

Titre: Méthode d'exploration visuelle de données de sondage destinée à un utilisateur novice
Title: un utilisateur novice

Auteur: Manal Maroua Boucenna
Author:

Date: 2020

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Boucenna, M. M. (2020). Méthode d'exploration visuelle de données de sondage destinée à un utilisateur novice [Master's thesis, Polytechnique Montréal].
Citation: PolyPublie. <https://publications.polymtl.ca/5389/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/5389/>
PolyPublie URL:

Directeurs de recherche: Thomas Hurtut
Advisors:

Programme: Génie informatique
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Méthode d'exploration visuelle de données de sondage destinée à un utilisateur
novice**

MANAL MAROUA BOUCENNA

Département de génie informatique et génie logiciel

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

Août 2020

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

**Méthode d'exploration visuelle de données de sondage destinée à un utilisateur
novice**

présenté par **Manal Maroua BOUCENNA**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Jinghui CHENG, président

Thomas HURTUT, membre et directeur de recherche

Farida CHERIET, membre

DÉDICACE

*À mes parents,
rien ne serait possible sans vous. . .*

REMERCIEMENTS

J'adresse mes remerciements à mon directeur de recherche, Thomas HURTUT, pour ses conseils et sa patience tout au long de ma maîtrise. Je te remercie pour le temps que tu as consacré à m'apporter les outils méthodologiques indispensables à la conduite de cette recherche. Ton exigence m'a grandement stimulée.

J'adresse mes sincères remerciements à tous les professeurs, intervenants, et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé mes réflexions et ont accepté de me rencontrer et répondre à mes questions durant mes recherches.

Je dédie tout ce travail à mes parents, mes modèles et source de motivation pour avoir cru en moi et pour tous leurs sacrifices : *merci mamoune, merci papoune* !

Je remercie Djazou, Hakim, Rym et Yasmine de m'avoir accueillie puis supportée (tant bien que mal) pendant ce confinement et pour tous les moments agréables et expérimentations qu'on a vécus ensemble.

Enfin, je remercie Sassou, Louloute, Mimoune, Katoussa, Katouss, Mimz et Yasmine. Votre soutien était inestimable !

Finalement, merci pour tout Bibouche. . .

RÉSUMÉ

Les sondages sont un moyen efficace de sonder l'opinion d'une population à partir d'un échantillon. Une fois le sondage administré, une analyse des réponses récoltées doit se faire en combinant des outils statistiques, de la programmation et de la visualisation de données. Cependant, avec l'accroissement du nombre de questions et la complexité des données, l'analyse de ces données peut vite devenir une tâche ardue pour un utilisateur novice ayant un bagage limité en visualisation de données et programmation.

L'étude suivante porte sur la conception d'un outil d'exploration visuelle destiné à un public novice. Cet outil pourrait, par exemple, être pertinent pour des étudiants du premier cycle en sciences politiques ayant des connaissances basiques en visualisation de données.

L'objectif de cette étude est de proposer un outil de visualisation qui permette à un utilisateur novice d'explorer efficacement ses données de sondage. Nous avons d'abord identifié les principales tâches d'exploration réalisées par notre public cible : distribution, relation entre les variables et évolution temporelle. Puis, nous avons analysé les diverses recommandations, disponibles dans la littérature, concernant la conception d'outils d'exploration adaptés à un public novice.

L'outil de visualisation développé se base sur trois principaux éléments de design : l'approche « *overview first, details on-demand* », la formulation semi-textuelle d'une requête en définissant explicitement la tâche et la question de sondage, et la génération automatique de visualisations. Cet outil intègre également des interactions dynamiques et intuitives telles que le filtrage et le glisser-déposer (*Drag-n-Drop*).

Une étude pilote a été réalisée auprès de quatre étudiants en science politique dans le but d'évaluer l'atteinte de nos objectifs. L'évaluation a permis de recueillir diverses rétroactions et recommandations. Les résultats préliminaires montrent un vif intérêt, de la part des participants, envers notre outil. Les participants ont jugé que l'outil et ses interactions diverses étaient faciles d'utilisation et donc adaptés à un utilisateur novice. Les visualisations générées ont été, pour la plupart, jugées pertinentes. Plusieurs améliorations ont été proposées par nos participants en termes d'interaction et de gestion de données manquantes.

De manière générale, les participants ont réussi, à l'aide de notre outil, à explorer les résultats d'un sondage. Cependant en raison du nombre réduit de participants, les résultats ne sont pas statistiquement significatifs pour infirmer ou confirmer nos hypothèses. Néanmoins, les observations récoltées durant cette évaluation nous ont permis de cerner les principaux besoins

de notre public cible en termes de visualisations et d'identifier les limites et améliorations à apporter dans un prochain travail de recherche.

ABSTRACT

Polls are an effective way to survey the opinion of a population from a sample. Once the survey is administered, an analysis of the responses collected must be done by combining statistical tools, programming and data visualization. However, with the increased number of questions and the complexity of the data, analyzing this data can quickly become a daunting task for a novice user with limited background in data visualization and programming.

The following study involves the design of a visual exploration tool for a novice audience. This tool could, for example, be relevant for undergraduate political science students with basic knowledge in data visualization.

The objective of this study is to provide a visualization tool that allows a novice user to effectively explore their survey data. We first identified the main exploration tasks performed by our target audience: distribution, relationship between variables and time evolution. Then, we analyzed the various recommendations, available in the literature, regarding the design of exploration tools suitable for a novice audience.

The visualization tool developed is based on three main design elements: the "*overview first, details on-demand*" approach, the semi-textual formulation of a request by explicitly defining the task and the question of survey, and the automatic generation of visualizations. This tool also integrates dynamic and intuitive interactions such as filtering and drag-and-drop.

A pilot study was carried out with four political science students to assess the achievement of our goals. The evaluation gathered various feedback and recommendations. The preliminary results show a keen interest on the part of the participants in our tool. Participants found the tool and its various interactions to be easy to use and therefore suitable for a novice user. The visualizations generated were, for the most part, considered relevant. Several improvements were proposed by our participants in terms of interaction and management of missing data.

In general, participants were able to explore the results of a survey using our tool. However, due to the small number of participants, the results are not statistically significant to invalidate or confirm our hypotheses. Nevertheless, the observations gathered during this evaluation allowed us to identify the main needs of our target audience in terms of visualizations and to identify the limits and improvements to be made in future research work.

TABLE DES MATIÈRES

DÉDICACE	iii
REMERCIEMENTS	iv
RÉSUMÉ	v
ABSTRACT	vii
TABLE DES MATIÈRES	viii
LISTE DES TABLEAUX	xi
LISTE DES FIGURES	xii
LISTE DES SIGLES ET ABRÉVIATIONS	xv
LISTE DES ANNEXES	xvi
CHAPITRE 1 INTRODUCTION	1
1.1 Problématique	2
1.2 Objectif général	3
1.3 Plan du mémoire	3
CHAPITRE 2 REVUE DE LITTÉRATURE	4
2.1 Processus de création d'une visualisation	4
2.2 Considérations conceptuelles pour le design d'une visualisation	5
2.2.1 Exploration visuelle de données	5
2.2.2 Outils de visualisation destinés aux utilisateurs novices	7
2.3 Revue des outils de visualisation de données de sondages	8
2.4 Revue des outils de visualisation de données multidimensionnelles	11
2.4.1 Visualisations à une seule vue (<i>Single view</i>)	12
2.4.2 Visualisations à vues multiples (<i>Multiple views</i>)	14
CHAPITRE 3 MÉTHODOLOGIE PROPOSÉE	18
3.1 Contexte d'application	18
3.1.1 Données	18

3.1.2	Utilisateurs cibles	20
3.2	Définition et abstraction des tâches	20
3.3	Recommandations pour le design d'un outil d'exploration visuelle destiné à un utilisateur novice	21
3.4	Hypothèses et objectifs	21
3.4.1	Objectifs de recherche	22
3.4.2	Hypothèses de recherche	23
3.5	Outil de visualisation : Design et fonctionnalités	24
3.5.1	Vue globale et détaillée d'un sondage	24
3.5.2	Génération automatique de visualisation	33
3.5.3	Interactivité de l'outil	39
3.6	Implémentation	42
CHAPITRE 4 PROTOCOLE D'ÉVALUATION DE LA VISUALISATION PROPOSÉE		45
4.1	Méthode d'évaluation	45
4.2	Données	45
4.3	Échantillon	46
4.4	Procédure d'évaluation de l'outil de visualisation	47
4.4.1	Étape 1 : Questions générales	47
4.4.2	Étape 2 : Présentation de l'outil	47
4.4.3	Étape 3 : Exploration libre	47
4.4.4	Étape 4 : Évaluation de l'outil	47
4.4.5	Étape 5 : Entrevues et questions ouvertes	47
4.5	Hypothèses de recherche et mesures d'évaluation associées	48
CHAPITRE 5 RÉSULTATS ET DISCUSSIONS		49
5.1	Présentation et discussion des résultats	49
5.1.1	Questionnaire sur l'exploration libre	49
5.1.2	Questionnaire sur la génération de visualisations	52
5.1.3	Questionnaire sur l'expérience utilisateur	55
5.2	Synthèse des recommandations proposées	57
5.2.1	Visualisations	57
5.2.2	Gestion de données et interactions	57
5.2.3	Exportation des vues	58
5.3	Limitations	58
5.3.1	Limitations de l'outil de visualisation	58

5.3.2 Limitations de l'expérience	59
CHAPITRE 6 CONCLUSION	61
6.1 Synthèse des travaux	61
6.2 Améliorations futures	62
RÉFÉRENCES	64
ANNEXES	70

LISTE DES TABLEAUX

Tableau 2.1	Tableau récapitulatif des principales recommandations de design d'un outil d'exploration visuelle pour utilisateur novice.	8
Tableau 3.1	Types de questions pris en considération dans notre étude	19
Tableau 3.2	Tableau récapitulatif des principales recommandations de design d'un outil d'exploration visuelle pour utilisateur novice.	22
Tableau 3.3	Association entre chaque élément de design retenu et ses objectifs. . .	23
Tableau 3.4	Exemple de questions de sondage et leurs représentations associées . .	27
Tableau 3.5	Exemple de réponses d'une question de type échelle de Likert. La moyenne arithmétique est de 3,5, ce qui se traduirait par «Ni en désaccord ni d'accord et demi». Cette valeur ne fait pas de sens statistiquement.	28
Tableau 3.6	Répartition des responsabilités lors du processus de création d'une visualisation avec la méthode proposée.	32
Tableau 3.7	Tableau récapitulatif des visualisations générées par type de tâches et questions de sondage.	35
Tableau 3.8	Exemple d'encodage pour échelle de Likert	44
Tableau 4.1	Exemple de données recueillies pour chaque participant	46
Tableau 4.2	Correspondance entre les hypothèses de recherche et les questionnaires de l'expérience	48
Tableau 5.1	Résumé des questions pour évaluer l'exploration	49
Tableau 5.2	Résumé des questions pour évaluer les visualisations générées	52
Tableau A.1	Liste de travaux analysés pour l'extraction de tâches abstraites. . . .	70
Tableau B.1	Visualisations générées et descriptions associées. Les descriptions sont reprises et adaptées des sites web : <i>seeingdata</i> [1] et <i>datavizproject</i> [2].	71

LISTE DES FIGURES

Figure 2.1	Représentation de données de sondages sous forme de glyphes. © 2014 Les auteurs Computer Graphics Forum © 2014 The Eurographics Association and John Wiley & Sons Ltd. Publié par John Wiley & Sons Ltd [3]	9
Figure 2.2	Visualisation des réponses d'un questionnaire sur la satisfaction des clients en utilisant une topologie de coordonnées parallèles. © 2003 IEEE [4]	10
Figure 2.3	Interface de l'outil permettant de visualiser les résultats d'un sondage réalisé auprès de citoyens. Droits gérés par Taylor et Francis [5] . . .	11
Figure 2.4	Interface de l'outil destiné aux utilisateurs novices, permettant de visualiser les résultats d'un questionnaire [6]. Capture d'écran réalisée par l'auteur.	12
Figure 2.5	Matrice de <i>Scatter Plot</i> en utilisant la base de données Iris [7]. Figure reproduite par l'auteur. [8]	13
Figure 2.6	Méthode des coordonnées parallèles en utilisant la base de données «Iris» [7]. © 1990, IEEE [9]	14
Figure 2.7	Méthode des <i>Parallel Sets</i> en utilisant la base de données «Titanic» [10]. © 2006, IEEE [11]	15
Figure 2.8	Glyphes de type visage de Chernoff. Droits gérés par Taylor et Francis [12]	16
Figure 2.9	Interface de <i>Keshif</i> présentant les données sur le nombre d'accidents mortels aux États-Unis en 2013. © 2018, IEEE [13]	16
Figure 2.10	Interface de <i>Lineup</i> présentant les 2000 meilleures compagnies. © 2013, IEEE [14]	17
Figure 2.11	Interface de <i>Domino</i> présentant des relations entre les variables d'une base de données musicale. © 2014, IEEE [15]	17
Figure 3.1	Figure résumant les différents types de sondage. De haut en bas : approche transversale, approche longitudinale à panel et approche longitudinale à cohorte.	19
Figure 3.2	Vue réduite de l'outil développé utilisant les données normalisées de l'étude comparative des systèmes électoraux [16]. Le panneau latéral (A) offre une vue globale de la distribution des questions du sondage. Le pied de page (B) est dans ce cas réduit. Le panneau central (C) présente la zone de formulation de requêtes et générations de visualisations. . .	25

Figure 3.3	Vue complète de l'outil développé utilisant les données normalisées de l'étude comparative des systèmes électoraux [16]. Le panneau latéral (A) offre une vue globale de la distribution des questions du sondage. Le pied de page (B) offre une vue globale de la distribution des métadonnées. Le panneau central (C) présente la zone de formulation de requêtes et générations de visualisations.	26
Figure 3.4	Représentation choisie par type de données	28
Figure 3.5	Exemple de pied de page comportant des métadonnées telles que le nombre de participants par pays, année ou type d'élection.	29
Figure 3.6	Exemple de visualisations générées dans le panneau central pour avoir une vue détaillée.	30
Figure 3.7	Décomposition d'une requête, selon la méthode de Munzner, en : «pourquoi», «quoi» et «comment». Les éléments «pourquoi» et «quoi» sont choisis par l'utilisateur (blocs bleus), tandis que le «comment» sera généré automatiquement (bloc vert) à ce dernier.	30
Figure 3.8	Zone de définition de requêtes : (A) Cases à cocher pour sélectionner une tâche parmi les 3 tâches proposées (« <i>Distribution</i> », « <i>Relationship</i> » et « <i>Time evolution</i> ») (B) Zone de saisie (avec autocomplétion) pour choisir une ou plusieurs questions de sondage ou métadonnées (C) Bouton de génération de visualisations selon de type de tâches et données choisi.	31
Figure 3.9	Exemple de bonne et mauvaise manipulation lors du choix de la tâche <i>Distribution</i> (<i>T1</i>).	31
Figure 3.10	Exemple de bonne et mauvaise manipulation lors du choix de la tâche <i>Time evolution</i> (<i>T3</i>).	32
Figure 3.11	Exemple de bonne et mauvaise manipulation lors du choix de la tâche <i>Relationship</i> (<i>T2</i>).	32
Figure 3.12	Exemple de la fenêtre générée comportant la visualisation.	33
Figure 3.13	Exemple de fenêtre contextuelle (<i>pop-up</i>) expliquant comment lire un graphique à barres.	34
Figure 3.14	Visualisations générées pour la tâche « <i>Distribution</i> »	37
Figure 3.15	Visualisations générées pour la tâche « <i>Relation</i> » (<i>Relationship</i>) dans le cas des relations bivariées.	38
Figure 3.16	Visualisations générées pour la tâche « <i>Relation</i> » (<i>Relationship</i>) dans le cas des relations multivariées.	38

Figure 3.17	Visualisations générées pour la tâche « Évolution temporelle » (<i>Time evolution</i>)	40
Figure 3.18	Illustration de l'interaction de sélection. L'utilisateur sélectionne la question sur le genre et décide de ne voir que les réponses des femmes. . .	40
Figure 3.19	Illustration de l'interaction de brosse sur un graphique de coordonnées parallèles.	41
Figure 3.20	Les 3 phases de l'interaction glisser-déposer. Phase 1 : Lors de leur génération, les visualisations sont, par défaut, indépendantes. Phase 2 : L'utilisateur peut choisir de lier deux vues grâce à l'interaction glisser-déposer, et ainsi créer un groupe de vues liées. Phase 3 : Les groupes peuvent être défaits au besoin.	42
Figure 3.21	Illustration de l'interaction glisser-déposer. Sur la gauche, les deux visualisations sont indépendantes. Sur la droite, les deux visualisations sont liées et peuvent interagir entre elles.	43
Figure 5.1	Résultats du questionnaire sur l'exploration libre	50
Figure 5.2	Résultats du questionnaire sur la génération de visualisations. La question Q4 a été ajoutée suite à notre première entrevue, d'où le nombre de 3 réponses seulement.	52
Figure 5.3	Résultats du questionnaire sur l'expérience utilisateur : utilité, facilité d'utilisation et satisfaction	55
Figure 5.4	Résultats du questionnaire sur l'expérience utilisateur	55

LISTE DES SIGLES ET ABRÉVIATIONS

AED	Analyse Exploratoire de Données
UdeM	Université de Montréal
ÉCSÉ	Étude Comparative des Systèmes Électoraux

LISTE DES ANNEXES

Annexe A	Analyse de travaux et abstraction de tâches	70
Annexe B	Guide de descriptions des visualisations générées	71
Annexe C	Courriel de recrutement	72
Annexe D	Formulaire d'information et de consentement	73
Annexe E	Questionnaire sur les caractéristiques du participant	79
Annexe F	Questionnaire sur l'exploration libre	80
Annexe G	Questionnaire sur la génération de visualisations	82
Annexe H	Questionnaire sur l'expérience utilisateur	83
Annexe I	Questions ouvertes et discussions	85

CHAPITRE 1 INTRODUCTION

Les sondages sont présents dans notre vie de tous les jours. Un sondage est un outil méthodologique d'observation composé d'une ou de plusieurs questions. Il peut être administré par téléphone, en ligne ou en personne. Il permet de sonder les caractéristiques d'une certaine population à partir d'un échantillon jugé statistiquement représentatif [17].

Les sondages sont utilisés dans plusieurs domaines et scénarios, par exemple :

- une entreprise qui désire collecter l'avis de ses clients pour améliorer ses services ;
- un professeur qui souhaite jauger la compréhension de ses étudiants à l'aide d'examens ;
- un parti politique qui veut connaître l'opinion publique afin de prendre des décisions stratégiques.

Lors de la réalisation d'un sondage, on passe par différentes étapes. Les auteurs du livre «*Designing Surveys*» présentent les 5 étapes suivantes [18] :

1. définir la population cible et la taille de l'échantillon ;
2. définir les objectifs du sondage ainsi que les questions à y intégrer ;
3. choisir la voie par laquelle le sondage sera administré et préparer le matériel requis ;
4. réaliser le sondage et recueillir les résultats¹ ;
5. trier et analyser les données puis publier les résultats.

Dans le cadre de ce projet, nous nous penchons sur l'exploration et l'analyse des données en utilisant des visualisations. En effet, l'analyse des résultats d'un sondage soutient la prise de décision en répondant aux questions comme : «Les résultats sont-ils statistiquement significatifs?», «Quelle est la tendance générale concernant la question x ?», «L'âge a-t-il une influence sur les réponses de la question y ?», etc. Cette analyse peut être effectuée, en partie, grâce à la visualisation de données. Une fois les éléments pertinents retenus, les résultats sont partagés avec les collaborateurs et preneurs de décision.

1. Dans certain cas, un prétest peut être effectué pour valider ou ajuster les étapes précédentes

1.1 Problématique

Grâce à la démocratisation des outils permettant la conception de sondages, il est devenu de plus en plus facile d'en réaliser. Nous retrouvons par exemple des outils tels que : *Google Form* [19], *SurveyMonkey* [20] et *SurveyGizmo* [21] qui offrent, en plus de la création et la diffusion de sondages, des visualisations basiques permettant de résumer les résultats d'un sondage.

Les principaux consommateurs de ce genre d'outils sont souvent des utilisateurs novices qui savent utiliser un ordinateur, mais qui n'ont pas beaucoup d'expérience en programmation et ne sont pas experts en interprétation de visualisations [22, 23]. Ce profil d'utilisateurs nécessite une attention particulière, car une étude analysant leurs comportements vis-à-vis de la visualisation a montré qu'ils choisissent souvent une visualisation en se basant sur leur expérience et familiarité [24]. En d'autres termes, ces derniers ne choisissent pas forcément la représentation la plus appropriée à leurs besoins, ce qui peut fausser leurs prises de décisions.

Les outils cités précédemment permettent principalement d'effectuer des analyses superficielles et répondre à des questions simples telles que : « Comment sont distribuées les réponses à la question x ? ». Dans le cas où un utilisateur novice souhaiterait effectuer une analyse ou exploration plus poussée, cela nécessiterait l'intervention d'un expert en analyse de données ou l'utilisation d'outils complémentaires. Cette approche est souvent coûteuse en termes de temps et d'argent.

Il est également important de présenter les défis relatifs à la nature des données. En effet, les sondages peuvent porter différents types de questions : questions à choix multiples, échelles d'évaluation, questions ouvertes, etc. Chaque cas de figure nécessite un type de réponse particulier : nominal, ordinal, textuelle, géographique, etc. De plus, nous rappelons que la façon dont le sondage est administré influe également sur la base de données finale. Par exemple, les sondages peuvent être administrés une seule fois (enquêtes transversales), ou de façon répétitive dans le temps ou l'espace (enquêtes longitudinales). Tous ces paramètres peuvent rapidement rendre l'exploration des données ambiguë pour un utilisateur novice.

Il y'a donc un écart entre les besoins d'un utilisateur novice en termes d'exploration et d'analyse de données de sondage, et les outils disponibles à cet effet. Il serait donc intéressant et pertinent de proposer un outil d'exploration visuelle offrant des possibilités d'analyse plus poussées.

Cet outil devra être intuitif et facile d'utilisation pour un utilisateur novice. Il devra l'assister dans son analyse et ses tâches d'exploration. De plus, l'outil devra supporter les principaux types de questions et enquêtes. Enfin, notre outil doit être évolutif afin de pouvoir supporter

des sondages à large nombre de questions et répondants, exemple : sondage de recensement.

1.2 Objectif général

L'objectif de cette recherche est de concevoir un outil qui permet à un utilisateur d'analyser et d'explorer visuellement ses données de sondage. Plus particulièrement, l'outil est destiné aux utilisateurs novices ayant une expérience limitée en programmation et visualisations de données.

Il serait pertinent, par exemple, pour des étudiants de premier cycle en sciences politiques. Ces derniers n'ont pas forcément les compétences suffisantes en programmation et en visualisations de données pour analyser des données de sondages à large nombre de questions et répondants. En utilisant notre outil, l'utilisateur novice devra être capable d'émettre ou d'évaluer des hypothèses sur ce type de données de manière efficace : rapide, facile et agréable.

Nous validons notre étude avec une expérience utilisateur réalisée auprès d'un groupe d'étudiants en sciences politiques.

1.3 Plan du mémoire

Le mémoire suivant est divisé en 6 chapitres. Le chapitre prochain (chapitre 2) englobe l'état de l'art en lien avec les outils d'exploration visuelle de sondage destinés aux utilisateurs novices. Dans le chapitre 3, on y présente les objectifs et hypothèses de recherche de ce travail. Aussi, on décrit la méthodologie suivie lors de la conception de notre outil. Le chapitre 4 englobe le protocole utilisé pour l'évaluation de la solution proposée. Dans le chapitre 5, les résultats de l'étude pilote sont présentés, suivis de discussions et recommandations. Enfin, le chapitre 6 conclut cette recherche en rappelant les idées principales, tout en proposant des pistes pour les travaux futurs.

CHAPITRE 2 REVUE DE LITTÉRATURE

Dans le chapitre qui suit, la revue de littérature des principaux travaux ayant guidé la conception de l’outil proposé est présentée. La première section rappelle les étapes de création et analyse d’une visualisation. La deuxième section survole les guides de la littérature concernant le design d’outils d’exploration destinés à des novices. La troisième section porte sur les outils permettant de visualiser les données de sondage. Enfin, la quatrième section présente de manière plus générale les outils de visualisation pour données multidimensionnelles.

