ENSEIGNEMENTS DU TEST UTILISATEUR D'UNE GÉOVISUALISATION DYNAMIQUE

Des améliorations possibles pour les expériences en cartographie

par Cécile Saint-Marc,[1] Marlène Villanova-Oliver,[1] Paule-Annick Davoine,[1] Cicely Pams-Capoccioni,[2] Dorine Chenier[2]

1. Université Grenoble-Alpes, CNRS, Grenoble-INP, LIG
700 avenue Centrale 38400 SAINT-MARTIN-D'HERES
cecile.saint-marc@univ-grenoble-alpes.fr, marlene.villanova-oliver@univ-grenoble-alpes.fr, paule-annick.davoine@univ-grenoble-alpes.fr

2. SNCF Ingénierie & Projets, LVE
6 avenue François Mitterrand 93210 LA PLAINE SAINT-DENIS
cicely.pams@reseau.sncf.fr, dorine.chenier@reseau.sncf.fr

Cet article présente les résultats d'une étude utilisateurs qui visait à comparer deux interfaces de géovisualisation (une interactive et l'autre interactive et animée) montrant les chaînes d'événements (effets dominos) survenus pendant les inondations majeures passées. L'analyse des résultats a révélé des limites dans l'expérimentation, qui concernaient les variables mesurées (taux de bonnes réponses et vitesse de réponse) et les tâches réalisées par les participants. Les résultats ont notamment montré que les deux variables les plus communément utilisées dans les expérimentations en cartographie ne semblent pas suffisantes pour révéler des différences entre les deux interfaces. D'autres métriques, extraites des traces d'utilisation des participants, ont été utilisées avec succès. Pour diminuer la variabilité entre les résultats des expérimentations cartographiques et pour améliorer la comparabilité des résultats entre études, nous émettons une proposition de solution : développer et partager des jeux de données et des jeux de tâches, de complexités cognitives connues et couvrant différentes thématiques, comme cela est déjà mis en œuvre dans d'autres champs de recherche.

Introduction

Les développements récents en cartographie ont montré un intérêt croissant pour les cartes animées, pour visualiser les dynamiques des phénomènes géographiques. Cependant, la recherche a échoué à déterminer si ce type de carte était efficace pour comprendre des phénomènes complexes, les résultats expérimentaux étant incohérents (Griffin *et al.* 2006). Ces incohérences dans les résultats des études sont souvent causées soit par le contexte dans lequel les cartes sont utilisées, par exemple des tâches qui sont plus faciles à remplir avec un type de cartes, soit par le fait que des approches cognitives différentes sont employées par les lecteurs selon que la carte est animée ou non, phénomène connu sous le terme de « non-équivalence computationnelle » des cartes (Fabrikant *et al.* 2008). Selon Tversky *et al.* (2002), les incohérences de résultats proviennent souvent d'erreurs dans les protocoles expérimentaux : par exemple si l'un des types de carte testé communique plus d'informations que les autres avec lesquels il est comparé ou encore si une question mal formulée conduit à un biais dans les réponses. Ces constats nous conduisent à nous demander quelles méthodes ou bonnes pratiques pourraient être mises en place pour éviter ces erreurs, qui continuent de se produire malgré l'attention souvent considérable accordée à la conception des protocoles expérimentaux.

Cet article présente des limites identifiées lors d'un test utilisateur que nous avons conduit. Ce test visait à évaluer l'efficacité d'un outil de géovisualisation pour explorer la chaîne des événements s'étant produits durant les inondations passées. A partir de l'analyse des résultats et de travaux menés dans d'autres champs

disciplinaires, nous proposons une solution qui pourrait diminuer les risques de commettre des erreurs de conception dans les expérimentations et qui pourraient améliorer la comparabilité des résultats entre études.

Tout d'abord, le contexte de l'étude et la méthode expérimentale sont décrits. Ensuite, les résultats principaux sont présentés, conduisant aux limites qu'ils ont révélées dans le protocole expérimental. Enfin, des propositions pour contrer ces limites sont discutées et des perspectives de recherche futures sont exposées pour améliorer le design de protocoles expérimentaux en cartographie.

