POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Supporting Transportation Decision-Makers with Tool Design and Data Uncertainty Visualizations

NASIM SHARBATDAR
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Septembre 2020

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Supporting Transportation Decision-Makers with Tool Design and Data Uncertainty Visualizations

présenté par **Nasim SHARBATDAR**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Farida CHERIET, présidente
Jinghui CHENG, membre et directeur de recherche
Catherine MORENCY, membre et codirectrice de recherche
Thomas HURTUT, membre

DEDICATION

To my beloved parents,
and my lovely sister...

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere gratitude to my supervisor, Professor Jinghui Cheng for giving me an opportunity to grow, and for encouraging and trusting me. I truly appreciate your support, honesty, and patience throughout my studies.

Also, I would like to thank my co-supervisor Catherine Morency for her guidance and tremendous support during my studies and helping in the several challenges faced in the mobility and transport domain.

I am very grateful and privileged to be part of the Ministry of Transportation of Québec's (MTQ) OSC^{2}AR project to develop and deploy the OSC^{2}AR tool for monitoring and analyzing traffic conditions on the road network.

I also acknowledge all my committee members: Professor Farida Cheriet and Professor Thomas Hurtut for sparing their precious time in order to review and evaluate my research work.

Finally, a special thanks to my family, to my parents who trusted me to follow my dreams, provided unconditional support throughout my education and devoted all their life to my success and happiness, and to my sister for all her support and love.

RÉSUMÉ

Les décideurs en matière de transport des agences gouvernementales jouent un rôle important dans la gestion des conditions du réseau de circulation, qui à leur tour ont un impact majeur sur le bien-être des citoyens. Les pratiques, les défis et les besoins de ce groupe de praticiens sont moins représentés dans la littérature HCI. D'autre part, il existe un manque d'outils permettant de répondre aux besoins des décideurs et de les aider à suivre et analyser les conditions de circulation sur le réseau routier. De plus, presque toutes les données de transport comportent une certaine incertitude, et une bonne communication de l'incertitude des données par le biais de visualisations peut avoir un effet important sur la qualité des décisions prises par les experts en transport. Cependant, il y a peu de connaissances sur la façon dont les décideurs en matière de transport gèrent les données incertaines et perçoivent les visualisations de l'incertitude des données.

Nous comblons ces lacunes dans un premier temps par une étude d'entrevue avec 19 praticiens du Ministère des Transports du Québec (MTQ), un organisme gouvernemental responsable des infrastructures de transport au Québec, Canada. Nous avons ensuite créé des directives de conception basées sur les résultats de ces entretiens et conçu une interface utilisateur pour l'analyse des conditions de circulation. La conception de l'interface utilisateur a été évaluée et répétée au moyen de courtes études sur les utilisateurs avec 6 praticiens travaillant au MTQ.

Pour aider les praticiens du transport à comprendre et à résoudre l'incertitude des données, nous avons proposé une nouvelle méthode et appliqué la divergence Jensen-Shannon (JSD) pour quantifier l'incertitude des données sur la vitesse de la circulation. Nous avons ensuite utilisé plusieurs techniques de visualisation telles que la couleur, la forme et la taille pour créer des visualisations d'incertitude qui présentent les niveaux d'incertitude des données en chevauchement avec les données de vitesse d'origine. Nous avons ensuite mené une étude sur les utilisateurs auprès de 11 décideurs experts et novices en matière de transport pour explorer comment les praticiens considèrent l'incertitude des données dans leur processus de prise de décision et comprennent également leurs préférences et leurs perceptions sur nos visualisations d'incertitude des données.

Dans notre recherche, nous avons constaté que les décideurs qui travaillent à la surveillance et à l'analyse des conditions de circulation sur le réseau routier peuvent le plus bénéficier de la recherche sur les outils et plates-formes d'analyse de données qui (1) fournissent des informations pour soutenir la sensibilisation à la qualité des données, (2) sont interopérables.

avec d'autres outils dans le flux de travail complexe des praticiens, et (3) prennent en charge une analyse visuelle intuitive et personnalisable. Ces implications peuvent également être informatives pour la conception d'outils soutenant d'autres tâches et domaines de prise de décision. De plus, les professionnels du transport ont trouvé que la conception de notre outil était utile pour les aider à comprendre et à analyser leurs données. Ils ont également souligné que le filtrage flexible des données, le choix de la sélection de visualisation et les fonctionnalités de comparaison de données sur une seule interface conviviale peuvent être très bénéfiques pour leur processus de prise de décision.

En outre, notre étude des utilisateurs sur l'incertitude des données a révélé les défis importants auxquels les praticiens sont confrontés lorsqu'ils traitent des données incertaines, tels que les biais dans la collecte de données, les erreurs humaines et d'appareils et le manque de confiance envers les fournisseurs de données tiers. Notre méthode de quantification de l'incertitude basée sur la JSD a été généralement bien perçue par les participants pour aider à obtenir une meilleure transparence et une meilleure connaissance de leurs données. De plus, les résultats de l'étude sur les utilisateurs ont indiqué que des facteurs tels que la familiarité avec les représentations, les connotations existantes de la visualisation basées sur la pratique et les objectifs d'utilisation des données ont influencé les préférences des participants sur les visualisations d'incertitude des données.

En résumé, les résultats de ce mémoire peuvent être utilisés pour créer de meilleurs outils d'analyse visuelle pour soutenir les praticiens dans la pratique de la prise de décision en matière de transport.

ABSTRACT

Transportation decision-makers from government agencies play an important role in addressing the traffic network conditions, which in turn, have a major impact on the well-being of citizens. The practices, challenges, and needs of this group of practitioners are less represented in the HCI literature. On the other hand, there is a gap for tools that can fulfill decision-makers' needs and help them monitor and analyze traffic conditions on the road network. Additionally, almost all transportation data have some uncertainty, and good communication of data uncertainty through visualizations can have a great effect on the quality of the decisions that transportation experts make. However, there is little knowledge about how transportation decision-makers deal with uncertain data and perceive data uncertainty visualizations.

We address these gaps first through an interview study with 19 practitioners from Transports Québec (MTQ), a government agency responsible for transportation infrastructures in Québec, Canada. We then created design guidelines based on the results of these interviews and designed a user interface for traffic conditions analysis. The user interface design was evaluated and iterated through short user studies with 6 practitioners working at MTQ.

To help transportation practitioners understand and address data uncertainty, we proposed a novel method and applied Jensen-Shannon Divergence (JSD) to quantify data uncertainty in traffic speed. We later used several visualization techniques such as color, shape, and size to create uncertainty visualizations that present the levels of data uncertainty in overlapping with the original speed data. We then conducted a user study with 11 expert and novice transportation decision-makers to explore how practitioners consider data uncertainty in their decision-making process and also understand their preferences and perceptions on our data uncertainty visualizations.

In our research, we found that decision-makers who work on monitoring and analyzing traffic conditions on the road network can most benefit from research about data analysis tools and platforms that (1) provide information to support data quality awareness, (2) are interoperable with other tools in the complex workflow of the practitioners, and (3) support intuitive and customizable visual analytics. These implications can also be informative to the design of tools supporting other decision-making tasks and domains. In addition, transportation practitioners found our tool design to be useful in helping them understand and analyze their data. They also pointed out that flexible data filtering, choice of visualization selection, and data comparisons features on a single user-friendly interface can be very beneficial for

their decision-making process.

Further, our user study about data uncertainty revealed the important challenges that practitioners faced when dealing with uncertain data, such as bias in data collection, human and device errors, and lack of trust in third-party data providers. Our JSD-based uncertainty quantification method was generally well perceived by the participants to help achieve better transparency and awareness of their data. Additionally, the user study results indicated that factors such as familiarity to the representations, existing connotations of the visualization based on practice, and the goals of data usage influenced the participants' preferences on data uncertainty visualizations.

In summation, the results of this thesis can be used to create better visual analysis tools to support practitioners in the transportation-decision making practice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

MTQ	Ministère des Transport(Transports Québec)
ERT	État des routes
UCD	User-Centered Design
CHI	Conference on Human Factors in Computing Systems
HMI	Human-Machine Interactions
KLD	Kullback-Leibler Divergence
JSD	Jensen-Shannon Divergence
TAZ	Traffic Analysis Zone
OD	Origin-Destination
GPS	Global Positioning System

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

1.1 Context and Motivation

Traffic congestion is a matter that brings concerns in most populated cities, posing various economic, environmental, and social issues [2] [3] [4]. For example, the annual estimated cost of congestion in Montréal in 2018 is $4.2 billion [4]. During the same year, congestion cost the United States a total of $87 billion and 6.3 billion hours of direct time loss [3]. Bad traffic conditions also cause accidents, increase air pollution, and result in energy waste. The traffic network conditions are largely determined by two interlocked human factors: (1) the behaviors and needs of traffic network users (e.g. drivers, cyclists, etc.) and (2) the perception and practices of transportation decision-makers. In Human-Computer Interaction (HCI) research, plenty of work has been done in understanding and supporting the first group of people (e.g. [5,6]), while the challenges and needs of the second group are less represented. We address this limitation and directly focus on understanding and supporting transportation decision-makers.

Montréal is the second largest city in Canada, with a population of around 1.7 million. As many major cities, Montréal experiences some severe traffic congestion, especially in the central business areas and in the road networks connecting the city with its suburbs. The traffic situation is also complicated by the frequent events that require traffic control and the severe winter weather conditions. The transportation management and planning practitioners often need to make difficult decisions, balancing short term traffic impacts and long term benefits and risks when planning road closures and road work.

We are particularly interested in understanding the practices, challenges, and needs of this group of decision-makers and examining how they use traffic data to support their work. To achieve this, we first conducted an in-depth interview study with transportation decision-makers. The results from these interviews also allowed us to create four personas [7] that summarized the challenges and needs of different types of transportation decision-makers and can be used to communicate these elements with the larger community who focuses on creating tools to support transportation decision-making. Leveraging this knowledge, we then designed the user interface of a traffic condition analysis tool for decision-makers. Our tool design was evaluated and iterated through a user-centered approach with the decision-makers.

On the other hand, transportation data often carries all sorts of uncertainty. Using raw data without being aware of the limitations and the quality of data will lead to biased decisions and results. Therefore bringing awareness of data uncertainty to decision-makers can help them make better, more accurate, and more reliable decisions. More specifically, transportation decision-makers frequently work with average speed data in their traffic conditions analysis process. Traditionally, they use the number of observations recorded for average speed data to assess its quality. However, there is no concrete measurement to show how many number of observations are enough to assess the uncertainty of speed data. Although there is no specific measurement of the uncertainty of their data, making comparisons using historical data would give them a better understanding of how to deal with their data and validate their decisions based on it. Methods such as relative entropy and Jensen-Shannon divergence can be explored to quantify uncertainty in transportation data based on historical data comparisons. Additionally, data uncertainty is an extremely important and complex concept to communicate through visualizations for experts and non-experts, as it can graphically present large data in such a way that details and characteristics of the data can be easily grasped. Good communication of data uncertainty through visualizations can have a great effect on the quality of the decisions that transportation experts make.

Thus in this thesis, we also focus on examining methods of quantifying and visualizing uncertainty in transportation data. Our uncertainty quantification method is based on the similarity of distributions between the data under analysis and the historical data, leveraging the concept of Jensen-Shannon divergence; this method avoids the reliance on the number of observations. Additionally, we also designed several ways to visualize the data uncertainty we calculated, overlapping with the original data. In a user study, we explored how transportation decision-makers considered uncertainty in transportation data and perceived our quantification and visualization methods.

1.2 Research Objectives

This master's thesis focuses on facilitating the decision making process for transportation decision-makers to achieve better and more accurate results. Making long term transportation decisions is an extremely important and difficult process and transportation decision-makers face many challenges along the way. Not being aware of the levels of uncertainty in the data used for decision-making can also have severe consequences. To address these issues, this thesis focus on the following objectives:

1. Identifying the practices, needs, and challenges of transportation decision-makers through an interview study.

2. Contributing to the designing of a user interface for decision-makers for their traffic conditions analysis.

3. Quantifying and visualizing data uncertainty and examining them through a user study with decision-makers.

To achieve these objectives, we first conducted interviews with transportation decision-makers to understand their needs. We then contributed to the design of the user interface of a new traffic condition analysis tool and created guidelines and wire-frames and evaluated them through a user study. We finally proposed and evaluated a method to quantify and visualize uncertainty in traffic data.

1.3 Thesis Outline

The rest of the thesis is organized as follows. In chapter 2 we present a review of the related literature; particularly, we focused on previous studies about user-centered design, personas, role-based visualizations, visualizations for traffic analysis, quantifying data uncertainty, and data uncertainty visualization. Next, chapter 3 (based on the paper we have published in CHI 2020 Late-Break Works [8]) discusses transportation decision-makers' needs in a traffic condition analysis tool. Chapter 4 contains the user-centered design process for a traffic condition analysis tool that focuses on addressing the needs of the transportation decision-makers. Chapter 5 reports on our new uncertainty quantification and visualization methods for transportation data and a user study evaluating our methods. Finally, we conclude the thesis by including the summary of our work, limitations, and potential future work in chapter 6.

CHAPTER 2 LITERATURE REVIEW

This work is linked to several areas of scientific literature focused on (1) the general approach to user-centered design, (2) the creation and use of personas, (3) interaction-based design on the role and (4) strategies, visualization objects relevant for traffic analysis, (5) quantifying data uncertainty, and (6) visualization techniques to represent data uncertainty. In the following sections, we briefly review the literature in each of these areas.

2.1 User-Centered Design

User-centered design (UCD), sometimes also called "human-centered design" or "design thinking", is a general term that describes a methodology that helps designers and developers to create better information and communication technologies to meet the needs of users. The UCD process essentially focuses on involving end-users throughout the design process, in activities ranging from occasional user advice to intensive participatory design. The concept originated from Donald Norman's research laboratory at the University of California at San Diego (UCSD) and was developed in Norman's founding book "The Design of Everyday Things" [9]. Norman [9] described UCD as a design process focusing on (1) solving the right problem and (2) a way of doing it that meets the needs and capabilities of humans [9]. The process involves an iterative cycle with four overlapping phases, described in Figure 2.1 [9].

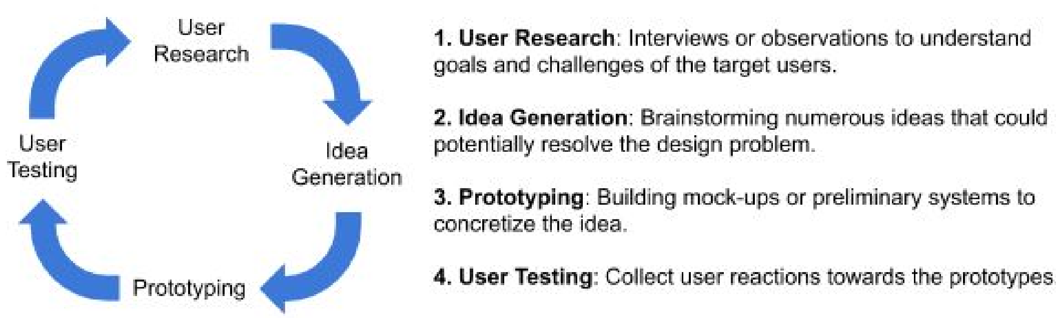


Figure 2.1 The iterative process of user-centered design

Following Norman's work, the researchers identified key features of the UCD. For example, Gould and Lewis [10] recommended three principles of system design that represent the essence of UCD. These principles are: (1) early focus on users and tasks, (2) empirical measurement of user interaction and (3) iterative design [10]. Gulliksen et al. [11] also examined existing theories and analyzed them in a large number of software projects in order to identify the practical principles of UCD. Ultimately, 12 UCD principles were proposed, which capture factors such as user participation, incremental development, prototyping, user evaluation and a user-centric attitude by the team development [11]. In order to support the application of these principles in UCD practice, the authors then built lists of activities (one for each principle) specifying the main techniques and methods enabling the principles to be satisfied [11]. For example, user interviews, focus groups, field studies and personas are recommended to support "user focus". For the ideation phase of UCD, many researchers have proposed general design guidelines to facilitate the creation of an appropriate interaction. The most influential are the eight Golden Rules by Shneiderman and the ten general heuristic methods by Nielsen for the design of human-machine interactions (HMI). We summarize the two sets of guidelines and heuristics below.

Based on the concept of direct manipulation, which focuses on leveraging physical metaphors in creating visual representations and user interactions [12], Shneiderman proposed the Eight Golden Rules of Interface Design [13], which we briefly discuss below.

1. **Strive for consistency**, which specifies that design elements, including layout, colour, terminologies, and sequences of actions should be used consistently throughout the system. Exceptions should be limited and readily comprehensible.

2. **Seek universal usability**, which requires designers to recognize the needs of diverse users (based on factors including technical expertise, culture, disabilities, age, and so on) and design for “plasticity.”

3. **Offer informative feedback**, which specifies that the system should provide appropriate feedback to user actions; for example, modest feedback and responses should be provided for frequent and minor actions, while substantial feedback should be given to infrequent and major actions.

4. **Design dialogues to yield closure**, which identifies that sequence of actions (based on tasks) should be organized as conversational dialogues that include a beginning, middle, and end.

5. **Prevent errors**, based on which the user interface should be designed so that users cannot make serious errors. Additionally, once an error was made, easy and meaningful recovery mechanisms should be provided.

6. **Permit easy reversal of actions**, based on which the user interface should offer undo and redo actions as much as possible.

7. **Keep users in control**, which requires the interface to be designed so that users can perform the desired task and surprises and annoyances are minimized.

8. **Reduce short-term memory load**, which specifies that designers should avoid interfaces in which users must remember information from one screen to another.

Jakob Nielsen and his colleagues have also proposed a set of guidelines for creating and evaluating user interface design, called "usability heuristics" [14]. Some of these guidelines overlap with Shneiderman's "golden rules". Here we discuss each heuristic and mention this overlap, if any.

1. **Visibility of the system state**. It overlaps with Shneiderman's 3rd "Golden Rules" directive and further identifies the importance of continuous visibility and updating of system state.

2. **Adequacy between the system and the real world**. It captures the overall objective of direct manipulation, by specifying that the system must use terms, concepts, and a sequence of actions familiar to users, preferably respecting real-world conventions rather than using terms system-oriented.

3. **User control and freedom**. This cuts across Shneiderman's Guidelines 6 and 7.

4. **Consistency and standards**. It overlaps with Shneiderman's first "Golden Rules" directive. It also emphasizes compliance with the design standards and conventions of the platform.

5. **Error prevention**. It overlaps with Shneiderman's 5th "Golden Rules" directive.

