Acronyme du projet	**DISCRET**
Titre complet	**D**émonstrateur d'**I**dentification de **S**ituations **C**ritiques via la **R**emontée de données multisources pour l'al**E**rte en **T**emps-réel
Axe thématique principal	AXE 2 : REMONTÉE D'ALERTE PAR LA POPULATION
Type de recherche	X Recherche industrielle • Développement expérimental

Aide totale demandée	487 387 €	Durée du projet	18 mois

Table des matières

Résumé du projet

Le projet DISCRET vise à démontrer qu'il est possible de détecter, en temps réel, des situations atypiques ou critiques, au travers de l'analyse des données d'un opérateur de téléphonie mobile (Orange), enrichies par des informations extraites du réseau social Twitter et à proposer un premier prototype de système d'avertissement destiné aux services de gestion de crises, de sécurité et de secours. L'hypothèse du projet, confirmée par plusieurs travaux de recherche récents, est que des événements significatifs engendrent localement une modification substantielle des flux et de la nature des communications. Ces anomalies, quasi concomitantes de l'événement, peuvent être détectées et localisées sur la base du réseau d'antennes relais. D'autre part, la connaissance anticipée de la localisation de l'événement et de la nature des communications (changement de nombre ou de fréquence de messages envoyés à travers différents canaux), permet d'envisager de collecter de manière plus efficace les données des médias sociaux, afin de caractériser et de contextualiser l'événement et donc de mieux valoriser les remontées d'informations par la population par des voies non spécifiquement dédiées à l'alerte.

Le projet constitue une contribution à l'axe 2 de l'appel d'offres : « *Remontée d'alerte par la population* ». Le projet présente l'originalité d'exploiter des informations transmises de manière passive par la population au travers de l'activité mesurée sur les réseaux de téléphonie mobile et les remontées d'informations actives par des canaux «généralistes» : les media sociaux (Twitter dans le cas du présent projet). Même si les médias sociaux ne sont pas spécifiquement dédiés aux remontées d'alertes, ils constituent un canal populaire et privilégié de diffusion d'informations événementielles.

Abstract

DISCRET aims at demonstrating the possibility to detect and locate, in real-time, unusual or critical situations in urban areas, based on the analysis of cell phone network data. This detection will be complemented with information extracted from social networks (i.e., Twitter in the context of the project). A prototype of a warning platform for security and emergency operators will be implemented. Several recent research works have shown that major events induce locally significant modifications of the amount and nature of cellular network communications. These anomalies, typically concomitant with the unusual event, may be detected and located based on the network of cell phone antennas. Moreover, the early detection and localization of the events, together with the knowledge of the associated communication activity, allow for a more effective retrieval of information from the social networks. That permits to provide elements of description and context for the detected event and, therefore, to increase the value of the information conveyed by the population via channels that are not explicitly conceived for alerting purposes.

DISCRET is a contribution to the second research axis listed in the call for proposals: "broadcasting private warnings". The originality of the project lies in the joint usage of information generated by the population in a passive way (i.e., through the cell-phone communication activity) and the one produced in an active way through non-specific channels (i.e., online social networks). Social networks are not specifically dedicated to the broadcast of warnings, but they represent popular and major event information and broadcasting media.

1 Pertinence de la proposition au regard des orientations de l'appel à projet

1. 1. Contexte et enjeux économiques et sociaux

L'intelligence Artificielle au coeur des réseaux mobiles : de nouvelles opportunités

Les réseaux de téléphonie mobile sont aujourd'hui l'un des vecteurs majeurs de circulation des informations, en période normale et en période de crise. Ils constituent la voie privilégiée de diffusion des informations par les particuliers via les réseaux sociaux notamment. De très nombreux services et applications, ainsi que plusieurs solutions proposées dans les travaux de recherche se fondent sur l'utilisation des téléphones mobiles comme réseaux de capteurs *in situ*. Cette utilisation peut être le fruit d'une démarche active du détenteur du téléphone (installation d'applications et renseignement volontaire de données) ou passive (transmission de la localisation du téléphone via une application par exemple). Le cas le plus emblématique de ce type d'usage est l'application de navigation GPS participative Waze. Elle reste cependant une exception par le succès qu'elle a rencontré auprès du public. La plupart des applications dédiées ont cependant du mal à s'imposer sur un marché à l'offre foisonnante et donc à produire des informations fiables et représentatives.

Les opérateurs de téléphonie mobile collectent pour leur part de nombreuses données sur les communications transitant par leurs réseaux et antennes relais : dates, durées, localisation, nature des appels (appels vocaux, SMS, applications…). Ces données constituent aujourd'hui une nouvelle source d'information d'une grande richesse, qui sous certaines conditions (agrégation, anonymisation), peuvent être analysées et exploitées pour répondre à des questions très diverses. L'Intelligence Artificielle couplée au *Big Data*, appliqués aux données mobiles, *Mobile Crowd Sensing* (MCS) en anglais, constitue un champ de recherche et d'innovation en évolution rapide. Les conférences NetMob (*scientific analysis of mobile phone datasets* - netmob.org), organisées depuis 2010 tous les deux ans, permettent en particulier de faire le point régulièrement sur les avancées scientifiques dans ce domaine. Le traitement en temps différé des jeux de données issues des réseaux de téléphonie mobile permet d'analyser les distributions et les déplacements des populations et d'en déduire par exemple les besoins de développement des infrastructures de transport ou les principales voies de propagation des épidémies, de quantifier les phénomènes migratoires, voire la dynamique de certains marchés ou de détecter l'émergence de pénuries alimentaires. Plusieurs programmes soutiennent ces recherches, notamment dans les pays en voie de développement : initiative *Global Pulse* de l'ONU ou concours « *Data for Development* » (D4D) organisés par Orange, par exemple. En France, l'analyse de ces données constitue la base de plusieurs projets et a déjà débouché sur des applications opérationnelles. On peut citer l'utilisation de cette source de données pour le recueil de d'informations sur le trafic routier [1] et sa gestion [2,3][1]. Une autre application consiste à reconstituer des matrices de déplacement à partir de ces données [4]. Elle est à l'origine de services commerciaux comme *Flux Vision* d'Orange ou *Geostatistics* de SFR.

Les perspectives du diagnostic temps réel pour la détection de situations anormales

Des travaux de recherche pionniers ont exploré les effets d'événements exceptionnels sur l'activité mesurée sur les réseaux de téléphonie mobile : attentats, crashs aériens, blackouts ou inondations [5, 6]. Il en ressort que ces événements sont tous accompagnés d'une nette modification (ici,

[1] Le projet STRIP, coordonné par l'Ifsttar, a par exemple montré une forte corrélation entre les incidents routiers et le volume d'appels (principalement sortants).

augmentation) du nombre et de la nature des communications sur une courte période (*pulse*) dans les secteurs touchés. Ces caractéristiques laissent espérer que ces modifications remarquables[2] pourraient être identifiées en analysant en temps réel les flux de données des réseaux de téléphonie mobile. Ceci permettrait d'identifier et de localiser précocement des incidents ou situations de crises et/ou de conforter des informations provenant d'autres sources (appels de numéros d'urgence, constats des services de police et de secours, vidéo-surveillance). C'est la voie que se propose d'explorer le projet DISCRET.

Au-delà des verrous scientifiques et techniques de l'approche proposée qui seront détaillés en partie 2 de ce document, le projet DISCRET présente plusieurs originalités par rapport aux travaux antérieurs précurseurs sur le même sujet :

- Les travaux antérieurs étaient fondés sur l'analyse des relevés de consommation des utilisateurs (CRA). C'est une source d'information qui sera aussi exploitée dans le projet DISCRET pour tester les approches proposées sur des événements de l'année 2019 ou qui se produiront en dehors des zones géographiques et des plages de temps choisies pour les périodes de test du projet. Le projet comprendra aussi une analyse directe du flux d'informations, issu du réseau de téléphonie mobile : données de signalisation réseau. Le projet prendra appui sur la méthodologie développée et mise en œuvre dans le cadre d'un autre projet ANR en cours, le projet CANCAN (*Content ANd Context based Adaptation in Mobile Networks*), auquel contribuent certains participants du projet DISCRET. L'objectif est de démontrer les possibilités offertes par l'approche lorsqu'elle est appliquée sur les données disponibles en temps réel. De plus, les données de signalisation réseau sont beaucoup plus riches que les compte-rendus d'appels (déplacement des téléphones, nature des applications utilisées) et ouvrent la possibilité de traitements plus avancés. En particulier, des informations potentiellement utiles pour augmenter l'efficacité de la collecte dans les réseaux sociaux pourront être extraites (cf. structure du projet).
- D'autre part, les événements analysés sont souvent exceptionnels par la concentration de population, comme le seront les Jeux Olympiques 2024 à Paris. Un travail approfondi de collecte des données et définition des signatures d'activité des réseaux de téléphonie mobile, signatures variables dans le temps, devra être entrepris en amont des analyses, afin de pouvoir déterminer des anomalies.

