

Des tweets pour l'observation des réponses sociales aux crues rapides de l'automne 2014 dans le sud-est de la France : quelles informations, quels traitements ?

Camille Cavalière

▶ **To cite this version:**

Camille Cavalière. Des tweets pour l'observation des réponses sociales aux crues rapides de l'automne 2014 dans le sud-est de la France : quelles informations, quels traitements ?. Hydrologie. 2015. dumas-01282012

HAL Id: dumas-01282012

https://dumas.ccsd.cnrs.fr/dumas-01282012

Submitted on 3 Mar 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Master 2 « Sciences du Territoire »

Mention « Systèmes Territoriaux, Aide à la Décision, Environnement » (STADE)

Des tweets pour l'observation des réponses sociales aux crues rapides de l'automne 2014 dans le sud-est de la France : quelles informations, quels traitements ?

Stage de recherche réalisé en partenariat avec le laboratoire PACTE et le LIG

Mémoire soutenu le 16 juin 2015

Par Camille CAVALIERE

devant un jury composé de :

Responsable : **Mme Céline Lutoff, Maître de conférences, Institut de Géographie Alpine, Université Joseph Fourier**

Examinateurs : **Mme Elise Beck, Maître de conférences, Institut de Géographie Alpine, Université Joseph Fourier**

Mme Paule-Annick Davoine, Maître de conférences, Laboratoire d'Informatique de Grenoble

Mme Céline Lutoff, Maître de conférences, Institut de Géographie Alpine, Université Joseph Fourier

Année universitaire 2014-2015

Remerciements

Je tiens tout d'abord à remercier mon maître de stage, Mme Paule-Annick Davoine, maître de conférences au Laboratoire d'Informatique de Grenoble (LIG), ainsi que mon responsable pédagogique, Mme Céline Lutoff, maître de conférences à l'Université Joseph Fourier de Grenoble, pour m'avoir proposé ce stage très enrichissant en termes de technique et de méthode. Je les remercie encore pour leurs conseils avisés dans l'orientation de la recherche et dans la rédaction de ce mémoire.

J'adresse également mes plus sincères remerciements à Isabelle Ruin, chargée de recherche au LTHE pour ses conseils et pour m'avoir fait découvrir les outils de fouille de texte. Un grand merci également aux membres de l'équipe SLIDE du LIG, et en particulier à Etienne Dublé, ingénieur de recherche au LIG, sans l'intervention duquel le travail effectué dans le cadre de ce stage n'aurait pas été possible.

Enfin, un grand merci à tous les membres de l'équipe Steamer du LIG pour leurs conseils et leur accueil toujours aussi cordial.

Liste des acronymes

ANR : Agence Nationale de la Recherche

API : Application Programming Interface

CGIAR-CSI : Consortium for Spatial Information

CSV : Comma Separated Value

GMT : Greenwich Mean Time

IGN : Institut Géographique National

INSEE : Institut National de la Statistique et des Etudes Economiques

LIG : Laboratoire d'Informatique de Grenoble

LTHE : Laboratoire d'Etude des Transferts en Hydrologie et Environnement

OHMCV : Observatoire Hydro-Météorologique Cévennes-Vivarais

PACTE : Politiques publiques, Actions politiques, Territoires

PPRI : Plan de Prévention des Risques d'Inondation

RESO : Risques, environnement et société

SIG : Système d'information géographique

SLIDE : Scalable Information Discovery and Exploitation

SQL : Structured Query Language

SRID : Spatial Reference System Identifier

SRTM : Shuttle Radar Topography Mission

STeamer : Spatio-temporal information, Adaptability, Multimedia and knowledge Representation

TU : Temps Universel

WKT : Well-Known Text

Sommaire

Remerciements	p.2
Liste des acronymes	p.3
Sommaire	p.4
Présentation du stage et de la structure d'accueil	p.5
Introduction	p.6
Partie 1 : Twitter, une nouvelle source de données ouvertes et massives : perspectives et contraintes	p.11
Partie 2 : Du tweet à la carte : extraire, structurer et représenter l'information	p.21
Partie 3 : Analyse et représentation des tweets : quelles informations pour quelles distributions ?	p.31
Conclusion	p.55
Références bibliographiques	p.57
Table des figures	p.59
Table des tableaux	p.61
Table des matières	p.62
Table des annexes	p.64
Annexes	p.65

Présentation du stage et de la structure d'accueil

Mon stage a été effectué sur le projet ANR MobiClimEx, au sein de deux laboratoires : le laboratoire PACTE, porteur du projet, et le LIG. Ces deux laboratoires de recherche sont structurés autour de différentes thématiques. Ancré sur l'étude de la contribution potentielle d'une nouvelle forme de données ouvertes, les médias sociaux et en particulier les tweets, à la production d'une information exploitable pour l'analyse des comportements des populations, ce stage s'articule autour de deux thèmes de recherche. D'une part, il s'inscrit dans le thème RESO (Risques, environnement et société) du laboratoire PACTE et d'autre part, dans le thème de traitement de données et de connaissances à grande échelle en ce qui concerne le LIG.

La thématique RESO[1] est focalisée sur l'étude des interactions entre environnement et société, en particulier sur les questions de l'évolution et de l'adaptation des sociétés à leur environnement. Ces questions s'orientent ainsi sur trois axes de recherche : le sensible, c'est-à-dire les perceptions de l'espace par les sociétés, la gestion et les trajectoires paysagères ainsi que les risques. Ce dernier axe étudie notamment l'adaptation des sociétés aux changements et phénomènes climatiques extrêmes, sur des temporalités courtes ou longues. L'approche sociale des risques occupe une place majeure dans l'analyse : il s'agit en effet d'étudier l'évolution de l'exposition et de la vulnérabilité des populations en fonction de contextes politiques et sociaux spécifiques, mais encore d'analyser les réponses sociales face à un événement, au travers d'enquêtes post-crise, de retours d'expérience ou encore de l'évaluation de la planification de reconstruction.

L'axe de traitement de données et de connaissance à grande échelle du LIG est structuré autour de la géomatique, de l'extraction et de la représentation de connaissances, de la gestion et de la fouille de données massives, et de l'information issue des réseaux sociaux. J'ai ainsi intégré l'équipe Steamer[2] dont les travaux s'appuient sur le développement de formalismes et de méthodes de conception, de mise en œuvre et d'utilisation de systèmes d'information spatio-temporelle appliqués aux domaines de la cartographie interactive, des SIG mobiles et de l'étude des risques naturels. Le sujet du stage a également requis un travail collaboratif avec l'équipe SLIDE[3] du LIG, dont les travaux s'articulent autour de la collecte, de la fouille et du traitement de données massives issues de diverses sources. Ces chercheurs ont notamment mis au point l'infrastructure de collecte et de nettoyage qui m'a permis d'obtenir l'échantillon de tweets, indispensable au déroulement du stage.

Les principales missions de ce stage de quatre mois consistent ainsi, d'une part, à extraire de l'échantillon original une information qualifiée d'utile, qui permette, plus tard, d'apporter des éléments de réponse sur la question comportementale face à un événement extrême. D'autre part, il s'agit de résumer cette information et de comparer sa temporalité avec celle de l'événement, c'est-à-dire que nous chercherons à savoir s'il existe une simultanéité entre le contenu de l'information et la dynamique de l'événement.

[1] http://www.pacte-grenoble.fr/risque-environnement-et-societe-reso/
[2] https://www.liglab.fr/presentation/equipes/steamer
[3] https://www.liglab.fr/presentation/equipes/slide

Introduction

Présentation du projet MobiClimEx

Le projet ANR MobiClimEx[4] – Dynamique des mobilités quotidiennes et résidentielles face aux extrêmes météorologiques en contexte de changement climatique – est centré sur l'étude de l'exposition humaine face aux événements extrêmes d'origine hydrométéorologique, en particulier les crues rapides provoquées par les épisodes cévenols dans les régions méditerranéennes.

L'objectif du projet est d'analyser l'évolution des interactions entre les facteurs environnementaux et les facteurs sociaux aux échelles spatiales et temporelles. En effet, les fluctuations climatiques ont tendance à modifier les paramètres des aléas hydrométéorologiques ; de même, l'exposition des populations varie en fonction du temps, de l'espace, des comportements individuels ou collectifs, et du cadre législatif. Par exemple, la mobilité des individus en temps de crise ainsi que l'extension des zones résidentielles et l'installation de populations néorurales constituent des facteurs influençant la vulnérabilité des populations.

MobiClimEx s'intéresse donc, dans un contexte de changement climatique, à l'exposition des sociétés au travers de ces deux mobilités – quotidienne et résidentielle – ainsi qu'à leur évolution spatio-temporelle. L'évaluation des réponses sociales face aux dynamiques des crues rapides s'effectue alors à différentes échelles : il s'agit dans un premier temps d'identifier les effets de la survenue de l'événement dans l'environnement quotidien des individus, c'est-à-dire d'observer les modifications éventuelles que les individus apportent à leurs activités quotidiennes, à leur mobilité et donc à leur exposition. A plus long terme, l'objectif consiste également à diagnostiquer les évolutions résidentielles et notamment l'impact des politiques publiques mises en œuvre pour réguler l'exposition face aux crues rapides.

Le projet requiert par conséquent l'application d'une analyse intégrant les éléments des processus physiques et des phénomènes sociaux en jeu lors des crues rapides, ainsi que la mise en œuvre d'une approche permettant de croiser échelle spatiale et échelle temporelle. Le stage est, quant à lui, fondé sur l'exploitation d'une source de données récente, les médias sociaux, présentant des caractéristiques particulières (qui seront décrites dans les sections suivantes) et sa potentielle contribution à fournir des informations sur la place d'une crise dans les préoccupations et activités des populations concernées.

L'épisode cévenol

Les épisodes cévenols[5] se manifestent par des orages particulièrement violents et locaux : ils affectent en général un bassin versant précis et déversent, en un temps relativement restreint, des quantités d'eau considérables. Ces pluies intenses et durables, dont les cumuls peuvent atteindre plusieurs centaines de millimètres d'eau en quelques heures, ont un effet double : d'une part, elles saturent rapidement les sols et le ruissellement des versants provoque alors inondations, coulées boueuses, glissements de terrain et effondrement de chaussées ; d'autre part, elles gonflent les lits de petits cours d'eau et sont par conséquent à l'origine de crues rapides pouvant entraîner des dégâts

[4] http://www.agence-nationale-recherche.fr/suivi-bilan/editions-2013-et-anterieures/environnement-et-ressources-biologiques/societes-et-changements-environnementaux/fiche-projet-soc-env/?tx_lwmsuivibilan_pi2[CODE]=ANR-12-SENV-0002

[5] http://pluiesextremes.meteo.fr/episodes-mediterraneens_r48.html

considérables : en témoigne la crue de l'Ouvèze à Vaison-La-Romaine qui dévaste le centre-ville le 22 septembre 1992 (Figure 1).



Figure 1 : L'Ouvèze en crue sous le Pont Romain de Vaison-la-Romaine, le 22 septembre 1992 ***(Météo France pluies extrêmes)***

Ces événements résultent de conditions météorologiques particulières, qui surviennent en début de l'automne, lorsque les eaux de surface de la Méditerranée sont encore suffisamment chaudes. Ils se déclenchent en effet lorsque des masses d'air chaud et humide, en provenance de la Méditerranée et proches du sol, rencontrent des masses d'air froid situées en altitude : la confrontation de ces deux masses crée une instabilité orageuse forte. Enfin, les nuages se trouvent bloqués par les contreforts de massifs montagneux, en particulier les Cévennes, et des pluies torrentielles se déversent alors sur des surfaces relativement réduites.

Ces orages affectent principalement les départements de l'Ardèche, du Gard, de l'Hérault et de la Lozère, mais peuvent également déborder sur l'Aude, les Bouches-du-Rhône, ainsi que sur les Préalpes du Var et du Vaucluse. Ils surviennent fréquemment à l'automne, entre septembre et novembre et durent généralement entre 24 et 76 heures. Leur effet principal, sur les bassins versants, est la modification du régime hydrologique des rivières qui subissent alors des variations importantes de débit et provoquent de fortes crues accompagnées d'inondations conséquentes. Ainsi, pour reprendre l'exemple de la catastrophe de Vaison-la-Romaine en 1992, l'Ouvèze, dont le début moyen est estimé entre 10 et 12 m^3/s a atteint, en quelques heures, un débit de pointe de 1 200 m^3/s pour une hauteur d'eau de 17 mètres par rapport au lit mineur[6].

Localement, les crues cévenoles sont souvent nommées par le toponyme du cours d'eau, auquel est généralement ajouté le suffixe « -ade » : les crues du Lez à Montpellier sont ainsi qualifiées de « lézardes », les crues du Vidourle à Nîmes, les « vidourlades », etc.

Qu'est-ce-que Twitter ?

Créé en 2006, Twitter[7] est un service de microblogage, concept directement dérivé du blog, à partir duquel les utilisateurs peuvent communiquer via des messages courts, les tweets, limités à 140 caractères, auxquels peuvent être ajoutées photographies, vidéos ou URL. La motivation des concepteurs de la plateforme étant la constitution d'un réseau international ouvert, chaque tweet émis appartient au domaine public et est par conséquent visible par tout internaute, qu'il soit utilisateur de Twitter ou non. L'identification d'un tweet relatif à un sujet particulier peut s'effectuer en ajoutant un *hashtag*, c'est-à-dire un mot précédé du signe # ; ces *hashtags* sont créés quotidiennement, en particulier par les médias et institutions, en fonction de l'actualité.

[6] http://www.vaison-la-romaine.com/IMG/pdf/11-crue_de_1992.pdf
[7] http://fr.wikipedia.org/wiki/Twitter

Twitter dispose par ailleurs d'une interface de requête permettant de filtrer les tweets référencés en fonction de *hashtags* ou d'autres variables comme le contenu textuel, le nom d'utilisateur, le nom de la ville d'émission du tweet ou encore la date, ainsi que d'une API (interface de programmation informatique) permettant de télécharger un jeu de tweets correspondant à des critères sélectionnés.

Grâce à ses fonctions de tri, de filtrage et d'archivage, la plateforme peut donc être assimilée à un outil de stockage et de gestion de l'information massive. En outre, la cartographie des *hashtags* est souvent utilisée dans des buts précis : suivre les effets d'une catastrophe ou la propagation d'une épidémie, ou, à une autre échelle, visualiser la diffusion des informations dans le monde. En revanche, si Twitter permet de diffuser rapidement l'information, l'aspect non professionnel de la source reste très controversé (Goodchild & Glennon, 2010), notamment en raison de l'absence de régulation et de formalisme cadrant la production de l'information.

L'information envoyée sur Twitter a donc plusieurs propriétés : acquise en temps réelle par des observateurs, elle peut receler davantage de détails sur un événement précis ; elle est hétérogène, autant en forme qu'en contenu et provient de sources variées (Schade *et al.*, 2011). L'acquisition de connaissances à partir de cette source d'information nécessite en revanche de s'interroger sur sa pertinence et de prendre en compte les éventuels inconvénients liés à la limitation de 140 caractères en termes de langage (abréviations, phonétique, etc.)

Phénomènes hydrométéorologiques et inondations : entre géographie du risque et géomatique

L'ancrage de la géomatique et de la cartographie comme outils de prévention des risques naturels, et en particulier des aléas liés aux phénomènes hydrométéorologiques et aux inondations, tend à s'affirmer, non seulement dans la législation française mais encore dans les besoins des acteurs locaux dont l'enjeu est de concilier aménagement raisonné du territoire et prévention des risques (De Blomac, 2015).

Les événements extrêmes survenus ces dernières années – citons par exemple la tempête Xynthia et les inondations ayant frappé le Var en 2010 - ont entraîné la modification du cadre législatif qui s'appuie, dès lors, davantage sur un socle d'information géographique et l'intervention des outils de la géomatique, qu'il s'agisse de constructions de cartographies d'événements historiques, d'aléas, de diagnostics d'enjeux et de vulnérabilités, ou encore de modélisation de scénarii (De Blomac, 2015). La géomatique occupe donc une part active dans la construction de la mémoire des risques, elle-même généralement considérée comme l'instrument essentiel de la prévention (Dollfus & D'Ercole, 1996). Dans la géographie des risques, la mémoire est en effet appréhendée comme construction humaine et sociale qui capitalise les expériences, permet d'identifier les lieux et processus générateurs de dangers, tout en intégrant l'étude des comportements sociétaux face aux aléas sévissant dans un milieu anthropisé (Bailly, 1996). Dans le cas particulier des médias sociaux, l'analyse se situe à l'interface entre les outils informatiques de traitement de données nombreuses, les outils de la géomatique, l'événement et l'information individuelle qui lui est liée et qui peut potentiellement révéler des comportements face à la crise, ainsi que la place qu'elle occupe chez les populations affectées. Le travail du géomaticien consiste alors à organiser ces informations nombreuses pour représenter leur contenu ainsi que la variabilité spatio-temporelle de ce contenu ; au-delà, il s'agirait de savoir si cette information individuelle participative, peut, au même titre que des cartes d'aléas ou de vulnérabilité, s'intégrer à la mémoire collective et servir d'appui aux planificateurs en situation de retour d'expérience.

Problématisation : extraire et représenter l'information utile contenue dans des jeux de données nombreuses

En 2013, le nombre moyen de tweets postés quotidiennement dans le monde entier est estimé à 340 millions (Andrienko *et al*., 2013). Ainsi se pose la première difficulté qui consiste à extraire, lorsqu'on travaille sur une thématique particulière, sur une période ou encore sur un territoire précis, un jeu de tweets correspondant aux critères adéquats. Dans notre étude, nous travaillons sur les huit départements du sud-est de la France (Figure 2), qui sont fréquemment affectés par les phénomènes hydrométéorologiques décrits précédemment : il s'agit de l'Ardèche, de la Drôme, du Gard, de la Lozère, de l'Hérault, des Bouches-du-Rhône, du Var et du Vaucluse. Notre terrain d'étude comprend donc la vallée du Rhône ainsi que les départements dont certains territoires, situés aux pieds de contreforts montagneux – Cévennes, Vivarais, Préalpes, Baronnies, Maures, Estérel et Luberon – sont fréquemment affectés par les orages cévenols, comme les vallées du Lez, du Gard, de l'Ouvèze ou encore de l'Argens. Par ailleurs, la période temporelle sur laquelle le stage s'appuie couvre les mois d'octobre 2014 à mars 2015. Nous allons ainsi collecter un jeu de tweets bruité, c'est-à-dire que l'information contenue à l'intérieur est très variée et concerne avant tout le quotidien des utilisateurs ; de plus, ce jeu se présentera sous un format brut qu'il sera nécessaire de nettoyer et de structurer avant son importation dans les outils de traitement statistiques et cartographiques appropriés.

Les enjeux essentiels autour desquels le mémoire s'articule se déclinent donc de la manière suivante : nous devrons, dans un premier temps, supprimer l'information bruitée, c'est-à-dire le tweet du quotidien. Dans un second temps, il s'agira de définir les différentes étapes qui permettent d'évoluer, depuis un jeu de tweets bruts, vers un jeu de données construit, structuré et exploitable. Pour autant, le questionnement essentiel auquel le mémoire se tâchera d'apporter des éléments de réponse est le suivant : la perturbation cévenole est-elle un fait social suffisamment important pour être enregistrée dans les tweets ? Comment pouvons-nous progresser d'une information individuelle vers la caractérisation de comportements collectifs, c'est-à-dire comment résumer une information individuelle de contenu variable pour apporter des réponses sur les comportements des personnes en période de crise ? Pouvons-nous comparer la dynamique événementielle au flux et au contenu de l'information des personnes ? Le fait naturel et le fait social indiquent-ils des temporalités identiques ?

Nous pouvons d'ores-et-déjà soumettre les hypothèses suivantes : en termes d'utilisation de la plateforme de microblogage, nous supposons l'existence d'un lien étroit entre la distribution de la population et la distribution des tweets que nous pourrons cartographier ; ainsi, les espaces les plus densément peuplés sont probablement les lieux d'émission les plus importants. Pour autant, les tweets ne concernent qu'une partie restreinte de la population : les médias sociaux sont en effet majoritairement utilisés par une population jeune et habituée à communiquer via les médias et réseaux sociaux. En termes d'échange d'informations liées à une perturbation, nous pouvons supposer que le tweet enregistre la survenue d'un événement et que les flux de tweets augmentent pendant les périodes de crise. Nous pouvons également penser qu'il existe une simultanéité entre la dynamique événementielle et le contenu textuel du tweet ; de même, le tweet peut traduire le ressenti de l'individu vis-à-vis de l'événement et peut encore servir de référence pour mesurer la sévérité d'une crise, notamment au travers des flux enregistrés.

Le mémoire est donc structuré autour de trois parties : dans un premier temps, nous présenterons d'une part, les caractéristiques de l'information issue des médias sociaux autour des principaux concepts qui la théorisent et, d'autre part, les méthodes d'extraction et de classification de l'information mises en exergue dans la littérature. La deuxième partie s'attachera à la description des données, outils, et méthodes mises en œuvre pour constituer les différents corpus de tweets. Enfin, une dernière partie présentera les résultats des analyses et proposera un retour critique sur les choix méthodologiques retenus.

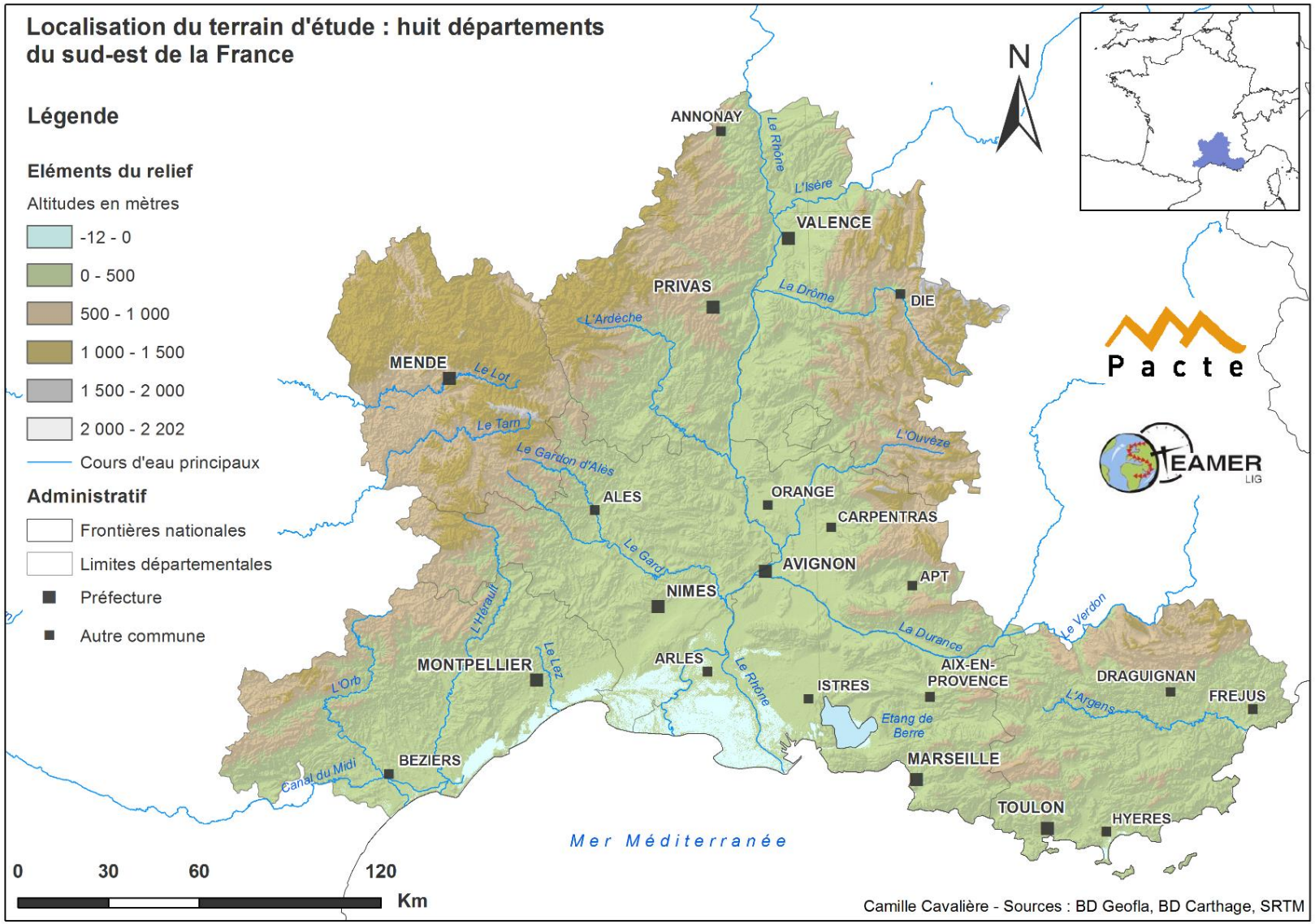


Figure 2 : Carte de présentation du terrain d'étude

1. Twitter, une nouvelle source de données ouvertes et massives : perspectives et contraintes

Cette première partie propose un état de l'art synthétique des travaux réalisés à partir de données issues de différentes plateformes de production d'information collective du web social, en particulier le réseau Twitter. Il s'agit dans un premier temps de recontextualiser l'utilisation du web 2.0 et des médias sociaux dans la communauté scientifique à travers les définitions des principaux concepts et de présenter les caractéristiques de cette source d'information. Nous exposerons ensuite les outils de visualisation et de collecte fréquemment utilisés, dans le cas spécifique de Twitter, avant de décrire les principales pistes d'exploitation destinées à construire un jeu de données formalisées.

1.1. Des concepts-clés autour du web social

Le Web 2.0 ou Web social fait référence à l'ensemble des technologies permettant d'introduire un certain degré d'interactivité entre une interface et un utilisateur qui, même si ses compétences en informatique sont très réduites, est apte à créer un contenu, c'est-à-dire une information, à la diffuser auprès d'autres utilisateurs et à enrichir le contenu web créé par un autre internaute (Millerand *et al*., 2010) . Le web social envisage ainsi Internet comme un lieu participatif où l'utilisateur a pour rôle de créer, de partager et de remanier l'information (Millerand *et al*., 2010) ; cette interactivité s'inscrit en particulier dans les médias sociaux.

1.1.1. Quelques définitions autour des médias sociaux

Les médias sociaux désignent alors toute interface sur laquelle un groupe d'internautes construit, indexe, organise et diffuse une information qui peut être modifiée ou commentée ; en dehors des réseaux sociaux, les exemples les plus connus et fréquemment consultés par les internautes restent les encyclopédies collaboratives, les blogs et microblogs (Twitter) ou encore les plateformes de stockage et de partage des photographies et de vidéos.

Plus récemment, la démocratisation et l'accessibilité à des outils de géolocalisation (GPS) ou directement connectés à Internet (Smartphone) ont accru la contribution de non-spécialistes à la production de données massives, c'est-à-dire de données acquises en temps réel, rapidement et en très grande quantité (Schade *et al*., 2013). Ces outils offrent un nouveau potentiel pour la recherche scientifique en termes de suivi de certains phénomènes, dans la mesure où il existe des milliers de contributeurs connectés (Schade *et al*., 2013). Dans le domaine de la géomatique, nous nous intéressons plus particulièrement à deux concepts :

- La *Volunteered Geographic Information* (VGI)

La communauté scientifique caractérise la VGI comme toute information géolocalisée, créée par un utilisateur et partagée gratuitement sur Internet via l'utilisation de différentes plateformes (Schade *et al*., 2013). Ces plateformes peuvent supporter une information quotidienne quelconque, comme Twitter, ou être destinée à une collecte de données dans un domaine précis, via des contributeurs volontaires : il s'agit par exemple de la plateforme Isibat-online, développée dans le cadre de l'ANR URBASIS[8], qui permet à tout individu enregistré et équipé d'un Iphone de participer à la collecte de données relatives à la vulnérabilité sismique des bâtiments (Davoine, 2014).

En tant qu'information liée aux médias sociaux, la VGI hérite de leurs principales propriétés : elle est donc massive, inépuisable, acquise et partagée en temps réel, hétérogène dans sa forme et dans

[8] Sismologie urbaine : évaluation de la vulnérabilité et des dommages sismiques par méthodes innovantes

son contenu, variable dans sa forme et dans sa crédibilité, et gratuite puisqu'elle entre dans le domaine public (Schade *et al*., 2013).

- Le *Crowdsourcing*

Le *crowdsourcing* caractérise l'acquisition d'un contenu via la sollicitation d'un groupe d'individus ou d'une communauté d'utilisateurs d'une plateforme. En outre, ce concept repose sur l'hypothèse selon laquelle un groupe de personnes peut s'avérer plus efficace qu'un expert pour résoudre un problème, et ce malgré l'existence d'un manque de professionnalisme (Goodchild & Glennon, 2010). Il implique ainsi deux postulats : d'une part, l'information créée par un groupe composé de multiples observateurs reflète davantage la réalité qu'une information créée par une seule personne. D'autre part, l'information collectée et diffusée par des citoyens volontaires qui coopèrent et ont un intérêt commun dans le domaine en question est plus précise et plus fiable qu'une information provenant d'une minorité extérieure (Goodchild & Glennon, 2010).

- Le citoyen, un nouveau producteur de connaissances

VGI et *crowdsourcing* considèrent l'individu comme « *citizen-as-sensor* » (Goodchild, 2009). Le citoyen-capteur appréhende ainsi l'individu comme l'élément fondamental de la connaissance de son territoire. En effet, l'individu perçoit son environnement et, au travers de ses pratiques quotidiennes de l'espace, il est capable de détecter tout changement intervenu. Qualifier le citoyen de capteur implique alors qu'un tel individu, équipé d'un appareil de géolocalisation, est apte à constituer un jeu de données spatio-temporelles révélant davantage de précision qu'un jeu de données expert : ces données peuvent ensuite servir de socle pour la production d'une connaissance scientifique (Goodchild, 2009).

1.1.2. L'affirmation des médias sociaux dans la communauté scientifique

Nous pouvons alors nous interroger sur les raisons de l'accroissement de l'utilisation d'une source de données témoignant d'une forte variabilité au niveau de sa forme et de sa qualité ou encore sur les métadonnées disponibles quant à son acquisition.

Dans la communauté scientifique, l'ancrage de la VGI et du *crowdsourcing* tend en effet à s'affirmer, notamment au travers de leur utilisation comme outils de collecte d'informations pour la gestion de crise. Ainsi, l'exploitation de cette information et la production de cartographies de crise ont d'ores-et-déjà fait leurs preuves lors des catastrophes naturelles récentes (McDougall, 2012) : Fukushima, les inondations du Queensland en Australie ou encore le suivi des incendies du maquis californien dans la région de Santa Barbara (Goodchild & Glennon, 2010).

A chaque événement évoqué, la mise en œuvre rapide de plateformes permettant la collecte instantanée de l'information citoyenne émise depuis téléphones portables et médias sociaux – comme Twitter – a favorisé une réponse rapide des autorités en termes d'organisation des secours selon la gravité des dommages affectant populations et territoires, puis de reconstruction.

Pour autant, l'utilisation de cette information ne se limite guère à la gestion d'une crise en cours : en effet, pendant les différentes phases d'un événement, les individus s'orientent vers les plateformes du web social pour échanger des informations sur l'évolution de la situation et expriment fréquemment leur point de vue sur la qualité des informations qu'ils ont pu recevoir (vigilance, alerte) ainsi que leur ressenti personnel vis-à-vis de l'efficacité de l'assistance aux victimes et des dispositifs de sauvegarde (Roy Chowdhury *et al*., 2013). Mise en relation avec d'autres données - photographies, imagerie aérienne ou satellite - l'analyse post-crise de cette information constitue alors un socle de connaissances pour effectuer un retour d'expérience, c'est-à-dire identifier les éléments ou comportements ayant accru la vulnérabilité ou les dysfonctionnements éventuels de la planification antérieure à l'événement (Dashti *et al*., 2014).

Enfin, l'exploitation de ces informations fournit une source de données alternative et avantageuse par rapport aux organismes et institutions traditionnels : il s'agit d'une information citoyenne et participative (McDougall, 2012) dont les producteurs sont déjà répartis sur le territoire ; en conséquence, tout individu équipé de l'outil adéquat et disposant d'un accès à Internet, devient un producteur potentiel d'information intéressante en cas de survenue d'un événement inhabituel. Chaque individu peut ainsi apporter instantanément une information immédiatement disponible, alors que des données plus élaborées produites par des organismes officiels et experts seront plus tardivement accessibles et payantes (McDougall, 2012). Par ailleurs, selon Dashti *et al*. (2014), l'information capturée par un observateur sur place au moment de l'événement peut s'avérer plus riche qu'une donnée experte, puisqu'il existe en général un délai entre la survenue de l'événement et l'arrivée des experts.

1.1.3. Quel degré de confiance attribuer à cette source d'information ?

L'exploitation de cette source de données massives, et notamment les tweets, se heurte à divers obstacles dont le *leitmotiv* reste la qualité et la crédibilité qu'on peut accorder à l'information disponible (Schade *et al*., 2013). Tout d'abord se pose la question légitime de l'objectivité de l'information : en effet, considérer l'individu comme capteur qui observe et perçoit son environnement peut représenter une prise de risques en termes de fiabilité, l'information pouvant être biaisée par l'état émotionnel de la personne ou d'autres facteurs ; certaines informations peuvent alors être subjectives. Le second obstacle provient du succès même et de la facilité d'accès des plateformes des médias sociaux : dans le cas précis de Twitter, le nombre d'informations – tweets – potentiellement exploitables ne cesse de croître ; à l'échelle globale, le nombre de tweets postés quotidiennement est ainsi estimé, en 2013, à 340 millions (Andrienko *et al*., 2013). Pour un chercheur travaillant sur un thème précis, les enjeux qui découlent de cette abondance d'informations sont donc l'élimination du bruit de fond[9] et la définition d'un protocole efficace d'extraction de l'information utile qui concerne une thématique spécifique.