6. **Recognition rather than recall**. It is superimposed on Shneiderman's 8th "golden rules" directive and offers a generic solution to reduce memory load in the short term.

7. **Flexibility and efficiency of use**. It overlaps with Shneiderman's 2nd directive.

8. **Aesthetic and minimalist design.** Aesthetic and minimalist design. It specifies that user interfaces should be organized logically and should not contain irrelevant information.

9. **Help users recognize, diagnose and recover from errors**. This cuts across Shneiderman's Guidelines 6 and 7.

10. **Help and documentation**, which requires designers to provide succinct and informative help and documentation in support of the user's task.

In this project, we follow a UCD approach in the design of the information tool for the analysis of traffic conditions on the road network. In particular, we used an interview method to research users and took advantage of the concept of personas (see next section) to capture user needs. We will use the Shneiderman and Nielsen design guidelines to support the development of the user interface.

2.2 Personas

The term persona was borrowed from an ancient Greek word meaning "theater masks". A persona can be defined as a virtual user who synthesizes a group of real users with regard to the type of interaction they wish to have with a system. Cooper was one of the first to introduce this term in the context of UCD to describe "hypothetical archetypes of real users", that is, groups of real users sharing needs and common characteristics, represented by fictitious individuals who are both concrete and alive [15]. In the context of UCD, personas are the primary outcome of the first stage of user research. Each persona usually contains a realistic name, a representative photo, and descriptions of their typical roles, tasks, objectives, motivations, and challenges. Since its introduction, the concept and the term persona have been widely adopted in the field of UCD and integrated into the design process of many companies to help the design and development team describe the characteristics of their users target and develop empathy towards the end- users of their systems. For example, Nielsen and Hansen [16] surveyed 13 Danish companies to investigate the use of persona. They found that personas were created early in the design process and used throughout the communication of user needs to make design decisions [16]. Miaskiewicz and Kozar [1] have summarized 22 benefits of using personas according to the opinions of 19 design experts [1]. We summarize the top five benefits in Table 2.1.

Many previous studies have reported cases where the use of personas has supported the UCD process and the design of successful products. For example, Pruitt and Grudin [17] described their process using personas at Microsoft. Once the personas were created, they were used throughout the development process, especially during the design, prioritization of functionality, documentation, and tests [17]. For example, the team assessed the importance of each characteristic of the product in relation to each persona in order to justify the priorities of the functionalities. Dotan et al. provided a detailed report on how personas were used when redesigning a learning platform [18]. In their project, the personas were created by means of interviews and observations with the target users, then supplemented by meetings with the development team. Once created, the personas were first discussed in a "familiarization session" during which the product developers made general hypotheses for each persona.

Benefit	Description
Audience focus	Focus product development on users/customers and their goals (rather than the specific limitations or opportunities presented by technology)
Product requirements prioritization	Prioritize product requirements and help to determine if the right problems are being solved
Audience prioritization	Prioritize audiences and bring about a focus on the most important audience(s)
Challenge assumptions	Bring to the surface and challenge long-standing (and often incorrect) organizational assumptions about the users/customers
Prevention of self-referential design	Help individuals realize how the users/customers are different from themselves

Table 2.1 Benefits of personas in UCD according to Miaskiewicz and Kozar's study [1]

It was then followed by a formative revision session of the prototype, during which the previous version of the product was examined in relation to the objectives and needs of the personas. Finally, a list of specific requirements for each persona was generated during this process. Margaret Burnett and her colleagues have also worked extensively on the role of personalities in promoting gender mainstreaming in software design. In their recent study, Burnett et al. (2016) found that GenderMag, a persona-driven software inspection method they developed, has successfully helped software practitioners to detect gender inclusion issues in their product [19].

The use of personas has been studied more formally in the context of software requirements engineering. For example, Aoyama (2005) proposed a persona-based approach to requirements engineering, called the Hanako method [20]. This method includes two main activities: (1) Persona analysis, in which the personas and their corresponding scenarios (ie idealized stories about the use of the system from a persona point of view) were created based on the grouping of users and their needs, and (2) the analysis of the value of requirements, in which the “hot spots” of the requirements and the “extension points” of the product were identified on the basis of an analysis of the interaction between personas and scenarios. Almaliki et al. (2015) proposed PAFA, a persona-based method for "adaptive feedback acquisition" [21]. This method focused on the use of personas in the analysis and adaptation of user comments in order to develop a software system. It included four iterative phases: (1) derive one or more scenarios according to each persona, (2) create objective models to capture the requirements of each persona according to their scenarios, (3) designing use cases based on the goal modelings to concretize desired user interaction, and (4) using these artifacts to analyze the variability and commonalities of functionality for different personas in order to adapt user feedback. Previous research has also identified various limitations of personas. Blomquist and Arvola (2002) presented a longitudinal observational study on the use of personas in a medium-sized business. They found that, even if the personas were available, they were not integrated into the design process of this business, mainly because the team members other than the interaction designers did not have sufficient knowledge of this concept [22]. This indicates that it may take a lot of time and effort to familiarize all stakeholders with this concept in order to realize its values. Matthews et al. (2012) conducted an interview study with 12 design experts from a global business based in North America, to study their practical use of personas. They found that in this endeavor, personas were primarily used to communicate the needs of end-users between stakeholders, rather than to fuel the design of their interactions [23]. The authors, therefore, argued that this was due to the fact that their personas were somewhat detached from the user study data and were, therefore, neither realistic nor precise.

In summary, personas have long been introduced and used in the UCD process. They should first be seen as a communication and needs clarification tool. If created and used correctly, they can help the design team stay focused on the needs of their target users and prioritize the different features of the system. In general, the creation of personas must be based on real data from user studies. There is no rule as to the ideal number of personas. Ideally, each persona will synthesize a group of users, but how the grouping will be done will depend on design objectives. Using personas requires knowledge and buy-in from the entire development team.

2.3 Role-Based Visualization

Previous studies have also explored role-based visualization techniques. For example, Wu et al. proposed a prototype of information visualization based on maps for emergency management, taking into account different roles in the teams [24]. Because of the scope, their system has adopted a collaborative design that allows users with different roles to annotate on the map to communicate their reasoning and judgments. Mckenna et al. (2015) reported having adopted a UCD approach to design a cybersecurity dashboard for different types of users [25]. They used two personas (a network manager and a cyber analyst) created as part of an interview study with 12 potential users to identify the needs of critical roles. They then used a data sketching technique to explore potential design solutions; this process focused on quickly creating visualizations based on a small-scale dataset in order to get quick feedback from users. Tokola et al. (2016) designed a manufacturing dashboard system representing three levels of hierarchy based on user roles: one for managers, one for supervisors, and another for operators. A survey method with different manufacturing experts was used to identify the key performance indicators that different types of users appreciated [26]. Mahmoodpour et al. (2018) used these indicators to create a role-based visualization for industrial systems based on the Internet of Things [27].

In our study, we use the approaches of previous works. In particular, we generated personas (through an interview study) to understand the needs (what the tool must meet) and the objectives (what the users aim to achieve with the tool) of important roles (functions of a group of users) of potential users (individuals who use the tool) of the MTQ. Our next step is to develop the key performance indicators for each persona and explore the visualization and design of dashboards with a data drawing technique.

2.4 Visualization for Traffic Analysis

Data visualization for traffic analysis is a difficult and active area of research. Many previous works have studied visualization techniques and tools in this area [28] [29] [30]. For example, Guo et al. have created a system to support interactive inspection of traffic data from three perspectives (spatial, temporal, and multidimensional views) to analyze microscopic traffic patterns and abnormal behaviors [31]. Ferreira et al. proposed a visualization model that allows users to query iteratively and exploratively traffic data, for different hours, using taxi data [32]. Likewise, Pu et al. have developed an interactive system using several advanced visualization techniques to monitor and analyze complex traffic conditions in large cities [33]. Cruz and Machado (2016) proposed a new technique for visualizing the circulation using semantic metaphors inspired by pulsed blood vessels [34]. Gomes et al. (2017) also studied the representation of circulation oscillation models by a technique inspired by the visualization of vector fields [35]. However, in many previous works, the analysis of the tasks that users have to perform with traffic data is insufficient. In particular, knowledge of the needs of planners and traffic experts on viewing traffic data is extremely limited. Our work fills this gap using a UCD approach. It must be understood that the UCD is both a research tool and a development tool; it's also a relatively new approach. This may explain why it has not been possible to find much public information around the users of tools as developed in this project.

2.5 Quantifying Data Uncertainty

The usage of raw data without being aware of their accuracy and limitations will lead to a biased estimation of models and decisions. Although the overall data availability has increased in recent years, this does not necessarily mean that the accuracy level of the data is high. All spatio-temporal data is limited by spatial, temporal, and thematic precision and accuracy [36].

Usually, uncertainty is quantitatively described by values like probability, standard deviation, distance, or error percentage. Thomson et al. use standard deviation to express concepts like error, precision, completeness, consistency, lineage, credibility, subjectivity and interrelatedness [37].

Representing uncertainty may have many advantages but is difficult in many contexts, and misunderstandings are quite common. Although producing data representations is easy, perceiving the data uncertainty correctly is a challenging process, especially for non-expert data consumers [38].

Several methods have been used to measure and quantify uncertainty in data. One common method is using relative entropy also known as Kullback-Leibler Divergence (KLD). Bencomo et al. have used KLD to measure degrees of uncertainty and deviations of self-adaptive systems [39]. Dai et al. proposed a concept called rough decision entropy to measure the uncertainty of an interval-valued decision system [40]. Chen et al. studied the uncertainty measurement problem about neighborhood systems, and proposed neighborhood entropy to measure the uncertainty of a neighborhood system [41].

Although awareness of uncertainty is highly important and required in transportation decision-making, to the best of our knowledge, no uncertainty quantification has been done in the area of speed data in traffic condition analysis. Our aim is to fill this gap by proposing a novel method and applying Jensen-Shannon Divergence (JSD) to quantify data uncertainty in traffic speed.

2.6 Data Uncertainty Visualization

Representations of information play an important part in transportation decision-making. The challenge of creating effective and clear visualizations to communicate uncertainty is an ongoing process. Visual variables were first identified by Bertin [42]. He has proposed seven different visual variables including position, size, shape, color brightness, color hue, orientation, and grain [42]. Other visual variables such as color saturation and arrangement

[43] and fuzziness, resolution, and transparency [44, 45] were added to this list.

Transportation decision-makers work with data that usually has some level of uncertainty along with it. They will need these uncertainty levels to be visualized to help them make more accurate and precise decisions. Various research has been done in the area of uncertainty visualization techniques [46, 47]. MacEachren et al. determined common visual variables for their intuitiveness of representing uncertainty, and evaluated the accuracy of brightness and fuzziness [48]. Sanyal et al. evaluated methods to visualize 1D and 2D data uncertainty using color-mapping or glyph sizes with traditional error bars [49]. Boukhelifa et al. have considered a new visual variable called sketchiness that uses line-based marks for presenting data uncertainty and then evaluated its effectiveness compared to blur, grayscale, and dash [50].

In our work, we have focused on the effectiveness of three main visual variables: color, shape, and size to fill the gap of having data uncertainty visualizations for traffic experts and to support transportation decision-making on the road network.

CHAPTER 3 ANALYZING THE NEEDS IN A TRAFFIC CONGESTION ANALYSIS TOOL

This chapter is based on a paper published in CHI 2020 Late-Break Works [8].

3.1 Introduction

Transportation decision-makers have an important role in long term changes in the structure of a city. One of the main focuses of these decision-makers is to control and analyze traffic congestion which has a strong effect on the well-being of the citizen. To facilitate the process of decision-making in these areas we first have to understand how decision-makers are currently analyzing traffic conditions, what challenges they face in their work, and what needs they have. In this chapter, we plan to understand these factors to later design a traffic congestion analysis tool for them. To achieve our goal we conducted an interview study with transportation decision-makers.

3.2 Methods

We conducted interviews with 19 participants from Montréal working at Transports Québec who focus on various aspects of transportation management and planning.

3.2.1 The Participants

Two employees of Transports Québec supported us in recruiting the participants to cover diverse job functions in the agency. Our participants included 12 males and 7 females. Their professional experience ranged from two years to more than 30 years (median = 7 years). The participants focused on different aspects of transportation management and planning, including (1) transportation modeling, (2) planning and sustainable mobility, (3) road network design and traffic analysis, (4) managing construction projects focused on improving traffic conditions on main arterial roads. Table 3.1 summarizes our participants.

3.2.2 Interview Method

The interviews were conducted in December 2018; each session lasted approximately 45 minutes. During the interviews, we asked the participants about: (1) their roles in Transports

Québec and their main practice and expertise, (2) their needs in using traffic data for decision-making, (3) the challenges they experienced when using such data, and (4) their expectations of tools for analyzing traffic conditions on the road network. All interviews were conducted in French, the official language in Québec. The interviews were recorded in audio and fully transcribed for analysis. The transcripts were later translated into English. For more details about the interview questions, see appendix A.

ID	Expertise / Role
P1	Sustainable transport
P2	Pavement and traffic expertise
P3	Road design
P4	Transport systems modeling
P5	Pavement and traffic expertise
P6	Transport systems modeling
P7	Project delivery
P8	Sustainable transport
P9	Major bridge projects
P10	Transport planning
P11	Major tunnel projects
P12	Transport systems modeling
P13	Sustainable transport
P14	Transport systems modeling
P15	Transport systems modeling
P16	Pavement and traffic expertise
P17	Pavement and traffic expertise
P18	Transport systems modeling
P19	Sustainable transport

Table 3.1 Participants' expertise and role.

3.2.3 Data Analysis

We adopted an inductive thematic analysis approach to analyze the interview data [51]. First, three researchers independently coded the interview data to identify initial ideas and themes. Then we carried out an affinity diagram activity to organize the codes in a hierarchy that reveals the common practices, challenges, and needs of the participants in the context of using the traffic data for supporting decision-making. After reviewing our codes and transcripts during several meetings with our group, we finalized a coding schema illustrating an overview of our findings.

Based on participants' responses to our coding schema, we then categorized the participants and created four personas to concretize and communicate the different characteristics of transportation management and planning practitioners.

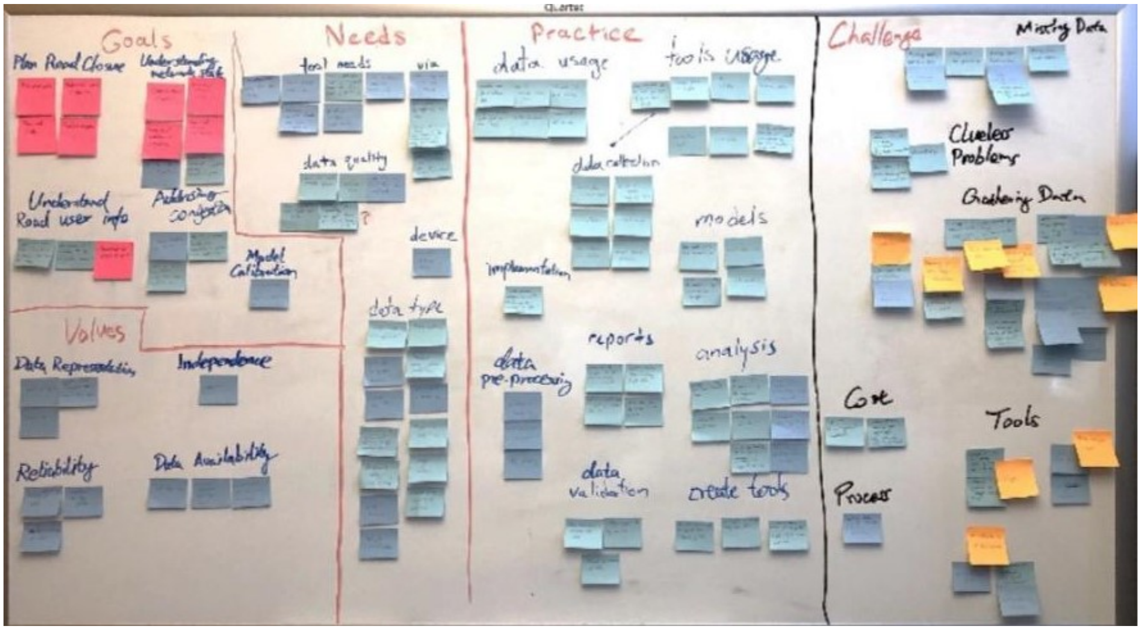


Figure 3.1 Affinity diagram

3.3 Results

3.3.1 Themes identified

We generated themes about the use of traffic data for decision-making in the following three main categories.

Practice

We considered participants' practices in using traffic data for decision-making as contexts for understanding their challenges and needs. Our participants collectively reported the use of four main types of data:

- Traffic counts and flow, ideally classified by the type of vehicle (e.g. trucks vs cars);
- Travel time or speed data based on road sections;
- Travel demand among various origins and destinations;
- Contextual data, varied according to the tasks to be performed, including information about road closures, accidents, weather, lane geometry, the socio-economic profile of geographic areas and users, pedestrian and bicycle counts, land use and residential development, public transit usage, etc.

Participants reported the use of a wide spectrum of recency and granularity of the data. At one end, tasks such as road work management mainly require recent or real-time traffic data with a fine temporal and spatial granularity (e.g. precise segments of the road in increments of 15 minutes). On the other end, tasks such as project planning require historical data with a more coarse temporal and spatial granularity (e.g. seasonal data of a region).

To collect, validate, and analyze this data, participants reported using a variety of tools. Some used more basic such as spreadsheets or Google Maps, while others applied complex ones such as MapInfo, SQL, or R. Participants also used advanced data analysis tools developed internally at Transports Québec or externally by a third party. Using these tools, participants sometimes created 2D/3D graphics to observe the evolution of congestion over time and space. Some participants also produced daily reports and/or made visualizations to justify an intervention.

We also noted two main types of preferences for participants' choice of data and tools, associated with their job responsibilities. Several participants mentioned that they prefer to have access to the raw data to calculate the indicators they needed for analysis. This level of access gives them better control and better knowledge about the quality of the data that goes into their model. However, several others preferred tools that offer graphical analyses and pre-calculated indicators. For these people, traffic data are often secondary inputs to their tasks (e.g. a project manager who needs to quickly validate or transmit information).

Challenges

Participants have discussed several major challenges regarding data collection and use that are interesting to designers of decision-making tools.

Many participants mentioned the challenge concerning the availability of data. For example, usually, only the most important road sections are equipped with permanent vehicle counting devices. Even for those sections, the device fails and severe weather conditions can postpone data collection. For some practitioners, especially those who focus on a more coarse data granularity, the lack of data would not be a major issue. For others, especially those who focus on the management of road closures and construction planning, missing or lack of precision in data can cause unforeseen situations and problems.