Un contexte favorable pour le déploiement de solutions opérationnelles

Le déploiement de la téléphonie mobile de 5ème génération (5G) va engendrer une très nette augmentation des flux de données transitant par les réseaux et de leur variabilité. De nouvelles approches technologiques comme le *Beamforming*[3] vont permettre de limiter la consommation énergétique tout en focalisant la diffusion radio entre une antenne et un mobile, un mobile pouvant alors être en itinérance sur plusieurs antennes d'un même relais lors du déplacement de l'utilisateur. Cette augmentation des données et de la variabilité ne pourra par ailleurs pas être uniquement prise en charge par une augmentation de la capacité nominale des réseaux, mais nécessitera la mise en œuvre de dispositifs d'ajustements dynamiques et surtout automatiques des réseaux à la variabilité

[2] On cherchera dans le projet DISCRET à détecter des écarts (mesure d'excentricité) par rapport à des signatures de référence pour chaque antenne des réseaux de téléphonie mobile. Ces écarts peuvent porter sur le nombre de communications, le type de communications (appels vocaux, sms, utilisations d'applications…), la durée moyenne des appels, le rapport entre des appels entrants et sortants, les vitesses de déplacement des téléphones… Une large partie de la tâche 2 du projet sera consacrée à l'identification des critères pertinents pour la définition des signatures de référence et la mesure des excentricités.

[3] Le Beamforming (appelé aussi filtrage spatial, formation de faisceaux ou formation de voies) est une technique de traitement du signal utilisée dans les réseaux d'antennes et de capteurs pour l'émission ou la réception directionnelle de signaux.

spatio-temporelle du nombre de sollicitations et des flux transférés. A cette fin, les opérateurs de téléphonie mobile développent actuellement des algorithmes permettant de scruter en temps réel l'activité des réseaux et d'optimiser leur structure (i.e., les voies de transfert des données). Le développement de telles applications est aussi l'objet du projet CANCAN cité plus haut et de plusieurs projets européens importants qui l'ont précédé. Il devient dans ce contexte tout-à-fait réaliste d'envisager la mise en œuvre d'autres applications basées sur l'analyse en temps réel des données de signalisation réseau, comme celle qui est proposée dans le projet DISCRET à l'horizon 2024. Ces applications pourraient être intégrées dans un ensemble plus large d'algorithmes d'exploration en temps réel de l'activité des réseaux.

Une application utile bien au-delà des Jeux Olympiques

La détection précoce d'incidents potentiels, au travers de l'activité des réseaux de téléphonie mobile, pourrait trouver de nombreux champs d'application au-delà des Jeux Olympiques. Elle pourrait constituer à terme un service d'appui, proposé par les opérateurs de téléphonie mobile, pour aider au diagnostic et à la gestion d'événements de natures diverses. La possibilité de localiser les événements grâce à la densité et la topologie des réseaux d'antennes et de détecter rapidement des écarts, rend cette solution particulièrement intéressante pour des événements à évolutions rapides dans le temps et dans l'espace :

- incidents et désordres en marge de manifestations,
- attentats,
- catastrophes naturelles et en particulier crues soudaines,
- perturbations dans les transports.

1.2 Retombées pour les Jeux Olympique de Paris 2024

La plateforme que le projet DISCRET se propose de développer compléterait utilement d'autres dispositifs d'observation et de remontée d'informations : appels des particuliers aux services de secours, constats directs des services de sécurité ou de secours, réseaux de vidéo-surveillance, informations saisies via des applications smartphone dédiées[4]. Les avertissements basés sur l'observation de l'activité des réseaux de téléphonie mobile aideraient en particulier à :

1. Attirer l'attention des PC de sécurité sur des secteurs particuliers, renforcer la vigilance par rapport aux autres sources d'information et inciter à la recherche d'informations complémentaires,
2. Conforter rapidement la détection et la localisation d'incidents détectés via d'autres sources d'information,
3. Proposer rapidement des éléments de diagnostic de situation au travers du moissonnage guidé des informations postées sur les réseaux sociaux comme Twitter.

La plateforme DISCRET pourrait s'avérer particulièrement utile dans le cas d'incidents se produisant en dehors des secteurs qui feront l'objet d'une surveillance renforcée (enceintes sportives, fan zones). On pense ici typiquement aux abords des stades, des fan zones et aux quartiers fréquentés par les touristes en marge des épreuves sportives. La proposition du projet DISCRET est par ailleurs complémentaire de celle du projet "*C-Life Innovation*", déposé par Orange Business Services en réponse à l'appel à manifestations d'intérêt (AMI) "Sécurité JO2024". La proposition de plate-forme de gestion de crises et d'alerte "*C-Life Guard*" pourrait bénéficier directement des résultats des travaux engagés dans DISCRET.

[4] Des applications dédiées aux Jeux Olympiques seront certainement développées – typiquement applications dédiées à la billetterie et aux informations pratiques, qui permettront aussi des remontées de signalements.

Du point de vue des performances, le projet DISCRET vise des délais de détection de quelques minutes[5] et une précision géographique de quelques décamètres à une centaine de mètres en zone urbaine, compte tenu de la densité des réseaux d'antennes relais. Un **niveau de TRL 6** est visé, le prototypage du système étant l'objet du Lot 4. Il est prévu de calibrer et de tester l'approche proposée durant une période de neuf mois de traitement des flux des données issues des réseaux de téléphonie mobile dans deux zones géographiques distinctes : la Région Parisienne et la région de Nice. Par ailleurs, le traitement des données archivées[6] de facturation permettront de tester l'approche si des événements significatifs intéressants se produisent en dehors de ces zones géographiques et de la période d'observation choisie : attentat, accident grave, orage ou crue soudaine localisée, par exemple.

Plusieurs projets en cours déjà cités - développement de dispositifs de traitement des informations des réseaux de téléphonie mobile en temps réel pour la téléphonie 5G ou le projet *C-Life Guard* - permettent d'envisager raisonnablement le passage à un niveau de TRL9 à l'horizon 2024 en cas de résultats probants du projet DISCRET.

Enfin, le projet DISCRET est centré sur la détection précoce d'anomalies et la recherche de leur sémantique dans les médias sociaux numériques. L'analyse en temps réel des données opérationnelles des réseaux de téléphonie mobile pourrait déboucher sur d'autres services, pistes qui pourront être explorées en marge du projet DISCRET. On peut penser à l'estimation de densité instantanée de présence dans les secteurs touchés par des incidents afin de mieux calibrer les procédures d'intervention et d'évacuation. On peut aussi imaginer le suivi des déplacements et des mouvements des foules, information intéressante pour comprendre le comportement collectif lors des évacuations - le niveau de précision spatiale (décamétrique à hectométrique) ne permettra cependant pas d'étudier les déplacements à petit échelle des foules dans des espaces confinés.

2 Positionnement et objectifs de la proposition

Le projet DISCRET aborde la problématique de la détection, de la classification et de la notification rapides d'incidents, avec un haut niveau de précision spatiale et temporelle, à partir des données collectées par les réseaux de téléphonie mobile, complétées par des informations collectées via d'autres sources et notamment les médias sociaux[7]: Twitter dans le cas du projet.