En ce qui concerne Twitter, nous pouvons souligner trois contraintes essentielles :

- L'absence de formalisme dans la production de l'information

Contrairement à une donnée produite par un organisme expert, formalisée selon un certain nombre de normes et fournie en métadonnées assurant la qualité et la traçabilité de l'information (Goodchild & Glennon, 2010), le contenu partagé sur la plateforme de microblogage ne répond à aucune contrainte structurelle ou textuelle, d'où son contenu et sa forme hétérogènes ; à ce problème peuvent s'ajouter les effets de la limitation des 140 caractères, c'est-à-dire l'utilisation volontaire et fréquente d'un langage SMS et d'abréviations.

- Qui utilise Twitter ?

Il est également nécessaire de s'interroger sur la représentativité statistique des tweets comme source de données destinée à analyser des faits sociaux concernant des populations aux profils différents : en effet, les habitants susceptibles d'être affectés par une crue rapide ou une inondation ne disposent certainement pas tous d'un compte Twitter qu'ils utilisent régulièrement. De même, tous les habitants ne disposent pas d'une même qualité de connectivité à Internet ou à d'autres réseaux mobiles.

Enfin, le profil démographique des utilisateurs français[10] indique un total de 2,3 millions de comptes actifs pour un âge moyen estimé à 22 ans ; 61% des utilisateurs ont moins de 35 ans, la catégorie

[9] Tout tweet n'ayant aucun rapport avec le thème de recherche, ou dont le contenu est trop imprécis pour être validé

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

des 15-24 ans représente 51% des utilisateurs ; cependant, la fréquentation – visite de la plateforme, sans émission de tweet - des personnes de plus en 55 ans est en pleine progression (chiffres de 2013).

- Les tweets géolocalisés

La question des coordonnées géographiques introduit la problématique de la précision du géoréférencement et donc de la localisation de l'information. En effet, un utilisateur tweetant depuis son ordinateur n'est pas contraint de saisir une information géographique : il peut, s'il active la fonction de géolocalisation sur son compte, saisir manuellement un nom de ville. Au final, seuls les utilisateurs tweetant depuis leur Smartphone et ayant activé la géolocalisation offrent une information dont l'incertitude liée au géoréférencement dépend uniquement de la marge d'erreur du GPS de l'appareil.

Les études effectuées à différentes échelles spatiales révèlent cependant un très maigre pourcentage de tweets incluant des coordonnées géographiques, par rapport au flux global : celui-ci est généralement compris entre 1% et 2% (Brovelli *et al*., 2014).

- La question de l'incertitude

Par conséquent, l'information contenue dans les tweets, ou qui leur est simplement liée, peut révéler plusieurs degrés d'incertitude auxquels nous pouvons être confrontés. La taxonomie de cette incertitude se décline de la manière suivante (Shu *et al*., 2003) :

- l'incertitude géographique, c'est-à-dire l'incertitude liée à la localisation de l'entité ;
- l'inexactitude : l'information diffusée est inexacte voire fausse ;
- l'inconsistance : les propos sont incohérents entre eux ;
- l'imprécision : le propos est vague, ambigu et prête à confusion ;
- l'incomplétude : suite à une erreur de manipulation de l'utilisateur, le message peut être tronqué ou vide.

Dans le cas des tweets géoréférencés, l'incertitude géographique est liée à la précision du GPS du Smartphone et peut donc varier en fonction de la couverture des satellites et de la situation géographique de l'individu, notamment s'il se trouve dans une zone boisée ou en montagne. L'incertitude liée au contenu des tweets, que nous sommes susceptibles de rencontrer, peut relever de l'imprécision, lorsqu'un tweet évoque une situation qui peut être subséquente à une perturbation cévenole mais qu'il ne fournit pas suffisamment de détails pour être validé. Nous pouvons également rencontrer des propos inexacts si des personnes évoquent l'existence possible de victimes ou si elles sont mal informées de l'évolution de la situation.

1.2. Visualiser et extraire les tweets

Twitter dispose de plusieurs outils à partir desquels un internaute peut générer l'affichage de tweets correspondant à des critères de recherche, ainsi que d'une interface de programmation (API) permettant de collecter un jeu de tweets.

1.2.1. Les outils de visualisation

La cartographie interactive *The one million tweet map*[11] propose un suivi en temps réel des tweets géolocalisés ; elle se présente sous la forme d'un planisphère et de cercles dont la taille est proportionnelle au nombre de tweets géolocalisés émis pour une zone géographique précise (clusters).

[11] http://onemilliontweetmap.com/

L'utilisateur peut zoomer et filtrer l'affichage des tweets selon des mots-clés, des *hashtags* ou encore en fonction de l'heure (Figure 3).

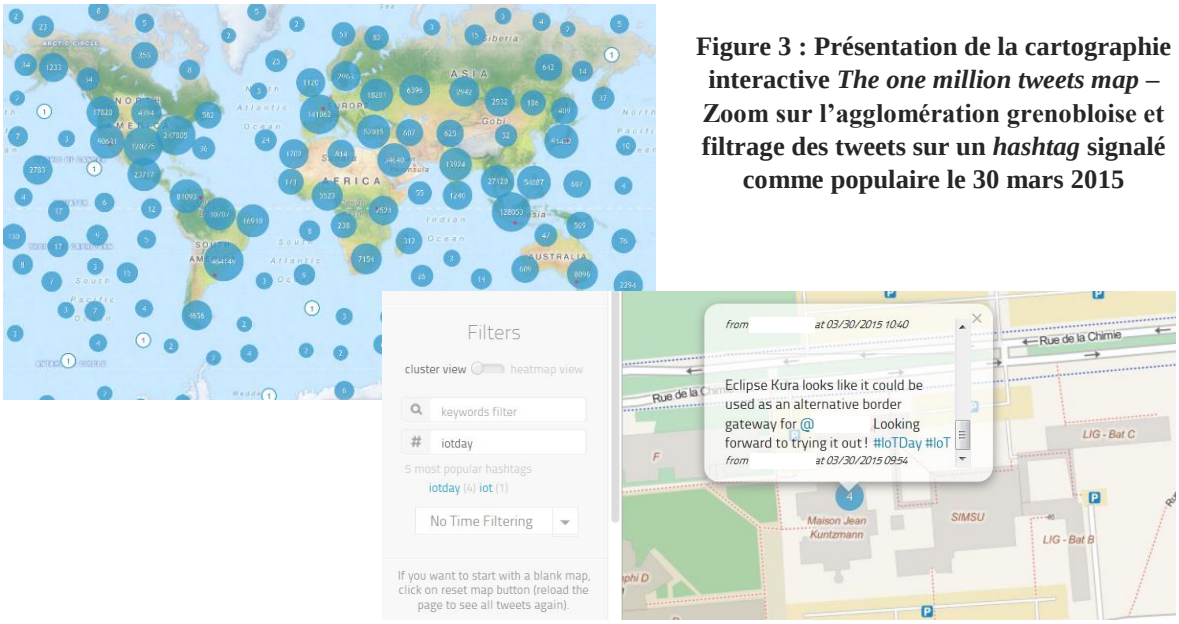


Figure 3 : Présentation de la cartographie interactive *The one million tweets map* – Zoom sur l'agglomération grenobloise et filtrage des tweets sur un *hashtag* signalé comme populaire le 30 mars 2015

Une requête manuelle peut également être effectuée à partir d'une interface plus complète de recherche avancée, accessible sur le site Internet de Twitter. Elle soumet à l'utilisateur davantage de critères de recherche, ne se restreint pas aux tweets géolocalisés et n'est pas limitée dans le temps (alors que l'affichage de la cartographie interactive précédemment mentionné est limité aux six dernières heures). Ces critères de recherche sont : textuels (choix de mots-clés ou de *hashtags*, choix de la langue), orientés sur les personnes (compte à l'origine du tweet, compte destinataire ou mentionné dans le tweet), et incluent la possibilité de sélectionner un lieu ainsi qu'une période temporelle. Une recherche de tweets contenant le mot-clé « inondation », postés à proximité de la ville d'Alès dans le Gard entre le 1er septembre et le 31 décembre 2014 donne le résultat ci-dessous (Figure 4).

Si l'utilisateur passe la souris sur la date (rectangle bleu), l'heure exacte de l'envoi s'affiche ; de même, en passant le pointeur sur l'élément encadré par le rectangle rouge, le nom de la ville où le tweet a été émis apparaît à son tour dans une infobulle.



Figure 4 : Exemple de recherche de tweets par mot-clé

1.2.2. L'outil d'acquisition de l'information brute

Twitter dispose de sa propre interface de programmation, l'*API Streaming* (Herfort *et al*., 2014), qui offre un accès à l'information publique (les tweets) archivée[12]. L'exploitation de cette interface requiert une étape d'implémentation, c'est-à-dire que l'utilisateur oriente son développement en fonction de ses besoins ; il doit configurer un total de onze paramètres pour définir les modalités d'acquisition et de stockage du jeu de données entre son ordinateur, le serveur et Twitter. Les principaux paramètres de l'API sont :

- `filter_level` : sélectionne les tweets à transmettre en fonction d'un nombre minimum d'attributs (date, heure, nom de l'utilisateur, lieu, coordonnées géographiques, nombre de retweets, de favoris, etc.)
- `language` : choisit la langue
- `track` : définit les modalités d'acquisition du texte du tweet, en particulier pour les caractères spéciaux, les URL, les noms de destinataires précédés de « @ » et les *hashtags*
- `location` : l'utilisateur saisit des coordonnées géographiques pour constituer une *bounding box*[13] ; seuls les tweets géolocalisés et postés dans la zone sélectionnée seront transmis au serveur.

1.3. Sélectionner et structurer l'information liée à un événement particulier

Dans la communauté scientifique, l'information *crowdsource*, qu'elle soit issue de réseaux sociaux (Facebook), de services de microblogage (Twitter), ou encore de plateformes de stockage de photographies ou de vidéos (Flickr et YouTube), peut être exploitée dans deux situations différentes. En cas de crise, l'événement, la collecte et l'exploitation de l'information sont simultanés, de manière à organiser l'aide aux sinistrés selon la gravité des dégâts. En situation de retour d'expérience, les méthodes cherchent davantage à étudier les interactions entre populations affectées et aléas naturels. Nous ne sommes, par conséquent, pas confrontés aux mêmes enjeux qu'en situation de crise. Néanmoins, dans les deux cas, l'objectif consiste à filtrer l'information en fonction de son rattachement à l'événement et de sa qualité.

1.3.1. Le *crowdsourcing* et la gestion de crise

En période de crise, l'exploitation de l'information *crowdsource* nécessite la mobilisation constante d'équipes d'informaticiens qui collectent, filtrent et organisent cette information pour la représenter sur des cartographies interactives, les *crowdmaps*, mises en ligne via des plateformes de traitement, comme *Ushahidi* (McDougall, 2012). Créée en 2007 suite aux violences post-électorales au Kenya, cette plateforme de collecte s'appuie sur un serveur *open-source* permettant de maîtriser l'afflux de données en période de crise, de filtrer et de vérifier la véracité de l'information citoyenne participative émise via les services de microblogage, les SMS ou encore les flux RSS.

Son exploitation pendant les crises majeures récentes – séisme d'Haïti en 2010 et catastrophe de Fukushima en 2011 - a notamment rendu possible l'existence d'un partage d'informations en double sens : *bottom-up*, lorsque les habitants diffusent l'information aux autorités via la plateforme afin que celles-ci organisent les secours en fonction de l'urgence ; *top-down*, qui consiste pour les autorités, à informer les sinistrés via les cartographies interactives, les réseaux sociaux ou services de microblogage,

[12] https://dev.twitter.com/streaming/overview
[13] Rectangle ou polygone dans lequel les tweets sont extraits

au sujet des différents postes où ils pourront accéder au ravitaillement, à l'eau potable ou encore aux soins médicaux (McDougall, 2012).

Plus récemment, après la série de séismes qui ont frappé le Népal sur les mois d'avril et de mai 2015, l'ICCROM[14] a mis en ligne une *crowdmap*[15] sur laquelle est répertoriée, sous forme de clusters, toute l'information collectée concernant les dommages subis par les bâtiments et monuments patrimoniaux (Figure 5) :

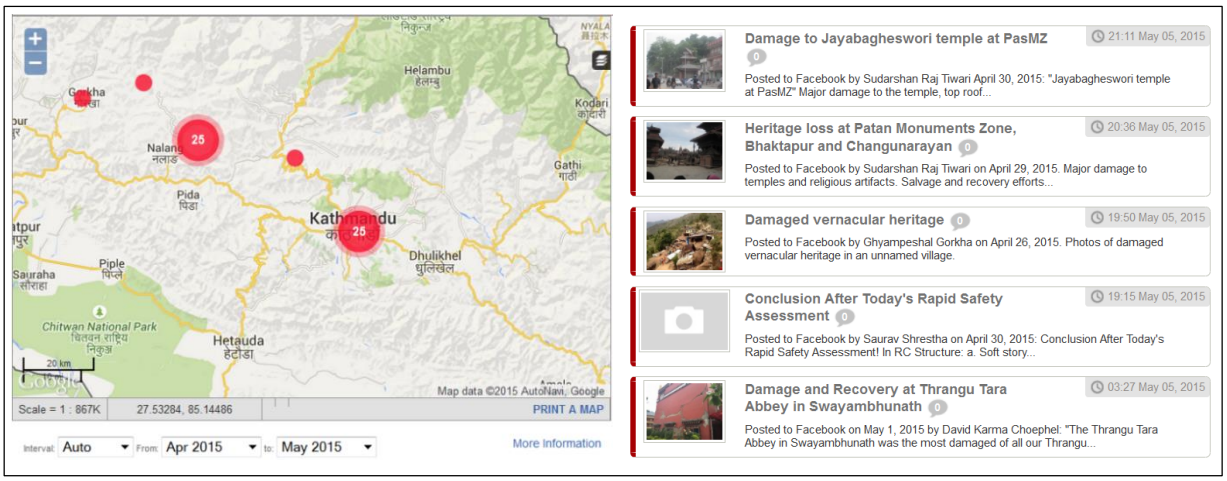


Figure 5 : *Crowdmap* et information envoyée à l'ICCROM pour l'inventaire des dégâts affectant le patrimoine népalais *(ICCROM)*

1.3.2. Le cas de Twitter : extraire un corpus de tweets pertinents

En situation d'analyse post-crise, nous cherchons des méthodes permettant de construire un jeu de données élaborées afin de transformer l'information brute en données exploitables. Ces méthodes consistent à effectuer une segmentation des tweets fondée sur l'examen de leur contenu lexical, c'est-à-dire déterminer si un tweet est lié ou non à l'événement en question (Herfort *et al*., 2014).

- L'extraction lexicale

L'analyse lexicale peut reposer sur deux approches différentes : une approche automatique (qualifiée de *data-driven retrieval* dans la littérature anglophone), ne suppose aucune intervention de l'utilisateur : on utilise un logiciel de fouille de données qui analyse et trie le vocabulaire contenu dans un corpus de texte (Imran *et al*., 2013).

L'approche supervisée est, quant à elle, dirigée par les connaissances de l'utilisateur (Roy Chowdhury *et al.*, 2013). Tout d'abord, celui-ci établit un glossaire de mots-clés répertoriant un vocabulaire susceptible d'être employé par les utilisateurs de Twitter qui produisent l'information recherchée ; ce glossaire peut être constitué dans différentes langues, en utilisant des supports particuliers de recherche d'un vocabulaire simple et spécifique à l'événement. La constitution de cette liste peut donc, d'une part, s'appuyer sur l'utilisation du dictionnaire, au travers des définitions des mots essentiels (comme inondation ou crue) afin de trouver le vocabulaire associé à ces mots, et, d'autre part, sur le vocabulaire relayé par les médias (Herfort *et al*., 2014).

[14] International Center for the Study of the Preservation and Restoration of Cultural Property
[15] https://kathmanduculturalemergency.crowdmap.com/main

Dans les deux cas, la première partition est suivie par la définition de règles de formalisation qui permettent de discriminer les tweets considérés comme valables, du bruit de fond éventuellement extrait à l'étape précédente (Roy Chowdhury *et al.*, 2013) : nous pouvons en effet rencontrer des tweets qualifiés de faux-négatifs, c'est-à-dire des tweets utilisant le ou les termes cibles mais sans rapport avec le thème recherché ainsi que des faux-positifs, qui correspondent à des tweets imprécis dont l'information est ambiguë, et dont nous ne pouvons pas valider la conformité (Goodchild & Glennon, 2010).

Ces tâches peuvent être sous-traitées via la plateforme *Crowdflower* qui offre un service de nettoyage des jeux de données *crowdsource* incomplets ou bruités (Imran *et al*., 2013) : les étapes préalables consistent, pour les chercheurs, à fournir le jeu contenant les informations collectées ainsi que des instructions précises qui indiquent aux collaborateurs les règles d'extraction de l'information adéquate. Dans le cas précis des tweets, il convient d'établir une première liste d'exemples de tweets ciblant les messages dont le rapport avec un événement est explicite, puis de créer un ensemble de questions auxquelles chaque tweet doit répondre pour être validé (Imran *et al*., 2013).

Dans notre cas, nous pourrions ainsi soumettre les tweets à au moins trois questions successives : le tweet est-il émis sur une journée de crise hydrométéorologique ? Le tweet fait-il mention d'une perturbation météorologique ou d'une crise hydrologique qui se déroule au moment où l'utilisateur émet son message ? Le tweet contient-il une information liée aux conséquences de l'événement sur des personnes, leurs activités, des biens ou des infrastructures ? Prenons l'exemple d'un tweet comme « *Enorme orage sur Nîmes* » émis le 28 novembre 2014 (date qui correspond à une journée de crise) ; ce tweet satisfait les exigences des deux premières questions, mais pas de la dernière : il n'est donc pas conservé dans le corpus final. En revanche, un tweet comme « *Gros orage sur #Grabels, 34. Eclairs en continu. Quartier Prédimau-Montalet complètement disjoncté, élec coupée* » émis à la même date, sera conservé : la date d'émission correspond à une journée de crise, nous avons de plus une information concernant la météo et une information sur les conséquences de cette perturbation. Cette méthode est donc très sélective et ne conserve que les informations les plus précises.

- La question de la pertinence des tweets

Cependant, cette méthode d'extraction par mots-clés ne s'avère que partiellement satisfaisante, en particulier si l'on souhaite extraire les tweets les plus pertinents en diminuant la part des différents degrés d'incertitude évoqués précédemment. En outre, elle repose uniquement sur une analyse textuelle et est, par conséquent, déterritorialisée, c'est-à-dire qu'elle dissocie l'information géographique du territoire et des conditions environnementales qui conditionnent sa création (Herfort *et al*., 2014). Ces auteurs proposent ainsi une méthode d'extraction multicritère en couplant l'information extraite en fonction du vocabulaire contenu dans les tweets et la géographie du phénomène étudié.

Cette méthode, testée sur les inondations de la vallée de l'Elbe en Allemagne en juin 2013 (Herfort *et al*., 2014), consiste à superposer les tweets à diverses couches d'information géographique : modèles numériques de terrain, zones inondables ou encore images aériennes capturées pendant les différentes phases de l'événement. L'extraction de l'information considérée comme pertinente repose ensuite sur deux postulats : « la première loi de la géographie » de Tobler, citée dans Herfort *et al*., qui implique alors que plus le tweet est proche d'un espace affecté par l'événement, plus il y a de chances qu'il soit lié à l'événement et que son contenu soit riche en information utile, ainsi que le postulat selon lequel les tweets sont émis par des locaux sur un événement local. En conséquence, les tweets liés à l'événement et considérés comme pertinents sont extraits sur un critère de proximité des zones inondées.

Les méthodes de tri des tweets en fonction de la pertinence et de la qualité de l'information sont encore peu définies et développées (Herfort *et al*., 2014). Cependant, les jeux de tweets peuvent être hiérarchisés, par examen lexical, en fonction des critères suivants : la présence de noms de lieux précis ou de *hashtags* permettant de rattacher le tweet à un ou plusieurs thèmes, l'insertion d'URL renvoyant

à des photographies, vidéos, reportages, articles en ligne, etc., permettant de vérifier la cohérence des propos (Herfort *et al*., 2014). Cette hiérarchisation peut plus simplement être établie en fonction du nombre de mots-clés présents dans chaque tweet ou encore du nombre d'informations différentes et partagées dans chaque message, en admettant l'hypothèse que plus le tweet contient d'informations, plus il est pertinent (Dashti *et al*., 2014).

1.3.3. Classification thématique d'un jeu de tweets

Cette étape de prétraitement est destinée à structurer l'information lexicale des tweets en thèmes particuliers. En fonction des objectifs, deux principaux types de classification thématique peuvent être envisagés, fondées sur la temporalité de la crise ou sur les thèmes précis évoqués par les tweets.

- Les différentes phases de la crise

L'étude des effets de l'annonce et de l'arrivée d'un événement sur les populations peut être fondée sur une classification des tweets en fonction de la temporalité de la crise à laquelle ils se rattachent (Roy Chowdhury *et al.*, 2013) : les tweets peuvent être ainsi annotés en fonction des périodes antérieures à la crise, simultanées et postérieures à la crise. La détermination s'effectue en fonction de l'examen de critères précis : des noms comme « vigilance » ou « alerte » peuvent ainsi se retrouver typiquement dans les tweets postés avant la crise ; il est néanmoins indispensable d'examiner la présence d'adverbes de temps et d'étudier la conjugaison des verbes : une personne peut en effet évoquer au passé une période simultanée à la crise alors que celle-ci est terminée.

- Le thème de l'information

Si l'on souhaite étudier les variations spatio-temporelles du contenu précis des tweets, il est possible d'effectuer une série de classements progressifs par thèmes. Imran *et al*. (2013) proposent une ontologie[16] automatisée fondée sur deux classifications successives. La première étape consiste à évaluer la source et le (ou les) destinataire(s) sur le thème de la connaissance de l'événement. Les tweets sont annotés en fonction de cinq critères :

- *Personal Only* : le tweet ne concerne que l'auteur et son entourage immédiat ; il n'a aucun intérêt pour des personnes extérieures ;
- *Direct Informative* : le tweet peut concerner des personnes extérieures à l'entourage de l'auteur, et celui-ci tweete sur un événement auquel il assiste ;
- *Indirect Informative* : le tweet peut concerner des personnes extérieures à l'entourage de l'auteur qui relaie une information obtenue via divers médias (l'auteur doit spécifier la source)
- *Informative* : le tweet peut concerner des personnes extérieures à l'entourage de l'auteur mais l'information est insuffisante pour déterminer les conditions de sa production.
- *Other* : le message ne peut être classé.

La seconde partition est établie à partir des cinq critères suivants :

- *Caution and advice* : le tweet relaie une information liée au déclenchement d'une alerte, ou à un risque ;
- *Casualties and damage* : le tweet évoque les victimes ou dommages liés à un événement ;
- *Donations of money or goods* : le tweet évoque une campagne de dons ;

[16] En informatique, outil qui permet de représenter un corpus de connaissances structurées en concepts organisés dans un graphique qui indique les relations existantes entre les différents concepts, ainsi que leur type (sémantique, héritage, etc.).

- *People missing, found or seen* : le tweet évoque une personne disparue, retrouvée ou encore la visite d'une personnalité politique après l'événement ;
- *Information source* : tout tweet contenant une adresse URL faisant référence à une photographie, une vidéo ou tout autre média (radio, télévision, journal, etc.)

Cette première partie a décrit le cadre général dans lequel s'inscrit le mémoire. Nous avons mis en exergue les difficultés liées aux caractéristiques des informations issues des médias sociaux, en particulier l'information provenant de Twitter qui est massive, hétérogène en forme et en qualité de contenu. Nous nous sommes alors intéressés à la recherche de références pour extraire et organiser un jeu de tweets bruts en données élaborées. Les principales références sur ce thème sont anglophones. Elles présentent des méthodes qui visent à diriger l'extraction des tweets en fonction du vocabulaire contenu dans les messages. Le travail qui sera mis en œuvre dans le cadre de ce stage pourra être complété par l'accomplissement des étapes de détermination de la pertinence de l'information et de sa classification thématique.

2. Du tweet à la carte : extraire, structurer et représenter l'information

Cette deuxième partie s'articule autour de la présentation des différentes données acquises auprès du LIG et du LTHE, et de la description des méthodes mises en œuvre destinées à : dans un premier temps, extraire et structurer un corpus de tweets dont le contenu lexical est en rapport avec l'événementiel hydrométéorologique et, dans un second temps, analyser le contenu de cette information afin de la quantifier et d'en mesurer les variations spatio-temporelles. L'objectif principal consiste donc à définir une méthodologie d'analyse permettant d'une part, d'extraire des tweets au travers de l'examen de leur contenu et, d'autre part, de représenter la variabilité spatio-temporelle de l'émission et du contenu des tweets, de manière à pouvoir ensuite exploiter conjointement l'information issue des tweets et la temporalité de l'événement.

2.1. Présentation des données exploitées dans le cadre du projet

Nous disposons de deux sources principales de données : les jeux de tweets issus d'une infrastructure de collecte et les données permettant de caractériser les hauteurs de précipitation ainsi que la sévérité des crises hydrométéorologiques.

2.1.1. Infrastructure de collecte et tweets

Le jeu de tweets bruts sur lequel les analyses reposent est issu de la base de données `twitterdb` de l'équipe SLIDE du LIG, mise en place pour les besoins du projet *Crowdhealth*[17] ; les tweets sont collectés, traités et stockés sur un serveur. L'infrastructure de collecte des tweets est organisée de la manière suivante (Figure 6) :

- les tweets sont collectés via un serveur connecté à l'API Streaming de Twitter : ce serveur a été configuré pour capter, de façon continue et en temps réel, le flux de tweets géolocalisés émis dans le monde entier ;
- les tweets subissent un premier filtrage : les retweets, c'est-à-dire les messages postés une première fois par un utilisateur et rediffusés par ses abonnés, sont automatiquement supprimés ;
- les tweets sont stockés sous la base de données PostGreSQL `twitterdb`, à partir de laquelle il est possible d'extraire un jeu de tweets sur une période et une zone géographique précises.

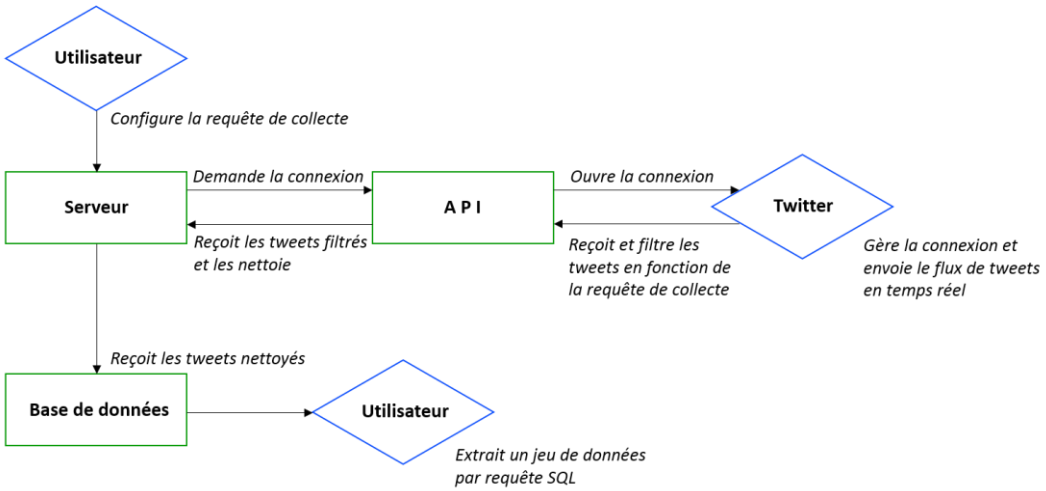


Figure 6 : Schématisation de l'infrastructure de collecte des tweets

[17] http://slide-apps.liglab.fr/~crowdhealth/

Le serveur a été mis en service le 8 octobre 2014 ; cependant, il s'est déconnecté de l'API le 12 octobre 2014 - par conséquent, nous ne disposons d'aucun tweet pour cette date - puis a été volontairement arrêté du 12 décembre 2014 au 5 janvier 2015. Il est également nécessaire de préciser les limites de cette collecte en ce qui concerne les tweets et échelles géographiques : tout d'abord, nous rappelons que les tweets géolocalisés ne représentent que 1% à 2% du flux mondial de tweets émis quotidiennement. Il s'agit en effet des tweets postés depuis un smartphone si l'utilisateur a activé la fonction de géolocalisation. Ensuite, Twitter a limité la collecte gratuite des tweets via l'API à 15% du flux émis sur la zone géographique demandée, dans les premiers mois. A partir de mars 2015, le serveur de l'équipe étant toujours en service, Twitter a baissé ce flux entrant à 7% des émissions totales. Si l'on souhaite alors compléter les jeux de données, les tweets non collectés peuvent être achetés auprès de l'entreprise.

Ce plafonnement a deux conséquences principales sur les jeux de tweets ; tout d'abord, nous n'obtenons pas la totalité des tweets géoréférencés émis sur notre terrain d'étude ; ensuite, si l'on raisonne en termes d'échelle spatiale, nous pouvons déduire qu'une collecte organisée à l'échelle mondiale restreint davantage les informations dont nous pouvons potentiellement disposer sur une échelle spatiale plus réduite comme c'est le cas de ce projet qui concerne huit départements français seulement.

La base de données `twitterdb`, dans laquelle est stockée l'information traitée par le serveur, est structurée en cinq tables portant sur les thèmes suivants : tweets (table `public.tweet`), utilisateurs, statistiques d'utilisation, langues et pays ; dans le cadre de l'extraction d'un jeu de tweets, seule la table `public.tweet` est utilisée. Cette table est structurée selon le modèle indiqué sur la figure 7.

Depuis cette table est extrait un jeu de tweets comportant les champs suivants :

- `tweet_id` : numéro d'identification du tweet
- `user_id` : numéro d'identification de l'utilisateur
- `text` : chaîne de caractères qui correspond au message de l'utilisateur
- `created at` : date et heure d'émission du tweet au format année/mois/jour 00 :00 :00 GMT
- `gps` : coordonnées gps du tweet, au format WKT, c'est-à-dire POINT(X Y)

```
                Table "public.tweet"
    Column      |           Type           | Modifiers
----------------+--------------------------+-----------
 tweet_id       | bigint                   | not null
 user_id        | bigint                   |
 text           | text                     |
 favourite_count | integer                  |
 retweet_count  | integer                  |
 created_at     | timestamp with time zone |
 captured_at    | timestamp with time zone |
 user_tweet_num | integer                  |
 saved_at       | timestamp with time zone |
 gps            | geometry(Point,4326)     |
```

Figure 7 : Schéma de la table `public.tweet` *(LIG, SLIDE)*

Les tweets sont extraits dans une zone géographique suffisamment large afin d'englober la totalité de la superficie des huit départements choisis (Ardèche, Drôme, Lozère, Gard, Hérault, Var, Vaucluse et Bouches-du-Rhône) ; les coordonnées des points encadrant la zone sont récupérées sous le logiciel Google Earth (Annexe 1).

Notre jeu de tweets bruts est ainsi extrait sur la période du 8 octobre 2014 au 1er avril 2015 ; exporté en format CSV, il enregistre un total de 1 322 212 lignes, correspondant à autant de tweets différents contenant les informations pour chacun des cinq champs mentionnés ci-avant. Nous élargissons volontairement la plage temporelle d'extraction des tweets afin de comparer l'automne avec des périodes marquées par d'autres événements et des périodes normales.

2.1.2. Les crises hydrométéorologiques

La seconde source de données concerne les crises hydrométéorologiques. Elle se présente sous deux formes principales :

- Les cumuls pluviométriques

Les cumuls pluviométriques sont fournis par l'OHMCV[18] et sont disponibles sous format raster, en fichier .asc, exportable sous un SIG ; une dalle est produite pour chaque événement et couvre une zone géographique précise (Figure 8). Pour l'instant, nous n'avons obtenu qu'une seule dalle raster, qui couvre la crise du 8 (16 heures TU) au 11 (6 heures TU) octobre 2014.

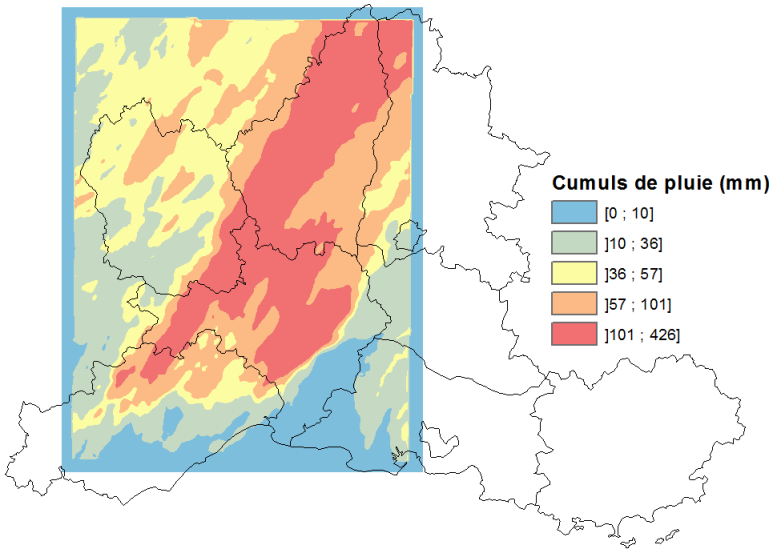


Figure 8 : Raster du cumul de pluie sur un événement *(OHMCV, LTHE)*

- Les caractéristiques des crises hydrométéorologiques

Les caractéristiques et la sévérité des crises sont identifiées à partir de données fournies par le LTHE, issues de la base de données CatNat[19] qui collecte et archive, depuis 2001, toute l'information relative aux événements dommageables ayant pour origine divers aléas naturels – inondations, mouvements de terrain, séismes, incendies, tempêtes, etc. – survenus en France et dans le reste du monde. L'information proposée, fournie en format CSV, se présente sous forme de dix-huit champs dont les principaux sont : les dates de début et de fin d'événement, le type de risque (sélectionné par l'utilisateur : dans notre cas, il s'agit du risque d'inondation), la nature de l'événement, sa localisation,

[18] http://www.ohmcv.fr/
[19] http://www.catnat.net/donneesstats/bd-catnat?view=catalogue

les nombres respectifs de personnes évacuées, blessées ou décédées pendant l'événement, le coût des dommages et une échelle de gravité (numérotée de 1 à 5 pour les crises les plus dommageables).