Expérimentation

Contexte et objectifs

Dans le cadre d'un projet de recherche, en partenariat avec le propriétaire et gestionnaire du réseau ferroviaire français (SNCF Réseau), nous avons développé une méthode cartographique pour visualiser les récits d'inondations historiques et les dommages causés au système ferroviaire. La carte est incluse dans un « environnement de géovisualisation », qui peut être défini comme une interface de visualisation interactive, permettant aux utilisateurs d'explorer et d'analyser les relations entre toutes les dimensions de l'information géographique, notamment à travers l'usage de fenêtres synchronisées, montrant chacune différents aspects de la donnée (espace, temps, description). Notre interface de géovisualisation intégrait une carte centrale, une ligne de temps située sous la carte, et des fonctions de filtres et de détail de l'information sur demande (fig. 1). Les détails concernant cet outil de géovisualisation, ses fondements scientifiques et les choix de design réalisés sont décrits par Saint-Marc *et al.* (2016).

L'application visait à montrer les chaînes d'événements, également appelées « effets dominos » (Provitolo 2005), survenues durant les inondations passées et qui contribuent bien souvent à aggraver les dommages. A titre d'illustration, durant l'inondation de l'hiver 1947 survenue dans le nord-est de la France, de fortes précipitations, combinées à une fonte des neiges rapide, ont causé le débordement simultané de nombreux cours d'eau descendant du massif des Vosges puis l'inondation d'une vaste zone, conduisant elle-même à de multiples dommages sur le réseau ferroviaire. Les relations de causalité entre les événements sont figurées dans la carte par des lignes pleines reliant un événement cause à sa ou ses conséquences.

Les recherches passées n'ont pas permis d'établir clairement si les visualisations animées étaient vraiment efficaces pour comprendre des phénomènes complexes (Tversky *et al.* 2002), ni si elles étaient adaptées à des experts d'un domaine qui sont novices dans l'usage d'environnements de géovisualisation. Par conséquent, deux versions de l'interface ont été proposées :

- Une version comportant une carte interactive (fig. 1). Tous les événements survenus durant l'inondation, à une échelle locale, sont affichés par des symboles sur la carte. Les lignes de causalité liant un événement à ses causes et conséquences sont affichables sur demande grâce au clic.
- Une version incluant une carte animée et interactive, dite carte *dynamique*. Les événements y apparaissent les uns après les autres au cours d'une animation temporelle. Chaque événement reste affiché jusqu'à la fin de l'animation en périphérie de son point d'apparition, dans un cartouche blanc perpendiculaire à la voie ferrée qu'il a impactée. De cette manière, la dernière scène de l'animation correspond au cas de l'interface interactive. Cette version diverge du cas uniquement interactif en deux autres points : d'une part, les lignes de causalité apparaissent automatiquement au cours de l'animation et, d'autre part, une fonction supplémentaire de contrôles de l'animation est introduite dans l'interface, pour jouer l'animation, la mettre en pause et changer de scène manuellement (avance ou retour dans le temps). A part ces trois points, les deux interfaces étaient identiques (ex : organisation générale de l'interface, contenu de la légende, fonctionnalités,…).

L'expérimentation visait à évaluer l'efficacité de ces deux modalités d'interface auprès d'utilisateurs experts de la SNCF, étant amenés à travailler sur les inondations dans leur mission. L'efficacité est définie comme la capacité à atteindre un objectif tout en dépensant le moins de ressources pour l'atteindre. Le protocole expérimental est décrit dans la section suivante.

Protocole

Vingt-deux experts du système ferroviaire (ingénieurs et docteurs confrontés aux inondations dans leur travail) ont participé à l'expérimentation. Ces experts avaient des niveaux variés en cartographie et étaient en grande majorité novices dans l'usage d'environnements de géovisualisation.

Notre hypothèse initiale était que l'interface dynamique serait plus efficace pour percevoir les relations de causalité entre événements que l'interface interactive. L'expérimentation a suivi un design *within-subject* : chaque participant testait les deux types de cartes. L'ordre de test des cartes était alterné selon les participants, en suivant un plan carré latin, de façon à contrôler un éventuel effet de fatigue.

L'hypothèse a été testée à travers la réalisation par les participants de deux jeux de tâches correspondant à deux objectifs différents : le premier correspondait à des questions de recherche visuelle et le second à des questions de lecture des relations de causalité entre événements survenus pendant une inondation donnée. Ces tâches avaient respectivement pour but de révéler : 1) si les participants parvenaient à répondre grâce à l'interface de géovisualisation et 2) si la méthode utilisée pour figurer les événements et leurs relations de causalité était lisible et adaptée aux utilisateurs. Comme l'interface de géovisualisation proposée pouvait paraître complexe, en particulier pour des utilisateurs novices, chaque jeu de tâche débutait par un didacticiel expliquant les caractéristiques et les fonctionnalités de l'application, puis par une tâche d'entraînement, dont le résultat était commenté par l'interviewer. Ensuite, chaque jeu de tâches consistait d'abord en deux questions d'échauffement, qui n'étaient pas destinées à être incluses dans l'analyse mais constituaient un entraînement supplémentaire, puis en trois « véritables » questions qui étaient monitorées.

