Correlation Patterns



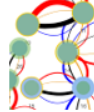

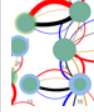


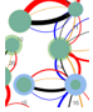
Separation in Decision Spaces of Individual 2 and i^{th} Individual	Individual 1: sub basins {10, 11, 14, 15}	Individual 2: sub basins {10, 11, 14, 15}	Individual 3: sub basins {10, 11, 14, 15}	Individual 4: sub basins {10, 11, 14, 15}	Individual 5: sub basins {10, 11, 14, 15}	Individual 6: sub basins {10, 11, 14, 15}	Individual 7: sub basins {10, 11, 14, 15}
							
Average Separation in cover crops decisions	0.154298	0	0.076644	0.146552	0.065592	0.17029	0.137597
Average Separation in filter strips decisions	0.005373	0	0	0.006651	0.006651	0.001803	0.131888

Figure 6.13: Average separation in the decision space between individual 2 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

Compared to the non Interactive Genetic Algorithm (IGA) based session, we notice a slight decrease in the correlation between users relative to the cover crop decisions that reaches values as low as 0.704 and as high as 0.930.

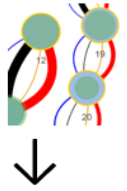







Separation in Decision Spaces of Individual 3 and i^{th} Individual	Individual 1: sub basins {12,19,20}	Individual 2: sub basins {12,19,20}	Individual 3: sub basins {12,19,20}	Individual 4: sub basins {12,19,20}	Individual 5: sub basins {12,19,20}	Individual 6: sub basins {12,19,20}	Individual 7: sub basins {12,19,20}
							
Average Separation in cover crops decisions	0.073976	0.105258	0	0.061155	0.181845	0.088071	0.050686
Average Separation in filter strips decisions	0.283194	0.021268	0	0.112142	0.112142	0.060347	0.016438

Figure 6.14: Average separation in the decision space between individual 3 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

As for the objective space variables, the inter-user correlation is characterized by a very high positive correlation of 0.99 that agrees with the visual conclusion drawn earlier in this section and the conclusions drawn from the non IGA based session. Compared to the previous discussion [Section 6.1], individual 7 still shows the least amount of agreement with the rest of the group when it comes to filter strip width.

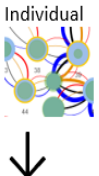
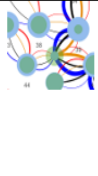
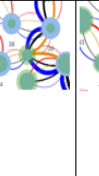
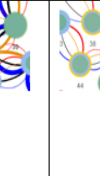
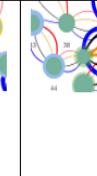


Separation in Decision Spaces of Individual 4 and i^{th} Individual	Individual 1: sub basins {38,39,44}	Individual 2: sub basins {38,39,44}	Individual 3: sub basins {38,39,44}	Individual 4: sub basins {38,39,44}	Individual 5: sub basins {38,39,44}	Individual 6: sub basins {38,39,44}	Individual 7: sub basins {38,39,44}
							
Average Separation in cover crops decisions	0.061303	0.124313	0.306022	0	0.166667	0.217558	0.122427
Average Separation in filter strips decisions	0.000484	0.180289	0.520621	0	0	0.519524	0.220211

Figure 6.15: Average separation in the decision space between individual 4 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

In the upcoming paragraph, we run some separation analysis in the decision space between the sub-basins of interest in order to identify the presence of potential separation values that could be interpreted as disagreement in the decision space. Similarly to the analysis performed in [Section 6.1], [Figure 6.13 to 6.18] represent the average separation in the decision space relative to the sub-basins of interest of each user.








Separation in Decision Spaces of Individual 5 and i^{th} Individual	Individual 1: sub basins {53, 58, 59, 61}	Individual 2: sub basins {53, 58, 59, 61}	Individual 3: sub basins {53, 58, 59, 61}	Individual 4: sub basins {53, 58, 59, 61}	Individual 5: sub basins {53, 58, 59, 61}	Individual 6: sub basins {53, 58, 59, 61}	Individual 7: sub basins {53, 58, 59, 61}
							
Average Separation in cover crops decisions	0.145115	0.153486	0.16943	0.137931	0	0.210832	0.257518
Average Separation in filter strips decisions	0.001107	0.296368	0.257153	0.079876	0	0.416792	0.330246

Figure 6.16: Average separation in the decision space between individual 5 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

Compared to the non IGA based session, there is a general increase in the decision space separation [Figure 6.13 to Figure 6.18] relative to the respective sub-basins of interest of each user. This separation increase is justifiable by the fact that the IGA specifically tailors its designs to each individual according to his or her previous preferences. This IGA characteristic is thus more likely to enhance existing differences between the participants.