Ce jeu de données permet en premier lieu d'identifier les dates des épisodes cévenols de l'automne 2014 ayant affecté les départements du sud-est de la France et pour lesquels nous disposons de tweets : il s'agit des épisodes du 9 au 13 octobre, des 3 et 4 novembre, du 9 au 11 novembre, des 14 et 15 novembre, du 24 au 27 novembre, et enfin, du 27 au 30 novembre. En outre, il permet de repérer les informations susceptibles d'être retrouvées dans les tweets comme le nombre de victimes, les évacuations, le coût des dégâts ou les effets des précipitations (crues, ruissellement urbain, etc.). Un extrait de ces données, sur les champs principaux, est disponible en annexe 2.

2.1.3. Autres données géographiques

D'autres données sont également utilisées pour l'extraction et l'analyse spatiale, pour l'étude de la répartition de la population ou encore pour constituer les fonds de carte. Il s'agit principalement de :

- Données vectorielles et matricielles

Les données vectorielles sont issues, d'une part, de la BD GEOFLA® de l'IGN, qui fournit les communes et départements ; les cours d'eau et autres surfaces hydrographiques proviennent, quant à elles, de la BD CARTHAGE®, également diffusée par l'IGN. Les autres données raster concernent des modèles numériques de terrain en dalles SRTM, d'une résolution spatiale de 90 mètres, qui sont téléchargées depuis le site du CGIAR-CSI[20] pour constituer des fonds de carte.

- Données carroyées sur la population de l'INSEE : résolution à 200 mètres

Le carroyage INSEE retenu propose un découpage du territoire français métropolitain en carrés de 200 mètres de côté ; à chaque carré est associé le nombre d'habitants recensés à l'intérieur (chiffres de 2010). Cependant, seules les entités géographiques contenant au moins onze ménages sont prises en compte. Les carreaux comptabilisant des effectifs trop faibles sont agglomérés en rectangles de taille plus importante. La méthodologie d'acquisition et de construction des données ainsi que la procédure détaillée permettant à l'utilisateur d'obtenir l'effectif total de la population, en répartissant les habitants localisés dans les rectangles sur chaque carreau, sont consultables et téléchargeables sur le site de l'INSEE[21].

Le carroyage nous permet d'étudier les densités de population, pour les mettre en relation avec les densités de tweets, sur une échelle plus fine que celle de la ville et de son nombre global d'habitants ; en revanche, ces données offrent une vision générale de la répartition de la population car les foyers fiscaux non soumis à la taxe d'habitation ne sont pas référencés et elles ne prennent pas en compte les mobilités susceptibles d'exister au sein d'un même foyer fiscal (les étudiants sont, par exemple, localisés au domicile de leurs parents).

2.2. De l'information brute à la constitution d'un corpus de tweets

Les premières étapes de traitement mises en œuvre après l'acquisition du jeu de tweets issu de la base de données `twitterdb`, présentée au point 2.1.1., requièrent l'utilisation de divers outils d'organisation et de traitement de l'information.

[20] http://srtm.csi.cgiar.org/SELECTION/inputCoord.asp
[21] http://www.insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=donnees-carroyees&page=donnees-detaillees/donnees-carroyees/donnees_carroyees_diffusion.htm

2.2.1. La base de données, un outil de nettoyage et de structuration

La première étape consiste donc à créer une nouvelle base de données spatiale, indépendante de `twitterdb`, en utilisant le logiciel PostGreSQL et l'extension spatiale PostGIS[22]. L'objectif de cette nouvelle base de données `Twitter` est de préparer, à partir du fichier CSV extrait de la base de données `twitterdb`, une table nettoyée et structurée qui constituera la future couche des tweets géolocalisés et intégrée dans un SIG. Cette table servira également de référence sur laquelle seront effectuées les différentes requêtes SQL destinées à extraire l'information nécessaire à la mise en œuvre des analyses statistiques.

Les différentes étapes mises en œuvre consistent donc à importer le fichier CSV dans une table, structurer cette table, puis créer la couche d'information géographique en établissant une connexion de la base de données vers un SIG. La procédure technique appliquée est décrite en annexe 3.

2.2.2. Construire un corpus de tweets relatifs à l'événementiel hydrométéorologique

La seconde phase de prétraitement de l'information consiste à extraire un jeu de tweets dont l'information textuelle est liée aux perturbations et crises hydrométéorologiques. Nous choisissons d'explorer deux procédés complémentaires, fondés sur la recherche de mots-clés contenus dans le texte des tweets ; la première méthode relève d'une expertise alors que la seconde s'appuie sur une approche logicielle. L'objectif est d'établir une liste précise de tous les mots que nous pourrons rechercher et trouver dans les tweets.

- L'extraction expertisée

L'extraction expertisée consiste à établir une bibliothèque de mots-clés liés aux types d'événements et d'aléas à partir d'un vocabulaire potentiellement employé dans les tweets. L'information envoyée par un utilisateur sur un événement peut en effet être très hétérogène, celui-ci pouvant annoncer, dans son tweet, la simple arrivée de la pluie jusqu'à l'inondation de sa maison ou de son quartier. Il est donc nécessaire de s'informer sur plusieurs critères : tout d'abord, le nom des cours d'eau concernés par les crues cévenoles ; le nom des villes, voire des quartiers, fréquemment soumis aux risques de ruissellement urbain et d'inondation, au travers notamment des PPRI ; le vocabulaire employé par les médias et bulletins météorologiques, qui peut être relayé par les utilisateurs dans leurs tweets. Enfin, une réflexion sur les termes du langage courant employés et les types d'informations diffusées par les utilisateurs sur l'événement doit être menée. Un utilisateur peut en effet mentionner une route coupée, une maison privée d'électricité sans pour autant parler d'inondation, d'orage ou encore d'intempérie. De même, un utilisateur peut évoquer un cours d'eau qui déborde au lieu d'employer le terme de crue.

La bibliothèque ainsi constituée contient, à ce stade, une base de quatre-vingts mots-clés, classés en fonction des cours d'eau, des villes, de la météo, et des différentes périodes de crise (Annexe 4) ; par exemple, nous avons inscrit les mots « *Orb* », « *Alès* », « *Orage* », et « *Crue* », qui s'intègrent dans les catégories énumérées ci-avant.

- L'extraction automatisée

Cette seconde méthode consiste à analyser, sans intervention de connaissance sur les crises, le contenu lexical des tweets au travers d'un logiciel de fouille de texte, afin de trouver et de repérer le vocabulaire fréquemment employé par les utilisateurs lorsqu'ils évoquent les perturbations hydrométéorologiques. Ce type d'analyse permettra d'une part, d'enrichir la liste des mots-clés

[22] http://www.postgresql.org/

précédemment établie ainsi que de valider leur usage et, d'autre part, de vérifier l'orthographe des utilisateurs sur certains mots.

Le logiciel de fouille de texte utilisé pour la réalisation de cette étape est KH Coder[23] ; à partir d'un corpus de texte importé sous différents formats (XLS, CSV, TXT, etc.), celui-ci peut accomplir diverses tâches d'analyse lexicale dont les principales sont : le calcul de l'occurrence de chaque mot rencontré, la production de diagrammes de cooccurrence (Figure 9) et de classifications ascendantes hiérarchiques ; ces deux derniers types de rendu permettent également de visualiser les relations entre les mots les plus fréquents. Dans les deux cas – diagrammes ou classifications – l'utilisateur paramètre deux seuils : un seuil minimal et un seuil maximal de cooccurrence (nombre de fois où deux mots apparaissent ensemble). La création de diagrammes nécessite également de définir le nombre de cooccurrences qui seront affichées à l'écran, alors que la classification n'est pas limitée.

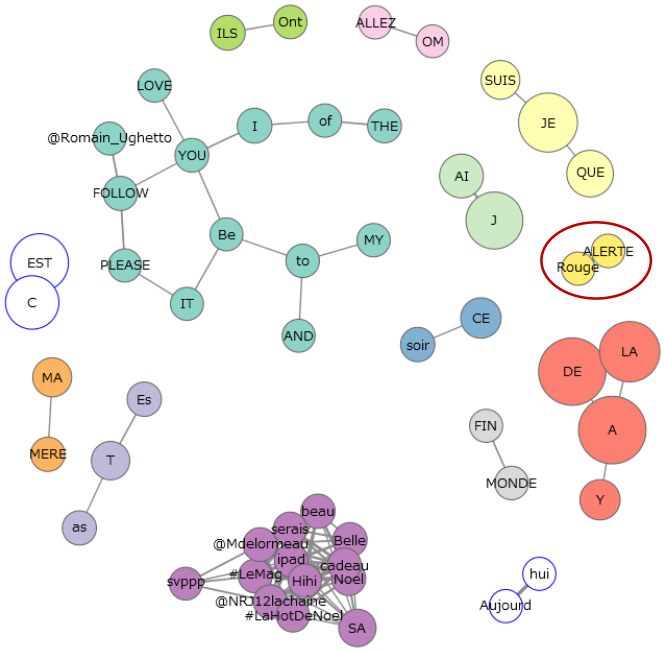


Figure 9 : Diagramme cartographique pour la journée du 28 novembre 2014 (seuils de cooccurrence compris entre 8 et 10)

Ce diagramme met ainsi en évidence plusieurs phénomènes : certains utilisateurs évoquent Noël (violet), d'autres un match de foot (rose) ; l'événementiel hydrométéorologique de la journée ressort au travers de la cooccurrence des mots alerte et rouge (et éventuellement fin et monde). On distingue également un groupe de mots qui peut s'apparenter à un spam, provenant d'un utilisateur anglophone qui demande à ce qu'on s'abonne à son compte.

Il sera alors nécessaire, pour l'efficacité de cette analyse, de choisir une période temporelle qui permette la mise en évidence d'un vocabulaire lié aux perturbations : des essais peuvent ainsi être menés sur un mois entier de l'automne, sur une période de crise complète (cf. dates indiquées au point 2.1.2), voire sur une seule journée si les résultats ne sont pas satisfaisants.

Les opérations permettant d'extraire un jeu de tweets, depuis la base de données, émis sur des dates précises, sont détaillées en annexe 5.

- Extraire les tweets

L'extraction est réalisée sous QGIS par requêtes SQL : nous recherchons en effet, depuis la table attributaire, les lignes dont le champ `text` contient au moins l'un des mots-clés de la liste dressée à

[23] http://khc.sourceforge.net/en/

partir de connaissances acquises et complétée par l'analyse de fouille de texte ; depuis le menu « *Propriétés* » de QGIS, les requêtes sont construites de la manière suivante : `« text » LIKE '%vigilance%'`. L'extraction par requête doit néanmoins, pour être complète, prendre en compte trois facteurs essentiels (Tableau 1) :

Fautes d'orthographe courantes
`« text » LIKE '%vigilance%' OR « text » LIKE '%vigilence%'`
Omission des accents
`« text » LIKE '%déluge%' OR « text » LIKE '%deluge%'`
Mots susceptibles d'être employés comme noms, verbes ou adjectifs
`« text » LIKE '%inond%'OR « text » LIKE '%innond%'`

Tableau 1 : Règles de syntaxe des requêtes d'extraction des tweets

Notons que les mots séparés par le signe « + » sur l'annexe 4 sont recherchés ensemble ; l'opérateur utilisé est alors `AND` : `« text » LIKE '%eau%' AND « text » LIKE '%niveau%'`. De même, les mots séparés par des espaces sont recherchés indépendamment sur le plan lexical mais regroupés en une seule sélection. L'opérateur utilisé est `OR` : `« text » LIKE '%route%' OR « text » LIKE '%pont%'`.

A chaque sélection, une nouvelle couche shapefile est exportée à partir des entités sélectionnées et est nommée en fonction du ou des mots-clés auxquels elle correspond.

- L'élimination du bruit

La dernière étape, qui consiste à éliminer le bruit, est effectuée manuellement. Le bruit est défini comme tout tweet extrait lors de l'étape précédente, mais qui n'a aucun rapport avec les événements hydrométéorologiques réels (Figure 10). Ces tweets sont extraits pour deux raisons principales : le mot-clé à partir duquel ils ont été recherchés est très général (maison, panne, urgence, etc.) et peut concerner des situations très variées ; le mot-clé est contenu dans un autre mot (crue est contenu dans cruel, éclair dans éclairer ou éclairage etc.).

Fausse alerte incendie. Finale du double néanmoins interrompue... http://t.co/zARRTPXeSR
Il y aurait bien eu l'entrainement pour me relaxer et évacuer tout le reste, malheureusement je suis blessé.
En plus la au lieu d'être au chaud chez moi je subit les courants d'air des cruelles salles de fac..
Und jetzt singen wir ... " Sur le pont ... on y danse ... "

Figure 10 : Information bruitée extraite

L'élimination de ces enregistrements nécessite au préalable la définition de règles qui posent un cadre dont l'objectif est de nous aider à établir une distinction entre l'information considérée comme valable et l'information bruitée à écarter du jeu de tweets final. Est ainsi considéré comme information utile :

- tout tweet en rapport avec un événement réel annoncé et attendu : *« Vigilance orange en Ardèche. Les Cévennes pourraient être touchées […] »*
- tout tweet en rapport avec un événement réel qui se déroule au moment où l'utilisateur émet son message : *« L'apocalypse ! L'eau monte. Aux cabanes de l'Arnel, des maisons sont isolées […] »*

- tout tweet en rapport avec un événement réel qui s'est déroulé et auquel l'utilisateur a assisté : *« On a eu des inondations y'a une semaine. »*
- tout tweet en rapport avec les conséquences d'un événement sur des personnes, des biens ou des infrastructures : *« Route d'entraînement inondée à Alès […] »*
- tout tweet traduisant la préoccupation, l'inquiétude ou l'appréhension d'une personne vis-à-vis d'un événement : *« […] j'avais pas eu d'eau, je crains pour les dégâts en zone inondable, j'espère que ça va aller », « Il y a des trams ? j'ai pas envie de rester bloqué si c'est inondé »*
- tout tweet évoquant le comportement d'une personne face à un événement risquant d'affecter ses activités : *« Il fait nuit, il pleut, c'est l'alerte orange mais je fais les magasins », « J'ai préféré rester prudent ce soir avec l'alerte météo en cours et ne pas aller au concert […]*
- tout tweet traduisant la lassitude d'une personne face à la répétitivité des événements : *« La pluie j'en ai marre, alerte rouge tous les deux jours […] »*
- tout tweet émettant un jugement sur l'annonce et la gestion de crise : *« […] merci pour cette alerte rouge qui n'a servi à rien. Pas une goutte de pluie […]*

Les tweets à caractère subjectif évoquant une situation irréelle sont considérés comme bruit et sont par conséquent supprimés : *« Petite alerte rouge pour demain, ça passerait bien »* ; de même, les messages incompréhensibles dus à un recours abusif du langage SMS sont écartés *: « waaaa comen i pleu c 1 truc de ouf mdr miskin lé gen si c re inonder »*.

La génération de la couche finale des tweets liés à l'événementiel hydrométéorologique nécessite encore l'application de deux outils de traitement SIG : tout d'abord, les shapefiles dont le bruit a été supprimé doivent être fusionnés ; dans un second temps, la fonction *dissolve* doit être appliquée à la nouvelle couche fusionnée, afin de supprimer les tweets qui seront présents plus d'une fois, un tweet étant susceptible de contenir plusieurs mots-clés recherchés : par exemple, un tweet contenant les mots « *orage* » et « *pluie* » sera présent deux fois dans le corpus intermédiaire.

2.3. Exploiter les jeux de tweets : quels usages pour quelle distribution des données ?

Les deux jeux de tweets élaborés au point précédent, c'est-à-dire la couche des tweets sans distinction sémantique inclus dans le terrain d'étude, et la couche des tweets en rapport à l'événementiel hydrométéorologique, constitueront les deux jeux de données destinés à caractériser et à comparer les variations de flux ainsi que la distribution spatio-temporelle des tweets en fonction des périodes et des perturbations. La méthodologie d'analyse est donc fondée sur trois principaux axes : l'étude de l'usage du réseau social sur la période extraite, la visualisation des variations spatio-temporelles de la distribution des tweets ainsi que l'étude spatio-temporelle du contenu des tweets sur une journée de crise.

2.3.1. L'usage quantitatif de Twitter

La première partie de l'analyse est fondée sur deux axes : il s'agit en premier lieu de caractériser l'usage quantitatif de Twitter chez ses utilisateurs, puis de mettre en exergue les variations de la distribution temporelle des tweets.

- Du tweet à l'utilisateur

Dans un premier temps, nous nous intéressons aux habitudes des utilisateurs de Twitter ; nous cherchons ainsi à connaître le nombre d'utilisateurs enregistrés dans la base de données mais encore à estimer le nombre moyen de tweets postés pour chaque utilisateur. Il est également pertinent de connaître

la proportion d'utilisateurs contribuant à la production d'une information liée aux perturbations et le nombre moyen de tweets par utilisateur ainsi que les variations quantitatives de tweets émis en fonction des utilisateurs.

L'objectif de cette analyse est d'une part, de caractériser l'usage du réseau social comme outil d'échange d'informations : les utilisateurs sont-ils plutôt ponctuels, réguliers ou *Twitter-addicted* ? D'autre part, il s'agit d'évaluer la contribution des utilisateurs à la production de l'information utile à notre étude, de manière à évaluer Twitter en tant que source de données pour observer les crises hydrométéorologiques à partir de notre échantillon.

La procédure explicitant les étapes qui permettent de connaître le nombre de tweets postés par chaque utilisateur est détaillée en annexe 6.

- La distribution temporelle des tweets

La distribution temporelle des tweets a pour objectif de caractériser leur répartition quantitative afin de mettre en évidence la variabilité temporelle des émissions, que les tweets soient liés ou non aux événements, et de mesurer les effets d'une crise sur les flux. Cette distribution peut être analysée sur deux échelles différentes : une analyse menée sur une échelle quotidienne consiste à calculer les totaux de tweets émis pour chaque journée enregistrée ; une analyse fondée sur une échelle horaire permet quant à elle d'établir le profil d'une journée normale (ou d'une journée de crise) en calculant les flux horaires d'émission sur vingt-quatre heures.

L'objectif de ces analyses se décline ainsi sur trois aspects principaux : tout d'abord, nous cherchons à caractériser les flux de tweets quotidiens totaux, regroupés par mois, afin de pouvoir calculer une moyenne mensuelle de tweets émis et de vérifier la possible existence de flux supérieurs en période perturbée. Nous souhaitons ensuite connaître et comparer les flux moyens par tranche horaire en période normale et en période de crise, afin de tester l'hypothèse selon laquelle le flux de tweets dépend de la dynamique de l'événement. Enfin, nous étudions l'attitude et les préoccupations des utilisateurs au travers de la distribution des tweets en rapport aux perturbations : il s'agit alors d'observer l'existence éventuelle d'une correspondance entre les dates de crises extraites depuis CatNat (cf. 2.1.2) et les pics de tweets ainsi que de quantifier les tweets émis en dehors de ces périodes.

Les étapes permettant de calculer le nombre de tweets en fonction des jours ou des heures sont décrites en annexe 7.

2.3.2. Visualisation et analyse des variations spatio-temporelles d'émission des tweets

L'utilisation des deux jeux de tweets sous SIG nous permet, afin de poursuivre les analyses, de représenter et de visualiser la distribution géographique des tweets, ainsi que leur variabilité spatio-temporelle. Les tweets constituant des fichiers particulièrement volumineux, nous utilisons la cartographie de densité pour représenter l'information. Ces cartes seront créées sous format raster grâce à l'extension *« Carte de chaleur »* de QGIS.

- La cartographie globale des tweets

Nous cherchons dans un premier temps à connaître la répartition globale des tweets, sans distinction particulière, sur l'ensemble de la période étudiée ; il s'agit, au-delà de la vérification de la corrélation logique entre densité de population et densité de tweets, d'observer l'existence éventuelle de foyers locaux qui peuvent concentrer un grand nombre de tweets en dehors des zones les plus peuplées. Il est également question, si ces foyers existent, d'observer leur durabilité dans le temps. C'est pourquoi des cartes de densité sont également créées sur des plages de temps mensuelles : nous retenons pour cette étape les mois d'octobre, de novembre et décembre 2014, de janvier et février 2015.

- Rechercher et confirmer des anomalies

L'existence de ces foyers en dehors des zones très peuplées peut être mise en évidence en comparant la densité de tweets à la densité de population (dont la cartographie est établie à partir du carroyage INSEE, cf. 2.1.3) ; cette étape consiste à créer, à partir de rasters de densité de tweets et de densité de population, un nouveau fichier exprimant le rapport tweets/population, en utilisant la calculatrice raster.

Le raster résultant du calcul du rapport entre tweets et population est interprété de la manière suivante : les cellules dont la valeur est comprise entre 0 et 1 indiquent soit que le nombre de tweets est inférieur à l'effectif de la population, soit que le nombre de tweets enregistrés sur une cellule est proportionnel à la population (valeur égale à 1). Les résultats supérieurs à 1 sont plus précisément recherchés car ils correspondent à des zones où le nombre de tweets émis est plus important que la densité de population. Il s'agit donc d'identifier avec précision ces foyers, leur existence en fonction du temps, les communes qu'ils concernent et, en sélectionnant les tweets postés sur ces zones, de caractériser leur profil, c'est-à-dire le nombre d'utilisateurs, le nombre total de tweets et le nombre de tweets liés aux événements.

- La cartographie des tweets liés aux phénomènes

Les tweets liés à l'événementiel hydrométéorologique font également l'objet de cartographies de densité d'échelle mensuelle. L'objectif consiste en premier lieu à spatialiser et à quantifier l'information liée à la crise afin de mettre en évidence les espaces affectés par les perturbations ; il s'agit ensuite de vérifier si les variations spatio-temporelles et quantitatives des émissions de tweets peuvent représenter un indicateur du nombre d'événements et de leur sévérité.

Il convient également de vérifier si les foyers isolés, identifiés lors de l'étape précédente, correspondent à d'importantes densités de tweets en rapport avec les événements, en superposant les rasters de densité mensuelle. A partir de ces résultats, nous pouvons mener une analyse spatio-temporelle du contenu lexical des tweets sur une échelle plus fine, zoomant sur l'événement. Cette analyse consiste, à partir des tweets enregistrés sur les jours de crise concernés, à représenter, heure par heure, la thématique de chaque tweet. L'objectif est de reconstituer et de visualiser, par tranche horaire, l'évolution du contenu textuel des tweets et la mobilité éventuelle des utilisateurs. La finalité consiste ainsi à comparer cette représentation avec des données d'évolution des précipitations, dont nous disposerons plus tard. Cette analyse implique néanmoins de créer un champ supplémentaire dans la table attributaire de la couche shapefile et d'affecter un thème général à chaque tweet. Il sera également nécessaire de veiller à ce que l'information véhiculée dans le tweet concerne l'espace dans lequel se trouve l'utilisateur : par exemple, une personne se trouvant dans la Drôme pourrait évoquer un événement survenu dans le Gard.

Dans cette partie, nous avons défini une méthodologie destinée à constituer, en combinant traitements SIG et SQL, deux corpus de tweets : les tweets émis sur le terrain d'étude, c'est-à-dire les huit départements du sud-est de la France affectés par les épisodes cévenols de l'automne 2014, et le corpus de tweets liés à ces événements et, de façon générale, liés aux perturbations hydrométéorologiques survenues d'octobre 2014 à mars 2015. Ce dernier a été uniquement construit à partir d'une analyse textuelle des tweets. Nous avons ensuite proposé une méthodologie d'analyse fondée sur deux aspects : une première phase d'analyse quantitative est destinée à évaluer la représentativité des tweets et à étudier leur distribution temporelle. La seconde phase est, quant à elle, davantage centrée sur la représentation spatio-temporelle des tweets afin d'observer les relations entre densité de tweets et événementiel hydrométéorologique.

3. Analyse et représentation des tweets : quelles informations pour quelles distributions ?

Cette partie expose les premiers résultats obtenus à partir de la méthodologie d'analyse exposée au point 2.3.2. Ces résultats fourniront les premières pistes de réponse aux principales interrogations évoquées précédemment. Il sera ainsi question d'évaluer progressivement les usages quantitatifs de Twitter, en fonction du temps et des utilisateurs, puis d'observer les distributions temporelles et spatio-temporelles des tweets. L'objectif final est de proposer une première évaluation des liens susceptibles d'exister entre tweets et paramètres des crises, comme la durée, l'intensité ou encore la dynamique événementielle.

3.1. Analyse du corpus de tweets enregistrés sur la période étudiée

Les sections suivantes présentent les résultats des analyses menées sur le corpus de tweets enregistrés et contenus dans le terrain d'étude, sans distinction sémantique. Elles sont structurées autour de trois points : la présentation du flux global de tweets et de l'utilisation de Twitter, l'analyse de la distribution temporelle des tweets et enfin, la représentation cartographique de la variabilité spatio-temporelle des émissions. Une carte des noms de communes qui seront évoquées est consultable en annexe 8.

3.1.1. L'utilisation de Twitter en chiffres

La période étudiée, du 8 octobre 2014 au 31 mars 2015, enregistre un effectif total de 1 153 664 tweets géolocalisés sur 151 jours de collecte. Le nombre moyen de tweets envoyés quotidiennement sur cette même période est de 7 639 pour un écart-type de 3 094, c'est-à-dire que les émissions de tweets enregistrent une certaine variabilité. Le nombre d'utilisateurs différents ayant tweeté sur la même période, c'est-à-dire ayant diffusé un message sur le réseau social au moins une fois, est de 28 691.

- L'usage de Twitter

L'utilisation du réseau social comme outil de communication est très variable d'un individu à l'autre : en effet, les calculs des totaux de tweets émis par chaque utilisateur enregistré sur les six mois extraits, effectués dans la table `users` témoignent d'une amplitude variant de 1 à 14 196 tweets par utilisateur. En moyenne, un utilisateur a émis 40 tweets géolocalisés sur toute la période, soit environ 0,3 tweet par jour. L'usage du réseau via les Smartphones est donc davantage ponctuel, la majorité des utilisateurs n'émettant pas quotidiennement.

La courbe des fréquences cumulées (Figure 11) confirme cette hypothèse : 80% des utilisateurs ont envoyé tout au plus vingt tweets géolocalisés en six mois ; 3% des utilisateurs ont émis entre 100 et 200 tweets et ont donc tweeté depuis leur Smartphone environ une fois par jour. Seuls 7% des utilisateurs ont émis plus de 500 tweets, c'est-à-dire qu'ils ont diffusé au moins trois messages par jour. Le tableau des données relatives au nombre d'utilisateurs se trouve en annexe 9.

L'utilisation de tweets géolocalisés paraît donc réductrice, dans la mesure où la majorité des utilisateurs, donc de contributeurs potentiels à la production d'une information utile, n'utilisent pas le téléphone comme outil de communication via Twitter. Ces habitudes pourraient néanmoins évoluer lorsqu'un utilisateur assiste à un événement qui perturbe une activité en cours ou s'il souhaite partager une information par le biais de la diffusion d'une photographie.

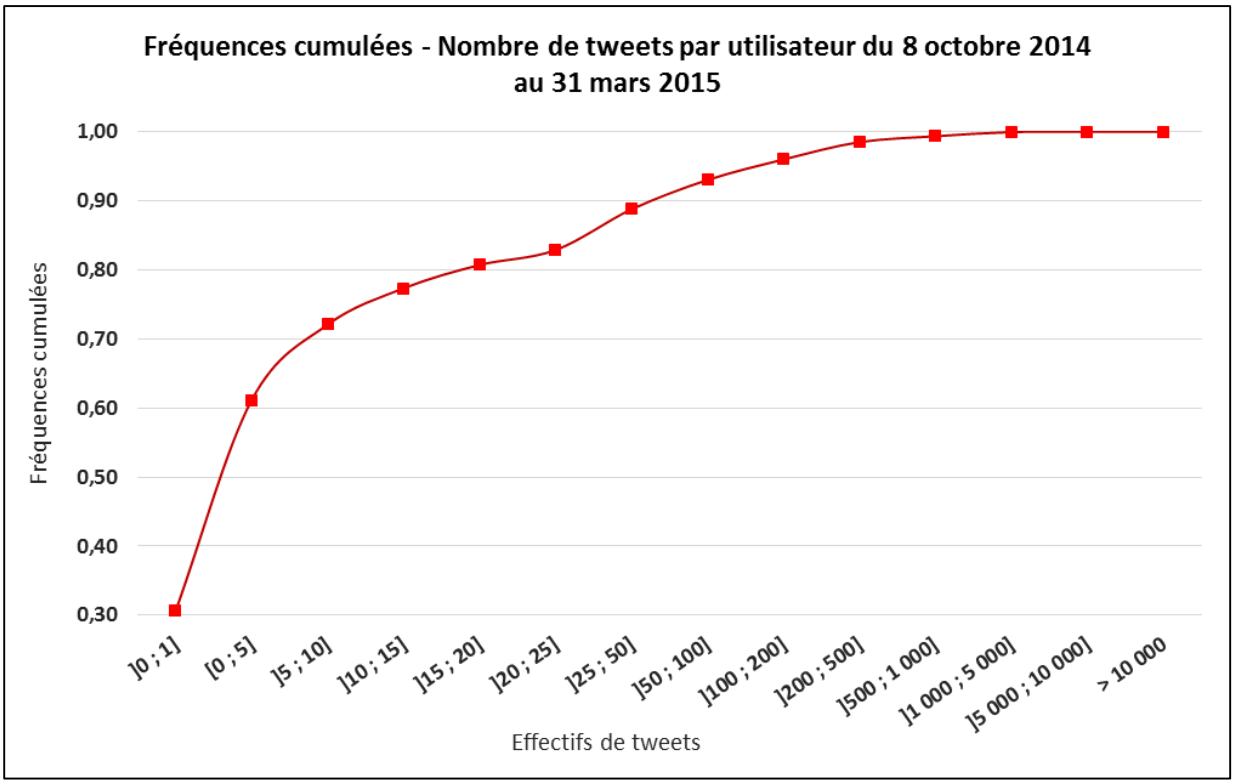


Figure 11 : Courbe des fréquences cumulées – nombre de tweets envoyés par utilisateur

- La géographie des tweets

La distribution spatiale de l'ensemble des tweets géolocalisés émis sur la période totale (Figure 12, page suivante) suit quatre logiques principales :

- les plus fortes concentrations de tweets correspondent à trois grands axes qui suivent les voies de communication principales et la répartition régionale de la population : le premier descend la vallée du Rhône ; la distribution des tweets se diffuse ensuite sur l'axe Avignon-Nîmes-Montpellier-Béziers le long du littoral languedocien, et sur l'axe Avignon-Marseille-Toulon le long du littoral de la région PACA.
- les principales vallées concentrent également les tweets : il s'agit en particulier des vallées de l'Argens (Var), de la Durance (Vaucluse), de l'Hérault et de l'Orb (Hérault).
- d'autres villes moins importantes ou espaces ruraux situés à proximité de grands centres urbains peuplés témoignent d'une utilisation importante de Twitter : Alès, Aubenas, extrêmes nord de la Drôme et de l'Ardèche ou encore les communes rurales situées à l'ouest de Montpellier.
- les tweets suivent également les axes de communication secondaires : Privas-Aubenas-Alès-Nîmes, Avignon-Apt ou encore Nîmes-Arles-Aix-Brignoles-Fréjus.