L'utilisation des données de téléphonie mobile dans le cadre de la gestion de crise est une question de recherche récente et stimulante, encore peu abordée dans la littérature, mais avec un potentiel d'applications opérationnelles important et la possibilité d'exploiter les dernières avancées issues de domaines connexes, comme par exemple la détection d'anomalie via des comptes rendus d'appels (CRA) et les données GPS, les solutions d'apprentissage automatique en temps réel, les études et modélisation et de la réponse humaine aux catastrophes naturelles, etc. Ce corpus de recherche permet d'envisager une application rapide dans un contexte opérationnel et industriel, comme celui des Jeux Olympiques 2024.

[5] Le nombre d'appels agrégés sur chaque antenne et chaque réseau devrait permettre une détection d'anomalies significatives à partir d'informations cumulées à l'échelle de la minute.

[6] Les données de facturation (CRA) sont archivées durant un an pour toute la France. Il sera donc possible d'étudier des événements de l'année 2019 au cours du projet si nécessaire. Notons, que les données de flux ont été extraites et archivées dans le cadre du projet ANR CANCAN pour une période couvrant l'incendie de Notre Dame de Paris. Cet événement pourra donc être étudié dans le cadre du projet DISCRET.

[7] D'après l'étude annuelle de Médiametrie sur l'état de l'Internet en France (février 2019), les réseaux sociaux (avec 30 millions d'usagers journaliers) représentent la première activité sur Internet, soit ⅛ du temps passé sur Internet. D'après cette même étude ⅓ du surf des jeunes est consacré aux réseaux et médias sociaux. En terme d'usages, on assiste à l'ère du *mobile-only* après celle du *mobile-first*, renforçant le changement de paradigme qui fait de chaque acteur des réseaux sociaux sur mobile un "capteur" (*human-as-sensor*).

2.1 Positionnement par rapport à l'état de l'art

Les recherches actuelles sur la dynamique humaine se sont principalement limitées aux données recueillies dans des conditions normales et stationnaires et à des analyses visant à identifier les activités quotidiennes régulières ainsi que les motifs récurrents de mobilité ou de présence de groupes de personnes au sein de zones spécifiques [7,8,9,10,11,12,13], avec des résolutions spatio-temporelles limitées en général.

Cependant, **il existe un besoin spécifique d'identifier des situations anormales, liées à la présence et à la mobilité de masses de personnes dans des environnements urbains,** et de comprendre comment les personnes modifient leur comportement lorsqu'ils sont exposés à des conditions en mutation rapide ou inconnues [5,14]. Ces anomalies, et les réponses qui en résultent, peuvent être causées par des aléas naturels, technologiques ou sociaux, tels que des intempéries, des émeutes urbaines, des catastrophes, des mouvements de panique, des pannes d'infrastructures, etc. Ces phénomènes ont de profondes répercussions sur les modèles et les procédures opérationnelles conçus essentiellement *sur* et *pour* des situations stationnaires et donc naturellement susceptibles de s'effondrer dans de circonstances atypiques [15].

Par extension, les considérations susmentionnées s'appliquent également aux événements urbains rares ou programmés à l'avance, tels que les événements sportifs, culturels ou sociaux, qui attirent de grandes masses de personnes. Ceci peut créer des situations de stress et perturber, souvent profondément, le fonctionnement normal des infrastructures d'une ville ou d'un pays. Pendant ces situations, les foules urbaines représentent à la fois des entités fragiles exposées aux différents types d'anomalies urbaines et un potentiel élevé de désordres majeurs et d'effets en cascade difficiles à gérer, mais également un indicateur potentiel et efficace de détection de ces anomalies.

Comme le suggère le corpus des recherches récentes sur le ***Mobile Crowd Sensing*** **(MCS), la détection des anomalies urbaines et la caractérisation des interventions en cas de catastrophe semblent de plus en plus réalistes** [16,17]. Le MCS est défini comme un nouveau paradigme de détection qui permet aux citoyens ordinaires de faire remonter des données massives générées par leurs usages [18]. La possibilité d'étudier, en temps réel, les changements du "métabolisme urbain" via le MCS a vu le jour grâce à l'utilisation généralisée des téléphones mobiles, qui permettent de suivre à la fois la mobilité des utilisateurs [19] et les communications en temps réel le long des liens du réseau social sous-jacent [20]. MCS permet donc de collecter des données en continu, c.-à-d. relativement aux activités quotidiennes normales, lors de situations critiques ainsi que pendant les phases de suivi de différents types de catastrophes, tels que les tremblements de terre [21,22,23], les inondations [24] et les attaques terroristes [25].

Spécifiquement, **l'originalité du projet DISCRET réside dans l'idée de coupler deux des sources d'information les plus pertinentes dans les études MCS pour la détection et la classification des anomalies** : *i)* **les données passives des téléphones mobiles** qui incluent les données de facturation, comptes-rendus d'appels (CRA), et le données de signalisation réseau plus nombreuses et précises; *ii)* **les messages (*posts*) publiés sur les médias sociaux** par leurs utilisateurs, considérés ici comme des "capteurs humains" (*human-as-sensor*).

Au cours de la dernière décennie, de nombreuses recherches ont été menées sur l'utilisation de données passives collectées et anonymisées par les opérateurs de téléphonie mobile, notamment les CRA, afin d'identifier des modèles de mobilité et de présence humaines, étant donné la possibilité d'analyser des traces, individuelles ou agrégés, à une résolution spatio-temporelle sans précédent [26].

Dans leur travail fondateur, Bagrow *et al.* [5] ont prouvé que **les CRA volumineuses contiennent des indications précieuses sur les différentes façons dont les personnes réagissent aux urgences majeures** (attentats à la bombe, accidents d'avion, tremblements de terre et pannes majeurs

d'infrastructures urbaines), **ainsi qu'aux événements urbains majeurs mais non désastreux** (festivals de musique et événements sportifs). Plus précisément, il ont montré que les situations à haut risque entraînent une forte augmentation de l'activité de communication (nombre d'appels sortants et de messages de texte par rapport aux utilisations pendant les jours normaux), à proximité physique de l'événement. Cela confirme que les données mobiles agissent comme une sorte de « sociomètre » des perturbations externes. Le volume des appels commence à diminuer immédiatement après l'urgence, ce qui suggère que la propension à échanger est la plus forte au début des événements. En revanche, les événements festifs, qui attirent toutefois une foule nombreuse, montrent une croissance de l'activité de communication plus graduelle, une tendance très différente par rapport à celle de type *jump-decay* observée dans des situations d'urgence menaçantes. En ce qui concerne l'étendue spatiale des événements anormaux analysés, l'amplitude de l'anomalie de communication est la plus forte près de l'événement et diminue rapidement avec une décroissance exponentielle liée à la distance de l'épicentre et à la nature de l'événement. En particulier, les événements mettant la vie en danger ont un impact observable sur le volume des appels à plusieurs dizaines ou centaines de kilomètres, tandis que les autres anomalies restent confinées à une échelle urbaine ou périurbaine (par exemple, moins de 10 km).

En s'appuyant sur les conclusions de Bagrow *et al.* [5], d'autres travaux ont exploré le potentiel de CRA pour caractériser les conséquences d'événements tragiques, généralement à très grande échelle (c.-à.-d., un pays entier). Par exemple, Lu *et al.* [21] ont analysé les mouvements d'environ 2 millions d'utilisateurs de téléphones portables d'Haïti afin d'évaluer la prévisibilité de leur mobilité vers des destinations plus sûres en cas de séisme et de quantifier la diminution attendue de la population dans les secteurs touchés au cours des mois suivant la catastrophe. Les analyses au fil du temps de plusieurs indicateurs de mobilité et de présence, calculés à partir du mobile (p.ex., variation du rayon de giration estimé, entropie des lieux fréquemment visités) confirment qu'une meilleure compréhension de la façon dont les personnes réagissent aux catastrophes est possible grâce aux données de la téléphonie mobile. Cette compréhension augmentée peut potentiellement aider les décideurs à simuler et à prévoir le nombre d'évacués dans les zones urbaines avec un temps de calcul et un coût réduits. Des conclusions similaires et encourageantes peuvent être trouvées en ce qui concerne les tremblements de terre [27], les inondations [28] et les événements à grande échelle, tels que des concerts ou des célébrations et des manifestations civiles [29].

