# Analyser des tweets géolocalisés pour explorer les réponses sociales face aux phénomènes météorologiques extrêmes

Camille Cavalière, Paule-Annick Davoine, Céline Lutoff, Isabelle Ruin

# Analyser des tweets géolocalisés pour explorer les réponses sociales face aux phénomènes météorologiques extrêmes

## *Réflexions épistémologiques et verrous méthodologiques*

**Camille Cavalière[1], Paule-Annick Davoine[1], Céline Lutoff [2], Isabelle Ruin[3]**

*1. Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes*
*700 Avenue Centrale, F-38401 Saint-Martin-d'Hères, France*
*{Camille.Cavaliere,Paule-Annick.Davoine}@imag.fr*

*2. Laboratoire PACTE-Territoires, Université Grenoble Alpes*
*14 bis, Avenue Marie Reynoard, F-38100 Grenoble, France*
*Celine.Lutoff@univ-grenoble-alpes.fr*

*3. Laboratoire LTHE, Université Grenoble Alpes*
*CS 40700, F-38058 Grenoble, France*
*Isabelle.Ruin@univ-grenoble-alpes.fr*

*RESUME. En automne, de violentes perturbations météorologiques frappent régulièrement le sud-est de la France ; pendant de tels événements, habitants et autorités locales diffusent de l'information via les médias sociaux. Cette information constitue une source de données gratuite et utile pour l'analyse sociale de l'événement, de la manière dont les personnes l'ont vécu. Cet article repose sur l'étude des tweets géolocalisés émis dans huit départements à l'automne 2014 : nous proposons de traiter l'information selon une posture de recherche originale : le raisonnement abductif.*

*ABSTRACT. In fall, the south-east of France is regularly affected by heavy weather disturbances. During those natural disasters, citizens and authorities turn to social media platforms to share information. They can therefore provide free and relevant in-field information. This paper aims at exploring georeferenced tweets created within eight departments in the fall 2014: we intend to process the twitter data using an original research approach, called abductive reasoning.*

*MOTS-CLES : MEDIAS SOCIAUX · TWITTER · PHENOMENES EXTREMES · ABDUCTION*
*KEYWORDS: SOCIAL MEDIA · TWITTER · EXTREME PHENOMENA · ABDUCTION*

## 1. Introduction

Au cours de la dernière décennie, les géographes ont assisté à la transformation de la nature, de la disponibilité et du volume de l'information géographique. La rencontre entre les systèmes de géolocalisation, dont l'usage s'est largement répandu grâce aux *GPS* et aux *Smartphones*, et le *Web 2.0*, constitue le facteur décisif ayant favorisé l'explosion de la disponibilité de données géographiques publiques (Miller, 2007). L'environnement de travail du chercheur est désormais constitué d'un monde extrêmement riche de données (Kitchin, 2013). Twitter est une plateforme du *Web 2.0* via laquelle les utilisateurs, nommés *tweeters*, créent et échangent quotidiennement de l'information relative à leurs activités ou opinions, tout en ayant la possibilité d'ajouter une information géographique, sous forme de coordonnées ou de *hashtags* spatiaux (villes, quartiers, etc.). Créé en 2006, Twitter a bénéficié d'un succès rapide et, en 2015, la plateforme enregistre 304 millions d'utilisateurs mensuels actifs pour un effectif quotidien de tweets émis estimé à 500 millions[1]. Le service fournit ainsi une source intarissable de données numériques centrées sur l'individu, ses opinions et ses interactions avec son environnement.

Ces dernières années, les travaux fondés sur l'étude de tweets géolocalisés ont été entrepris dans des thématiques variées : l'étude de la répartition spatiale des *tweeters* et la définition de profils sociaux (Li *et al.*, 2013), la mobilité internationale (Hawelka *et al.*, 2014), les relations entre individus, temps et espace (Andrienko *et al.*, 2013), ou encore l'étude de la dynamique et du cycle de vie des *hashtags* (Lin *et al.*, 2013). L'objectif de ces études consiste à identifier des structures régulières et répétitives dans le temps et dans l'espace (Erwig, 2004), les changements qui peuvent survenir, ou encore des comportements de *tweeting* qui révèlent les pratiques et habitudes individuelles ou collectives des *tweeters* (Andrienko *et al.*, 2013). D'autres auteurs se sont intéressés à l'exploration des réponses sociales lors de la survenue de phénomènes extrêmes, sociaux ou environnementaux : les réactions du public face à des événements exceptionnels qui affectent des milliers d'individus à l'échelle d'un pays (Lee Hughes et Palen, 2009), les réactions du public face à la criminalité (Kounadi *et al.*, 2015) ou encore les risques naturels (Dashti *et al.*, 2014). Les travaux effectués dans ce domaine ont démontré la pertinence de ces données particulières comme source d'information spatio-temporelle pour la détection et la caractérisation d'événements naturels en temps réel (De Longeville *et al.*, 2009 ; Dittrich et Lucas, 2014).