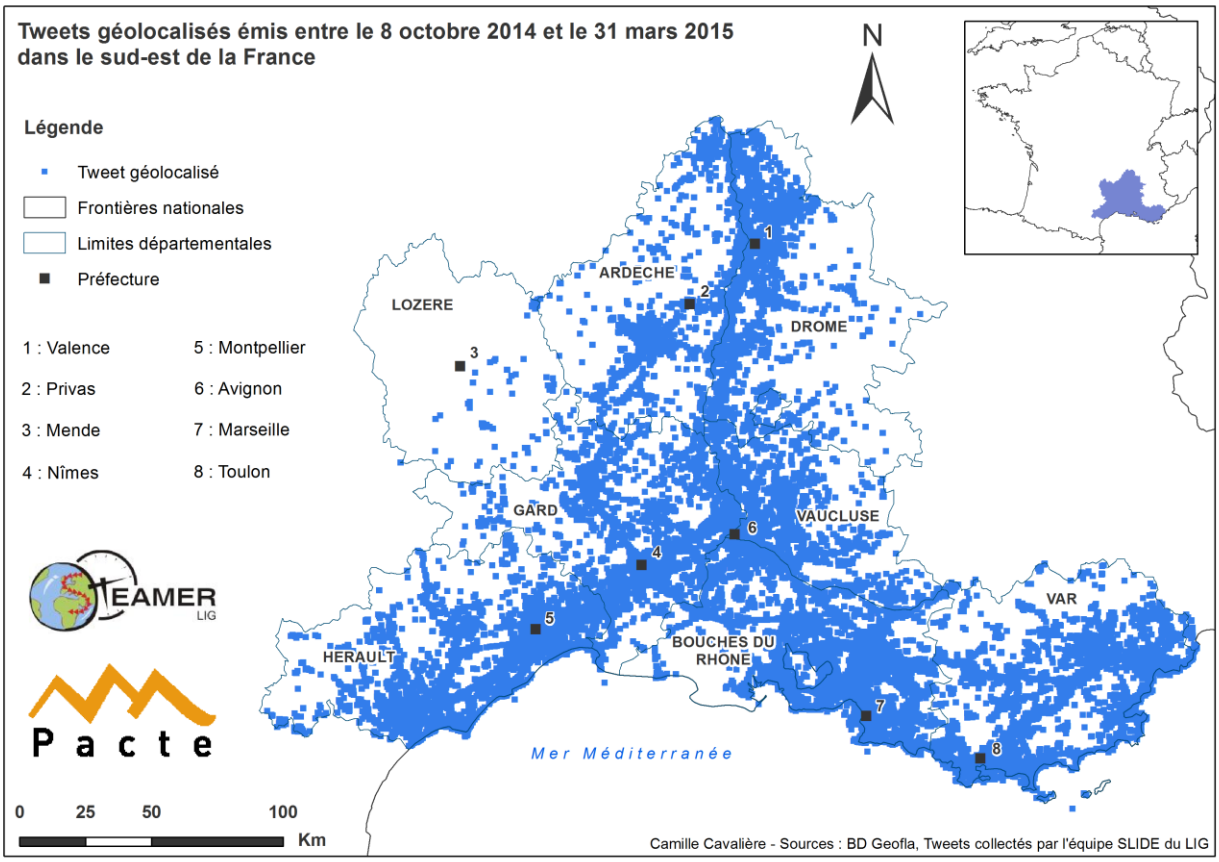


Figure 12 : Carte de localisation des tweets géolocalisés émis sur la période totale

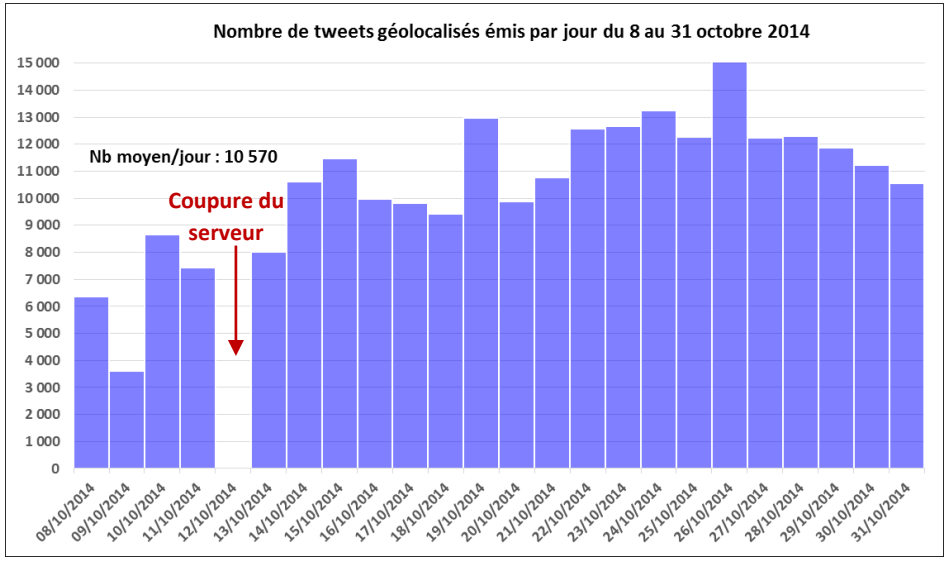
3.1.2. La distribution temporelle des tweets

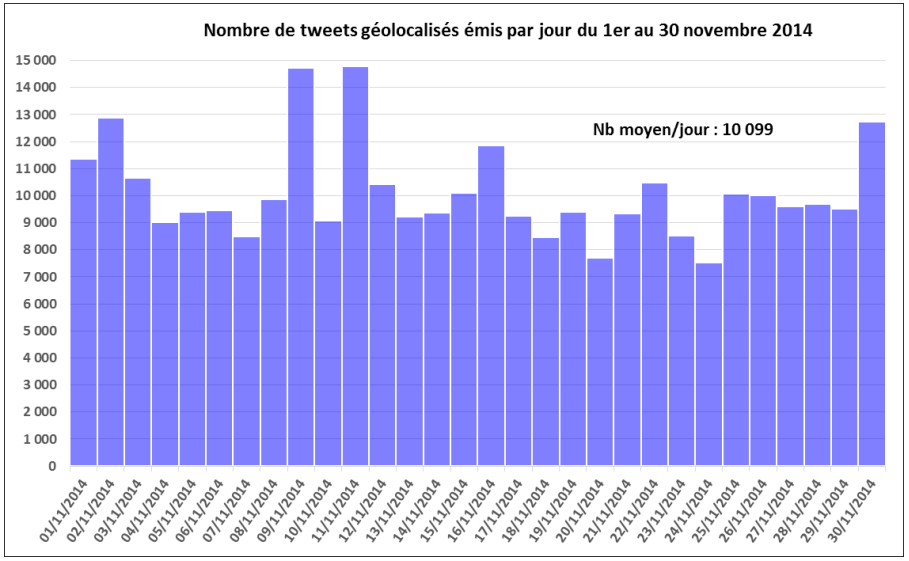
Cette distribution a été étudiée sur deux échelles temporelles : l'échelle quotidienne, qui représente l'effectif de tweets géolocalisés émis sur chaque journée enregistrée et l'échelle de l'heure, afin de caractériser les flux en fonction de la journée.

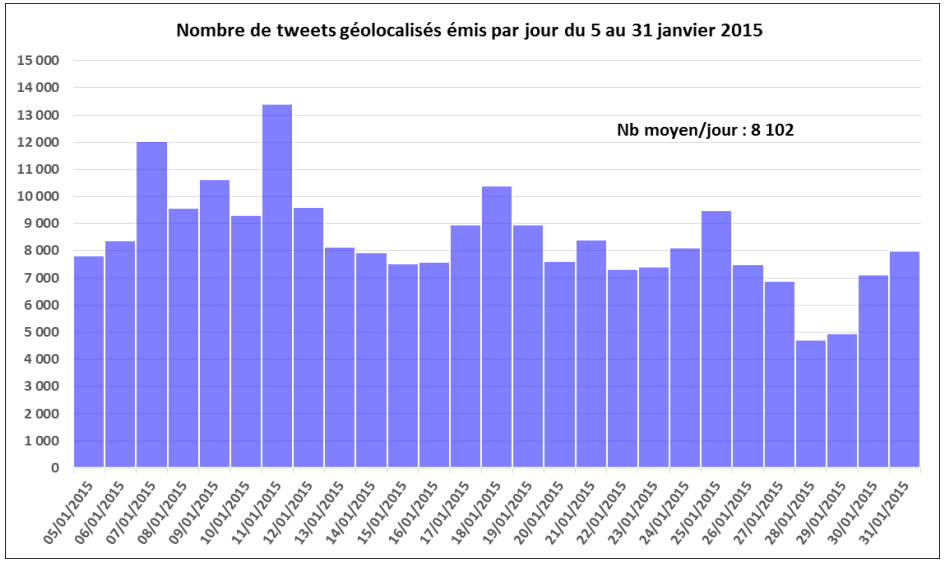
- La distribution quotidienne des tweets

Les effectifs de tweets géolocalisés ont été calculés pour chaque date enregistrée entre le 8 octobre 2014 et le 31 mars 2015 : la figure 13 permet de comparer le nombre de tweets émis quotidiennement pour les quatre mois sur lesquels nous avons un maximum de données, en prenant en compte les coupures du serveur et la diminution du flux de tweets collectés imposée par Twitter à partir de mars 2015 : il s'agit donc des mois d'octobre et de novembre 2014, de janvier et de février 2015.

Les données chiffrées, pour ces résultats, sont disponibles en annexe 10.







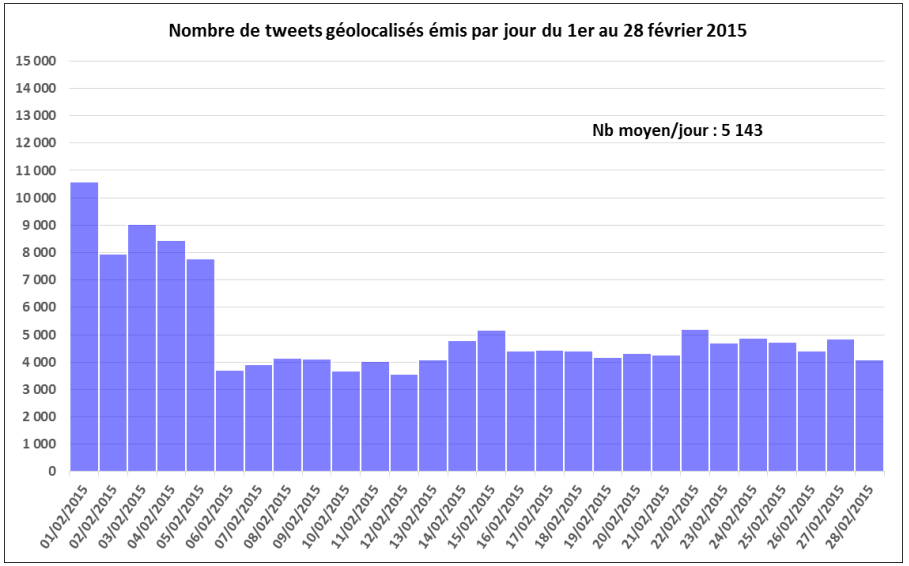


Figure 13 : Comparaison mensuelle des tweets géolocalisés émis quotidiennement

Les mois de l'automne présentent en effet des effectifs totaux plus conséquents que les deux mois de l'hiver affichés sur la figure précédente : sur ces quatre mois, la moyenne quotidienne de tweets émis est de 8 500 alors que cette même valeur, pour les mois d'octobre et de novembre, correspond respectivement à 10 570 et 10 099 tweets par jour. Par ailleurs, sur un total de 860 244 tweets enregistrés sur cette même période, les mois d'octobre et de novembre concentrent à eux seuls plus de 63% de ce flux, ce qui corrobore l'observation d'un déséquilibre temporel de la distribution quantitative des tweets.

Nous pouvons par ailleurs répertorier, sur ces deux mois, quarante-sept jours pendant lesquels le nombre de tweets émis dépasse la valeur moyenne, sur un total de cinquante-trois jours enregistrés (soit 88%) ; ces valeurs baissent sensiblement en janvier et en février ; seuls treize jours dépassent cette valeur seuil, sur un total de cinquante-cinq jours (soit 24%). Ces fortes valeurs se retrouvent sur les dates de crise, toutefois avec une variabilité quantitative assez marquée – 8 652 tweets le 10 octobre et 14 732 le 9 novembre 2014 - et également sur la première moitié du mois de janvier, l'événementiel ayant été marqué par les attentats de Paris survenus entre le 7 et le 9 janvier 2015. Notons néanmoins que les plus forts pics de tweets de l'automne ne coïncident pas nécessairement avec les dates de crise relevées : en effet, aucun événement météorologique n'est signalé le 26 octobre alors qu'on observe un pic à 15 062 tweets ce même jour.

Pour conclure, le surplus de tweets observé à l'automne, s'il est réellement le résultat des intempéries, semble indiquer que les tweets liés aux événements s'ajoutent au tweets des conversations habituelles – dans le cas contraire, nous observerions probablement une distribution homogène. Par ailleurs, la comparaison des valeurs moyennes entre l'automne et janvier indiquerait que l'événementiel local est davantage resté ancré dans les préoccupations qu'un événement ayant marqué l'actualité nationale et internationale.

- La distribution horaire des tweets

Les flux par tranche d'heure (Figure 14) ont également été calculés à partir de toutes les dates enregistrées sur la même période qu'indiquée précédemment.

Sans surprise, ces flux sont standardisés et représentatifs de l'activité quotidienne d'une personne : la communication sur le réseau s'amorce à partir 6 heures le matin et augmente progressivement pour atteindre un premier pic pendant la pause du déjeuner. Les flux diminuent ensuite pendant les premières heures de l'après-midi pour s'accroître de nouveau à partir des tranches horaires 16 heures – 18 heures, c'est-à-dire après la fin des cours ou à la sortie du travail. Les flux maximum sont enregistrés entre 21 heures et 23 heures, qui comptabilisent respectivement 706 et 682 tweets par heure (pour un flux horaire moyen de 316 tweets). L'activité diminue ensuite dans la dernière heure de la journée et sur les premières heures du matin.

Les données chiffrées de ce graphique sont disponibles en annexe 11.

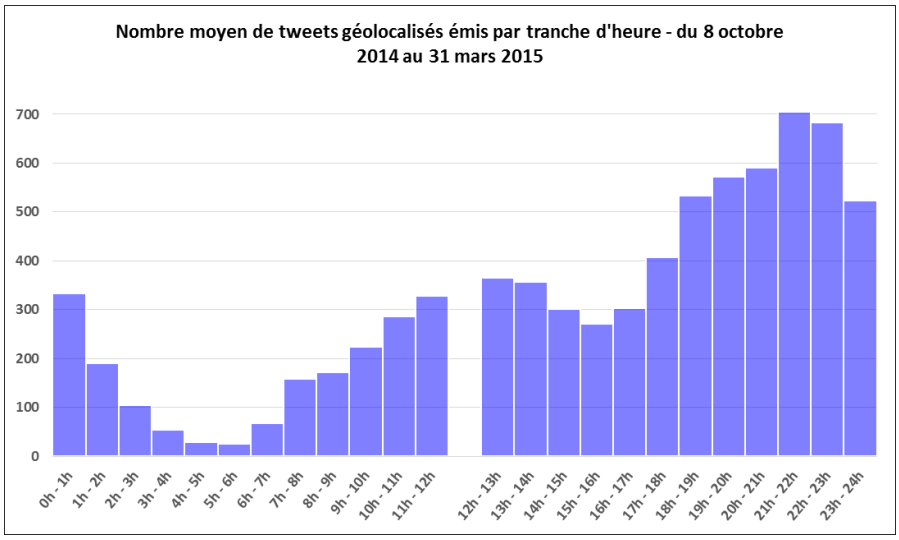


Figure 14 : Nombre moyen de tweets émis par tranche horaire

3.1.3. Les variations spatio-temporelles de la distribution des tweets

La variabilité de la distribution spatio-temporelle des tweets a été étudiée au travers de la représentation globale et mensuelle des densités de tweets ; les mois retenus pour les analyses d'échelle mensuelle sont octobre, novembre, décembre 2014, janvier et février 2015 (Figure 15). Nous avons également choisi un rayon de lissage de 5 000 mètres, pour obtenir une cartographie précise.

La fenêtre représentant la densité de tweets émis sur la période globale corrobore l'observation de la concentration maximale des tweets selon les trois axes de diffusion mis en évidence précédemment (cf. 3.1.1). Pour autant, plusieurs pôles d'émission importants, dont les densités dépassent mille tweets géolocalisés, se distinguent nettement dans des zones plus rurales : les communes du Cheylard, d'Aubenas et de Chomerac en Ardèche, la moitié sud de l'Hérault autour de Béziers ainsi que le Var, autour des communes de Salernes, Brignoles et Vinon-sur-Verdon.

L'analyse des fenêtres mensuelles indique que les villes ayant un statut administratif particulier – préfecture ou sous-préfecture – c'est-à-dire les villes les plus peuplées, concentrent les plus fortes densités de tweets, mais sur des superficies variables. Par ailleurs, si l'on s'intéresse aux couronnes périurbaines et zones rurales, les superficies les plus importantes affichant les plus fortes densités de tweets se retrouvent – dans l'ordre décroissant - sur les mois de novembre, de janvier et d'octobre, avec respectivement des surfaces de 7 764 km², 6 451 km² et 6 384 km² concentrant entre 100 et 10 000 tweets[24].

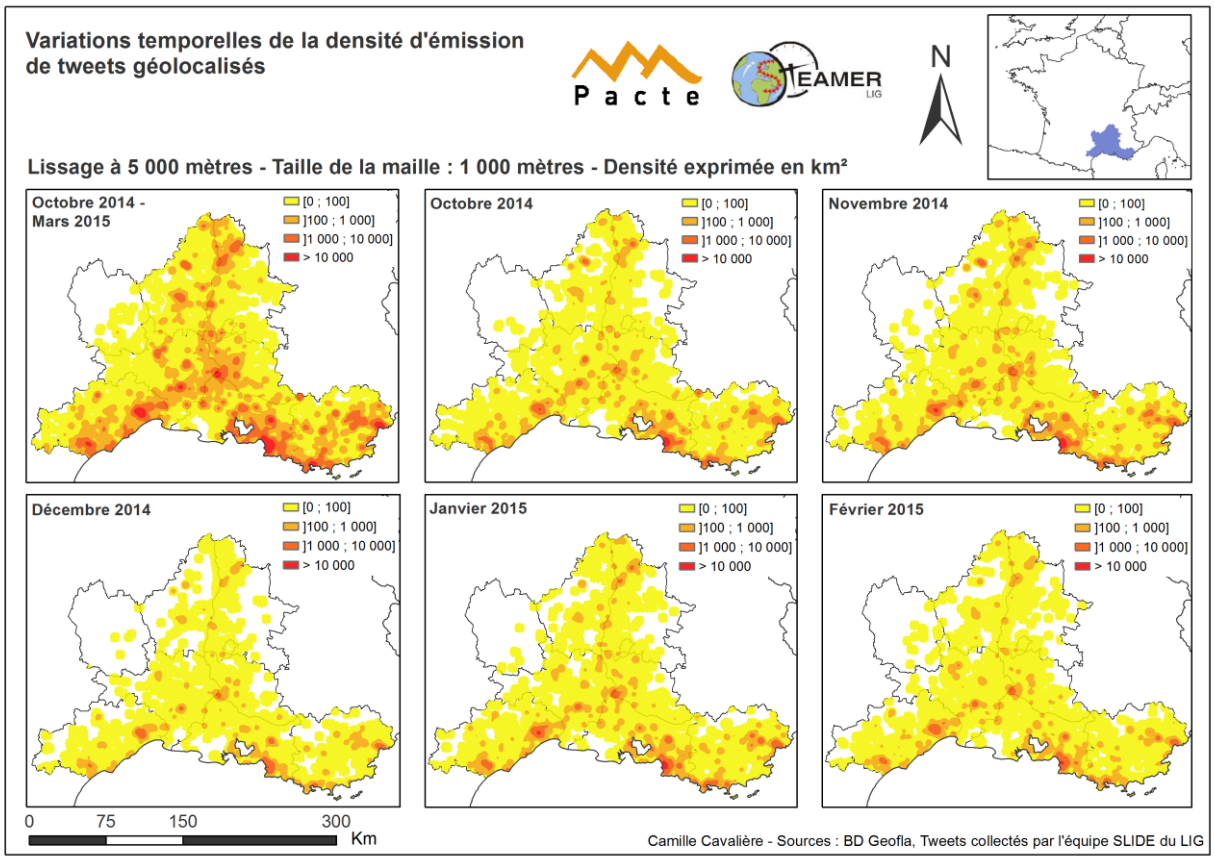


Figure 15 : Carte des variations spatio-temporelles de densité d'émission des tweets

[24] Les calculs de superficie des rasters sont effectués sous SIG à l'aide de la calculatrice raster

Certains foyers des milieux périurbains se distinguent particulièrement sur les périodes concernées par des événements conséquents :

- en Ardèche, on note deux foyers particulièrement actifs à l'automne, qui s'atténuent sur les autres mois : il s'agit d'îlots situés sur les communes du Cheylard et de Saint-Martial ;
- au nord et au sud d'Avignon, se concentrent plusieurs foyers très localisés qui semblent plus actifs sur novembre, octobre et janvier ;
- l'axe Draguignan-Fréjus est particulièrement actif sur octobre et novembre ;
- un dernier foyer plus local apparaît uniquement sur octobre et novembre, dans la vallée de la Durance, au point où se rencontrent Var, Vaucluse et Bouches-du-Rhône, sur la commune de Vinon-sur-Verdon.

3.1.4. Détection des anomalies de densités

Enfin, nous avons également choisi d'aborder la question de l'identification d'espaces dans lesquels nous pourrions observer des densités de tweets plus élevées que les densités de populations. La Figure 16 présente une collection de cartes représentant la répartition des tweets rapportée à la population sur les mêmes fenêtres temporelles qui ont été utilisées précédemment. La cartographie des densités de population, établie à partir des données carroyées de l'INSEE, est disponible en annexe 12.

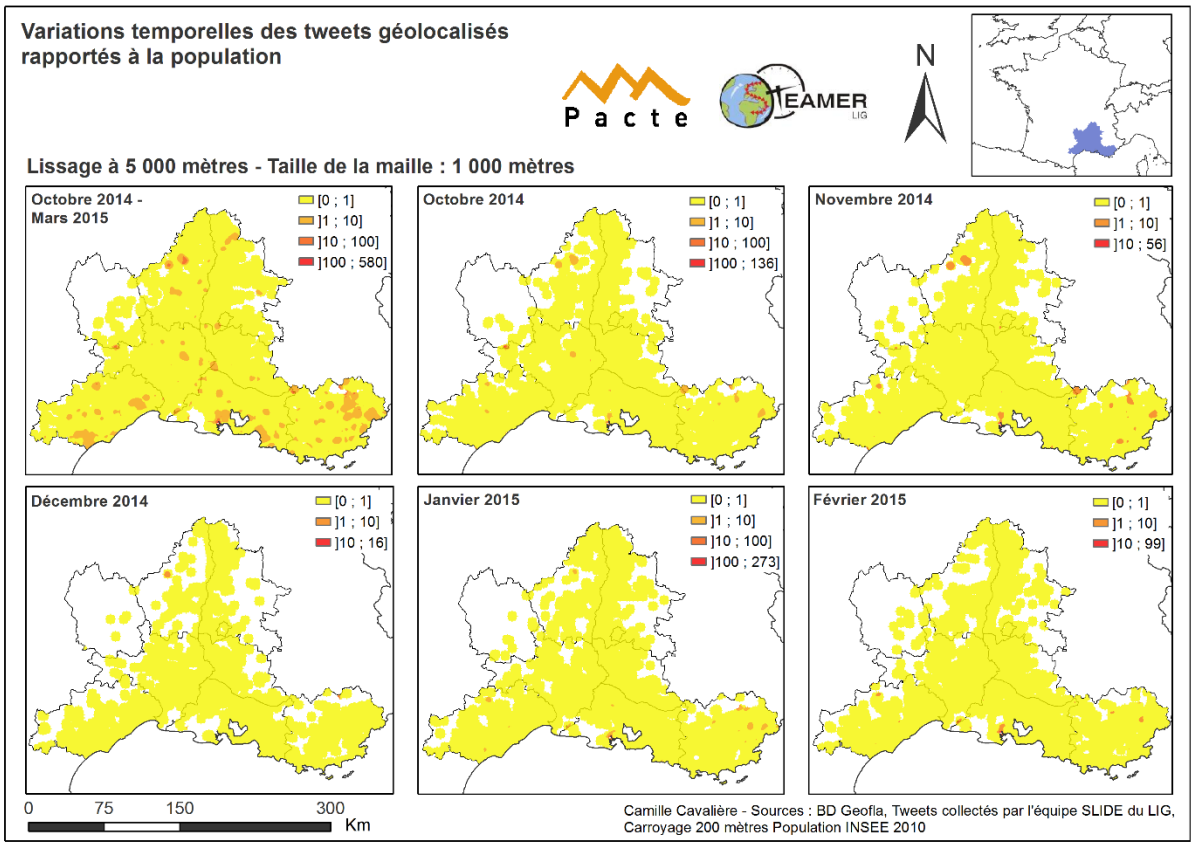


Figure 16 : Carte des variations spatio-temporelles des tweets rapportés à la population

Sur ces différentes fenêtres, la classe [0 ; 1] correspond à des rapports inférieurs ou équilibrés, c'est-à-dire que le nombre de tweets émis est sous-représenté, ou proportionnel (valeur 1) à la population de l'espace. Les principaux centres urbains témoignent de rapports équilibrés, sauf sur la période globale, pendant laquelle les excédents restent relativement restreints, compris entre un et dix tweets. Pour autant, on distingue plusieurs foyers de zones rurales ou périurbaines, actifs pendant l'automne sur les départements suivants :

- en Ardèche, sur les communes du Cheylard et de Saint-Martial ;
- à la limite entre le Gard et la Lozère, sur les communes de Valleraugue et Bassurels ;
- dans le Gard, sur la commune de Saint-Quentin-la-Poterie ;
- dans l'Hérault, sur les communes de Saint-Etienne-de-Gourgas et de Fontès ;
- dans les Bouches-du-Rhône, sur la commune de Fos-sur-Mer ;
- dans le Var, sur les communes de Vinon-sur-Verdon, Salernes, Collobrières, Fréjus, Puget, Figanières et Châteaudouble.

L'éventuelle correspondance entre ces foyers et la présence de fortes densités de tweets liés aux perturbations hydrométéorologiques sera vérifiée.

3.2. Analyse du corpus de tweets liés à l'événementiel hydrométéorologique

Nous exposons maintenant les résultats des analyses effectuées sur le corpus de tweets extraits, liés à l'événementiel hydrométéorologique. Sa présentation s'articule autour de trois points : la caractérisation de l'utilisation du réseau en période de crise, l'étude de la distribution temporelle des tweets et de leur contenu, et la visualisation des variations spatio-temporelles de la distribution des tweets.

3.2.1. L'utilisation de Twitter pendant des périodes de crise

En combinant les deux méthodes d'extraction précédemment exposées, nous avons construit un corpus de tweets filtrés enregistrant un total de 3 457 tweets émis entre le 8 octobre 2014 et le 31 mars 2015. Ces tweets sont répartis sur 123 dates différentes et couvrent donc 81% de la période totale étudiée. Sur ces 123 jours, le nombre moyen de tweets postés est de 28, celui-ci pouvant bien entendu varier en fonction de l'événement (la valeur de l'écart-type est de 51). Par ailleurs, ce second corpus enregistre 1 490 utilisateurs différents, soit 5% de l'effectif total des utilisateurs répertoriés sur les six mois.

- Utilisateurs et perturbations

Le poids des utilisateurs ayant émis au moins une information utile à l'étude des aspects sociaux relatifs à un événement hydrométéorologique est donc très faible : l'exploitation du réseau pour véhiculer ce type d'information n'est donc pas prioritaire ou les utilisateurs modifient peu leurs habitudes lorsqu'un événement survient. L'amplitude des tweets postés par chaque utilisateur inscrit dans ce jeu de tweets varie en effet de 1 à 43 ; le nombre moyen de tweets postés par chaque utilisateur sur les 123 dates enregistrées est de 2.

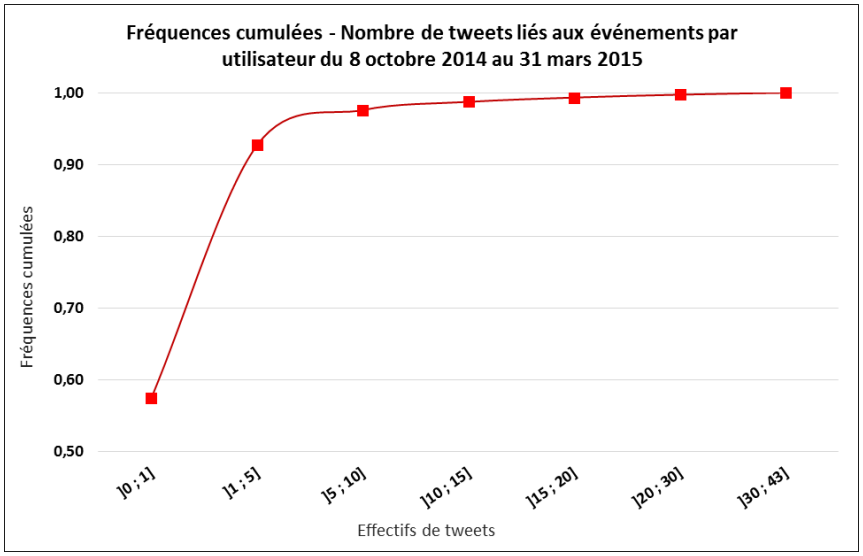


Figure 17 : Fréquences cumulées des utilisateurs en période de crise

La courbe des fréquences cumulées (Figure 17) indique que 57% de ces utilisateurs ont envoyé un seul tweet et 92% d'entre eux en ont postés moins de cinq.

Le tableau des données relatives au nombre d'utilisateurs se trouve en annexe 13.

- La géographie de crise de Twitter

La distribution spatiale des tweets liés à l'événementiel hydrométéorologique adopte un schéma identique à celui du corpus de tweets émis sur l'ensemble de la période (Figure 18), bien que les densités d'émission soient plus éparses : nous retrouvons en effet une diffusion de l'information le long du couloir rhodanien et sur le littoral méditerranéen. Une nouvelle fois, l'information est concentrée autour des principaux centres urbains régionaux ; pour autant, nous retrouvons certains foyers identifiés précédemment comme potentiellement liés à l'événementiel hydrométéorologique de l'automne 2014 (cf. 3.1.3) : il s'agit des foyers ardéchois, de la vallée de l'Argens et de l'îlot identifié sur la commune de Vinon-sur-Verdon, auxquels peut s'ajouter Béziers et sa couronne périurbaine.

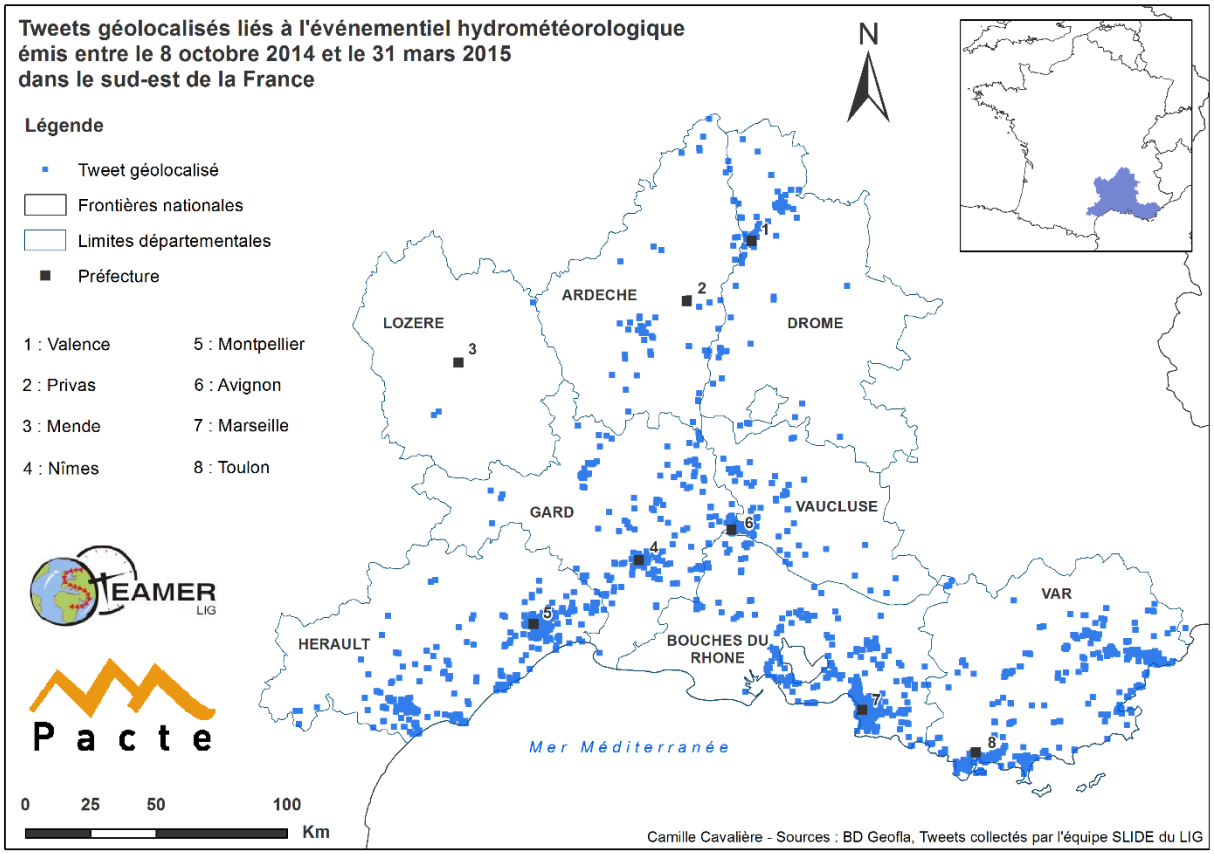


Figure 18 : Carte de localisation des tweets liés aux perturbations hydrométéorologiques

3.2.2. Etude de la distribution lexicale des tweets

Ce second corpus de tweets est donc le résultat de la combinaison des deux méthodes décrites dans la partie précédente : l'extraction globale a été menée, après complément de la liste établie par la connaissance experte (cf. annexe 4), sur une base de 91 mots-clés. Avant l'application de la fonction SIG *Dissolve*, destinée à supprimer les tweets présents plus d'une fois dans le jeu intermédiaire filtré, celui-ci contenait un total de 4 319 tweets. Ce jeu intermédiaire est ici exploité car nous nous intéressons davantage au nombre de tweets mentionnant un mot précis, plutôt qu'à leur distribution spatio-temporelle.

- Bilan de la recherche dirigée

L'extraction dirigée de tweets à partir d'un vocabulaire fondé sur la connaissance des phénomènes a permis d'extraire 3 568 tweets ; elle a donc fourni 98% du jeu intermédiaire. La répartition

des tweets en fonction des différents mots-clés recherchés indique un nombre moyen de 62 tweets par mot-clé (ou association de mots-clés) ainsi qu'un fort écart type, de 170. Treize mots-clés comptabilisent un effectif supérieur à cette moyenne : « *pluie* », « *pleut* », « *orage* », « *inondation* », « *alerte* », « *intempérie* », « *déluge* », « *tempête* », « *Gard* », « *Gardon* », « *route* », « *pont* » et « *éclair* » ; ces mots-clés représentent 83% de l'information totale. Le tableau 2 récapitule les effectifs de tweets pour les cinq mots-clés les plus fréquemment employés :

Pluie	1 073 tweets
Pleut	950 tweets
Orage	357 tweets
Inondation	295 tweets
Alerte	245 tweets

Tableau 2 : Effectifs de tweets enregistrés pour les cinq mots-clés les plus utilisés

Les résultats précis pour chaque mot-clé sont disponibles en annexe 14.

- Bilan de la recherche fondée sur l'examen automatisé du contenu lexical des tweets

Cette analyse complémentaire a nécessité de mener plusieurs essais à partir des jeux de tweets extraits sur des temporalités différentes avant d'aboutir à des résultats exploitables : nous avons en effet été confronté au problème du poids très restreint des tweets liés aux événements par rapport au flux total, et ce, même en période de crise, comme nous pourrons le constater dans les sections suivantes. Trois essais ont successivement été menés : le premier consistait à extraire, depuis la base de données `Twitter`, trois jeux correspondant successivement au champ `text` de l'ensemble des tweets émis pour les mois d'octobre, de novembre et de décembre, dont les effectifs de tweets postés sont respectivement 243 118, 302 972 et 102 680 ; le second essai a été fondé sur un jeu de tweets restreint qui contenait uniquement les messages envoyés sur les dates de crises CatNat, soit un total de dix-huit jours pour un jeu de 174 549 tweets. L'analyse textuelle de ces deux jeux n'a pas permis de mettre en exergue des associations lexicales liées aux événements pour lesquels les mots n'étaient pas répertoriés dans la liste principale.