Cependant, il convient de noter que **les approches basées sur le CRA souffrent toutes d'une résolution spatio-temporelle souvent grossière et ne peuvent être conduite qu'à *posteriori*. Elles sont donc insuffisantes pour être appliquées dans des contextes urbains et temps réel**, comme ceux ciblés par le projet DISCRET.

Par ailleurs, toutes **les études mentionnées traitent de la réponse humaine aux anomalies en comparant les comportements de communication à des événements pour lesquels la nature, le moment et le lieu sont déjà connus**. Il s'agit d'une limitation fondamentale lorsque l'objectif est de concevoir une méthodologie pour la détection des anomalies en temps réel, sans aucune hypothèse sur les perturbations ou les événements critiques à venir.

Partant de ce constat, Dobra *et al.* [6] sont parmi les premiers auteurs à proposer un cadre méthodologique complet pour détecter automatiquement les situations d'urgence au moyen de données CRA, à grande échelle. Les auteurs décrivent un système complet de détection d'événements critiques qui, similairement aux idées proposées dans le projet DISCRET, identifie d'abord les comportements de routine, associés aux situations typiques, puis, sur la base de telles connaissances, détecte les périodes de comportements inhabituels d'appels ou de mobilités, ainsi que des informations sur l'emplacement et l'étendue géographique de ces perturbations. D'un point de vue plus technique, l'approche utilise la méthode des résidus pondérés normalisés pour estimer les probabilités d'appels/mobilité inhabituels provenant d'utilisateurs aléatoires et pour identifier les jours avec un

volume anormal de communications. Les auteurs ont pu vérifier que certains jours présentant des augmentations anormales d'appels et de comportements de mobilité correspondent bien à des événements exceptionnels. Mais, chose encore plus intéressante, les journées où les appels et/ou la mobilité sont en baisse correspondent également à des événements atypiques. Les écarts importants par rapport à une signature moyenne (nombre moyen d'appels) sont bien révélateurs d'une situation exceptionnelle.

De manière similaire, Dong *et al.* [30] proposent un cadre de détection d'évènements inhabituels à partir de données CRA individuelles et à grande échelle, où une anomalie est caractérisée par cun nombre important de personnes qui présentent le même comportement de mobilité inhabituel. La méthodologie de détection des anomalies urbaines repose donc sur une approche en trois étapes : *i)* le profil typique est construit pour chaque utilisateur via les données CRA historiques, en identifiant les antennes plus visitées par plage horaire ; *ii)* une approche de clustering est appliqué pour détecter des foules d'utilisateurs effectuant des appels à proximité d'un même ensemble de stations de base sur une fenêtre temporelle donnée ; *iii)* enfin, des foules inhabituelles sont détectées comme celles dont les utilisateurs présentent une dissimilarité moyenne significative par rapport à leurs profils typiques. L'approche de Dong et al. présente des limites évidentes : il est nécessaire d'identifier et de caractériser, à l'intérieur des foules, les profils historiques individuels. En d'autres termes, les utilisateurs doivent être suivis individuellement sur de longues séquences de jours, ce qui pose problème vis-à-vis de la protection de la vie privée.

2.2 Caractère innovant de la proposition

Bien que prometteurs, les travaux actuels en matière de détection automatique des anomalies sont des investigations préliminaires avec des limitations claires, notamment: *i)* ils ont une granularité spatiale grossière (par exemple, des anomalies sont détectées au niveau de régions urbaines macroscopiques avec des surfaces couvrant plusieurs km^2), ne détectent donc que des changements extrêmement marqués de mobilité ou présence humaines; *ii)* leur capacité à distinguer automatiquement les comportements atypiques est très limitée et souvent absente, ils exigent souvent une connaissance a priori des événements critiques surveillés, *iii)* ils sont fondés sur des analyses *a posteriori* et ne visent pas une détection en temps réel; *iii)* ils ne fournissent pas d'informations contextuelles sur le type d'anomalie détectée.

Pour surmonter ces limitations, **le projet DISCRET propose d'améliorer et d'étendre les méthodologies existantes de détection d'anomalies**, proposées dans le cadre de la recherche sur les CRA [6,30] et, plus généralement, sur les séries temporelles [31], **aux données de signalisation réseau**, qui seront collectées et préparées de façon conforme au RGPD[8] par le partenaire Orange [32]. En effet, les données de facturation téléphonique ne sont disponibles qu'à posteriori et ont, de plus, un contenu informatif plus limité : seuls les appels vocaux et les messages textes sont répertoriés. Pour des raisons techniques, les opérateurs de réseaux mobiles collectent aussi des données de signalisation réseau en déployant des équipements dédiés qui surveillent tous les messages de contrôle échangés entre les téléphones mobiles et l'infrastructure d'accès radio. Ces sondes collectent aussi des événements autres que les appels et les messages texte (p.ex., messages de protocole IP, données de sessions d'application, *hand-overs*, rafraîchissement de position, etc.), augmentant ainsi la fréquence d'échantillonnage spatio-temporelle et le contenu de l'information. Les données de signalisation capturent donc les positions horodatées d'une grande partie de la population (20 à 40% pour les opérateurs actuels), au niveau de l'antenne (généralement en améliorant d'un facteur trois la précision spatiale par rapport aux CRA) et à une fréquence de plusieurs ordres de grandeurs plus élevée que les

[8] Règlement général sur la protection des données (RGPD), règlement 2016/679 de l'Union européenne dont les dispositions s'appliquent depuis le 25 mai 2018.

CRA. ils ont donc le potentiel de dépasser une partie des limitations des sources précédemment utilisées.

Concernant l'identification automatique des comportements atypiques, **DISCRET s'appuiera sur les travaux antérieurs des partenaires du projet concernant la détection automatique des signatures urbaines à partir des données de téléphone mobile**. Furno *et al.* [11,12] ont récemment développé une approche basée sur des techniques d'apprentissage automatique non supervisé pour récupérer des descriptions informées et dynamiques d'activités typiques se déroulant dans différentes zones d'une ville, en analysant les données de téléphonie mobile (CRA) d'Orange. En résumé, l'approche permet de segmenter automatiquement la ville en différentes zones, en fonction de la similarité entre les signaux qui décrivent l'activité de communication agrégée (appels et SMS) typiquement observée auprès de chaque station de base du réseau mobile de la ville monitorée. Chaque zone est donc synthétisée par un profil temporel hebdomadaire (heure par heure), qui s'est avéré être bien corrélé avec l'utilisation socio-économique prépondérante à l'intérieur de la zone (par exemple, centre d'affaires, zone résidentielle, zone de loisir, zone mixte, etc.). L'approche permet également de repérer avec précision des tissus urbains très particuliers, associés par exemple aux centres principaux de transport multimodal (p.ex., gares, échangeurs d'autoroute, etc.), aux zones de culte et touristiques, aux centres d'activités sportifs, etc. L'approche propose une nouvelle façon dynamique de décrire les utilisations typiques au moyen de multiples séries temporelles au lieu de cartes géographiques grossières et statiques. Par ailleurs, les solutions proposées ont été récemment étendues aux données multi-sources [13].

Pour saisir automatiquement les comportements inhabituels des comportements typiques au niveau des antennes, DISCRET se propose donc de lever les verrous suivants : ***i)*** **les solutions existantes de récupération automatique des signatures de trafic mobile seront étendues aux données de signalisation du réseau les plus riches**. La définition de signature sera étendue afin de prendre en compte d'autres dimensions au-delà des appels et des messages de texte, telles que les différents types d'applications Internet utilisées par les usagers du mobile. En particulier, **pour la phase d'entraînement (construction des signatures urbains : lot 2) on s'appuiera sur des données historiques fournies par Orange et on ciblera plus particulièrement des zones spécifiques** (qui seront particulièrement intéressantes pour les JO), tels que les stades, les centres de transport et les zones de supporteurs; ***ii)*** **les signatures récupérées serviront de base à la détection en ligne des dérives pertinentes des événements de signalisation réseau observés**. Des méthodes probabilistes et statistiques [6] et des techniques de clustering [31] permettront d'identifier des anomalies potentielles. En particulier, afin de tester la capacité des solutions proposées à détecter des situations atypiques, les données de signalisation réseau collectées par Orange (dans le cadre du projet CANCAN) lors d'événements spéciaux connus (manifestations de mai 2019, incendies de Notre-Dame, matches de football) pourront être analysées dès le début du projet à titre d'étude de cas ; ***iii)*** **les techniques de détection d'anomalie en ligne doivent être suffisamment efficaces pour permettre une mise en œuvre en temps réel**. A cet égard, des solutions d'apprentissage automatique distribué sur des données en continu seront adoptées et testées en temps quasi-réel.