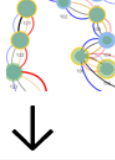
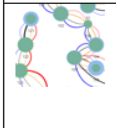
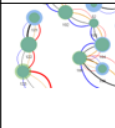
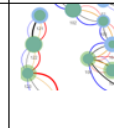
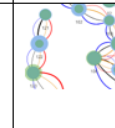
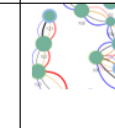
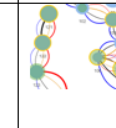
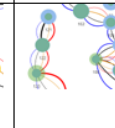

Separation in Decision Spaces of Individual 6 and i^{th} Individual 	Individual 1: sub basins {103, 105, 106, 121, 122}	Individual 2: sub basins {103, 105, 106, 121, 122}	Individual 3: sub basins {103, 105, 106, 121, 122}	Individual 4: sub basins {103, 105, 106, 121, 122}	Individual 5: sub basins {103, 105, 106, 121, 122}	Individual 6: sub basins {103, 105, 106, 121, 122}	Individual 7: sub basins {103, 105, 106, 121, 122}
							
Average Separation in cover crops decisions	0.354523	0.223188	0.082051	0.245927	0.198301	0	0.113111
Average Separation in filter strip decisions	0.377433	0.364464	0.219803	0.434959	0.377671	0	0.35837

Figure 6.17: Average separation in the decision space between individual 6 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

Similar to section 6.1 findings, the lowest separation is still relative to the filter strip values of individual 2 sub-basins of interest [Figure 6.13] with and overall filter strip average separation of 2.539%. On the other hand, the least amount of cover crop separation is relative to individual 3 sub-basins of interest similarly to the non IGA based session with an overall average cover crop separation percentage of 9.349%.

Moreover, considering all the separation analysis [Figure 6.13 to Figure 6.18], we notice that the average separation percentage of cover crop is 14.67%. In addition, the average filter strip separation analysis all figures included [Figure 6.13 to Figure 6.18] is equal to 17.46%.

The average disagreement between users decision space preferences relative to design generated using IGA, is less than 18%. Even though this separation values is almost two times higher than the one relative to the non IGA designs decisions, it is still a relatively small disagreement linked to subjective preferences in terms of decisions implementation.


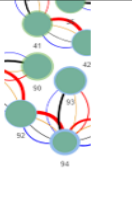
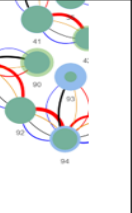
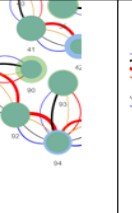
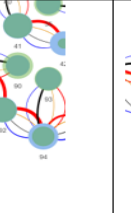
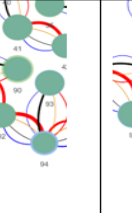
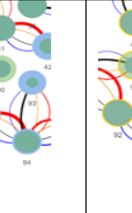
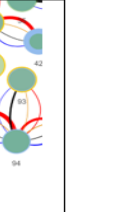
Separation in Decision Spaces of Individual 7 and i^{th} Individual	Individual 1: sub basins {41, 90, 92, 93}	Individual 2: sub basins {41, 90, 92, 93}	Individual 3: sub basins {41, 90, 92, 93}	Individual 4: sub basins {41, 90, 92, 93}	Individual 5: sub basins {41, 90, 92, 93}	Individual 6: sub basins {41, 90, 92, 93}	Individual 7: sub basins {41, 90, 92, 93}
							
Average Separation in cover crops decisions	0.19186	0.118849	0.057394	0.103949	0.146585	0.059235	0
Average Separation in filter strip decisions	0.058144	0.168532	0.072623	0.057499	0.057499	0.1562	0

Figure 6.18: Average separation in the decision space between individual 7 sub-basins of interest and the i^{th} individual decision space values for the same sub-basins.