La dernière analyse s'est donc appuyée sur des jours précis de crise, extraits indépendamment ; à partir de la représentation quantitative quotidienne des tweets inventoriés par la méthode expertisée, trois jeux de tweets ont été extraits de la base de données. En émettant l'hypothèse que les jours concentrant les principaux pics de tweets sont susceptibles de contenir le maximum d'information liée à la crise et, par conséquent, de nous fournir de nouveaux mots-clés, nous choisissons les dates en fonction des effectifs maximum quotidiens du jeu extrait par la méthode expertisée (Annexe 15). Les trois dates retenues sont le 10 octobre, les 4 et 28 novembre 2014 qui contiennent respectivement 8 652, 9 011 et 9 694 tweets.

Pour chacune de ces dates, deux classifications ascendantes hiérarchiques ont été réalisées, en choisissant des seuils minimaux/maximaux de cooccurrence restreints, de 2 à 3 puis de 3 à 4. Ces analyses ont permis de mettre en évidence plusieurs associations lexicales probablement liées aux événements hydrométéorologiques, dont la figure 19 en présente un échantillon.

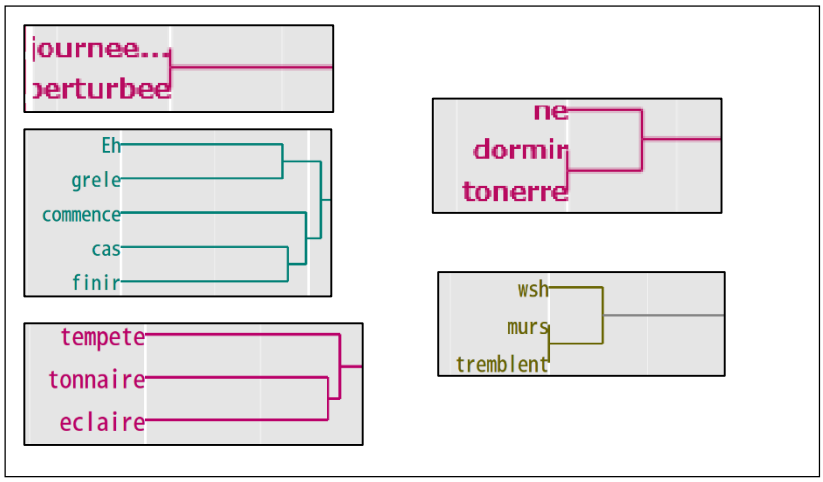


Figure 19 : Associations lexicales et mots-clés extraits depuis KH Coder

Cette méthode a permis de repérer un total de deux associations - « *journée perturbée* » et « *murs tremblent* » – et de sept nouveaux mots-clés : « *Anduze* », « *apocalypse* », « *Espeisses* », « *grêle* » (ou « *grele* »), « *tonerre* », « *tonnaire* » et « *trombe* ». Cette recherche complémentaire a contribué, après élimination du bruit, à enrichir le jeu de tweets intermédiaires évoqué ci-dessus de quatre-vingts tweets. Le tableau 3 ci-dessous affiche la distribution de tweets par mot-clés :

Text Mining	Grêle	30
	Apocalypse	29
	Trombe	8
	Murs Tremblent	4
	Tonnaire	4
	Tonerre	3
	Anduze	1
	Espeisses	1
	Journée Perturbée	0

Tableau 3 : Nombre de tweets par mot-clé extrait de l'analyse lexicale

- Evaluation de la distribution des tweets en fonction des principaux mots-clés

Pour terminer l'analyse de la distribution du contenu lexical des tweets, nous choisissons de visualiser et de comparer, sur un axe temporel, le nombre de tweets émis contenant les quatre principaux mots-clés présents dans les tweets : les mots-clés « *pluie* » et « *pleut* » sont regroupés sous une même entrée ; les autres mots utilisés sont « *orage* », « *alerte* » et « *inondation* ». Le graphique suivant (Figure 20) a été construit à partir des shapefiles intermédiaires créés à l'étape 2.2.2 qui stockent les tweets filtrés pour les mots-clés cités ci-dessus ; il présente les effectifs de tweets pour chaque mot-clé.

Nous observons tout d'abord une correspondance entre les dates des événements indiqués par CatNat et les pics de tweets. En outre, les quatre courbes présentent des variations analogues, hormis sur le début du mois de décembre où seule la courbe représentant les tweets liés à la « *pluie* », connaît un pic. Par ailleurs, ces courbes peuvent d'ores-et-déjà nous renseigner sur les habitudes des utilisateurs : il semblerait qu'ils ne communiquent pas avant la crise et que l'éventuelle phase de préparation n'apparaît, par conséquent, pas dans les tweets. En revanche, les utilisateurs partagent l'information dès le début de la crise et, de manière plus générale, tweetent sur la pluie dès que la météo est perturbée, comme le montre le graphique qui représente l'ensemble de ces tweets sur la période totale (Annexe

16). Sur ce point, il serait pertinent d'étudier le contenu textuel des tweets sur une échelle horaire pour les jours de crise, de sorte à observer l'évolution temporelle des termes liés à l'événement.

En ce qui concerne plus précisément les tweets émis pendant les crises, les dates d'octobre témoignent d'effectifs totaux de tweets, par mot-clé, très proches : ces tweets peuvent alors évoquer simultanément ces termes, ce qui indique que les utilisateurs ont assisté aux quatre événements. Ce phénomène pourrait constituer un indicateur de sévérité de l'événement : on distingue d'ailleurs les pics les plus importants de la période sur les courbes représentant les mots « *alerte* » et « *inondation* ». Les crises de novembre ont tendance à mettre en exergue l'intensité des pluies ; de même, l'événement survenu le 5 décembre concerne essentiellement la pluie.

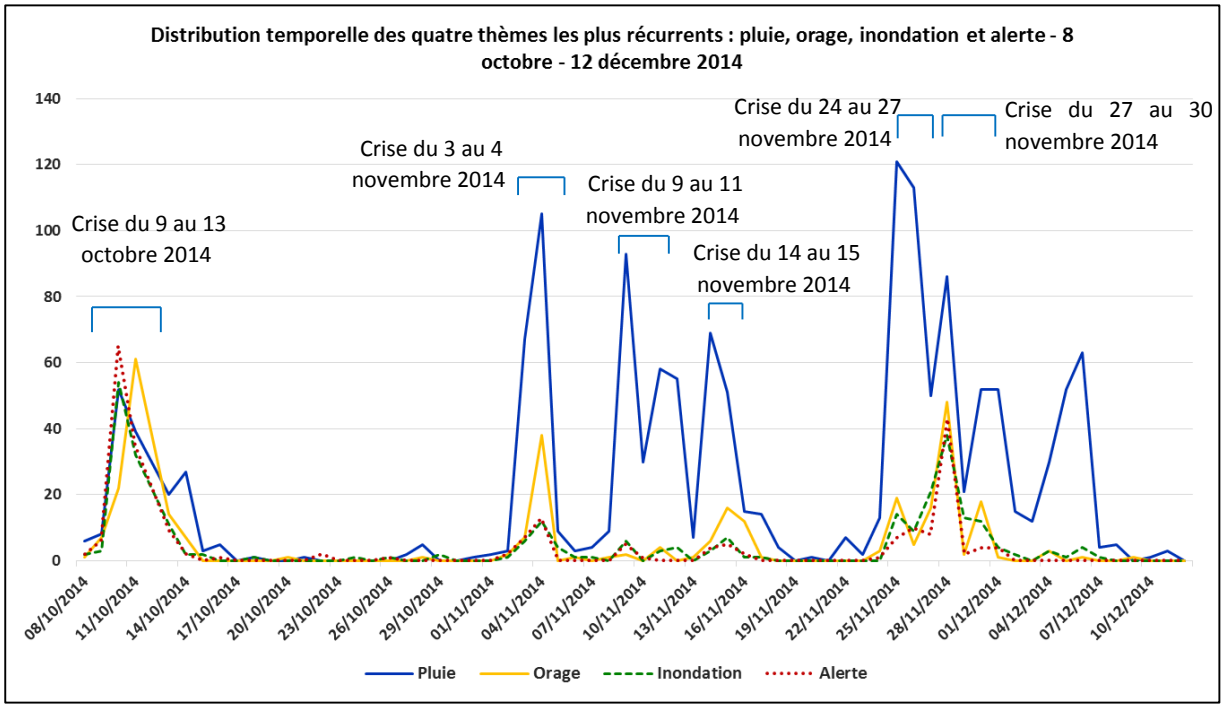


Figure 20 : Distribution temporelle des tweets contenant les mots-clés principaux

3.2.3. La distribution temporelle des tweets liés aux perturbations

Cette distribution a, une nouvelle fois, été étudiée sur deux échelles temporelles : alors que l'échelle quotidienne permet de visualiser la répartition globale des tweets, l'échelle horaire est destinée à estimer le degré de simultanéité entre la dynamique de l'événement et le moment d'émission des tweets.

- La distribution quotidienne des tweets liés aux perturbations

Les effectifs quotidiens ont été calculés pour les 123 dates enregistrées, à partir du corpus de tweets final, c'est-à-dire après la fusion des différentes couches de mots-clés et l'application de la fonction *Dissolve*. Nous rappelons que le nombre moyen quotidien de tweets émis est de 28 mais que celui-ci peut fortement varier en fonction des jours : ainsi, pendant les dates de crise de l'automne 2014, répertoriées par CatNat, ce nombre est de 122 tweets par jour.

Le graphique suivant (Figure 21, page 44) représente la contribution quotidienne des tweets du corpus événementiel, c'est-à-dire le pourcentage de tweets liés aux perturbations par rapport au flux total de la journée. On observe de nouveau la nette corrélation entre les dates de crise et les pics, c'est-à-dire les journées où les utilisateurs se sont davantage mobilisés sur l'événementiel météorologique.

Ces dates concentrent à elles seules plus de 63% de l'information collectée. Toutefois, cette participation à la production d'une information utile reste marginale puisque ces contributions sont comprises entre 0,01% et, pour le maximum enregistré en date du 28 novembre 2014, 3,22%. En outre, la contribution de notre jeu de 3 457 tweets par rapport au nombre total de tweets postés sur les 123 dates enregistrées est très faible, de l'ordre de 0,36%.

Deux autres informations ressortent de ce graphique : tout d'abord, la corrélation entre la contribution de tweets liés à la crise et la sévérité n'est pas vérifiée. Ce phénomène peut être dû à deux causes principales : l'événement a concerné une partie assez restreinte de la zone d'étude dans laquelle nous avons peu d'utilisateurs géolocalisés, ou les utilisateurs habituels de Twitter n'ont pas été directement affectés par la crise. Ensuite, les utilisateurs emploient le vocabulaire recherché en dehors des dates de crise indiquées : en effet, certains se mobilisent dès qu'il pleut, par exemple le 20 janvier et le 21 février 2015, comme il a été établi au point 3.2.1.

Les données ayant été utilisées par la production de ce graphique sont disponibles en annexe 17.

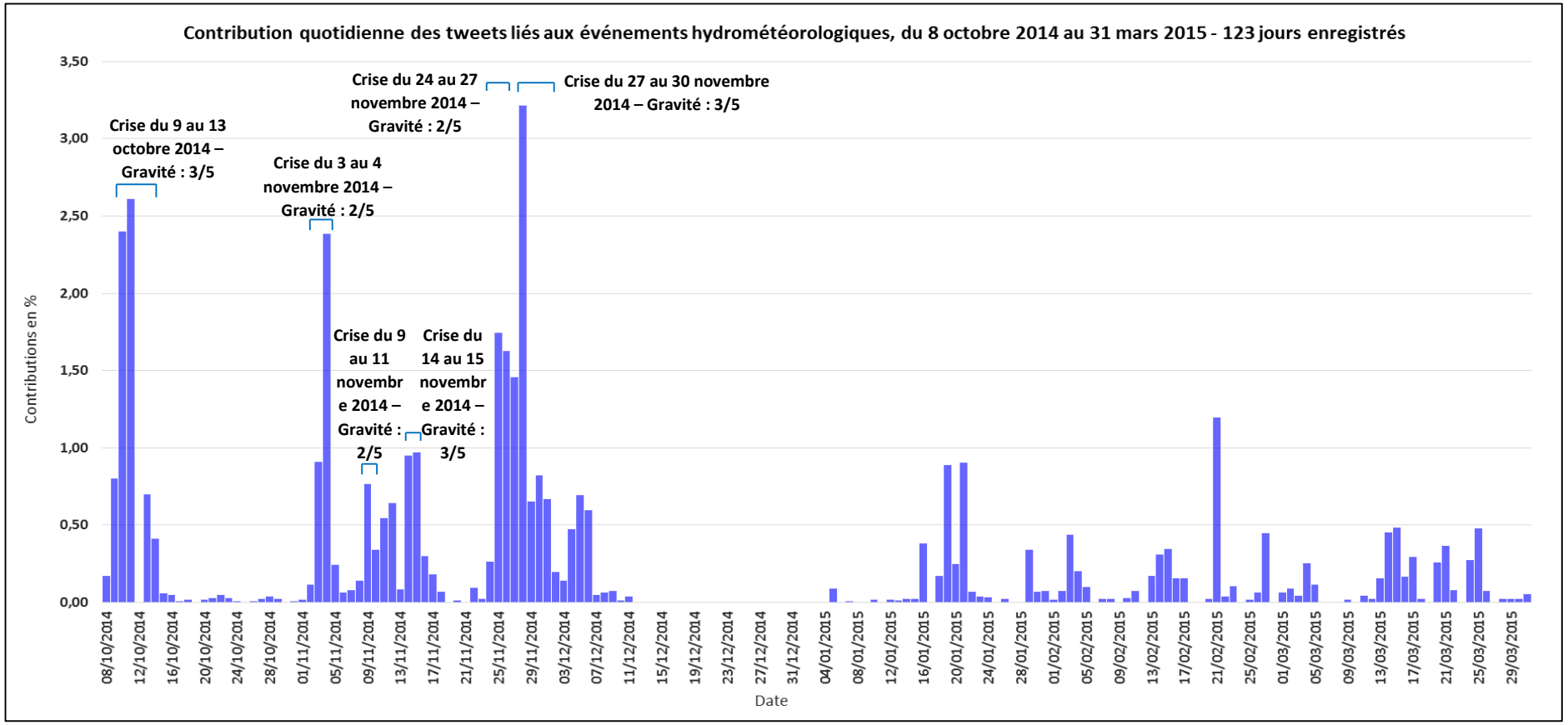


Figure 21 : Distribution quotidienne des tweets liés aux événements et comparaison avec les dates de crises hydrométéorologiques

- La distribution horaire des tweets au travers des crises

Nous avons dans un premier temps, cherché à comparer la distribution horaire du jeu de tweets extraits sur toute la période et la totalité des tweets émis, sans distinction textuelle, pour les dates de crise indiquées. Le graphique résultant (Figure 22) montre que les flux horaires de tweets émis en période de crise témoignent des mêmes variations que le jeu global. Pour autant, on constate que les flux des jours perturbés sont supérieurs à ceux d'une journée moyenne : en période de crise, le flux horaire moyen est de 375 tweets (316 pour une journée normale) ; de même, on calcule un excédent moyen horaire de 59 tweets enregistré sur ces jours particuliers. Les données sont disponibles en annexe 18.

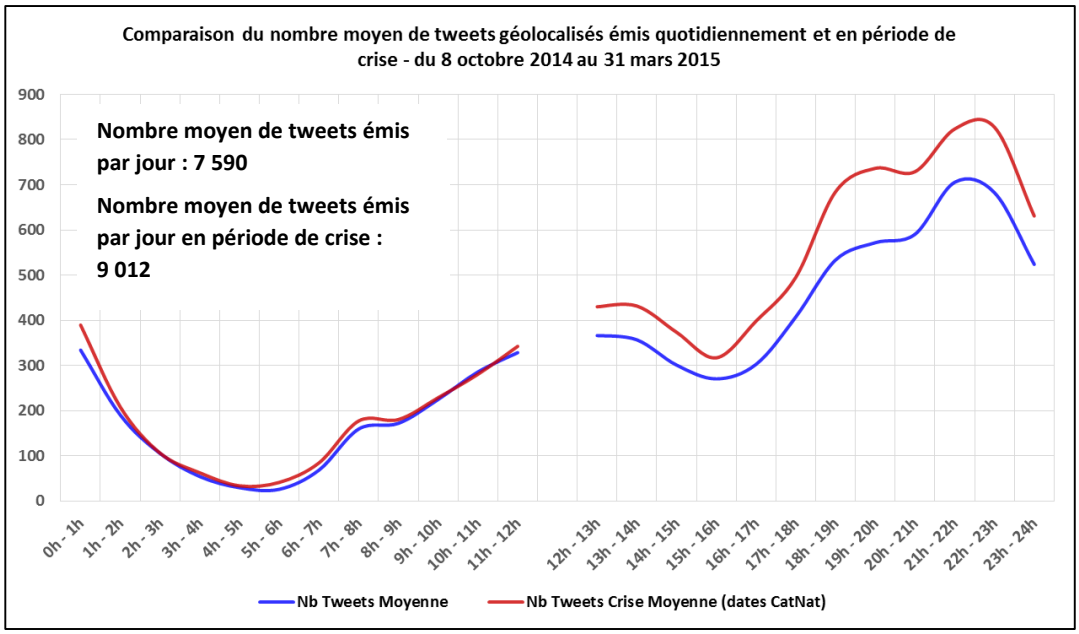


Figure 22 : Comparaison des flux horaires de la période globale (bleu) et des journées de crise (rouge)

Nous nous sommes ensuite attachés plus précisément à la représentation du corpus en fonction de plusieurs échelles temporelles : la distribution horaire globale des 3 457 tweets (Figure 23) indique des variations temporelles analogues à celle de la distribution moyenne analysée sur la période globale (cf. Figure 14). Néanmoins, les flux les plus importants sont enregistrés entre 7 heures et 10 heures le matin ; les forts pics de la soirée ne se retrouvent pas dans cette répartition.

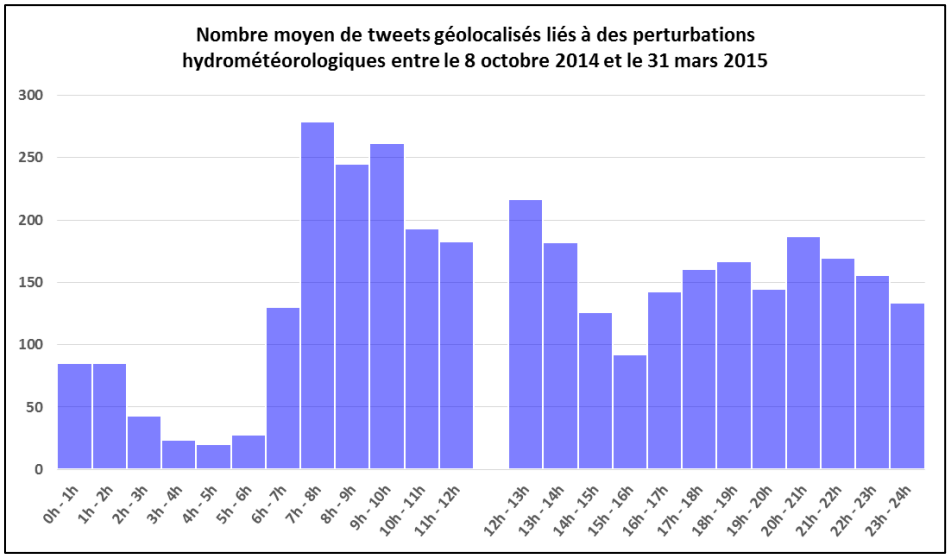


Figure 23 : Distribution horaire du corpus de tweets liés aux perturbations

Enfin, nous avons étudié plus en détail les flux caractéristiques d'une journée de crise, de sorte à évaluer l'existence de la simultanéité entre l'événementiel hydrométéorologique et l'émission de tweets qui lui sont liés. Nous avons effectué un essai sur la date de crise qui présente la plus forte contribution de tweets exprimant un événement (cf. Figure 21) : il s'agit du 28 novembre 2014. Sur la figure 24, on constate en effet que les flux de tweets liés aux perturbations ne suivent pas les variations des flux globaux moyens : on observe un premier pic survenant dans la nuit, le second entre 7 heures et 8 heures et le dernier entre 20 heures et 21 heures. L'analyse de l'évolution temporelle des quatre thèmes (Figure 25) est analogue à cette distribution : le pic des premières heures de la matinée est dû à l'orage ; la pluie est fréquemment mentionnée tout au long de la journée ; la distribution des termes « *inondation* » et « *alerte* » est plus variable, et indique deux pics simultanés entre 1 heure et 2 heures, puis entre 20 heures et 21 heures.

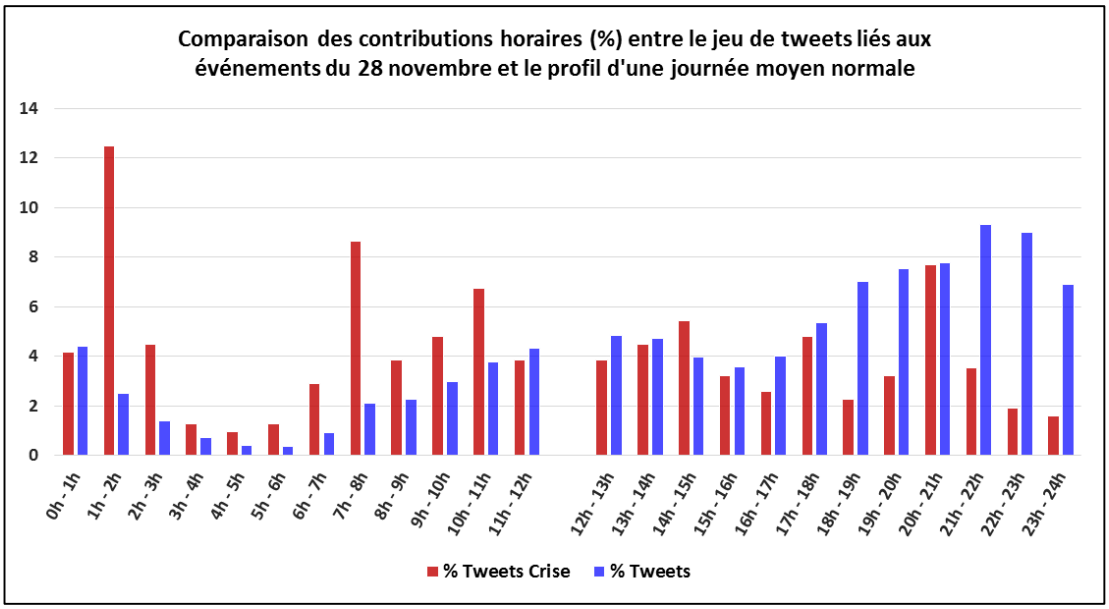


Figure 24 : Comparaison des flux horaires (%): période normale (bleue), tweets évoquant la crise (rouge)

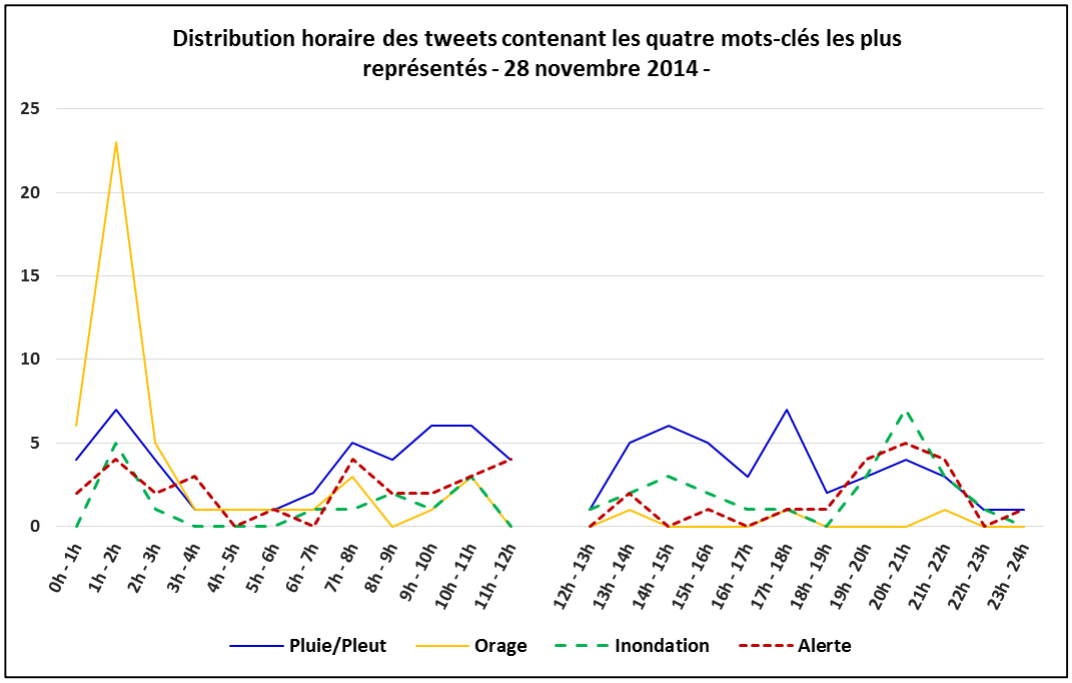


Figure 25 : Distribution horaire des tweets contenant les principaux mots-clés, le 28 novembre 2014

Ces graphiques peuvent être révélateurs de plusieurs phénomènes : les tweets évoquant des événements météorologiques, comme l'orage, suivent certainement leur temporalité. Néanmoins, il est plus difficile de se prononcer, sans données de hauteurs d'eau, sur les tweets évoquant les inondations ou les alertes. Le pic observé entre 7 heures et 8 heures peut ainsi être lié à des personnes qui prennent connaissance de l'état d'alerte avant de quitter leur domicile ; en revanche, le pic du soir pourrait correspondre au moins à deux situations différentes : un état d'alerte inondation est relancé ou les personnes évoquent les événements survenus dans la journée.

Les données à partir desquelles ces graphiques ont été construits sont disponibles en annexes 19 et 20.

3.2.4. La variabilité spatio-temporelle des tweets liés aux événements

Le corpus de tweets liés aux perturbations hydrométéorologiques a également fait l'objet d'une cartographie de densité (Figure 26) afin de visualiser et d'analyser les variations mensuelles des principaux foyers de localisation, tout en vérifiant leur correspondance éventuelle avec les îlots actifs de l'automne 2014, mis en évidence précédemment.

Cette cartographie indique que la distribution spatio-temporelle des tweets suit toujours les principaux axes identifiés ci-avant : les plus fortes concentrations se trouvent dans les préfectures, sous-préfectures et communes des couronnes péri-urbaines, dans la vallée de l'Argens et dans les îlots d'ores-et-déjà signalés. Les mois de l'automne, et en particulier novembre 2014, présentent les plus fortes densités et superficies couvertes par l'information utile. Si l'on visualise, pour les deux mois de l'automne, les rasters de densités de tweets du corpus global et du corpus de tweets en rapport avec les perturbations, nous obtenons tout d'abord une correspondance avec les îlots situés sur les communes suivantes : Le Cheylard et Aubenas en Ardèche, Brignoles, Salernes et Vinon-sur-Verdon dans le Var. Dans un second temps, d'autres îlots péri-urbains ou ruraux, se distinguent plus précisément comme pôle d'émission de tweets pendant les crises. Il s'agit des communes de :

- Régusse et Collobrières dans le Var
- Mouries et Tarascon dans les Bouches-du-Rhône
- Valleraugue, Bagnols-sur-Cèze et Sommières dans le Gard
- Sète, Agde, Laurens et Pézenas dans l'Hérault.

Nous pouvons également constater que certains îlots correspondent avec les espaces où les rapports entre densité de tweets et densité de population se sont révélés élevés : il s'agit des communes de Fréjus, Puget, Figanières, Châteaudouble, Salernes, Vinon-sur-Verdon, Fos-sur-Mer, Saint-Martial et Le Cheylard.

Cette cartographie permet ainsi, après crise et sans lien direct avec le territoire concerné et sa population, de contribuer à l'identification de zones ayant été affectées par les perturbations cévenoles.

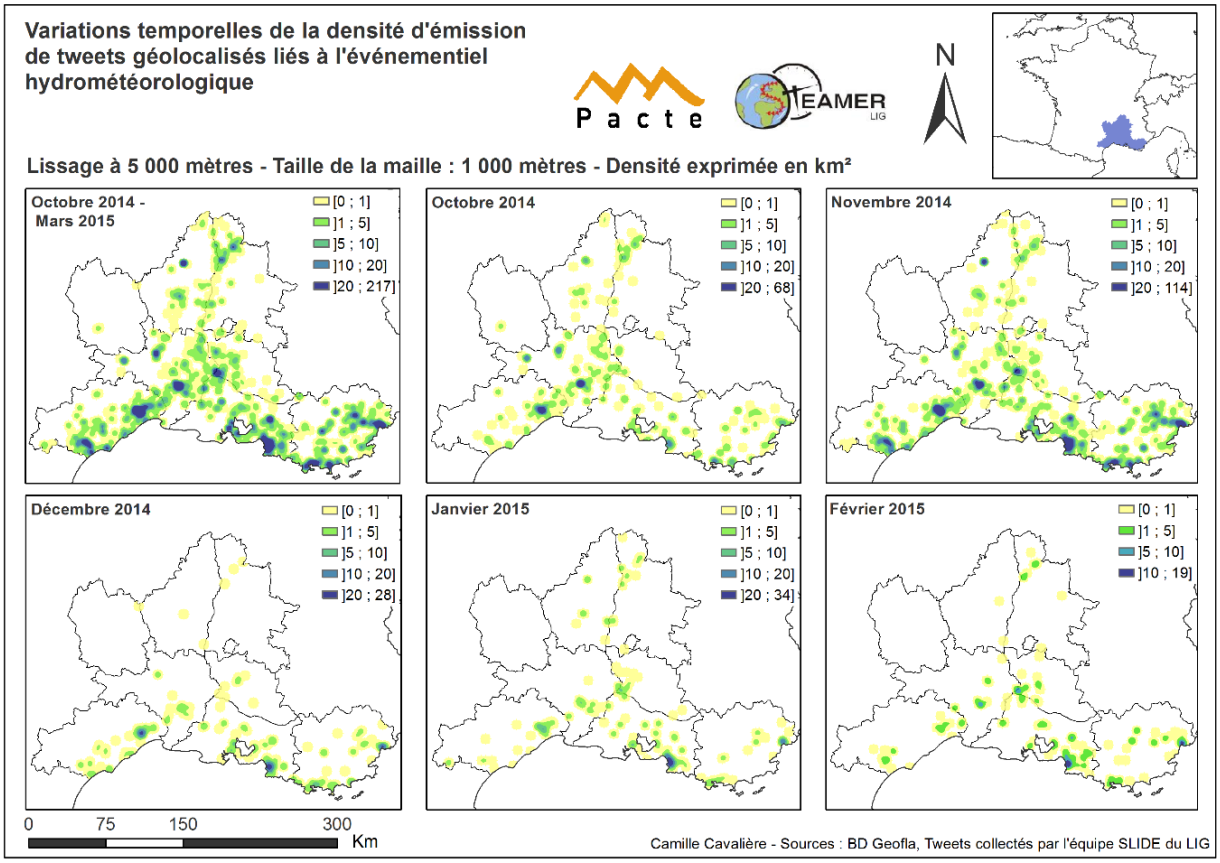


Figure 26 : Carte des variations spatio-temporelles de densité d'émission des tweets liés aux perturbations cévenoles

3.3. Etude d'espaces remarquables

Dans cette section, nous nous concentrons sur l'analyse des espaces particuliers mis en évidence précédemment, dans lesquels nous avons observé, d'une part, une densité de tweets plus élevée que celle de la population et, d'autre part, des foyers d'émission pendant les crises de l'automne. En outre, nous nous focalisons sur les départements pour lesquels nous disposons de données de cumuls de pluie sur octobre, c'est-à-dire l'Ardèche, la Lozère, le Gard et l'Hérault. Deux zones croisent ces critères : les communes de Valleraugue et Bassurels, situées dans le Gard et en Lozère, ainsi que Le Cheylard en Ardèche.

3.3.1. Le profil des espaces particuliers

Nous avons cherché à vérifier, pour ces deux foyers et sur un critère de proximité ou de superposition, l'existence possible d'un lien entre la forte densité d'émission de tweets liés aux perturbations et le fort rapport constaté entre tweets et population ; en d'autres termes, il s'agit de détecter si ces anomalies sont le résultat de flux inhabituels de tweets liés aux événements, ou si elles sont le fruit d'un usage intensif du réseau par certains utilisateurs. L'objectif consiste donc à caractériser le profil de ces zones, en termes d'effectif de tweets et de pratiques d'utilisation.

La carte suivante (Figure 27) représente, pour les trois départements concernés, l'ensemble des tweets géolocalisés émis sur octobre 2014, les densités de tweets liés aux événements (échelle jaune-rouge) et le rapport tweets/population (échelle vert-bleu), pour le même mois. Le premier foyer, localisé sur la commune de Bassurels, a été retenu en raison de sa proximité avec l'îlot lié aux événements (situé à Valleraugue), en admettant l'hypothèse qu'une perturbation affecte simultanément ces deux espaces et que des utilisateurs tweetent sur le même événement, quel que soit le lieu. Le second foyer est situé sur la commune du Cheylard : ici, les îlots de concentration de tweets liés aux événements et de tweets généraux se superposent.