Pour autant, l'observation des réponses sociales aux phénomènes extrêmes, qu'il s'agisse de catastrophes naturelles ou d'attentats, soulève l'interrogation de nos capacités à appréhender les caractéristiques de la donnée tweet afin d'identifier les obstacles qui contraignent l'analyse et l'interprétation des données. En effet, si les tweets peuvent constituer une source de données utile à une nouvelle et meilleure compréhension des phénomènes sociaux et des comportements des individus en période de crise, leur analyse requiert une réflexion approfondie sur la nature des données, les problèmes méthodologiques et l'approche scientifique à adopter : en

[1] http://www.blogdumoderateur.com/chiffres-twitter/

effet, un certain nombre de paramètres peuvent biaiser la pertinence de la donnée comme l'état psychologique de la personne qui crée l'information au moment où elle vit l'événement, l'incomplétude ou l'imprécision de l'information produite ou encore la distance entre l'individu qui crée la donnée et l'événement (Hertfort *et al.*, 2014). Dans un second temps, la donnée tweet appartient à ce que l'on désigne aujourd'hui sous le nom de *Big Data*, en référence aux données qui dépassent nos capacités et méthodes d'analyse traditionnelles (Miller et Goodchild, 2015). Il ne s'agit plus de collecter des données pour répondre à une question préalablement identifiée, mais de collecter d'abord pour identifier ensuite les questions auxquelles les données peuvent répondre (Miller et Goodchild, 2015). Cette posture se traduit par la primauté de l'observation, suivie par la formulation d'hypothèses.

Dans cet article, nous nous intéressons aux tweets géolocalisés produits dans le contexte des précipitations extrêmes et plus particulièrement des pluies et inondations cévenoles. Nous proposons ainsi une méthodologie originale d'approche et de traitement des données, destinée à favoriser la découverte de phénomènes par le raisonnement abductif qui s'inscrit dans la posture précédemment évoquée. Nous menons cette recherche dans le but d'estimer la pertinence du tweet comme indicateur de survenue d'une crise et de détecter les problèmes sous-jacents à son utilisation. Notre méthodologie est appliquée à l'événementiel hydro-météorologique - les épisodes cévenols - qui a affecté huit départements du sud-est de la France entre octobre et décembre 2014. Ces événements, nommés crues rapides, se caractérisent par des orages convectifs qui provoquent de courts épisodes de pluies torrentielles entraînant un ruissellement intense au sein de tous les bassins versants et l'élévation subite du niveau des rivières.

L'article est organisé de la manière suivante : la partie 2 présente les challenges associés à l'exploitation de telles données, l'ouverture que propose l'approche abductive et l'application que nous soumettons dans le cadre de cette recherche. La partie 3 présente les étapes successives de l'analyse des données. Finalement, la partie 4 conclut par une évaluation globale de la recherche et propose de prochains travaux envisageables.

## 2. L'abduction comme méthode pour l'exploration de la donnée tweet

Cette partie introduit les problèmes posés par la donnée tweet dans le cas particulier des phénomènes extrêmes et décrit la posture de recherche adoptée vis-à-vis des données. Elle présente les réflexions préalables à leur traitement, illustrées par une démarche destinée à favoriser la découverte de motifs intéressants.

### *2.1. Regard sur la donnée tweet dans un contexte événementiel*

Les tweets géolocalisés sont considérés comme des capteurs sociaux qui détectent et enregistrent les perturbations survenues (De Longeville *et al.*, 2009). La communauté scientifique a montré un intérêt croissant, notamment au travers de la *Volunteered Geographic Information* (VGI) et du *crowdsourcing*, pour les données

produites par des individus équipés d'appareils géolocalisables, qui communiquent en cas de survenue d'événement exceptionnel (Dittrich et Lucas, 2014). Néanmoins, contrairement à la VGI, les tweets ne sont pas forcément créés dans l'intention de diffuser une information ultérieurement utilisée par un tiers. En d'autres termes et dans la plupart des cas, les tweets ne sont pas créés dans un objectif de satisfaction d'exigences scientifiques. Par conséquent, l'information contenue est potentiellement non structurée, confuse, voire vague par rapport à notre objet d'étude. En outre, ce type de donnée est issu d'une réalité perçue par l'utilisateur : le tweet traduit le regard de l'individu vis-à-vis de l'événement : il révèle ainsi une représentation, qui peut varier sémantiquement (qualité du discours) mais également en fonction de l'expérience vécue de l'individu (l'événement est-il nouveau ou déjà connu ?) et de son degré d'objectivité (exagère-t-il un fait observé ?).

Enfin, l'existence du tweet de crise géolocalisé dépend d'une multitude de facteurs : il est le produit d'un contexte environnemental précis qui interagit avec la sphère sociale, et est donc lié à la réactivité et à l'implication des individus. Ces tweets résultent d'une motivation instantanée induite par la participation du *tweeter* à un événement inhabituel qui rompt avec la routine de sa sphère quotidienne. Cela implique que l'information envoyée concerne l'utilisateur dans sa sphère environnementale et sociale directe. En revanche, nous ne pouvons pas affirmer que la donnée tweet capture l'événement dans sa globalité : le *tweeting* ne concerne que certaines catégories de la population, notamment en fonction de l'âge[2] ; toute la population qui *tweete* ne dispose pas de la même qualité de réseau et nous supposons que l'implication des *tweeters* dans la création d'information utile est le résultat de facteurs imbriqués comme l'expérience vécue, précédemment évoquée, ou encore la situation et le comportement du *tweeter* à l'instant *t*.