The second most frequently discussed challenge is associated with the quality and transparency of the data. This is particularly problematic for data that come from the private sector, as P18 explains: "*I have a big concern for transparency. Especially with big data today. We are talking about Google data among others. It's a big black box. We do not know* *their quality [...] Often we have numbers, we use them. It's very dangerous. [...] If we have bad data, garbage in, garbage out, we will not have good models.*"

Several participants were also concerned about the difficulty of interpreting the data in the absence of contextual variables. For example, P17 mentioned: "*We validate the vehicle counting data, but sometimes we do not know why there has been a decline. It was observed, but we do not know why (...) is it because of the temperature, was there an accident somewhere, is it simply a seasonal decline?*"

Participants also explained that a lot of efforts had to be put into cleaning and validating the data. For example, P4 said: "*The data, like Bluetooth [for capturing traffic speed], contain a lot of noise. We must remove the noise. ... You have to remove the days that do not work.*" Some participants mentioned having conducted long-term studies to collect and validate data, sometimes cross-referencing data from other agencies.

With regard to the analytical tools, having to move frequently from one platform to another has posed a major problem. For example, P10 said: "*The fact that you have to go from one application to another does not make it easy to work. True, we are going to get our information everywhere. [...] But I would prefer that everything be in one place in the end.*" This challenge is originated from the fact that there is a lack of an integrated platform for collecting and analyzing traffic data; the interoperability of the current tools is not satisfiable (i.e. transferring data from one tool to another can be difficult, if not impossible).

Needs

With respect to the data, participants mentioned a number of needs that remain unfulfilled:

- Having data independent of the private sector, so that it can be cross-checked;
- Having more complete historical data, mostly related to contextual variables that would allow better data interpretation;
- Information on the method of data collection, the sample sizes and the accuracy of the estimates;
- Real-time data for managing road closures or construction sites.

The participants summarized the ideal tools for transportation decision-making as being easy to use, flexible, and multi-functional. They should include contextual variables and allow importing/exporting to ensure interoperability with other tools. For example, P1 emphasized the importance of having data to differentiate congestion types (e.g. caused by accident or weather condition). The tools should also provide information on the quality, availability, and accuracy of the data.

Data visualization is regarded as an important feature in such tools. Participants mentioned that they needed tools to visualize summary information to facilitate data analysis and to speed up reporting. For example, P11 mentioned: "*We often need to do repetitive things - that is, reports for the month before, the month after, the week, whatever ... So it's important to have a visualization that is quick to do, attractive and informative.*" Direct manipulation when analyzing and visualizing the data is also a common need, as P1 mentioned: "*What would be interesting is to have access to a heatmap, for example, according to the hours of the day and different points on the network, with a color representation of the traffic conditions with precise speeds.*"

3.3.2 The Personas

Informed by these themes, we created four personas that captured different roles of practitioners within the transportation decision-making agency: (1) **Mona**, the modeling expert, who focuses on the creation and analysis of transportation system models; (2) **Tom**, the transport planner, which focuses on the analysis of the impacts of network intervention requests; (3) **Sam**, the safety expert, who focuses on various aspects of road safety including signaling and speed limits; and (4) **Caroline**, the coordinator, who supervises various projects and conducts opportunity studies. An overview of the personas can be accessed at: https://github.com/HCDLab/TDMPersonas.

The following are the images of the four personas created through the interview study that captured diverse roles of transportation decision-makers, as well as their associated practices, challenges, and needs.

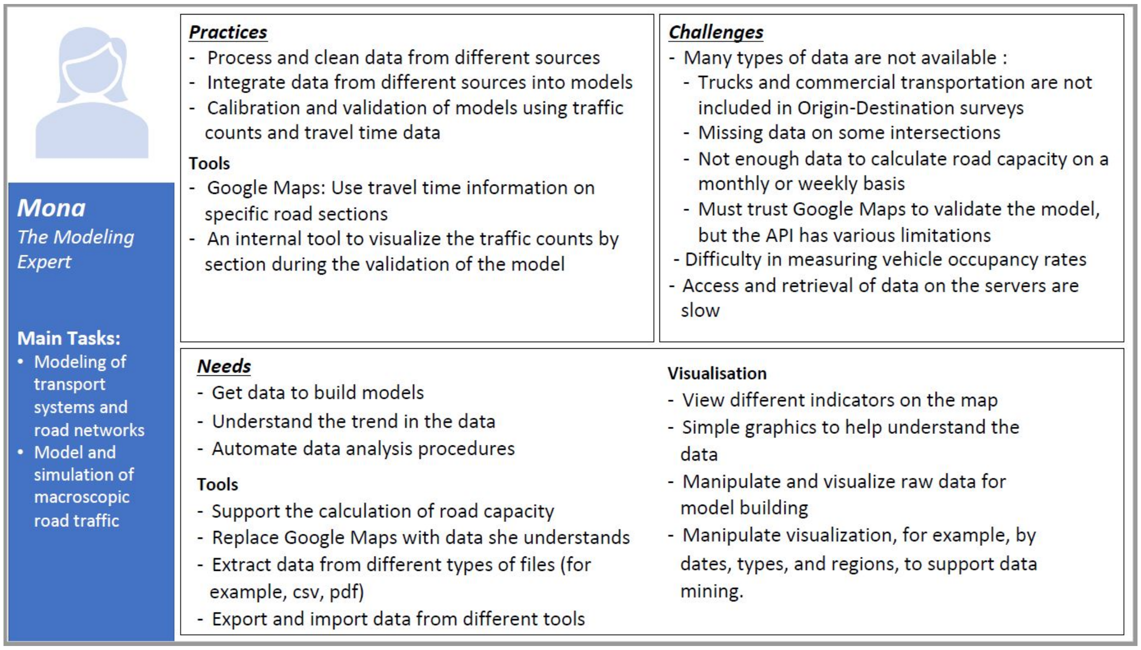


Figure 3.2 Mona, the modeling expert persona

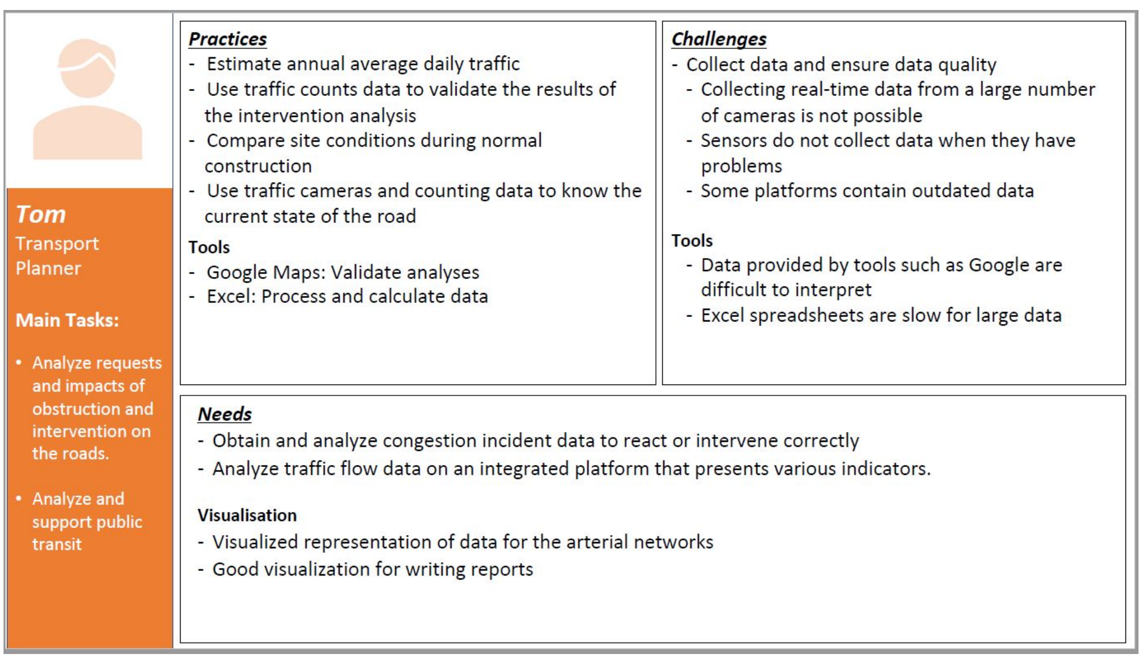


Figure 3.3 Tom, the Transport planner persona

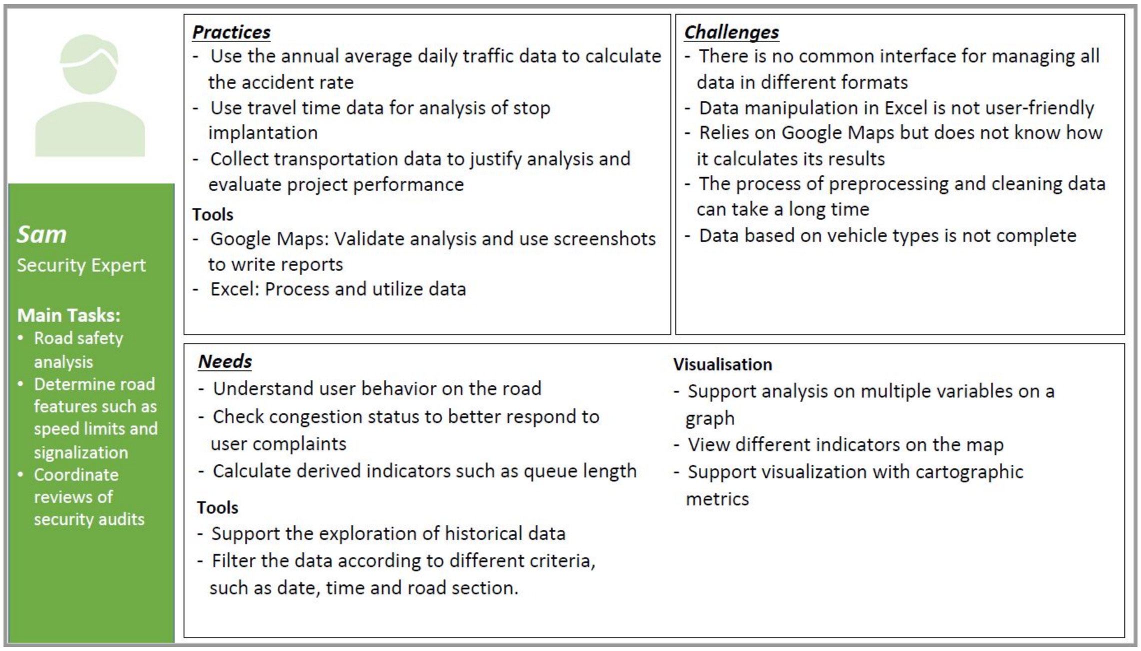


Figure 3.4 Sam, the security expert persona

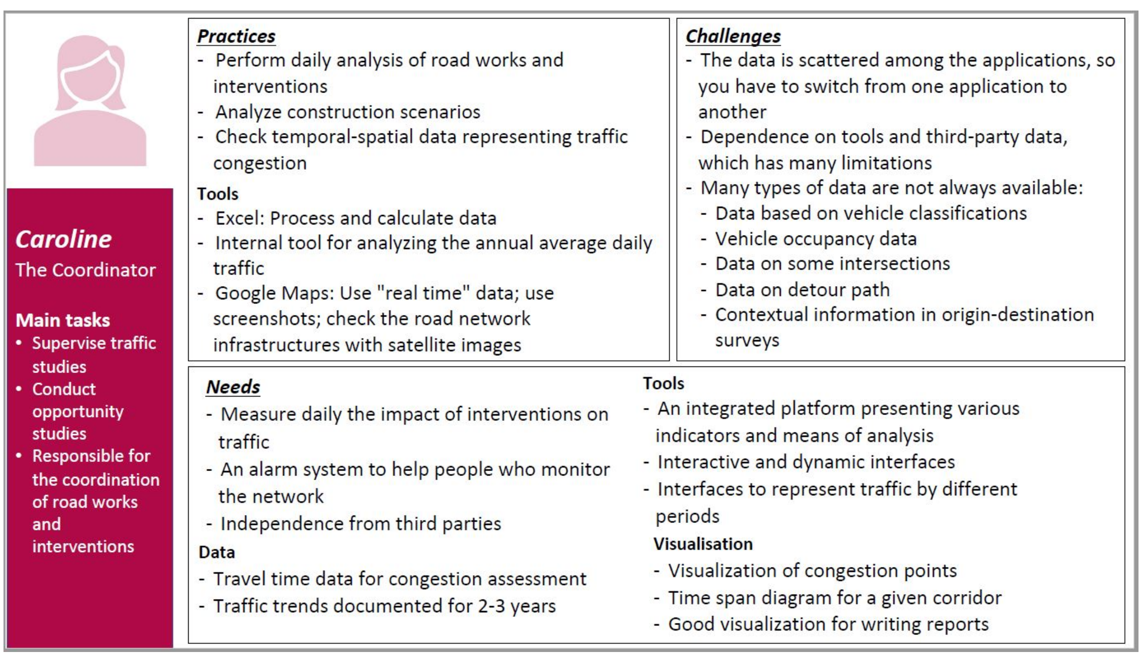


Figure 3.5 Caroline, the coordinator persona

3.4 Discussion

Through this interview study, we identified several opportunities for designing data analysis and decision support tools in the context of transportation management and planning. We are currently in the process of creating such a tool. Because many of these takeaways concern tool support for understanding data quality and interpreting data, they may also be informative to the design of tools supporting other decision-making tasks and domains.

3.4.1 Support Quality-Aware Data Analysis

One of the main takeaways in our studies is that traffic decision-makers frequently face the problem of determining the quality and reliability of their data. While their decision-making work requires an understanding of data quality, such information is usually not evident. This problem manifested in multiple aspects. The data gathered from third-party platforms usually do not contain information about data quality. The participants also frequently experience missing data, which resulted in varied data quality in temporal, spatial, and contextual dimensions. The tools they currently use also do not provide sufficient support for analyzing data quality and reliability. As a result, participants needed to spend a lot of extra effort to validate and clean their data. In situations where they are not or can not get informed about data quality, their decision-making process and outcomes will be largely affected.

This challenge can be addressed from several angles. First, quality-aware data analysis tools need to be developed to support the practitioners to understand the reliability of their decision-making outcomes. Particularly, techniques for representing uncertainty in data and analysis results need to be employed and adapted to support the work of this group of decision-makers. Second, updating and expanding the data collection infrastructure can partially address the missing data problem. Collaboration with private sectors (e.g. taxi companies) and research institutes can also increase data availability.

3.4.2 Support Interoperability Among Different Analysis Tools

Our participants reported that they leveraged data from diverse sources in their work. As a result, they often had to use multiple tools to collect and analyze data. A big problem they experienced when using these tools is the lack of interoperability. In other words, it is difficult to switch among different tools, transferring the data and analysis results. Context loss is an important effect reported by our participants when they used multiple tools. Dependence on third-party tools and platforms exacerbated this problem.

To alleviate this problem, tool makers need to better consider the workflow of the transportation decision-makers, supporting them to easily switch among tools. Additionally, contextual information created and used during data analysis also needs to be stored and be transferable. For example, tools can be designed to record the traces of the steps involved in data analysis that eventually lead to a certain conclusion. Such information can support transparency and communication about the decision-making process.

3.4.3 Supporting Quick and Informative Data Visualization

Our participants have mentioned the importance of data visualization in multiple contexts. Particularly, they valued both bite-sized visualizations to quickly explore the data and more sophisticated on-demand visualizations based on their immediate temporal, spatial, and contextual interests in the analysis. In addition to visual analytics purposes, the practitioners also required intuitive visualizations for communicating their analysis results with other stakeholders in the agency, as well as with the general public.

Designing a customizable visualization system, based on the tasks and roles of the different types of practitioners (e.g. based on the personas we created), would then be a useful and interesting next step. Because of the diverse tasks that this group of practitioners performs with visualization aids and their unique characteristics, balancing flexibility and ease of use is an important factor. Providing an appropriate level (i.e. sufficient and not overwhelming) of control during visual analytics and visual representation would be an important and challenging future work.

The results obtained from the interview study of this chapter were later used to design a traffic condition analysis tool for transportation decision-makers (See Chapter 4) and also inspired us to explore data uncertainty in transportation data and investigate data uncertainty visualization techniques that can support the work of transportation decision-makers (See Chapter 5).

CHAPTER 4 USER INTERFACE DESIGN FOR THE TRAFFIC CONDITION ANALYSIS TOOL

This chapter focuses on the user interface design for the traffic condition analysis tool. We leveraged the results of our interview study with 19 transportation decision-makers, mentioned in the previous chapter, to design this tool.

4.1 Lessons Learned From the Interviews and Design Guidelines

From the interview with the 19 practitioners at MTQ, we have noted the following points about their needs in a new tool for transportation decision making.

1. **Intuitive and Flexible Filtering**. As a basic requirement, all participating transportation decision-makers expressed the need to look into data from specific days, months, years, and hours on different segments to analyze and validate their results. On the road network, a road segment is considered a specific representation of a portion of the road having it's own unique characteristics such as length, location, etc. Participants have also described various scenarios where filtering is needed. Some examples include wanting to show the data between two selected dates, choosing data for weekdays, or weekends, selecting data based on seasons, and showing average speed data calculated every 15 minutes, 30 minutes or every hour of the day. As a result, we included a filtering feature in our tool that gives the users the flexibility to select their desired data for analysis.

2. **Support Comparison of Data**. Participants have also mentioned the need to frequently compare their data from different sources to validate data accuracy when using them to make decisions. They also frequently compare the current data with historical data of the same source to track trends and to identify any potential issues in the data. To fulfill the need for frequent data comparison of the practitioners, we would allow a side-by-side display of different visualizations of selected data on one screen. This would give practitioners the flexibility of being able to perform comparisons more easily and efficiently.

3. **Support Both Analysis and Reporting**. Participants usually need to do a lot of report writing in their work. As a result, they want visualizations that are quick to create, attractive, informative, and visualizes summary information accurately to facilitate data analysis and to speed up reporting. To simplify their report writing process and considering the importance of communicating their analysis results with other stakeholders in the agency, we have considered these needs when creating visualizations and we would include the option

to download and export the charts and data in our tool.

4. **Customizable Workspace**. Since the practitioners have different roles and needs, they want a tool that allows them to customize and use their required features of the tool based on the particular tasks of their projects. As each practitioner might have different needs based on their tasks and projects, we would provide features to allow users to customize how they want their data to be presented by choosing the appropriate type of visualizations and charts.