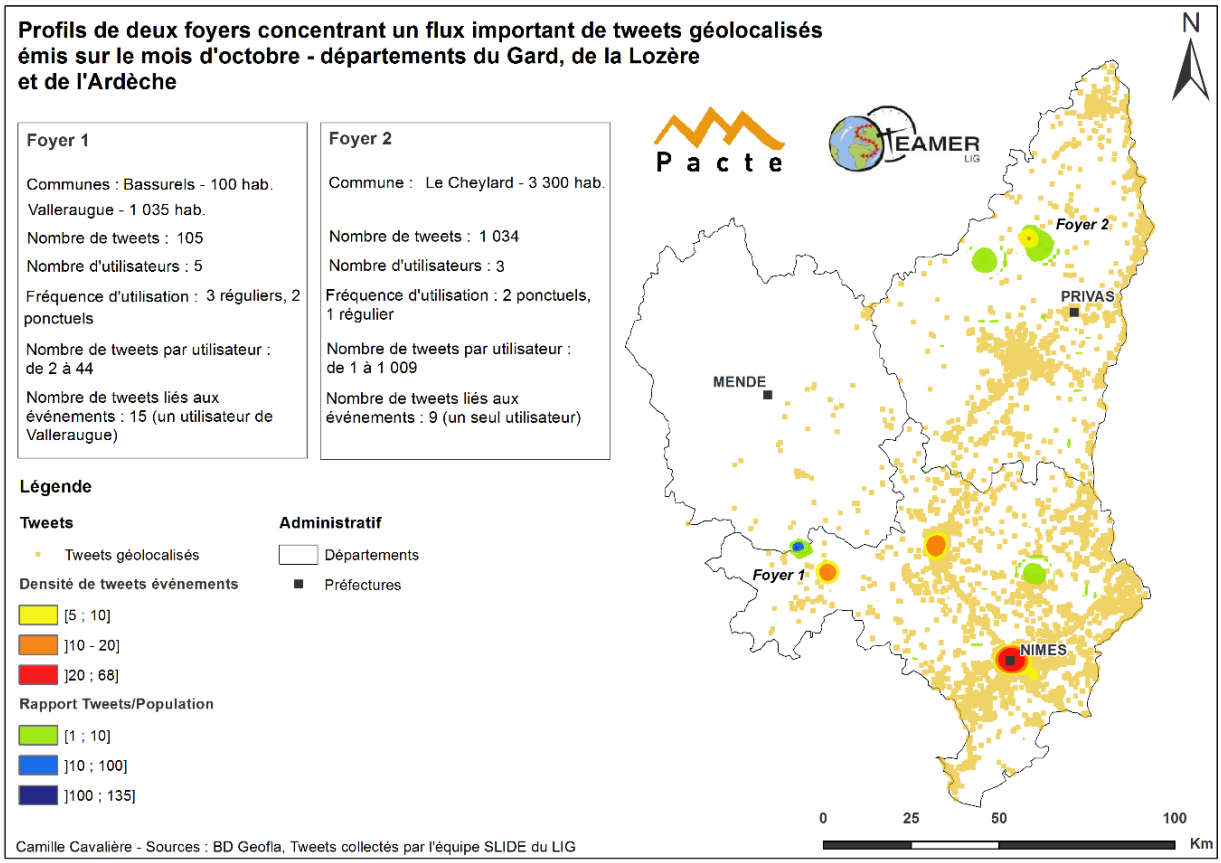


Figure 27 : Localisation et profils de deux espaces remarquables

L'analyse du nombre d'utilisateurs et de tweets enregistrés semble révéler, dans les deux cas, que la forte valeur du rapport de densités tweets/population est plutôt due à la présence d'un utilisateur isolé dont l'utilisation du réseau est fréquente voire intensive. Ainsi, sur les 105 tweets émis sur les communes de Valleraugue et de Bassurels, seuls 15 tweets sont directement liés à l'événementiel hydrométéorologique ; par ailleurs, les utilisateurs ayant contribué à la production de cette information ne se retrouvent pas sur l'îlot situé à la limite entre les deux départements. Après une étude plus précise de la zone, il s'avère que l'îlot en question est situé sur le Mont Aigoual et que les tweets enregistrés sur cette zone proviennent de trois utilisateurs différents dont deux randonneurs. Le troisième a émis quarante-trois tweets, que nous considérons comme *spams*, les tweets se présentant toujours sous forme d'une suite de points d'interrogation.

Le foyer du Cheylard présente un profil plus intéressant : 1 034 tweets ont été envoyés en octobre 2014, dont 97% sont imputés à un seul utilisateur qui est, par ailleurs, l'unique utilisateur de l'espace à contribuer à la production de l'information utile et également celui qui, sur toute la période étudiée, a posté le plus de tweets liés à l'événementiel hydrométéorologique (43 tweets). Toutefois, sur octobre, seuls 9 de ses tweets, sur les 1 009 qu'il a émis ce même mois, évoquent un événement

hydrométéorologique, soit moins de 1%. Au final, les forts rapports de densités observés sont davantage le reflet d'une importante émission d'information bruitée.

3.3.2. Le Cheylard et la crise

Nous nous sommes tout de même intéressés au profil de cet utilisateur, par le biais de l'analyse de l'information utile partagée dans ses tweets pendant les mois d'octobre et de novembre 2014 : la distribution temporelle de ses tweets indique que cet échantillon contenant, sur ces deux mois, un effectif de 40 tweets, s'avère représentatif de la distribution globale. Le graphique suivant (Figure 28) représente le pourcentage de tweets émis à chaque date, par rapport au nombre total de tweets enregistrés sur le corpus de l'utilisateur du Cheylard, et sur le corpus global de tweets filtrés. Le tableau des données est disponible en annexe 21.

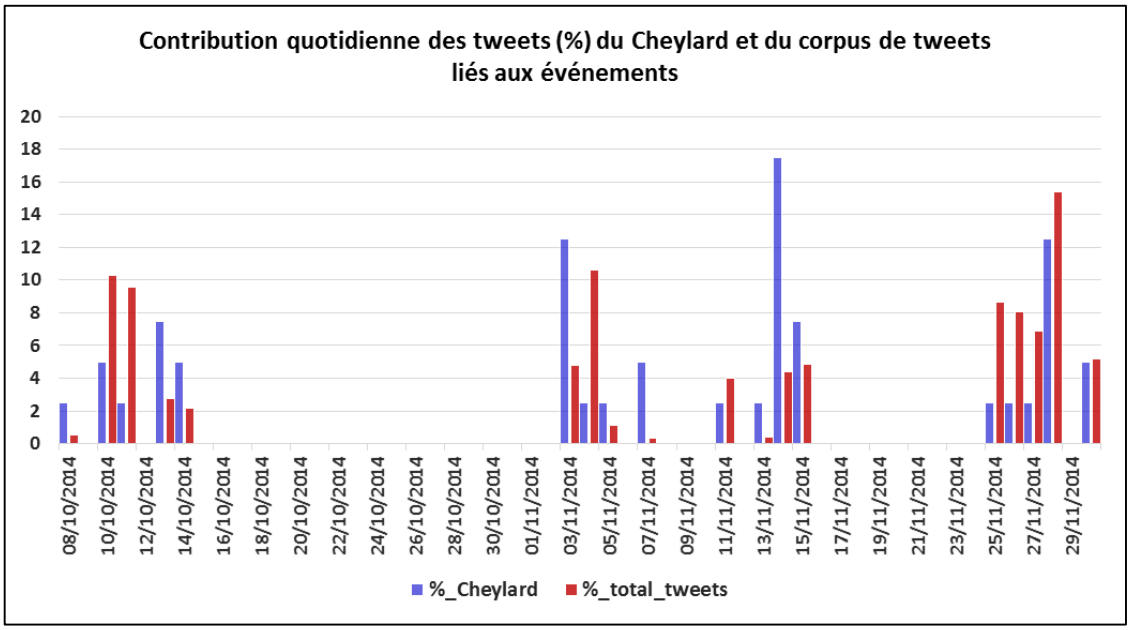


Figure 28 : Comparaison de la contribution quotidienne des tweets liés aux perturbations pour le Cheylard et le corpus filtré global

Les pics de tweets de l'utilisateur du Cheylard correspondent en effet aux dates de crises indiquées par CatNat ; ces principaux pics se retrouvent les 3, 14 et 28 novembre 2014. En revanche, la comparaison avec le corpus global indique l'existence de décalages journaliers, notamment sur les crises des 3 et 4 novembre.

Finalement, l'analyse de la thématique des tweets de cette personne révèle une situation assez paradoxale : le graphique en barres, ci-après (Figure 29), indique le nombre de tweets quotidiens émis par l'utilisateur, en fonction de quatre thèmes : le thème information regroupe tout type d'information liée à une crise qui ne correspond pas à un événement local, c'est-à-dire que cette personne n'assiste pas à l'événement au moment où elle tweete ; les trois autres thèmes – météo (qui regroupe toute information liée à la pluie ou l'orage), inondation et vigilance, évoquent un événement qui la concerne. Par ailleurs, nous avons pu établir, grâce au fichier raster représentant les cumuls de précipitations entre le 8 et le 11 octobre 2014 (cf. Figure 8), que cette personne se situe sur la zone représentée en rouge, qui indique un cumul compris entre 101 et 426 mm d'eau. Si l'on examine, sur ces dates, le contenu textuel de ses tweets liés aux perturbations, on constate qu'elle ne diffuse aucune information sur l'événementiel local ; celui-ci se retrouve uniquement sur les crises de novembre. Les tweets liés à la météo représentent

65% de ses émissions, et les tweets diffusant une information liée à une perturbation qui affecte d'autres territoires en constituent 28%. Le tableau de données est disponible en annexe 22.

Nous avons donc l'exemple d'un utilisateur qui a fortement contribué à produire une information utile, mais qui s'avère finalement peu variée en contenu et peu détaillée sur les événements qui le concernent.

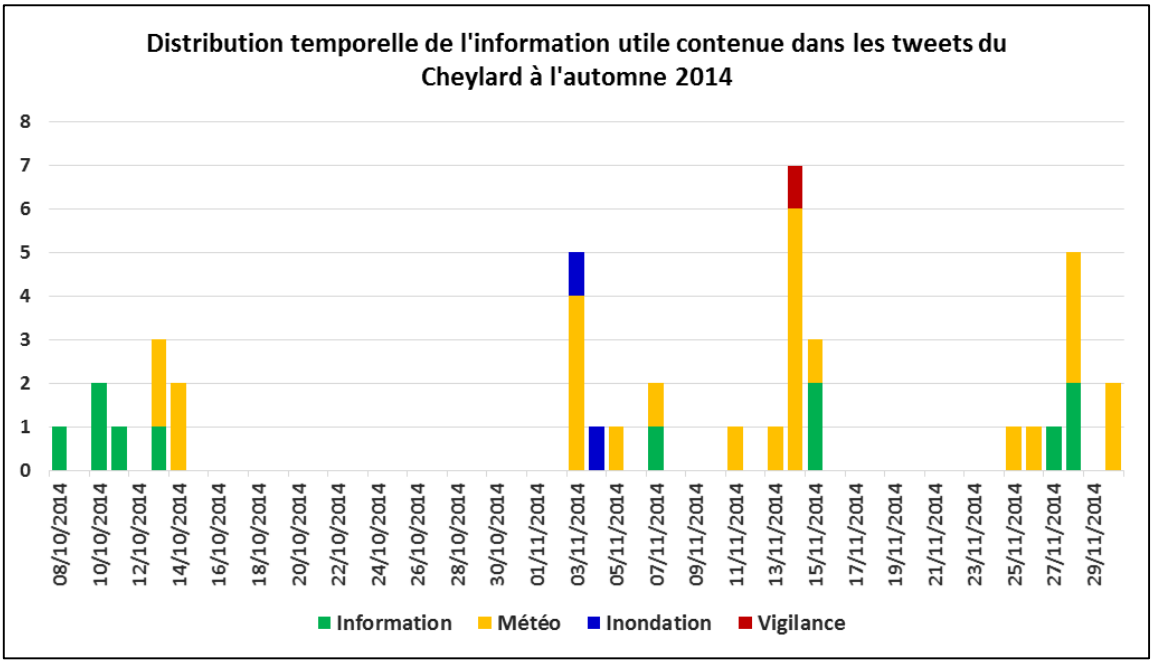


Figure 29 : Répartition des tweets filtrés de l'utilisateur du Cheylard en fonction du thème des tweets

3.4. Au-delà des résultats : retour critique sur les méthodes employées

Cette dernière section adopte un regard critique vis-à-vis des méthodes d'extraction et d'analyse employées pour la construction et l'exploitation de corpus de tweets : elle s'articule autour des aspects essentiels qui nécessitent d'être pris en compte par rapport aux résultats présentés ici et pour l'amélioration de la méthodologie de constitution et d'analyse de l'information.

3.4.1. La part de l'interprétation personnelle dans la sélection des tweets

En premier lieu, nous rappelons que le filtrage des tweets, c'est-à-dire l'extraction en fonction de la présence d'un vocabulaire particulier et l'élimination de l'information bruitée, a été entièrement réalisé de façon manuelle. Si l'exploitation d'un logiciel de fouille de texte a permis de vérifier l'utilisation, dans les tweets, de certains mots déterminés par expertise, la définition de règles destinées à cibler l'information peut d'ores-et-déjà constituer une interprétation des messages ; en outre, il arrive que ces règles ne permettent pas de pallier certains flous. Ainsi, nous avons choisi de considérer qu'un tweet évoquant une situation non accomplie, qui n'existe donc pas au moment où l'utilisateur tweete, était appréhendé comme information bruitée à écarter. Cependant, nous avons rencontré au moins deux types d'information différente pouvant correspondre à cette situation : « *Météo France, une alerte rouge s'il vous plaît, j'ai pas envie d'aller en cours* » ou « *s'il y a l'alerte rouge demain, je vais pas en cours* ». Le premier type de message, qui exprime un souhait, est considéré comme bruit alors que le second évoque un comportement de protection et est, par conséquent, conservé. Or, ces deux utilisateurs peuvent se trouver dans la même situation au moment où ils tweetent : il est en effet possible qu'ils

assistent à un orage ou à des précipitations violentes qui constituent l'élément déclencheur de ces tweets ; ces deux types d'information pourraient donc être considérés comme liés aux perturbations.

A l'étape d'élimination du bruit, nous pouvons également rencontrer des informations qui semblent liées à une perturbation mais dont le contenu textuel n'est pas suffisant pour valider ce lien : il s'agit des tweets qu'on peut qualifier de faux-positifs (Goodchild & Glennon, 2010), c'est-à-dire que l'information contenue dans la donnée ne permet pas de vérifier sa conformité avec le profil recherché. Nous pouvons citer deux types de tweets correspondant à cette situation : *« un carnage sur la route pour rentrée, mdrr »*. Ce tweet est éliminé car il ne contient aucun autre type d'information qui nous permettrait d'établir un lien avec l'événementiel, comme un hashtag – par exemple, #inondation – ou une URL renvoyant à une photographie offrant un aperçu de la situation au moment où l'utilisateur tweete. De plus, en ce qui concerne ce tweet en particulier, la date d'émission (le 22 octobre 2014), qui pourrait constituer un dernier élément de validation, ne correspond pas à l'une des dates de crise CatNat. En revanche, des tweets comme *« Gardon d'Alès au pont vieux @ Centre Ville Alès http://t.co/Xas3eGIaQT »* ou encore *« GRAND AVIGNON | Il y a des jours comme aujourd'hui où l'on apprécierai vraiment un deuxième #pont sur le #Rhône. #avignon #inondations »*, sont conservés dans le jeu final, car ils contiennent respectivement une adresse URL renvoyant à une photographie de la rivière en crue et un *hashtag*.

3.4.2. Améliorer l'extraction de l'information

Nous avons par ailleurs constaté, en calculant les pourcentages de tweets liés à l'événementiel par rapport au flux total, que la participation des utilisateurs à la production d'une information utile via l'utilisation du Smartphone, reste ponctuelle, voire marginale, même pour les jours de crise (cf. Figure 15). L'un des problèmes majeurs reste donc la très forte proportion de bruit, et ce, même après la réalisation de la première étape d'extraction des tweets contenant des mots-clés ou des radicaux précis. Or, ce bruit peut provenir d'une activité intensive du réseau par quelques utilisateurs isolés : nous avons vu, au travers de la construction de courbes de fréquences cumulées, d'une part, que 80% des utilisateurs via le Smartphone ont envoyé tout au plus vingt tweets sur les six mois enregistrés et, d'autre part, que seule une poignée d'utilisateurs envoient plusieurs milliers de tweets. Par exemple, l'utilisateur qui a tweeté le plus sur toute la période (14 196 tweets) n'a envoyé que trois tweets évoquant un événement météorologique. De même, parmi les 180 utilisateurs ayant envoyé plus de 1 000 tweets en six mois, 140 d'entre eux ont tweeté sur une information utile, ce qui représente 771 tweets soit 22% du corpus et, par conséquent, une part non négligeable de l'information finale extraite. Si l'on s'intéresse aux utilisateurs ayant tweeté au moins 3 000 fois, nous en répertorions 24 dont 22 d'entre d'eux ont envoyé 194 tweets, soit 6% du corpus, liés aux perturbations.

A partir de ces données, nous pourrions fixer un seuil limite de tweets enregistrés pour chaque utilisateur, afin de réduire le bruit initial et d'améliorer l'efficacité de l'utilisation du logiciel de fouille de texte.

3.4.3. Réorienter la méthodologie d'analyse

Au travers d'une méthodologie d'analyse des variations spatio-temporelles des émissions de tweets, nous avons cherché à mettre en évidence des espaces où les densités de tweets dépasseraient les densités de population, à observer si ces phénomènes se produisaient pendant les mois concernés par les perturbations cévenoles et s'ils coïncidaient avec des fortes densités de tweets liées à l'événementiel hydrométéorologique. La méthode s'est finalement avérée restrictive, puisque nous avons essentiellement mis en exergue des espaces ruraux sur lesquels nous avons poursuivi les dernières analyses. Celles-ci nous ont finalement montré que ces anomalies de densités étaient davantage le fruit d'utilisateurs isolés témoignant d'un usage très fréquent du réseau et ne participant pas nécessairement

à la production d'une information utile. Toutefois, elles ont pu confirmer l'hypothèse selon laquelle un utilisateur peut partager une information liée à une crise qui ne le concerne pas localement.

Les analyses menées sur ces espaces ne permettent donc pas de mesurer le poids réel de l'information utile lorsqu'on compare la densité de tweets liés à l'événement et la forte valeur du rapport de densités tweets/population. En outre, elles nous ont fourni peu de diversité lexicale, ce qui ne facilite pas la mise en œuvre d'une étude plus approfondie de la variabilité spatio-temporelle du contenu des tweets et la comparaison avec des données pluviométriques.

Les seuils de classes de la cartographie exprimant des rapports entre densités de tweets et densités de population devraient ainsi être révisés : en effet, nous nous sommes uniquement focalisés, jusqu'à présent, sur les espaces dans lesquels ces valeurs sont supérieures à 1. Or, il serait plus pertinent, pour la poursuite des analyses, de concentrer les observations sur la classe]0 ; 1], ce qui permettrait de mettre en évidence des espaces plus peuplés dans lesquels nous pourrions trouver davantage d'utilisateurs contribuant à la production d'une information utile et diversifiée.

3.4.4. Bilan de l'information apportée par les premiers résultats sur l'utilisation du réseau

Les analyses mises en œuvre permettent néanmoins d'apporter quelques éléments de réponse au questionnement initial. Tout d'abord, l'extraction fondée sur la recherche de mots-clés potentiellement employés par les utilisateurs est efficace : les analyses textuelles nous ont en effet confirmé l'utilisation, dans les tweets, des principaux mots-clés qui avaient été inventoriés dans la liste. Par ailleurs, cette dernière pourrait être enrichie au travers de l'ajout de mots dans les langues étrangères, voire de vocabulaire occitan : nous avons en effet uniquement extrait les tweets contenant le radical « *flood* » ou encore le mot « *storm* » pour une recherche de vocabulaire anglais, qui a enrichi le jeu de tweets de 34 tweets différents.

La répartition des effectifs totaux de tweets émis quotidiennement est plus difficile à comparer : en effet, le mois de décembre est incomplet (mais dans le cas contraire, il aurait été marqué par les fêtes de fin d'année), le mois de janvier est perturbé par les attentats et le mois de mars n'est quantitativement pas comparable, en raison de la diminution de moitié du flux de tweets collectés par le serveur. Au final, seul le mois de février peut servir de référence ; nous pouvons confirmer, également au travers des calculs de flux moyens quotidiens sur les différents mois, qu'octobre et novembre 2014 ainsi que janvier 2015 concentrent des flux bien plus conséquents que les douze premiers jours de décembre et le mois de février. De même, si l'on compare les flux d'émission horaires moyens entre le corpus de tweets global et l'ensemble des tweets émis sur les journées de crise, on constate, pour ces jours particuliers, des surplus de tweets sur chaque tranche horaire. Cependant, nous n'avons pas vérifié si ces anomalies étaient le résultat de l'enregistrement de l'événementiel dans les conversations des utilisateurs ; les excédents constatés pourraient ainsi être le résultat du fait suivant : la survenue d'un événement ne modifie pas les habitudes et thèmes habituels des conversations mais se surajoute à ceux-ci.

En ce qui concerne la simultanéité entre le contenu textuel du tweet et de la dynamique événementielle, nous avons pu constater, par le biais de l'analyse de la journée du 28 novembre 2014, l'existence d'une variabilité des flux horaires par rapport à une journée moyenne ; de plus, l'étude du vocabulaire le plus représenté dans les tweets tend à confirmer l'hypothèse selon laquelle l'utilisateur diffuse une information au moment où il assiste à l'événement et, sans doute, au moment où il reçoit une information relative à la crise (comme un département placé en alerte, via un bulletin météorologique). Le manque de données météorologiques ne nous permet pas encore de valider ce résultat.

Enfin, la spatialisation de l'information liée aux crises met en évidence deux phénomènes : certains foyers d'émission sont dus à des utilisateurs isolés et assidus, qui n'apportent pas une grande

diversité lexicale et dont l'information, qui provient ainsi d'une seule personne (donc une seule source) ne peut être corroborée par les tweets d'autres utilisateurs. Par ailleurs, les plus fortes concentrations de tweets se retrouvent dans les grands centres urbains qui, au regard de la faible participation des utilisateurs à la production d'informations utiles, sont susceptibles de nous fournir davantage de renseignements variés, autant en contenus qu'en utilisateurs. L'étude de ces centres nous permettrait en outre d'aborder la question de la mobilité des individus en temps de crise et de la diffusion de l'information en fonction des espaces affectés.

Conclusion

Synthèse de l'exploitation de l'information issue de Twitter

Ce mémoire s'est focalisé sur l'étude de l'information issue des médias sociaux, plus précisément de la plateforme de microblogage Twitter, à travers des événements particuliers, les crises hydrométéorologiques qui affectent le sud-est de la France et qui se manifestent par des orages violents et des crues rapides. Nous avons ainsi axé ce mémoire autour de la production d'un jeu de données exploitables, de l'évaluation des caractéristiques de l'information contenue dans ce jeu et des indications qu'elle était susceptible de nous fournir quant à la dynamique et à la sévérité d'un événement.

Au terme de la présentation de la méthodologie d'analyse et des résultats soumis dans ce mémoire, nous pouvons ainsi retenir plusieurs points essentiels. Tout d'abord, la distribution temporelle de l'information contenue dans le jeu de tweets original que nous avons obtenu et qui inclut des mois en dehors de l'automne, indique une corrélation entre l'existence d'événements et de flux de tweets importants. Nous ne pouvons pas pour autant en conclure que ces forts flux soient le résultat direct des perturbations : nous avons en effet uniquement travaillé avec des tweets géolocalisés, ce qui a pour conséquence de réduire l'information dont nous pourrions disposer. Néanmoins, nous avons pu constater que les flux horaires des journées de crise diffèrent des flux horaires moyens : cette observation provient sans doute du fait qu'un utilisateur tweetant depuis le Smartphone capture et émet une information instantanée, liée à un phénomène exceptionnel qui se déroule devant lui. Dans ce cas, il est fort probable, comme nous avons pu le constater en analysant la distribution horaire de tweets contenant un vocabulaire particulier, que le flux de tweet lié à un événement particulier suive sa dynamique.

Nous avons donc pu constater que la perturbation cévenole est enregistrée dans le contenu textuel des tweets ; en outre, nous avons observé une distribution régulière, sur toute la période, de tweets liés aux simples perturbations météorologiques, ce qui signifierait que celles-ci préoccupent certains utilisateurs.

En ce qui concerne le diagnostic de la sévérité des crises par les tweets et surtout, de sa correspondance avec l'échelle de gravité de CatNat, les résultats s'avèrent plus mitigés : si la cartographie de densité et la distribution quantitative quotidienne des tweets liés aux perturbations semblent indiquer que le mois de novembre a été le plus perturbé, la crise d'octobre 2014, survenue du 8 au 11 et d'une échelle de gravité de 3/5, est bien moins évidente si l'on se réfère uniquement aux tweets. L'étude des variations spatio-temporelles des tweets peut donc se révéler aléatoire : nous avons en effet mis en évidence que la plupart des contributeurs à l'information utile, via Smartphone, sont très ponctuels et n'envoient qu'un ou deux tweets sur toute la période enregistrée. En revanche, certains utilisateurs sont très producteurs et peuvent générer un îlot de forte densité de tweets en un point précis de l'espace : nous pouvons donc nous trouver dans une situation telle que : un espace particulièrement affecté par une crise est peu fourni en informations car peu de personnes tweetent via Smartphone ; un espace peu affecté est fourni car un ou quelques utilisateurs tweetent énormément. Ces cartes ne sont donc pas suffisantes pour servir de base solide afin de valider l'hypothèse selon laquelle les tweets enregistrent la sévérité de la crise. La méthodologie mise en œuvre doit donc être réorientée.

Perspectives d'orientation de la méthodologie

La poursuite des travaux menés dans le cadre du stage peut se structurer autour de deux axes principaux : une étude couplée, entre différents types de données, de leurs variations spatio-temporelles ainsi que la poursuite de la structuration du corpus afin d'en extraire une information ciblée sur des critères prédéfinis.

Nous pourrions donc mener, à l'échelle d'une ville - voire d'un bassin versant - dans laquelle nous disposons suffisamment d'information, variée en termes de contenu et provenant de divers utilisateurs, une étude spatio-temporelle conjointe du contenu des tweets géolocalisés et de la dynamique des précipitations. L'objectif d'une telle analyse consisterait à comparer l'évolution du contenu des tweets et la dynamique de l'événement, par tranche horaire. Cette étude pourrait, par ailleurs, être fondée sur l'ensemble des tweets confondus, c'est-à-dire sans distinction sémantique, de manière à visualiser la place occupée par le tweet lié à l'événement pendant sa survenue, parmi le tweet du quotidien. Une seconde étude serait, quant à elle, focalisée sur le corpus de tweets liés à l'événement, afin de visualiser la variabilité du vocabulaire spécifique à la perturbation, ainsi que le flux de tweets par rapport à la dynamique événementielle. Une telle analyse pourrait également nous permettre d'évaluer l'éventuelle existence de processus particuliers de diffusion de l'information via le réseau (en comparant l'information contenue dans le texte du tweet). Il serait également pertinent de mener plusieurs analyses, pour un même espace, en fonction de dates de crises différentes, ce qui nous permettrait de savoir si les mêmes utilisateurs sont actifs et d'évaluer l'évolution textuelle des tweets d'une crise à l'autre. Ce type d'analyse implique cependant d'effectuer, au préalable, une classification thématique des tweets.

Nous pourrions en effet, pour construire un jeu de données élaborées et finalisées, reprendre ce point évoqué dans les références bibliographiques consultées mais écarté de la méthodologie mise en œuvre dans le cadre de ce travail. Il s'agirait en effet de poursuivre la structuration du corpus des 3 457 tweets liés aux événements hydrométéorologiques en définissant une typologie fondée sur des critères précis destinés à rattacher les tweets à un thème : les tweets pourraient ainsi être classés selon qu'ils évoquent un événement météorologique, un événement hydrologique, une phase de préparation à la crise, des comportements de protection ou de mise en danger, des dégâts occasionnés, etc. et être de nouveau comparés à la temporalité de l'événement afin d'observer une éventuelle correspondance entre les différentes phases de la crise évoquées dans les tweets et la dynamique réelle de l'événement.

Références bibliographiques

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*, Volume 15, n°3, mai 2013, pp. 72-82.

Bailly A. S. (1996). Environnement, risques naturels, risques de sociétés. in A. S. Bailly (dir.), Risques naturels, risques de sociétés, *Economica*, p.1-5.

Brovelli M. A., Zamboni G., Muñoz C. A., Bonetti A. (2014). Exploring Twitter georeferenced data related to flood events : an initial approach. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, juin 2014. Publié en ligne sur : http://www.agile-online.org/Conference_Paper/cds/agile_2014/agile2014_147.pdf

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 Conference on Information Systems for Crises response and Management (ISCRAM)*. Publié en ligne sur : http://iscram2014.ist.psu.edu/sites/default/files/misc/proceedings/p24.pdf

Davoine P-A. (2014). « Contributions géomatiques pour la gestion des risques naturels : modélisation, géovisualisation, acquisition ». Habilitation à diriger des recherches, Université de Grenoble, 201 pages.

De Blomac F. (2015). Inondations, quoi de neuf ? in V. Baculard (dir.), *Décryptagéo – Le Mag*, n°167, mai 2015, pp. 7-14.

Dollfus O. et D'Ercole R. (1996). Les mémoires des catastrophes au service de la prévision et de la prévention des risques naturels. in A. S. Bailly (dir.), Risques naturels, risques de sociétés, *Economica*, p.7-18.

Goodchild M. (2009). Citizens as sensors : the world of volunteered geography. *GeoJournal*, Volume 69, n°4, août 2007, pp. 211-221.

Goodchild M., Glennon J. (2010). Crowdsourcing geographic information for disaster response : a research frontier. *International Journal of Digital Earth*, Volume 3, n°3, 2010, pp. 231-241.

Herfort 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. In Huerta J., Schade S, Granell C. (eds) : *Connecting a Digital Europe through Location and Place*. Volume 1, Springer International Publishing, pp. 55-71.

Imran M., Elbassuoni S., Castillo C., Diaz F., Meier P. (2013). Extracting information nuggets from disaster-related messages in social media. *Proceedings of the 10th International Conference on Information Systems for Crises response and Management (ISCRAM)*. Publié en ligne sur : http://qcri.org.qa/app/media/1843

McDougall K. (2012). An assessment of the contribution of volunteered geographic information during recent natural disasters. *Spatially enabling government, industry and citizens : research and development perspectives.* Needham : GSDI Association Press, pp. 201-214.

Millerand F., Proulx S., Rueff J. (2010). Web social. Mutation de la communication. Presses de l'université du Québec, 396 pages.

Roy Chowdhury S., Amer-Yahia S., Castillo C. (2013). Tweet4act : using incident specific profiles for classifying crises-related messages. *Proceedings of the 10th International Conference on Information Systems for Crises response and Management (ISCRAM)*. Publié en ligne sur: https://www.cs.auckland.ac.nz/~asghar/papers/ISCRAM13-Tweet4act.pdf

Schade S., Diaz L., Ostermann F., Spinsanti L., Luraschi G., Cox S., Nuñez M., De Longeville B. (2011). Citizen-based sensing of crisis event : sensor web enablement for volunteered geographic information. *Applied Geomatics*, Volume 5, n°1, mars 2013, pp. 3-18.