2.1 Processus de création d’une visualisation

Cette section présente le processus de création d’une visualisation. Parmi les taxonomies existantes concernant la conception d’interface de visualisation, on présente celle retenue dans notre travail de recherche : la taxonomie proposée par Tamara Munzner [25]. Cette dernière, grandement utilisée dans la littérature [26–28], s’adapte à notre situation, car elle permet d’abstraire des tâches spécifiques afin de mener une analyse indépendamment du contexte d’application.

Avant la conception d’une visualisation, certaines contraintes de design (humaines et matérielles) doivent être identifiées. Nous devons par exemple définir l’utilisateur cible. Considérons le cas d’un utilisateur débutant, il risque d’avoir des besoins et attentes différents de ceux d’un expert dans son domaine. Le support utilisé (écran) pour présenter la visualisation est, quant à lui, une contrainte matérielle. En effet, selon la taille de l’écran, certaines visualisations peuvent être plus pertinentes que d’autres. Une fois ces contraintes identifiées et fixées, elles guideront nos choix futurs.

Lors du choix d’un design, certaines questions doivent être posées afin de structurer l’analyse du problème. Afin de définir une visualisation, Tamara Munzner [25] propose un cheminement de questions défini par un :

- Quoi : «Quelles sont mes données?» ;
- Pourquoi : «Quel est mon objectif en présentant mes données?» ;
- Comment : « Quelle est la meilleur façon de les présenter ? ».

En répondant à ces questions, le concepteur de la visualisation peut :

1. comprendre son jeu de données et définir explicitement sa nature ;
2. définir clairement les objectifs de la visualisation et de l’utilisateur cible ;

3. définir l'ensemble de tâches abstraites réalisées par l'utilisateur ;
4. trouver le meilleur moyen de représenter ses données afin de répondre aux objectifs identifiés.

Il est à noter que ce processus est effectué de façon itérative tout au long du design de la visualisation.

Dans ce travail de recherche, nous suivons cette décomposition du processus de création d'une visualisation. Dans le chapitre suivant, nous définirons les contraintes de design (section 3.1.2), puis nous répondrons aux : «Quoi» (section 3.1.1), «Pourquoi» (section 3.2) et «Comment» (section 3.5).

2.2 Considérations conceptuelles pour le design d'une visualisation

Pour ce projet, deux principaux aspects de design sont pris en considération. Notre objectif est de réaliser un outil d'**exploration visuelle** destiné aux **utilisateurs novices**. Dans cette section, nous présentons les guides et taxonomies disponibles dans la littérature concernant : (1) la conception d'outils d'exploration visuelle et (2) la conception d'outils de visualisation adaptés aux utilisateurs novices.

2.2.1 Exploration visuelle de données

Dans cette sous-section, on définit d'abord l'analyse exploratoire de données (AED) et ses piliers. Ensuite, nous survolons les taxonomies existantes dans la littérature concernant la conception d'une interface d'exploration visuelle de données.

L'analyse exploratoire de données (AED), proposée par John W. Tukey en 1977 [29], consiste en l'utilisation de représentations visuelles et approches statistiques afin d'extraire de l'information pertinente des données [30]. L'AED permet d'identifier des tendances ou relations entre les variables, en plus de formuler ou valider une hypothèse.

Elle se base sur 4 principaux piliers [31, 32] :

1. la visualisation de données ;
2. l'analyse résiduelle ;
3. la redéfinition ou transformation des données ;
4. les statistiques de base et non paramétriques.

Dans le cadre de ce projet, il sera question principalement de l'aspect visualisation de données. En effet, cette dernière joue un rôle primordial dans l'exploration de données, car elle

permet de synthétiser visuellement les données et ainsi faciliter la détection de tendances ou relations. Contrairement aux approches statistiques classiques qui mobilisent uniquement le système cognitif, la visualisation de données permet d'exploiter le cerveau humain en utilisant efficacement le système perceptif et cognitif [33].

Afin de tirer profit au maximum de cette force, la conception de l'outil de visualisation doit être bien pensée. Dans le cas de l'exploration visuelle, plusieurs taxonomies complémentaires ont été proposées dans la littérature. Ces taxonomies présentent les dimensions «Pourquoi» et «Comment» définies précédemment par Munzner [25]. Elles définissent l'ensemble de tâches abstraites qui seront réalisées par l'utilisateur (Pourquoi), suivies d'un ensemble de recommandations concernant les choix de conception (Comment).

Tâches abstraites de l'AED

Ben Shneiderman propose l'une des premières taxonomies. Selon lui il existe 7 principales tâches abstraites : «*overview, zoom, filter, details-on-demand, relate, history, and extracts*» [34]. Amar et coll. identifient ces tâches comme étant des actions telles que : «trier, filtrer, agréger, corréler, grouper, et dériver des attributs» [35].

Enfin, Tufféry [30] définit les tâches de l'analyse exploratoire visuelle comme étant :

1. la distinction d'attributs,
2. l'analyse univariée des données,
3. la détection d'interaction entre les variables (bivariée et multivariée),
4. la détection de données manquantes et aberrantes,
5. l'extraction de caractéristiques (*feature engineering*)

Recommandations de design pour un outil d'exploration de données

Ben Shneiderman introduit le *mantra* de la visualisation de données stipulant : «*Overview first, zoom and filter, details on demand*» [34]. En d'autres termes, il faut d'abord introduire le contexte et une vue générale des données, offrir la possibilité à l'utilisateur de zoomer et filtrer, et enfin lui laisser l'option d'aller plus en détail dans son exploration.

Dans leur étude, Ghosh et coll. [36] ont analysé une cinquantaine d'outils d'exploration visuelle. Ils en ont extrait les points à considérer dans la conception d'outils exploratoires pour les données multidimensionnelles. On retient les recommandations de design suivantes :

1. l'outil doit permettre de réaliser les principales tâches d'exploration : analyser la distribution et les relations entre les variables,
2. l'outil doit offrir la possibilité d'interagir avec les données en filtrant, ordonnant et groupant ses données,

3. l'outil doit être facile d'utilisation, accueillant et adapté pour des novices.

2.2.2 Outils de visualisation destinés aux utilisateurs novices

Cette sous-section définit d'abord l'utilisateur novice en informatique puis dans le contexte de la visualisation de données. On présente ensuite les recommandations existantes dans la littérature concernant le design d'un outil de visualisation pour ce profil d'utilisateur.

L'utilisateur novice en informatique est un utilisateur ayant de l'expérience avec un ordinateur, mais dont les compétences en programmation sont basiques ou inexistantes [22]. Plus particulièrement, l'utilisateur novice en visualisation de données est défini comme tout individu non expert en analyse visuelle de données [24]. Il a des connaissances basiques en visualisation qui se limitent principalement aux représentations utilisées dans sa vie de tous les jours.

Une première étude, classifiant les utilisateurs selon leurs compétences en visualisation de données [22], a indiqué qu'un utilisateur novice agit souvent comme consommateur. Ce dernier réalise rarement de l'exploration visuelle poussée et détaillée de ses données. De plus, l'étude de Grammel et coll. [24] a montré que lorsqu'un utilisateur novice construit ses visualisations, il rencontre beaucoup de difficultés à traduire et décomposer une hypothèse ou question en attributs de données et tâches abstraites. De plus, ce dernier a tendance à choisir les visualisations qui lui sont familières même si ces dernières ne sont pas optimales.

Les deux études précédentes [22, 24] offrent quelques recommandations à prendre en considération dans la conception d'une visualisation pour novices, entre autres :

- l'approche suggestive automatique de visualisations afin de familiariser l'utilisateur avec des visualisations alternatives plus optimales ;
- l'utilisation de requêtes textuelles pour la sélection d'attributs ;
- la définition de paramètres visuels par défaut (couleur, taille, échelle, etc.) ;
- la contextualisation de la visualisation en utilisant les éléments visuels appropriés : titre, légende, etc.

D'autres études [37–39] ont été réalisées dans le but de comparer le comportement des utilisateurs novices versus celui des experts dans la création de tableaux de bord. Ces résultats appuyaient les travaux précédemment cités en ajoutant quelques recommandations telles que :

1. la simplification du processus de création de visualisation en optant pour des approches utilisant un minimum de programmation, ex. : l'utilisation d'interactions de type *Drag-n-Drop* ;
2. la possibilité de modifier ses figures tout en ayant la flexibilité d'annuler une action ;

3. disponibilité des vues multiples pour comparaison ;
4. possibilité de sauvegarder et partager le tableau de bord et ses interactions avec les pairs.

Le tableau Tableau 2.1 offre une synthèse des principales recommandations à suivre quant à la conception d'un outil d'**exploration visuelle** pour **utilisateur novice** :

Catégorie	Sigle	Recommandations
Exploration visuelle	<i>R1</i>	Vue globale et détaillée des données.
	<i>R2</i>	Possibilité d'effectuer les principales tâches d'exploration.
	<i>R3</i>	Utilisation d'interactions permettant de filtrer, ordonner et grouper les données.
Utilisateur novice	<i>R4</i>	Pas ou très peu d'efforts en programmation lors de la création de visualisations.
	<i>R5</i>	Utilisation de paramètres par défaut (taille, couleur, légende, etc.)
	<i>R6</i>	Génération automatique ou suggestion de visualisation.
	<i>R7</i>	Simplification du processus de définition de requête.

Tableau Tableau 2.1 Tableau récapitulatif des principales recommandations de design d'un outil d'exploration visuelle pour utilisateur novice.

2.3 Revue des outils de visualisation de données de sondages

Les données de sondages sont souvent visualisées sous forme de tableaux et analysées en utilisant des outils statistiques [40]. Cependant, plusieurs taxonomies et travaux concernant la conception d'outils de visualisation spécifiques aux données de sondage existent dans la littérature. Dans cette section, nous présentons certains de ces travaux.

Une équipe de recherche brésilienne propose une taxonomie qui se base sur la littérature [41]. Elle permet de choisir efficacement une visualisation selon le type de questions. Par exemple, l'étude préconise l'utilisation du graphique à barres empilées au lieu du diagramme circulaire pour les questions à choix multiples ayant plus de 6 options de réponses. Ce travail de synthèse est pertinent, cependant cette taxonomie ne considère pas le cas d'analyses bivariées ou multivariées des données.

Une étude propose d'utiliser des glyphes pour visualiser les réponses d'un questionnaire [3]. Chaque réponse individuelle est représentée par une polyligne. Ces dernières sont ensuite superposées et saturées (figure Figure 2.1). Cette méthode permet de visualiser les réponses d'un individu, le profil type des répondants et la diversité de réponses. Cependant, elle n'est applicable que pour un type spécifique de questionnaires ayant toujours le même choix de réponses (ex. : grille à choix multiples).

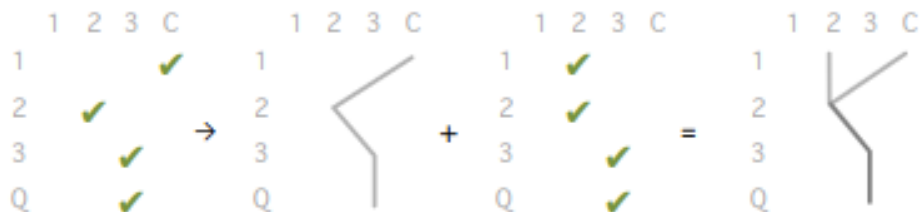


Figure Figure 2.1 Représentation de données de sondages sous forme de glyphes. © 2014 Les auteurs Computer Graphics Forum © 2014 The Eurographics Association and John Wiley & Sons Ltd. Publié par John Wiley & Sons Ltd [3]

Brodbeck et coll. proposent un outil permettant de représenter les résultats d'un questionnaire de satisfaction ayant 80 questions [4]. Cette approche propose de subdiviser les questions en sous-groupes et les visualiser sous forme de graphique des coordonnées parallèles (figure Figure 2.2). Ceci est un exemple représentatif d'outil d'exploration personnalisé. Il répond à des besoins spécifiques et nécessite une certaine expertise humaine pour subdiviser les questions.

Nous retrouvons également dans la littérature des solutions sous forme de tableaux de bord. Par exemple, l'outil *SurveyVis* permet de visualiser les résultats d'un sondage de satisfaction afin d'évaluer l'efficacité du système de santé publique en Inde [42]. Il utilise plusieurs visualisations connectées telles que :

1. des graphiques à barres et des cartes de chaleurs (*Heatmap*) pour avoir une vue générale des réponses ;
2. une carte pour visualiser la position géographique ;
3. un *slider* pour déterminer la fenêtre temporelle ;
4. un graphique des coordonnées parallèles (figure Figure 2.6) pour l'exploration des réponses individuelles.

Une équipe de recherche britannique a également proposé un outil de visualisation des résultats d'une enquête réalisée auprès de citoyens [5]. Cet outil exploratoire (figure Figure 2.3) permet de répondre à 3 principales questions : qui, où et comment. Tout cela en étant informatif, agréable à utiliser et accessible à plusieurs profils. Trois visualisations ont été choisies :

1. un graphique à barres verticales pour les questions utilisant l'échelle de Likert ¹ ;

1. L'échelle de Likert permet de sonder, selon une échelle comprenant 5 à 7 choix, le degré d'accord ou de désaccord relatif à une affirmation.

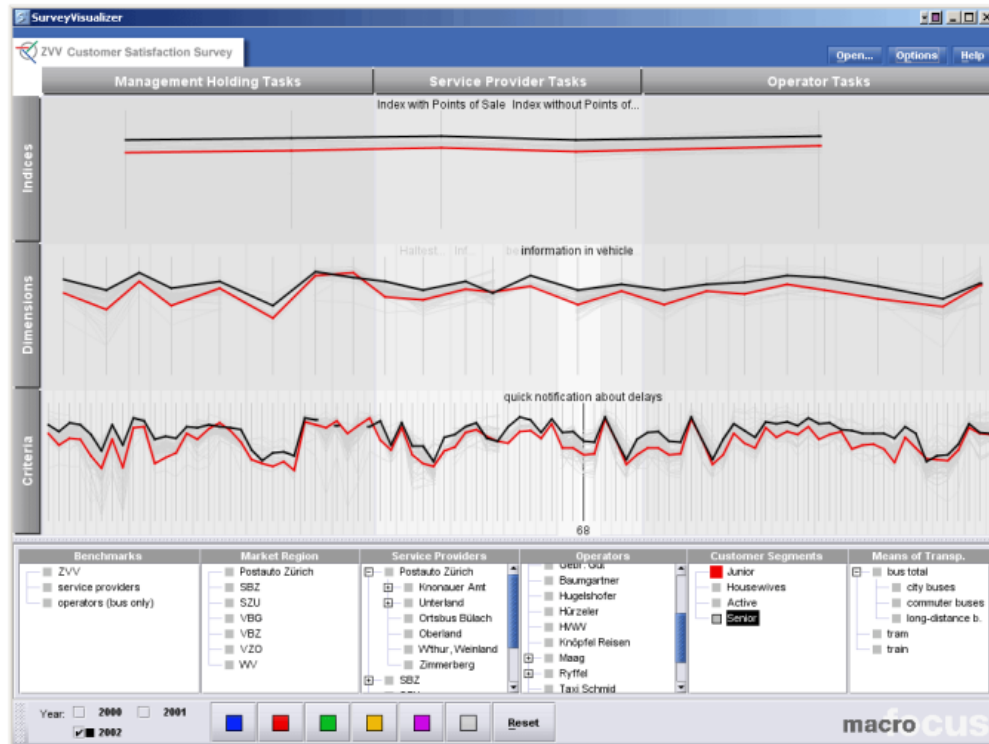


Figure Figure 2.2 Visualisation des réponses d'un questionnaire sur la satisfaction des clients en utilisant une topologie de coordonnées parallèles. © 2003 IEEE [4]

2. un graphique à barres horizontales pour les questions démographiques ;
3. une carte proportionnelle² (*Treemap*) pour représenter la position géographique.

Les deux derniers exemples sont inspirants, car ils visent un public d'utilisateurs novices en visualisation de données. Cependant, contrairement à notre approche qui se veut être plus généraliste en termes d'exploration, ces exemples répondent à des tâches et données spécifiques.

Enfin, Zheng a développé un outil, destiné aux utilisateurs novices, permettant de visualiser les résultats de questionnaires [6]. Cet outil est plus générique que les deux précédents. Il génère automatiquement des graphiques en se basant sur le type et le nombre de questions sélectionnées (figure Figure 2.4). Zheng a réalisé une étude pilote afin de valider ses hypothèses. Cette dernière a montré que les utilisateurs de son outil arrivent à créer des graphiques complexes plus facilement et ainsi allouer plus de temps à l'analyse et l'exploration de données. Cependant, cette même étude pilote a soulevé quelques limitations qui seront prises en considération dans la conception de notre outil :

2. La carte proportionnelle permet d'encoder des données hiérarchiques en utilisant des rectangles imbriqués de différentes tailles et couleurs.



Figure Figure 2.3 Interface de l'outil permettant de visualiser les résultats d'un sondage réalisé auprès de citoyens. Droits gérés par Taylor et Francis [5]

1. avec uniquement six choix de visualisations, les utilisateurs se trouvent limités en termes d'explorations ;
2. les tâches prises en considération sont la distribution et la corrélation. Or, l'étude pilote a montré que l'analyse de l'évolution temporelle d'une question était une tâche primordiale, particulièrement dans le cas d'études comparatives ;
3. l'outil manque de support pour un utilisateur novice. L'auteur recommande l'utilisation d'annotations interactives ou d'un guide afin d'aider l'utilisateur à mieux s'approprier l'outil.

2.4 Revue des outils de visualisation de données multidimensionnelles

En vue du nombre réduit de résultats concernant les outils de visualisations de données de sondages, nous avons jugé pertinent d'examiner la littérature traitant des outils d'exploration visuelle de données multidimensionnelles. En effet, les sondages peuvent être considérés comme étant des données multidimensionnelles pour deux raisons : (1) les sondages comportent généralement plus que deux questions, et (2) différents types de réponses peuvent être attendues (numériques, catégoriques, textuelles, etc.) selon le type de question posée. Dans cette section, nous présenterons les outils proposés par la littérature. Nous prendrons

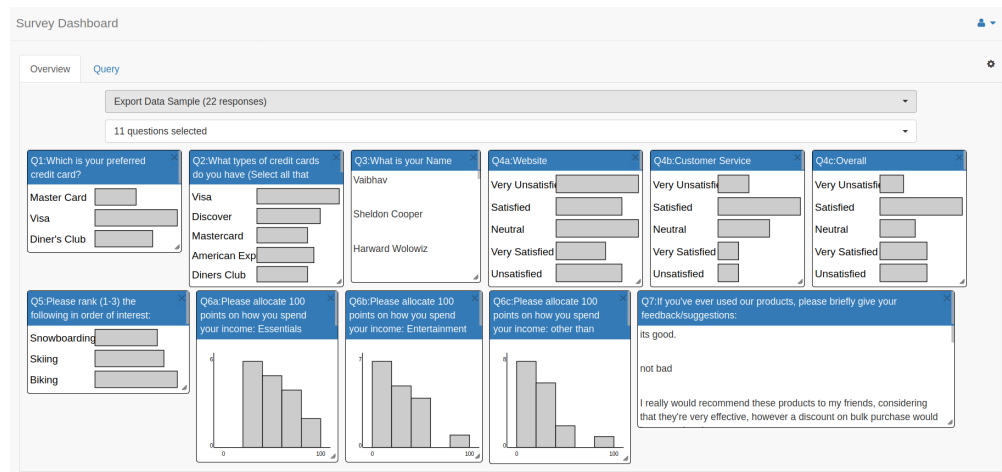


Figure Figure 2.4 Interface de l'outil destiné aux utilisateurs novices, permettant de visualiser les résultats d'un questionnaire [6]. Capture d'écran réalisée par l'auteur.

en considération les visualisations à une seule vue (*single view*) et celles à vues multiples (*multiple views*).

2.4.1 Visualisations à une seule vue (*Single view*)

Un système à une seule vue permet d'afficher une donnée selon une représentation (tabulaire, graphique, etc.) unique et distincte [43].

Dans cette catégorie, nous retrouvons par exemple la matrice de *Scatter plots* (SPLOM) [8]. Cette méthode consiste à visualiser sous forme de matrice toutes les combinaisons possibles de graphes. Cette méthode est pertinente quand il faut détecter des tendances et corrélations entre deux variables (figure Figure 2.5). Cependant, une limitation à cette approche est son évolutivité, en effet, plus le nombre de dimensions augmente et plus le nombre de combinaison augmente de façon quadratique. Plusieurs travaux, tels que «Corrgram» [44], présentent à cet effet des techniques de réarrangement automatique de l'ordre des dimensions afin de faciliter l'interprétation. Une autre limitation de cette méthode est que l'on ne peut visualiser que les relations bivariées. De plus, il y'a une redondance de certaines représentations.

Une autre méthode très répandue pour visualiser des données multidimensionnelles est le graphique des coordonnées parallèles [9]. Cette représentation consiste à visualiser plusieurs dimensions à travers des axes parallèles. Chaque élément est représenté par une ligne polygonale (figure Figure 2.6). Contrairement à la matrice de *Scatter plots*, cette approche permet de visualiser les relations entre plus que deux dimensions à la fois. En effet, pour n -dimensions nous avons $n!$ permutations ou ordonnancements possibles. Cela est également une des limi-

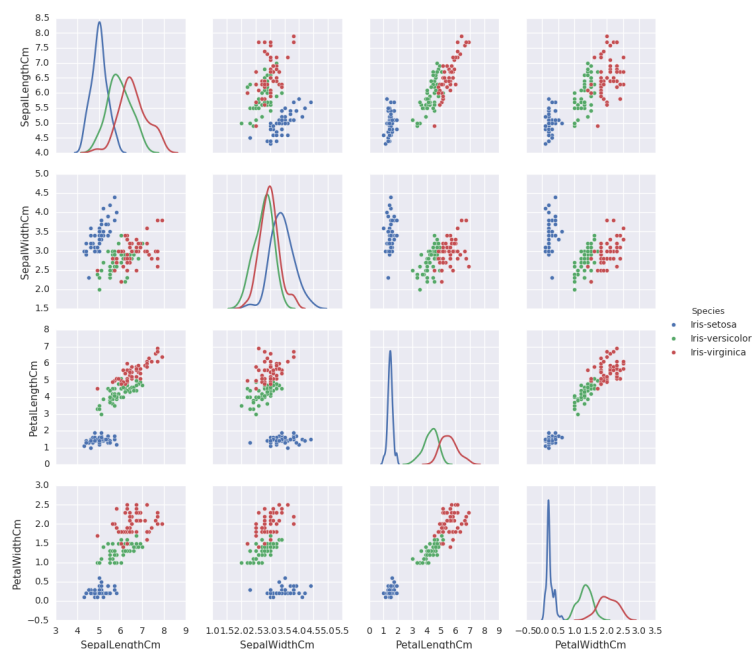


Figure Figure 2.5 Matrice de *Scatter Plot* en utilisant la base de données Iris [7]. Figure reproduite par l'auteur. [8]

tations de cette approche, car l'ordre peut fortement biaiser l'information visualisée [45, 46]. De plus, le bruit visuel (*visual clutter*) [45] augmente avec le nombre de dimensions ce qui rend l'interprétation des données plus compliquée. Une variante aux coordonnées parallèles est le *Parallel sets* [11]. Cette dernière s'applique aux données catégoriques (figure Figure 2.7).

L'utilisation de glyphes est une option intéressante dans le cas de données multidimensionnelles. Cela consiste à visualiser des symboles graphiques (visage, fleurs) en jouant sur les caractéristiques du symbole (forme, taille, couleur, etc.) afin d'encoder l'information, par exemple : le visage de Chernoff [12] (figure Figure 2.8). Les glyphes permettent de faire ressortir des tendances et regrouper les éléments similaires. Cependant, ils sont difficiles à interpréter avec l'augmentation du nombre de dimensions.

Toutes ces solutions *single view* rencontrent des limitations en termes de résolutions d'écran. Elles deviennent complexes à interpréter et explorer pour un utilisateur novice quand le nombre de dimensions augmente. Cela peut être évité grâce à l'utilisation d'interactions ou de vues multiples [47, 48].