5. **Interoperability With Other Tools.** Our participants also expressed needs to incorporate the various features of different tools they use for analysis to avoid frequent switching of tools. This would improve efficiency and prevent context loss in the analysis and data transferring process for the decision-makers. They require the tool to be multi-functional and give them the flexibility to be able to import their existing data from different sources and also export to different tools if needed for further analysis. To achieve this, we would allow them to import their existing data and filter it to visualize their needed data.

6. **Example-Based Interaction.** Also, we are aware that not all practitioners have expertise in using such a tool. Particularly, some participants expressed that, although they felt that the tool could be useful, they have little experience in visual analysis tools and do not know exactly what to expect from it. For these participants, demonstrating the possible features of the tool in an intuitive way is important. So instead of having a wizard-style page that users start by applying filters to data, we decided to have an example-based interaction on the first page, showing some visualizations of filtered data so the users can have an initial idea of what they can do and what to expect. In addition to this, a project cloning option is also added so new users can explore different filtering and visualization options of project examples. Experienced users can also use this cloning feature to make changes, either big or small, to their existing projects without having to lose their previous version.

4.2 Design Process

After setting the design guidelines for our tools based on the needs of the practitioners, we conducted the user interaction design of the tool through the following process.

We first created the initial conceptual design of different features based on the design guidelines. Figures 4.1, 4.2, 4.3 show these designs. The main features we focused on during this step are (1) filtering where users could specify the type of data and the date and time and then select the desired location of interest on the map by searching and/or area selection (Figure 4.2); (2) comparing visualizations where users could compare two graphs side-by-side

(Figure 4.3); (3) data set selection where users could update or choose the data to use; (4) showing realtime videos from cameras located on some intersections (Figure 4.1); and (5) showing trends in historical data).

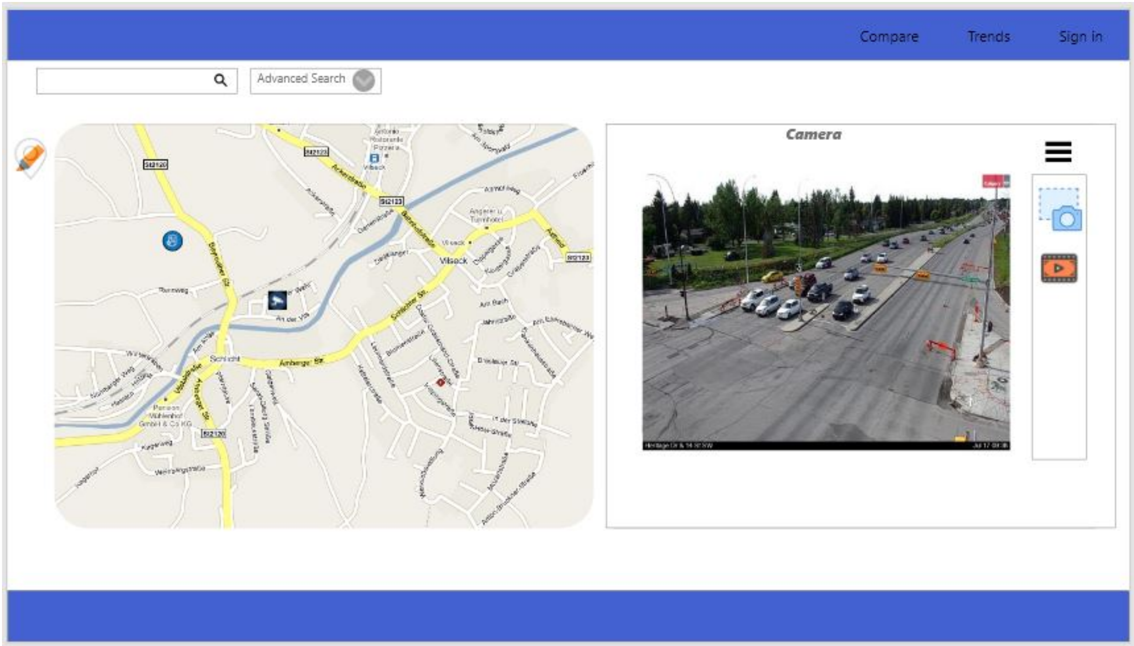


Figure 4.1 The map and live camera

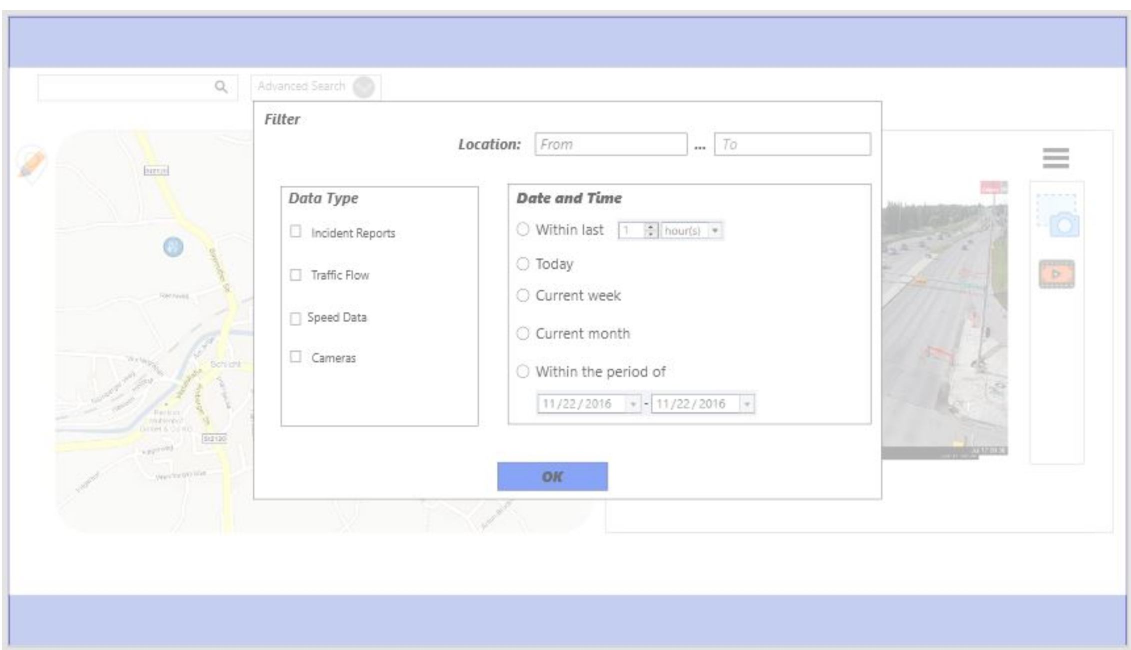


Figure 4.2 Data filtering options

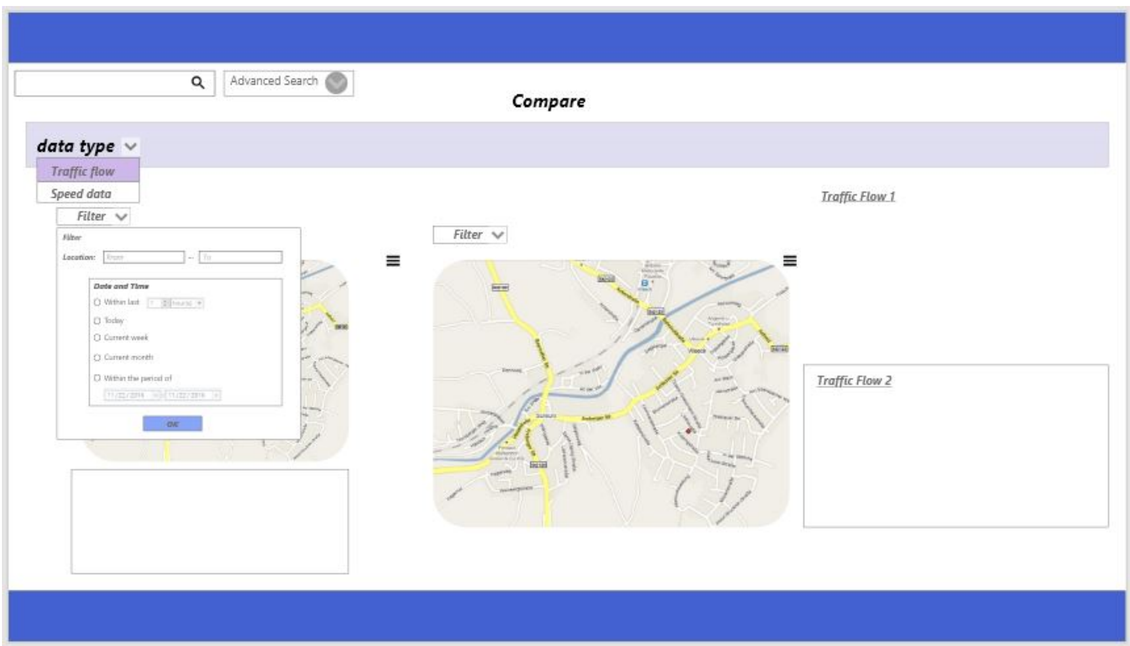


Figure 4.3 Comparison

In the second step of our design process, we held two main meetings and several smaller meetings with our team members in the transportation department to discuss the tool design. The goals of these meetings were to (1) develop scenarios of consultation interfaces to meet the needs of partners including main indicators and possible visualization objects, spatial and temporal resolution levels, possible aggregations, flexibility in the choice of ranges, and objects of analysis. 2) Define the architecture of the data tables and the storage methods, for example, the data describing travel speeds and times meaning what kind of data is recorded, at what level(spatial/temporal) and for how. This part also contains identifying explanatory and contextual data, the format of the data, its contribution, and how it can be used in the interface. In addition, we discussed what will be required by the MTQ where data storage is also done.

Through these meetings and also the needs of the practitioners from the interviews, we identified the following top 10 data visualizations for our tool design. Note that these visualizations are all dependent upon the selection of date/time and traffic network segments.

1. Average speed for each hour of the day, with indications of interesting percentiles and the number of observations.

2. The ratio of speeds for each hour of the day, with indications of interesting percentiles and the number of observations. The speed ratio is calculated by dividing average speed by the speed limit and the number of observations for a certain speed in a time frame is the number of vehicles that have been reported to have that speed at the same time frame.

3. Distribution of average hourly counts of vehicles for weekdays, weekends, and holidays. Traffic counts are the number of vehicles passing by a certain road or intersection. It is undertaken either automatically by electronic traffic count devices or recorded manually on handheld devices by human observations.

4. Statistics about road incidents, including accidents, road surface conditions (e.g. snow or rain), signal conditions, etc.

5. Speed ratio represented as a heatmap that covers a range of time and space.

6. A map indicating traffic conditions based on speed ratio and average speed.

7. A map indicating travel time and variability by Traffic Analysis Zone (TAZ) and validate the number of observations. Traffic Analysis Zone is the area described precisely by local transportation officials for organizing traffic data and is used in transportation planning models.

8. Propagation of the congestion. This chart shows the spread of congestion on a series of connected roads and how the road segments are impacted by congestion as in how many

minutes/hours in a day was congested.

9. An estimation model showing the probability of congestion for different time frames of the day considering different forecast variables such as good visibility, snowfall, icy pavements.

10. Highlights and trends (highs and lows) for the events. This visualization shows at what times of the day the traffic is high or low in correlation with different events such as road closures, construction, etc.

We further discussed what other features we can add to improve our tool.

1. *Tabs*: Practitioners usually work on several projects at the same time and it becomes very frustrating for them to constantly switch between different tools and projects to do their analysis. To give flexibility and increase focus on working with several ongoing projects, we decided to add tabs so users could create different projects and work on them at the same time.

2. *Project duplication and save button*: We realized that decision-making could be quite a long process and every time the practitioners have to put a lot of energy to bring up the data needed for their analysis to work on. We wanted to help users speed up their analysis process by being able to save their projects in our tool and open it back up to exactly where they left off with the visualizations. Therefore we added a save button so they could save their project to work on later. Also, we considered adding a project duplication and project saving buttons where users could duplicate a project to increase efficiency and save time when they want to make some changes to an existing project.

3. *Number of visualizations per screen*: In addition, the number of charts and visualizations per screen was another focus of our meeting. Having several visualizations on one screen can help with data comparisons and also to clearly have different data presented for a certain section. We wanted to give users the option of having different visualizations on one screen but at the same time, they should be able to read and use the visualizations for them to be useful. Considering the size of the visualizations we planned for a maximum of 4 visualizations per screen, the users can maximize any of these visualizations if they find it too small to view the details.

4. *Visualization download button*: The practitioners are constantly involved in report writing and sharing there analysis and reports with their colleagues, therefore we added a download option to the visualizations that they create in our tool so they could use it for their reports.

5. *Segment selection on map*: Transportation decision-makers are quite familiar with the geographical location of the data they are working with. Having that in mind, they prefer to visually pick their selected road/segment on a map and then see the available data for it. In our traffic condition analysis tool apart from being able to specify a location by typing the

name of the road/segment, another option we added was segment selection. Users are able to select one or several segments and have details of the location data, speed limit, average speed, length, etc visualized either on the map or on other visualization types.

6. *Fixed filter box*: Next the design of the filter box was mentioned. Users would constantly need to make small changes in data filtering to review different scenarios of data to make decisions. Also having several visualizations might confuse the users on what filtering options they have chosen to create each one. So instead of having pop-ups to present the filtering options and having them disappear after the user finishes the data filtering process, we decided on having a fixed filter box. This would avoid any confusion in the projects, the users could always take a look at what data they have filtered out when having doubts in their resulted visualizations, and also make changes faster.

We created the following mock-ups of the traffic condition analysis tool based on the discussed features (see Figure 4.4, 4.5, 4.6, 4.7). These mock-ups were used as a reference by other team members of this project to help implement the the tool.

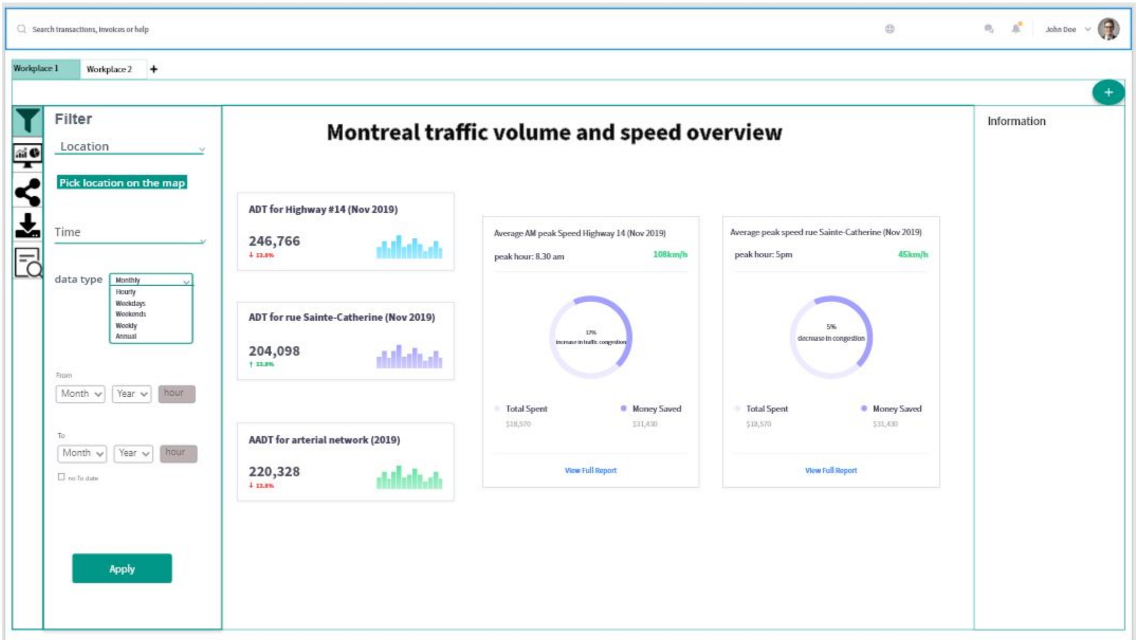


Figure 4.4 Mock up of the first page of the tool

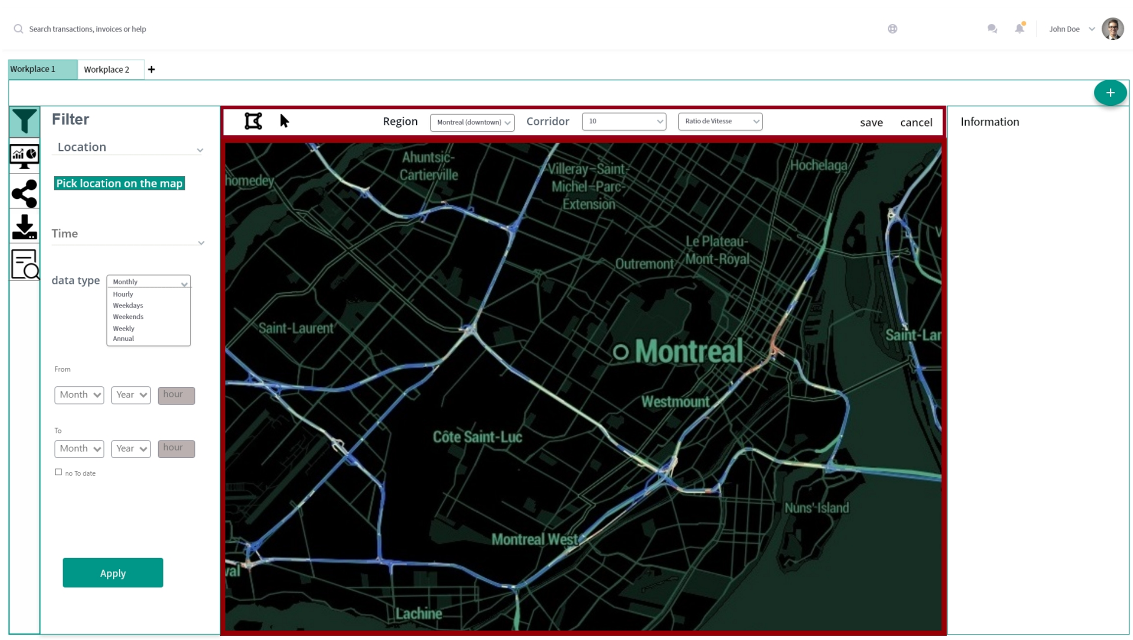


Figure 4.5 Map and filtering section

Figure 4.6 Adding new visualizations

Figure 4.7 Representations of having several visualizations on one screen

4.3 The Current Version of the Design

After several rounds of discussions and iterations, the current design of the tool which was programmed and implemented by another team member is of the following (see Figures 4.8, 4.9, 4.10, 4.11). Note that this tool is currently still under implementation and will be modified and improved in future versions.