Moreover, although, the decision space separation percentage relative to IGA design sessions is higher than the ones relative to the non IGA design session, the objective space remains unchanged. That indicate that the GA algorithms generates designs that are more user specific while searching for an optimal common solution objective wise.

To conclude, in both the non IGA and the IGA case, the users are not reaching a 100% agreement when it comes to the decision space, the small separation in their preferred conservation plan practices represents a tolerable difference that is due to the subjective human preference. From both the visual and the analytical observations, the participant are converging toward a rather similar conservation plan with identical objective space values and relatively similar decision space values.

6.3 Inter-Session Comparison

Comparing non IGA related designs and IGA related designs from both a visual and analysis observation, the participants are maintaining similar patterns in these two sessions.

In order to verify the hypothesis that the participants are maintaining similar preferences when moving from one type of session to another, we compute the correlation between the two different types of sessions for each individual and each variable [Figure 6.19].

Correlation between interactive and non-interactive sessions	Cover Crops	Filter Strip	Peak Flow Reduction	Cost Reduction	Sediment Reduction	Nitrate Reduction
Individual 1	0.890162	0.841617	0.999576	0.995125	0.997678	0.999997
Individual 2	0.918207	0.868033	0.999779	0.995896	0.996636	0.999992
Individual 3	0.908046	0.758737	0.999838	0.996076	0.995567	0.999992
Individual 4	0.854339	0.765479	0.999843	0.992658	0.997153	0.999995
Individual 5	0.905433	0.907441	0.999702	0.991966	0.994567	0.999992
Individual 6	0.859009	0.81588	0.999541	0.990539	0.995918	0.999986
Individual 7	0.859009	0.81588	0.999541	0.990539	0.995918	0.999986

Figure 6.19: Correlation between interactive and non interactive sessions for each individual and each decision and objective space variable.

From [Figure 6.19] we first notice that the highest close to 1 correlation values for each individual are relative to the objective space variables. This emphasizes the fact that the participant are remaining consistent with their conservation plan objective when moving from one session type to another. In addition, this very high between sessions similarity in the objective space also highlights the already noticed visual agreement between individuals objective space wise.

Moreover, analyzing the between session correlation in the decision variable space, we notice that the lowest correlation values are the ones related to the filter

strip variable with the lowest value reaching 0.7587 for individual 3. Even though, both correlation relative to filter strips and cover crops reach values as low as 0.7587 and 0.8543, these values still indicate a positive high correlation between the two types of sessions decision space wise.

Finally, [Figure 6.19] numerically justifies that the participant remained fairly consistent with the decision and objective space choices when moving from a non IGA based session to IGA based sessions. These findings agree with the visual assessments and provide decision stakeholder with more confidence in the conservation plan data collected.