Une force des interfaces de géovisualisation est de combiner de multiples vues sur le jeu de données exploré : vues spatiales (carte), vues temporelles (graphiques) et vues thématiques (description textuelle et sémiologie) (Davoine *et al.* 2009). Les questions posées aux participants se sont basées sur ces trois dimensions présentes dans notre interface : des questions sur la symbologie (quoi ?), sur l'espace (où ?), sur le temps (quand ?) et des questions combinant temps et espace. Deux questions sur la sémiologie ont également été intégrées, pour tester la lisibilité des symboles et encourager les participants à explorer la légende. Les questions sont listées dans le tableau 1.

Les questions d'échauffement étaient toujours posées avant les véritables questions mais, à l'intérieur de ces deux catégories, l'ordre des questions était tiré aléatoirement pour chaque participant. Les questions de recherche d'information étaient posées lors de la première série de questions et celles de lecture des causalités étaient posées dans la deuxième série ; cet ordre avait pour but de permettre aux participants de comprendre les informations affichées et de s'habituer à l'interface, avant d'interpréter les relations de causalités, questions qui sont par ailleurs moins habituelles.

Comme cela a été mentionné dans l'introduction, les cartes animées et non-animées ne sont probablement pas équivalentes dans la manière dont elles sont traitées par le cerveau (non-équivalence computationnelle) (Fabrikant *et al.* 2008). Par conséquent, les deux mesures d'efficacité les plus utilisées en cartographie, nommément le taux de bonnes réponses et la durée

Type de tâche	Code	Question	Dimension
Echauffement	TF1	Quelle date correspond à la couleur orange dans la légende ?	Sémiologie
Echauffement	TF4	Quel est le symbole d'un incident affectant un remblai ?	Sémiologie
Echauffement	TF5	A quelle date le premier événement de travaux apparaît-il ?	Temps
Echauffement	TF3	A quel pk [point kilométrique] est survenu l'événement du 21 novembre ?	Espace x Temps
Recherche d'information	T2	Le passage à niveau 264 a-t-il été submergé par le débordement du cours d'eau ?	Espace
Recherche d'information	T1	Combien d'événements du type travaux sont survenus le 14 novembre ?	Temps
Recherche d'information	T3	A quelle date les enrochements du viaduc de l'Orbieu ont-ils été emportés ?	Espace x Temps
Lecture des causalités	T5	Quelle est la conséquence de l'événement qui est survenu au pk 391.760 ?	Espace
Lecture des causalités	T8	Quelle est la date de fin de la pose des nouveaux poteaux caténaires ?	Temps

Tableau 1 : Liste de questions auxquelles les participants devaient répondre au cours de l'expérimentation et dimensions de l'information questionnée

de réponse (Kinkeldey *et al.* 2014, 377), peuvent potentiellement être complétées par d'autres types de mesures, dans le but d'expliquer et d'enrichir les résultats. Des mesures qualitatives peuvent également s'avérer utiles. Au cours de cette expérimentation, nous avons utilisé deux méthodes qualitatives : l'observation directe des actions des participants et l'enregistrement de leurs traces d'utilisation, c'est-à-dire la séquence de clics réalisée dans l'interface. Sur la base des traces d'utilisation, deux autres variables quantitatives ont été analysées pour chaque question :

- la durée avant le premier clic, qui peut indiquer un moment de réflexion ou une hésitation avant de répondre à la question ;
- l'écart entre le nombre de clics réalisé et le nombre minimal de clics nécessaires pour répondre, qui peut informer sur le caractère intuitif de l'interface pour identifier des entités d'intérêt et leurs caractéristiques utiles à la réponse. Le nombre absolu de clics dans l'interface dépend du type de carte utilisé : les contrôles de l'animation génèrent inévitablement plus d'interactions avec la carte dynamique, qu'avec la carte interactive dans laquelle ils ne sont pas présents. C'est pourquoi l'écart au nombre de clics optimal semblaient être un indicateur plus fiable pour analyser les différences entre types de tâches.