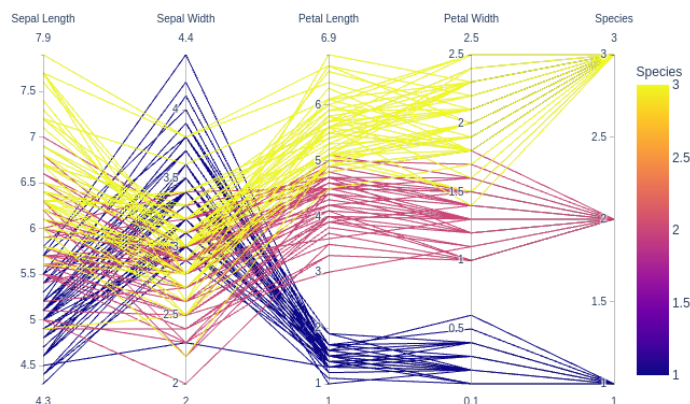


Figure Figure 2.6 Méthode des coordonnées parallèles en utilisant la base de données «Iris» [7]. © 1990, IEEE [9]

2.4.2 Visualisations à vues multiples (*Multiple views*)

Un outil de visualisation à vues multiples consiste à assembler plusieurs représentations complémentaires, synchronisées et connectées entre elles [43]. Ce type d'approche est très utilisé dans le cas de l'exploration visuelle de données multidimensionnelles hétérogènes volumineuses [49, 50].

L'utilisation de vues multiples est souvent accompagnée d'interactions comme :

- la sélection synchronisée entre les vues ;
- le filtrage selon une ou plusieurs dimensions ;
- la manipulation de dimensions : l'utilisateur peut ajouter, enlever ou changer l'ordre d'une dimension, etc.

Nous présenterons trois exemples d'outils à vues multiples : *Keshif*, *Lineup* et *Domino*.

L'outil *Keshif* [13] permet de visualiser sous forme de tableau de bord (figure Figure 2.9) toutes les variables d'une base de données. Le tableau de bord est généré automatiquement selon le type de variable (temporelle, géographique, numérique, etc.).

Cet outil offre plusieurs interactions permettant d'effectuer du filtrage synchronisé, de comparer plusieurs filtres en même temps grâce à une option de verrouillage, etc.

L'outil *Lineup* [14] démontre qu'il est possible de visualiser des données sous forme de tableau si cela est fait avec la bonne visualisation. Chaque variable est visualisée dans sa colonne selon un encodage spécifique. L'outil offre également la possibilité de regrouper et agréger les données avec un système de filtrage avancé (figure Figure 2.10). La limitation à cet outil

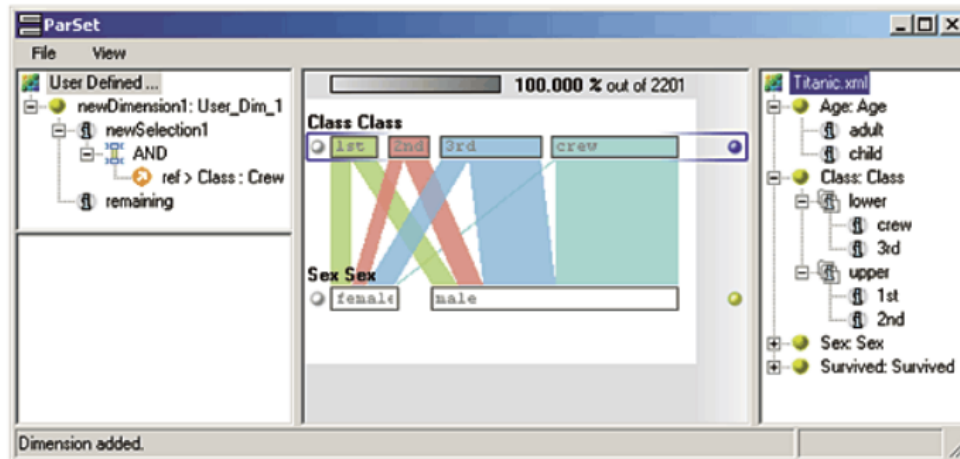


Figure 2.7 Méthode des *Parallel Sets* en utilisant la base de données «Titanic» [10].
© 2006, IEEE [11]

est qu'en gardant une forme tabulaire, toutes les lignes de données sont visibles. Dans le cas d'une grande base de données (ex. : sondage de recensement), ceci peut vite poser des problèmes en termes de bruit visuel et ressources informatiques.

Enfin, l'outil *Domino* [15] permet d'arranger et combiner, tels des blocs, des données multidimensionnelles hétérogènes (figure Figure 2.11). Cet outil est particulièrement pertinent pour un utilisateur novice, car il utilise une interaction de type glisser-déposer et ne nécessite aucune connaissance en programmation ni en visualisation, car les visualisations sont générées automatiquement. Cependant, *Domino* se limite à définir uniquement des relations entre les variables. Cette dernière ne représente qu'une tâche parmi d'autres tâches d'exploration.

Cette liste non exhaustive nous permet de déterminer les éléments principaux dans le design d'un outil de visualisation pour données multidimensionnelles.

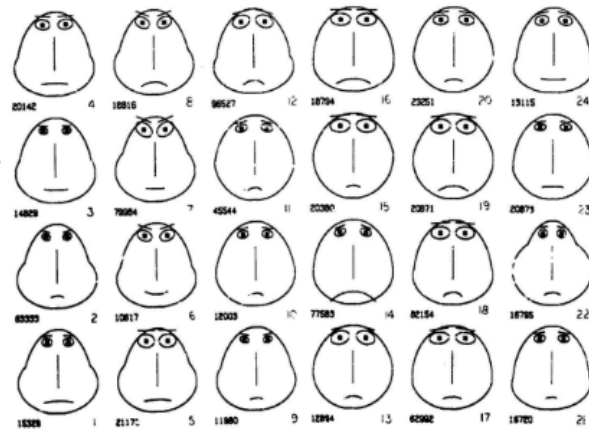


Figure Figure 2.8 Glyphes de type visage de Chernoff. Droits gérés par Taylor et Francis [12]

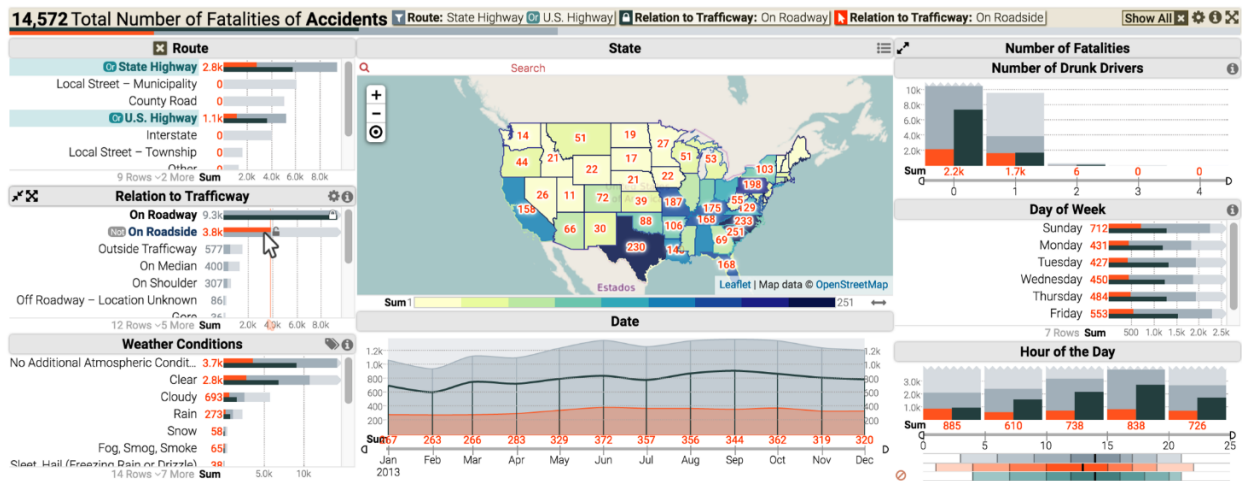


Figure Figure 2.9 Interface de *Keshif* présentant les données sur le nombre d'accidents mortels aux États-Unis en 2013. © 2018, IEEE [13]



Figure Figure 2.10 Interface de *Lineup* présentant les 2000 meilleures compagnies. © 2013, IEEE [14]

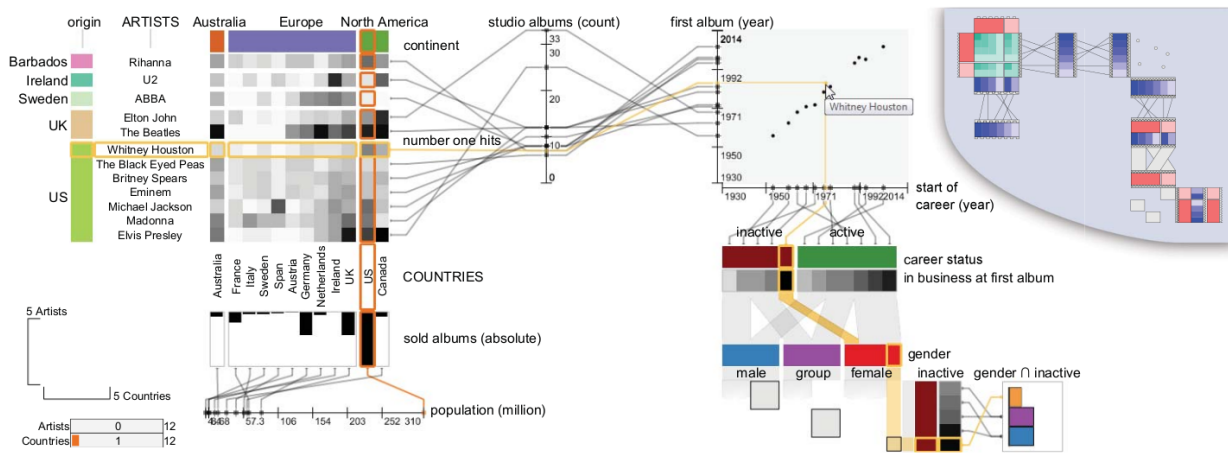


Figure Figure 2.11 Interface de *Domino* présentant des relations entre les variables d'une base de données musicale. © 2014, IEEE [15]

CHAPITRE 3 MÉTHODOLOGIE PROPOSÉE

Dans ce chapitre, on expose la méthodologie suivie dans le développement de notre outil interactif d'exploration de données de sondages. Nous adopterons le processus de création d'une visualisation proposé par Munzner (quoi, pourquoi et comment). D'abord, nous présenterons le contexte de notre application : le type de données (quoi) et l'utilisateur cible. Ensuite, nous définirons les principales tâches d'explorations que l'utilisateur cible souhaiterait effectuer (pourquoi). Aussi, nous rappellerons les recommandations, retenues de la littérature, pour la conception d'un outil d'exploration visuelle destiné à un utilisateur novice. Une fois ces paramètres fixés, nous soumettrons nos hypothèses et objectifs de recherche. Puis, le modèle de visualisation et ses fonctionnalités seront décrits en détail (comment). Finalement, les choix d'implémentation de l'outil de visualisation seront explicités.

3.1 Contexte d'application

Le contexte d'application permet de fixer les paramètres et contraintes qui guideront la conception de notre outil. Nous définirons, dans cette section, la nature de nos données puis le public visé.

3.1.1 Données

Dans le cadre de ce projet, l'attention s'est portée sur les données de sondages. Nous allons d'abord survoler les différents types de sondages puis les différents types de questions de sondage. Enfin, nous parlerons des métadonnées qui permettent une meilleure compréhension du contexte d'un sondage.

Types de sondages

Selon l'objectif de l'étude, il existe deux principaux types de sondages [51] : l'approche **transversale** et l'approche **longitudinale**. L'approche transversale consiste à faire passer le sondage une seule fois dans le temps. L'approche longitudinale permet de faire un suivi dans le temps en effectuant plusieurs observations ou rencontres. Ce suivi peut être fait avec exactement les mêmes participants (panel) ou avec des participants partageant certaines caractéristiques (cohorte). La figure Figure 3.1 permet de résumer les différents types de sondages.

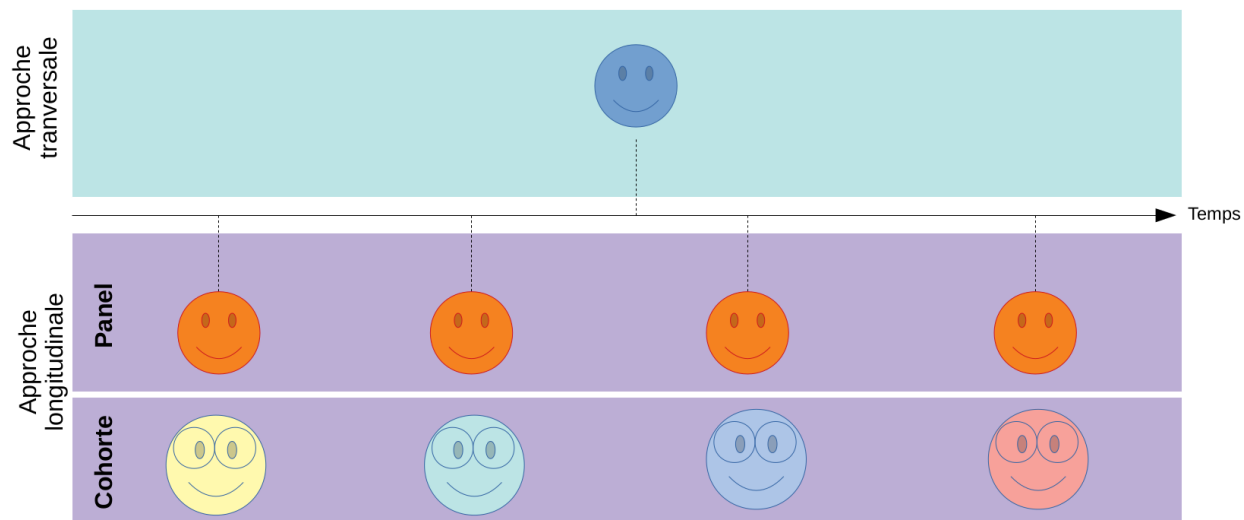


Figure Figure 3.1 Figure résumant les différents types de sondage. De haut en bas : approche transversale, approche longitudinale à panel et approche longitudinale à cohorte.

Types de questions de sondages

Lors de l'administration d'un sondage, il existe plusieurs types de questions possibles [52]. Chaque question nécessite un type de réponse particulier. Le tableau Tableau 3.1 résume les types de questions de sondage pris en considération, dans un premier temps, pour notre étude.

Type de questions	Exemple de question	Exemple de réponses	Type de variable
Questions fermées à choix unique	Quelle est votre parfum de glace préféré ?	Vanille, Chocolat ou Fraise	Nominale
Échelle d'évaluation	Sur une échelle de 1 à 10, combien noteriez-vous la qualité de notre service ?	De 1 à 10	Ordinale
Échelle de Likert	Êtes-vous satisfait par la qualité de nos services ?	De «Très déçu» à «Très satisfait»	Ordinale
Questions ouvertes à valeurs numériques	Combien de vols avez-vous effectués au cours des 10 dernières années ?	Un nombre entier positif	Numérique

Tableau Tableau 3.1 Types de questions pris en considération dans notre étude

Les métadonnées

Les métadonnées sont des éléments essentiels dans la compréhension d'un sondage. Elles permettent de définir le contexte d'administration du sondage : l'année, le lieu, l'organisme, etc. [53]. Elles sont particulièrement importantes dans le cas d'études longitudinales ou comparatives.

3.1.2 Utilisateurs cibles

Les sondages sont utilisés par une variété de profils pouvant aller du marqueteur à l'employeur en passant par le sociologue. Ces derniers ont souvent recourt à des analyses statistiques, parfois réalisées par un tiers parti, pour interpréter les résultats de leur étude.

Pour ce projet de recherche, nous ciblons les utilisateurs qui étudient régulièrement les données de sondage, mais qui ont peu ou pas de connaissances en visualisation de données. Ils ne sont donc pas experts ou n'ont suivi aucune formation en visualisation de données.

Cela pourrait, par exemple, être des étudiants en sciences politiques en premier cycle d'études universitaires. Ces derniers étudient les données de sondages, mais ont cependant un bagage limité en visualisation de données, car cela ne fait pas forcément partie de leur programme d'étude.

De manière générale, nous visons les utilisateurs novices dans le but d'améliorer et enrichir leurs connaissances en visualisation de données. En effet, l'une des caractéristiques des utilisateurs novices est qu'ils ont tendance à choisir leurs visualisations en fonction de leur familiarité avec celle-ci [24], même si celle-ci n'est pas optimale pour la tâche qu'ils visent à accomplir. En utilisant notre outil, nous générerons aux utilisateurs novices des visualisations basées sur l'état de l'art [41] et un processus rigoureux de création de visualisation [25] afin d'éduquer et enrichir leurs connaissances en visualisation de données.

3.2 Définition et abstraction des tâches

Maintenant que nous avons cerné la nature de nos données («Quoi») et l'utilisateur cible, nous devons définir la liste de tâches que l'utilisateur aura à effectuer («Pourquoi»).

Étant dans le contexte de l'exploration visuelle, il a été conclu de la littérature [30] que les principales tâches seraient :

1. la distinction d'attributs ;
2. l'analyse univariée des données ;
3. la détection d'interaction entre les variables (bivariée et multivariée) ;

4. la détection de données manquantes et aberrantes ;
5. l'extraction de caractéristiques (*feature engineering*).

Nous avons conjointement travaillé avec des chercheurs de la chaire de recherche en études électorales de l'UdeM. Ces derniers utilisent régulièrement les sondages dans leurs travaux afin d'analyser le comportement des citoyens. Au total, nous avons effectué 4 entrevues (5 h au total) avec un professeur et deux étudiants de la chaire. L'objectif était de valider ou compléter la liste de tâches, tirée de la littérature, en discutant avec eux des tâches d'exploration réalisées le plus fréquemment dans leurs analyses.

À travers nos rencontres, les motivations du projet furent exposées et acquiescées par nos collaborateurs. Cependant, au fur et à mesure des entrevues, nous avons remarqué qu'il était difficile pour eux de décrire leurs tâches habituelles de manière moins technique ou abstraite. Afin d'y remédier, l'initiative a été prise de revoir certains de leurs travaux utilisant les données de sondage [54–64], dans le but d'extraire les tâches abstraites les plus récurrentes.

La méthode de Munzer [25] fut utilisée ici afin d'abstraire et traduire une tâche spécifique telle que : «Je veux savoir si les femmes votent plus que les hommes.» en tâche abstraite telle que : «Relation bivariée entre deux variables». Cela permet de dissocier la tâche à effectuer du domaine d'expertise et ainsi permettre une généralisation de la visualisation. L'analyse complète est disponible dans l'annexe A.

De cette analyse, trois principales tâches sont ressorties :

- (*T1*) : l'analyse de la distribution d'une variable ;
- (*T2*) : l'analyse de la relation entre des variables ;
- (*T3*) : l'évolution temporelle d'une variable.

3.3 Recommandations pour le design d'un outil d'exploration visuelle destiné à un utilisateur novice

Comme il a été vu dans la revue de littérature, nous rappelons dans cette section les principales recommandations retenues pour le design d'un outil d'exploration visuelle destiné à un utilisateur novice. Ces recommandations nous guideront dans nos choix de design (tableau Tableau 3.2).

3.4 Hypothèses et objectifs

Les paramètres suivants sont pris en considération dans la conception de l'outil proposé :

Catégorie	Sigle	Recommandations
Exploration visuelle	<i>R1</i>	Vue globale et détaillée des données.
	<i>R2</i>	Possibilité d'effectuer les principales tâches d'exploration.
	<i>R3</i>	Utilisation d'interactions permettant de filtrer, ordonner et grouper les données.
Utilisateur novice	<i>R4</i>	Pas ou très peu d'efforts en programmation lors de la création de visualisations.
	<i>R5</i>	Utilisation de paramètres par défaut (taille, couleur, légende, etc.)
	<i>R6</i>	Génération automatique ou suggestion de visualisation.
	<i>R7</i>	Simplification du processus de définition de requête.

Tableau 3.2 Tableau récapitulatif des principales recommandations de design d'un outil d'exploration visuelle pour utilisateur novice.

1. l'utilisateur est novice et a des connaissances limitées en programmation et visualisation de données ;
2. les deux types de sondages (longitudinal et transversal) sont pris en considération ;
3. les questions fermées à choix unique, échelle d'évaluation, échelle de Likert et questions ouvertes à valeurs numériques sont prises en considération dans notre étude ;
4. nous considérons la présence de métadonnées.

À partir des points cités précédemment et la revue de littérature, nous avons défini certains objectifs et hypothèses à ce projet de recherche.

3.4.1 Objectifs de recherche

L'outil de visualisation a été conçu pour répondre à un ou plusieurs des objectifs suivants :

(O1) : Générer des visualisations adaptées au type de tâches et questions de sondage choisies par l'utilisateur novice ;

(O2) : Traduire des hypothèses de recherche en visualisations de manière simple et intuitive ;

(O3) : Effectuer les principales tâches d'exploration.

Tenant compte de ces objectifs, nous proposons de développer un tableau de bord qui offre une vue globale et détaillée des données.

Cet outil permettra de définir facilement une requête pour générer automatiquement une visualisation prédéfinie selon le type de tâche et variable choisi (question de sondage). En effet, l'une des caractéristiques de l'utilisateur novice en visualisation de données est qu'il rencontre souvent des difficultés à traduire une question de recherche en visualisation. Pour

y parvenir, des taxonomies, telles que le processus de conception de Tamara Munzner [25], sont généralement utilisées par des experts. Nous présentons donc une manière originale de définir les requêtes, comme le ferait un expert en visualisation de données, qui soit adaptée aux novices.

Enfin, il utilisera des interactions telles que le filtrage pour améliorer l'expérience utilisateur. Le tableau Tableau 3.3 représente l'association entre chaque élément de design retenu et nos objectifs.

Recommandation	Design retenu	Objectifs		
		<i>O1</i>	<i>O2</i>	<i>O3</i>
<i>R1, R2</i>	Vue globale et détaillée d'un sondage			✓
<i>R4, R5, R6, R7</i>	Génération automatique de visualisation	✓	✓	
<i>R3</i>	Interactivité (filtrage, <i>Drag-n-Drop</i> , etc.)			✓

Tableau Tableau 3.3 Association entre chaque élément de design retenu et ses objectifs.

3.4.2 Hypothèses de recherche

L'hypothèse scientifique générale est que notre outil permet à un utilisateur novice d'effectuer son analyse et ses tâches d'exploration de manière intuitive et simple.

Une étude utilisateur (*user-study*) formelle avec un échantillon de participants humain aura pour but de confirmer ou infirmer les sous-hypothèses suivantes en rapport avec nos choix de design :

(*H1*) : Une vue globale dans la visualisation montrant toutes les questions de sondages et les métadonnées permet à l'utilisateur d'émettre de nouvelles hypothèses ;

(*H2*) : Un choix explicite de tâches et questions de sondage permet à un utilisateur novice de traduire ses hypothèses en visualisation ;

(*H3*) : La prédéfinition de visualisations par tâches et types de variable permet l'utilisateur d'explorer efficacement son jeu de données ;

(*H4*) : La présentation de potentielles nouvelles visualisations à l'utilisateur novice lui permet d'améliorer ses connaissances en visualisations de données ;

(*H5*) : Des interactions de type glisser-déposer et filtrage améliorent l'expérience d'un utilisateur novice.

3.5 Outil de visualisation : Design et fonctionnalités

Dans la section qui suit, nous présenterons en détail chaque élément de design retenu et cité dans le tableau Tableau 3.3. Nous commencerons par l'organisation de notre plate-forme permettant une vue globale et détaillée des données. Ensuite, nous exposerons notre approche de génération automatique de visualisations. Finalement, nous parlerons des différentes interactions utilisées dans notre outil.

3.5.1 Vue globale et détaillée d'un sondage

Le premier aspect à prendre en considération dans le design de notre outil était sa composition : la manière de présenter ou distribuer l'information sur l'outil. En se basant sur le principe de Shneiderman [34] stipulant : « *overview first, zoom and filter, details on demand* », il a été décidé de diviser l'outil de visualisation en 3 sections (figures Figure 3.2 et Figure 3.3) : **un panneau latéral**, **un panneau central** et **un pied de page**. Chaque section permet de visualiser des types de données ou niveaux de détails particuliers. En d'autres termes, nous avons utilisé des vues multiples (*multiple views*).



Figure Figure 3.2 Vue réduite de l'outil développé utilisant les données normalisées de l'étude comparative des systèmes électoraux [16]. Le panneau latéral (A) offre une vue globale de la distribution des questions du sondage. Le pied de page (B) est dans ce cas réduit. Le panneau central (C) présente la zone de formulation de requêtes et générations de visualisations.



Figure 3.3 Vue complète de l'outil développé utilisant les données normalisées de l'étude comparative des systèmes électoraux [16]. Le panneau latéral (A) offre une vue globale de la distribution des questions du sondage. Le pied de page (B) offre une vue globale de la distribution des métadonnées. Le panneau central (C) présente la zone de formulation de requêtes et générations de visualisations.