Shu H., Spaccapietra S., Parent C. (2003). Uncertainty of geographic information and its support in MADS. Proceedings of the International Symposium on Spatial Data Quality. Publié en ligne sur : http://infoscience.epfl.ch/record/99139/files/GIUncertainty_HK.pdf&sa=U&ei=YoFiU-bnLeX-ygPauoLgDQ&ved=0CDYQFjAF&usg=AFQjCNHuxf-MS1ut99cmv_8gAEjWwjmagQ

Table des figures

Figure 1 : L'Ouvèze en crue sous le Pont Romain de Vaison-la-Romaine, le 22 septembre 1992 *(Météo France pluies extrêmes)* 7

Figure 2 : Carte de présentation du terrain d'étude 10

Figure 3 : Présentation de la cartographie interactive *The one million tweets map* – Zoom sur l'agglomération grenobloise et filtrage des tweets sur un *hashtag* signalé comme populaire le 30 mars 2015 15

Figure 4 : Exemple de recherche de tweets 15

par mot-clé 15

Figure 5 : *Crowdmap* et information envoyée à l'ICCROM pour l'inventaire des dégâts affectant le patrimoine népalais *(ICCROM)* 17

Figure 6 : Schématisation de l'infrastructure de collecte des tweets 21

Figure 7 : Schéma de la table `public.tweet` *(LIG, SLIDE)* 22

Figure 8 : Raster du cumul de pluie sur un événement *(OHMCV, LTHE)* 23

Figure 9 : Diagramme cartographique pour la journée du 28 novembre 2014 (seuils de cooccurrence compris entre 8 et 10) 26

Figure 10 : Information bruitée extraite 27

Figure 11 : Courbe des fréquences cumulées – nombre de tweets envoyés par utilisateur 32

Figure 12 : Carte de localisation des tweets géolocalisés émis sur la période totale**Erreur ! Signet non défini.**

Figure 13 : Comparaison mensuelle des tweets géolocalisés émis quotidiennement 34

Figure 14 : Nombre moyen de tweets émis par tranche horaire 35

Figure 15 : Carte des variations spatio-temporelles de densité d'émission des tweets 36

Figure 16 : Carte des variations spatio-temporelles des tweets rapportés à la population 37

Figure 17 : Fréquences cumulées des utilisateurs en période de 38

Figure 18 : Carte de localisation des tweets liés aux perturbations hydrométéorologiques 39

Figure 19 : Associations lexicales et mots-clés extraits depuis KH Coder 41

Figure 20 : Distribution temporelle des tweets contenant les mots-clés principaux 42

Figure 21 : Distribution quotidienne des tweets liés aux événements et comparaison avec les dates de crises hydrométéorologiques 44

Figure 22 : Comparaison des flux horaires de la période globale (bleu) et des journées de crise (rouge) 45

Figure 23 : Distribution horaire du corpus de tweets liés aux perturbations 45

Figure 24 : Comparaison des flux horaires (%): période normale (bleue), tweets évoquant la crise (rouge) 46

Figure 25 : Distribution horaire des tweets contenant les principaux mots-clés, le 28 novembre 2014 46

Figure 26 : Carte des variations spatio-temporelles de densité d'émission des tweets liés aux perturbations cévenoles 48

Figure 27 : Localisation et profils de deux espaces remarquables 49
Figure 28 : Comparaison de la contribution quotidienne des tweets liés aux perturbations pour le Cheylard et le corpus filtré global 50
Figure 29 : Répartition des tweets filtrés de l'utilisateur du Cheylard en fonction du thème des tweets 51

Table des tableaux

Tableau 1 : Règles de syntaxe des requêtes d'extraction des tweets 27
Tableau 2 : Effectifs de tweets enregistrés pour les cinq mots-clés les plus utilisés 40
Tableau 3 : Nombre de tweets par mot-clé extrait de l'analyse lexicale 41

Table des matières

Remerciements 2
Liste des acronymes 3
Sommaire 4
Présentation du stage et de la structure d'accueil 5
Introduction 6

1. Twitter, une nouvelle source de données ouvertes et massives : perspectives et contraintes 11
1.1. Des concepts-clés autour du web social 11
1.1.1. Quelques définitions autour des médias sociaux 11
1.1.2. L'affirmation des médias sociaux dans la communauté scientifique 12
1.1.3. Quel degré de confiance attribuer à cette source d'information ? 13
1.2. Visualiser et extraire les tweets 14
1.2.1. Les outils de visualisation 14
1.2.2. L'outil d'acquisition de l'information brute 16
1.3. Sélectionner et structurer l'information liée à un événement particulier 16
1.3.1. Le *crowdsourcing* et la gestion de crise 16
1.3.2. Le cas de Twitter : extraire un corpus de tweets pertinents 17
1.3.3. Classification thématique d'un jeu de tweets 19

2. Du tweet à la carte : extraire, structurer et représenter l'information 21
2.1. Présentation des données exploitées dans le cadre du projet 21
2.1.1. Infrastructure de collecte et tweets 21
2.1.2. Les crises hydrométéorologiques 23
2.1.3. Autres données géographiques 24
2.2. De l'information brute à la constitution d'un corpus de tweets 24
2.2.1. La base de données, un outil de nettoyage et de structuration 25
2.2.2. Construire un corpus de tweets relatifs à l'événementiel hydrométéorologique 25
2.3. Exploiter les jeux de tweets : quels usages pour quelle distribution des données ? 28
2.3.1. L'usage quantitatif de Twitter 28
2.3.2. Visualisation et analyse des variations spatio-temporelles d'émission des tweets 29

3. Analyse et représentation des tweets : quelles informations pour quelles distributions ? 31
3.1. Analyse du corpus de tweets enregistrés sur la période étudiée 31
3.1.1. L'utilisation de Twitter en chiffres 31

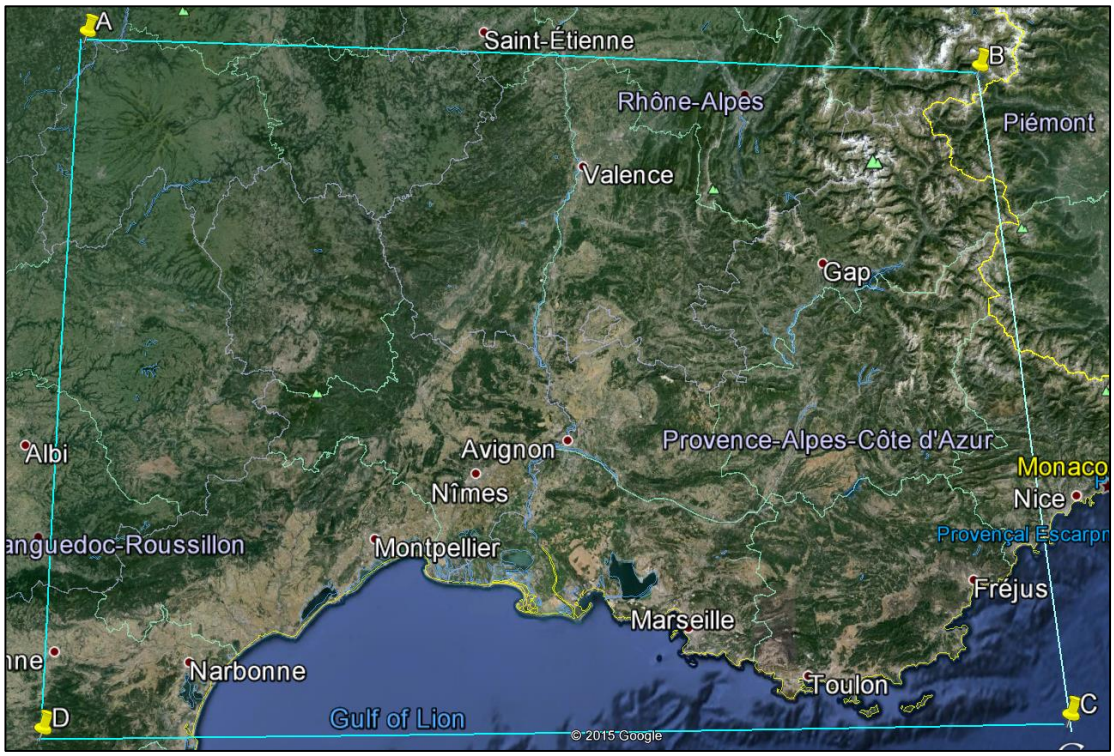
3.1.2. La distribution temporelle des tweets 33
3.1.3. Les variations spatio-temporelles de la distribution des tweets 36
3.1.4. Détection des anomalies de densités 37
3.2. Analyse du corpus de tweets liés à l'événementiel hydrométéorologique 38
3.2.1. L'utilisation de Twitter pendant des périodes de crise 38
3.2.2. Etude de la distribution lexicale des tweets 39
3.2.3. La distribution temporelle des tweets liés aux perturbations 42
3.2.4. La variabilité spatio-temporelle des tweets liés aux événements 47
3.3. Etude d'espaces remarquables 48
3.3.1. Le profil des espaces particuliers 48
3.3.2. Le Cheylard et la crise 50
3.4. Au-delà des résultats : retour critique sur les méthodes employées 51
3.4.1. La part de l'interprétation personnelle dans la sélection des tweets 51
3.4.2. Améliorer l'extraction de l'information 52
3.4.3. Réorienter la méthodologie d'analyse 52
3.4.4. Bilan de l'information apportée par les premiers résultats sur l'utilisation du réseau 53

Conclusion 55
Références bibliographiques 57
Table des figures 59
Table des tableaux 61
Table des matières 62
Table des annexes 64

Table des annexes

Annexe 1 : Localisation et coordonnées de la zone extraite 65
Annexe 2 : Extrait du tableur CatNat sur les risques d'inondation et de ruissellement 66
Annexe 3 : Description de la préparation du je de tweets original 67
Annexe 4 : Liste de mots-clés recherchés dans les tweets 68
Annexe 5 : Extraction de tweets pour des dates choisies 69
Annexe 6 : Etapes pour l'analyse des utilisateurs 70
Annexe 7 : Etapes mises en œuvre pour l'analyse des flux 71
Annexe 8 : Carte de localisation des villes principales 72
Annexe 9 : Fréquences cumulées des utilisateurs de Twitter 72
Annexe 10 : Nombre de tweets émis en fonction des jours 73
Annexe 11 : Nombre de tweets émis en fonction de l'heure 75
Annexe 12 : Carte des densités de population 76
Annexe 13 : Fréquences cumulées des utilisateurs en période de crise 76
Annexe 14 : Nombre de tweets extraits et filtrés par mots-clés recherchés 77
Annexe 15 : Distribution quantitative et temporelle des tweets filtrés par la méthode expertisée 78
Annexe 16 : Distribution des effectifs de tweets contenant les mots-clés principaux 79
Annexe 17 : Distribution temporelle des tweets du corpus liés aux événements 81
Annexe 18 : Comparaison des flux horaires de la période globale et des journées de crise 84
Annexe 19 : Comparaison des pourcentages d'émission horaires des tweets liés à la crise du 28 novembre 2014 et du jeu de tweets global 85
Annexe 20 : Distribution horaire des tweets contenant les principaux mots-clés employés par les utilisateurs, le 28 novembre 2014 86
Annexe 21 : Comparaison de la contribution quotidienne des tweets liés aux perturbations pour le Cheylard et le corpus filtré global 86
Annexe 22 : Distribution quotidienne des tweets du Cheylard en fonction des thèmes 87

Annexe 1 : Localisation et coordonnées de la zone extraite



Coordonnées géographiques des points (WGS 84, degrés décimaux) :

A (45, 402843 ; 2,295547)
B (45,226058 ; 6,913276)
C (42,918327 ; 7,135171)
D (42,907016 ; 2,305051)

Annexe 2 : Extrait du tableur CatNat sur les risques d'inondation et de ruissellement

Date début	Date fin	Risque	Nature de l'événement	Epicentre	Nbre évacués	Nbre blessés	Nbre victimes	Dommages (coût estimé en G€)	Echelle de gravité	commentaires
27/11/2014	30/11/2014	INON	Nouvel épisode d'intempéries pluvio-orageuses et d'inondations pour les départements méditerranéens : 5 morts	Méditerranée	3500	3	5	250 S	3/5	Inondation et coulée de boue; Sous-peril: débordement de plaine
24/11/2014	27/11/2014	INON	Inondations localisées en région Languedoc-Roussillon et dans le Var	Languedoc-Roussillon et dans le Var	1500			0,00	2/5	Ruissellement urbain - inondations et coulée de boue
14/11/2014	15/11/2014	INON	Nouvel épisode pluvio-orageux intense dans le Sud-Est de la France : 6 morts	Sud-Est	0		6	0,00	3/5	Crue & débordement torrentiel
09/11/2014	11/11/2014	INON	Pluies diluviennes et inondations dans le Sud-Est de la France	Sud-Est	0		1	0,00	2/5	Crue & débordement torrentiel

Annexe 3 : Description de la préparation du jeu de tweets original

Après l'installation de PostGreSQL et de l'extension PostGIS, la base de données `Twitter` est créée sous le modèle de la base `template_postgis`, qui permet alors d'activer des fonctionnalités de géométrie. Le codage choisi est l'UTF-8 qui assure l'affichage des caractères spéciaux comme les accents ou autres symboles insérés dans les tweets. Une table `tweets` est ensuite créée : cinq champs sont ajoutés, de manière à respecter les colonnes du fichier CSV :

- le champ `tweet_id` de type numérique
- le champ `user_id` de type numérique
- le champ `text` de type chaîne de caractères
- le champ `emitted` qui regroupe date et heure d'émission ; ce champ est volontairement créé en type texte afin de pouvoir, pour les besoins des analyses ultérieures, le tronquer en deux colonnes distinctes : date et heure
- le champ `gps` de type géométrie.

Le fichier CSV est importé dans la table `tweets` à l'aide de la requête `COPY CSV File` :

```
COPY tweets
FROM 'g:/stage/tweets_extraction/1_donnees_brutes/tweets_pgis.txt' DELIMITER
','
```

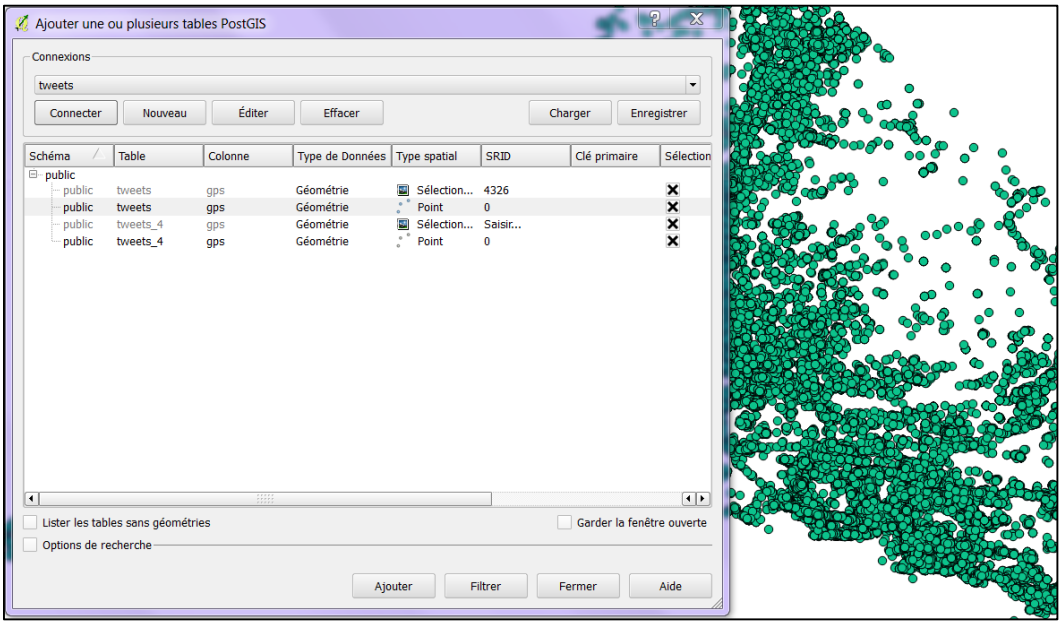
La seconde étape de structuration consiste à tronquer le champ `emitted` en deux nouveaux champs `date` et `heure` et à supprimer le fuseau horaire : la fonction `SUBSTR(string, start, lengh)` est alors appliquée deux fois afin d'effectuer deux troncatures, respectivement entre la date et l'heure, puis entre l'heure et le fuseau horaire ; ce dernier champ étant inutile, il ne sera pas conservé dans la structure de la table finale :

```
CREATE TABLE tweets_2 AS
SELECT tweet_id, user_id, text, gps, SUBSTR(emitted, 1, 11)
CREATE TABLE tweets_3 AS
SELECT tweet_id, user_id, text, gps, SUBSTR(emitted, 12, 8)
```

	tweet_id [PK] numeric	user_id numeric	text character varying	gps geometry	date text	heure text
1	519840665472335872	1337099178	Normalement ce week j'ai les air force one pers	01010000001A6CEA3C2A3E0A4008CDA	2014-10-08	15:24:07
2	519840740093231104	815724566	Pour dire à quel point j'ai raté les ds, j'ai m	0101000000471E882CD20408409BE44	2014-10-08	15:24:25
3	519840785618202624	613346063	@caredluke moi aussi □□	01010000004A09C1AA7A991340C8091	2014-10-08	15:24:36
4	519840795403505664	895839127	Vidaaaaa http://t.co/WLHNEj358A	010100000055336B29204D134052651	2014-10-08	15:24:38
5	519840804672905217	1594402945	Ouais mais j'cours pas derrière les pd comme to	0101000000F1129CFA40721440997FF	2014-10-08	15:24:40
6	519840819239747584	107723347	@Babineuh @Laura sc11 ton grd a devrait avoir	010100000004ADC090D58D15403CA1D	2014-10-08	15:24:44
7	519840846678872064	82117490	Etihad launches 'alliance' across six airlines	01010000004087F9F2021C1C40508F6	2014-10-08	15:24:50
8	519840890794557440	510966796	jsp pq mais ds ma tête Clara c ma sœur :/	0101000000E3DF675C38C01B40EDB60	2014-10-08	15:25:01
9	519848639439859712	225761033	"@RemballerparSMS: http://t.co/dVQ1vnOZX6"	010100000019ADA3AA09D21B40E1B54	2014-10-08	15:55:48
10	519848726828167168	1140574464	The clocks run out, time's up, OVER BLAHH	01010000002997C62FBC2214403354C	2014-10-08	15:56:09
11	519848768116903936	944449992	si c vrai la tracklist de the weeknd…vs allez m	0101000000C2C1DEC4905C0D40D940B	2014-10-08	15:56:19
12	519848848370704384	125136689	Il y a 3 façons de disparaitre. La première est	0101000000664E97C5C4861B4020F0C	2014-10-08	15:56:38
13	519848908504449024	944449992	ca va etre pire que days before rodeo que jecou	0101000000E4BB94BA645C0D409160A	2014-10-08	15:56:52
14	519848955463868416	2759086891	@drugxmadison tu est d ou	0101000000088BA0F406A031640A9A5B	2014-10-08	15:57:03
15	519849042491478016	847992836	Corleone l'album qui est au dessus	0101000000A8E3310395D116403E7AC	2014-10-08	15:57:24
16	519849051127549952	1137803312	Grosse folle ma mère	0101000000BBD573D2FBE609404E0E9	2014-10-08	15:57:26
17	519849051765096449	815724566	Moh les copines qui m'attendent pour faire chic	0101000000F92EA52E19070840D1E97	2014-10-08	15:57:26
18	519849143905562624	1568827795	Je vois des Alyson partout Mdrrr □	0101000000E41071732A3916403BFF7	2014-10-08	15:57:48
19	519849187761213440	455152174	Tournage de Cain #portautonome #Marseille: des	01010000009605137F14651540E275F	2014-10-08	15:57:59

Extrait de la table `tweets` après troncature des champs

La couche d'entités ponctuelles géoréférencées est alors générée sous le SIG QGIS : une connexion est créée entre le logiciel et la base de données `Twitter`, à partir de l'outil *« Ajouter une couche PostGIS »* ; elle permet à QGIS de lister les tables contenant une colonne géométrique. L'utilisateur sélectionne le type de géométrie ainsi que le SRID de la couche - respectivement ponctuelle et 4326 pour le système WGS 84 dans notre cas– puis le logiciel affiche une couche temporaire qui sera enregistrée au format esri.



Connexion de la base de données vers QGIS pour la création de la couche d'entités ponctuelles

Cette nouvelle couche doit ensuite être projetée dans un système de coordonnées approprié, le Lambert 93, à l'aide de l'outil « *Reproject Layer* » de la boîte de traitement de QGIS ; enfin, la couche contenant les points (donc les tweets) enregistrés sur notre terrain d'étude est obtenue par requête spatiale entre la couche des huit départements étudiés – issue des départements de la BD GEOFLA® – et la couche de tweets générée lors de l'étape précédente.

Annexe 4 : Liste de mots-clés recherchés dans les tweets

Météo	Cévenol
	Déluge
	Eclair
	Flotte
	Foudre
	Intempérie
	Orage
	Pleut
	Pluie
	Précipitations
	Storm
	Tempête
	Tonnerre

Cours d'eau	Agout
	Argens
	Cèze
	Dourbie
	Drôme
	Gard Gardon
	Lez
	Mosson
	Nartuby
	Tarn
	Rhône
	Vidourle

Villes	Alès
	Anduze
	Baume
	Calmette
	Grabels
	Nîmes
	Pérols
	Sète
	Uzès
	Vigan

Période Pré-Crise	Alerte
	Evacuation
	Prudence
	Vigilance

Période Post-Crise	Arbre
	Assurance Assureur
	Catastrophe
	Dégâts Dommages
	Déblayer
	Nettoyer
	Sinistre

Période Syn-Crise	Boue
	Circulation
	Crise
	Critique
	Crue
	Débordement
	Digue Canal Chenal Bassin Egout
	Eau + Monte
	Flood
	Inondation
	Maison Foyer Logement Habitation
	Niveau + Eau
	Panne Electricité Coupure
	Potable
	Rivière
	Route Pont
	Ruissellement
	Secours Pompiers
	Torrent
	Urgence

Annexe 5 : Extraction de tweets pour des dates choisies

L'extraction des tweets émis aux dates qui sont retenues requiert, dans un premier temps, l'exportation de la table attributaire du shapefile des tweets postés dans la zone d'étude, en un fichier CSV (voir annexe 3). Ce fichier doit ensuite être importé dans une nouvelle table `tweets_4` créée dans la base de données `Twitter`, selon la procédure habituelle. La base de données contient alors la table des tweets enregistrés dans les huit départements. L'exécution successive de deux requêtes permettra de créer une troisième table à partir de la sélection du champ `text` de la table `tweets_4`, puis d'exporter cette nouvelle table en fichier CSV, qui peut être importé dans KH Coder :

- Sélectionner les tweets d'un jour précis :

```
CREATE TABLE crise_j AS
SELECT text FROM tweets_4
WHERE date LIKE '2014-10-10' ;
COPY crise_j TO 'g:/crise_j.csv' DELIMITER ','
```

- Sélectionner les tweets d'un mois entier :

```
CREATE TABLE crise_m AS
```

```
SELECT text FROM tweets_4 WHERE date LIKE '2014-10-%' ;
COPY crise_m TO 'g:/crise_m.csv' DELIMITER ','
```

Annexe 6 : Etapes pour l'analyse des utilisateurs

L'utilisation de la base de données `Twitter` est de nouveau nécessaire pour la mise en œuvre de cette analyse ; dans un premier temps, nous allons créer une table `tweets_crise` dans laquelle sera importé un fichier CSV généré à partir de la table attributaire de la couche des tweets en rapport avec les perturbations. Puis nous allons appliquer deux requêtes de sorte à créer deux nouvelles tables, `users` et `crise_users`, correspondant respectivement à la table de tous les utilisateurs de la période et des utilisateurs répertoriés dans les tweets liés aux perturbations ; dans ces deux tables seront indiqués les numéros d'utilisateurs et le nombre de tweets qu'ils ont émis, grâce à la combinaison des fonctions SQL `Sum` et `Group By`. Pour ce faire, il est au préalable indispensable d'associer, à chaque ligne des tables `tweets_crise` et `tweets_4`, la valeur 1 (chaque ligne correspondant à un tweet). Un nouveau champ « `valeur` » de type numérique est donc créé puis l'application de la fonction `Update` permet d'affecter à chaque ligne la valeur souhaitée. Après l'application des fonctions `Sum` et `Group By`, des fichiers CSV sont exportés et analysés sous tableur.

- Syntaxe des requêtes :

```
UPDATE tweets_4
SET valeur = '1' ;
UPDATE tweets_crise
SET valeur = '1' ;
CREATE TABLE users AS
SELECT user_id, SUM(valeur)FROM tweets_4
GROUP BY user_id
ORDER BY sum DESC ;
CREATE TABLE crise_users AS
SELECT user_id, SUM(valeur)FROM tweets_crise
GROUP BY user_id
ORDER BY sum DESC ;
COPY users TO 'g:/users.csv' DELIMITER ','CSV HEADER ;
COPY crise_users TO 'g:/crise_users.csv' DELIMITER ','CSV HEADER
```

	user_id numeric	sum numeric
1	2455260391	14196
2	819893718	9841
3	2171002048	8883
4	1612639134	8722
5	2236609791	7142
6	609380794	6892
7	1118142301	6663
8	2169617979	6050
9	1321303915	5431
10	331827739	4983
11	1396625233	4489
12	124970015	4397
13	1101256837	3822
14	1940458094	3805
15	579033099	3542

	user_id numeric	sum numeric
1	124970015	43
2	749686110	38
3	249822976	32
4	223660979	27
5	615004763	26
6	137355603	24
7	111814230	24
8	897182526	23
9	332306489	21
10	287108415	21
11	457778883	19
12	495703548	19
13	751263060	18
14	217129346	18
15	427301182	17

Extraits des tables* `users` *(gauche) et* `crise_users` *(droite)

Annexe 7 : Etapes mises en œuvre pour l'analyse des flux

Le calcul des effectifs totaux de tweets émis en fonction des jours requiert une intervention dans la base de données `Twitter` : deux nouvelles tables, `time` et `time_crise`, doivent être créées dans lesquelles seront enregistrés, à l'aide des fonctions `Sum` et `Group By`, les effectifs quotidiens. Les tables seront ensuite exportées dans un fichier au format CSV :

```
CREATE TABLE time AS
SELECT date, SUM(valeur)FROM tweets_4
GROUP BY date ;
CREATE TABLE time_crise AS
SELECT date, SUM(valeur)FROM tweets_crise
GROUP BY date ;
COPY time TO 'g:/time.csv' DELIMITER ','CSV HEADER ;
COPY time_crise TO 'g:/time_crise.csv' DELIMITER ','CSV HEADER
```

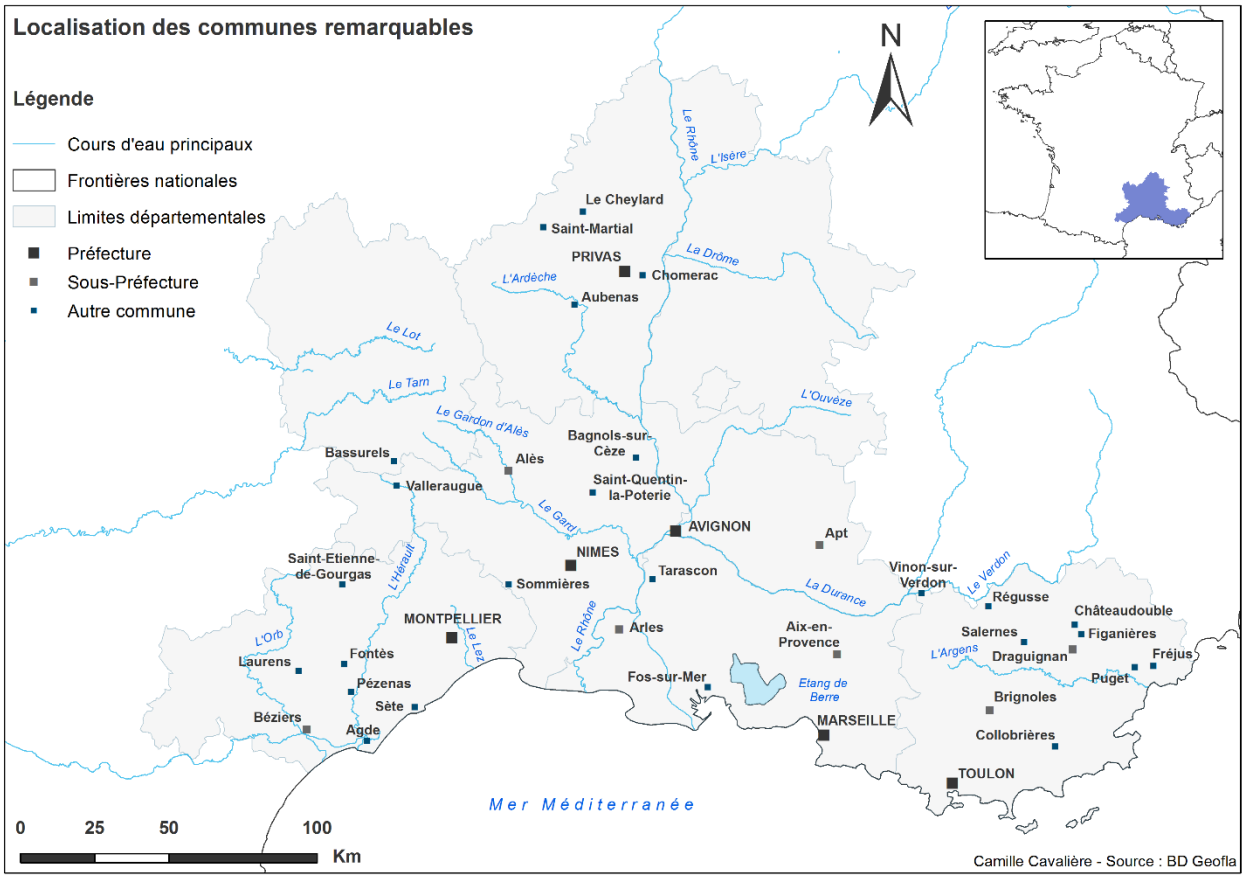
	date text	sum numeric
1	2015-01-07	12039
2	2015-03-22	4830
3	2014-11-30	12733
4	2015-02-24	4872
5	2014-11-13	9230
6	2015-02-19	4179
7	2014-11-02	12892
8	2015-02-09	4115
9	2015-02-27	4866
10	2015-01-15	7526
11	2014-11-01	11356
12	2015-03-23	4112
13	2015-03-28	3757
14	2014-11-27	9608
15	2015-01-25	9485

	date date	sum numeric
1	2014-10-08	11
2	2014-10-09	29
3	2014-10-10	208
4	2014-10-11	194
5	2014-10-13	56
6	2014-10-14	44
7	2014-10-15	7
8	2014-10-16	5
9	2014-10-17	1
10	2014-10-18	2
11	2014-10-20	2
12	2014-10-21	3
13	2014-10-22	6
14	2014-10-23	4
15	2014-10-24	1

Extraits des tables* `time` *(gauche) et* `time_crise` *(droite)

Les calculs des flux horaires sont effectués sur les tables `tweets_4` et `tweets_crise`, à partir de requêtes simples, de type : `SELECT Count(*) FROM tweets_4 WHERE heure LIKE '00%'` lorsque l'on souhaite connaître le nombre de tweets de la table correspondante postés entre minuit et une heure.

Annexe 8 : Carte de localisation des villes principales



Annexe 9 : Fréquences cumulées des utilisateurs de Twitter

Nombre de tweets	Effectif	Fréquence	Fréquences cumulées
]0 ; 1]	8 797	0,30661	0,30661
[0 ; 5]	8 751	0,30501	0,61162
]5 ; 10]	3 135	0,10927	0,72089
]10 ; 15]	1 510	0,05263	0,77352
]15 ; 20]	968	0,03374	0,80726
]20 ; 25]	621	0,02164	0,82890
]25 ; 50]	1 691	0,05894	0,88784
]50 ; 100]	1 217	0,04242	0,93026
]100 ; 200]	855	0,02980	0,96006
]200 ; 500]	719	0,02506	0,98512
]500 ; 1 000]	245	0,00854	0,99366
]1 000 ; 5 000]	173	0,00603	0,99969
]5 000 ; 10 000]	8	0,00028	0,99997
> 10 000	1	0,00003	1,00000
TOTAL	**28 691**		

Annexe 10 : Nombre de tweets émis en fonction des jours

Octobre :

Date	Cnt_Date
08/10/2014	6 372
09/10/2014	3 617
10/10/2014	8 652
11/10/2014	7 423
13/10/2014	8 008
14/10/2014	10 627
15/10/2014	11 463
16/10/2014	9 974
17/10/2014	9 813
18/10/2014	9 435
19/10/2014	12 986
20/10/2014	9 894
21/10/2014	10 765
22/10/2014	12 587
23/10/2014	12 684
24/10/2014	13 260
25/10/2014	12 268
26/10/2014	15 062
27/10/2014	12 257
28/10/2014	12 289
29/10/2014	11 889
30/10/2014	11 246
31/10/2014	10 547
TOTAL	**243 118**
MOYENNE	**10 570**

Novembre :

Date	Cnt_Date
01/11/2014	11 356
02/11/2014	12 892
03/11/2014	10 652
04/11/2014	9 011
05/11/2014	9 409
06/11/2014	9 459
07/11/2014	8 489
08/11/2014	9 856
09/11/2014	14 732
10/11/2014	9 080
11/11/2014	14 789
12/11/2014	10 414
13/11/2014	9 230
14/11/2014	9 363
15/11/2014	10 089
16/11/2014	11 863
17/11/2014	9 248
18/11/2014	8 450
19/11/2014	9 408
20/11/2014	7 711
21/11/2014	9 336
22/11/2014	10 476
23/11/2014	8 526
24/11/2014	7 519
25/11/2014	10 064
26/11/2014	10 012
27/11/2014	9 608
28/11/2014	9 694
29/11/2014	9 503
30/11/2014	12 733
TOTAL	**302 972**
MOYENNE	**10 099**

Décembre :

Date	Cnt_Date
01/12/2014	10 183
02/12/2014	9 549
03/12/2014	9 274
04/12/2014	8 641
05/12/2014	7 902
06/12/2014	11 495
07/12/2014	12 462
08/12/2014	9 433
09/12/2014	7 746
10/12/2014	7 552
11/12/2014	7 986
12/12/2014	457
TOTAL	**102 680**
MOYENNE	**8 557**

Janvier :

Date	Cnt_Date
05/01/2015	7 834
06/01/2015	8 388
07/01/2015	12 039
08/01/2015	9 580
09/01/2015	10 637
10/01/2015	9 326
11/01/2015	13 414
12/01/2015	9 610
13/01/2015	8 149
14/01/2015	7 926
15/01/2015	7 526
16/01/2015	7 590
17/01/2015	8 957
18/01/2015	10 396
19/01/2015	8 975
20/01/2015	7 629
21/01/2015	8 407
22/01/2015	7 335
23/01/2015	7 417
24/01/2015	8 126
25/01/2015	9 485
26/01/2015	7 499
27/01/2015	6 885
28/01/2015	4 735
29/01/2015	4 950
30/01/2015	7 128
31/01/2015	8 000
TOTAL	**170 139**
MOYENNE	**8 102**

Février :

Date	Cnt_Date
01/02/2015	10 601
02/02/2015	7 950
03/02/2015	9 050
04/02/2015	8 443
05/02/2015	7 774
06/02/2015	3 727
07/02/2015	3 929
08/02/2015	4 166
09/02/2015	4 115
10/02/2015	3 671
11/02/2015	4 044
12/02/2015	3 564
13/02/2015	4 086
14/02/2015	4 787
15/02/2015	5 164
16/02/2015	4 418
17/02/2015	4 434
18/02/2015	4 428
19/02/2015	4 179
20/02/2015	4 337
21/02/2015	4 260
22/02/2015	5 194
23/02/2015	4 699
24/02/2015	4 872
25/02/2015	4 731
26/02/2015	4 419
27/02/2015	4 866
28/02/2015	4 107
TOTAL	**144 015**
MOYENNE	**5 143**

Mars :

Date	Cnt_Date
01/03/2015	4 714
02/03/2015	4 361
03/03/2015	4 469
04/03/2015	4 729
05/03/2015	4 355
06/03/2015	4 509
07/03/2015	4 546
08/03/2015	5 688
09/03/2015	4 707
10/03/2015	4 667
11/03/2015	4 610
12/03/2015	3 903
13/03/2015	3 752
14/03/2015	3 968
15/03/2015	5 543
16/03/2015	4 150
17/03/2015	4 045
18/03/2015	4 121
19/03/2015	3 552
20/03/2015	3 841
21/03/2015	4 361
22/03/2015	4 830
23/03/2015	4 112
24/03/2015	4 010
25/03/2015	4 148
26/03/2015	3 992
27/03/2015	3 296
28/03/2015	3 757
29/03/2015	4 492
30/03/2015	3 818
31/03/2015	3 708
TOTAL	**132 754**
MOYENNE	**4 282**

Nombre total de tweets postés sur 151 jours enregistrés : **1 153 664**
Nombre moyen de tweets émis par jour : **7 639**

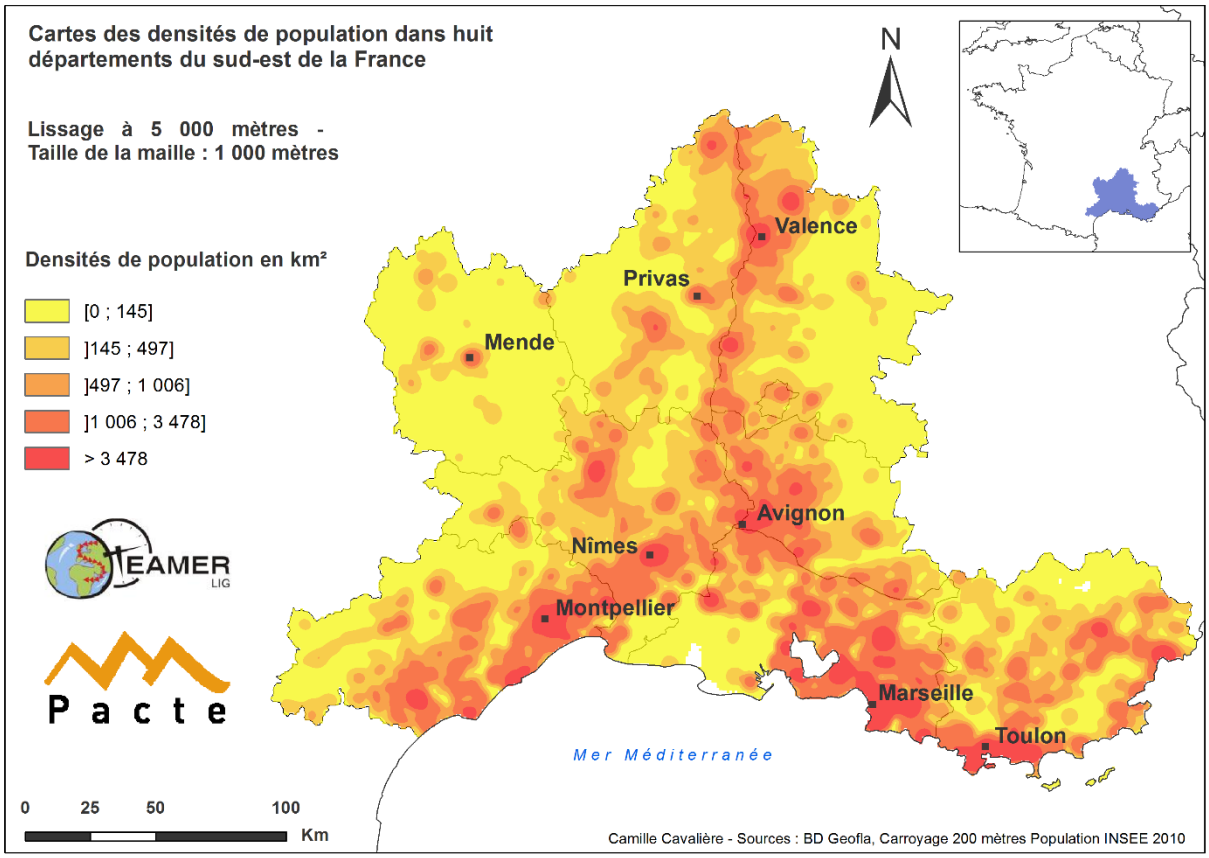
Annexe 11 : Nombre de tweets émis en fonction de l'heure

Heure	Nb Tweets	Nb Tweets Moyenne
0h - 1h	50 796	334
1h - 2h	28 882	190
2h - 3h	15 988	105
3h - 4h	8 229	54
4h - 5h	4 490	30
5h - 6h	3 941	26
6h - 7h	10 415	69
7h - 8h	24 239	159
8h - 9h	26 140	172
9h - 10h	34 183	225
10h - 11h	43 438	286
11h - 12h	49 924	328
12h - 13h	55 662	366
13h - 14h	54 224	357
14h - 15h	45 712	301
15h - 16h	41 106	270
16h - 17h	45 993	303
17h - 18h	61 944	408
18h - 19h	81 078	533
19h - 20h	86 917	572
20h - 21h	89 753	590
21h - 22h	107 266	706
22h - 23h	103 730	682
23h - 24h	79 617	524

Les résultats de la colonne **Nb Tweets Moyenne** sont obtenus en divisant le nombre total de tweets enregistrés par heure (**Nb Tweets)** par le nombre total de jours (soit 151).