Les analyses quantitatives menées à partir des données mesurées ont consisté en des tests statistiques univariés et bivariés, nommément le test du Khi-deux, l'ANOVA et la régression linéaire, avec un risque d'erreur ?? fixé à 5%. Les traces d'utilisation ont été analysées qualitativement (analyse des comportements de réponse) grâce au logiciel Undertracks (Mandran *et al.* 2015) et quantitativement (nombre et types d'interactions).

Résultats et limites

Résultats principaux

Les variables indépendantes évaluées dans l'expérimentation présentée ici étaient donc : la modalité d'interface (interactive ou dynamique), le type de tâche (recherche d'information ou lecture des relations de causalité) et la dimension de l'information interrogée (espace, temps, espace + temps). Les facteurs secondaires, susceptibles d'influencer les résultats et qui ont été contrôlés, étaient : les caractéristiques individuelles des participants (caractéristiques démographiques, familiarité avec le thème et avec les technologies utilisées) et l'ordre des questions (effet de fatigue, effet d'apprentissage). Ces différentes variables mesurées ont donné lieu à des analyses croisées présentées dans cette section.

Effet du type de tâche

Un premier résultat est que le taux moyen de bonnes réponses est plus élevé pour les tâches de recherche d'information (86%) que pour les tâches de lecture des causalités (62%). Cette différence est statistiquement significative (intensité de relation moyenne : 28%). La durée moyenne de réponse est plus courte pour les tâches de recherche d'information (104") que pour les tâches de lecture de causalités (79") mais cette différence n'est pas significative. La durée de réponse est légèrement, mais pas significativement, liée à la validité de la réponse donnée (intensité de relation faible : 4,7%) : les participants mettent en moyenne plus de temps à répondre lorsque leur réponse finale est fausse. La carte interactive est plus efficace que la carte dynamique (taux de bonnes réponses plus élevé, durée de réponse plus faible), quel que soit l'objectif de la tâche (fig. 2), mais les différences de résultats entre les deux cartes ne sont pas statistiquement significatives.

Ainsi, l'application de géovisualisation testée semble adaptée aux utilisateurs experts du domaine pour retrouver des informations, mais la manière d'afficher les relations de causalités pourrait être améliorée. Les relations de causalités semblent être plus faciles à lire lorsque les utilisateurs peuvent les afficher sur demande (cas interactif) que lorsqu'elles apparaissent pendant l'animation (cas dynamique).

Effet de la dimension de l'information interrogée

Le taux de bonnes réponses est le plus faible pour les questions portant à la fois sur l'espace et le temps (70%), suivi par les questions portant uniquement sur le temps (84%) puis par les questions sur l'espace ou la sémiologie qui ont le meilleur score (plus de 95%) (fig. 3). Cette différence est statistiquement significative (intensité de relation moyenne : 31%). La durée de réponse est également significativement plus longue pour les questions portant à la fois sur l'espace et le temps (intensité de relation faible : 25%). Elle est ensuite plus longue pour les questions portant sur l'espace (83" en moyenne) que pour les questions portant sur le temps uniquement (67") ou la sémiologie (29") (fig. 4).

La durée avant le premier clic suit la même tendance : elle est plus élevée pour les questions portant uniquement sur l'espace (5,5" à 7,5" de plus comparées aux deux autres types). Cette différence n'est significative que lors de l'usage de la carte dynamique (intensité de relation faible : 18%). Cette durée plus longue avant le premier clic pour les questions portant sur l'espace peut

s'expliquer par le fait que ces questions nécessitent un temps de recherche visuelle dans la carte pour trouver la localisation concernée par la question, avant d'interagir pour chercher la réponse. La carte dynamique conduit à un taux de bonnes réponses plus faible et une durée de réponse plus élevée que la carte interactive quelle que soit la dimension interrogée, mais ces écarts ne sont pas significatifs.

Sur la base de ces résultats, il est possible de conclure que les questions portant à la fois sur le temps et l'espace sont plus difficiles pour les utilisateurs que les autres questions. Cela n'est pas surprenant car ces tâches interrogent deux dimensions de l'information au lieu d'une seule ; elles combinent donc probablement le temps de réponse plus long propre aux questions portant sur l'espace et le taux de bonnes réponses plus faible propre aux questions sur le temps.

L'analyse a montré que la dimension de l'information interrogée n'était pas le seul facteur expliquant les différences de résultats. Par exemple, les participants ont eu des résultats significativement plus mauvais à la question T7 qu'aux autres tâches du même type (fig. 5). Ce résultat a été exploré plus avant et est expliqué dans la section 3.2.