### *2.2. La place des tweets dans une nouvelle démarche d'analyse abductive*

Les travaux précédents (cf. introduction) ont révélé la complexité du contenu des tweets et la nécessité de considérer toutes les facettes de cette information. La donnée tweet présente en effet trois dimensions susceptibles d'introduire une part conséquente d'hétérogénéité : elle est spatiale (De Longeville *et al.*, 2009), mais de précision variable (*hashtag*, coordonnées GPS). Elle est temporelle et peut donc nous permettre de suivre les séquences d'un événement (De Longeville *et al.*, 2009). Enfin, elle est sémantique (Dashti *et al.*, 2014) mais variable dans sa forme (texte, *hashtags*, images, URL) et en fonction des deux premiers paramètres : le contenu des tweets varie donc en fonction du temps et de l'espace.

Ces propriétés ont incité les chercheurs à soulever l'épineuse question des capacités à traiter et à exploiter ces données (Dashti *et al.*, 2014). Ainsi, l'évolution récente des données, qu'elles soient produites par des individus ou par des machines,

---

[2] En 2015, le nombre d'utilisateurs mensuels actifs en France est estimé à 2,3 millions. Source : http://www.blogdumoderateur.com/chiffres-twitter/

a favorisé l'émergence d'un nouveau paradigme, celui des sciences guidées par les données, ou *data-driven sciences* (Miller et Goodchild, 2015). La perspective d'une nouvelle compréhension des phénomènes géographiques se heurte à des masses de données spatio-temporelles et sémantiques hétérogènes, qu'il est difficile d'explorer de manière approfondie. Ainsi, les chercheurs peuvent difficilement entrevoir la richesse que les données peuvent potentiellement receler (Anderson, 2008). Cette richesse, il faut la découvrir : or, Kitchin (2013) note que les méthodes de traitement et d'analyse de données ont peu évolué depuis les années 1990. Ce même auteur affirme même que les *new forms of data science are in their infancy*. C'est pourquoi cette évolution s'est plutôt traduite, dans les sciences humaines, par un questionnement épistémologique sur les possibilités de donner du sens aux données en prenant en compte leurs dimensions géographique et sociale (Kitchin, 2013). Les prémices des *data-driven sciences* ont ainsi préconisé un rejet systématique de l'inférence déductive afin de privilégier l'attention portée à l'émergence et à la détection de motifs inattendus, à la formulation d'hypothèses *a posteriori* (Kell et Oliver, 2003) et à leur test.

Le raisonnement abductif a ainsi été récemment introduit comme inférence clé dans l'exploration des données et la découverte de phénomènes intéressants cachés dans des masses de données (Miller et Goodchild, 2015). Il s'agit d'une démarche explicative fondée sur la perception et la juxtaposition de faits observés pour remonter aux causes qui les produisent (Walters, 2012) et destinée à prévoir le comportement d'un système (Hoffmann, 1999). Elle fait donc référence à la capacité de générer des hypothèses face à un résultat inattendu (Hoffmann, 1999). Cette nouvelle approche a pour but de construire progressivement une connaissance ancrée sur les données (Kitchin, 2013). Nous proposons ici de voir de quelle manière elle nous permet d'étudier les questionnements qui émergent concernant les réponses sociales aux phénomènes naturels extrêmes.

### *2.3. Description de la démarche de recherche adoptée*

La figure 1 illustre la démarche théorique de l'exploration des données que nous souhaitons tester. Nous associons les inférences déductive et abductive. A partir d'un constat établi dans l'état de l'art, nous formulons l'hypothèse suivante (Phase 1) : s'il existe une étroite correspondance entre densités de populations et densités de tweets (Lin *et al.*, 2013), nous supposons que les tweets liés à un phénomène extrême échappent à cette logique et peuvent ainsi constituer le marqueur social de la localisation, voire de l'intensité de ce phénomène, quelle que soit sa nature. Nous focalisons donc notre attention sur les anomalies de *tweeting*, c'est-à-dire les espaces où l'activité de *tweeting* est plus importante que la population, et pensons qu'ils sont davantage présents dans les aires urbaines. Nous faisons néanmoins le pari que les résultats permettront de détecter des motifs non conformes à ces attentes (B).

La phase 2 correspond à l'inférence abductive : l'observation intéressante détectée à la fin de l'étape précédente fait l'objet d'une nouvelle hypothèse (A) fondée sur l'information apportée par la donnée tweet. Cette hypothèse est à son tour soumise à l'expérience. Enfin, si l'existence d'un lien entre (A) et (B) semble se

confirmer, il convient, en phase 3, d'explorer la validité de ce lien pour confirmer ou nuancer l'exactitude de l'hypothèse (A).