Enfin, pour enrichir l'information produite et fournir des éléments contextuels sur la nature des événements atypiques détectés, **DISCRET collectera et analysera, en temps réel, des données issues du réseau social Twitter.** Earle *et al.* [33] ont évalué l'utilisation de Twitter pour la détection des tremblements de terre en cartographiant la zone touchée au travers de tweets générés après le tremblement de terre de Morgan Hill, Californie, le 30 mars 2009. De même, les caractéristiques temporelles et spatiales d'un séisme survenu aux États-Unis sont analysées par Crooks *et al.* [34] en utilisant des messages Twitter. Ces messages sont considérés comme une forme hybride d'un système de capteurs distribués pouvant être utilisé pour identifier la zone d'impact du séisme. Sakaki *et al.* [35,36] ont étudié les tremblements de terre en temps réel via Twitter et ont proposé un algorithme

pour la détection d'événements critiques à l'aide de l'analyse des tweets. Afin de traiter les informations fournies par les "capteurs sociaux", ils ont mis au point un classificateur de tweets. Pour traiter les informations spatiales, ils ont produit un modèle spatio-temporel probabiliste permettant de localiser le centre de l'événement. En tant qu'application, ils ont développé un système de compte rendu de tremblement de terre à utiliser au Japon. Vieweg *et al.* [37] ont proposé d'améliorer la compréhension du contexte pendant des situations d'urgence (événements d'inondations au Mississipi et en Oklahoma au printemps 2009) grâce aux informations de micro-blogging afin d'améliorer la perception de la situation et de son évolution et de permettre un déploiement efficace de services de secours.

Afin d'étudier la dynamique et l'évolution des communautés dans les réseaux sociaux en réponse à des situations d'urgence, Lu et Brelsford [38] ont collecté des données sur Twitter avant et après le séisme de 2011 au Japon. Ils ont montré que la dynamique et l'évolution des réseaux sociaux sont nettement modifiées en situations d'urgence. De façon similaire, une série d'incendies survenus récemment à Santa Barbara, aux États Unis, ont permis d'examiner les opportunités liées à l'information géographique communiquée volontairement par la population via des médias sociaux (p.ex., Flickr, Google MyMaps et Twitter) et son rôle potentiel dans la gestion des crises et des catastrophes [39, 40].

En conclusion, s'appuyant sur des jeux de données de signalisation réseau et sur des méthodologies existantes pour la détection et la caractérisation d'anomalies**, le projet DISCRET vise à développer un système complet, capable de détecter et de classer précisément, avec un bon niveau de confiance, de multiples types d'anomalies urbaines générant des flux d'activité atypiques.**

DISCRET cible une solution fonctionnant en temps réel et capable d'identifier des situations atypiques au niveau de zones urbaines de 0,1 km^2 et un pas de temps d'échantillonnage d'une minute, indispensables pour le suivi en temps réel d'événements à forte dynamique spatio-temporelle. En atteignant son objectif, DISCRET constituerait une avancée majeure en matière de surveillance urbaine, démontrant pour la première fois la viabilité, sur la base de scénarios réels, d'une approche de détection de situations atypiques sur la base de données de téléphonie mobile complétées par des informations issues du réseau social Twitter.

3 Programme scientifique et technique, organisation du projet

3.1 Programmation et organisation du projet

Le projet DISCRET est organisé en quatre lots (cf. organigramme de la figure 1) :

1. La coordination, le reporting et la valorisation du projet (Lot 1).
2. La constitution et la préparation des jeux de données (Lot 2). L'analyse de ces données et la détermination des signatures spatio-temporelles de l'activité du réseau de téléphonie mobile est l'étape clé du projet, qui permettra de d'identifier les variables déterminantes et les critères de détection des anomalies (excentricités).
3. La proposition de procédures de détection automatique d'anomalies via les données de téléphonie mobile et d'extraction d'informations contextualisées à partir du flux Twitter (Lot 3) et l'évaluation des performances de ces procédures par comparaison avec les observations issues du terrain. Ce lot est le coeur du projet.
4. Le développement d'une plateforme de démonstration qui permettra le rejeu d'événements majeurs analysés au cours du projet et une étude de faisabilité pour préparer le déploiement opérationnel des solutions proposées (Lot 4).

Le projet prendra appui sur une période d'observation intensive du flux d'information issu du réseau de téléphonie mobile (plage rouge sur le diagramme de Gantt, figure 2). Les séries de données collectées actuellement dans le cadre du projet ANR CANCAN (mars - juin 2019)[9], ainsi que les relevés de facturation (CRA), archivés durant un an, seront aussi exploitées et permettront d'engager les analyses dès le début du projet. Le contenu détaillé, les livrables et jalons des lots, ainsi que les risques et les solutions de repli correspondantes sont décrits succinctement ci-après.

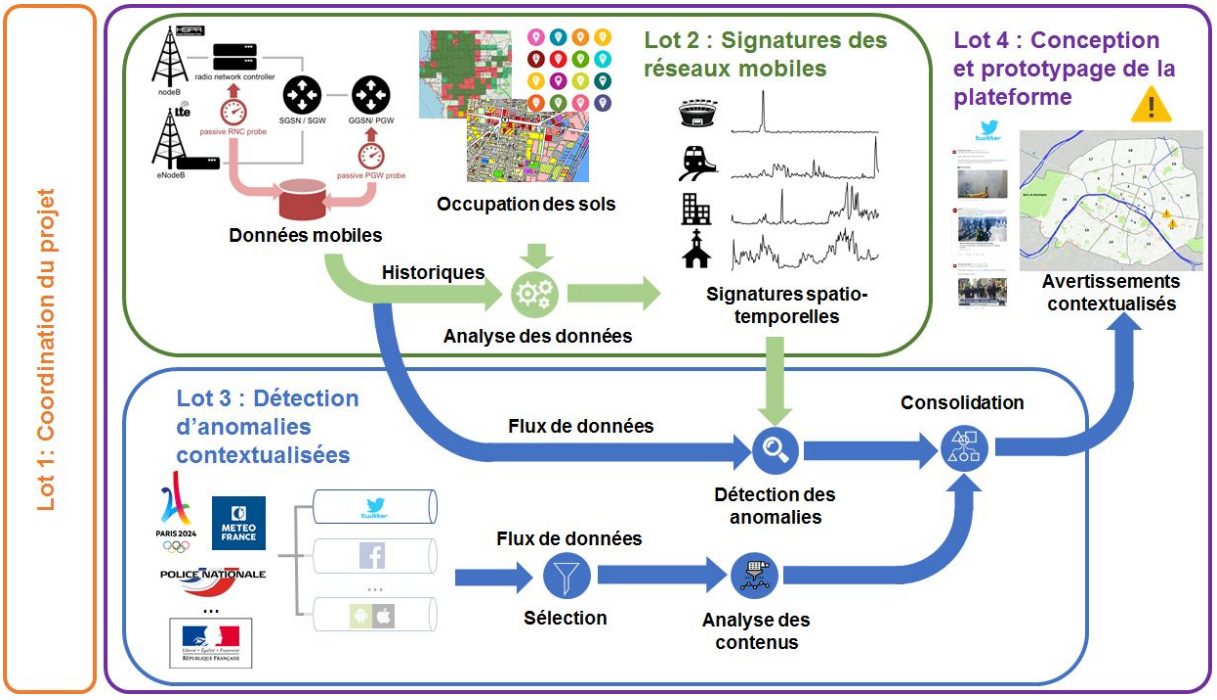

Figure 1 : Organisation du projet

[9] Ce jeu de données contient notamment les informations relatives à plusieurs événements majeurs (incendie de Notre Dame, manifestation du 1er mai à Paris, marathon de Paris).