Effets de fatigue et d'apprentissage

Une diminution de la performance des utilisateurs au cours du test, qui peut traduire un effet de fatigue, a été observée : la durée de réponse augmente au cours de chaque jeu de tâches mais cette augmentation n'est statistiquement significative que dans la deuxième série de tâches sur la lecture des causalités, l'intensité de cette relation étant plus élevée avec la carte dynamique (28%) qu'avec la carte interactive (11,6%). Dans la première série de questions, c'est-à-dire les tâches de recherche d'information, le taux de bonnes réponses diminue légèrement au cours du test (fig. 6). Durant les tâches de lecture de causalité, le taux de bonnes réponses diminue avec la carte dynamique, tandis qu'il reste à peu près constant avec la carte interactive.

Les résultats n'ont pas montré d'effet d'apprentissage significatif : la performance des participants n'augmentait pas durant les séries de questions, pas plus qu'elle ne mettait en évidence que les participants s'habituaient à l'interface au cours de leurs réponses. Cet effet était en fait intentionnellement évité par la mise en place d'une session de formation à l'usage de la carte avant de répondre aux questions via un didacticiel, puis une question d'entraînement et deux questions d'échauffement. Le seul effet d'apprentissage observé l'a été dans la seconde série de questions, dans le cas où la carte dynamique était testée avant la carte interactive (voir fig. 6 à gauche, les tâches positionnées de 8 à 10). Ce résultat laissait penser qu'il existait une influence de l'ordre de test des cartes, ce qui a été exploré plus avant.

Lorsque la carte dynamique est utilisée en première position, la durée avant le premier clic est significativement plus longue que lorsque la carte interactive est utilisée en premier, et cela dans les deux séries de tâches (fig. 7). Si la durée avant le premier clic est interprétée comme un manque de confiance ou un besoin de réfléchir avant d'agir, alors cela signifierait qu'utiliser la carte dynamique en première modalité rendrait les participants moins à l'aise avec l'interface de géovisualisation. Cette hypothèse semble être renforcée par les résultats sur la confiance des participants en leur réponse, qui était évaluée à la fin de chaque série de questions : les participants utilisant la carte dynamique en premier et la carte interactive en second voient leur confiance en leur réponse augmenter (de 36% à 50% de participants « très confiants »), tandis que les participants utilisant d'abord la carte interactive puis la carte dynamique voient leur confiance diminuer (de 63% à 18% de participants « très confiants »).

L'écart au nombre de clics optimal suit une tendance inverse : quand la carte dynamique est utilisée en premier, l'écart au nombre de clics optimal est plus faible, quelle que soit la série de questions, que lorsque la carte interactive est utilisée en premier. Cela signifie que l'usage de la carte interactive en première modalité favorise un plus grand nombre d'interactions avec l'interface.

En suivant notre hypothèse explicative, cela peut signifier qu'un nombre de clics plus élevé lors de l'usage de la carte dynamique en première modalité indiquerait que l'utilisateur se sent plus confiant dans l'interface. Cependant, cette interprétation de l'écart au nombre de clics optimal n'est pas totalement cohérente avec un autre résultat, montrant qu'il existe une corrélation positive entre un écart au nombre de clics optimal élevé et un taux de bonnes réponses faible (intensité de relation très faible : 6%). L'interprétation de l'écart au nombre de clics optimal devrait donc faire l'objet d'investigations complémentaires dans d'autres tests.

Limites de l'expérimentation

L'analyse des résultats a conduit à identifier deux limites qui remettaient en question certains aspects du protocole expérimental. Elles sont détaillées ci-dessous.

Questions imprécises

Il n'était pas possible de répondre de façon précise à deux questions, T2 (portant sur l'espace) et T8 (portant

sur le temps), en utilisant les informations disponibles dans l'interface de géovisualisation. Cela était dû à une méconnaissance du domaine thématique par le cartographe ayant élaboré le protocole expérimental. La formulation de ces deux questions présentait une ambiguïté et était donc sujette à interprétation par les experts. Il était donc impossible de déterminer si les réponses des participants étaient correctes ou non car cela dépendait de leur compréhension de la question. Cette imprécision n'avait pas été remarquée malgré un pré-test du protocole mené auprès de quelques experts. Elle a été découverte grâce à l'observation directe des participants durant le test.

Face aux questions imprécises, les participants modifiaient leur comportement de réponse : une exploration significativement plus longue des données (par exemple, ils cliquaient sur tous les événements près de la localisation où la réponse était susceptible de se trouver) et/ou une réponse aléatoire (par exemple « oui » ou « non » selon le participant, pour la question T2).