6.4 Hot Spot Map and Network Map Visualization

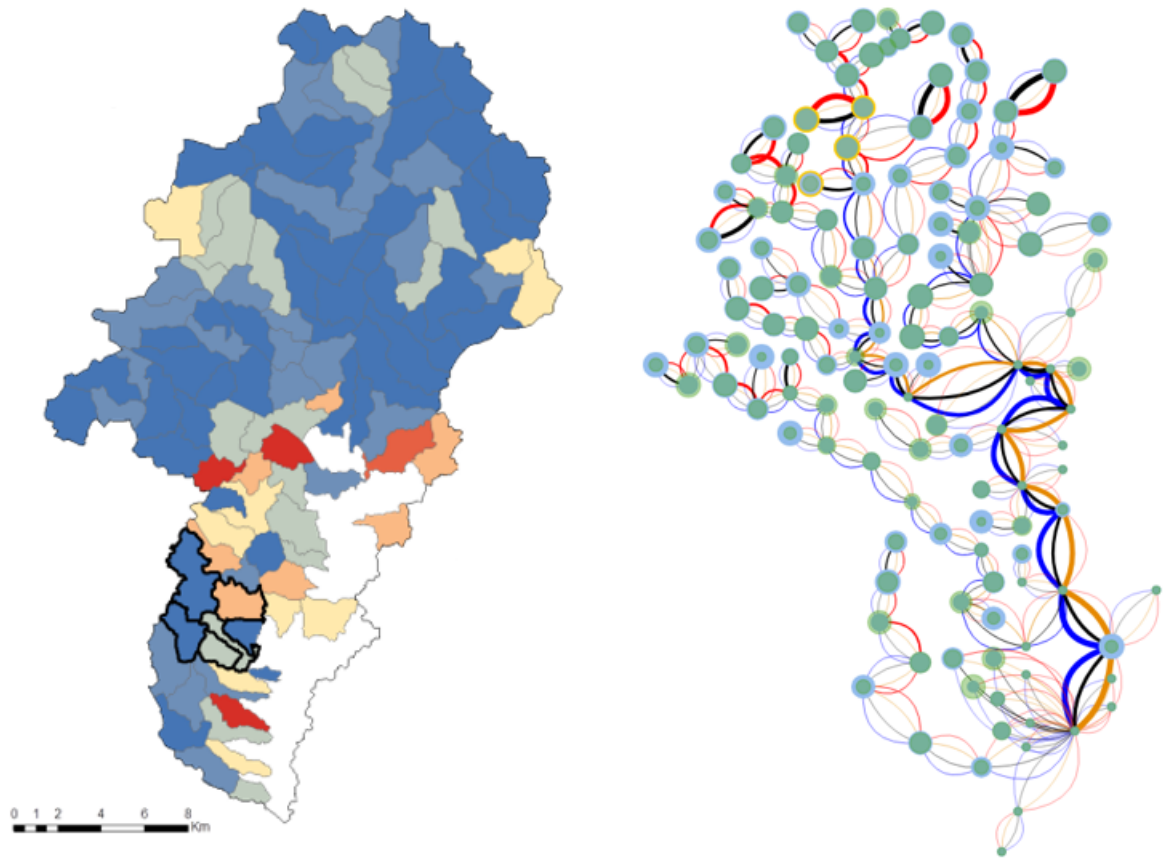


Figure 6.20: A side by side comparison between a single variable (cover crops) hot spot map representation and our holistic network map visualization.

Figure 6.20 shows a side by side comparison between a hot spot map used for the eagle creek watershed conservation plan visualization and our novel network map visualization model.

The hot spot map is a single variable representation thus it only provides

knowledge about one variable independently of its interaction and influence on other variables. In fact, if we consider our data set as a set of records M and dimensions N [28], the number of values used in the hot spot map visualization is equal to $130 * 1$. On the other hand, our network map visualization is a multi-dimension visualization that uses $130 * 6$ values to generate the representation. There is more information encapsulated in the second network map visualization than in the hot spot map. Moreover, visualizing multiple variables simultaneously adds the information of the relationship between dimensions $\frac{N*(N-1)}{2}$ [28], in our case 15 pairwise relationships between the dimensions are represented in our visualization not counting 3-ways, 4-ways or N-ways relationships between the different dimensions.

Moreover, due to the small amount of information visualized, the hot spot map visualization does not provide the user with a selective information-content [27]. Additionally, mentally linking a set of uni-variate hot spot maps in order to get an understanding of the "big picture" and the underlying patterns can become cognitively challenging and could end up being misleading.

Since having a holistic complete visualization [8] plays a crucial role in pattern discovery, the goal of our novel visualization methodology is to provide a holistic view with carefully chosen graphical attributes that are both easy to discriminate [12] and reflective of the connection between the two spaces of the data set. In fact, our network map visualization model enabled the discovery of interesting pattern in the pilot watershed conservation plan study. The discovered patterns that were further analytically validated provide a global and local understanding to guide

stakeholder in their posterior decision making process.

Chapter 7: Conclusion

In this thesis, we presented our novel multidimensional decision and objective space graph visualization and demonstrated its usage in a watershed conservation plan context. The main goal of this visualization model is to provide the user with a holistic view of all the variables in a single 2D visual in a way that favors the discovery of meaningful insights and patterns. In fact, this holistic complete view is important especially when visualizing decision and objective space related variables that happen to carry a semantic relationship between the two spaces. Preserving that inter-variable connection facilitates the pattern and trends discovery. Moreover, along with the holistic view, our visualization also provides a way of separately visualizing each dimension or a combination of dimensions. The demonstrated simultaneous visualization in the watershed conservation plan context facilitated the process of patterns discovery and resulted in an understanding of inter-variable and inter-user relationship that is in accordance with the results of data analysis. As demonstrated in the results section, this visualization enabled us to gain interesting insight that would be helpful for posterior decision making process.