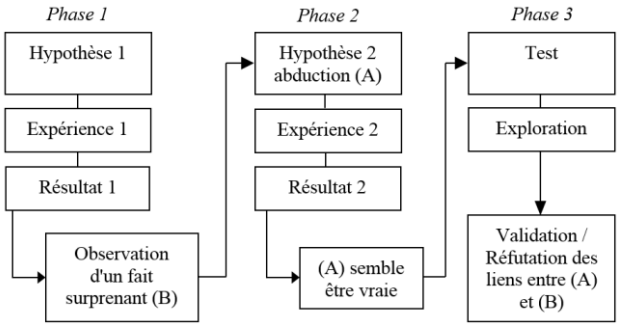



*Figure 1. Démarche théorique de construction de la connaissance*

## 3. Exploration et analyse des jeux de données

Nous avons travaillé sur un jeu de données brut, constitué de 597 741 tweets géolocalisés émis entre le 8 octobre et le 12 décembre 2014 dans les départements suivants, qui sont fréquemment affectés par les épisodes cévenols : Drôme, Ardèche, Gard, Lozère, Hérault, Bouches-du-Rhône, Var et Vaucluse. Ils ont été collectés par un serveur connecté à l'*API Streaming* de Twitter. Ce jeu constitue notre base pour l'identification de l'information de crise et l'analyse de données.

### *3.1. Construction d'un jeu de tweets de crise*

Les tweets de crise, qui mentionnent une perturbation cévenole, peuvent être sélectionnés en fonction de leur contenu (Hertfort *et al.*, 2014). La sélection de l'information de crise est alors fondée sur la recherche de mots-clés à travers l'élaboration d'un glossaire. Nous ne nous focalisons pas en particulier sur les tweets qui contiennent des *hashtags* (comme par exemple *#inondation* ou *#intempéries*) : nous craignons en effet qu'il s'agisse d'un critère de sélection trop rigoureux qui n'extrairait qu'une infime partie du jeu de tweets de crise potentiel.

#### *3.1.1. Construction du glossaire d'analyse*

Le jeu de tweets de crise est construit en respectant deux contraintes : il doit être le plus complet possible en termes d'effectif et le plus précis possible sur le plan de la qualité de l'information, et de la minimisation du bruit résiduel.

Dans un premier temps, la sélection de mots-clés requiert une étape de *brainstorming*, pendant laquelle nous essayons d'inventorier un maximum de vocabulaire (ou associations lexicales) diversifié sur l'événement considéré. Dans le cas étudié, ces mots-clés ont été recherchés de la manière suivante :

- le vocabulaire événementiel contextuel : *orage, éclair, tonnerre, foudre, pluie, inondation* ;

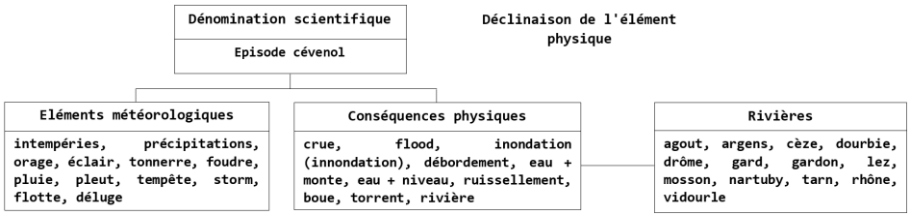
- le vocabulaire officiel ou scientifique qui décrit les phénomènes : *épisode cévenol, intempéries, précipitations*, ou encore le vocabulaire officiel de prévention des risques : *alerte, vigilance*. On peut facilement le relever en consultant la presse en ligne ;

- les noms de communes fréquemment affectées ainsi que les noms des rivières soumises aux crues rapides ;

- le vocabulaire informel et plus ou moins familier que les *tweeters* sont susceptibles d'employer pour évoquer l'événement : *flotte* à la place de *pluie*, *déborder* pour signaler un cours d'eau en crue.

La difficulté consiste ensuite à identifier le vocabulaire spécifique aux effets de ces événements, c'est-à-dire d'inventorier les éléments susceptibles d'être affectés : *l'électricité coupée, la route fermée, la maison isolée, l'intervention des pompiers, l'arbre déraciné*, etc. L'objectif est de mettre en évidence les tweets qui ne mentionnent pas explicitement des éléments physiques (comme l'orage ou l'inondation) mais qui seraient focalisés sur les effets de l'événement. En procédant de cette manière, nous nous attendons à ce que des tweets comme "*j'ai plus d'électricité chez moi"* et *"j'ai plus d'électricité chez moi #orage #intempéries"* soient conservés dans le jeu de données de crise final. Cette première recherche a permis de collecter un ensemble de 80 mots-clés et associations lexicales.

Dans un second temps, nous avons procédé à un examen interne des données, afin de mettre en évidence du vocabulaire employé par les *tweeters* et de compléter la liste précédemment dressée. Nous avons eu recours à des outils de *text mining* (KH Coder) afin de clustériser les mots utilisés dans les tweets et de repérer : du vocabulaire identifié comme directement lié à l'événementiel hydro-météorologique auquel nous n'avions pas songé, comme *grêle* ou *apocalypse* ; des mots que nous avons inscrits dans notre liste de recherche, mais dont l'orthographe est erronée dans les tweets (*tonnaire)* ; des mots ou associations lexicales potentiellement liés aux événements, et dont la validité doit être testée : les mots *murs* et *tremblent* apparaissent dans un cluster contenant le terme *orage*. En revanche, l'association *journée perturbée* n'a identifié que quelques tweets sans lien avec notre problématique. A l'issue de ce double travail, dix catégories de mots-clés sont proposées. Elles sont présentées dans la figure 2 et distinguent le vocabulaire des phénomènes physiques d'une part, ses répercussions sur l'individu et la sphère sociale d'autre part, puis les mots supplémentaires mis en évidence par *text mining*.