Les questions T2 et T8 ont donc été exclues des résultats présentés dans la section 3.1 et, pour augmenter le nombre de questions analysées, les tâches d'échauffement ont été intégrées dans l'analyse (mise à part, évidemment, dans l'analyse de l'effet du type de tâche).

Impact de la complexité cognitive des questions

Les résultats n'étaient pas corrélés uniquement à la dimension de l'information interrogée dans les questions. Nous avons déterminé qu'ils étaient également très corrélés à leur complexité cognitive. Dans le champ des sciences de l'apprentissage, Anderson and Krathwohl (2001) ont proposé six types de processus cognitifs impliqués durant la résolution d'exercices et demandant des efforts croissants : *mémoriser*, *comprendre*, *appliquer*, *analyser*, *évaluer* et *créer*. Le tableau 2 décrit ces concepts.

Dimension	**Définition**
Mémoriser	Retrouver, reconnaître et se remémorer les connaissances pertinentes à partir de la mémoire à long-terme
Comprendre	Construire du sens à partir de messages oraux, écrits ou graphiques, en interprétant, exemplifiant, classifiant, résumant, inférant, comparant et expliquant
Appliquer	Conduire ou utiliser une procédure en exécutant ou implémentant
Analyser	Découper le matériel en parties constituantes, déterminer comment les parties sont reliées les unes aux autres et avec une structure ou un objectif global, en différenciant, organisant et attribuant
Evaluer	Emettre des jugements basés sur des critères et des standards en vérifiant et critiquant
Créer	Rassembler des éléments pour former un tout cohérent et fonctionnel ; réorganiser les éléments en un nouveau motif ou une nouvelle structure, en générant, planifiant ou produisant

Tableau 2 : Taxonomie de la complexité cognitive des questions (Anderson and Krathwohl 2001; éditée par Wu et al. 2012)

Nous n'avons découvert ces travaux qu'après la fin de l'expérimentation. Nous n'avons donc pu étudier qu'à posteriori les différences de complexité cognitive entre les questions de notre test (tableau 3). Nous avons remarqué en particulier que la question T7, qui a conduit à des résultats significativement inférieurs aux autres (fig. 5), était celle avec le niveau de complexité le plus élevé dans notre échantillon (« analyse »). La prise en compte de ce facteur dans notre étude a permis d'obtenir de nouveaux résultats.

Les résultats des participants sont significativement plus mauvais pour les questions de complexité « analyse » (taux de bonnes réponses 60% plus faible que les questions de complexité « comprendre », intensité de relation forte : 56%) et leur durée de réponse est significativement plus longue (42" plus longue que pour les questions de complexité « compréhension », intensité de relation moyenne : 31%). Les résultats des participants ne sont pas significativement différents entre les niveaux de complexité « compréhension » et « mémorisation ». L'écart au nombre de clics optimal est significativement plus élevé pour les questions « analyse » seulement dans le cas où la carte dynamique est utilisée (fig. 8). Sur la base des résultats précédents, la carte dynamique semble donc plus complexe à utiliser que la carte interactive. On peut poser l'hypothèse que cette carte provoque une charge cognitive extrinsèque plus élevée pour les utilisateurs, ce qui, combiné à des questions avec une charge cognitive intrinsèque élevée telles que les questions « analyse », peut entraîner une surcharge cognitive (Khalil *et al.* 2005).

Question	Complexité cognitive	Dimension
T1	Compréhension	Temps
T2	Mémorisation	Espace
T3	Compréhension	Espace x Temps
T5	Compréhension	Espace
T7	Analyse	Espace x Temps
T8	Compréhension	Temps
TF1	Mémorisation	Sémiologie
TF3	Compréhension	Espace x Temps
TF4	Mémorisation	Sémiologie
TF5	Compréhension	Temps

Tableau 3 : Complexité cognitive des questions dans notre expérimentation (les questions imprécises ont été exclues de l'analyse finale)

Les dimensions de l'information interrogées dans les questions ne sont pas statistiquement indépendantes de la complexité cognitive des questions, ce dernier paramètre n'ayant pas été contrôlé lors de la conception du protocole. Nous avons donc conduit une nouvelle analyse de la dimension de l'information, pour un niveau de complexité cognitive homogène fixé sur « compréhension ». En considérant un niveau de complexité cognitive égal, les dimensions de l'information interrogées n'ont pas d'impact significatif sur le taux de bonnes réponses. En revanche, comme auparavant, elles ont bien un impact sur la durée de réponse (intensité de relation faible : 7%) : les questions impliquant l'espace ou à la fois l'espace et le temps correspondent à un temps de réponse plus élevé (respectivement 84" et 99") que les questions portant seulement sur le temps (67") (fig. 9, gauche). La durée avant le premier clic est également légèrement plus élevée pour les questions impliquant la dimension spatiale. Avec la carte interactive, l'écart au nombre de clics optimal est plus faible pour les questions sur l'espace et plus élevé pour les questions sur le temps (fig. 9, droite).