7.1 Limitations and Future Work

Simultaneously visualizing multivariate data with all its inter-variable connections is a challenging task that can easily result in a cluttered, cognitively challenging visualization. For this reason, one of the limitation of our visualization is the limited number of attributes in the decision space. In fact, in the demonstrated watershed conservation plan visualization, the number of decision space variables is equal to two when the number of objective space variables visualized reaches four variables. We limit the number of visualized decision space variables to a maximum of three variables in order to not exceed seven variables simultaneously visualized and avoid cognitive overload [20]. This seven simultaneously visualized variables limit is relative to the usage of color coding that should not exceed seven colors [5] in order for the user to be able to easily discriminate the different attributes[20]. Moreover, due to the high dimensionality of the data set visualized, we sacrifice the ability to show finely grained quantitative details of each variable. In fact, we can still visualize important high and low point patterns of each variables but this visualization doesn't provide a representation of the actual quantitative values of each data point. In addition, in general the patterns and correlation between the attributes of a data set are not already known by the user and the goal of data visualization is to provide that knowledge gain by unveiling meaning insight about the data. However, since we don't have a preliminary knowledge about the trends and patterns encapsulated by the data, we cannot confidently assess the effectiveness of a particular visualization [34]. In thesis, we validated our data

visualization to a certain degree by using data analysis in general and correlation analysis in particular. However, reaching a complete confident validation would require information visualization to be governed by a clearly defined theory or set of theories [28] which is still an open area of research.

Bibliography

- [1] *WESTORE Website*. Accessed: 11/20/2017.
- [2] Meghna Babbar-Sebens, Robert C Barr, Lenore P Tedesco, and Milo Anderson. Spatial identification and optimization of upland wetlands in agricultural watersheds. *Ecological Engineering*, 52:130–142, 2013.
- [3] Meghna Babbar-Sebens and Barbara S Minsker. Interactive genetic algorithm with mixed initiative interaction for multi-criteria ground water monitoring design. *Applied Soft Computing*, 12(1):182–195, 2012.
- [4] Meghna Babbar-Sebens, Snehasis Mukhopadhyay, Vidya Bhushan Singh, and Adriana Debora Piemonti. A web-based software tool for participatory optimization of conservation practices in watersheds. *Environmental Modelling & Software*, 69:111–127, 2015.
- [5] Simone Bianco, Francesca Gasparini, and Raimondo Schettini. Color coding for data visualization. In *Encyclopedia of Information Science and Technology, Third Edition*, pages 1682–1691. IGI Global, 2015.
- [6] Ayan Biswas, Soumya Dutta, Han-Wei Shen, and Jonathan Woodring. An information-aware framework for exploring multivariate data sets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2683–2692, 2013.
- [7] Cynthia Brewer. *Designing better Maps: A Guide for GIS users*. ESRI press, 2015.
- [8] Winnie Wing-Yi Chan. A survey on multivariate data visualization. *Department of Computer Science and Engineering. Hong Kong University of Science and Technology*, 8(6):1–29, 2006.
- [9] Peter Eades. *Drawing free trees*. International Institute for Advanced Study of Social Information Science, Fujitsu Limited, 1991.

- [10] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. *Advances in knowledge discovery and data mining*, volume 21. AAAI press Menlo Park, 1996.
- [11] Thomas Gorko, Calvin Yau, Abish Malik, Matt Harris, Jun Xiang Tee, Ross Maciejewski, Cheryl Qian, Shehzad Afzal, Bryan Pijanowski, and David Ebert. A multi-scale correlative approach for crowd-sourced multi-variate spatiotemporal data. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.
- [12] Christopher G Healey et al. Perception in visualization. *Retrieved February*, 10:2008, 2007.
- [13] Ivan Herman, Guy Melançon, and M Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on visualization and computer graphics*, 6(1):24–43, 2000.
- [14] Jonathan D Herman, Harrison B Zeff, Patrick M Reed, and Gregory W Characklis. Beyond optimality: Multistakeholder robustness tradeoffs for regional water portfolio planning under deep uncertainty. *Water Resources Research*, 50(10):7692–7713, 2014.
- [15] Alfred Inselberg. The plane with parallel coordinates. *The visual computer*, 1(2):69–91, 1985.
- [16] Bernhard Jenny, Daniel M Stephen, Ian Muehlenhaus, Brooke E Marston, Ritesh Sharma, Eugene Zhang, and Helen Jenny. Force-directed layout of origin-destination flow maps. *International Journal of Geographical Information Science*, 31(8):1521–1540, 2017.
- [17] Bernhard Jenny, Daniel M Stephen, Ian Muehlenhaus, Brooke E Marston, Ritesh Sharma, Eugene Zhang, and Helen Jenny. Design principles for origin-destination flow maps. *Cartography and Geographic Information Science*, 45(1):62–75, 2018.
- [18] Joseph Robert Kasprzyk, Patrick Michael Reed, Brian R Kirsch, and Gregory W Characklis. Managing population and drought risks using many-objective water portfolio planning under uncertainty. *Water Resources Research*, 45(12), 2009.