When users log in to the system, they will first see a blank screen with an add project icon (1 in Figure 4.8) which gives the users the flexibility to create as many projects as they need for their analysis. A project in our context contains the analysis of traffic conditions on highways, main roads for different regions of the city. Users can also delete an existing project (2 in Figure 4.8) or duplicate an existing project (3 in Figure 4.8) if needed. The projects will be shown on the left of the screen as tabs (4 in Figure 4.8). After creating a new project they will see plus icons (5 in Figure 4.8) on the screen. By clicking a plus icon, users can visualize their data by creating new visualizations for their analysis. They would then see the visualization selection menu (6 in Figure 4.9) on the right. In this section, users have the option to choose their visualization type based on their needs and tasks. Once a visualization type is selected, a new dashboard appears on the right (7 in Figure 4.10) that consists of the project name (8 in Figure 4.11) where users can name their projects, and selection of the number of visualizations (9 in Figure 4.11) that they want to have presented on a single screen. In this tool, the users can select up to 4 visualizations to represent data on a single screen. Once the type of visualization is selected, users can see more detail on the location selected on the map and the segment that they want to analyze (10 in Figure 4.11). Then the period picker (11 in Figure 4.11) filter allows users to select a date/time duration on the calendar, pick a special day of the week, weekends, or weekdays, pick a time period and also choose a time interval which for example will show the recorded data for every 30minutes, or 1 hour of that time period. After selecting all the required options, the user can submit the request by clicking on the submit button (12 in Figure 4.10). Then the selected data will be presented on their selected visualization. The users could also modify a visualization by first clicking on the visualization on the screen and then change the filtering options. Their modifications will then be finalized when clicking the submit button and therefore the results will be updated on the visualization. If they already have created their visualizations but want specific visualizations to appear next to each other, they can swap their locations by dragging the visualization to their preferred place on the screen.

Figure 4.8 First screen of tool

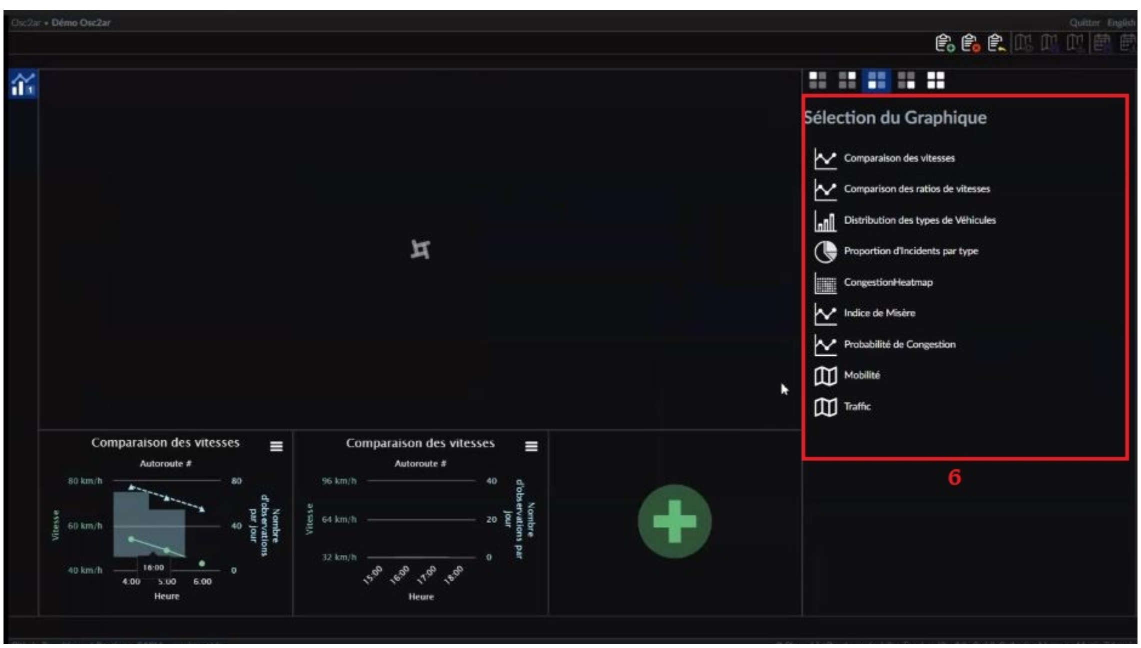


Figure 4.9 Visualization selection

Figure 4.10 Maximum of 4 visualization

Figure 4.11 Map selection

4.4 Users' Feedback on the New Design

We have conducted a short user study with six practitioners from MTQ who have participated in our original interview study to collect user feedback. The studies were conducted over Zoom and each study took about 20 minutes to complete. During the study, we showed screenshots of the traffic condition analysis tool and explained a scenario of how a first-time user would start creating a project and using the options to create visualizations for their analysis. Then we asked the participants to let us know a few positive and negative points about the design and the features of the tool. Finally, we asked participants to give us any suggestions on what features they would like to be added to make the tool more useful and efficient for them.

All participants expressed satisfaction with the tool design. For example, P3 said: 'At a glance, the design seems very modern and gives a lot of flexibility to us'. Participants also mentioned that it is nice that there are several charts and visualization options to choose from, as depending on their roles they have different needs. They were interested to see maps and also to choose from different time periods depending on their analysis.

The following are the main categories derived from the participants' feedback on our traffic conditions analysis tool.

Usefulness in Workflow: The practitioners have mentioned how visualizing their data in such a tool gives a better and visual understanding of their data, and will definitely also speed up their analysis. More particularly, practitioners specialized in transportation planning and modeling who work with the impact of dynamic traffic and regional analysis have said that this tool can provide additional sources of information about the traffic that is not usually easy to retrieve. One participant stated "It is great that we can understand the causes of traffic demand, such as the events, incidents, the weather conditions like when it's raining or snowing. These parameters will all help us in our analysis to make decisions and also simulate such situations so we can be aware of what kind of service we can provide to people using our network". Also, having these options without switching tools seemed quite interesting and efficient for the practitioners as it will eliminate the frustration of the process and help them compare their different results in a single user-friendly platform.

Features and Functionalities: Although having several ongoing projects to work on all in one tool seemed very interesting and helpful, not knowing the limit towards how many tabs and projects they can have may cause confusion for them and they would also want to have symbols to show which projects are currently active.

The practitioners were glad to see that they have different options for visualizations to choose from as different people have different needs to do their analysis. Between the different visualizations, having a map seemed to be a special option as being able to find a segment on a map is much easier and the option of selecting segments and have the data for that section presented on the map will be a great help for them.

Another feature that they found helpful was the ability to download the visualizations with their selected data as image files, so they could save them to either look back later or use them for the reports.

Also considering that the data they work with has some uncertainty, one participant suggested having a checkbox that allows them to know what the uncertainty of the data from that source is. To explain more he then added, "If I take one source, maybe the level of uncertainty is high for it but if I have another source with less uncertainty I will use that one for the time range I selected and I will be more comfortable with using the data from that source and doing analysis with it."

We also asked the participants what other data they would like to see visualized so we could later add more visualization types to the newer version of our tool. Two participants stated that they would look to see the percentage of vehicles and trucks on the road, meaning what percentage of the total vehicle counts were trucks and what percentage were regular vehicles. As trucks take more space on the road, this data would be very helpful for them to analyze traffic conditions. Also, they thought showing the average speed, counts, and number of observations all in one chart would be interesting.

Filtering Dashboard: Practitioners were very interested in the fact that they are able to choose many days instead of a single day and different time periods too for their data to see the visualizations. One participant wanted to know a little more detail of how the data is aggregated and was concerned how the data of several days is all shown on one chart and stated "We have the number of observation for one day shown on charts and files, but what if we select several days? we would like to know what is being shown on charts and the calculations behind it.". She later suggested, "Maybe there could be a chart for speed data of all the days selected and the number of observations would just be shown for a specific day so we could see if there is a day that doesn't have any number of observations, and if so we would like to have that information."

About the date and time selection, some participants suggested adding more options in the date selection. One participant mentioned, " Sometimes we analyze data for just weekdays or just weekends or for examples all Mondays of a season, so it would be great to have that kind of option".

User Interface Layout: One participant said that having the activity bar (filter box) fixed on the right of the screen wasn't what he is normally used to seeing but he surprisingly prefers it designed this way. In contrast, another participant thought that the filtering section was taking too much space and he preferred to have this section minimized to have more space for the visualizations. Also they found having several visualizations on a single screen very interesting and said it could definitely help them see the data they have more clearly now that they are all together.

Visual Design: In terms of visual design, some practitioners stated that they liked the black theme of the tool, while some others found it hard for the eyes. A few practitioners mentioned that the font is too small and sometimes it is hard to see the details of the charts because of this. Collectively, participants suggested having an option where the user can customize the font, scale, and the background colors of the tool. They stated that the icons perfectly reflected their actions.

We are currently incorporating these comments to improve the traffic condition analysis tool.

4.5 Discussion

The positive feedback of our user study shows that all practitioners were interested in our traffic condition analysis tool design. Their feedback also confirms the usefulness of the design guidelines we considered based on their needs in a traffic condition analysis tool (from interview study results). These guidelines included intuitive and flexible filtering, supporting comparison of data, supporting both analysis and reporting, creating a customizable workspace, and supporting interoperability with other tools.

The designed traffic condition analysis tool can help decision-makers in several ways. More specifically, the visualizations can help communicate large datasets in a more visual and easier way, speeds up the decision-making process for practitioners, and overall lead to better and more accurate long term decisions on the road network.

Although we covered most of the practitioners' needs in our interface design, we didn't consider some needs such as realtime data, video coverage of intersections. At the time, we had two main reasons that lead to this decision. First, there were data collection limitations and therefore limited data to present in our tool. Also, the architecture of this tool and the storage limitations didn't allow such a feature as it would slow down the tool.

We will use the comments that the practitioners gave to improve our traffic conditions analysis tool. For future research, when the implementation of this tool is completed, another user study can be done to understand how practitioners use this tool for their analysis, evaluate the effectiveness of the tool, and also the decisions made by this tool.

CHAPTER 5 DATA UNCERTAINTY QUANTIFICATION AND VISUALIZATION

5.1 Introduction

We have understood from our interview study in chapter 3 that the practitioners are not completely confident about the data that is provided to them through different sources. More specifically, they have mentioned that counts, speed data, and the number of observations that they have for their analysis may not be accurate and they need to verify them by either looking at the same type of data from other sources or comparing it with historical data. In fact, almost all transportation data have some uncertainty. It is thus extremely important for decision-makers to be aware of the level of uncertainty in the data they work with. Ignoring this fact could create a lot of difficulty in their decisions and create bias in their analysis. However, there is little knowledge that is established about how transportation decision-makers deal with uncertain data and how they would make decisions based on their awareness of the level of uncertainty in the data. Additionally, visualizations have always been a great way to efficiently communicate data. In our interviews, transportation practitioners have also mentioned the need for visualizing their data to help them with their decisions and analysis. Therefore, visualizing uncertainty in data can be a technique to quickly let the practitioners know the level of uncertainty in their data.

Therefore, in this chapter, our goal is to investigate data uncertainty visualization techniques that can support the work of transportation decision-makers. Particularly, we first examine a novel method for calculating the level of uncertainty for transportation data. This method, as we will see in section 5.2.2, is inspired by the fact that transportation practitioners often use historical data to assess the quality and uncertainty of the current data. We then apply data uncertainty visualization techniques to visualize these different levels of data uncertainty on top of the regular data the transportation decision-makers are using. These visualization techniques were examined in a user study that focused on not only evaluating the practitioners' perceptions and preferences in the various visualizations but also on exploring how transportation practitioners consider and use information about data uncertainty in their decision-making process.

5.2 Methods

In this section, we first describe our novel method for uncertainty quantification of transportation data. We then present the various visualization designs we created using this quantification method and the user study examining these visualization techniques.

5.2.1 Quantifying Uncertainty

Transportation decision-makers often work with average speed data in their traffic conditions analysis process. In our study we define data uncertainty as the difference of the current data compared to a larger amount of historical data on the same time period. In order to calculate the levels of uncertainty in average speed data that is recorded for every hour, we considered evaluating the number of observations for speed data of every hour for each highway segment.

While having high numbers of observations for an average speed may demonstrate higher certainty to the speed data for decision-makers, the opposite case is not always true. Meaning that lower numbers of observations don't necessarily result in high uncertainty in data, because the number of observations can be affected by factors such as the geometry of the road, and the time of the day or the season. Hence, the number of observations depend on the variability of the number of cars passing and this remains a challenge for decision-makers. Therefore, transportation practitioners usually rely on other sources of speed data to understand the uncertainty in their data.

Inspired from the fact that many of our interview participants used historical data to validate and evaluate the uncertainty of their current data, we propose the use of Jensen-Shannon Divergence (JSD) as a technique for quantifying uncertainty in transportation data. Intuitively, this technique measures the difference between two probability distributions (in our case, the distribution of the historical data and that of the current data), leveraging the concept of entropy.

In information theory, Shannon proposed entropy as a measure of disorderliness and uncertainty in data [52]. Shannon entropy (H) for event X with n possible outcomes and probabilities p_1,...,p_n is defined as:

$$H(X) = H(p_1, ..., p_n) - \sum_{i=1}^{n} p_i log p_i$$

Shannon entropy is widely used to quantify information because it satisfies certain important criteria. Particularly, the Shannon entropy reaches maximal for uniform distributions. It is also continuous, non-negative, and additive for independent events (H(x,y)= H(x)+ H(y)).

For our study, to calculate the difference between two distributions, we first leveraged the concept of Kullback-Leibler Divergence (KLD). KLD, also known as relative entropy, measures the difference between two probability distributions. Considering p(x) and q(x) as two different probability distributions over X which is a random variable, the KLD or relative entropy D(p||q) is shown as below:

$$D(p||q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)}$$

KLD is a popular metric used to quantitatively measure the distance between two probability distributions. This metric has been used in various domains such as machine learning [53] [54] [55], transportation and traffic congestion [56, 57], Biology [58, 59]. For example, KLD has been previously applied to healthcare applications using machine learning algorithms [60]. Since usually large data is handled through machine learning applications, KLD helps the users to keep track of reality by identifying the differences in data distributions. In the transportation domain, Chan et al. (2005) modeled traffic flow in videos and encoded the underlying motions separately into two distributions. They used KLD to measure the distances of these probability distributions; their distance measures were then used to retrieve traffic videos similar to a query [56]. In another approach Zeroual et al. (2018) detected and estimated abnormal traffic congestion by using KLD to measure between probability density functions of congestion-free residuals and actual residuals of their models; where a KLD closer to zero indicates a congestion-free scenario and a significant difference from 0 indicates important changes from normal conditions [57]. .

Although KLD is always non-negative, it is not symmetric, meaning D(p||q) does not always equal D(q||p). It also does not have an upper bound. As such, it is difficult to directly use it to represent the different levels of uncertainty in the transportation data.

Therefore, we decided to use another measurement of uncertainty that measures the difference between two probability distributions called Jensen-Shannon Divergence (JSD) which is derived from Kullback-Leibler divergence and is a smoother and bounded version of it. Contrary to KLD, JSD is a symmetrical function and is bounded between zero (when two distributions are identical) to one (when for all possible values of the variable, the probability in one and only one distribution equals zero). The formula to calculate the Jensen-Shannon Divergence is:

$$JSD(P||Q) = \frac{1}{2} D(P||M) + \frac{1}{2} D(Q||M)$$

Where P and Q are the two probability distributions, M is defined as $M = (P + Q)/2$, $D(P||M)$ is the KLD between P and M, and $D(Q||M)$ is the KLD between Q and M.

In previous research Rhun et al. proposed different energy consumption models based on the aggregated data to estimate the energy consumption of the vehicles in the road network. They used JSD to compare the estimated energy consumption distributions to the measured ones [54]. Murakami proposed a method to model a probability of being located in each region, and then re-identifies traces(time-series location data) based on the JSD between two probability distributions for each user [61]. Also in modeling real-time human mobility that combines the advantages of transportation data and mobile phone signaling data, JSD is used to measure similarities between probability distributions of users' location [62].

In summary, JSD handles the limitations of KLD and creates a bounded and smoother measure to evaluate the difference between two probability distributions. That is why we chose to use JSD to quantify the uncertainty of data in our study. Particularly, the uncertainty of speed data is measured by calculating the JSD against historical and estimated data distributions. For example, to determine the uncertainty of speed data for a specific date and hour, JSD calculates the differences of distributions of the number of observations of the average speed of that day and the historical distribution of the number of observations for that day.

5.2.2 Dataset

The dataset we focused on for our visualization includes data from the taxi register, set up by the Montreal taxi office, which collects GPS data from nearly 4,000 car taxis, and whose data extraction is automated.

For the purpose of analysis, we extracted the data for highway 40 of Montreal; we have chosen Highway 40 also known as Metropolitan Autoroute within Montreal, because it is one of the busiest highways in Quebec and also, one of the main focuses of transportation decision-makers and their traffic condition analysis. Data for all hours of weekdays of February 2020 were extracted for highway 40. In this dataset, each datapoint represents information on a certain section of the highway called a road segment; we have derived these segments from the État des routes (ERT) system. ERT segments are quite long, between 500 meters and 46 kilometers. The majority of segments are between 5 kilometers and 20 kilometers in length. Also, for each datapoint, data is presented in different temporal granularity, such data recorded every 15 minutes, 30 minutes, and every hour. Each datapoint consists of information about the road segment number, average speed, date and time of the recorded data, the number of observations, the speed limit of the segment, and the length of the segment. Using this dataset, we focus on evaluating the uncertainty of the speed data.

For every segment, we have aggregated the number of observations for each hour of all individual weekdays together that were within the month of February 2020. Therefore, we displayed the aggregated data as distributions of the number of observations within 10 speed bins. As an example, all the number of observations for hour 20 of all Mondays of February 2020 were aggregated. We applied this aggregation method for the rest of the weekdays and every hour of a weekday.

To understand the uncertainty of the speed data presented for every hour of a day, we wanted to compare the distribution of the number of observations of that specific day and hour, with its aggregated data distribution of all the data over the month on the same day of the week and the same hour. This would give us an understanding of how different the data presented for a specific date and hour is compared to what is normally recorded within the same day and hour of a month. We then applied JSD to make the comparisons and measure the difference between two distributions; this measure will demonstrate the level of uncertainty of the average speed data recorded for a specific day and hour. We concluded that the more similar the distributions were, the less uncertain the speed data for that day and hour is. The higher the difference between the distributions, the JSD would be closer to 1 which shows a higher uncertainty level for the speed data.