Le temps de réponse n'étant pas significativement différent entre les deux types de cartes, il semble que la carte interactive requière moins d'efforts pour lire l'espace (moins de clics) mais plus d'effort pour lire le temps, alors que la tendance inverse est observée pour la carte dynamique (plus d'effort pour lire la dimension spatiale).

A cause de la dépendance entre les variables étudiées et du faible nombre de questions réparti dans chaque catégorie de tâches, les résultats de cette expérimentation ne peuvent que constituer des pistes, qui nécessiteront d'être approfondies dans des expériences futures. La conclusion principale de ce travail est qu'il est important de mieux identifier les variables indépendantes liées aux questions, car elles ont un impact significatif sur les résultats. Cela interroge également la reproductibilité et la comparabilité des résultats expérimentaux en cartographie, par exemple dans les cas où la complexité cognitive des questions n'a pas été caractérisée. Des solutions existent dans d'autres champs d'études, qui pourraient constituer des sources d'inspiration pour répondre à ces problématiques.

Conclusion et perspectives

La comparaison des deux modalités de cartes a montré que la carte dynamique est plus difficile à utiliser pour des experts du domaine. Elle génère une plus grande fatigue, moins d'interactions avec l'interface et moins de confiance des participants dans leurs réponses. Cependant, il semble qu'elle nécessite moins d'efforts pour lire la dimension temporelle de l'information (moins grand nombre de clics) mais plus d'effort pour lire la dimension spatiale que la carte interactive.

Cette expérimentation, comme un certain nombre dans le passé, a révélé la difficulté de construire un protocole expérimental robuste dans lequel toutes les variables indépendantes sont contrôlées. La représentation cartographique ou l'interface de géovisualisation ne sont pas les seuls facteurs à avoir une influence sur les résultats : les caractéristiques des tâches à réaliser durant le test jouent aussi un grand rôle.

Les résultats ont aussi confirmé que, même si le taux de bonnes réponses et la durée de réponse sont probablement nécessaires pour analyser l'efficacité des cartes, ils peuvent être enrichis par d'autres variables qui apportent des conclusions supplémentaires ou plus détaillées sur les résultats. En particulier, les observations qualitatives, telles que l'observation des participants en cours de test, l'enregistrement des traces d'utilisation ou les mesures oculométriques ont démontré leur utilité pour expliquer les résultats quantitatifs bruts (Çöltekin *et al.* 2010 ; Fabrikant *et al.* 2008 ; Ooms and De Maeyer 2015).

Une solution idéale pour éviter les erreurs dans la conception de protocoles serait de réutiliser les protocoles passés. Cependant, les protocoles et les tâches testées sont souvent très spécifiques à un domaine d'application et il est difficile de savoir dans quelle mesure ils peuvent être généralisés à d'autres domaines (Kinkeldey *et al.* 2014, 375).

Une autre solution serait de créer des jeux de données de test ouverts, dans une diversité de champs thématiques et de zones géographiques, et des jeux de tâches généralisables associés, sur lesquels les évaluations futures pourraient s'aligner. De tels jeux de données et jeux de tâches sont déjà utilisés pour maximiser la comparaison des résultats entre études dans le domaine de la recherche d'informations (RI), qui étudie notamment l'usage et les interactions avec les moteurs de recherche. Les jeux de données et de tâches standardisés mis en œuvre couvrent plusieurs domaines d'application (santé, commerce, loisir, etc.) et tous les niveaux de complexité cognitive (Wu *et al.* 2012). Ils sont ensuite utilisés pour comparer l'efficacité des moteurs de recherche et les comportements d'utilisateurs en conditions contrôlées (ex : Moffat *et al.* 2013).