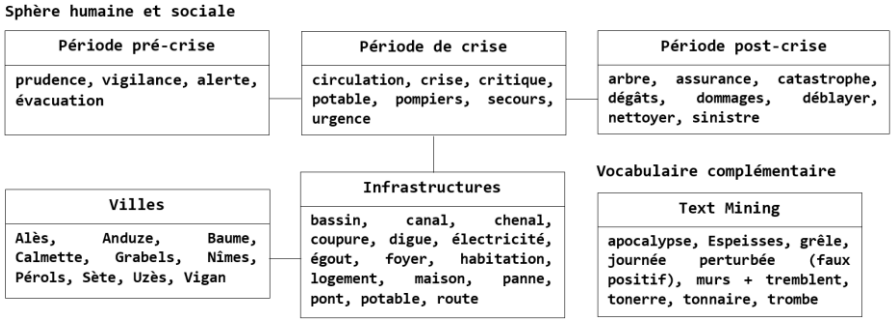



*Figure 2. Liste des mots-clés inventoriés par catégories*

### *3.1.2. Elimination du bruit résiduel et constitution du corpus final*

Un certain nombre de mots-clés pré-définis ou découverts dans le corpus de tweets peuvent se retrouver dans des contextes variés. C'est le cas par exemple des termes *pompiers*, *secours*, *alerte*, *inondation*. Nous avons donc jugé essentiel de vérifier la correspondance entre les tweets que nous avons extraits et le contexte événementiel considéré. Cette étape de relecture et de « nettoyage » des données a été effectuée manuellement. Étant donné que de nombreux tweets peuvent être sujets à l'interprétation, nous avons défini un ensemble de règles de validation des tweets considérés comme pertinents (tableau 1). Pour être sélectionné, un tweet doit être centré sur l'événement (ses manifestations physiques et/ou ses conséquences) ou sur les réactions de l'individu.

*Tableau 1. Liste des conditions et exemples de tweets sélectionnés*

| Thème du tweet | Exemple de tweet |
|---|---|
| bulletin ou alerte météorologique | *#Nîmes : arrêt des précipitations depuis 40 min. L'épisode devrait décroître d'ici la fin de la matinée* |
| événement en cours ou survenu | *Demain j'ai pas cours, mon lycée a été inondé !!* |
| conséquences (individus, biens, infrastructures) | *Le pont qui mène en ville est plein d'eau, j'suis coupé du monde ! ??* |
| inquiétude / peur | *Il pleut de dingue chez moi omg j'ai peur* |
| comportement | *Il fait nuit, il pleut, c'est l'alerte orange mais je fais les magasins* |
| opinion (gestion de la crise, secours) | *@meteofrance merci pour cette #alerterouge qui n'a servi à rien. Pas une goutte de pluie ici.* |

Les tweets dont le contenu s'est avéré inintelligible, ambigu, ou exprimait un souhait (un élève suppliant l'annonce d'une alerte afin que ses cours soient annulés) n'ont pas été retenus. Finalement, sur les 10 750 tweets présents dans le jeu intermédiaire après exécution des requêtes, nous n'en avons conservé que 2 789, soit 26%. Au final, le jeu de tweets de crise ne représente que 0,47% du jeu de tweets brut initial pour l'ensemble des événements de la période observée.

### *3.2. Analyse spatio-temporelle : application de la démarche de recherche*

Cette partie teste la démarche présentée au paragraphe 2.3. Nous étudions la distribution spatio-temporelle des tweets dans l'objectif de mettre en évidence des structures spatiales répétitives à travers le temps (échelle mensuelle).

#### *3.2.1. Détection des hotspots d'activité (Phase 1)*

La distribution spatio-temporelle des tweets a été analysée par l'étude de densités normalisées : l'objectif consiste à mettre en évidence les foyers (ou *hotspots*) potentiels d'activité en gommant les effets des densités de population, à partir de l'étude du jeu de tweets brut. Nous utilisons la densité de Kernel pour générer des cartes mensuelles des densités de tweets à partir du jeu de données brut. Une carte des densités de population est créée à partir des données carroyées de l'INSEE[3]. Les cartes finales sont générées en calculant le ratio entre les densités de tweets mensuelles et les densités de population. La figure 3 représente la distribution spatiale des ratios calculés pour chacun des mois collectés, et met en évidence les espaces présentant des anomalies.

Les zones où la valeur des pixels est comprise entre 0 et 1 correspondent à des espaces où l'activité peut être qualifiée de standard, c'est-à-dire que les densités de tweets émis sont inférieures ou égales à la densité de population. Les zones où ce ratio est supérieur à 1 forment nos zones d'intérêt et désignent les espaces qui concentrent plus de tweets que d'habitants.