Panneau latéral

Le panneau latéral offre une vue globale de toutes les données du sondage et cela de manière à suivre la recommandation *R1* (voir tableau Tableau 3.3). Il permet de voir le nom de toutes les questions du sondage et la distribution de leurs réponses sous forme de petits graphiques. La motivation derrière ce choix de design est qu'un utilisateur qui découvre un sondage pour la première fois pourrait avoir besoin de savoir combien de questions son sondage comporte et comment les réponses sont distribuées dans l'ensemble.

Afin de garder une certaine homogénéité, deux représentations ont été choisies pour visualiser les données sur le panneau latéral (figure Figure 3.4) :

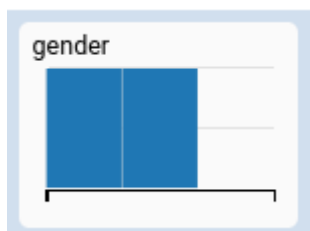
1. les données nominales telles que les questions fermées à choix unique (ex. : « *Quel est votre genre ?* ») sont représentées sous forme de graphique à barres ;
2. les données ordinales ou numériques telles que les échelle d'évaluation, échelle de Likert et questions ouvertes à valeurs numériques (ex. : « *Sur une échelle de 1 à 10, combien noteriez-vous l'efficacité de ce produit ?* ») nécessitant des valeurs qui peuvent être ordonnées, sont représentées sous forme de boîte à moustaches.

Le tableau Tableau 3.4 résume les représentations choisies.

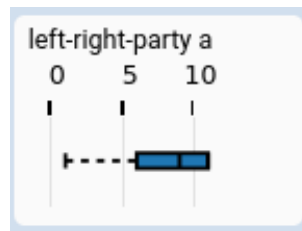
Question	Type	Domaine de réponses	Représentations
Quel est votre genre ?	Nominal	{Homme, Femme, Autre}	Graphique à barres
Sur une échelle de 1 à 10, combien noteriez-vous l'efficacité de ce produit ?	Ordinal	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	Boîte à moustaches
Quel âge avez-vous ?	Numérique	$[0, n]$ où n est l'âge maximal	Boîte à moustaches

Tableau Tableau 3.4 Exemple de questions de sondage et leurs représentations associées

Cependant, l'utilisation de la boîte à moustaches pour représenter les questions de type échelle de Likert présentait quelques divergences dans la littérature [65–67]. En effet, certains jugent qu'il ne faut pas considérer les réponses d'une question de type échelle de Likert comme une valeur ordinale. Le principal argument des détracteurs est que sa moyenne arithmétique n'a pas de sens statistique [67] et donc ne peut être représentée avec une boîte à moustache. Le tableau Tableau 3.5 présente un exemple de réponses d'une question de type échelle de Likert. La moyenne arithmétique de cet exemple est de 3,5. Selon les détracteurs de cette pratique, une moyenne de 3,5 se traduirait par «Ni en désaccord ni d'accord et demi». Ce



(a) Graphique à bandes pour données nominales



(b) Boîte à moustaches pour données ordinales/numériques

Figure Figure 3.4 Représentation choisie par type de données

qui ne fait statistiquement pas de sens.

Question	Total	(1) Tout à fait d'accord	(2) D'accord	(3) Ni en désaccord ni d'accord	(4) Pas d'accord	(5) Pas du tout d'accord
J'aime le chocolat.	10	1	1	3	2	3

Tableau Tableau 3.5 Exemple de réponses d'une question de type échelle de Likert. La moyenne arithmétique est de 3,5, ce qui se traduirait par «Ni en désaccord ni d'accord et demi». Cette valeur ne fait pas de sens statistiquement.

D'un autre côté, les partisans de cette approche proposent un ensemble de recommandations à suivre [65] et proposent d'opter pour d'autres mesures pour analyser ce type de données tels que la moyenne géométrique [66]. Afin de garder une certaine homogénéité, nous avons décidé, pour cette première version, de considérer les questions de type échelle de Likert comme données ordinales et de les représenter à l'aide de boîtes à moustaches. Nous avons néanmoins suivi les recommandations proposées et avons utilisé des mesures complémentaires : moyenne géométrique, minimum, maximum, etc.

Pied de page

La fonction et l'objectif de la zone de pied de page sont similaires à ceux du panneau latéral. Cependant, cette zone prend en charge les métadonnées (figure Figure 3.5). Contrairement au panneau latéral, celui-ci n'est pas fixe et peut être caché, car les métadonnées sont des éléments complémentaires pour la compréhension d'un contexte et donc moins prioritaires que les données de sondage.

Pour garder une homogénéité dans les représentations, la convention suivante a été suivie :

1. Données temporelles : graphique en ligne ;
2. Données nominales : graphique à barres ;

3. Données numériques et ordinales : boîte à moustache.

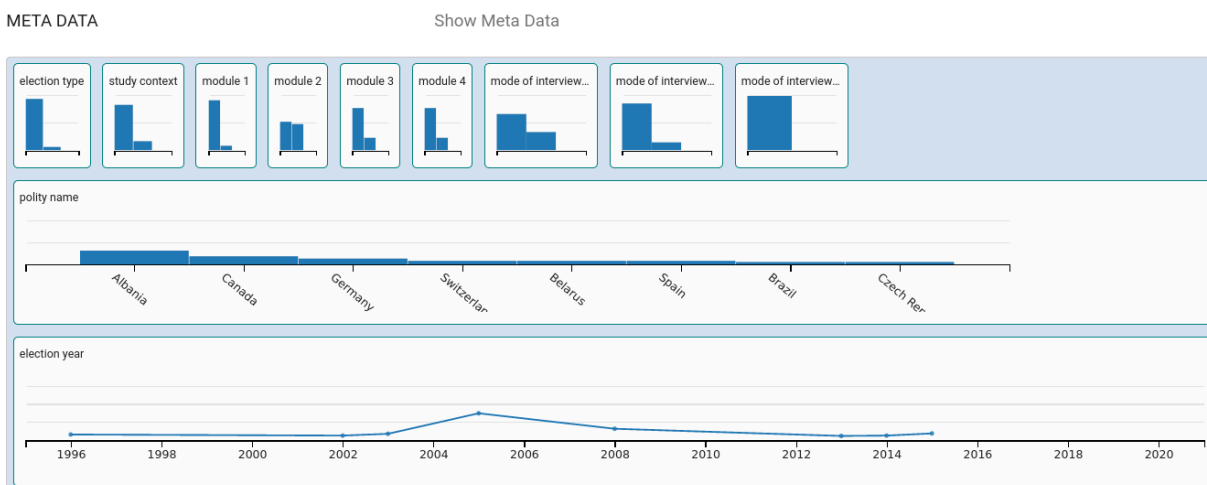


Figure Figure 3.5 Exemple de pied de page comportant des métadonnées telles que le nombre de participants par pays, année ou type d'élection.

Panneau central

Le panneau central (figure Figure 3.6) offre une vue plus détaillée des données. Ce panneau a été conçu pour répondre à 3 objectifs :

1. il doit permettre l'utilisateur d'effectuer des tâches d'analyse plus spécifiques : $T1$, $T2$, $T3$;
2. il doit permettre un utilisateur novice de traduire aisément une question de recherche en visualisation ;
3. il doit faciliter le processus de création en générant automatiquement une visualisation selon les besoins de l'utilisateur (ce processus sera détaillé dans la section 3.5.2).

Ce panneau utilise le principe des vues multiples pour offrir à l'usager la flexibilité de créer différents scénarios. Il est subdivisé en une zone de définition de requêtes et une zone de création des visualisations. Il prend en considération les recommandations $R2$, $R4$, $R5$, $R6$ et $R7$ du tableau Tableau 3.3.

Zone de définition de requêtes

La zone de définition de requêtes propose une approche qui offre la possibilité à l'utilisateur de présenter ses requêtes sous forme semi-textuelle. En temps normal, on exprime une requête de la forme : « *Je veux voir la distribution de la question sur l'âge* ». Nous nous inspirons de cette syntaxe pour décomposer une requête en « *quoi* » et « *pourquoi* » (figure Figure 3.7).

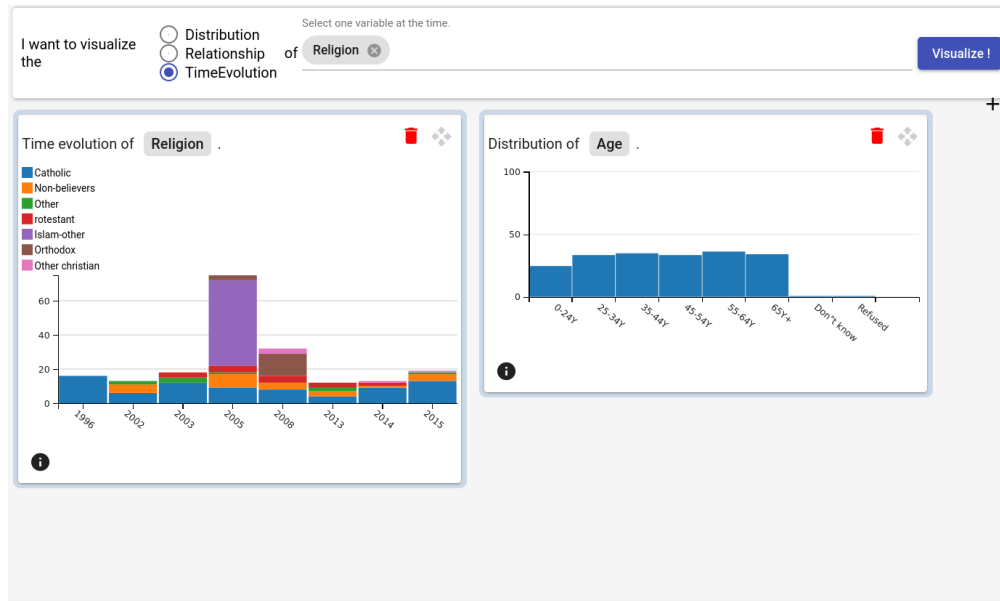


Figure Figure 3.6 Exemple de visualisations générées dans le panneau central pour avoir une vue détaillée.

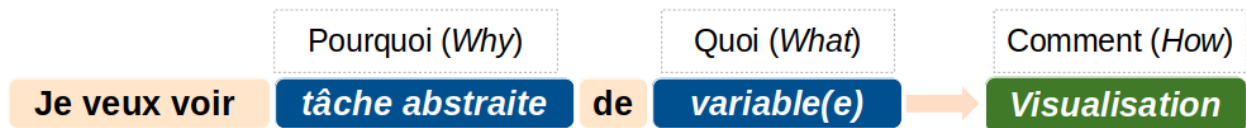


Figure Figure 3.7 Décomposition d'une requête, selon la méthode de Munzner, en : «pourquoi», «quoi» et «comment». Les éléments «pourquoi» et «quoi» sont choisis par l'utilisateur (blocs bleus), tandis que le «comment» sera généré automatiquement (bloc vert) à ce dernier.

Dans cette approche, l'utilisateur définit sa requête de manière explicite. Il choisit le type de tâche qu'il souhaite réaliser (pourquoi) et les données qu'il souhaite visualiser (quoi) (figure Figure 3.8). Cette méthode, citée précédemment dans l'état de l'art [22, 24], fait partie des bonnes pratiques recommandées dans le design d'un outil de visualisation pour un utilisateur novice (voir la recommandation *R7* du tableau Tableau 3.3). En effet, il est important de simplifier à un usager novice l'étape de définition de requêtes qui peut être floue dans le cas de l'exploration.

Cette pratique est réalisée, à l'aide de deux interactions :

- des cases à cocher permettant de choisir la tâche abstraite à accomplir. Ces tâches sont, au préalable, définies dans la section 3.2. Le menu comporte 3 choix : distribution, relation (*Relationship*) et évolution temporelle (*Time evolution*);

- une zone de saisie (avec autocomplétion) offrant la possibilité de choisir une ou plusieurs questions de sondage ou métadonnées.

Figure Figure 3.8 Zone de définition de requêtes : (A) Cases à cocher pour sélectionner une tâche parmi les 3 tâches proposées (« *Distribution* », « *Relationship* » et « *Time evolution* ») (B) Zone de saisie (avec autocomplétion) pour choisir une ou plusieurs questions de sondage ou métadonnées (C) Bouton de génération de visualisations selon de type de tâches et données choisi.

Afin de garder une cohérence sémantique, certaines règles ont été fixées lors de la définition d'une requête :

1. quand l'utilisateur sélectionne la tâche *Distribution* ($T1$) ou *Time evolution* ($T3$), il ne peut choisir qu'une seule question de sondage ou métadonnée à visualiser à la fois (figures Figure 3.9 et Figure 3.10) ;

Figure Figure 3.9 Exemple de bonne et mauvaise manipulation lors du choix de la tâche *Distribution* ($T1$).

2. quand l'utilisateur sélectionne la tâche *Relationship* ($T2$), il doit choisir au minimum deux variables (questions de sondage ou métadonnées) à visualiser (figure Figure 3.11).

Cette information est donnée à l'utilisateur sous forme de texte au-dessus de la zone de saisie (figure Figure 3.8). Par exemple, lors du choix des tâches *Distribution* ou *Time evolution*, le message suivant apparaît : « *Select one variable at the time.* ». Dans le cas contraire, un message d'avertissement informera l'utilisateur des corrections à apporter.

La zone de création des visualisation

I want to visualize the Distribution Relationship TimeEvolution of Select one variable at the time. ✓

I want to visualize the Distribution Relationship TimeEvolution of Select one variable at the time. ✗

Figure Figure 3.10 Exemple de bonne et mauvaise manipulation lors du choix de la tâche *Time evolution* ($T3$).

I want to visualize the Distribution Relationship TimeEvolution of Select at least two variables. ✓

I want to visualize the Distribution Relationship TimeEvolution of Select at least two variables. ✗

Figure Figure 3.11 Exemple de bonne et mauvaise manipulation lors du choix de la tâche *Relationship* ($T2$).

La zone de création des visualisations est le résultat des choix faits par l'utilisateur dans l'espace de définition de requêtes. En effet, quand l'utilisateur appuie sur le bouton « *Visualize* » (figure Figure 3.8), une visualisation est générée sous forme de fenêtre (cette opération sera détaillée dans la section 3.5.2).

Il y a donc une répartition de tâches entre l'utilisateur et l'outil de visualisation (tableau Tableau 3.6). Cette approche, également citée dans l'état de l'art [22, 24], fait partie des pratiques recommandées lors du design d'un outil de visualisation pour un utilisateur novice (voir les recommandations $R4$, $R5$, $R6$ et $R7$ du tableau Tableau 3.3). En effet, un usager n'ayant pas d'expérience pertinente en analyse de visualisations risque de ne pas choisir la bonne visualisation selon ses besoins («quoi» et «pourquoi»).

Étapes de Munzner	Qui choisit ?	Résultat
Quoi ?	L'utilisateur	Une donnée
Pourquoi ?	L'utilisateur	Une tâche
Comment ?	Notre outil	Une visualisation

Tableau Tableau 3.6 Répartition des responsabilités lors du processus de création d'une visualisation avec la méthode proposée.

La visualisation générée se présente sous forme de fenêtre dans le panneau principal (figure Figure 3.12). La fenêtre comporte les éléments suivants :

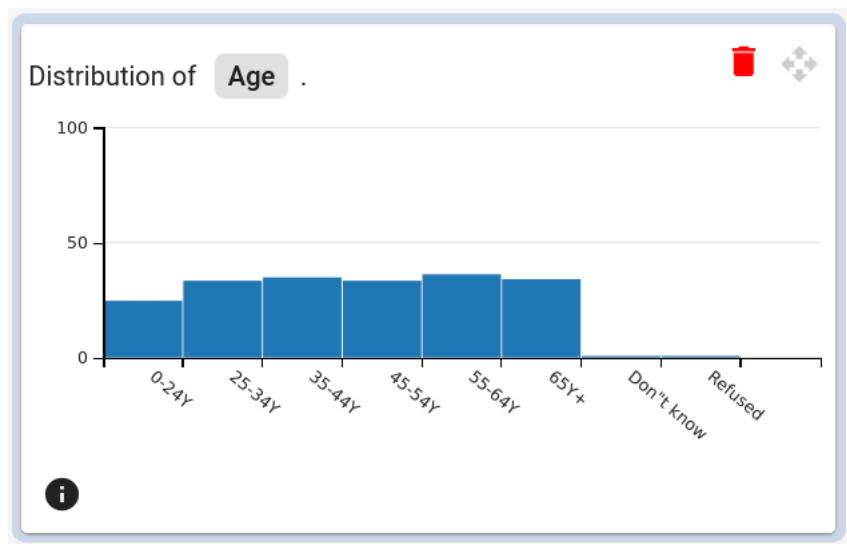





Figure Figure 3.12 Exemple de la fenêtre générée comportant la visualisation.

1. la visualisation générée ;
2. une courte description rappelant ce qui est visualisé (quoi) et pour quelle tâche (pourquoi) ;
3. un bouton de suppression  afin de fermer les fenêtres non utilisées ;
4. un bouton glisser-déposer  permettant de déplacer les fenêtres et les réarranger. Ainsi, l'utilisateur est capable de grouper et séparer des vues. De plus, il peut réarranger le panneau central comme son tableau de bord personnalisé ;
5. un bouton d'information  . Considérant le manque d'expérience de l'utilisateur, il a été considéré de lui apporter assistance dans le cas où ce dernier ne connaîtrait pas la visualisation générée. Cette assistance se présente sous forme de fenêtre contextuelle (*pop-up*) lui indiquant comment interpréter la figure à travers une brève description (figure Figure 3.13). Les descriptions complètes figurent dans l'annexe B.

3.5.2 Génération automatique de visualisation

Tenant compte de la sixième recommandation (*R6*) concernant les usagers novices, un système de génération automatique de visualisations a été implémenté. D'après nos hypothèses préliminaires deux bénéfiques peuvent ressortir de son utilisation :

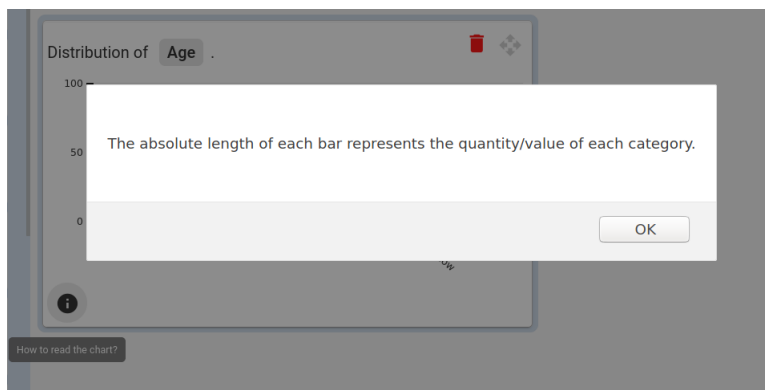


Figure 3.13 Exemple de fenêtre contextuelle (*pop-up*) expliquant comment lire un graphique à barres.

1. la génération automatique permettra l'utilisateur de se concentrer sur l'exploration et l'interprétation des données,
2. en utilisant notre outil, l'utilisateur novice enrichira sa littérature en visualisation de données.

Avant la génération d'une visualisation, notre système prend en compte deux paramètres choisis par l'utilisateur dans la zone de définition de requêtes : **La tâche à effectuer** ($T1$, $T2$ et $T3$) et **la question de sondage choisie**. Une fois les deux paramètres choisis, une visualisation sera générée.

Le tableau Tableau 3.7 offre une synthèse de toutes les visualisations choisies par type de tâches et variables. Puis, nous développerons et expliquerons plus en détail nos choix.

Nous avons principalement suivi la taxonomie disponible dans la littérature [41] concernant les techniques de représentation visuelle de résultats de sondages. En complément de cela, nous avons utilisé le guide proposé par Tamara Munzner dans son livre *Visualization Analysis and Design* [25].

Tâche 1 : Distribution

Par définition, une distribution signifie la répartition d'un élément selon un critère particulier [68]. Dans le cas d'un sondage, cela représente la répartition des réponses à une question posée.

La visualisation générée dépendra du **type de variable** (question de sondage) choisi :

1. Variable nominale :

On retrouve dans cette catégorie les questions fermées à choix unique.

Ex. : *Question* : «Quel est votre parfum de glace préféré?», *Réponses possibles* : Cho-

Tâche	Nombre de questions sélectionnées	Type de questions	Visualisation générée	Figure
<i>Distribution</i>	1	Nominal	Graphique à barres	Figure 3.14a
		Ordinal	Graphique à barres + boîte à moustache	Figure 3.14b
		Numérique	boîte à Violon	Figure 3.14c
<i>Relationship</i>	2	Numérique/Nominal	boîte à Violon	Figure 3.15c
		Numérique/Ordinal	Nuage de points	Figure 3.15b
		Nominal	Nuage de carrés	Figure 3.15a
	2 et plus	Numérique/Numérique	Graphique de coordonnées parallèles	Figure 3.16b
		Numérique/Nominal/Ordinal	Graphique de coordonnées parallèles	Figure 3.16b
		Nominal	Graphique des ensembles parallèles	Figure 3.16a
<i>Time evolution</i>	1	Nominal	Graphique à barres compilées	Figure 3.17a
		Ordinal	boîte à moustache	Figure 3.17b
		Numérique	boîte à moustache	Figure 3.17b

Tableau 3.7 Tableau récapitulatif des visualisations générées par type de tâches et questions de sondage.

colat, Fraise ou Vanille.

Pour ce cas, la littérature [25, 41] encourage d'utiliser un **graphique à barres**, car :

- (a) le canal visuel de la longueur reste le meilleur pour représenter une valeur quantitative (le nombre de répondants) ;
- (b) le canal visuel de la position spatiale sur un axe horizontal reste le meilleur pour représenter une valeur qualitative (le choix de réponse).

Cette visualisation (figure Figure 3.14a) permet de chercher une valeur particulière ou comparer des valeurs selon la hauteur.

2. Variable ordinale :

On retrouve dans cette catégorie les questions à échelle de Likert ou d'évaluation.

Ex. : *Question* : «Sur une échelle de 1 à 10, combien noteriez la glace au chocolat ?»,
Réponses possibles : 1 à 10.

Une combinaison de deux visualisations a été retenue. Vu que le nombre de réponses possibles reste dénombrable, un **graphique à barres** permettra de voir une distribution individuelle des réponses. En complément, une **boîte à moustache** permettra d'avoir les principaux paramètres statistiques tels que la moyenne, la variance, le minimum, etc.

Afin de représenter l'échelle de Likert, un encodage a été établi pour transformer une valeur catégorique ordinale en valeur numérique représentable sur une boîte à moustache (voir tableau Tableau 3.8).

Cette visualisation (figure Figure 3.14b) permet de :

- (a) chercher une valeur particulière ;
- (b) comparer des valeurs selon la hauteur ;
- (c) extraire des valeurs dérivées telles que la moyenne et la variance ;
- (d) détecter les valeurs aberrantes.

3. Variable numérique :

On retrouve dans cette catégorie les questions ouvertes à valeurs numériques.

Ex. : *Question* : «Quel âge avez-vous ?», *Réponses possibles* : Nombre entier positif.

La visualisation retenue est une variante de la boîte à moustache : **la boîte à violon**. Cela a été choisi, car :

- (a) cette variante a les avantages d'un graphique à densité et d'une boîte à moustache ;
- (b) elle offre une vue globale de la distribution et une vue détaillée, car les données ne sont pas agrégées.

Cette visualisation (figure Figure 3.14c) permet de :

- (a) voir la distribution de la variable sur un intervalle continue ;
- (b) extraire des valeurs dérivées telles que la moyenne et la variance ;
- (c) détecter les valeurs aberrantes.

Tâche 2 : Relations (*Relationship*)

Une relation peut être définie comme une connexion, une corrélation ou un lien entre deux variables. Cette dernière peut être bivariée ou multivariée.