La façon de transposer cette pratique dans la cartographie devra être défini et exploré dans des recherches futures. Les jeux de tâches pourraient aussi adresser une diversité d'autres facteurs susceptibles d'influencer le comportement des utilisateurs et leurs réponses, comme la dimension de l'information questionnée, que nous avons étudiée ici, ou les classes d'analyses de Bertin (élémentaire, intermédiaire, global) (Bianchin 2012). Le partage de jeux de données ouverts et de questions d'évaluation adaptées à la lecture et à l'analyse de cartes pourrait promouvoir une meilleure comparabilité entre les études utilisateurs cartographiques et pourrait enfin permettre de déterminer si les cartes animées ont réellement un intérêt cognitif pour lire et comprendre les processus spatio-temporels.

Remerciements

Nous remercions les ingénieurs de SNCF Réseau pour leur temps et leur implication dans cette recherche. Un grand merci également à Nadine Mandran du LIG, pour ses idées et son soutien durant la conception de cette expérimentation et l'analyse des résultats.

Bibliographie

Anderson, L. W., & Krathwohl, D. A. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York (USA): Longman.

Bianchin, A. (2012). "Actualité de l'approche de Jacques Bertin dans l'ensignement de la cartographie". *CFC*, *212*. http://www.lecfc.fr/new/articles/212-article-2.pdf

Çöltekin, A., Fabrikant, S. I., & Lacayo, M. (2010). "Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings". *International Journal of Geographical Information Science*, *24*(10), 1559–1575. doi:10.1080/13658816.2010.511718

Davoine, P.-A., Moisuc, B., Gensel, J., & Arnaud, A. (2009). GenGHIS : environnement pour le développement d'applications de géovisualisation à données géo-référencées multidimensionnelles. In *Festival International de Géographie*. Saint-Dié-des-Vosges.

Fabrikant, S. I., Rebich-Hespanha, S., Andrienko, N., *et al.* (2008). "Novel Method to Measure Inference Affordance in Static Small-Multiple Map Displays Representing Dynamic Processes". *The Cartographic Journal*, *45*(3), 201–215. doi:10.1179/000870408X311396

Griffin, A. L., Maceachren, A. M., Hardisty, F., *et al.(*2006). "A Comparison of Animated Maps with Static Small-Multiple Maps for Visually Identifying Space-Time Clusters". *Annals of the Association of American Geographers*, *96*(4), 740–753.

Khalil, M. K., Paas, F. G. W. C., Johnson, T. E., & Payer, A. F. (2005). "Design of interactive and dynamic anatomical visualizations: The implication of cognitive load theory". *The Anatomical Record (Part B: New Anatomist)*, *286*(2), 15–20. doi:10.1002/ar.b.20078

Kinkeldey, C., Maceachren, A. M., & Schiewe, J. (2014). "How to Assess Visual Communication of Uncertainty ? A Systematic Review of Geospatial Uncertainty Visualisation User Studies". *The Cartographic Journal*, *51*(4), 372–386. doi:10.1179/1743277414Y.0000000099

Mandran, N., Ortega, M., Luengo, V., & Bouhineau, D. (2015). "DOP8_Cycle: Merging both data and analysis operators life cycles for Technology Enhanced Learning". In *LAK '15*. Poughkeepsie (NY, USA). http://projet-undertracks.imag.fr/wp-content/uploads/2014/04/LAK15_submission15_mandran.pdf

Moffat, A., Thomas, P., & Scholer, F. (2013). "Users Versus Models : What Observation Tells Us About Effectiveness Metrics". In *CIKM '13*. San Francisco (USA): ACM. doi:10.1145/2505515.2507665

Ooms, K., & De Maeyer, P. (2015). "Georeferencing eye tracking data on interactive cartographic products". In *27th International Cartographic Conference (ICC 2015)*. Rio de Janeiro (Brazil).

Provitolo, D. (2005). "Un exemple d'effets de dominos : la panique dans les catastrophes urbaines". *Cybergeo : European Journal of Geography*, 19. doi:10.4000/cybergeo.2991

Saint-Marc, C., Davoine, P.-A., Villanova-Oliver, M. *et al.*. (2016). "Géovisualisation de récits de catastrophes naturelles et de leurs impacts : Application aux inondations ayant impacté le système ferroviaire". In *Sageo 2016*. Nice (France).

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). "Animation : can it facilitate ?" *International journal of human-computer studies*, 247–262. doi:10.1006/ijhc.1017

Wu, W., Kelly, D., Edwards, A., & Arguello, J. (2012). "Grannies , tanning beds , tattoos and NASCAR : Evaluation of search tasks with varying levels of cognitive complexity". In *Proceedings of the 4th Information Interaction in Context Symposium (IIIX'12)* (pp. 254–257). Nijmegen (The Netherlands): ACM. doi:10.1145/2362724.2362768