- [19] Kenneth L Kelly. Twenty-two colors of maximum contrast. *Color Engineering*, 3(26):26–27, 1965.
- [20] George Miller. Human memory and the storage of information. *IRE Transactions on Information Theory*, 2(3):129–137, 1956.
- [21] Susan L Neitsch, Jeffrey G Arnold, Jim R Kiniry, and Jimmy R Williams. Soil and water assessment tool theoretical documentation version 2009. Technical report, Texas Water Resources Institute, 2011.
- [22] Niels-Peter Vest Nielsen, Jens Michael Carstensen, and Jørn Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805(1-2):17–35, 1998.
- [23] Tuan Pham, Rob Hess, Crystal Ju, Eugene Zhang, and Ronald Metoyer. Visualization of diversity in large multivariate data sets. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1053–1062, 2010.
- [24] Adriana D Piemonti, Meghna Babbar-Sebens, and E Jane Luzar. Optimizing conservation practices in watersheds: Do community preferences matter? *Water Resources Research*, 49(10):6425–6449, 2013.
- [25] Adriana Debora Piemonti. *Interactive Genetic Algorithms for Watershed Planning : An Investigation of Usability and Human-centered Design*. Oregon State University, Corvallis, Oregon], 2016.
- [26] Adriana Debora Piemonti, Meghna Babbar-Sebens, Snehasis Mukhopadhyay, and Austin Kleinberg. Interactive genetic algorithm for user-centered design of distributed conservation practices in a watershed: An examination of user preferences in objective space and user behavior. *Water Resources Research*, 53(5):4303–4326, 2017.
- [27] Helen C Purchase, Natalia Andrienko, TJ Jankun-Kelly, and Matthew Ward. Theoretical foundations of information visualization. In *Information Visualization*, pages 46–64. Springer, 2008.
- [28] Helen C Purchase, Robert F Cohen, and Murray James. Validating graph drawing aesthetics. In *International Symposium on Graph Drawing*, pages 435–446. Springer, 1995.

- [29] Edward M. Reingold and John S. Tilford. Tidier drawings of trees. *IEEE Transactions on software Engineering*, (2):223–228, 1981.
- [30] Bernice E Rogowitz and Lloyd A Treinish. Data visualization: the end of the rainbow. *IEEE spectrum*, 35(12):52–59, 1998.
- [31] Jinwook Seo and Ben Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information visualization*, 4(2):96–113, 2005.
- [32] Robert Simmon. *Use of Color in Data Visualization*. Accessed: 04/28/2018.
- [33] Soil and NY Water Conservation District Wayne County. *What is a watershed?* Accessed: 05/08/2018.
- [34] Jarke J Van Wijk. Views on visualization. *IEEE transactions on visualization and computer graphics*, 12(4):421–432, 2006.
- [35] John Q Walker. A node-positioning algorithm for general trees. *Software: Practice and Experience*, 20(7):685–705, 1990.
- [36] Matthew O Ward and Kevin J Theroux. Perceptual benchmarking for multivariate data visualization. In *Scientific Visualization Conference, 1997*, pages 314–314. IEEE, 1997.
- [37] Edward J Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411):664–675, 1990.
- [38] Wikipedia. *Flow Map*. Accessed: 03/15/2018.
- [39] Wikipedia. *Graph Drawing*. Accessed: 03/13/2018.
- [40] Matthew J Woodruff, Patrick M Reed, and Timothy W Simpson. Many objective visual analytics: rethinking the design of complex engineered systems. *Structural and Multidisciplinary Optimization*, 48(1):201–219, 2013.