The derived results from JSD were all between 0 and 1. So next to help decision-makers have a clear understanding of these measures and easily identify different levels of uncertainty in their data, we create categories of levels of uncertainty. We put our JSD results in a histogram to see the distribution of the samples. We computed the mean JSD and used a standard deviation bell curve to create new categories based on JSD distributions. Hence, for our first category, we considered the range of ± 0.5(standard deviations) far from the mean because we wanted to keep the mean in the middle of a range instead of at the border. More specifically, the bounds derived were [-2.5SD to-1.5SD), [-1.5SD to -0.5SD), [-0.5SD to +0.5SD), [+0.5SD to +1.5SD), [+1.5SD to +2.5SD), [+2.5SD to +3.5SD]. As a result, the following 6 categories were extracted based on our mean and SD calculations: [0-0.04), [0.04-0.17), [0.17, 0.29), [0.29-0.42), [0.42-0.55), [0.55-1]. These categories are used for representing different levels of uncertainty for average speed data in our visualizations. Although we put our uncertainty measures into 6 categories, these categories need to be recalculated for each dataset individually, since JSD outputs will differ for each segment and the new mean and standard deviation will be used to identify the number of categories needed to present uncertainty levels and also the upper and lower bounds of each category.

5.2.3 Data Uncertainty Visualization Design

Now that we have a measure of data uncertainty, the next step is to visualize it for decision-makers in transportation. Based on the literature review we explored the use of three main visual dimensions to visualize data uncertainty: size, shapes, and color. We applied these methods on top of the original average speed data that is presented in line charts used in our traffic condition analysis tool mentioned in chapter 4. After a discussion with transportation experts, we also decided to apply these methods on stepline charts, as it shows a more precise view of the data for every hour.

The five resulted visualizations are presented in Figures 5.1- 5.5. In these visualizations, bigger shapes, thicker areas, and the red color show higher uncertainty in the data, while smaller shapes, thinner areas, and the green color represent lower uncertainty in data.

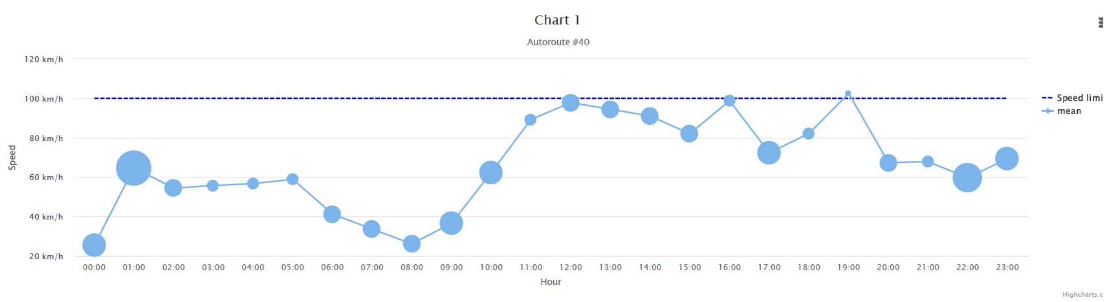


Figure 5.1 Using different sizes of shapes as a data uncertainty visualization technique

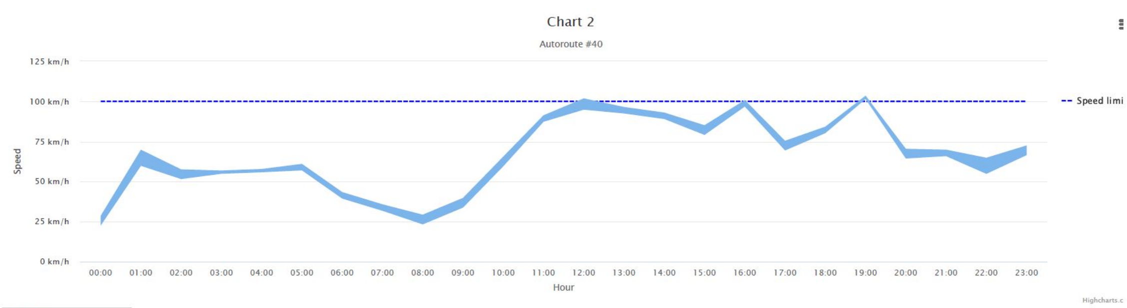


Figure 5.2 Using ribbon as a data uncertainty visualization technique

5.2.4 User Study

We conducted a user study with eleven participants: six experienced practitioners from Montréal working at Transports Québec (MTQ) as transportation decision-making experts, and

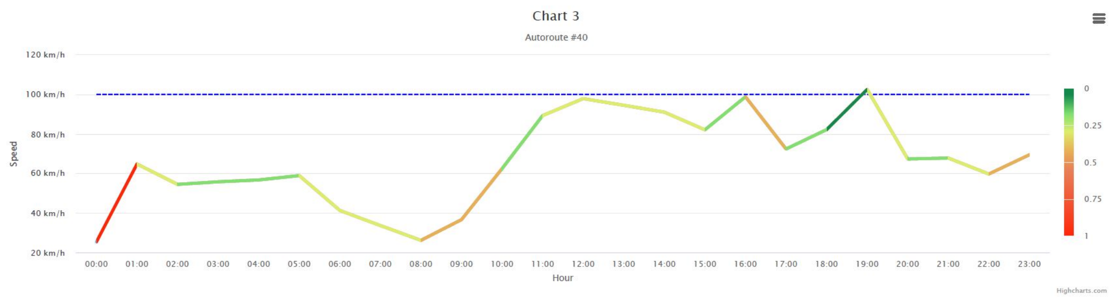


Figure 5.3 Using color as a data uncertainty visualization technique

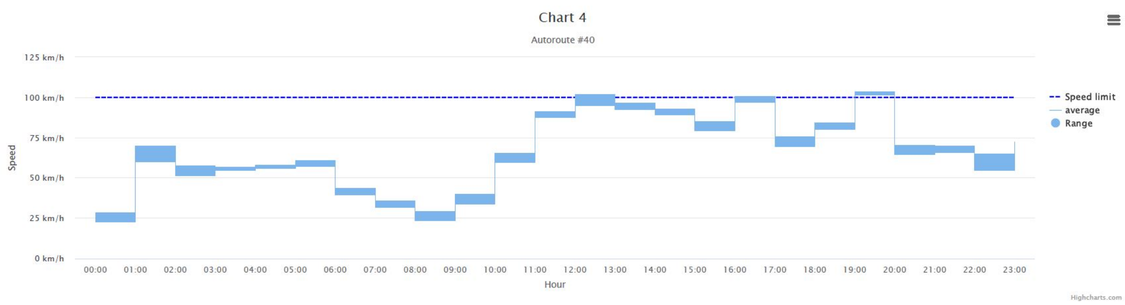


Figure 5.4 Using width range on a stepline as a data uncertainty visualization technique

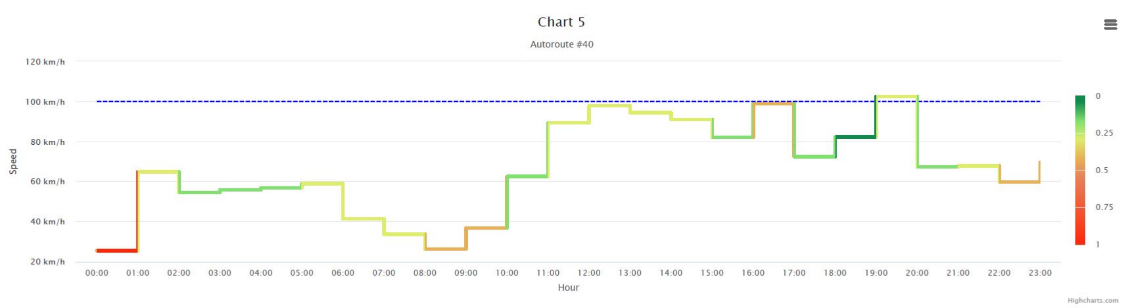


Figure 5.5 Using color on a stepline as a data uncertainty visualization technique

five students working on transportation projects as novice users in the area of transportation decision-making.

Our expert participants included five males and one female. Their professional experience ranged from two years to 28 years. The participants focused on different aspects of transportation management and planning, including (1) transportation modeling, (2) planning and sustainable mobility, (3) road network design and traffic analysis, and (4) managing construction projects focused on improving traffic conditions on main arterial roads.

The novice participants included two females and three males. Their experience with transportation projects ranges between one year to five years. Their projects vary from calculating indicators for taxi trips, evaluating the effectiveness of sensors installed in cars, using video recordings for safety analysis of cyclists and pedestrians on intersections, and developing dashboards for public transportation data analysis.

The interviews with these participants were conducted in June and July 2020 over online meetings; each session lasted approximately one hour. Each user study with the professional transportation decision-makers contained three main sections.

In the first section of the study we first asked the participants about their role and experiences, and the data they work with, in their projects. Then we showed them two line charts that represent the speed data of February 2020 from a segment on highway 40 of Montreal on every hour for Mondays and Fridays, respectively; the speed data is also overlaid with a bar chart representing the number of observations; see Figure 5.6 for an example. This is a typical visualization that the practitioners are used to working with. Looking at the two charts we asked the participants how they would assess the trends of congestion for the two weekdays and what additional information they would need to perform this assessment.

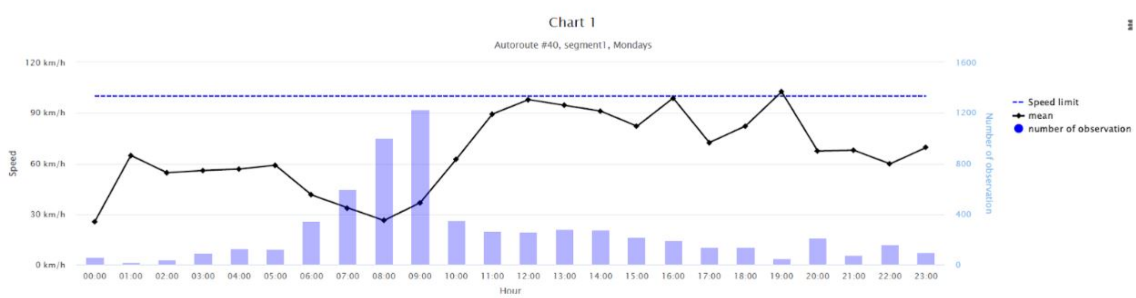


Figure 5.6 A line chart representing the average speed and number of observations.

The second part of the study is focused on data uncertainty in transportation. We asked the participants to describe the kinds of uncertain data they work with and the causes of uncertainty in their data.

The third section involves showing participants the five visualizations representing data uncertainty (see Figures 5.1-5.5) and asking for their opinions and preferences on these visualizations. In this section, we are interested in understanding how decision-makers perceive and can benefit from these representations.

User studies with novice users contained only the third section as described above since they do not yet have the experience assessing congestion or addressing uncertain data. For more details about the user study and the protocols, see appendix B. The user studies were recorded in video and then fully transcribed for analysis. We then coded the interview data to identify themes within the transportation decision-makers' responses.

5.3 Results

In this section, we first report the approaches transportation decision-makers take to assess the level of congestion. We then present the challenges they face while working with different sources of data uncertainty and their needs for addressing data uncertainty in their analysis. Finally, we provide practitioners' preferences and feedback on our data uncertainty visualizations.

5.3.1 Approaches to Assess the Level of Congestion

To assess the level of congestion, the practitioners required several data sources such as average speed, the number of observations, and vehicle counts for both directions of a segment. The separate counts for vehicles and trucks are considered as well because trucks take more space on the road; this information is used to calculate the density of the roads/highways, which later affects the congestion levels. Practitioners working at the MTQ all mentioned that speed less than 60% of the speed limit is often considered as congestion. So, to assess the level of congestion they pay close attention to where the speed is under that range and also how many hours in a day that area is congested. Another approach is to separate Mondays and Fridays from the rest of the weekdays, this creates a more realist analysis in congestion levels for typical and exceptional days of the week. Additional factors are also considered when analyzing congestion, such as the geometry of the road and slopes. Transportation decision-makers who are familiar with the structure of the roads and highways will know that at certain areas cars cannot go beyond a certain speed due to these matters, and the reason behind it is not congestion.

5.3.2 Challenges

The participants have mentioned several reasons that create uncertainty in the data they work with. Some were already mentioned as challenges in our previous interview studies but during this study, they provide more details on the origin of uncertainty in data and how it can affect their decision-making. The following are the main challenges of practitioners working with different sources of data uncertainty in the transportation domain:

Bias in Data Collection. Participants frequently mentioned the uncertainty resulted from the data collection process; many gave the example of the origin-destination (OD) surveys.

In Montreal, the OD surveys are done every 5 years through phone calls, mails, and online forms. The surveys gather information about peoples' trips from the day before either made by foot, car, bicycle, or public transportation and the responses help improve planning for public transit, road systems and also enhance urban development plans for the Greater Montréal. Our participants have mentioned that although they have methods to have more certain answers and results in OD surveys, they believe survey results may be biased. For example, a lot of people don't have landline phones or they don't always pick up cell phone calls. However, when someone does respond, they are asked to provide details of the trips made by other household members as well. While it is assumed that the data provided is true, the decision-makers are not sure of how aware the respondent is of his/her household member's trips. Another point is that it can be hard to remember the exact time of their trips, so people tend to give average answers or the time they used as a routine, which might not be exact. Also, the younger generation doesn't seem to be interested in participating in these surveys. So, there is a lack of respondents from this category. There is also a cultural bias in OD surveys and it is harder to reach certain communities that speak a different language than English or French. Moreover, people who use public transportation are often more cooperative and concerned in responding to OD surveys compared to people who own a car and drive. All these details have a great impact in the quality of the data that leads to uncertainty in the decision-making process.

Lack of Trust in Third-Party Data Providers. As mentioned by the participants, they always have uncertainty towards the data that is provided to them through third parties. In order to avoid this kind of uncertainty in the data, the decision-makers want to know in detail the methodologies used in the process of gathering the data. They require complete transparency in the data in order to make accurate decisions. For example, P2 mentioned that they would like to assess the source of data that third parties provide, saying "We must know how confident we should be on the data representing reality. Many companies sell information and they give us a black box and we cannot do anything with it. They have to be honest with how they collect the data and what sources they use. Sometimes they say they will give us data for a highway, and they use only a connected device installed on particular cars like Mercedes, BMW, which are high-level cars, this way it seems they provide a good sample with many observations but we can't say it represents all the drivers with different cars on the streets. The level of uncertainty in these kinds of data samples are not related to the mean speed or the number of observations, but the real issue is with the accuracy of how it represents different cars on the highway and not just a specific type."

Data Bias Caused by Unpredictable Drivers Behaviour. Transportation decision-makers find the unpredictable behaviour in drivers and cyclists creates uncertainty when analyzing the data they receive. As stated by P2, “weather conditions could cause an illusion for people that they should drive with slower speed or entering the highway with an inappropriate speed and that affects the speed data that is captured making it seem like low speed was a result of congestion.” The same goes for lane changes and how the drivers pay attention to the signs and follow road rules such as speed limit.

Human Errors in Data Input. Human errors in manual entries and missing data create uncertainty for participants. For example, P8 said, “There are lots of mistakes in my data. Lots of entries are done manually by drivers. They have to change the status manually, and sometimes forget to do it. If the taxi is free they have to say if it’s free, occupied, oncoming, or unavailable if doing a particular trip like to a hospital.” There are also problems with determining the beginning and end of the trips. It is mentioned by one participant, “There isn’t any information for when the client enters a taxi or to say that I now have a new client and when there’s a change. So, by not having this data it’s hard to say if it was a long single ride or one ride after another.” Similarly, P4 also mentioned, “although we know that the cars have passed a certain place, but we don’t know their starting point and destination to be able to further analyze the data.”

Device Errors Traffic data such as speed, number of observations, and counts are normally collected by sensors, satellites, GPS, cameras, and other devices. Data collected through these devices are accurate as long as the devices function properly throughout the year. However, our participants mentioned that there are always times that the devices stop working or report erroneous results. As a result, participants either end up with gaps in their data or uncertainty in the data provided. Moreover, when the decision-makers want to do their analysis on a certain section that doesn’t have the devices installed, they would have to look for the available counts data near that area and use that data to make their uncertain estimations. Sometimes the reason behind uncertainty in data is actually because the technology is not updated and older devices have lower precision in data gathering compared to the updated devices. Participants mainly mentioned that this results in imprecision in data related to location; e.g. P7 mentioned, “I work with GPS data but the location data is not sufficiently precise. The point they give us has a 200m precision. If you point on the map you can see the user is completely on another road.” Also in public transits that passengers use smart cards to enter, it is possible that sometimes not all of the data is recorded by the devices. This kind of missing data will leave gaps in the dataset and cause uncertainty for decision-makers.

5.3.3 Needs for Addressing Data Uncertainty

The practitioners are aware that there is uncertainty in all transportation data but they require a level of reliability and trust towards data providers and the quality of the data. This would include giving more details on the background of how the data was collected, what devices or methods were used, what data did they replace missing data with, and how they dealt with errors and limitations. By knowing the details behind the data that they are given and the level of uncertainty of the data, they would then decide on how to use that data, what part to use and compare with different sources and which parts to pay more attention to. If the level of uncertainty is provided, the decision makers want to know what is considered for uncertainty in their data and how it has been calculated. Having these kind of details in the data creates trust and reliability towards humans, devices, and data providers. In sum, when the decision-makers could make their decisions based on complete awareness of their data, these decisions would lead to more accurate modeling and planning.

5.3.4 Uncertainty Visualization Feedback

In general, the most popular visualization between the participants was the ribbon and stepline width charts (Figures 5.2 and 5.4) and the least popular graph was the color stepline (Figure 5.5). We now present the details of the participants' perceptions towards these visualizations.

Usefulness of the Visualizations

Participants have all given positive feedback on how the data uncertainty visualizations can influence their decision-making processes. They have mentioned that these visualizations could help them pay closer attention to data that has high uncertainty and data that is more accurate and precise. With this knowledge, they can also use data from other sources that have lower uncertainty to avoid any errors in their analysis. One participant mentioned the importance of visualizing data uncertainty when the data and the decisions are linked to security and safety measures, saying "If the uncertainty of data for a certain section and hour is really high and is linked to security decisions it can be a real problem and have severe consequences in the future." Also, these visualizations are said to be a great help for transportation modelers. Seeing the hours with the highest and lowest data uncertainty will help them consider certain accurate data in their model calibrations; and if the data is too uncertain they will know to discard them from their models and reports to increase the accuracy of their results.