3.2 Description des travaux par tâche

Lot 1 : Coordination du projet

Leader	**Ifsttar**			
Calendrier	M1→M18			
Efforts (p.m)	**Ifsttar**	**Orange**	**UTT**	
	5	1	1	
Objectifs	Le Lot 1 est consacré à la coordination du projet ainsi qu'à la valorisation de ses résultats.			
Indicateur de succès	Tenu du calendrier du projet, tenu des réunions de travail et d'un séminaire de clôture, diffusion et valorisation des résultats			
Risques & solutions	Les risques de ce lot sont essentiellement liés aux retards possibles d'exécution du planning des tâches. Des réunions régulières sont prévues au cours du projet afin de favoriser les échanges et de réaménager le programme de travail si besoin.			
Livrables et jalons	D1.1 Kick-off : séminaire de démarrage du projet (M1) D1.2 Accord de consortium (AC) (M3) D1.3 Plan de gestion des données (M6) D1.4 Rapport d'avancement intermédiaire (M9) D1.5 Réunions intermédiaires de revue de projet (M6 et M12) D1.6 Colloque de clôture (M18) D1.7 Rapport final du projet (M18)			

Tâche 1.1 - Pilotage du projet et valorisation (M1→M18) - Ifsttar, Orange, UTT
La coordination du projet sera assurée par l'Ifsttar. Le projet reposant fortement sur la disponibilité des données d'Orange (cf. Lot 2), les procédures de demande d'accès aux données de téléphonie mobile au sein d'Orange ainsi que les discussions sur l'accord de consortium seront engagées dès le mois de juin 2019. Il est prévu de tenir quatre réunions du consortium afin d'assurer le suivi de l'avancement du projet et de faciliter les échanges scientifiques. Par ailleurs un webinar mensuel de suivi rassemblera les responsables de tâche durant toute la durée du projet.
Tâche 1.2 - Administration et Reporting (M1→M18) - Ifsttar, Orange, UTT
Cette tâche comprend la coordination de la rédaction des rapports d'avancement et final du projet (M9 et M18), la rédaction et l'instruction de l'Accord de Consortium (M3) et le plan de gestion des données (M6). Par ailleurs, l'équipe de coordination assurera le lien avec l'ANR et le SGDSN et participera à l'ensemble des réunions qui seront organisées dans le cadre de l'appel à projets JO 2024.
Tâche 1.3 - Communication et valorisation (M12→M18) - Ifsttar, Orange, UTT
Le projet de recherche est court (18 mois) : la plupart des actions de communication et de valorisation se concentreront donc à la fin du projet. Il est notamment prévu un colloque de clôture rassemblant 50 à 100 personnes et de présenter les résultats du projet à la conférence Netmob qui se tiendra en 2021. Plusieurs publications scientifiques pourront être envisagées à l'issue du projet. S'ils sont probants, **les résultats du projet DISCRET devraient aussi pouvoir donner lieu au développement d'une application opérationnelle par Orange, dont l'étude de faisabilité constitue la tâche T4.2 du projet DISCRET.**

Lot 2 : Collecte, analyse et extraction des signatures des réseaux mobiles

Leader	**ORANGE**			
Calendrier	M1→M16			
Efforts (p.m)	**Ifsttar**	**Orange**	**UTT**	
	22	18	1	
Objectifs	Le Lot 2 porte sur la constitution des jeux de données pour le projet et l'élaboration des signatures à partir des **flux des sondes réseaux 2G, 3G et 4G,** ainsi que des historiques des données de la facturation (CRA). De nouveaux ensembles de données très détaillées sur le comportement des usagers de mobile seront recueillis, ainsi que des informations détaillées sur la nature, l'occurrence et la localisation des incidents éventuels au cours des événements observés (T3.3). Des **signatures spatio-temporelles** de l'activité du réseau (agrégée au niveau des antennes) seront définies.			
Indicateur de succès	Construction de la typologie spatio-temporelle des signatures et mise en forme des données relatives aux événements d'intérêt. Mise à disposition de l'information nécessaire pour les travaux des Lots 3 et 4.			
Risques & solutions	Les risques dans ce Lot sont liés au processus de collecte des données et aux autorisations nécessaires pour partager ces données avec les partenaires du projet dans le respect de la RGPD. Le premier risque est atténué par l'expérience importante du partenaire Orange sur ce point (ex. collecte de données pour deux éditions du *Challenge Data for Development* (D4D) [41], et dans le cadre des ANR ABCD et CANCAN). Le second risque est lié à l'anonymisation et à la finalité des traitements. Tous les traitements de données personnelles (agrégations par antenne) seront effectués sur l'infrastructure du contrôleur des données Orange et sous contrôle de son *Data Protection Officer* (DPO). Seuls les agrégats et éléments anonymisés pourront quitter les espaces sécurisés. La durée de conservation des données personnelles chez Orange (12 mois) est suffisante pour réaliser les tâches des Lots 3 et 4.			
Livrables et jalons	D2.1 Inventaire de données mobile disponibles pour projet (M1) D2.2 Rapport technique méthodologie pour les signatures type (M12) J2.1 Mise à disposition des données formatées (M4) J2.2 Mise à disposition des signatures type simples (M5) J2.3 Mise à disposition des signatures type multidimensionnelles (M11) J2.4 Mise à disposition d'une base de données issue des enquêtes post-événements (M16)			

Tâche 2.1 - Collecte et formatage des données de réseaux mobiles (M1→M15) : <u>Orange</u>, Ifsttar
Les données de signalisation mobile seront collectées et préparées sur l'infrastructure d'Orange. Il faudra fusionner les données provenant de deux sondes réseaux. La mise en cohérence et nettoyage des données seront réalisés sur la plateforme sécurisée Big Data d'Orange Labs (cluster Hadoop) qui offre tous les outils Big Data avancés nécessaires. Par ailleurs, les données similaires sont actuellement collectées et traitées sur cette plateforme pour le projet ANR CANCAN et les outils informatiques développés pourront être réutilisés pour DISCRET.
Une des difficultés de ce lot consiste à collecter les données décrivant l'activité urbaine lors d'événements qui attirent de grandes masses de personnes comme lors des Jeux Olympiques. Pour pallier ce problème, une double technique d'atténuation des risques sera adoptée: i) les données historiques seront collectées (historique des CRA disponible avec une granularité grossière); ii) les

données de signalisation réseau fines collectées dans le cadre du projet CANCAN seront utilisées pour la construction des signatures lors d'événements importants déjà survenus (ex. incendie de Notre Dame, manifestation du 1er mai 2019 à Paris, Marathon de Paris 14.04.2019). DISCRET collectera les données pour les événements à venir comme la Fête de la Musique, 14 juillet, et sur les événement sportifs importants (Roland-Garros, Triathlon de Paris, Tour de France...) afin d'observer différentes configurations spatiales de ce type d'événements en ville.

Tâche 2.2 - Extraction des signatures type (M2→M12) : Ifsttar, Orange

Cette tâche concerne la transformation des données brutes extraites du réseau mobile en signatures d'événements par antenne afin de supporter la phase de détection d'anomalies/situations atypiques (Lot 3). Elle nécessite des ressources importantes (mémoire et puissance de calcul) et sera réalisée sur l'infrastructure d'Orange. Il s'agit de sélectionner et d'implémenter le filtrage sur des périodes et lieux, et d'extraire les signatures temporelles du trafic téléphonique et d'usage de l'internet mobile. La méthode d'extraction de signatures utilisera une version étendue de l'approche initialement proposée dans [11]. L'extension de cette approche doit permettre une **définition multi-événements (c.-à.-d, multidimensionnelles) et plus fine de l'activité mobile typique par antenne.**

S'agissant de la granularité temporelle de la signature avec les données de signalisation réseau, cette tâche étudiera la possibilité de construire des signatures avec une résolution à la minute (des périodes d'agrégation plus longues: 5, 10 ou 15 min seront également examinées). Pour la dimension contextuelle, c'est-à-dire la nature de l'activité sur le réseau mobile, chaque antenne sera décrite par une combinaison de plusieurs séries temporelles, chacune associée à un ensemble d'événements : (i) réseau, (ii) communications, (iii) usages de services Internet spécifiques, (iv) mobilité entre zones.