La visualisation générée dépendra du **type de variable** (question posée) et des **combinaisons réalisées** (bivariées ou multivariées). Nous retrouvons les cas de figure suivants :

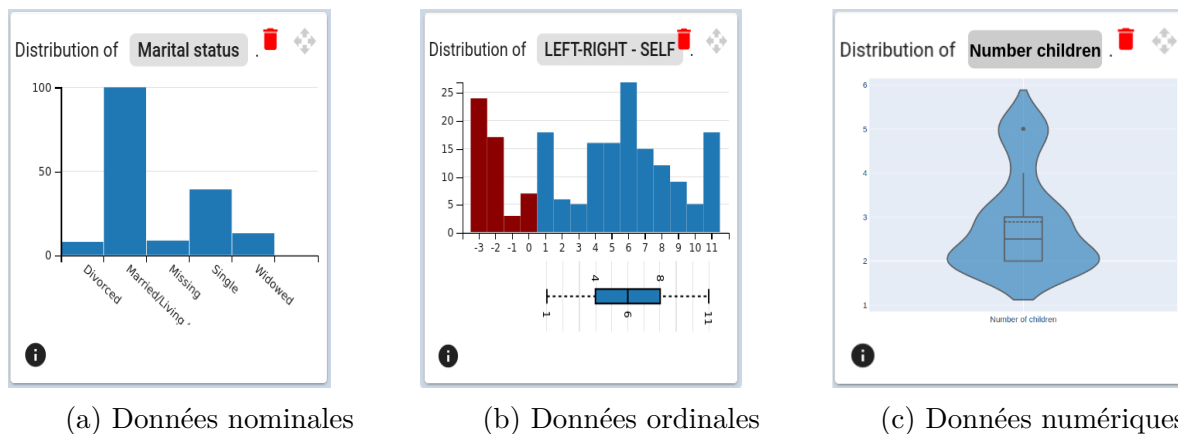


Figure Figure 3.14 Visualisations générées pour la tâche « Distribution »

1. Cas 1 : Relation bivariée

(a) Relation entre deux variables nominales :

Pour ce genre de situation, le **nuage de carrés** a été retenu (figure Figure 3.15a). Cette visualisation permet de :

- comparer et estimer la relation entre les réponses de sondage ;
- détecter les valeurs aberrantes.

(b) Relation entre deux variables numériques ou ordinales :

Le **nuage de points** a été retenu (figure Figure 3.15b), car il permet de :

- comparer et estimer la relation entre les réponses de sondage.
- détecter les valeurs aberrantes.

(c) Relation entre une variable numérique ou ordinale **et** nominale : La **boîte à violon** a été retenue (figure Figure 3.15c), car il permet de :

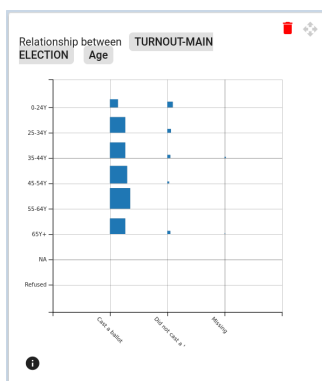
- estimer la distribution de la variable numérique ou ordinale selon la variable catégorique.
- comparer les distributions par rapport aux différentes valeurs catégoriques,
- détecter les valeurs aberrantes.

2. Cas 2 : Relation multivariée

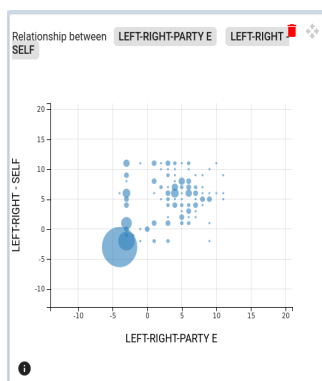
(a) Relation entre des variables nominales :

Le **graphique des ensembles parallèles** «*Parallel sets*» a été retenu, car il permet de :

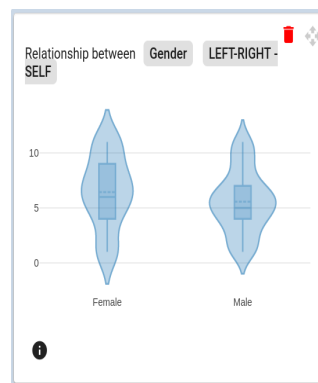
- visualiser la distribution de chaque dimension ;



(a) Données catégoriques



(b) Données numériques



(c) Données numériques et catégoriques

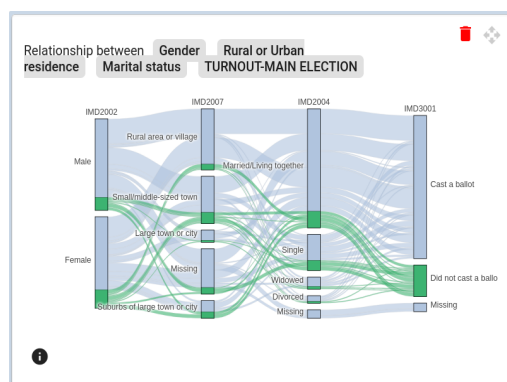
Figure Figure 3.15 Visualisations générées pour la tâche « Relation » (*Relationship*) dans le cas des relations bivariées.

— estimer l'évolution d'une distribution selon plusieurs facteurs.

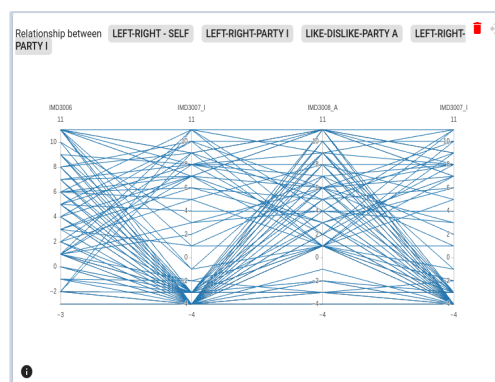
(b) Relation entre des variables numériques ou ordinales ou nominales :

Le **graphique de coordonnées parallèles** a été retenu, car il permet de :

- estimer la corrélation entre deux ou plusieurs facteurs ;
- estimer l'évolution d'une distribution selon plusieurs facteurs ;
- détecter les corrélations positives et négatives.



(a) Données catégoriques



(b) Données numériques/nominales/ordinales

Figure Figure 3.16 Visualisations générées pour la tâche « Relation » (*Relationship*) dans le cas des relations multivariées.

Tâche 3 : Évolution temporelle (*Time evolution*)

Une évolution temporelle représente une relation spécifique entre une variable et le temps. Elle permet de détecter une tendance (*pattern*) ou un évènement particulier dans le temps.

La visualisation générée dépendra du **type variable** (question posée) choisi :

1. Variable nominale :

Pour ce genre de situation, le **graphique à barres empilées** a été retenu, car :

- (a) le canal visuel de la longueur sur un axe non-aligné est bon pour représenter une valeur quantitative (le nombre de répondants) ;
- (b) les canaux visuels de la couleur et la position spatiale sur un axe horizontal restent les meilleurs pour représenter les valeurs qualitatives (le choix de réponse + l'année).

Cette visualisation (figure Figure 3.17a) permet de :

- (a) chercher une valeur particulière ;
- (b) comparer les changements d'une réponse par année ;
- (c) comparer les proportions de chaque catégorie de réponse.

2. Variable ordinale et numérique :

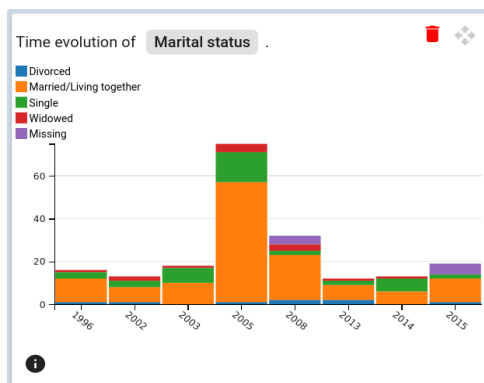
Pour ce genre de situation, les **boîtes à moustaches** ont été retenues. Cette fois-ci, elles sont alignées côte à côte pour permettre une comparaison directe.

Cette visualisation (figure Figure 3.17b) permet de :

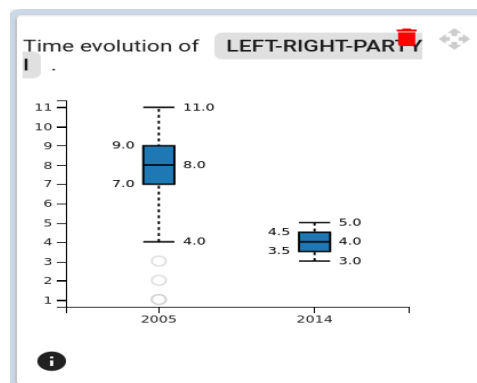
- (a) comparer la distribution d'une variable sur plusieurs valeurs temporelles ;
- (b) extraire des valeurs dérivées telles que la moyenne et la variance ;
- (c) détecter les valeurs aberrantes.

3.5.3 Interactivité de l'outil

Yi, Ji Soo, et coll. définissent une interaction comme étant une fonctionnalité permettant de manipuler (directement ou indirectement) et interpréter une visualisation [69]. Dans un outil de visualisation à vues multiples, l'utilisation d'interactions est plus qu'indispensable pour effectuer certaines tâches telles que le filtrage, le réarrangement des vues, etc. De plus, elles représentent un moyen puissant pour améliorer l'expérience de l'utilisateur et son engagement [25]. Nous décrivons dans cette section les principales interactions adoptées dans notre outil.



(a) Données nominales



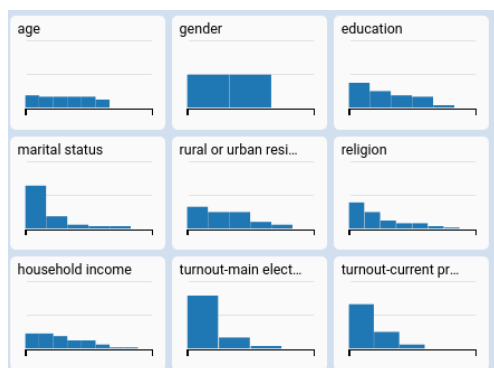
(b) Données ordinales ou numériques

Figure Figure 3.17 Visualisations générées pour la tâche « Évolution temporelle » (*Time evolution*)

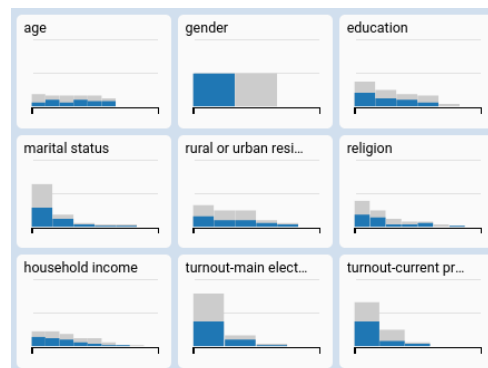
La sélection

Cette interaction, également implémentée sur l'outil de visualisation *Domino* [15] (figure Figure 2.11), permet de sélectionner une réponse particulière à une question de sondage. Cette sélection génère un filtre qui se propage sur l'ensemble des autres questions du sondage. Cette interaction est utilisée dans les visualisations permettant une sélection comme le graphique à barres, les nuages de points, etc.

La figure Figure 3.18 illustre un exemple où l'utilisateur choisit la question sur le genre et décide de ne voir que les réponses des femmes. Une telle interaction est réversible en re cliquant sur le même élément.



(a) Avant sélection



(b) Après sélection

Figure Figure 3.18 Illustration de l'interaction de sélection. L'utilisateur sélectionne la question sur le genre et décide de ne voir que les réponses des femmes.

Les info-bulles

Les info-bulles sont présentes dans toutes les visualisations. Elles apparaissent en survolant un élément. Elles sont particulièrement pertinentes dans les petites visualisations telles que celles du panneau latéral ou pied de page. Elles offrent des informations complémentaires, ex. : la valeur d'une réponse choisie, les paramètres statistiques (boîte à moustache), etc.

Brossage (*Brushing*)

Le brossage étant différent de la sélection, il permet de choisir un intervalle particulier. Cette interaction est utilisée dans le cas du graphique de coordonnées parallèles (figure Figure 3.19). Le brossage réduit le bruit visuel en ne gardant que les valeurs pertinentes à notre analyse.

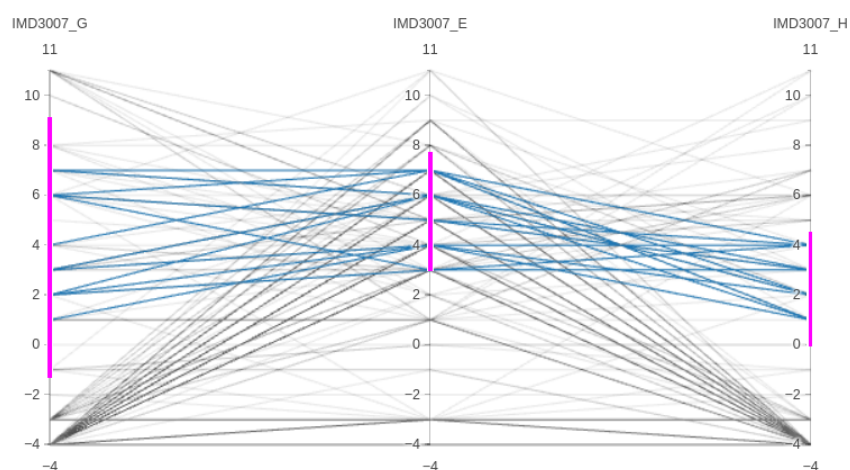



Figure Figure 3.19 Illustration de l'interaction de brossage sur un graphique de coordonnées parallèles.

Glisser-déposer (*Drag-n-drop*)

En utilisant des vues multiples, ces dernières peuvent être liées, en termes d'interaction, ou complètement indépendantes. La coordination entre les vues est essentielle pendant une analyse exploratoire, car elle permet de déceler des relations cachées entre les vues.

Dans notre approche, nous proposons l'utilisation de l'interaction glisser-déposer dans le but de lier ou séparer des vues. Cette interaction nécessite peu d'efforts de la part de l'utilisateur novice. Les étapes sont les suivantes (voir figure Figure 3.20) :

1. Une fois une visualisation générée, cette dernière est, par défaut, indépendante, c'est-à-dire qu'elle n'interagit avec aucune autre visualisation générée (figure Figure 3.21a).
2. L'utilisateur sélectionne et fait glisser une des visualisations en utilisant le bouton . Ensuite, il dépose la visualisation sélectionnée sur celle qu'il souhaite connecter (figure Figure 3.21b). Un système implicite de groupe est donc utilisé. Un groupe peut contenir deux ou plus visualisations ensembles. Dans un même groupe les vues et interactions sont liées.
3. Cette interaction est réversible et les groupes peuvent être séparés. Pour cela, l'utilisateur sélectionne la visualisation à séparer et la fait glisser en dehors du groupe.

À noter que, par manque de temps, cette interaction n'a été implémentée que pour les visualisations de la tâche *Distribution*.

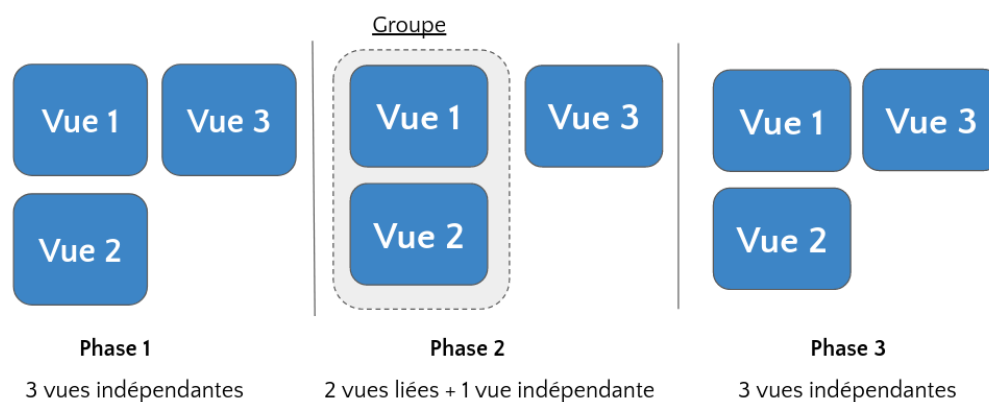
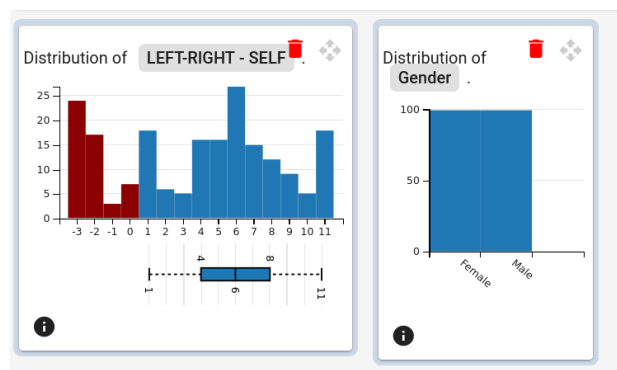


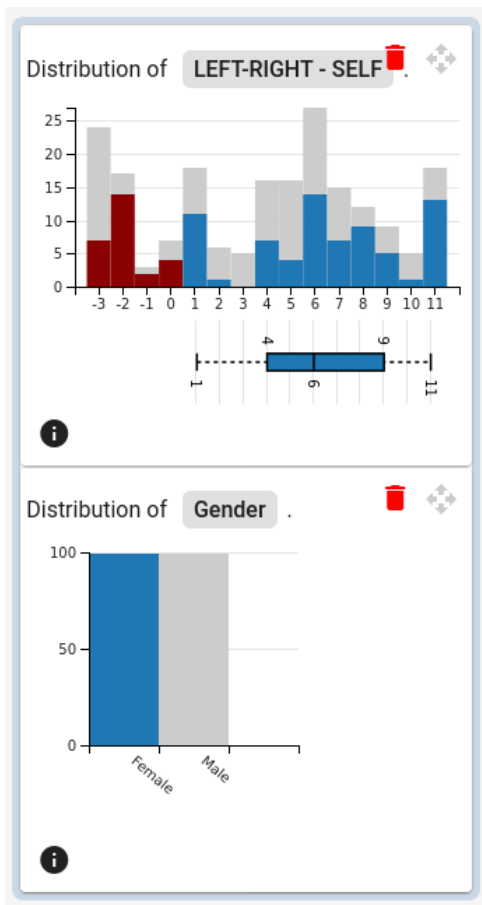
Figure Figure 3.20 Les 3 phases de l'interaction glisser-déposer. Phase 1 : Lors de leur génération, les visualisations sont, par défaut, indépendantes. Phase 2 : L'utilisateur peut choisir de lier deux vues grâce à l'interaction glisser-déposer, et ainsi créer un groupe de vues liées. Phase 3 : Les groupes peuvent être défaits au besoin.

3.6 Implémentation

L'outil a été développé sous forme de plate-forme Web afin d'être plus accessible pour les utilisateurs. Nous avons donc utilisé des technologies Web (HTML, CSS et TypeScript) avec le cadre d'application Angular 7. Les visualisations et interactions ont été implémentées grâce aux bibliothèques de visualisation : D3.js [70], DC.js [71], crossfilter.js [72] et Plotly [73]. Notre outil traite actuellement un sondage contenant 60 variables et 200 réponses collectées. Les données de sondage ont été stockées sur la base de données orientée document MongoDB. En termes d'évolutivité, nous avons remarqué une latence avec l'augmentation du nombre de questions et répondants. Cette latence est provoquée lors de la génération des visualisations



(a) Avant interaction



(b) Après interaction

Figure 3.21 Illustration de l'interaction glisser-déposer. Sur la gauche, les deux visualisations sont indépendantes. Sur la droite, les deux visualisations sont liées et peuvent interagir entre elles.

ou l'interaction de filtrage. Nous proposons donc comme amélioration pour les versions futures de garder en cache les visualisations et interactions déjà créées afin de réduire le temps de calcul.

L'outil est accessible sur le lien suivant :

https://manelboucenna.github.io/Survey_Visualization_project/.

Valeurs de l'échelle de Likert	Encodage
Tout à fait d'accord	4
D'accord	3
Ni en accord, Ni en désaccord	2
Pas d'accord	1
Tout à fait en désaccord	0

Tableau Tableau 3.8 Exemple d'encodage pour échelle de Likert

CHAPITRE 4 PROTOCOLE D'ÉVALUATION DE LA VISUALISATION PROPOSÉE

Ce chapitre décrit le protocole détaillé d'évaluation de l'outil développé dans le cadre de ce projet. Il introduit la méthode d'évaluation choisie et les données utilisées pour la démonstration. Ensuite, il présente l'échantillon de participants pour notre évaluation. Puis, il décrit le protocole suivi qui a été approuvé par le comité d'éthique de recherche (CÉR) de Polytechnique Montréal. Enfin, il rappelle les hypothèses de recherche à valider et leurs métriques respectives.

4.1 Méthode d'évaluation

Quand vient l'étape d'évaluation d'un outil de visualisation, deux approches principales sont possibles : en laboratoire ou en ligne. En raison des circonstances sanitaires, dues à la COVID-19, il a été préférable de se tourner vers des entrevues en vidéoconférence de 50 minutes. Cette option est une approche hybride des évaluations en ligne et en laboratoire, car : (1) dans le cadre d'une étude préliminaire, telle que la nôtre, elle nous permet d'échanger avec les participants et de collecter des retours pertinents, mais (2) elle nous limite en matière de contrôle sur certains paramètres de l'expérience tels que la taille de l'écran, le type de navigateur, etc.

Une évaluation sous forme de vidéoconférence nécessite la présence des responsables du projet et du participant simultanément. Cela nécessite une plage horaire fixe, ce qui peut être contraignant pour les participants et ainsi réduire leur nombre. Un total de quatre à six participants a été estimé pour cette étude. Ce nombre a été choisi en comparant notre étude à d'autres travaux similaires dont la moyenne était de cinq participants tels que : *Cerebral* [74], *MizBee* [75], *Keshif* [13], etc.

4.2 Données

Dans le cadre de l'évaluation, nous avons visualisé une enquête *longitudinale à cohorte* avec métadonnées. Plus particulièrement, nous avons utilisé le sondage de l'étude comparative des systèmes électoraux (ECSE) [16]. Cette base de données suscite un grand engouement auprès des chercheurs en sciences politiques et sociales. Elle est composée d'un ensemble de questions administrées depuis 20 ans à travers 55 pays avec près de 281000 répondants.

Au total, nous avons retenu **60 variables**, dont 50 questions de sondages et 10 métadonnées,

composées de questions de type :

1. questions fermées à choix unique ;
2. échelle d'évaluation ;
3. échelle de Likert ;
4. questions ouvertes à valeurs numériques.

Pour chaque répondant sont fournies ses données de sondage (réponses aux questions posées) ainsi que les métadonnées qui lui sont associées. Le tableau Tableau 4.1 est un exemple de données disponibles pour une personne.

Dans le but de réduire le biais de confirmation¹, seule une portion de **200 réponses** choisies aléatoirement a été utilisée.

Type de données	Exemple
Données de sondage	Quel âge avez-vous ?
	Que pensez-vous du président actuel ?
	Comptez-vous voter ?
Métadonnées	Pays
	Type d'élection
	Année

Tableau Tableau 4.1 Exemple de données recueillies pour chaque participant

4.3 Échantillon

Dans le cadre de cette étude, nous avons choisi un échantillonnage non probabiliste de convenance² composé d'étudiants en sciences politiques. Ces derniers représentent un échantillon intéressant, car ils analysent régulièrement dans leurs travaux des données de sondage et ne sont pas particulièrement experts en visualisations de données. Ce qui correspond au profil d'utilisateur novice désiré.

Les participants de l'étude sont recrutés par mail (annexe C) et doivent signer le formulaire de consentement (annexe D). Au total, nous nous sommes entretenus avec 4 participants : 2 femmes, 2 hommes, 2 âgés entre 18-24 ans, 1 âgé entre 25-34 ans et 1 âgé entre 35-44 ans. Nous nous sommes entretenus avec deux étudiants en maîtrise et deux étudiants au doctorat.

1. Tendence à chercher des éléments permettant de confirmer des connaissances antérieures [76].

2. Échantillon choisi pour des raisons de coûts et d'accessibilité, ex. : étudiants, clients, etc.

4.4 Procédure d'évaluation de l'outil de visualisation

L'expérience s'est déroulée en 5 étapes. Tout au long de ces étapes, la responsable du projet restait présente pour répondre aux questionnements des participants.

4.4.1 Étape 1 : Questions générales

Il a été demandé au participant de fournir quelques informations générales : genre, groupe d'âge et niveau d'éducation. Ces informations nous permettent de mieux évaluer la diversité de notre échantillon de recherche et interpréter nos résultats. Le questionnaire complet est disponible dans l'annexe E.