Users' Perception and Direct Feedback on the Visualizations

When looking at the visualizations, all participants thought that the thicker sections on Figures 5.2 and 5.4 and/or bigger shape sizes in Figure 5.1 and/or the sections with red colors in Figures 5.3 and 5.5 have higher uncertainty. However, they had different guesses as to how we calculated the uncertainty values. Some were thinking we used variance and standard deviation to show the uncertainty in data.

Also, they thought that having different levels of uncertainty shown for the data didn't necessarily mean that they couldn't use the data for their analysis, as one participant said "I can see the uncertainty of data on different periods and I'll try to see if I can use the info. Maybe I can use all of it but with just a warning that I must pay attention to the data being used."

In terms of the visual dimension used to represent uncertainty, transportation data analysts, and novice safety analysts found the colors to be more visual and easier to understand and the different levels of uncertainty are more identifiable in colors rather than on the ribbon or width charts, while the transportation modelers and planners preferred not using colors representing uncertainty as they saw the colors as indications of congestion levels.

All participants liked the stepline charts compared to the regular line charts to show the original speed data. They mentioned that the step line charts showed them a clearer representation of the data for every hour and then showing the uncertainty as width and color made more sense to them since the uncertainty of every hour was more identifiable and they could easily compare different hours with each other.

As we presented our data uncertainty representations, one of the participants mentioned an interesting idea to combine the stepline width and the stepline color representations for the same data (Figure 5.7). In his idea, this would give a more visual representation that would catch the users' eye faster. We added this new chart to our study to get feedback from other participants. Although the participants were surprised by the combination, they all gave positive views on how it is easy to see the trends and more uncertain areas that they need to pay attention too. The only doubt was that if it is necessary to have both methods showing the same level of uncertainty on one chart.

Factors Influencing the Users' Visualization Preferences

Although participants had different preferences in how they like to have data uncertainty presented, the considerations they had when choosing so were common. The following are important factors that determine the reasons behind which visualization techniques were

Figure 5.7 Combination of color and width as a data uncertainty visualization technique

preferred over others.

Familiarity of Representations. When we ask what participants thought about the visualizations, we saw a trend in how they preferred certain representations because they have either seen them before in their reports or the methods seemed familiar. For example, one participant mentioned that "The ribbon visualization is really common, I am already used to seeing this type of graphic so I'm really confident to use it."

Representations Linked to Other Meanings. Some participants mentioned how looking at some representations have reminded them of other meanings they have seen before, which are not about data uncertainty and can be confusing. For example, the visualization with different sizes of shapes reminded them of line charts with the shape sizes representing the amount of population. Colors also have other meanings to transportation experts; most of them mentioned that they are used to seeing colors referring to levels of congestion on the roads and highways, with red meaning high congestion and green meaning not congested. So, for the experts working daily with traffic conditions, having colors showing data uncertainty would create confusion. One participant suggested having different shades of a single color such as blue or black to show levels of uncertainty to avoid this confusion.

Uses of Visualizations. As participants were looking at the visualizations and describing what they understood from the different representation methods, we realized that the participants have different overviews when looking and analyzing the charts. Depending on their roles some mentioned they prefer to have visualizations that would quickly show them where there is high uncertainty so they pay attention to the data for those hours in their analysis. This group of people was mostly interested in color representations as they considered it as a salient stimuli. Other participants were more interested in the uncertainty of data in peak hours (6 am to 9 am and 3 pm to 6 pm) and focused more on details of the levels of uncertainty for those periods. These participants who were more focused on the details were interested in the ribbon and stepline width charts, as the transition of the line thickness between different hours seemed to provide more detail of different uncertainty levels for them to consider for their analysis without actually needing to focus on the map legend.

Preference Differences Between Experts and Novice Users

In this study, we have considered the feedback of both transportation experts with over four years of experience and novice users such as students in the field of transportation. Although their overall views were similar to how errors in the data collection process lead to uncertainty in the data they work with, we found some differences in how they deal with uncertainty in the data and their preference in data uncertainty visualizations.

The major difference between these groups was the level of awareness of the causes of uncertainty and how to deal with them for analysis. A common way to find errors and inaccuracy in data for experts is to compare their data with either historical data or data from different sources. This usually gives them an overall understanding of where the data is uncertain and if they should avoid using it or replacing it with other data. Novice users, on the other hand, usually do not have access to other sources of data and are forced to deal with the data they are given. This creates confusion for them especially when they have missing data due to data collection errors and they often need to discard data when the missing parts lead to misunderstanding and they can't fill the gaps.

For their visualization preferences for data uncertainty, the experts were more interested in the details of the levels of uncertainty and the ranges on the charts such as the ribbon chart and the stepline width. On the contrary, the novelists found color and shape to be a more obvious visualization, easier to understand, and straightforward; they were a bit confused on how the ribbon and ranges should be interpreted for decision-making.

5.4 Discussion

Based on our interview study we understood that all transportation data has some level of uncertainty. Using the original data without knowing their underlying uncertainty leads to wrong decisions. Our JSD method shows quantification of data uncertainty for transportation analysis by comparing two data distributions and calculating the differences between them.

More specifically our main focus in this study was the uncertainty of speed data. It is quite difficult to judge the uncertainty of an average speed based on its number of observations and there is no justification as to how many numbers of observations are required to believe that an average speed data is certain or uncertain. However, our analysis confirms that JSD can be used to calculate the uncertainty in speed data. This method calculates the data uncertainty based on comparison of the distribution of the data in analysis with the distribution of the historical data.

In addition to confirming the usefulness of our JSD-based uncertainty quantification and visualization method, our user study results demonstrate that the transportation decision-makers face many challenges working with uncertain data. These challenges and their needs can give us an overview of the practitioners' understanding of what they consider as uncertainty, and how they prefer it to be visualized. This information can help us create more unique and customized guidelines to visualize levels of data uncertainty for not only traffic conditions analyzers, but for other groups of transportation decision-makers as well.

CHAPTER 6 CONCLUSION

This chapter summarises the findings of this thesis, discusses the major limitations of our approach, and finally highlights some future directions of research.

6.1 Summary of Works

In this study, we focus on understanding how transportation decision-makers work with traffic data and further supporting their work by tool and visualization design. We first conducted an interview study with 19 practitioners working at MTQ to understand their practices, needs, and challenges in transportation decision-making. We then summarized the needs and challenges of different types of transportation decision-makers and created personas to represent them.

The next aim of our study was to design a traffic condition analysis tool for transportation decision-makers based on the results of our interview study. After several discussions with transportation experts, we created design guidelines and later mockups of our traffic conditions tool. We later conducted a user study to get feedback from transportation-decision makers of MTQ to improve our traffic conditions analysis tool.

We also noticed that transportation decision-makers deal with a lot of uncertainty in their data. To understand the accuracy of their data, they either compare data with different sources or with historical data. To solve this problem, we have introduced a novel method to calculate the level of uncertainty of speed data using JSD. This measure of data uncertainty can bring awareness to transportation decision-makers and help them make more accurate decisions in the future.

In addition, we investigated data uncertainty visualization techniques to visualize different levels of uncertainty on top of the regular speed data that transportation decision-makers are using. We conducted a user study with 11 people including experts and novice transportation decision-makers. The aim of this study was to evaluate the practitioners' perceptions in various data uncertainty visualizations and also to explore how transportation decision-makers consider and use data uncertainty in their decision-making process.

6.2 Limitations

In our study, we have only focused on working with transportation decision-makers from one organization (i.e. MTQ) as we wanted to create a tool to facilitate traffic condition analysis for them.

Moreover, our study covers data analysis for road traffic so only experts in this area are able to use this tool to make traffic conditions analysis. Other areas such as analysis of public transportation, pedestrians, cyclists, etc are not covered in this study.

To calculate the levels of data uncertainty, we used the average speed data. So our work is limited to the uncertainty of data in this domain. Also we have not considered situations where speed limits change across a corridor.

In our data uncertainty visualization user study, the evaluation of our data uncertainty visualizations is only focused on the participants' preferences, but not on how efficiently the uncertainty information helped people to make decisions.

Our uncertainty visualization techniques were only applied to line charts to show the uncertainty of average speed data and the practitioners' feedback on the visualizations was limited to these charts in our user study.

6.3 Future Research

For future work, we would like to address our limitations. First, broadening our research to understand the needs and challenges of transportation decision-makers working in other organizations can help establish the external validity of our results.

We used a novel method using JSD to calculate levels of data uncertainty in speed data. This method can be further used to calculate data uncertainty in various domains both in transportation data and along with machine learning methods and also other domains such as bioinformatics which constantly deal with inaccuracy in data.

In addition, we plan to investigate more data uncertainty visualization techniques such as using animation, blurriness, and sketchiness. These techniques can also be applied to maps and heatmaps to communicate data uncertainty with decision-makers.

Through our user study, we have explored a graph with a combination of two different data uncertainty visualizations techniques based on one of the practitioners' suggestions. We gained positive feedback for this combination from other practitioners and plan to investigate and explore further combinations of visualization techniques to represent data uncertainty.

REFERENCES

[1] T. Miaskiewicz and K. A. Kozar, "Personas and user-centered design: How can personas benefit product design processes?" *Design Studies*, vol. 32, pp. 417–430, 2011.

[2] H. Sun, J. Wu, D. Ma, and J. Long, "Spatial distribution complexities of traffic congestion and bottlenecks in different network topologies," *Applied Mathematical Modelling*, vol. 38, no. 2, pp. 496 – 505, 2014.

[3] INRIX, "Inrix 2018 global traffic scorecard," 2019. [Online]. Available: http://inrix.com/scorecard/

[4] K. Lévesque, "Les coûts de la congestion évalués à 4,2 milliards pour 2018," *La Presse*, 2018.

[5] S. Balters, M. Bernstein, and P. E. Paredes, "On-road stress analysis for in-car interventions during the commute," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19. New York, NY, USA: ACM, 2019, pp. LBW2110:1–LBW2110:6.

[6] M. Obrist, D. Wurhofer, A. Krischkowsky, E. Karapanos, D. H. Wilfinger, N. Perterer, and M. Tscheligi, "Experiential perspectives on road congestions," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '13. New York, NY, USA: ACM, 2013, pp. 535–540.

[7] A. Cooper, *The Inmates Are Running the Asylum.* Indianapolis, IN, USA: Macmillan Publishing Co., Inc., 1999.

[8] N. Sharbatdar, Y. Lamine, B. Milord, C. Morency, and J. Cheng, "Capturing the practices, challenges, and needs of transportation decision-makers," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–7. [Online]. Available: https://doi.org/10.1145/3334480.3382864

[9] D. A. Norman, "The design of everyday things: Revised and expanded edition," 2013.

[10] J. D. Gould and C. Lewis, "Designing for usability: Key principles and what designers think," *Commun. ACM*, vol. 28, no. 3, p. 300–311, Mar. 1985. [Online]. Available: https://doi.org/10.1145/3166.3170

[11] J. Gulliksen, B. Göransson, I. Boivie, S. Blomkvist, J. Persson, and Å. Cajander, "Key principles for user-centred systems design," *Behaviour & Information Technology*, vol. 22, pp. 397 – 409, 2003.

[12] Shneiderman, "Direct manipulation: A step beyond programming languages," *Computer*, vol. 16, no. 8, pp. 57–69, 1983.

[13] B. Shneiderman, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 3rd ed. USA: Addison-Wesley Longman Publishing Co., Inc., 1997.

[14] J. Nielsen, "How to conduct a heuristic evaluation," *Nielsen Norman Group*, vol. 1, pp. 1–8, 1995.

[15] A. Cooper, *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity (2nd Edition)*. Pearson Higher Education, 2004.

[16] L. Nielsen and K. Storgaard Hansen, "Personas is applicable: A study on the use of personas in denmark," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 1665–1674. [Online]. Available: https://doi.org/10.1145/2556288.2557080

[17] J. Pruitt and J. Grudin, "Personas: practice and theory," in *Proceedings of the 2003 conference on Designing for user experiences*, 2003, pp. 1–15.

[18] A. Dotan, N. A. M. Maiden, V. Lichtner, and L. Germanovich, "Designing with only four people in mind? - a case study of using personas to redesign a work-integrated learning support system," in *INTERACT*, 2009.

[19] M. Burnett, A. Peters, C. Hill, and N. Elarief, "Finding gender-inclusiveness software issues with gendermag: A field investigation," 05 2016.

[20] M. Aoyama, "Persona-and-scenario based requirements engineering for software embedded in digital consumer products," in *13th IEEE International Conference on Requirements Engineering (RE'05)*, 2005, pp. 85–94.

[21] M. Almaliki, C. Ncube, and R. Ali, "Adaptive software-based feedback acquisition: A persona-based design," in *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 2015, pp. 100–111.

[22] Å. Blomquist and M. Arvola, “Personas in action: ethnography in an interaction design team,” in *Proceedings of the second Nordic conference on Human-computer interaction*, 2002, pp. 197–200.

[23] T. Matthews, T. Judge, and S. Whittaker, “How do designers and user experience professionals actually perceive and use personas?” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’12. New York, NY, USA: Association for Computing Machinery, 2012, p. 1219–1228. [Online]. Available: https://doi.org/10.1145/2207676.2208573

[24] A. Wu, X. Zhang, G. Convertino, and J. M. Carroll, “Civil: support geo-collaboration with information visualization,” in *Proceedings of the ACM 2009 international conference on Supporting group work*, 2009, pp. 273–276.

[25] S. McKenna, D. Staheli, and M. Meyer, “Unlocking user-centered design methods for building cyber security visualizations,” in *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)*. IEEE, 2015, pp. 1–8.

[26] H. Tokola, C. Gröger, E. Järvenpää, and E. Niemi, “Designing manufacturing dashboards on the basis of a key performance indicator survey,” *Procedia CIRP*, vol. 57, pp. 619–624, 2016.

[27] M. Mahmoodpour, A. Lobov, M. Lanz, P. Mäkelä, and N. Rundas, “Role-based visualization of industrial iot-based systems,” in *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*. IEEE, 2018, pp. 1–8.

[28] M. Elbaham, K. K. Nguyen, and M. Cheriet, “A traffic visualization framework for monitoring large-scale inter-datacenter network,” in *2016 12th International Conference on Network and Service Management (CNSM)*. IEEE, 2016, pp. 277–281.

[29] M. L. Pack, “Visualization in transportation: challenges and opportunities for everyone,” *IEEE computer graphics and applications*, vol. 30, no. 4, pp. 90–96, 2010.

[30] Z. Wang, T. Ye, M. Lu, X. Yuan, H. Qu, J. Yuan, and Q. Wu, “Visual exploration of sparse traffic trajectory data,” *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1813–1822, 2014.

[31] H. Guo, Z. Wang, B. Yu, H. Zhao, and X. Yuan, “Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection,” in *Proceedings of the 2011 IEEE Pacific Visualization Symposium*, 2011, pp. 163–170.

[32] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, 2013.

[33] J. Pu, S. Liu, Y. Ding, H. Qu, and L. Ni, "T-watcher: A new visual analytic system for effective traffic surveillance," in *2013 IEEE 14th International Conference on Mobile Data Management*, vol. 1, 2013, pp. 127–136.

[34] P. Cruz and P. Machado, "Pulsing blood vessels: A figurative approach to traffic visualization," *IEEE Computer Graphics and Applications*, vol. 36, no. 2, pp. 16–21, 2016.

[35] G. A. M. Gomes, E. Santos, and C. A. Vidal, "Interactive visualization of traffic dynamics based on trajectory data," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2017, pp. 111–118.

[36] A. Ballatore and A. Zipf, "A conceptual quality framework for volunteered geographic information," in *International Conference on Spatial Information Theory.* Springer, 2015, pp. 89–107.

[37] J. Thomson, E. Hetzler, A. MacEachren, M. Gahegan, and M. Pavel, "A typology for visualizing uncertainty," in *Visualization and Data Analysis 2005*, vol. 5669. International Society for Optics and Photonics, 2005, pp. 146–157.

[38] M. Brown, S. Sharples, J. Harding, C. J. Parker, N. Bearman, M. S. Maguire, D. Forrest, M. Haklay, and M. Jackson, "Usability of geographic information: current challenges and future directions." *Applied ergonomics*, vol. 44 6, pp. 855–65, 2013.

[39] N. Bencomo and A. Belaggoun, "A world full of surprises: Bayesian theory of surprise to quantify degrees of uncertainty," in *Companion Proceedings of the 36th International Conference on Software Engineering*, 2014, pp. 460–463.

[40] J. Dai, W. Wang, Q. Xu, and H. Tian, "Uncertainty measurement for interval-valued decision systems based on extended conditional entropy," *Knowledge-Based Systems*, vol. 27, pp. 443–450, 2012.

[41] Y. Chen, K. Wu, X. Chen, C. Tang, and Q. Zhu, "An entropy-based uncertainty measurement approach in neighborhood systems," *Information Sciences*, vol. 279, pp. 239–250, 2014.

[42] J. Bertin, "Semiology of graphics; diagrams networks maps," Tech. Rep., 1983.

[43] J. Morrison, "A theoretical framework for cartographic generalization with the emphasis on the process of symbolization," *International Yearbook of Cartography*, vol. 14, pp. 115–127, 01 1974.

[44] A. M. MacEachren, "Visualizing uncertain information," *Cartographic perspectives*, no. 13, pp. 10–19, 1992.

[45] A. MacEachren and C. P. Keller, "How maps work: representation, visualization & design," *Cartographica*, vol. 34, no. 4, p. 55, 1997.

[46] D. Spiegelhalter, M. Pearson, and I. Short, "Visualizing uncertainty about the future," *science*, vol. 333, no. 6048, pp. 1393–1400, 2011.

[47] H. Guo, J. Huang, and D. H. Laidlaw, "Representing uncertainty in graph edges: An evaluation of paired visual variables," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 10, pp. 1173–1186, 2015.

[48] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan, "Visual semiotics & uncertainty visualization: An empirical study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2496–2505, 2012.

[49] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, and R. Moorhead, "A user study to compare four uncertainty visualization methods for 1d and 2d datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1209–1218, 2009.

[50] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J.-D. Fekete, "Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2769–2778, 2012.

[51] G. Guest, K. M. MacQueen, and E. E. Namey, *Applied thematic analysis.* Sage Publications, 2011.

[52] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[53] S. Filippi, O. Cappé, and A. Garivier, "Optimism in reinforcement learning and kullback-leibler divergence," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010, pp. 115–122.

[54] A. Le Rhun, F. Bonnans, G. De Nunzio, T. Leroy, and P. Martinon, "A stochastic data-based traffic model applied to vehicles energy consumption estimation," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[55] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7893–7897.

[56] A. B. Chan and N. Vasconcelos, "Classification and retrieval of traffic video using auto-regressive stochastic processes," in *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.* IEEE, 2005, pp. 771–776.