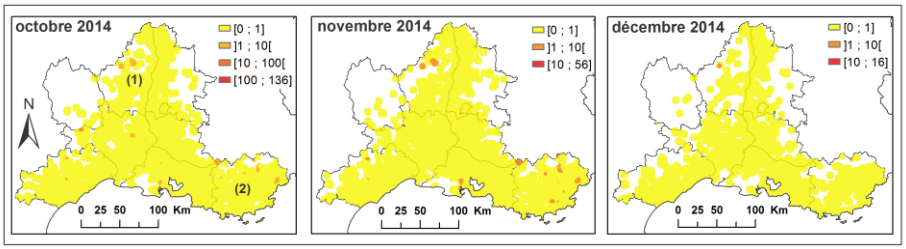



*Figure 3. Détection de hotspots mensuels d'activité*

[3] http://www.insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=donnees-carroyees

Les *hotspots* ainsi détectés sont principalement localisés dans les départements de l'Ardèche (1) et du Var (2), mais absents des agglomérations principales. Nous observons donc un phénomène résultant (B), les anomalies, et recherchons une hypothèse (A) qui pourrait l'expliquer.

### *3.2.2. Densités mensuelles de tweets de crise (Phase 2)*

Nous supposons que l'observation (B) résulte de la survenue d'un événement exceptionnel. Néanmoins, comme (B) apparaît dans des espaces péri-urbains ou ruraux, nous pensons que la perturbation à l'origine de (B) n'est pas liée à un événement d'origine sociale qui aurait la capacité de drainer un flux conséquent de tweets, comme un match de football ou toute autre manifestation. Nous supposons donc que ces anomalies proviennent de perturbations qui peuvent être d'origine hydro-météorologique.

Nous explorons alors les variations spatio-temporelles des densités mensuelles de tweets de crise (figure 4), afin de détecter l'existence éventuelle d'une relation entre les *hotspots* précédemment mis en évidence et les foyers d'émission de tweets de crise.

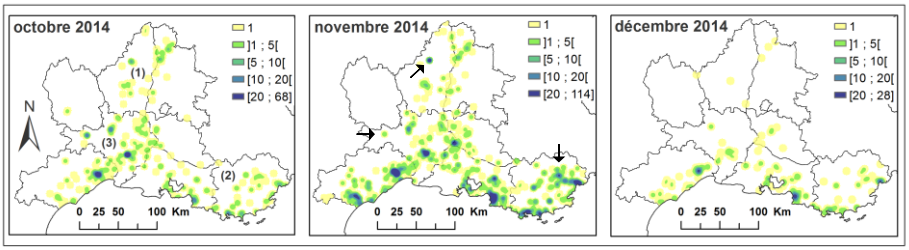



*Figure 4. Densités mensuelles de tweets de crise*

Sans surprise, les plus fortes densités de tweets sont rassemblées dans les agglomérations régionales et le long du littoral méditerranéen qui concentrent les plus fortes densités de population. Certains foyers d'émission de tweets de crise, qui constituent des motifs répétitifs dans le temps, en particulier dans le nord de l'Ardèche (1), l'ouest du Gard (3) et dans le Var (2) sont néanmoins identifiés en dehors de ces espaces et indiqués par des flèches sur la figure 4.

En comparant ces foyers à ceux mis en évidence lors de l'étape précédente, nous avons découvert des structures communes et choisi d'explorer plus en détail le *hotspot* localisé dans le nord de l'Ardèche afin de vérifier si l'anomalie détectée est bien liée à la manifestation de perturbations météorologiques. En d'autres termes, nous vérifions l'hypothèse selon laquelle des anomalies de *tweeting* permettent d'identifier des lieux particulièrement affectés par les événements hydro-météorologiques extrêmes, et constituent ainsi des marqueurs pertinents.

*3.2.3. Exploration du hotspot ardéchois (Phase 3)*

Nous souhaitons savoir ce qui est précisément arrivé en ce lieu particulier : en explorant ce motif plus en détail, il nous apparaît focalisé sur une commune rurale, le Cheylard, qui compte environ 3 000 habitants. La période étudiée enregistre 3 194 tweets bruts émis par quatre *tweeters*. Trois d'entre eux sont des utilisateurs occasionnels alors que le dernier peut être qualifié de *super tweeter* : il a en effet créé 3 106 tweets sur la période observée, dont 39 tweets de crise. Le foyer mis en évidence provenait donc de la super activité d'un seul utilisateur. Nous ne le considérons pas pour autant comme un *spammer*[4] : cet individu est en effet le *tweeter* pour lequel nous avons constaté la participation maximale à la création de tweets de crise (à titre comparatif, 90% des personnes enregistrées dans le jeu de tweets de crise ont créé moins de 5 tweets liés aux événements). Pour explorer les habitudes et le comportement de ce *tweeter*, nous avons sélectionné et analysé l'information contenue dans ses tweets de crise afin de les classer par thèmes : quatre ont été identifiés (tableau 2, les thème des tweets sont représentés en gras) :