Pour la granularité spatiale, les signatures seront d'abord définies pour chaque antenne du réseau mobile Orange. Une approche hiérarchique sera aussi considérée afin de reconfigurer le zonage d'événements selon des critères de similarité des signatures [11,12] ou de proximité spatiale des lieux d'intérêt [10]. Des méthodes de clustering ou de factorisation tensorielle pour les signatures multidimensionnelles par antenne seront mobilisées en s'inspirant des approches les plus récentes [42,43]. Cela permettra une implémentation efficace et distribuée des méthodes temps-réel de détection d'anomalie mises en oeuvre dans le Lot 3.

Les méthodologies proposées assurent la protection de la vie privée des usagers, i.e. *privacy by design*. La détection des événements est agrégée au niveau du point de collecte des données (antenne relais). L'interprétation d'anomalies concerne des comportements

collectifs - dans la version opérationnelle du projet il n'y a plus de traitement des traces individuelles.

En s'appuyant sur les données déjà collectées par Orange (dans le cadre du projet CANCAN), cette tâche fournira rapidement (à M5) un premier jeu de signatures simples (c.-à.-d., au niveau de l'antenne et relativement aux événements agrégés d'appels et SMS) au Lot 3. Un ensemble plus complet de signatures complexes (multidimensionnelles) sera fourni comme deuxième sortie à M11.

Tâche 2.3 - Documentation des événements (M2→M16) : Ifsttar, UTT

Il s'agit dans cette tâche de rassembler les informations sur les divers incidents ayant pu se produire durant les événements étudiés dans le projet DISCRET (lieux et instants d'occurrence) à partir de l'exploitation de diverses sources : médias, médias sociaux, mains courantes des services d'intervention et de secours, interviews de témoins sur le terrain. Ces informations permettront d'évaluer la pertinence des méthodes de détection d'anomalies proposées pour l'identification d'incidents (Lot 3, T3.3) et seront intégrées à la plateforme de démonstration (Lot 4). Le projet prendra appui sur l'expérience de l'Ifsttar en matière de retour d'expérience post-événements, développée pour l'étude des crues soudaines [44].

Lot 3 : Méthodes de détection d'anomalies contextualisées

Leader	**UTT**			
Calendrier	M1 →M16			
Efforts (p.m)	**Ifsttar**	**Orange**	**UTT**	
	13	20	40	
Objectifs	Le premier objectif de ce lot est de proposer des méthodes de détection de situations atypiques dans le flux de communication mobile à partir des signatures urbaines types provenant du Lot 2. Ces méthodes doivent pouvoir être implémentées en temps réel. Le deuxième objectif du lot est la collecte, l'agrégation et le traitement de données générées par les utilisateurs de réseaux sociaux (Twitter dans le cas présent). Le travail dans ce lot utilise des algorithmes de data mining et de traitement du signal (le flux Twitter étant considéré comme un signal) pour compléter la détection d'anomalies et contextualiser les anomalies à partir de l'analyse automatisée des contenus publiés. Le dernier objectif est le développement de techniques pour la consolidation et la fusion des méthodes de détection d'anomalie à partir les données de téléphonie mobile et Twitter, afin d'augmenter le niveau de confiance et les informations contextuelles disponibles pour qualifier les anomalies détectées et d'évaluer les approches proposées par confrontation des anomalies détectées avec les observations de terrain rassemblées dans la tâche 2.3.			
Indicateur de succès	Mise à disposition d'une méthode temps quasi réel pour la détection de comportements de communication mobile anormaux. Collecte et formatage des données de Twitter en modes passif et adaptatif (pour les événements d'intérêt). Prise en compte des sorties du Lot 2 dans la stratégie de collecte adaptative. Mise à disposition d'une méthode de géo-inférence pour la consolidation. Mise à disposition d'une méthode de fusion d'anomalies détectées à l'aide de Twitter et des données de téléphonie mobile. Evaluation des performances des approches proposées			
Risques & solutions	Le premier risque dans ce lot repose sur la capacité de disposer de données de Twitter exhaustives tant en mode passif qu'en mode adaptatif. Pour atténuer ce premier risque nous prévoyons de mettre en place une location de l'API complète de *Twitter Firehose* sur des événements ponctuels à côté de l'accès développeur à l'*API Streaming*. Le second risque réside dans la possibilité que les données de téléphonie ne soient pas prêtes au démarrage de la Tâche 3.1. Ce risque reste limité par les précaution prises dans le Lot 2.			
Livrables et jalons	D3.1: Protocole d'interfaçage avec les plateformes TweetCapt (M2) D3.2 Algorithme multi-indicateur de geo-inférence de tweets (non géolocalisés) (M14) D3.3 Rapport technique de l'algorithme de détection d'événements et d'anomalie (M16) J3.1 Méthode de détection des anomalies dans le flux de communication mobile (M14) J3.2 Mise à disposition de la base de tweets et des événement géo-inférés (M16) J3.3 Mise à disposition d'algorithmes de fusion et consolidation (M16)			

Tâche 3.1 Détection d'anomalies avec données de téléphonie mobile (M6→M15) - Ifsttar, Orange
Cette tâche étudiera les techniques de classification efficaces pour l'inférence de situations atypiques (augmentation du volume de l'activité de communication et de consommation de certains services, croissance soudaine d'événements liées à la mobilité, changement de forme du signal, etc.) par rapport aux signatures prototypiques de téléphonie mobile provenant du Lot 2. Cette tâche s'attachera également à la création d'un processus de mise à jour périodique de signatures afin d'adapter l'approche à des variations de fréquentation et donc de signatures typiques de certains lieux d'intérêt.
D'un point de vue méthodologique, le principal défi consiste à développer une méthode de classification qui puisse fonctionner en temps réel. Pour atteindre ce but, une méthode combinant l'intelligence artificielle (IA) et l'apprentissage statistique sera mise en oeuvre (méthode de classifications de courbes, méthode à noyaux ou méthode générative). Elle sera complétée par des tests statistiques séquentiels (test de Wald) pour détecter en temps réel les écarts par rapport à des prototypes (profils ou signatures types). Ces techniques visent donc à détecter explicitement des dynamiques différentes de celles capturées par des signatures de référence.
D'autre part, le volume de données à analyser étant important, la détection d'anomalies et la mise à jour périodique des signatures types devront être réalisées au plus près de la source des flux de données afin de minimiser la charge et la latence du réseau (*Mobile Edge Computing*, cf. Lot 4). Pour mettre à jour les signatures types, nous utiliserons une méthode d'apprentissage distribué comme le *Federated Learning* [45] implémenté au niveau du *Edge Computing* [46,47]. Cette approche permet une estimation d'un modèle de "classifieur" initial, calculé localement, qui partage ensuite ses paramètres avec les serveurs de même niveau (voisinage réseau). Le modèle global est ensuite recalculé à partir des différents modèles locaux.
Du point de vue technologique, la complexité informatique de l'exploration des données requiert l'utilisation de solutions appropriées pour l'implémentation temps-réel et évolutive. Dans le cadre du projet nous utiliserons les plateformes open source Apache Kafka et TensorFlow qui répondent à ce critère.