4.4.2 Étape 2 : Présentation de l'outil

La responsable de l'expérience fait une démonstration des principales fonctionnalités de l'outil. Ensuite, le participant teste l'outil pour la première fois, supervisé par la responsable de l'expérience qui répond à tous ses questionnements. Cette étape permet à l'utilisateur de se familiariser avec l'outil et confirmer sa compréhension de la démonstration faite précédemment.

4.4.3 Étape 3 : Exploration libre

Le participant utilise librement l'outil pendant 15 minutes. À partir de ce moment, la voix et l'écran du participant sont enregistrés dans le but d'étudier les remarques et comportements généraux adoptés par ce dernier.

4.4.4 Étape 4 : Évaluation de l'outil

Tous les participants répondent aux mêmes questions pour évaluer leur expérience utilisateur. Les questions ont été rédigées afin d'évaluer l'atteinte de nos objectifs. Le questionnaire complet est disponible dans les annexes F, G et H. À cette étape, la responsable de l'expérience se retire afin de ne pas biaiser l'utilisateur dans ses choix de réponses.

4.4.5 Étape 5 : Entrevues et questions ouvertes

Une fois que le participant répond au questionnaire, on lui pose 5 questions ouvertes (disponibles à l'annexe I). Cette dernière partie du questionnaire donne l'occasion au participant de communiquer verbalement ses recommandations et autres opinions inédites sur l'outil. Lors de cette étape, un enregistrement audio est fait afin de permettre sa retranscription.

4.5 Hypothèses de recherche et mesures d'évaluation associées

L'évaluation a pour but de confirmer ou infirmer les hypothèses en rapport avec nos choix de design :

(*H1*) : Une vue globale dans la visualisation montrant toutes les questions de sondages et les métadonnées permet à l'utilisateur d'émettre de nouvelles hypothèses.

(*H2*) : Un choix explicite de tâches et questions de sondage permet à un utilisateur novice de traduire son hypothèse en visualisation.

(*H3*) : La prédéfinition de visualisations par tâches et types de variable permet à l'utilisateur d'explorer efficacement son jeu de données.

(*H4*) : La présentation de potentielles nouvelles visualisations à l'utilisateur novice lui permet d'améliorer ses connaissances en visualisations de données.

(*H5*) : Des interactions de type glisser-déposer et filtrage améliorent l'expérience d'un utilisateur novice.

À cette fin, les participants de l'expérience ont répondu à des questionnaires d'évaluation (disponibles aux annexes F, G, H et I) en rapport avec les hypothèses de recherche. Le tableau Tableau 4.2 présente la correspondance entre les hypothèses et les différents questionnaires de l'expérience. Les résultats de cette évaluation seront présentés dans le chapitre suivant.

Questionnaire	Hypothèses de recherche				
	<i>H1</i>	<i>H2</i>	<i>H3</i>	<i>H4</i>	<i>H5</i>
Questionnaire sur l'exploration libre (annexe F)	✓	✓	✓		
Questionnaire sur la génération de visualisation (annexe G)		✓	✓	✓	
Questionnaire sur l'expérience utilisateur (annexe H)					✓
Questions ouvertes et discussions (annexe I)	✓	✓	✓	✓	✓

Tableau Tableau 4.2 Correspondance entre les hypothèses de recherche et les questionnaires de l'expérience

CHAPITRE 5 RÉSULTATS ET DISCUSSIONS

Les conclusions de l'évaluation décrite précédemment seront présentées dans ce chapitre. D'abord, les résultats aux différents questionnaires et les échanges avec les participants seront analysés et comparés aux hypothèses de recherche. Puis, une synthèse des recommandations proposées sera effectuée. Enfin, nous décrirons les différentes limitations de l'outil et du protocole d'évaluation.

5.1 Présentation et discussion des résultats

Dans cette section, nous présenterons les résultats des 3 questionnaires de l'évaluation (annexes F, G et H). Les résultats seront appuyés par les retours récoltés durant nos discussions avec les participants (annexe I).

5.1.1 Questionnaire sur l'exploration libre

Le questionnaire sur l'exploration libre a pour but de confirmer ou infirmer les hypothèses *H1*, *H2* et *H3*. Les questions sont résumées dans le tableau Tableau 5.1 pour faciliter la lecture. Les résultats du questionnaire sont illustrés dans la figure Figure 5.1.

Questions du questionnaire sur l'exploration libre
<i>Q1</i> : À l'aide de ce produit, j'ai fait plusieurs observations pertinentes.
<i>Q2</i> : À l'aide de ce produit, j'ai découverts des tendances et relations pertinentes entres les données.
<i>Q3</i> : À l'aide de ce produit, j'ai réalisé toutes mes tâches d'explorations.
<i>Q4</i> : Le panneau du haut m'a permis de traduire facilement une question que j'avais, en tête, en visualisation.
<i>Q5</i> : Grâce au panneau de gauche, j'ai pu émettre de nouvelles hypothèses.
<i>Q6</i> : Grâce au panneau central, j'ai pu infirmer ou confirmer une hypothèse.
<i>Q7</i> : Ne pas avoir à construire mes visualisations m'a permis de me concentrer sur mes tâches d'explorations.

Tableau Tableau 5.1 Résumé des questions pour évaluer l'exploration

Les résultats des questions *Q1* et *Q2* montrent que notre outil a permis à nos utilisateurs de faire plusieurs observations pertinentes sous forme de tendances et relations. Durant leur utilisation de l'outil, les participants avaient tendance à alterner régulièrement entre le panneau de gauche et le panneau central pour avoir une vue globale et détaillée de leurs données.

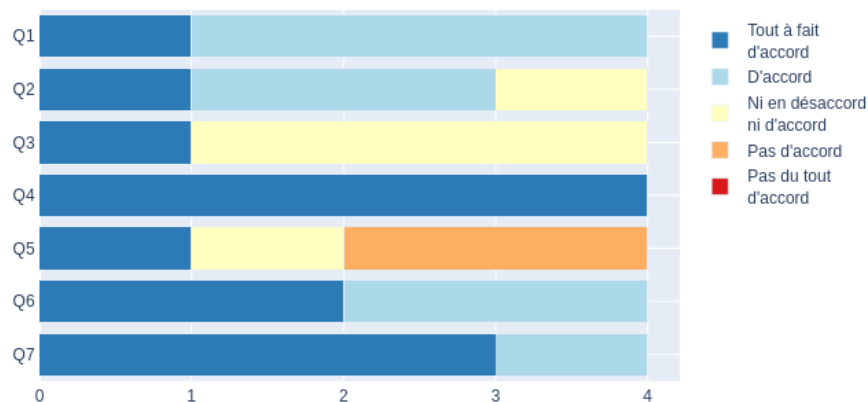


Figure Figure 5.1 Résultats du questionnaire sur l'exploration libre

Nous avons remarqué aussi que seul un participant sur quatre avait utilisé le pied de page pour analyser les métadonnées. L'une de nos suppositions serait la visibilité du pied de page, contenant les métadonnées, qui est par défaut masqué. Nous supposons également que cela pourrait être dû au manque de temps durant l'évaluation. En effet, en vue des 15 minutes de l'expérience et du nombre de fonctionnalités disponibles, il serait pertinent, pour les futures évaluations, d'accorder plus de temps aux participants afin de tester les différentes fonctionnalités de l'outil.

Pour la question Q3, trois participants sur quatre ont donné la mention «Ni en désaccord ni en accord» tandis qu'un participant a donné la mention «Tout à fait d'accord». Nous pensons que, dus au manque d'expérience en visualisations de données, les participants ne connaissent pas la définition exacte des tâches d'exploration comme définies dans la littérature (voir section 2.2.1). En effet, une des participantes a demandé s'il était possible de répondre «Je ne sais pas» à cette question. Contrairement à la question Q2 qui était plus précise, la question Q3 se voulait être plus généraliste en englobant d'autres aspects de l'exploration (filtrage, détection de valeurs aberrantes, etc.). Nous conseillons donc, pour les futures évaluations, la formulation d'une question plus explicite ou sa décomposition en plusieurs sous-questions.

La question Q4 concerne principalement la zone de définition de requêtes. Les résultats sont encourageants, car tous les participants ont accordé la mention «Tout à fait d'accord». L'un des participants a déclaré : «*J'aime que la requête soit écrite comme une phrase, c'est bien que l'utilisateur puisse mettre en phrase ce qu'il veut faire.*». Une autre utilisatrice a déclaré : «*Les requêtes sont faciles à faire, car il suffit de remplir les blancs pour voir sa visualisation.*». Un des participants a quand même précisé que la tâche *time evolution* n'était pas très pertinente pour lui, car il n'avait jamais étudié d'enquêtes longitudinales. Cette observation nous permet de revoir le choix de nos tâches en affichant uniquement les tâches

nécessaires selon le type d'enquêtes. Un autre participant a soulevé qu'il aurait été pertinent d'ajouter la tâche «Régression» en plus des 3 tâches citées précédemment : «*Les régressions de type logistiques ou linéaires sont souvent effectuées par les étudiants dans leurs analyses.*».

Pour la question Q5, les participants ont accordé la mention «Pas d'accord» (50%), «Ni en désaccord ni en accord» (25%) et «Tout à fait d'accord» (25%). Une des participantes a déclaré : «*J'ai beaucoup aimé la possibilité d'effectuer des filtres multiples sur toutes mes variables dans le panneau de gauche.*». Cependant, les participants se sont plaints de l'absence de description ou explication pour certaines questions de sondage. Cette remarque s'applique particulièrement à la visualisation «boîte à moustache» dont l'encodage a été omis. Enfin, deux de nos participants avaient eu le réflexe de faire glisser une des visualisations du panneau de gauche vers le panneau central pour l'agrandir. Les avis étant divergents, nous ne pouvons pas nous prononcer sur ce point.

Pour la question Q6, les résultats sont encourageants. Les participants ont accordé la mention «Tout à fait d'accord» (50%) et «D'accord» (50%). L'une des candidates a mentionné : «*J'ai aimé que mes visualisations générées soient à portée de main, je pouvais facilement passer d'une vue à l'autre.*». Un autre candidat a souligné : «*J'aurais aimé pouvoir exporter mes graphiques pour les ajouter dans un rapport [...] Pouvoir aussi changer la couleur de mes visualisations ou ajouter un titre et les sources.*». Ce point est intéressant, car il a été partagé par divers candidats, ce qui nous laisse croire que cela pourrait être une fonctionnalité pertinente pour les futures versions. D'autres interactions ont été proposées par les participants, par exemple :

1. ajouter un bouton «*Reset visualisation*» pour enlever les filtres appliqués à une visualisation ;
2. ajouter un bouton «*Clear dashboard*» pour supprimer toutes les visualisations générées ;
3. ajouter la possibilité d'agrandir la taille des visualisations.

Toutes ces interactions pourraient améliorer l'expérience utilisateur et faire gagner du temps aux utilisateurs.

Enfin, les résultats de la question Q7 ont montré que le fait de générer automatiquement une visualisation enlève un poids à l'utilisateur le laissant ainsi se concentrer sur son analyse et exploration. Cette conclusion confirme les recommandations retenues de l'état de l'art [22,24].

En raison du nombre réduit de participants, nous ne pouvons pas confirmer ou infirmer *H1* qui stipule qu'une vue globale montrant toutes les questions de sondages et les métadonnées permet à l'utilisateur d'émettre de nouvelles hypothèses.

Les résultats préliminaires des questions Q3 et Q5 divergent et ne sont pas probants. Cepen-

dant, les autres résultats portent à croire que la répartition choisie, « *overview first, details on-demand* » a permis l'utilisateur de faire des observations intéressantes ou d'évaluer ses hypothèses. Aussi, grâce à notre outil, les participants semblent avoir pu traduire facilement leurs hypothèses en visualisations. Enfin, les résultats portent à croire que grâce à notre outil, les utilisateurs novices ont pu se concentrer sur leur exploration. Toutefois, ces résultats préliminaires doivent être confirmés auprès d'un plus large panel.

5.1.2 Questionnaire sur la génération de visualisations

Le questionnaire sur la génération de visualisations *Q2* avait pour but de valider les hypothèses *H2*, *H3* et *H4*. Les questions sont résumées dans le tableau Tableau 5.2 pour faciliter la lecture. Les résultats du questionnaire sont illustrés dans la figure Figure 5.2.

Questions du questionnaire sur la génération de visualisations	
<i>Q1</i> :	Les visualisations générées étaient pertinentes.
<i>Q2</i> :	À l'aide de ce produit, j'ai compris l'utilisation de certaines visualisations.
<i>Q3</i> :	Ne pas pouvoir choisir mes visualisations m'a frustré(e).
<i>Q4</i> :	La quantité de visualisations différentes m'a perturbé(e).

Tableau Tableau 5.2 Résumé des questions pour évaluer les visualisations générées

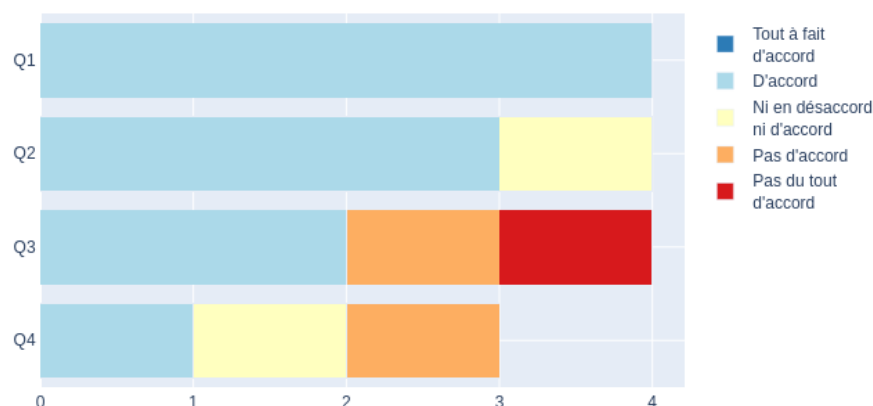


Figure Figure 5.2 Résultats du questionnaire sur la génération de visualisations. La question *Q4* a été ajoutée suite à notre première entrevue, d'où le nombre de 3 réponses seulement.

Les réponses de la question *Q1* indiquent que les participants sont tous « D'accord » avec le fait que les visualisations générées étaient pertinentes. L'une des participantes a attesté : « *Les modes de représentations sont pertinents, toutes les régressions ou les types d'analyses* »

que j'ai voulu faire, j'ai pu le faire [...] pour une première analyse ça à pas mal répondu à ce que je voulais faire ». Toutefois, quelques propositions revenaient souvent sur l'utilisation des visualisations, par exemple :

1. donner le choix à l'utilisateur de garder ou pas les valeurs manquantes dans un questionnaire («pas de réponse», «ne veux pas répondre», «ne sais pas», etc.) ;
2. pouvoir réordonner les catégories de réponses ;
3. pouvoir renommer les questions ;
4. pouvoir renommer les catégories de réponses ;
5. pouvoir choisir les catégories de réponses à afficher avant la création d'une visualisation ;
6. pouvoir grouper les catégories de réponses. Par exemple, l'utilisateur voudrait grouper les catégories «catholique» et «protestant» en une seule catégorie «chrétien».

Ces suggestions prouvent un besoin d'effectuer des actions plus poussées. En discutant avec une des participantes, elle nous déclara : *«Toutes ces interactions doivent se faire tout au long de l'exploration et non au prétraitement des données seulement. Car ces idées me viennent à l'esprit au fur et à mesure de mon analyse».*

Les réponses de la question Q2 indiquent que trois participants sur quatre sont «D'accord» avec le fait que notre outil les avait aidés à comprendre l'utilisation de certaines visualisations. Il est à noter que les utilisateurs avaient tendance à «jouer» avec une nouvelle visualisation pour comprendre son utilisation. Certaines représentations ne semblaient pas être parfaitement bien accueillies par les participants, nous citons : le graphique de Sankey, le graphique des coordonnées parallèles et la boîte à violon. D'ailleurs, une des participantes a déclaré en utilisant la boîte à violon et le graphique des coordonnées parallèles respectivement : *«J'aurais préféré quelque chose (visualisation) de plus conventionnel [...] Lorsqu'on choisit plus de deux variables, la visualisation n'est pas intuitive».* Ces visualisations sont connues auprès de nos participants, mais rarement utilisées.

De plus, seul un participant sur quatre a eu régulièrement recours à l'utilisation du bouton «How to read the chart». Cette observation pourrait nous laisser penser que la position du bouton «How to read the chart» devrait être mise plus en évidence. Une autre alternative serait se faire apparaître un mini-tutoriel, que l'utilisateur pourrait choisir d'ignorer, à chaque nouveau type de visualisation créée. D'ailleurs, l'un des participants avait souhaité trouver un glossaire plus complet et interactif des visualisations.

Les avis sur la question Q3 divergent. Les participants ont accordé la mention «D'accord» (50%), «Pas d'accord» (25%) et «Pas du tout d'accord» (25%). Certains participants affirment préférer avoir le choix entre la visualisation suggérée et d'autres alternatives. Cela

est dû à deux facteurs :

1. le type de visualisation qui, parfois, n'était pas intuitif pour un utilisateur novice ;
2. le besoin d'apprentissage par essai-erreur.

Un des participants a d'ailleurs soulevé : « *J'aurais aimé pouvoir choisir mes visualisations. Il y'a un certain côté intuitif à voir qu'une visualisation ne fonctionne pas, mais c'est nous qui l'avons créé ainsi* ». En contrepartie, un des participants a déclaré : « *Je n'avais pas vraiment d'attente en termes de visualisations, car je faisais l'exploration de l'outil en général* ». En d'autres termes, l'utilisateur était en mode découverte. Ce qui peut nous pousser à repenser les futures évaluations en les décomposant en sous-catégories pour évaluer uniquement ce phénomène.

Pour les réponses de question Q4, les participants ont accordé, de manière égale, les mentions : «D'accord», «Ni d'accord ni en désaccord» et «Pas d'accord». Il est à noter que cette question a été ajoutée suite à notre première entrevue, d'où le nombre de 3 réponses seulement. Les avis étant divergents, nous ne pouvons pas nous prononcer sur la question.

En raison du nombre réduit de participants, nous ne pouvons pas confirmer ou infirmer *H2* qui stipule qu'un choix explicite de tâches et questions de sondage permet à un utilisateur novice de traduire son hypothèse en visualisation.

Cependant, les résultats préliminaires semblent indiquer que notre système de génération de requêtes actuel permet de traduire facilement une hypothèse en visualisation.

En raison du nombre réduit de participants, nous ne pouvons pas confirmer ou infirmer *H3* qui stipule que la prédéfinition de visualisations par tâches et types de variable permet à l'utilisateur d'explorer efficacement son jeu de données.

Cependant, les résultats préliminaires portent à croire que le fait de générer automatiquement les visualisations pour l'utilisateur lui permet de se concentrer sur ses tâches d'analyse. Cependant, il a été observé qu'il pourrait être intéressant de laisser l'option à l'utilisateur de choisir et personnaliser ses visualisations (changer les couleurs, renommer les réponses, etc.).

En raison du nombre réduit de participants, nous ne pouvons pas confirmer ou infirmer *H4* qui stipule que la présentation de potentielles nouvelles visualisations à l'utilisateur novice lui permet d'améliorer ses connaissances en visualisations de données.

De plus, les résultats ne sont pas probants, car nous n'avons pas prouvé de lien direct entre la vue de nouvelles visualisations grâce à l'utilisation de notre outil et l'amélioration des connaissances du participant en termes de visualisations.

5.1.3 Questionnaire sur l'expérience utilisateur

Le questionnaire évaluant l'expérience de l'utilisateur a pour but de valider l'hypothèse $H5$ qui stipule que *des interactions de type glisser-déposer et filtrage améliorent l'expérience d'un utilisateur novice*. Les questions sont disponibles dans l'annexe H. Les résultats du questionnaire sont illustrés dans les figures Figure 5.3 et Figure 5.4.

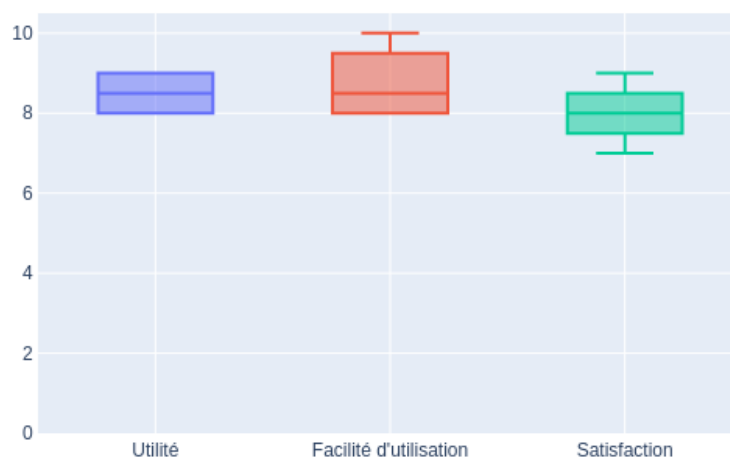


Figure Figure 5.3 Résultats du questionnaire sur l'expérience utilisateur : utilité, facilité d'utilisation et satisfaction

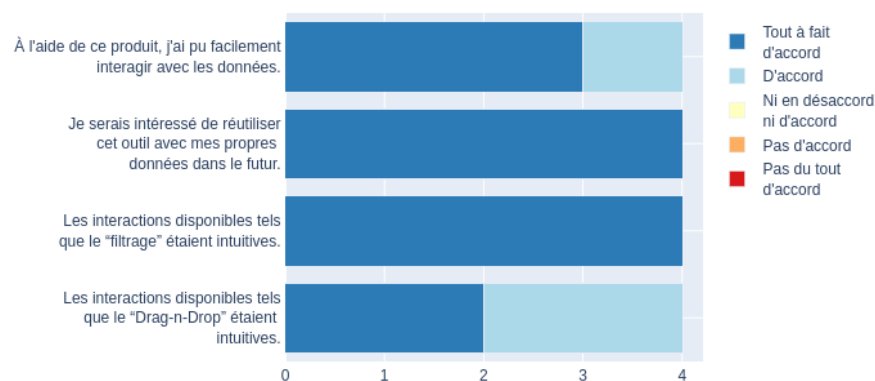


Figure Figure 5.4 Résultats du questionnaire sur l'expérience utilisateur

Premièrement, nous avons demandé aux participants de juger l'utilité de notre outil en y attribuant une note de 0 à 10. Au total, nous avons récolté une moyenne et une médiane de 8,5 sur 10. De plus, tous les participants ont accordé la mention «Tout à fait d'accord» au fait qu'ils souhaitaient réutiliser notre outil en utilisant leurs propres données. Cette observation nous encourage à explorer un peu plus cette voie et perfectionner notre outil.

Deuxièmement, nous avons demandé aux participants de juger la facilité d'utilisation de notre outil en y attribuant une note de 0 à 10. Au total, nous avons récolté une moyenne de 8,75 et une médiane de 8,5 sur 10. Cet aspect faisait partie des points appréciés par les participants. D'ailleurs, nous avons récolté les témoignages suivants : « *L'outil est assez intuitif et simple [...] j'ai pas du me casser la tête [...] c'est facile d'essayer différentes choses.* » et « *L'outil est simple d'utilisation et n'est pas encombré* ». De manière générale, les participants ont trouvé les interactions faciles d'utilisation. De plus, les interactions de type «filtrage» et «glisser-déposer» (*Drag-n-Drop*) ont été jugées intuitives. Ces deux interactions ont été citées par les participants comme faisant partie des points forts de l'outil. Comme prévu, l'interaction glisser-déposer a été utilisée par les participants (parfois combinée avec du filtrage) pour lier, réarranger et séparer les vues générées.

Les participants ont également utilisé l'interaction de filtrage dans le panneau de gauche comportant toutes les questions de sondages. Cependant, certains participants rencontraient quelques difficultés en utilisant ce panneau. Premièrement, lors de l'application d'un filtre (ex. : genre = femme), l'interaction qui fait apparaître l'info-bulle montrant la catégorie sélectionnée était lente (2 s). Certains participants ne remarquaient donc pas cette possibilité et ne savaient pas ce qui avait été sélectionné. Deuxièmement, lors de l'utilisation de filtres multiples (ex. : genre = femme, age = 25-34Y), l'un des participants a déclaré : "*j'aurais aimé avoir une zone de texte m'affichant quelles catégories ont été sélectionnées. Ainsi, je pourrai aussi faire attention si j'en sélectionne (une catégorie) par erreur*".

Enfin, nous avons demandé aux participants de juger leur satisfaction vis-à-vis de notre outil en y attribuant une note de 0 à 10. Au total, nous avons récolté une moyenne et une médiane de 8 sur 10. Cependant, il est important de préciser que durant les évaluations certains bogues informatiques ont été repérés. Les participants, étant plus experts dans leur domaine, ont effectué des analyses et combinaisons plus complexes que celles effectuées par nos soins durant nos tests.