*Tableau 2. Classement thématique et exemples de tweets*

| Thème du tweet | Exemple de tweet |
|---|---|
| **météo** : observation en temps réel | *Pluie et gris ici* |
| **inondation** : le *tweeter* a été inondé | *Là, ça y est c'est l'inondation* |
| **alerte** : l'Ardèche est en alerte ou vigilance météorologique | *noir ici, vigilance orange* |
| **information** : le tweet est lié à une crise mais ne concerne pas l'environnement local du *tweeter* | *Nouvel accident dans le Gard : un homme retrouvé noyé http://t.co/oBy8tM2X2C via BFMTV* |

La figure 5 révèle qu'une grande partie de l'information créée reste homogène sur le plan sémantique, et ne fournit, par conséquent, que peu d'information de terrain (qui concerne directement le *tweeter*) utile et pertinente : en effet, lorsqu'il *tweete* pour décrire un événement local, cet utilisateur crée une information relative à la météo, et plus particulièrement à une averse en cours. En outre, signalons que le thème que nous avons nommé *information* dans le tableau 2, qui représente une part non négligeable de l'effectif total de tweets de crise, se réfère à des événements qui ont affecté d'autres territoires, par exemple d'autres départements ou d'autres villes qu'il mentionne dans le tweet, alors que sa géolocalisation indique qu'il est resté dans son quartier de résidence. Nous constatons ainsi un autre problème désigné comme téléprésence asynchrone (Miller, 2007), qui qualifie l'émancipation de la présence physique de l'individu, dans le temps et dans l'espace, pour avoir connaissance d'un événement et s'en faire le témoin indirect.

[4] Un individu qui envoie une grande quantité de messages inutiles.

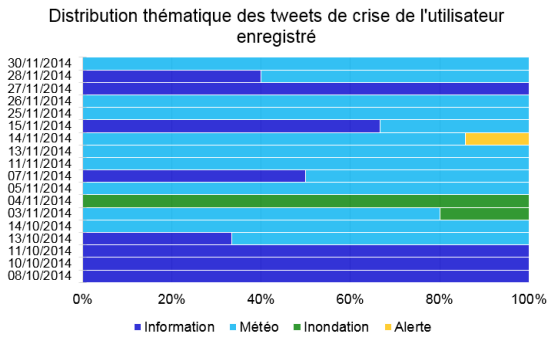



*Figure 5. Exploration des tweets de l'utilisateur du Cheylard*

Ainsi, la pertinence de la relation que nous cherchions à tester entre la concentration des tweets et la survenue d'un événement hydro-météorologique intense n'est pas complète. L'événement est susceptible d'entraîner une mobilisation locale, signalée par de fortes densités de tweets mais cet effet semble dépendant du comportement des individus et de leur activité sur la plateforme. Nous avons vérifié ce qui se passait en un autre point isolé qui correspond au foyer localisé sur la commune de Valleraugue (ouest du Gard), et qui s'étend jusqu'au Mont Aigoual. Le même phénomène est constaté : un seul *tweeter* crée de l'information de crise et celle-ci reste focalisée sur l'événement local et relative à la sphère sociale (mobilisation des secours, conséquences sur les infrastructures) mais ne fait aucune mention des caractéristiques physiques de l'événement.

L'information de crise d'un utilisateur n'est donc pas forcément exploitable pour qualifier un phénomène physique ou ses réponses sociales observables localement, puisque les tweets peuvent également contenir une information relative à d'autres lieux, parfois très éloignés de celui que l'on observe.

## 4. Conclusion

Cet article s'est attaché à présenter une démarche de recherche originale que nous avons testée dans l'objectif de problématiser la donnée tweet comme donnée géographique, mais également comme marqueur social d'événements extrêmes. Nous avons ainsi développé une méthodologie réutilisable d'exploitation des tweets, caractérisée par un double objectif : d'une part, en associant déduction et abduction, nous avons testé notre capacité à formuler des hypothèses pour expliquer des liens entre des observations, et d'autre part, nous avons mis en évidence les ambiguïtés du tweet comme marqueur de phénomènes. En vérifiant la correspondance spatiale et temporelle des anomalies de *tweeting* du jeu de tweets bruts et les foyers d'émission de tweets de crise, nous souhaitions savoir si ces anomalies étaient le résultat d'une perturbation d'origine hydro-météorologique et estimer si elles pouvaient constituer un indicateur pertinent. Les résultats ont montré qu'on ne pouvait ni confirmer, ni infirmer cette hypothèse : si une densité anormale de tweets peut être considérée

comme un marqueur d'événement, ces tweets ne constituent pas un indicateur fiable pour qualifier la dynamique physique et sociale de l'événement, en raison de trois facteurs : nous avons mis en évidence la forte activité d'individus isolés, et donc une nouvelle problématique comportementale, la téléprésence aux événements ainsi que l'incomplétude de l'information.