[57] A. Zeroual, F. Harrou, Y. Sun, and N. Messai, "Integrating model-based observer and kullback–leibler metric for estimating and detecting road traffic congestion," *IEEE Sensors Journal*, vol. 18, no. 20, pp. 8605–8616, 2018.

[58] T. Lenaerts, J. Ferkinghoff-Borg, F. Stricher, L. Serrano, J. W. Schymkowitz, and F. Rousseau, "Quantifying information transfer by protein domains: analysis of the fyn sh2 domain structure," *BMC structural biology*, vol. 8, no. 1, p. 43, 2008.

[59] J. Schug, W.-P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. Stoeckert, "Promoter features related to tissue specificity as measured by shannon entropy," *Genome biology*, vol. 6, p. R33, 02 2005.

[60] A. Clim, R. D. Zota, and G. TinicA, "The kullback-leibler divergence used in machine learning algorithms for health care applications and hypertension prediction: A literature review," in *EUSPN/ICTH*, 2018.

[61] T. Murakami, "A succinct model for re-identification of mobility traces based on small training data," in *2018 International Symposium on Information Theory and Its Applications (ISITA)*. IEEE, 2018, pp. 164–168.

[62] Z. Huang, X. Ling, P. Wang, F. Zhang, Y. Mao, T. Lin, and F.-Y. Wang, "Modeling real-time human mobility based on mobile phone and transportation data fusion," *Transportation research part C: emerging technologies*, vol. 96, pp. 251–269, 2018.

APPENDIX A INTERVIEW STUDY QUESTIONS

Responsibilities

1. What is your function?

2. Can you describe your responsibilities and duties?

3. How long have you been in this position?

Objectives, obstacles, and solutions in the use of data on traffic conditions on the road network

1. How do you use road traffic data in your work?

2. What are your 2 or 3 main objectives in the use of traffic data on the road network?

3. For each objective:

o Can you describe your approach?

o What are the biggest obstacles to this goal? o How did you overcome these obstacles?

Experience in researching traffic conditions on the road network

1. Can you tell us about a good experience (that is, a successful search) that you had when you were looking for traffic data on the road network?

o What are the key elements that made this research successful, in your opinion?

2. Can you tell us about a bad experience (that is, unsatisfactory research) that you had during your data search?

o What are the key elements that made this research unsuccessful, in your opinion?

Experience in using traffic data on the road network

1. Can you tell us a good experience (a successful use) that you had when using road traffic data?

o What are the key elements that made this enforcement work successful, in your opinion?

2. Can you tell us about a bad experience (that is, unsatisfactory use) you had when using road traffic data?

o What are the key elements that made this implementation unsuccessful, in your opinion?

Idea of tools

1. In a utopian world, what kind of tools for monitoring and analyzing traffic conditions on

the road network would you find most useful?

o What functions would you like to have in this kind of tools?

o What information would you like to have in this kind of tool?

2. What kind of road traffic monitoring and analysis tools did you use?

o What were the most useful functions? and the least useful?

APPENDIX B USER STUDY MATERIAL

Part 1: Assessing the level of congestion

In this part users were asked to first introduce themselves, and then assess the congestion level of Mondays and Fridays based on two graphs showing speed data.

1. What is your function and how many years of experience do you have working with transportation data?

We have the average speed, and the number of observations and the speed limit of autoroute 40 of Montreal, on Mondays and Fridays between 12 am and 23 pm on Monday, Feb 3rd, 2020, and Friday, Feb 28, 2020. The aggregated speed data of a single segment of Mondays and Fridays are shown on two line graphs.

2. Based on this data and the graphs, how would you assess the level of congestion between these two days?

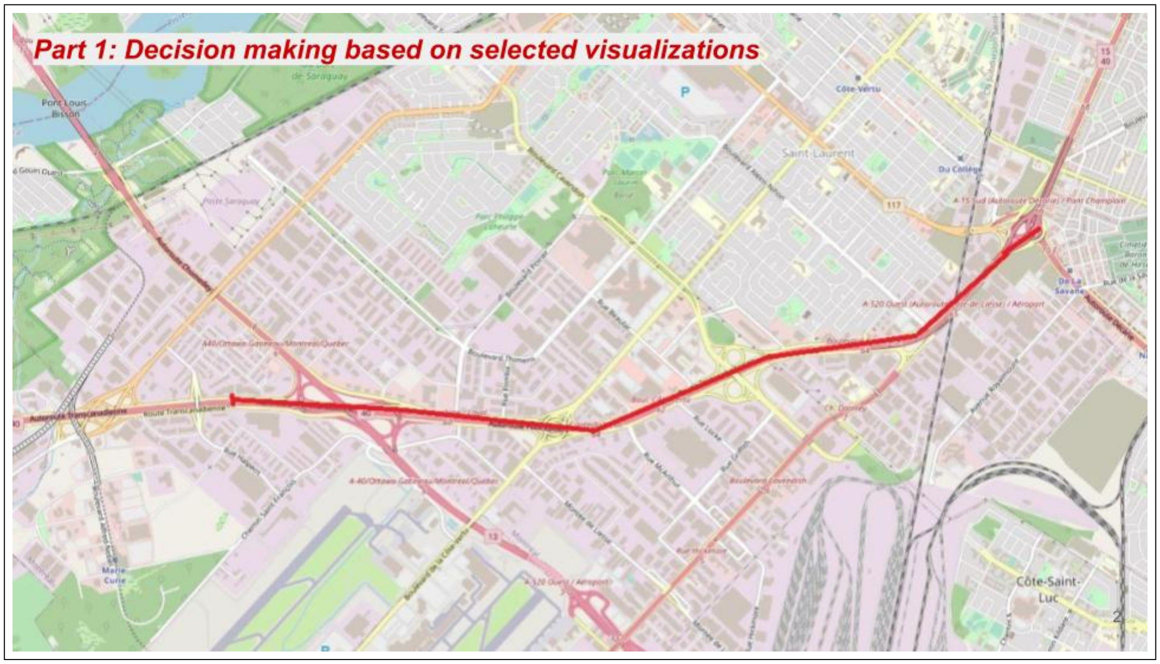
3. Is there any other data you would like to see that can help you assess this impact?

4. What information would you consider when you want to assess the level of congestion on Mondays and Fridays?

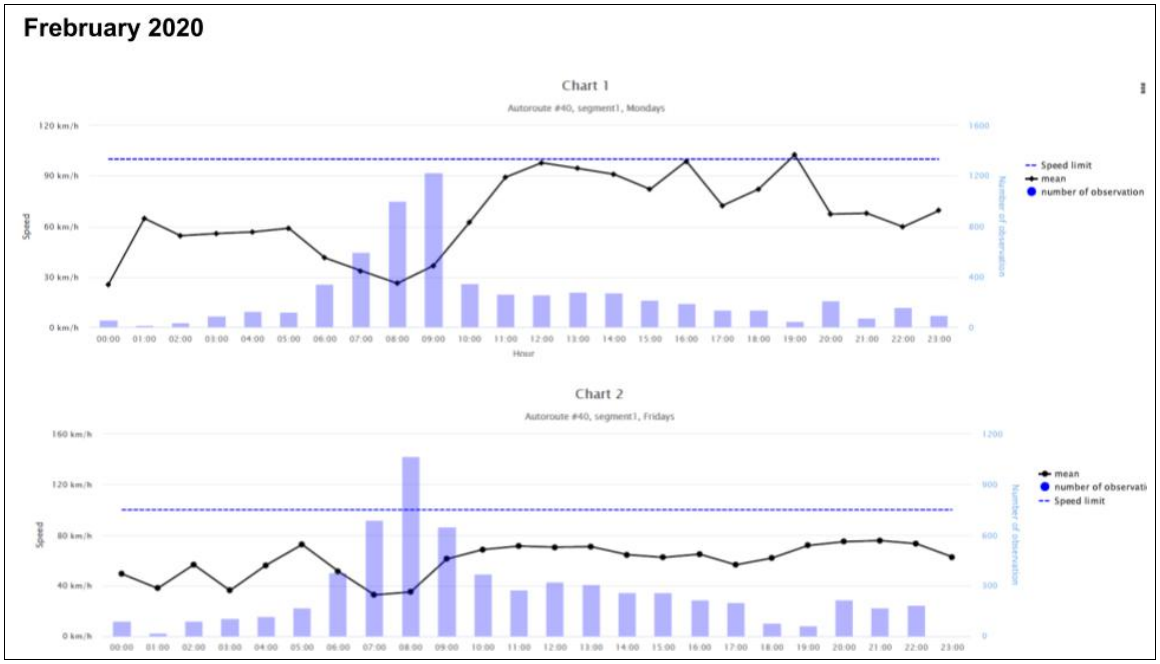
5. How do you estimate the difference in the level of congestion between different segments?

6. Do you think there is an increase in congestion on Fridays? Why?

7. What factors do you think will cause an increase in congestion?


Part 1: Decision making based on selected visualizations
Saint-Laurent
Côte-Vertu
Du Collège
Côte-Saint-Luc
2


A40/Ottawa-Gatineau/Montréal/Québec
Autoroute Transcanadienne
Route Transcanadienne
A-40/Ottawa-Gatineau/Montréal/Québec
A-520 Ouest / Aéroport
Marie Curie
3



Part 2: Data uncertainty in transportation data

In the second part of the study we asked the following questions about data uncertainty:

8. Do you need to work with uncertain data?

9. What are the different types of data uncertainty in your work?

10. What do you think causes uncertainty in speed data?

Part 3: Data uncertainty visualizations

In this part of our study we showed the uncertainty visualizations one by one and asked for participant's feedback. More specifically we asked the following questions for each visualization:

11. What do you think this kind of representation means?

12. Which parts do you think have the highest and lowest uncertainty?

We then showed all visualizations on a single page and asked them the following questions:

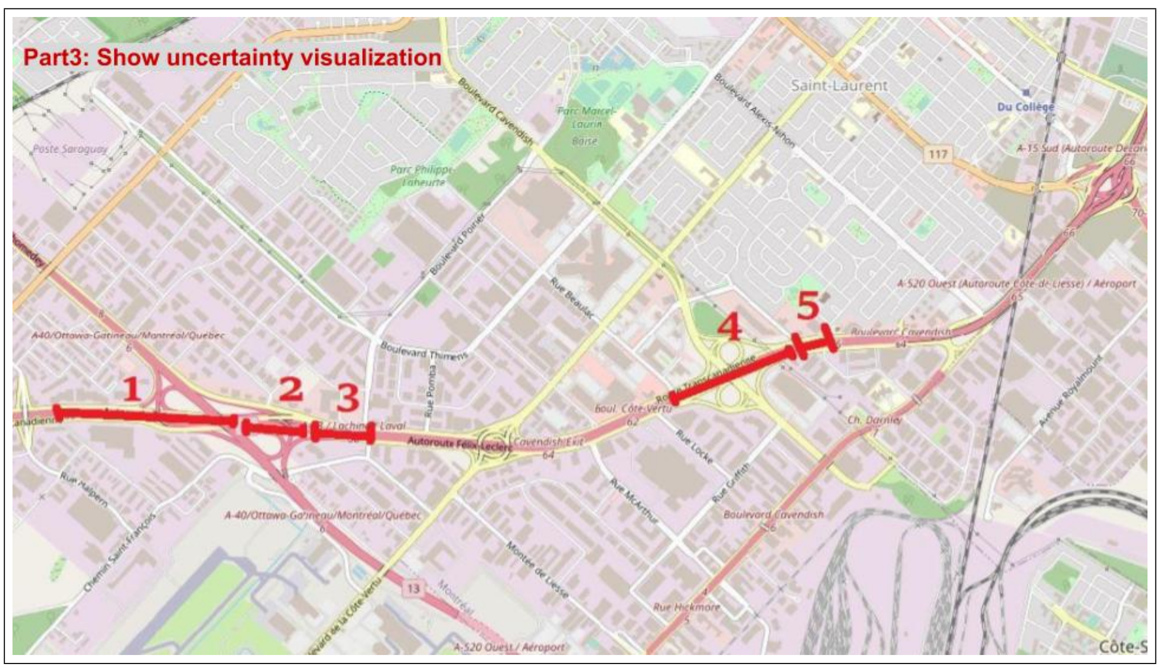
13. Which one do you think is the worst kind of representation for data uncertainty?

14. Which one do you think can be more useful for decision-makers?

15. How can decision-makers benefit from these data uncertainty visualizations?

16. If you had these visualizations to assess the level of congestion instead of the line charts in task 1, how would they affect your answers in assessing the level of congestion?

17. What are the consequences and risks of not being aware of uncertain data?



Visualization 1: ribbon

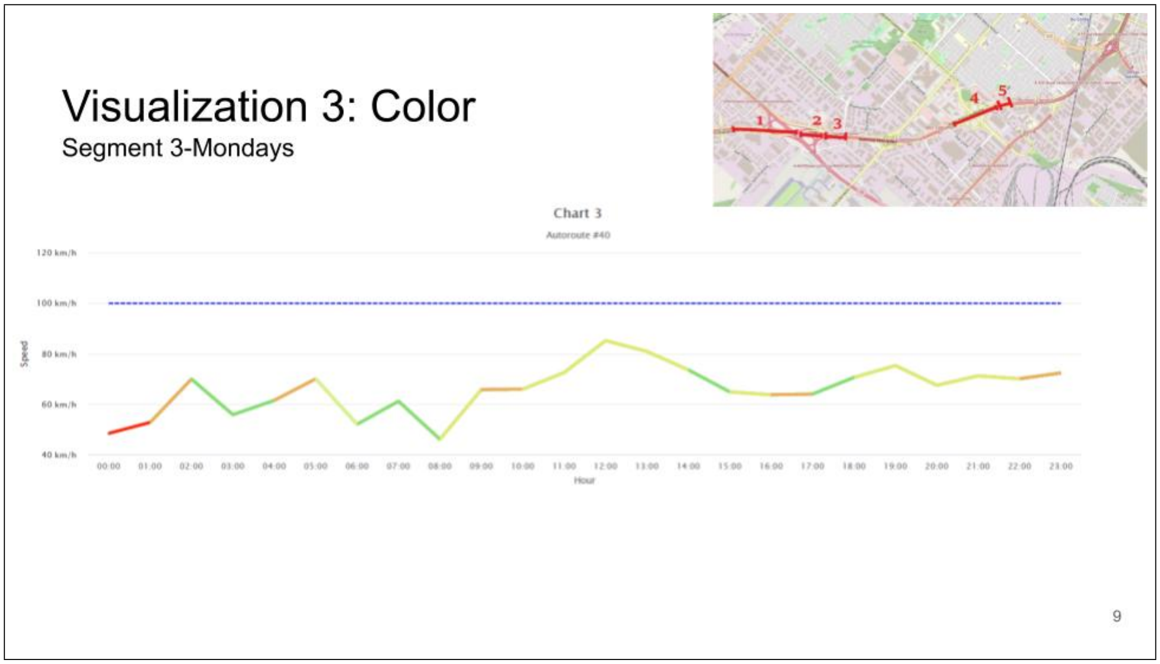
Segment 1-Mondays

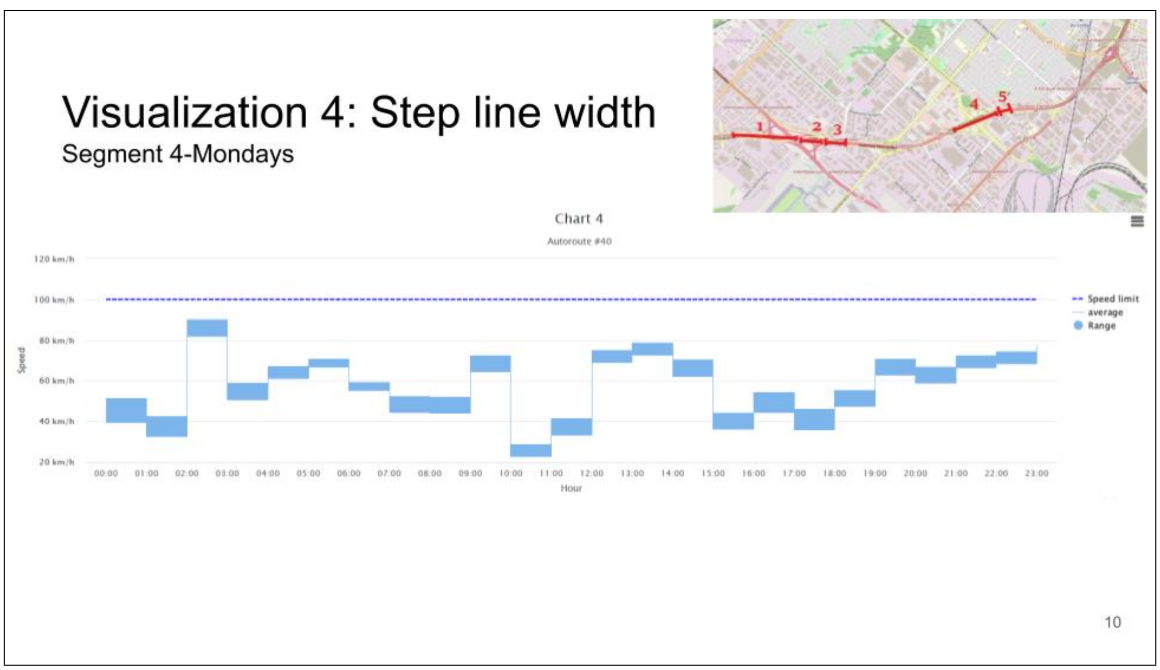
7

Visualization 2: Shapes

Segment 2 -Mondays

8


Visualization 3: Color
Segment 3-Mondays
1
2
3
4
5
Chart 3
Autoroute #40
120 km/h
100 km/h
80 km/h
60 km/h
40 km/h
Speed
00:00 01:00 02:00 03:00 04:00 05:00 06:00 07:00 08:00 09:00 10:00 11:00 12:00 13:00 14:00 15:00 16:00 17:00 18:00 19:00 20:00 21:00 22:00 23:00
Hour
9


Visualization 4: Step line width
Segment 4-Mondays
1
2
3
4
5
Chart 4
Autoroute #40
120 km/h
100 km/h
80 km/h
60 km/h
40 km/h
20 km/h
Speed
00:00 01:00 02:00 03:00 04:00 05:00 06:00 07:00 08:00 09:00 10:00 11:00 12:00 13:00 14:00 15:00 16:00 17:00 18:00 19:00 20:00 21:00 22:00 23:00
Hour
Speed limit
average
Range
10

Visualization 5: Step line color

Segment 5-Mondays

11

Visualization 6: Step line, color & width

12