En raison du nombre réduit de participants, nous ne pouvons pas confirmer ou infirmer *H5* qui stipule que *des interactions de type glisser-déposer et filtrage améliorent l'expérience d'un utilisateur novice*.

Néanmoins, les résultats préliminaires portent à croire que les interactions «glisser-déposer» et de filtrage améliorent l'expérience de l'utilisateur novice. Cependant, certaines corrections pourraient améliorer cette expérience.

5.2 Synthèse des recommandations proposées

Dans cette section, nous allons synthétiser les recommandations et améliorations proposées par nos participants au cours de l'évaluation. Ces recommandations ont été regroupées en trois catégories : visualisations, gestion de données et interactions, et exportation des vues.

5.2.1 Visualisations

La moitié des participants ont exprimé leur désir de choisir eux même les visualisations. Tandis que d'autres étaient satisfaits par les vues générées. Afin de satisfaire ces besoins, certains ont proposé de générer une visualisation tout en laissant le choix à l'utilisateur d'opter ou non pour une autre représentation. Ce compromis permettrait à l'utilisateur de tester plusieurs alternatives et se forger sa propre opinion finale. Dans des travaux futurs, il serait intéressant d'élargir notre panel de représentations disponibles.

Dans la même catégorie, certains participants ont proposé de présenter un guide des visualisations plus interactif et dynamique que celui présenté actuellement (bouton «*How to read the chart*»). Cela pourrait être sous forme de glossaire complet avec des démonstrations interactives.

5.2.2 Gestion de données et interactions

Un des principaux points réclamés par nos participants est l'ajout d'actions permettant d'interagir avec les données. Tout d'abord, cela concernait la gestion des données manquantes (absences de réponses, le participant ne veut pas répondre, etc.) qui était un point que nous avons omis. En effet, les participants ont souligné l'importance des données manquantes et donc souhaitaient avoir le choix de les garder ou les enlever durant leur analyse.

Aussi, plusieurs interactions, autres que le filtrage, avaient été sollicitées, par exemple :

1. réordonner des réponses ;
2. filtrer les réponses à afficher avant la création d'une visualisation ;
3. grouper plusieurs catégories de réponses.

Ces interactions, effectuées de manière itérative, élargissent le champ d'actions possibles durant une analyse exploratoire. Cependant, cela devra être fait de manière subtile et minimaliste afin de ne pas encombrer l'utilisateur novice.

De plus, certaines recommandations ont été proposées afin d'améliorer l'expérience utilisateur, par exemple :

1. avoir un rapport des différents filtres appliqués ;
2. pouvoir réinitialiser les filtres des visualisations générées ;
3. pouvoir supprimer d'un coup une ou plusieurs visualisations générées ;
4. pouvoir modifier la taille des visualisations générées.

5.2.3 Exportation des vues

Tous les participants avaient exprimé la nécessité d'exporter la visualisation ou l'ensemble de visualisations créées. Cela permettra à l'utilisateur de diffuser et partager ses résultats avec ses collègues ou ses professeurs. De plus, un besoin de personnaliser ses vues a été émis. Cela pourrait être fait à travers :

1. le choix des couleurs des visualisations ;
2. l'ajout d'éléments textuels décrivant une visualisation (titre, source...) ;
3. le renommage des réponses d'une question.

5.3 Limitations

Nous présenterons dans cette section les différentes limitations reliées à notre outil et celles reliées au protocole d'évaluation.

5.3.1 Limitations de l'outil de visualisation

Grâce à l'expérience utilisateur, nous avons pu déceler les limitations de notre outil.

Limitations de design

Une première limitation de notre outil réside dans la gestion des données manquantes dans un sondage. Cette opération semble, selon les témoignages de nos participants, être nécessaire dans l'exploration d'un sondage.

De plus, notre outil, dans sa version actuelle, ne prend en considération que 4 types de questions de sondage : questions fermées à choix unique, échelle d'évaluation, échelle de Likert et questions ouvertes à valeur numérique. D'autres types de questions doivent être reconsidérés dans les travaux futurs, ex. : questions ouvertes textuelles, date, position géographiques, etc.

Enfin, l'introduction d'une visualisation, jugée moins conventionnelle, à un utilisateur novice devrait être améliorée. En effet, la méthode actuelle utilisée (bouton « *How to read the chart* ») ne semble pas être intuitive pour tous les participants. Nous proposons de mettre plus en

évidence ce bouton ou bien l'apparition automatique d'une fenêtre contextuelle «*pop-up*» lors de la génération d'un nouveau type de représentation. Cette fenêtre pourrait comporter un mini tutoriel interactif que l'utilisateur pourrait choisir de suivre ou ignorer.

Limitations techniques

Lors du chargement initial des données, un temps d'attente est observé. En effet, à chaque utilisation de l'outil les données sont retéléchargées à partir de la base de données. Il serait intéressant de trouver un moyen de réduire ce temps d'attendre en gardant en cache les dernières données chargées afin de les réutiliser dans le futur.

Nous avons également remarqué une latence de 1 à 2 secondes pendant l'opération de filtrage. Cette latence pourrait s'aggraver avec l'augmentation du nombre de questions posées et réponses collectées. Cette opération de filtrage est principalement réalisée sur le panneau de gauche et est prise en charge par la librairie `crossfilter.js` [72]. Nous préconisons d'optimiser cette interaction afin d'améliorer l'expérience de l'utilisateur.

Enfin, des bogues informatiques ont été repérés et notés suite à l'expérience. Ces derniers sont principalement apparus dans l'interaction «glisser-déposer» pour lier les vues.

5.3.2 Limitations de l'expérience

Parmi les facteurs susceptibles de causer un biais dans les données observées, il y'a la taille de l'échantillon, la nature des données visualisées, la diversité de l'étude et l'environnement semi-contrôlé. Les facteurs les plus pertinents sont examinés ci-dessous.

Taille de l'échantillon

La principale limitation que nous décelons est la taille de notre échantillon. Les résultats obtenus ne sont statiquement pas significatifs pour infirmer ou confirmer certaines hypothèses. Néanmoins, le fait d'avoir un échantillon de petite taille nous a permis d'accorder plus de temps d'échange à chaque participant et ainsi récolté des retours plus qualitatifs. Nous recommandons cependant de réitérer l'expérience avec un plus grand nombre de participants.

Nature des données

L'outil de visualisation est destiné à tout type de sondages et d'enquêtes. Cependant, l'évaluation n'a été faite qu'avec une seule enquête : l'ÉCSÉ. Cela nous a, certes, permis d'unifier les paramètres de l'expérience pour tous les participants. Cependant, nous ne pouvons pas

généraliser l'efficacité de notre outil sur d'autres types d'enquêtes : transversales, longitudinales à panel, etc. Nous recommandons à l'avenir de réitérer l'expérience auprès de plusieurs sous-groupes de participants dont chacun aurait un type d'enquête différent.

Diversité de l'étude

Pendant cette première évaluation, plusieurs éléments de design ont été évalués en un temps réduit (15 minutes). Nous avons parcouru : les visualisations générées, le système de génération de requêtes, les interactions, etc. Néanmoins, chaque élément de design pourrait être évalué individuellement. Par exemple, une expérience pourrait être proposée afin d'évaluer uniquement le système de génération de requêtes.

Environnement semi-contrôlé

Réaliser une évaluation sous forme de vidéoconférence nous a permis de nous assurer du sérieux du participant tout au long de l'expérience. Cependant, nous pensons que la présence de la responsable de projet durant l'expérience pourrait avoir été un facteur influant sur les réponses des participants les poussant à être plus indulgents ou moins objectifs.

De plus, l'environnement étant partiellement contrôlé nous ne pouvions pas prédire la taille de l'écran du participant ou son fureteur par exemple. Les résultats peuvent donc avoir été influencés par ces paramètres.

Nous recommandons de réitérer l'expérience en laboratoire dans un environnement contrôlé et cela sans la présence de la responsable du projet.

CHAPITRE 6 CONCLUSION

Les sondages sont un moyen efficace de sonder l'opinion d'une population à partir d'un échantillon. Une fois le sondage administré, une analyse des réponses récoltées doit se faire en combinant des outils statistiques, de la programmation et de la visualisation de données. Cependant, avec l'accroissement du nombre de questions et la complexité des données, l'analyse de ces données peut vite devenir une tâche ardue pour un utilisateur novice ayant un bagage limité en visualisation de données et programmation.

6.1 Synthèse des travaux

Notre projet de recherche a porté sur la conception et l'évaluation d'un outil d'exploration visuelle de données de sondages destiné à un public novice. Cet outil pourrait, par exemple, être pertinent pour des étudiants du premier cycle en sciences politiques ayant des connaissances basiques en programmation et visualisation de données. L'objectif de notre outil est de permettre à l'utilisateur novice d'effectuer facilement ses analyses et tâches d'exploration (distribution, relation entre variables et évolution temporelle) sur des données de sondage.

La visualisation développée est basée sur différentes recommandations récoltées à partir de la littérature, par exemple : l'approche «*overview first, details on-demand*», la formulation semi-textuelle d'une requête et la génération automatique de visualisation. Ces éléments avaient pour but d'atteindre les objectifs suivants : 1) générer des visualisations adaptées au type de tâches et questions de sondage choisies par l'utilisateur novice, 2) traduire des hypothèses de recherche en visualisations de manière simple et intuitive, et 3) effectuer les principales tâches d'exploration.

Une étude a été réalisée auprès de quatre étudiants en science politique dans le but d'évaluer la validité de nos hypothèses de recherche :

(*H1*) : Une vue globale dans la visualisation montrant toutes les questions de sondages et les métadonnées permet à l'utilisateur d'émettre de nouvelles hypothèses.

(*H2*) : Un choix explicite de tâches et questions de sondage permet à un utilisateur novice de traduire ses hypothèses en visualisation.

(*H3*) : La prédéfinition de visualisations par tâches et types de variable permet l'utilisateur d'explorer efficacement son jeu de données.

(*H4*) : La présentation de potentielles nouvelles visualisations à l'utilisateur novice lui permet d'améliorer ses connaissances en visualisations de données.

(*H5*) : Des interactions de type glisser-déposer et filtrage améliorent l'expérience d'un utilisateur novice.

En raison du nombre réduit de participants, nos résultats ne sont statistiquement pas significatifs. Nous ne pouvons donc pas infirmer ou confirmer nos hypothèses. Cependant, les résultats préliminaires semblent indiquer que les hypothèses *H2*, *H3* et *H5* sont valides. Cela doit cependant être confirmé auprès d'un plus large échantillon.

L'évaluation a également permis de recueillir diverses rétroactions, recommandations et pistes pour des travaux futurs. Les résultats préliminaires montrent un vif intérêt, de la part des participants, envers notre outil. Les participants ont jugé que l'outil et ses interactions diverses étaient faciles d'utilisation et donc adaptés à un utilisateur novice. Les visualisations générées ont été pour la plupart jugées pertinentes. Plusieurs améliorations ont été proposées par nos participants en termes d'interactions, visualisations et de gestion de données manquantes.

6.2 Améliorations futures

Les résultats de l'étude préliminaire nous ont permis de cerner les principaux besoins de notre public cible en termes de visualisations. Les retours récoltés portent à croire que cette voie devrait être explorée plus en détail. Cependant, certains points pourraient être améliorés concernant l'outil de visualisation et son évaluation.

En ce qui concerne l'outil de visualisation, voici nos recommandations pour les versions futures :

- prendre en charge un plus large éventail de questions de sondages : questions ouvertes, dates, etc. ;
- effectuer une analyse plus complète sur les travaux des chercheurs en science politique afin d'en extraire d'autres tâches d'exploration,
- développer de nouvelles fonctionnalités permettant d'interagir avec les données (ordonner, grouper, etc.) ;
- améliorer l'aspect éducatif de notre outil de visualisation à travers des tutoriels ou glossaires plus interactifs et adaptés aux utilisateurs novices ;
- implémenter une fonctionnalité permettant l'exportation des visualisations générées ;
- implémenter une fonctionnalité permettant l'ajout des données propres de l'utilisateur.

L'évaluation de notre outil, et autres outils de visualisations similaires, pourrait être améliorée en considérant les points suivants :

- élargir la taille de l'échantillon afin d'avoir des résultats statistiquement plus significatifs ;

- effectuer l'évaluation sur différents types d'enquêtes (transversales, longitudinales, etc.) ;
- mener l'évaluation en laboratoire afin de mieux contrôler les paramètres de l'expérience : taille d'écran, fureteur, etc. ;
- mener l'évaluation sans la présence des responsables de projet ;
- concevoir des évaluations plus spécifiques de l'outil de visualisation. Ces dernières permettront d'évaluer rigoureusement chaque aspect de l'outil.

Plusieurs pistes de recherche peuvent découler de cette étude pour les travaux futurs. Par exemple, nous pourrions tester l'efficacité de notre dispositif en tant qu'outil pédagogique. Une étude pourrait suivre l'utilisation de notre outil par des professeurs avec leurs étudiants.

Aussi, ce travail de recherche a permis d'évaluer notre outil sur un public d'étudiants. Cependant, les sondages étant utilisés dans une multitude de domaines, il serait intéressant de présenter notre outil à un plus large éventail d'utilisateurs : médecins, marqueteur, professeurs et administrations scolaires, sociologues, etc. De futures recherches pourraient se concentrer sur ces différents publics afin de définir et comparer leurs besoins en termes de visualisation.

De plus, les données de sondages sont un cas particulier des données multidimensionnelles. Cependant, notre évaluation a soulevé le point sur certaines considérations propres aux données de sondage, par exemple : la gestion des données manquantes. Les futurs travaux devraient étudier plus en détail les considérations propres aux données de sondages. Dans la même optique, il serait intéressant d'élargir notre outil à des données plus générales en réévaluant les besoins en termes de tâches d'exploration et de visualisations.

Ensuite, notre outil actuel peut être utile pour les utilisateurs plus expérimentés dans la visualisation de données. Il peut être utilisé pour une analyse préliminaire par exemple. Cependant, si l'utilisateur souhaite créer des graphiques ou des interactions plus complexes, il pourrait se retrouver limité par l'outil. Une amélioration possible pour les travaux futurs peut être l'ajout d'un choix plus large de graphiques et d'interactions.

Finalement, la recherche exploratoire est un domaine suscitant un vif intérêt chez les chercheurs [77, 78]. L'analyse devrait être poussée afin d'améliorer la stratégie d'exploration des données de sondage (et données multidimensionnelles). À titre d'exemple, une étude pourrait porter sur le processus d'exploration visuelle horizontale d'une base de données puis l'immersion verticale pour une variable particulière [78]. Cela permettra de concevoir un système qui anticipe et répond aux besoins de l'utilisateur plus efficacement.

RÉFÉRENCES

- [1] “Seeing data,” <http://seeingdata.org/>, visité le : 2020-01-30.
- [2] “Data viz project,” <https://datavizproject.com/>, visité le : 2020-01-30.
- [3] A. Kachkaev, J. Wood et J. Dykes, “Glyphs for exploring crowd-sourced subjective survey classification,” *Computer Graphics Forum*, vol. 33, n^o. 3, p. 311–320, 2014.
- [4] D. Brodbeck et L. Girardin, “Visualization of large-scale customer satisfaction surveys using a parallel coordinate tree,” dans *International Conference on Information Visualization*. IEEE, 2003, p. 197–201.
- [5] A. Slingsby *et al.*, “Designing an exploratory visual interface to the results of citizen surveys,” *International Journal in Geographical Information Science*, vol. 28, n^o. 10, p. 2090–2125, 2014.
- [6] Z. Xu, “Interactive visualization of survey results,” Mémoire de maîtrise, Dép. de génie informatique, Université de Victoria, 2016. [En ligne]. Disponible : https://dspace.library.uvic.ca/bitstream/handle/1828/7221/Xu_Zheng_MSc_2016.pdf?sequence=1
- [7] E. Anderson, “The species problem in iris,” *Annals of the Missouri Botanical Garden*, vol. 23, n^o. 3, p. 457–509, 1936.
- [8] D. B. Carr *et al.*, “Scatterplot matrix techniques for large n,” *Journal of the American Statistical Association*, vol. 82, n^o. 398, p. 424–436, 1987.
- [9] A. Inselberg et B. Dimsdale, “Parallel coordinates : a tool for visualizing multi-dimensional geometry,” dans *Visualization’90*. IEEE, 1990, p. 361–378.
- [10] “Titanic data set (statlib),” 2006.
- [11] R. Kosara, F. Bendix et H. Hauser, “Parallel sets : Interactive exploration and visual analysis of categorical data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, n^o. 4, p. 558–568, 2006.
- [12] H. Chernoff, “The use of faces to represent points in k-dimensional space graphically,” *Journal of the American Statistical Association*, vol. 68, n^o. 342, p. 361–368, 1973.
- [13] M. A. Yalçın, N. Elmqvist et B. B. Bederson, “Keshif : Rapid and expressive tabular data exploration for novices,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, n^o. 8, p. 2339–2352, 2017.
- [14] S. Gratzl *et al.*, “Lineup : Visual analysis of multi-attribute rankings,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, n^o. 12, p. 2277–2286, 2013.

- [15] —, “Domino : Extracting, comparing, and manipulating subsets across multiple tabular datasets,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, n^o. 12, p. 2023–2032, 2014.
- [16] The Comparative Study of Electoral Systems, “Integrated module dataset,” Octobre 2019. [En ligne]. Disponible : www.cses.org
- [17] “Échantillon (statistiques),” dans *Wikipédia*, 13 juin 2020. [En ligne]. Disponible : [https://fr.wikipedia.org/wiki/%C3%89chantillon_\(statistiques\)](https://fr.wikipedia.org/wiki/%C3%89chantillon_(statistiques))
- [18] J. Blair, R. F. Czaja et E. A. Blair, *Designing surveys : A guide to decisions and procedures*. Sage Publications, 2013.
- [19] “Google forms,” (accédé le 29-Octobre-2019). [En ligne]. Disponible : <https://docs.google.com/forms>
- [20] “SurveyMonkey,” (accédé le 29-Octobre-2019). [En ligne]. Disponible : <https://fr.surveymonkey.com>
- [21] “SurveyGizmo,” (accédé le 29-Octobre-2019). [En ligne]. Disponible : <https://www.surveygizmo.com/>
- [22] J. Heer *et al.*, “Creation and collaboration : Engaging new audiences for information visualization,” dans *Information Visualization*. Springer, 2008, p. 92–133.
- [23] Z. Pousman, J. Stasko et M. Mateas, “Casual information visualization : Depictions of data in everyday life,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, n^o. 6, p. 1145–1152, 2007.
- [24] L. Grammel, M. Tory et M.-A. Storey, “How information visualization novices construct visualizations,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, n^o. 6, p. 943–952, 2010.
- [25] T. Munzner, *Visualization analysis and design*. CRC Press, 2014.
- [26] A. Afifi *et al.*, *Practical multivariate analysis*. CRC Press, 2019.
- [27] S. Liu *et al.*, “Visualizing high-dimensional data : Advances in the past decade,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, n^o. 3, p. 1249–1268, 2016.
- [28] A. Kirk, *Data visualisation : A handbook for data driven design*. Sage, 2016.
- [29] J. W. Tukey, *Exploratory data analysis*. Reading, MA, 1977, vol. 2.
- [30] S. Tufféry, *Data mining and statistics for decision making*. John Wiley & Sons, 2011.
- [31] P. F. Velleman et D. C. Hoaglin, *Applications, basics, and computing of exploratory data analysis*. Duxbury Press, 1981.

- [32] C. H. Yu, “Exploratory data analysis in the context of data mining and resampling.” *International Journal of Psychological Research*, vol. 3, n^o. 1, p. 9–22, 2010.
- [33] S. Few, “Data visualization for human perception,” dans *The Encyclopedia of Human-Computer Interaction*, 2^e éd. Denmark : Interaction Design Foundation, 2014, ch. 35.
- [34] B. Shneiderman, “The eyes have it : A task by data type taxonomy for information visualizations,” dans *International Conference on Visual Language*. IEEE, 1996, p. 336–343.
- [35] R. Amar, J. Eagan et J. Stasko, “Low-level components of analytic activity in information visualization,” dans *International Conference on Information Visualization*. IEEE, 2005, p. 111–117.
- [36] A. Ghosh *et al.*, “A comprehensive review of tools for exploratory analysis of tabular industrial datasets,” *Visual Informatics*, vol. 2, n^o. 4, p. 235–253, 2018.
- [37] M. Elias et A. Bezerianos, “Exploration views : understanding dashboard creation and customization for visualization novices,” dans *Conference on Human-Computer Interaction*. Springer, 2011, p. 274–291.
- [38] J. Heer et B. Shneiderman, “Interactive dynamics for visual analysis,” *Queue*, vol. 10, n^o. 2, p. 30–55, 2012.
- [39] M. Smuts, B. Scholtz et A. Calitz, “Design guidelines for business intelligence tools for novice users,” dans *Annual Research Conference on South African Institute of Computer Scientists and Information Technologists*, 2015, p. 1–15.
- [40] R. Sapsford, *Survey research*. Sage, 2006.
- [41] B. P. Caetano *et al.*, “A proposal for visualization techniques recommendation to represent survey data,” dans *11th Iberian Conference on Information Systems and Technologies*. IEEE, 2016, p. 1–6.
- [42] J. Sreevalsan-Nair *et al.*, “Collaborative design of visual analytic techniques for survey data for community-based research in public health,” dans *8th Workshop on Visual Analytics in Healthcare*. IEEE, Octobre 2017.
- [43] M. Q. Wang Baldonado, A. Woodruff et A. Kuchinsky, “Guidelines for using multiple views in information visualization,” dans *International Working Conference on Advanced Visual Interfaces*, 2000, p. 110–119.
- [44] M. Friendly, “Corrgrams : Exploratory displays for correlation matrices,” *The American Statistician*, vol. 56, n^o. 4, p. 316–324, 2002.
- [45] J. Heinrich et D. Weiskopf, “State of the art of parallel coordinates.” dans *Eurographics*. State of the Art Reports, 2013, p. 95–116.

- [46] M. Friendly et E. Kwan, “Effect ordering for data displays,” *Computational Statistics and Data Analysis*, vol. 43, n^o. 4, p. 509–539, 2003.
- [47] G. Grinstein, M. Trutschl et U. Cvek, “High-dimensional visualizations,” dans *Visual Data Mining Workshop, KDD*, vol. 2. Citeseer, 2001, p. 120.
- [48] G. Ellis et A. Dix, “A taxonomy of clutter reduction for information visualisation,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, n^o. 6, p. 1216–1223, 2007.
- [49] C. Chang *et al.*, “Evaluating perceptually complementary views for network exploration tasks,” dans *Conference on Human Factors in Computing Systems*. SIGCHI, 2017, p. 1397–1407.
- [50] R. Sciabassi, “Computational support for the development of medical multimedia applications,” *Center for Clinical Neurophysiology, University of Pittsburgh Medical Center*, 1993.
- [51] A. Blackstone *et al.*, *Principles of sociological inquiry : Qualitative and quantitative methods*. Saylor Foundation, 2012.
- [52] “Questionnaire,” dans *Wikipédia*, 24 janvier 2020. [En ligne]. Disponible : <https://fr.wikipedia.org/wiki/Questionnaire>
- [53] P. J. Lavrakas, *Encyclopedia of survey research methods*. Sage Publications, 2008.
- [54] A. Blais et C. H. Achen, “Civic duty and voter turnout,” *Political Behavior*, vol. 41, n^o. 2, p. 473–497, 2019.
- [55] A. Blais *et al.*, “What is the cost of voting?” *Electoral Studies*, vol. 59, p. 145–157, 2019.
- [56] D. Stockemer et A. Blais, “Voters and abstainers in national and european elections,” *European Review*, vol. 27, n^o. 2, p. 300–315, 2019.
- [57] A. Blais, F. Feitosa et S. Sevi, “Was my decision to vote (or abstain) the right one?” *Party Politics*, vol. 25, n^o. 3, p. 382–389, 2019.
- [58] S. Sevi *et al.*, “How do turks abroad vote?” *Turkish Studies*, vol. 21, n^o. 2, p. 208–230, 2020.
- [59] F. Kostelka, A. Blais et E. Gidengil, “Has the gender gap in voter turnout really disappeared?” *West European Politics*, vol. 42, n^o. 3, p. 437–463, 2019.
- [60] I. Lago et A. Blais, “Early voting and satisfaction with vote choice,” *Representation*, vol. 55, n^o. 2, p. 225–233, 2019.
- [61] S. Sevi, V. Arel-Bundock et A. Blais, “Do women get fewer votes ? no.” *Canadian Journal of Political Science*, vol. 52, n^o. 1, p. 201–210, 2019.
- [62] L. Rheault *et al.*, “Understanding people’s choice when they have two votes,” *Journal of Elections, Public Opinion and Parties*, p. 1–18, 2018.