Ces résultats nous amènent à considérer les remarques suivantes : il serait préférable d'explorer les *hotspots* d'espaces plus peuplés, comme la vallée de la Nartuby entre Draguignan et Le Muy (Var), qui peuvent offrir un meilleur potentiel d'information diversifiée. Dans un second temps, nous soulevons la question suivante : étant donné le maigre jeu de tweets de crise que nous avons extrait, et dont il resterait à évaluer le degré de complétude, nous pouvons nous demander ce que représente un événement majeur pour un *tweeter* dans le contexte des risques naturels, c'est-à-dire un événement qui est, sur le plan individuel ou collectif, suffisamment important pour devenir un sujet de discussion prépondérant.

Au final, après la mise en œuvre de ces premiers traitements, l'abduction pourrait nous aider à apporter des éléments de réponse à d'autres questions : quels facteurs poussent le *tweeter* à créer une information de crise ? qu'est-ce-qui explique la variabilité spatio-temporelle de l'information sémantique et de sa quantité en termes d'effectif ? Néanmoins, à ce stade et avec le peu de données collectées, il est difficile d'affirmer la possibilité de reconstituer l'histoire d'une crise à partir des tweets. Au vu des résultats, nous pensons que le tweet n'a pas la capacité de constituer une donnée autosuffisante et que la construction de connaissances ne doit pas être séparée du contexte de production de l'information, c'est-à-dire qu'elle doit prendre racine sur l'utilisation simultanée des tweets et de données complémentaires, rattachées au contexte physique de l'événement (cumuls pluviométriques, hauteur d'eau des rivières, etc.).

*Remerciements.*

*Le travail présenté dans ce papier s'inscrit dans le cadre du projet ANR MobiClimEx (ANR-12-SENV-0002-01) et du projet MISTRAL-HyMeX.*

*Les auteurs remercient l'équipe SLIDE du Laboratoire d'Informatique de Grenoble pour la mise à disposition des tweets via leur serveur de collecte.*

## Bibliographie

Anderson C. (2008). *The end of theory: The data deluge makes the scientific method obsolete*, http://www.wired.com/2008/06/pb-theory/

Andrienko G., Andrienko N., Bosch H., Ertl T., Fuchs G., Jankowski P., Thom D. (2013). Thematic patterns in georeferenced tweets trough space-time visual analytics. *Computing in Science and Engineering* vol. 15, n° 3, p. 72-82.

Dashti S., Palen L., Heris M. P., Anderson K. M., Anderson T. J., Anderson S. (2014). Supporting disaster reconnaissance with social media data : a design-oriented case study

of the 2013 Colorado floods. *Proceedings of the 11th International ISCRAM Conference*, University Park, USA.

De Longeville B., Smith R. S., Luraschi G. (2009). "OMG, from here, I can see the flames!" : a case use of mining Location Based Social Networks to acquire spatio-temporal data on forest fires. *Proceedings of the 2009 International Workshop on Location Based Social Networks*, Seattle, USA.

Dittrich A., Lucas C. (2014). Is this Twitter event a disaster ? *Proceedings of the AGILE'2014 International Conference on Geographic Information Science.* Castellón, Spain.

Erwig M. (2004). Toward Spatio-Temporal Patterns. *Spatio-Temporal Databases*, Berlin Heidelberg, Springer, p. 29-53.

Hawelka B., Sitko I., Beinat E., Sobolevsky S., Kazakopoulos P., Ratti C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* vol. 41, n° 3 , p. 260-271.

Hertfort B., Porto de Albuquerque J., Schelhorn S.-J., Zipf A. (2014). Exploring the geographical relations between social media and flood phenomena to improve situational awareness. *Connecting a digital Europe through Location and Place*, Springer International Publishing, p. 55-71.

Hoffmann M. (1999). Problems with Peirce's concept of abduction. *Foundations of Science*, vol. 4, n° 3, p. 271-305.

Kell D. B., Oliver S. G. (2003). Here is the evidence, now what is the hypothesis ? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, vol. 26, n° 1, p. 99-105.

Kitchin R. (2013). Big data and human geography. Opportunities, challenges and risks. *Dialogues in Human Geography*, vol. 3, n° 3, p. 262-267.

Kounadi O., Lampoltshammer T.J., Groff E., Sitko I., Leitner M. (2015). *Exploring Twitter to analyze the public's reaction patterns to recently reported homicides in London*, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0121848

Lee Hughes A., Palen L. (2009). Twitter adoption and use in mass convergence and emergency events. *Proceedings of the 6th International ISCRAM Conference*, Gothenburg, Sweden.

Li L., Goodchild M. F., Xu B. (2013). Spatial, temporal and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, vol. 40, n° 2. p. 61-77

Lin Y. R., Margolin D., Keegan B., Baronchelli A., Lazer D. (2013). #Bigbirds never die : understanding social dynamics of emergent hashtags. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, Cambridge, USA.

Miller H. J. (2007). Placed-Based versus People-Based Geographic Information Science. *Geography Compass*, vol. 10, n° 3, p. 503-535.

Miller H. J., Goodchild M. F. (2015). Data-driven geography. *Geojournal*, vol. 80, n° 4, p. 449-461.

Walters B. B. (2012). An event-based methodology for climate change and human-environment research. *Geografisk Tidsskrift-Danish Journal of Geography*, vol. 112, n° 2, p. 135-143.