Flux horaire moyen pour la période : 316 tweets émis par heure.

Annexe 12 : Carte des densités de population



Annexe 13 : Fréquences cumulées des utilisateurs en période de crise

Nombre de tweets	Effectif	Fréquence	Fréquences cumulées
]0 ; 1]	857	0,5748	0,57478
]1 ; 5]	525	0,3521	0,92689
]5 ; 10]	73	0,0490	0,97586
]10 ; 15]	17	0,0114	0,98726
]15 ; 20]	9	0,0060	0,99329
]20 ; 30]	6	0,0040	0,99732
]30 ; 43]	4	0,0027	1,00000
TOTAL	**1 491**		

Annexe 14 : Nombre de tweets extraits et filtrés par mots-clés recherchés

Cours d'eau	Agout	0
	Argens	2
	Cèze	4
	Dourbie	0
	Drôme	2
	Gard Gardon	104
	Lez	12
	Mosson	1
	Nartuby	3
	Tarn	1
	Rhône	20
	Vidourle	4

Météo	Cévenol	10
	Déluge	129
	Eclair	72
	Flotte	25
	Foudre	8
	Intempérie	136
	Orage	357
	Pleut	950
	Pluie	1 073
	Précipitations	6
	Storm	14
	Tempête	113
	Tonnerre	56

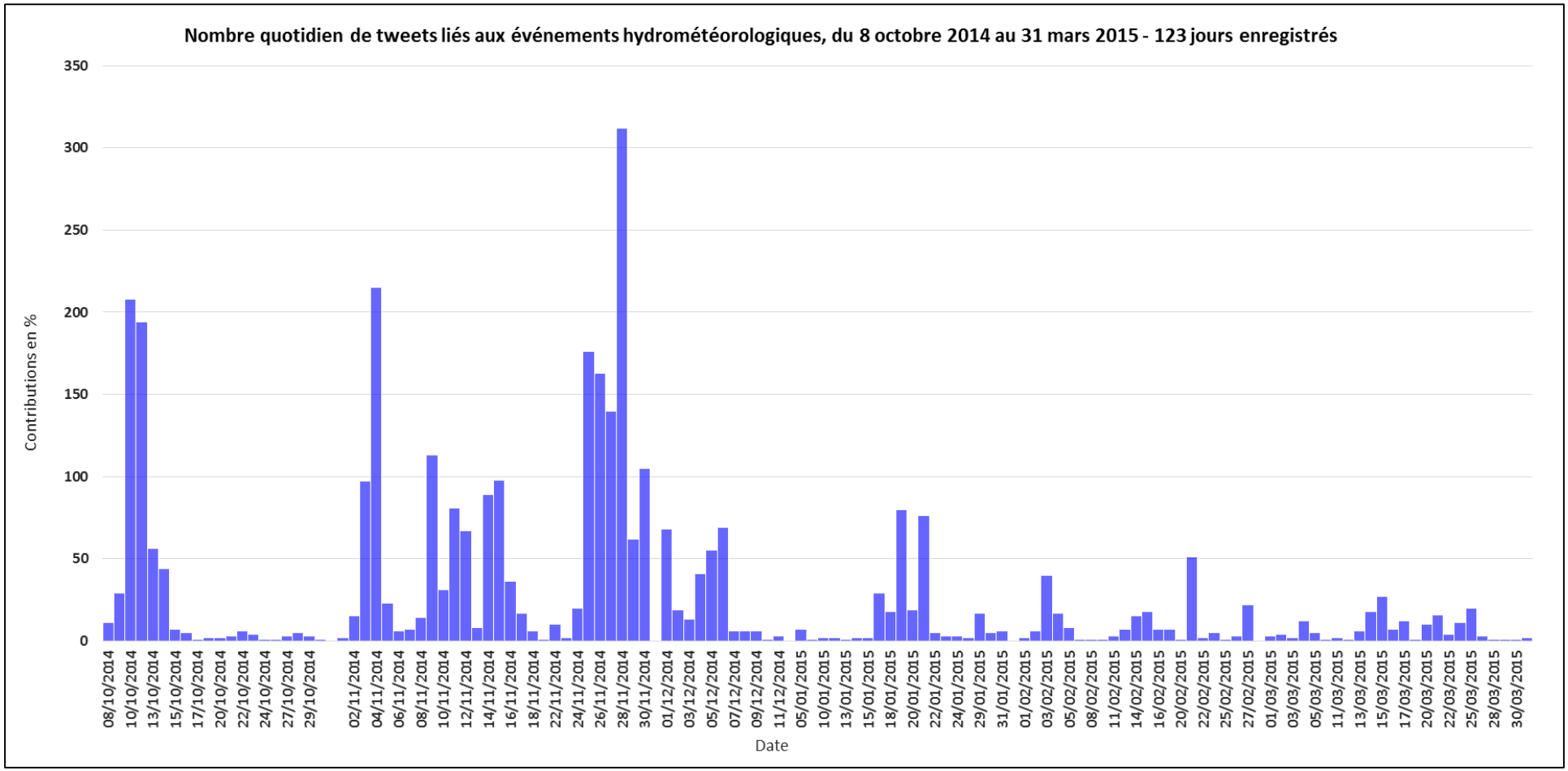
Villes	Alès	26
	Anduze	1
	Baume	0
	Calmette	2
	Grabels	2
	Nîmes	60
	Pérols	0
	Sète	4
	Uzès	2
	Vigan	2

Période Pré-Crise	Alerte	245
	Evacuation	24
	Prudence	15
	Vigilance	61

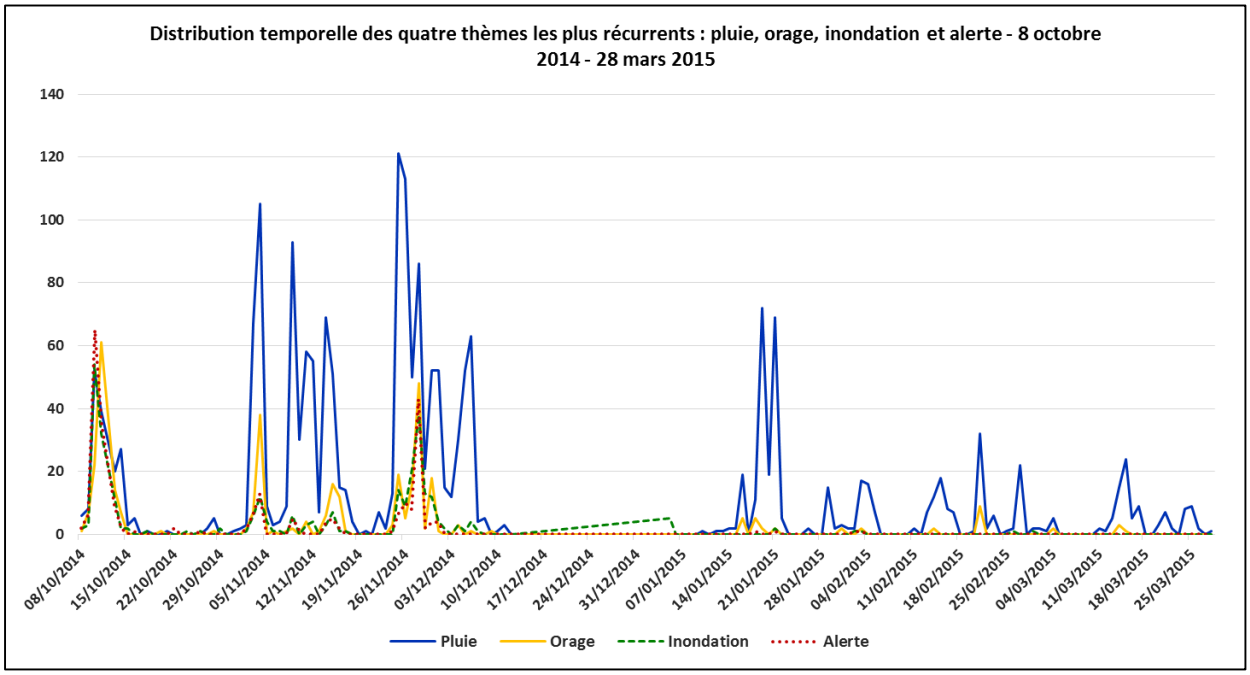
Période Syn-Crise	Boue	6
	Circulation	5
	Crise	6
	Critique	5
	Crue	20
	Débordement	37
	Digue Canal Chenal Bassin Egout	5
	Eau + Monte	5
	Flood	20
	Inondation	295
	Maison Foyer Logement Habitation	58
	Niveau + Eau	5
	Panne Electricité Coupure	48
	Potable	1
	Rivière	6
	Route Pont	94
	Ruissellement	3
	Secours Pompiers	11
	Torrent	16
	Urgence	5

Période Post-Crise	Arbre	9
	Assurance Assureur	1
	Catastrophe	9
	Dégâts Dommages	11
	Déblayer	0
	Nettoyer	4
	Sinistre	8

Annexe 15 : Distribution quantitative et temporelle des tweets filtrés par la méthode expertisée



Annexe 16 : Distribution des effectifs de tweets contenant les mots-clés principaux



Date	Pluie	Orage	Inondation	Alerte
08/10/2014	6	1	2	2
09/10/2014	8	7	3	6
10/10/2014	52	22	54	65
11/10/2014	39	61	32	35
13/10/2014	20	14	11	9
14/10/2014	27	7	2	2
15/10/2014	3	0	2	0
16/10/2014	5	0	0	1
17/10/2014	0	0	0	0
18/10/2014	1	0	1	0
19/10/2014	0	0	0	0
20/10/2014	0	1	0	0
21/10/2014	1	0	0	0
22/10/2014	0	0	0	2
23/10/2014	0	0	0	0
24/10/2014	0	0	1	0
25/10/2014	0	0	0	0
26/10/2014	0	0	1	1
27/10/2014	2	0	0	0
28/10/2014	5	1	0	0
29/10/2014	0	0	2	0
30/10/2014	0	0	0	0
31/10/2014	1	0	0	0

Date	Pluie	Orage	Inondation	Alerte
01/12/2014	52	1	4	4
02/12/2014	15	0	2	0
03/12/2014	12	0	0	0
04/12/2014	30	3	3	0
05/12/2014	52	0	1	0
06/12/2014	63	1	4	0
07/12/2014	4	0	1	0
08/12/2014	5	0	0	0
09/12/2014	0	1	0	0
10/12/2014	1	0	0	0
11/12/2014	3	0	0	0
12/12/2014	0	0	0	0

Date	Pluie	Orage	Inondation	Alerte
01/11/2014	2	0	0	0
02/11/2014	3	2	1	2
03/11/2014	67	7	6	7
04/11/2014	105	38	12	13
05/11/2014	9	0	4	0
06/11/2014	3	1	1	0
07/11/2014	4	0	1	0
08/11/2014	9	1	0	0
09/11/2014	93	2	6	5
10/11/2014	30	0	0	1
11/11/2014	58	4	3	0
12/11/2014	55	0	4	0
13/11/2014	7	1	0	0
14/11/2014	69	6	3	4
15/11/2014	51	16	7	5
16/11/2014	15	12	1	2
17/11/2014	14	1	1	0
18/11/2014	4	0	0	0
19/11/2014	0	0	0	0
20/11/2014	1	0	0	0
21/11/2014	0	0	0	0
22/11/2014	7	0	0	0
23/11/2014	2	0	0	0
24/11/2014	13	3	0	1
25/11/2014	121	19	14	7
26/11/2014	113	5	9	10
27/11/2014	50	16	21	8
28/11/2014	86	48	38	43
29/11/2014	21	2	13	2
30/11/2014	52	18	12	4

Date	Pluie	Orage	Inondation	Alerte
05/01/2015	0	0	5	0
06/01/2015	0	0	0	0
07/01/2015	0	0	0	0
08/01/2015	0	0	0	0
09/01/2015	0	0	0	0
10/01/2015	1	0	0	0
11/01/2015	0	0	0	0
12/01/2015	1	0	0	0
13/01/2015	1	0	0	0
14/01/2015	2	0	0	0
15/01/2015	2	0	0	0
16/01/2015	19	5	0	0
17/01/2015	0	0	0	0
18/01/2015	11	5	0	1
19/01/2015	72	2	0	0
20/01/2015	19	0	0	0
21/01/2015	69	2	2	1
22/01/2015	5	0	0	0
23/01/2015	0	0	0	0
24/01/2015	0	0	0	0
25/01/2015	0	0	0	0
26/01/2015	2	0	0	0
27/01/2015	0	0	0	0
28/01/2015	0	0	0	0
29/01/2015	15	0	0	0
30/01/2015	2	0	0	0
31/01/2015	3	2	0	0

Date	Pluie	Orage	Inondation	Alerte
01/02/2015	2	0	0	0
02/02/2015	2	0	1	1
03/02/2015	17	2	1	1
04/02/2015	16	0	0	0
05/02/2015	7	0	0	0
06/02/2015	0	0	0	0
07/02/2015	0	0	0	0
08/02/2015	0	0	0	0
09/02/2015	0	0	0	0
10/02/2015	0	0	0	0
11/02/2015	2	0	0	0
12/02/2015	0	0	0	0
13/02/2015	7	0	0	0
14/02/2015	12	2	0	0
15/02/2015	18	0	0	0
16/02/2015	8	0	0	0
17/02/2015	7	0	0	0
18/02/2015	0	0	0	0
19/02/2015	0	0	0	0
20/02/2015	1	0	0	0
21/02/2015	32	9	0	0
22/02/2015	2	0	0	0
23/02/2015	6	0	0	0
24/02/2015	0	0	0	0
25/02/2015	1	0	0	0
26/02/2015	2	0	1	0
27/02/2015	22	0	0	0
28/02/2015	0	0	0	0

Date	Pluie	Orage	Inondation	Alerte
01/03/2015	2	0	1	0
02/03/2015	2	0	0	0
03/03/2015	1	0	0	0
04/03/2015	5	2	0	0
05/03/2015	0	0	0	0
06/03/2015	0	0	0	0
07/03/2015	0	0	0	0
08/03/2015	0	0	0	0
09/03/2015	0	0	0	0
10/03/2015	0	0	0	0
11/03/2015	2	0	0	0
12/03/2015	1	0	0	0
13/03/2015	5	0	0	0
14/03/2015	15	3	0	0
15/03/2015	24	1	0	0
16/03/2015	5	0	0	0
17/03/2015	9	0	0	0
18/03/2015	0	0	0	0
19/03/2015	0	0	0	0
20/03/2015	3	0	0	0
21/03/2015	7	0	0	0
22/03/2015	2	0	0	0
23/03/2015	0	0	0	0
24/03/2015	8	0	0	0
25/03/2015	9	0	0	0
26/03/2015	2	0	0	0
27/03/2015	0	0	0	0
28/03/2015	1	0	0	0
29/03/2015	0	0	0	0
30/03/2015	0	0	0	0
31/03/2015	0	0	0	0

Annexe 17 : Distribution temporelle des tweets du corpus liés aux événements

Date	Tweets_évé	Nb_Total_Tweets	%_Tweets_évé
08/10/2014	11	6 372	0,17
09/10/2014	29	3 617	0,80
10/10/2014	208	8 652	2,40
11/10/2014	194	7 423	2,61
13/10/2014	56	8 008	0,70
14/10/2014	44	10 627	0,41
15/10/2014	7	11 463	0,06
16/10/2014	5	9 974	0,05
17/10/2014	1	9 813	0,01
18/10/2014	2	9 435	0,02
20/10/2014	2	9 894	0,02
21/10/2014	3	10 765	0,03
22/10/2014	6	12 587	0,05
23/10/2014	4	12 684	0,03
24/10/2014	1	13 260	0,01
26/10/2014	1	15 062	0,01
27/10/2014	3	12 257	0,02
28/10/2014	5	12 289	0,04
29/10/2014	3	11 889	0,03
31/10/2014	1	10 547	0,01

Date	Tweets_évé	Nb_Total_Tweets	%_Tweets_évé
01/11/2014	2	11 356	0,02
02/11/2014	15	12 892	0,12
03/11/2014	97	10 652	0,91
04/11/2014	215	9 011	2,39
05/11/2014	23	9 409	0,24
06/11/2014	6	9 459	0,06
07/11/2014	7	8 489	0,08
08/11/2014	14	9 856	0,14
09/11/2014	113	14 732	0,77
10/11/2014	31	9 080	0,34
11/11/2014	81	14 789	0,55
12/11/2014	67	10 414	0,64
13/11/2014	8	9 230	0,09
14/11/2014	89	9 363	0,95
15/11/2014	98	10 089	0,97
16/11/2014	36	11 863	0,30
17/11/2014	17	9 248	0,18
18/11/2014	6	8 450	0,07
20/11/2014	1	7 711	0,01
22/11/2014	10	10 476	0,10
23/11/2014	2	8 526	0,02
24/11/2014	20	7 519	0,27
25/11/2014	176	10 064	1,75
26/11/2014	163	10 012	1,63
27/11/2014	140	9 608	1,46
28/11/2014	312	9 694	3,22
29/11/2014	62	9 503	0,65
30/11/2014	105	12 733	0,82

Date	Tweets_évé	Nb_Total_Tweets	%_Tweets_évé
01/12/2014	68	10 183	0,67
02/12/2014	19	9 549	0,20
03/12/2014	13	9 274	0,14
04/12/2014	41	8 641	0,47
05/12/2014	55	7 902	0,70
06/12/2014	69	11 495	0,60
07/12/2014	6	12 462	0,05
08/12/2014	6	9 433	0,06
09/12/2014	6	7746	0,08
10/12/2014	1	7552	0,01
11/12/2014	3	7986	0,04

Date	Tweets_évé	Nb_Total_Tweets	%_Tweets_évé
05/01/2015	7	7 834	0,09
07/01/2015	1	12 039	0,01
10/01/2015	2	9 326	0,02
12/01/2015	2	9 610	0,02
13/01/2015	1	8 149	0,01
14/01/2015	2	7 926	0,03
15/01/2015	2	7 526	0,03
16/01/2015	29	7 590	0,38
18/01/2015	18	10 396	0,17
19/01/2015	80	8 975	0,89
20/01/2015	19	7 629	0,25
21/01/2015	76	8 407	0,90
22/01/2015	5	7 335	0,07
23/01/2015	3	7 417	0,04
24/01/2015	3	8 126	0,04
26/01/2015	2	7 499	0,03
29/01/2015	17	4 950	0,34
30/01/2015	5	7 128	0,07
31/01/2015	6	8 000	0,08

Date	Tweets_évé	Nb_Total_Tweets	%_Tweets_évé
01/02/2015	2	10 601	0,02
02/02/2015	6	7 950	0,08
03/02/2015	40	9 050	0,44
04/02/2015	17	8 443	0,20
05/02/2015	8	7 774	0,10
07/02/2015	1	3 929	0,03
08/02/2015	1	4 166	0,02
10/02/2015	1	3 671	0,03
11/02/2015	3	4 044	0,07
13/02/2015	7	4 086	0,17
14/02/2015	15	4 787	0,31
15/02/2015	18	5 164	0,35
16/02/2015	7	4 418	0,16
17/02/2015	7	4 434	0,16
20/02/2015	1	4 337	0,02
21/02/2015	51	4 260	1,20
22/02/2015	2	5 194	0,04
23/02/2015	5	4 699	0,11
25/02/2015	1	4 731	0,02
26/02/2015	3	4 419	0,07
27/02/2015	22	4 866	0,45

Date	Tweets_évé	Nb_Total_Tweets	%_Tweets_évé
01/03/2015	3	4 714	0,06
02/03/2015	4	4 361	0,09
03/03/2015	2	4 469	0,04
04/03/2015	12	4 729	0,25
05/03/2015	5	4 355	0,11
09/03/2015	1	4 707	0,02
11/03/2015	2	4 610	0,04
12/03/2015	1	3 903	0,03
13/03/2015	6	3 752	0,16
14/03/2015	18	3 968	0,45
15/03/2015	27	5 543	0,49
16/03/2015	7	4 150	0,17
17/03/2015	12	4 045	0,30
18/03/2015	1	4 121	0,02
20/03/2015	10	3 841	0,26
21/03/2015	16	4 361	0,37
22/03/2015	4	4 830	0,08
24/03/2015	11	4 010	0,27
25/03/2015	20	4 148	0,48
26/03/2015	3	3 992	0,08
28/03/2015	1	3757	0,03
29/03/2015	1	4492	0,02
30/03/2015	1	3818	0,03
31/03/2015	2	3708	0,05

3 457 tweets émis en rapport avec les perturbations

Total de 966 338 tweets émis les jours où des tweets en rapport avec des phénomènes ont été enregistrés

Nombre moyen de tweets émis relatifs à un événement sur les jours enregistrés

28

Nombre moyen quotidien de tweets émis (sans distinction de contenu) pour les mêmes jours enregistrés

7 856

% de tweets en rapport avec les événements (calculés à partir des deux valeurs ci-dessus)

0,36

Annexe 18 : Comparaison des flux horaires de la période globale et des journées de crise

Heure	Nb Tweets Normal	Nb Moyen Tweets Normal	Nb Tweets Crise	Nb Moyen Tweets Crise (dates CatNat)	Moyenne Crise - Moyenne Normale
0h - 1h	50 796	334	7 010	389	55
1h - 2h	28 882	190	3 739	208	18
2h - 3h	15 988	105	1 908	106	1
3h - 4h	8 229	54	1 124	62	8
4h - 5h	4 490	30	606	34	4
5h - 6h	3 941	26	741	41	15
6h - 7h	10 415	69	1 522	85	16
7h - 8h	24 239	159	3 193	177	18
8h - 9h	26 140	172	3 246	180	8
9h - 10h	34 183	225	4 118	229	4
10h - 11h	43 438	286	5 050	281	-5
11h - 12h	49 924	328	6 163	342	14
12h - 13h	55 662	366	7 739	430	64
13h - 14h	54 224	357	7 770	432	75
14h - 15h	45 712	301	6 721	373	73
15h - 16h	41 106	270	5 705	317	47
16h - 17h	45 993	303	7 171	398	96
17h - 18h	61 944	408	8 915	495	88
18h - 19h	81 078	533	12 321	685	151
19h - 20h	86 917	572	13 255	736	165
20h - 21h	89 753	590	13 122	729	139
21h - 22h	107 266	706	14 825	824	118
22h - 23h	103 730	682	14 892	827	145
23h - 24h	79 617	524	11 355	631	107

Annexe 19 : Comparaison des pourcentages d'émission horaires des tweets liés à la crise du 28 novembre 2014 et du jeu de tweets global

Heure	Nb Tweets Crise	% Tweets Crise	Nb Tweets	% Tweets
0h - 1h	13	4,17	334	4,40
1h - 2h	39	12,50	190	2,50
2h - 3h	14	4,49	105	1,39
3h - 4h	4	1,28	54	0,71
4h - 5h	3	0,96	30	0,39
5h - 6h	4	1,28	26	0,34
6h - 7h	9	2,88	69	0,90
7h - 8h	27	8,65	159	2,10
8h - 9h	12	3,85	172	2,27
9h - 10h	15	4,81	225	2,96
10h - 11h	21	6,73	286	3,77
11h - 12h	12	3,85	328	4,33
12h - 13h	12	3,85	366	4,82
13h - 14h	14	4,49	357	4,70
14h - 15h	17	5,45	301	3,96
15h - 16h	10	3,21	270	3,56
16h - 17h	8	2,56	303	3,99
17h - 18h	15	4,81	408	5,37
18h - 19h	7	2,24	533	7,03
19h - 20h	10	3,21	572	7,53
20h - 21h	24	7,69	590	7,78
21h - 22h	11	3,53	706	9,30
22h - 23h	6	1,92	682	8,99
23h - 24h	5	1,60	524	6,90
TOTAL	**312**	**100,00**	**7 590**	**100,00**

Annexe 20 : Distribution horaire des tweets contenant les principaux mots-clés employés par les utilisateurs, le 28 novembre 2014

Heure	Pluie/Pleut	Orage	Inondation	Alerte
0h - 1h	4	6	0	2
1h - 2h	7	23	5	4
2h - 3h	4	5	1	2
3h - 4h	1	1	0	3
4h - 5h	1	1	0	0
5h - 6h	1	1	0	1
6h - 7h	2	1	1	0
7h - 8h	5	3	1	4
8h - 9h	4	0	2	2
9h - 10h	6	1	1	2
10h - 11h	6	3	3	3
11h - 12h	4	0	0	4
12h - 13h	1	0	1	0
13h - 14h	5	1	2	2
14h - 15h	6	0	3	0
15h - 16h	5	0	2	1
16h - 17h	3	0	1	0
17h - 18h	7	1	1	1
18h - 19h	2	0	0	1
19h - 20h	3	0	3	4
20h - 21h	4	0	7	5
21h - 22h	3	1	3	4
22h - 23h	1	0	1	0
23h - 24h	1	0	0	1

Annexe 21 : Comparaison de la contribution quotidienne des tweets liés aux perturbations pour le Cheylard et le corpus filtré global

date	Cnt_date	%_Cheylard	Nb_Tweets	%_total_tweets
08/10/2014	1	2,50	11	0,54
10/10/2014	2	5,00	208	10,26
11/10/2014	1	2,50	194	9,57
13/10/2014	3	7,50	56	2,76
14/10/2014	2	5,00	44	2,17
03/11/2014	5	12,50	97	4,79
04/11/2014	1	2,50	215	10,61
05/11/2014	1	2,50	23	1,13
07/11/2014	2	5,00	7	0,35
11/11/2014	1	2,50	81	4,00
13/11/2014	1	2,50	8	0,39
14/11/2014	7	17,50	89	4,39
15/11/2014	3	7,50	98	4,83
25/11/2014	1	2,50	176	8,68
26/11/2014	1	2,50	163	8,04
27/11/2014	1	2,50	140	6,91
28/11/2014	5	12,50	312	15,39
30/11/2014	2	5,00	105	5,18
TOTAL	**40**	**100,00**	**2027**	**100,00**

Annexe 22 : Distribution quotidienne des tweets du Cheylard en fonction des thèmes

Date	Information	Météo	Inondation	Vigilance
08/10/2014	1	0	0	0
10/10/2014	2	0	0	0
11/10/2014	1	0	0	0
13/10/2014	1	2	0	0
14/10/2014	0	2	0	0
03/11/2014	0	4	1	0
04/11/2014	0	0	1	0
05/11/2014	0	1	0	0
07/11/2014	1	1	0	0
11/11/2014	0	1	0	0
13/11/2014	0	1	0	0
14/11/2014	0	6	0	1
15/11/2014	2	1	0	0
25/11/2014	0	1	0	0
26/11/2014	0	1	0	0
27/11/2014	1	0	0	0
28/11/2014	2	3	0	0
30/11/2014	0	2	0	0
TOTAL	**11**	**26**	**2**	**1**

Résumé du mémoire

Ce mémoire s'inscrit dans le cadre d'un stage de Master 2 effectué en partenariat avec le laboratoire PACTE et le LIG, sur le projet ANR MobiClimEx - Dynamique des mobilités quotidiennes et résidentielles face aux extrêmes météorologiques en contexte de changement climatique - qui s'appuie sur l'étude de l'évolution des vulnérabilités et des réponses sociales des populations face à des phénomènes hydrométéorologiques extrêmes. Le mémoire propose d'explorer une nouvelle source d'information, les médias sociaux et en particulier la plateforme de microblogage Twitter, comme outil d'acquisition de données susceptibles de fournir aux chercheurs des informations sur la place d'un phénomène dans la société et sur les comportements des individus face à ce phénomène.

L'information collectée auprès des médias sociaux s'articule autour de deux concepts : la *volunteered geographic information* et le *crowdsourcing*, qui désignent d'une part, l'information géolocalisée, créée et partagée par un utilisateur sur Internet via des plateformes de médias sociaux, et, d'autre part, la participation d'un groupe ou d'une personne à la création d'un contenu spécifique pouvant se révéler riche en connaissances. Bien que cette source d'acquisition de l'information géographique ait pu faire ses preuves pendant les catastrophes récentes, sa contrainte principale reste qu'elle est massive et hétérogène, en forme et en contenu, en raison de l'absence de règle de formalisation qui conditionnent sa production. Lorsqu'on travaille sur un thème précis, l'enjeu primordial de l'exploitation de cette source de données consiste donc à déterminer une méthodologie permettant d'extraire l'information correspondant au thème recherché et de supprimer le bruit, c'est-à-dire les tweets n'ayant aucun rapport avec ce thème.

Le mémoire se positionne dans une phase d'analyse post-crise de l'information envoyée par Smartphone sur le réseau Twitter, lors des épisodes cévenols et crues rapides ayant affecté le sud-est de la France à l'automne 2014. Il explore une méthodologie d'extraction fondée sur la recherche d'un vocabulaire précis employé par les utilisateurs pour évoquer ce type d'événement. Ce vocabulaire est déterminé à partir de deux approches complémentaires : l'expertise et la fouille de texte automatisée. Le questionnement s'articule autour des axes suivants : il s'agit dans un premier temps de déterminer si ce type d'événement est enregistré dans le contenu des tweets, puis d'évaluer l'éventuelle correspondance entre la dynamique des flux de tweets et la dynamique de l'événement. Enfin, nous cherchons à savoir si le tweet peut servir d'indicateur pour révéler la sévérité d'une crise.

Les analyses mises en œuvre visent donc à étudier et à comparer la distribution des tweets sur plusieurs échelles temporelles, à analyser la distribution spatio-temporelle des tweets par l'intermédiaire de cartographies des jeux de tweets, ainsi qu'à étudier les variations temporelles du contenu des tweets pendant une crise.