- [63] D. Bol *et al.*, “Voting and satisfaction with democracy in flexible-list pr,” *Electoral Studies*, vol. 56, p. 23–34, 2018.
- [64] J. Aldrich, A. Blais et L. B. Stevenson, *Many Faces of Strategic Voting : Tactical Behavior in Electoral Systems Around the World*. University of Michigan Press, 2018.
- [65] S. E. Harpe, “How to analyze likert and other rating scale data,” *Currents in Pharmacy Teaching and Learning*, vol. 7, n^o. 6, p. 836–850, 2015.
- [66] K. Achilleas, “On likert scales, ordinal data and mean values.” Fevrier 2013. [En ligne]. Disponible : <https://achilleaskostoulas.com/2013/02/13/on-likert-scales-ordinal-data-and-mean-values/#Likert>
- [67] J. Carifio et R. Perla, “Resolving the 50-year debate around using and misusing likert scales,” *Medical Education*, vol. 42, n^o. 12, p. 1150–1152, 2008.
- [68] “Distribution,” dans *Wikipédia*, 24 janvier 2019. [En ligne]. Disponible : <https://fr.wikipedia.org/wiki/Distribution>
- [69] J. S. Yi *et al.*, “Toward a deeper understanding of the role of interaction in information visualization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, n^o. 6, p. 1224–1231, 2007.
- [70] M. Bostock, V. Ogievetsky et J. Heer, “D³ data-driven documents,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, n^o. 12, p. 2301–2309, 2011.
- [71] W. Gordon *et al.* (2012) dc.js. [En ligne]. Disponible : <https://github.com/dc-js/dc.js>
- [72] Square Inc. (2012) crossfilter. [En ligne]. Disponible : <https://github.com/crossfilter/crossfilter>
- [73] Plotly Technologies Inc., “Collaborative data science,” Montreal, QC, 2015. [En ligne]. Disponible : <https://plot.ly>
- [74] A. Barsky *et al.*, “Cerebral : Visualizing multiple experimental conditions on a graph with biological context,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, n^o. 6, p. 1253–1260, 2008.
- [75] M. Meyer, T. Munzner et H. Pfister, “Mizbee : a multiscale synteny browser,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, n^o. 6, p. 897–904, 2009.
- [76] “Biais de confirmation,” dans *Wikipedia*, 29 avril 2020. [En ligne]. Disponible : https://fr.wikipedia.org/wiki/Biais_de_confirmation
- [77] R. W. White et R. A. Roth, “Exploratory search : Beyond the query-response paradigm,” *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 1, n^o. 1, p. 1–98, 2009.

- [78] M. Dörk, S. Carpendale et C. Williamson, “The information flaneur : A fresh look at information seeking,” dans *Conference on Human Factors in Computing Systems*. SIGCHI, 2011, p. 1215–1224.

ANNEXE A ANALYSE DE TRAVAUX ET ABSTRACTION DE TÂCHES

Travaux	Année	Tâches abstraites
Civic duty and voter turnout. [54]	2019	- Distribution des variables - Relation bi-variée
What is the cost of voting? [55]	2019	- Distribution des variable - Relation multi-variée
Voters and Abstainers in National and European Elections. [56]	2019	- Distribution des variables
Was my decision to vote (or abstain) the right one? [57]	2019	- Distribution des variable - Relation bivariée - Relation multi-variée
How do Turks abroad vote? [58]	2019	- Distribution des variable - Relation bivariée
Has the gender gap in voter turnout really disappeared? [59]	2019	- Distribution des variables - Relation multi-variée - Evolution temporelle
Early Voting and Satisfaction with Vote Choice. [60]	2019	- Distribution des variables - Relation multi-variée
Do Women Get Fewer Votes? No. [61]	2019	- Distribution des variables - Evolution temporelle
Understanding people's choice when they have two vote. [62]	2018	- Distribution des variables - Relation multi-variée
Voting and satisfaction with democracy in flexible-list PR. [63]	2018	- Distribution des variables - Relation bivariée
Voting strategically in two-vote elections. [64]	2018	Distribution des variables

Tableau Tableau A.1 Liste de travaux analysés pour l'extraction de tâches abstraites.

ANNEXE B GUIDE DE DESCRIPTIONS DES VISUALISATIONS GÉNÉRÉES

Visualisation	Description
Graphique à barres	« The absolute length of each bar represents the quantity/value of each category. »
Graphique à barres + Boite à moustache	« The absolute length of each bar represents the quantity/value of each category. A box plot displays median, higher/lower quartiles and maximum/minimum. Source : www.datavizproject.com »
Graphique à barres empilées	« Stacked chart uses the length of two or more stacked bars to represent the components of a total quantitative value across a range of different categorical values. Source : www.seeingdata.org »
Nuage de carrés	« The figure represents three dimensions. Each entity with its triplet (v_1 , v_2 , v_3) of associated data is plotted as a square. The square's xy location expresses two of the v_i values and the third v_i value is expressed through the size. »
Graphique de Violin	« The violin plot is similar to box plots, except that it also shows the probability density of the data at different values. Typically, violin plots will include a marker for the median of the data and a box indicating the interquartile range, as in standard box plots. Source : www.datavizproject.com »
Nuage de points	« This chart displays three dimensions of data. Each entity with its triplet (v_1 , v_2 , v_3) of associated data is plotted as a disk that expresses two of the v_i values through the disk's xy location and the third through its size. Source : www.datavizproject.com »
Graphique de Sankey	« Variables are assigned to vertical axes that are parallel. Categories are represented with blocks on each axis. The height of a block represents the size of the cluster and the height of a stream field represents the size of the components contained in both blocks connected by the stream field. Source : www.datavizproject.com »
Graphique de coordonnées paralleles	« Variables are assigned to vertical axes that are parallel. Each line represents the values of one data through all the categories. Considering different axis arrangements may be of interest. Source : www.datavizproject.com »

Tableau Tableau B.1 Visualisations générées et descriptions associées. Les descriptions sont reprises et adaptées des sites web : [seeingdata](http://seeingdata.org) [1] et [datavizproject](http://datavizproject.com) [2].

ANNEXE C COURRIEL DE RECRUTEMENT

Bonjour,

Dans le cadre de mon projet de maîtrise en visualisation de données, je travaille sur l'implémentation d'un outil de visualisation permettant d'explorer des données de sondage. Cet outil est destiné particulièrement à un utilisateur novice ayant des connaissances limitées en visualisation de données. Mon équipe de recherche a reçu l'autorisation du professeur André BLAIS afin de vous solliciter à participer à l'évaluation de mon outil.

Cette évaluation aura lieu au cours d'une vidéoconférence de 50 minutes, selon vos disponibilités au courant du mois de juin ou juillet 2020. Il est à noter que la participation à cette recherche sera enregistrée (voix et capture d'écran).

Un total de six (6) personnes seront retenues. Pour participer, vous devez être âgé de plus de 18 ans et être à l'aise avec l'utilisation d'un ordinateur. Vous devrez également vous munir du logiciel de vidéoconférence Zoom.

Votre participation est entièrement volontaire. De plus, vous pourrez en tout temps quitter la vidéoconférence si vous ne souhaitez plus participer à l'évaluation.

Nous offrirons aux personnes ayant participé à l'entrevue une carte-cadeau d'une valeur de 20\$ chez Renaud-Bray par voie postale.

Si vous souhaitez participer à l'évaluation, cliquez sur ce lien. Il vous redirigera vers le formulaire de participation afin de choisir vos disponibilités. Les informations personnelles que vous introduirez serviront à vous recontacter afin de confirmer votre participation. Elles seront ensuite détruites.

Je vous remercie à l'avance pour votre intérêt à l'évaluation de mon outil de visualisation.

Cordialement,

Manal Maroua BOUCENNA

Courriel : manal-maroua.boucenna@polymtl.ca

Étudiante à la maîtrise en génie logiciel

Polytechnique Montréal

ANNEXE D FORMULAIRE D'INFORMATION ET DE CONSENTEMENT

Titre de l'activité de recherche :

Méthode d'exploration visuelle de données de sondage destinée à un utilisateur novice.

Équipe de recherche :

Responsable de l'activité de recherche

Manal Maroua Boucenna

Candidate à la maîtrise en génie logiciel

École Polytechnique de Montréal

C.P. 6079, succ. Centre-ville

Montréal (Québec)

H3C 3A7

Tél. (514) 340-4711 poste 7109

Fax. (514) 340-5139

Adresse courriel : manal-maroua.boucenna@polymtl.ca

Sous la direction de :

Thomas Hurtut

École Polytechnique de Montréal

C.P. 6079, succ. Centre-ville

Montréal (Québec)

H3C 3A7

Tél. (514) 340-4711 poste 7109

Fax. (514) 340-5139

Adresse courriel : thomas.hurtut@polymtl.ca

Financement de l'activité de recherche

La présente activité de recherche est financée par Polytechnique Montréal.

Conflits d'intérêts

L'équipe de recherche n'est pas en situation de conflit d'intérêts dans le contexte de la présente activité de recherche.

Préambule

Nous vous invitons à participer à une activité de recherche qui vise à évaluer l'efficacité d'une nouvelle méthode d'exploration visuelle de données de sondages destinée aux utilisateurs novices en visualisation de données.

Cependant, avant d'accepter de participer à cette activité et de signer le présent formulaire d'information et de consentement, veuillez prendre le temps de lire l'information présentée.

Nous vous invitons à poser toutes les questions que vous jugerez utiles à la responsable ou au responsable de l'activité de recherche ou à tout autre membre de l'équipe de recherche et à leur demander de vous expliquer tout mot ou renseignement qui ne serait pas clair. Nous vous invitons également à prendre conseil auprès de toute autre personne de qui vous aimeriez obtenir un avis à propos de votre éventuelle participation.

Présentation générale du projet de recherche

L'objectif principal de ce projet de recherche est de développer une méthode de visualisation adaptée à un utilisateur novice lui permettant d'explorer les données de sondage. Plus particulièrement, cette méthode devra permettre de :

1. générer des visualisations adaptées au type de tâches et données choisies ;
2. traduire des hypothèses de recherche en visualisations de manière simple et intuitive ;
3. effectuer les principales tâches d'exploration.

Afin d'atteindre nos objectifs, nous avons développé un outil interactif et intuitif permettant d'explorer visuellement les données de sondage. Nous avons conçu un questionnaire qui permettra d'évaluer la capacité de cet outil à accomplir certaines tâches.

Critères d'inclusion et d'exclusion

Dans le cadre de cette activité, nous cherchons à recruter des personnes majeures qui sont suffisamment à l'aise avec un ordinateur pour effectuer une vidéoconférence et interagir avec les éléments du questionnaire fourni.

Nature et durée de votre participation à l'activité de recherche

Nous effectuerons une seule rencontre de 30 à 50 minutes en visioconférence sur l'application Zoom. L'entrevue aura lieu entre le mois de juin et juillet 2020 selon vos disponibilités (matin, après-midi ou soirée). Il n'y a aucune restriction sur le lieu de l'entrevue, cependant nous préconisons de choisir un lieu calme.

Des enregistrements audio et des captures d'écran auront lieu pendant une partie de l'entrevue.

Premièrement, vous serez invité à répondre à un questionnaire en ligne afin de fournir quelques informations générales : genre, groupe d'âge, etc. Ces informations nous permettront de mieux évaluer la diversité de notre échantillon de recherche et comprendre nos résultats. Cette étape ne devrait pas durer plus de 2 minutes.

Deuxièmement, une démonstration de l'outil en ligne vous sera faite par la responsable de l'expérience. Vous pourrez tester l'outil en ligne à votre tour et posez vos questions. Cette étape devrait durer 10 minutes.

Troisièmement, vous serez invité à utiliser l'outil en ligne pour une durée de 15 minutes et explorer les données d'un sondage. Nous enregistrerons l'activité sur votre écran ainsi que votre voix dans le but d'étudier les remarques et comportements généraux adoptés par les participants.

Quatrièmement, vous serez invité à répondre à un second questionnaire qui permettra d'évaluer votre expérience avec notre outil. Cette étape devrait durer 5 minutes.

Finalement, la responsable de l'expérience vous posera quelques questions ouvertes afin de recueillir votre ressenti général et recommandation. Nous enregistrerons votre voix dans le but de retranscrire vos réponses et les étudier.

À la fin de l'entrevue, nous vous demanderons de fournir votre adresse postale afin de vous faire parvenir votre carte-cadeau par voie postale.

Risques pouvant découler de votre participation à l'activité de recherche

Les risques de cette participation sont équivalents aux risques associés à l'utilisation d'un ordinateur et d'un logiciel de vidéoconférence. Afin de minimiser les risques liés à l'utilisation du logiciel de vidéoconférence Zoom, nous vous offrirons un lien et un mot de passe unique pour avoir accès à l'entrevue.

De plus, lors des enregistrements d'écran, il vous sera rappelé de ne pas naviguer sur une autre fenêtre Web que celle de notre outil. Cependant, dans le cas où vous vous tromperiez et changeriez de fenêtre pendant la période d'enregistrement, vous vous verriez exposer au risque d'avoir vos autres fenêtres enregistrées.

Inconvénients pouvant découler de votre participation à l'activité de recherche

L'inconvénient majeur est la perte de temps associée à l'expérience, qui sera entre 30 et 50 minutes.

Avantages pouvant découler de votre participation à l'activité de recherche Vous ne retirerez aucun bénéfice personnel de votre participation au présent projet de recherche. Toutefois, votre participation va permettre de valider une méthode de visualisation inédite,

qui pourra être utilisée dans de nombreux domaines où les données de sondages sont analysées. Elle pourra aider ou inspirer d'autres chercheurs dans le milieu de la visualisation à développer des outils d'exploration de données de sondage plus appropriés aux utilisateurs novices.

Compensation financière Suite à votre participation, une carte-cadeau d'une valeur de 20\$ chez Renaud-Bray vous sera transmise par voie postale approximativement un mois après l'entrevue.

Participation volontaire et possibilité de retrait

Votre participation à la présente activité de recherche est volontaire. Vous êtes donc libre de refuser d'y participer et pouvez à tout moment décider de vous en retirer sans avoir à motiver votre décision et sans risquer d'en subir de préjudice. Vous n'avez qu'à en informer la personne— ressource de l'équipe de recherche, et ce, par simple avis verbal.

En cas de retrait, vous pouvez demander la destruction des données vous concernant. Cependant, il sera impossible de retirer vos données ou votre matériel des analyses menées une fois ces dernières publiées ou diffusées.

Tout au long des activités de recherche, vous recevrez en temps opportun l'information pertinente en lien avec votre participation.

L'équipe de recherche et le comité d'éthique de la recherche se réservent le droit de vous retirer de l'étude si vous ne respectez pas les consignes, s'il existe des raisons administratives d'abandonner l'activité, ou pour toutes autres raisons concernant la faisabilité de l'étude. Si une telle situation survient, l'équipe de recherche vous en informera dès que possible.

Confidentialité et protection de vos données

L'équipe de recherche recueillera et consignera toutes vos données de manière sécuritaire de façon à en protéger le caractère confidentiel.

Voici comment nous protégerons vos données **lors de la collecte** :

- les données recueillies pour l'étude seront complètement anonymes ;
- lors des enregistrements audio, aucune information personnelle ne sera divulguée. Seul un identifiant sera cité afin de vous différencier ;
- lors de l'enregistrement de l'écran, vous serez invité à ouvrir une nouvelle fenêtre sur votre fureteur afin de protéger votre confidentialité. Seules les activités sur notre outil en ligne seront enregistrées ;
- vos données identificatoires (nom, courriel et les adresses postales) seront détruites une fois la compensation financière envoyée par voie postale.

Voici comment nous protégerons vos données **lors des analyses et du transfert des**

données entre les membres de l'équipe :

- les enregistrements audio seront retranscrits puis détruits un mois après l'entrevue,
- seuls les responsables du projet auront accès aux résultats du questionnaire en ligne à l'aide d'un mot de passe ;
- seuls les responsables du projet auront accès aux enregistrements d'écran et audios en s'identifiant avec un mot de passe ;
- aucune autre personne n'aura accès aux données brutes. Seules les conclusions seront partagées aux autres collaborateurs ;
- pendant l'analyse, les données seront conservées en sécurité sur un disque dur externe dans le local du directeur de recherche **Thomas Hurtut**.

Voici comment nous protégerons vos données **lors des publications** :

- les données présentées dans les publications ne feront jamais référence à des informations qui pourraient permettre de reconnaître une personne. Par ailleurs, les résultats obtenus ne seront en aucune façon associés à une personne précise. Des identifiants uniques seront accordés pour différencier les différentes personnes ;
- seules les conclusions seront publiées.

Enfin, voici comment nous protégerons vos données **après le projet de recherche** :

Les données seront conservées pendant 10 ans sur un serveur crypté du département de génie logiciel de Polytechnique Montréal. Suite à cela, les données seront détruites.

Vous avez le droit de consulter votre dossier de recherche pour vérifier l'exactitude des renseignements recueillis aussi longtemps que l'équipe de recherche ou Polytechnique Montréal détiendront ces informations. Cependant, afin de préserver l'intégrité scientifique du projet de recherche, certaines informations seront accessibles seulement à la fin du projet de recherche.

Diffusion des résultats de la recherche

Le chercheur responsable utilisera les données du projet de recherche pour les simples fins du projet de recherche. Les données du projet de recherche pourront être publiées dans des revues scientifiques ou partagées avec d'autres personnes lors de discussions scientifiques. Les enregistrements (voix et capture d'écran) seront retranscrits et analysés. Les données présentées dans les publications ne feront jamais référence à des informations qui pourraient permettre de vous reconnaître. En effet, aucune information identificatoire ne sera diffusée, seuls des identifiants seront accordés pour différencier les participants.

Si vous êtes intéressés par les conclusions de ce projet de recherche, nous pourrions vous envoyer une copie numérique du mémoire à la fin du projet. Il est à noter qu'il ne sera pas

possible pour l'équipe de recherche de faire un lien entre vos réponses et le courriel fourni à la fin du questionnaire.

Indemnisation en cas de préjudice et droits des participant(e)s

Si vous deviez subir quelque préjudice que ce soit par suite de votre participation à cette activité de recherche, vous ne renoncez à aucun de vos droits ni ne libérez les chercheurs, l'organisme de financement ou Polytechnique Montréal de leurs responsabilités légales et professionnelles.

Personnes-ressources

Si vous avez des questions concernant le projet de recherche, vous pouvez communiquer avec Thomas Hurtut au (514) 340-4711, poste 7109 ou encore par courriel à thomas.hurtut@polymtl.ca.

Pour toute préoccupation sur vos droits ou sur les responsabilités de l'équipe de recherche concernant votre participation à ce projet, vous pouvez contacter le Comité d'éthique de la recherche de Polytechnique Montréal au (514) 340-4711, poste 4420 ou encore par courriel à ethique@polymtl.ca

Consentement à la participation au projet de recherche

1. J'ai pris connaissance de la documentation ci-jointe, décrivant la nature et le déroulement du projet de même que les risques et les inconvénients qui pourraient survenir ;
2. Je comprends que j'ai droit à des réponses satisfaisantes aux questions que je poserais quant à mon implication dans ce projet tout au long de ma participation ;
3. Je consens à participer librement à ce projet, après avoir obtenu et pris le temps d'y réfléchir à ma satisfaction et sans avoir subi de pression à cet effet ;
4. Je comprends qu'en participant à ce projet de recherche, je ne renonce à aucun de mes droits ni ne dégage les chercheurs de leurs responsabilités ;
5. Je comprends que je peux consulter le dossier que l'équipe de recherche constitue sur moi ;
6. Je pourrai à tout moment, sur simple avis de ma part, revenir sur ma décision de participer et serai alors immédiatement libéré de mon engagement ;
7. J'ai reçu une copie du présent document.

Veuillez répondre à l'email de recrutement reçu par « **J'ai lu le formulaire d'information et de consentement et j'accepte de participer** » afin de confirmer votre inscription à l'étude.

ANNEXE E QUESTIONNAIRE SUR LES CARACTÉRISTIQUES DU PARTICIPANT

Pour chaque question, le participant a sélectionné une seule option.

1. Quel est votre genre ?
 - Homme
 - Femme
 - Autre
 - Ne préfère pas répondre
2. Quel est votre groupe d'âge ?
 - 18 à 24 ans
 - 25 à 34 ans
 - 35 à 44 ans
 - 45 à 54 ans
 - 55 à 64 ans
 - 65 à 74 ans
 - 75 ans et plus
 - Ne préfère pas répondre
3. Quel est votre niveau d'éducation ?
 - Diplôme d'études secondaires (DES)
 - Diplôme d'études professionnelles (DEP)
 - Diplôme d'études collégiales (DEC)
 - Baccalauréat
 - Maîtrise
 - Doctorat
 - Autre
 - Ne préfère pas répondre
4. Avez-vous suivi une formation spécialisée en visualisation de données ?
 - Oui
 - Non
 - Ne préfère pas répondre

ANNEXE F QUESTIONNAIRE SUR L'EXPLORATION LIBRE

1. À l'aide de ce produit, j'ai fait plusieurs observations pertinentes.
 - Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord
2. À l'aide de ce produit, j'ai découverts des tendances et relations pertinentes entre les données.
 - Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord
3. À l'aide de ce produit, j'ai réalisé toutes mes tâches d'explorations.
 - Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord
4. Le panneau du haut m'a permis de traduire facilement une question que j'avais en tête, en visualisation.
 - Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord
5. Avoir une vue globale de mes données et méta données m'a permis d'émettre de nouvelles hypothèses.

- Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord
6. Avoir une vue détaillée de mes données (panneau central) m'a permis d'infirmier ou confirmer une hypothèse.
- Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord
7. Ne pas avoir à construire mes visualisations m'a permis de me concentrer sur mes tâches d'explorations.
- Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord

ANNEXE G QUESTIONNAIRE SUR LA GÉNÉRATION DE VISUALISATIONS

1. Les visualisations générées étaient pertinentes.

- Tout à fait d'accord
- D'accord
- Ni en désaccord ni d'accord
- Pas d'accord
- Pas du tout d'accord

2. À l'aide de ce produit, j'ai compris l'utilisation de certaines visualisations.

- Tout à fait d'accord
- D'accord
- Ni en désaccord ni d'accord
- Pas d'accord
- Pas du tout d'accord

3. Ne pas pouvoir choisir mes visualisations m'a frustré.

- Tout à fait d'accord
- D'accord
- Ni en désaccord ni d'accord
- Pas d'accord
- Pas du tout d'accord

ANNEXE H QUESTIONNAIRE SUR L'EXPÉRIENCE UTILISATEUR

1. À l'aide de ce produit, j'ai pu facilement interagir avec les données.
 - Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord
2. Les interactions disponibles telles que le «*Drag-n-Drop*» étaient intuitives.
 - Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord
3. Les interactions disponibles, telles que le filtrage, étaient intuitives.
 - Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord
4. Sur une échelle de 1 à 10, considérez-vous cet outil comme étant utile ?
1 aura l'indication «très inutile» et 10 «très utile».
(1) O O O O O O O O O O (10)
5. Je serais intéressé de réutiliser cet outil avec mes propres données dans le futur.
 - Tout à fait d'accord
 - D'accord
 - Ni en désaccord ni d'accord
 - Pas d'accord
 - Pas du tout d'accord

6. Sur une échelle de 1 à 10, avez-vous trouvé cet outil facile à utiliser ?

1 aura l'indication «très difficile» et 10 «très facile».

(1) O O O O O O O O O O (10)

7. Sur une échelle de 1 à 10, êtes-vous satisfait par cet outil ?

1 aura l'indication «Très insatisfait» et 10 «Très satisfait».

(1) O O O O O O O O O O (10)

ANNEXE I QUESTIONS OUVERTES ET DISCUSSIONS

Le participant donne verbalement son avis.

1. Pourriez-vous nous parler des points que vous avez le plus appréciés ?
2. Pourriez-vous nous parler des points que vous avez le moins appréciés ?
3. Avez-vous découvert de nouvelles visualisations à travers cet outil ?
4. Comment interprétez-vous ce graphique ?
5. Quelles fonctionnalités auriez-vous souhaité trouver dans cet outil et qui n'était pas présentes ?