Tâche 3.2 Détection d'anomalies potentielles via Twitter (M1→M16) - UTT
Cette tâche se décline en trois grandes étapes : La première étape repose sur la mise en place d'une **veille continue et adaptative de Twitter,** qui permettra de collecter les tweets sous leur format natif (JSON). Le module de collecte TweetCapt®, développé par l'UTT sera mis en oeuvre. Ce module permet de stocker le contenu des Tweets, les médias échangés, les URLs partagés, tout en générant des statistiques sur les différentes entités (#hashtags, @mention, etc.) et en construisant différents graphes d'interactions: entité-entité, entité-ressources, etc. Dans le cadre de la plateforme DISCRET, trois stratégies seront mises en œuvre: (i) une **collecte passive permanente** en utilisant les propriétés de l'Interfaces de Programmation Applicative (API) *Streaming* de Twitter, qui délivre un flux en continu; (ii) une **collecte ciblée,** suivant des paramètres définis à priori (mots clés, comptes institutionnels ou événementiels officiels, zones géographiques d'intérêt); (iii) une **collecte contextuelle** qui se déclenche sur la base des sorties de l'analyse des données d'usage de téléphonie mobile et de leurs applications associées. L'objectif final est d'avoir une stratégie de collecte à large spectre, qui puisse garantir la prise en compte non seulement la volumétrie, mais aussi la variété et surtout la variabilité des phénomènes observés (les 3V du Big Data).
La seconde étape, qui concerne la détection d'événements et d'anomalies, se fera en trois temps. Premièrement il s'agit de détecter (globalement) l'occurrence par l'analyse du flux de tweets (considéré ici comme un signal spatio-temporel). La problématique revient alors à un problème classique de traitement de signal pour déterminer des sauts inhabituels dans la fréquence de publication en général ou en rapport avec certains entités clefs (#hashtags, @profils, médias, URLs, etc.). Deuxièmement, une approche probabiliste d'analyse conjointe des sauts dans ces différents

"signaux" permettra de fusionner ces observations et d'identifier les écarts. Troisièmement, il s'agira de qualifier la nature normale ou anormale de l'événement détecté. Pour cela on aura recours à l'analyse du graphe d'activités qui n'est autre que le graphe multi-mode qui intègre les différentes types d'interactions dans Twitter (mention, retweet, reply) et le partage des contenus (redirection vers les serveurs d'hébergement des contenus - images, vidéos, liens, etc). La détection d'anomalies dans la structure de ce graphe dynamique (dans le temps), croisée avec la détection d'événements, permettra d'enrichir la base de connaissance de la plateforme DISCRET.
Pour finir, les sorties de la détection d'événements et d'anomalies sur les données de Twitter seront utilisées pour documenter les événements de la base de connaissances de DISCRET. Ces événements seront ainsi enrichis par des éléments, extraits des tweets sélectionnés, en constituant une base de métadonnées spatio-temporelles [48] de descriptions contextualisées (statistiques, dataviz, etc.) plus fines de "l'environnement de l'événement": période, zone géographique, profil du flux des échanges, contenu des échanges…, afin d'assurer la phase de documentation des événements (Tâche 2.3).

Tâche 3.3 Fusion des méthodes et évaluation (M10→M16)- **Orange,** Ifsttar, UTT

Cette tâche intégrera et consolidera les anomalies détectées par les écarts aux signatures type de téléphonie mobile (T3.1) et les alertes détectées via Twitter (T3.2). Les données d'historique des signatures de l'activités (mobile et Twitter) seront mises en forme pour une présentation dynamique de type séries temporelles, permettant de générer différents niveaux d'agrégation d'information si besoin (alignement géographique et temporel).
Tout d'abord, face au faible pourcentage de tweets géolocalisés, nous proposons un algorithme de géo-inférence des tweets comme un autre axe de consolidation des détections de la Tâche 3.2. Cet algorithme exploite plusieurs indicateurs [48] déduit des métadonnées et du contenu des tweets pour enrichir les détections par une composante spatiale très souvent manquante. Pour un événement donné, l'algorithme utilise les tweets géolocalisés nativement pour constituer une zone de référence à partir de laquelle on joue une combinaison d'un bootstrap (dans le cas des tweets émis depuis un location-based service) et d'un Monte-Carlo (pour les autres tweet).
Sachant que plusieurs alertes - chacune avec un niveau de confiance différent - pourront être produites par les deux sources de données (téléphonie mobile et Twitter) et pour les mêmes zones géographiques (recueillies par géolocalisation ou géo-inférence), le deuxième objectif de cette tâche est de proposer un formalisme de fusion compatible avec le temps réel, dans lequel les différentes décisions hétérogènes seront fusionnées pour une prise de décision : avertissement ou non.
Pour fusionner les résultats de deux types des alertes, nous utiliserons la théorie de l'évidence de Dempster-Shafer pour opérer la fusion des différents indicateurs [49,50,51,52,53]. Une implémentation quasi temps-réel basée sur les travaux de [51,54] sera développée pour combiner les croyances individuelles associé aux anomalies provenant des sources individuelles, en utilisant les scores de confiance fournis comme poids. Cette approche devrait permettre de réduire le nombre de fausses alarmes et fournir un liste d'anomalies significatives contextualisées avec un niveau de confiance accru.
Une fois la fusion effectuée et les séries spatio-temporelles des anomalies détectées seront comparées pour plusieurs événements - une dizaines a priori - aux incidents observés sur le terrain et documentés pour le projet DISCRET au travers des enquêtes post-événements conduites dans la tâche 2.3. Cette comparaison permettra une évaluation qualitative des performances de la méthode de détection d'anomalies. L'ensemble des séries temporelles seront intégrées au sein de la plateforme de démonstration et de visualisation pour permettre le rejeu des événements.

Lot 4 : Conception et prototypage du plateforme : intégration et exploitation

Leader	**Orange**			
Calendrier	M3→M18			
Efforts (p.m)	**Ifsttar**	**UTT**	**Orange**	
	1	4	27	
Objectifs	Le Lot 4 se focalise sur la conception et le prototypage d'une plateforme d'***Intelligence Artificielle*** permettant de monitorer de manière réaliste les événements qualifiés d'excentricités (anomalies). Ce lot comprendra deux tâches : (i) la conception d'une interface de visualisation et de démonstration qui sera alimentée par les résultats du Lot 3 et (ii) la réalisation d'une étude de faisabilité pour préparer le déploiement opérationnel des méthodes proposées dans le Lot 3.			
Indicateur de succès	Livraison d'une plateforme de visualisation et de démonstration permettant le rejeu d'un certain nombre d'événements étudiés dans le cadre du projet DISCRET. Etude de faisabilité identifiant les principales contraintes de mise en oeuvre opérationnelle des approches proposées, leur compatibilité avec l'architecture du réseau et identifiant les grands choix méthodologiques et technologiques de conception et de mise en oeuvre, compte tenu des contraintes liées au volume de donnée à traiter et à la fréquence de calculs souhaitée.			
Risques & solutions	Les risques de ce lot sont liés aux retards éventuels de livraison des résultats du Lot 3. Des échanges réguliers avec les acteurs du Lot 3 et le pilotage de la Tâche 3.3 par Orange permettront d'intégrer les avancées et résultats du Lot 3 progressivement et donc d'avancer dans la réalisation du Lot 4, sans attendre la livraison de l'ensemble de ses résultats du Lot 3 en M16.			
Livrables et jalons	D4.1 Plateforme de visualisation et de démonstration (web-service) (M18) D4.2 Etude de faisabilité pour le déploiement opérationnel (M18) J4.1 Cahier des charges de la plateforme de visualisation et démonstration (M4)			

Tâche 4.1 Plateforme de visualisation et de démonstration (M3→M18) - Orange, Ifsttar
Dans cette tâche on propose une interface de visualisation et de démonstration simple de type web-service. Compte tenu de la durée du projet, il ne s'agit pas d'un produit final mais d'un prototypage. La plateforme ne sera pas conçue à ce stade pour traiter les données en temps réel. Elle permettra le rejeu d'un certain nombre d'événements étudiés dans le cadre du projet DISCRET en offrant une interface de visualisation (i) des séries temporelles d'anomalies localisées, (ii) des contenus issus de la sélection des tweets et de les comparer aux (iii) déroulements effectifs des événements, reconstitués à partir des enquêtes post-événements (Tâche 2.3). L'interface proposée permettra aux utilisateurs potentiels de visualiser ce que pourrait être une future plateforme d'avertissement, basée sur l'analyse de l'activité des réseaux de téléphonie mobile. L'interface proposée **donnera la possibilité à l'utilisateur de valider ou d'invalider les avertissements** (intervention humaine), pour simuler un fonctionnement opérationnel où il disposera éventuellement d'autres sources d'informations.

Tâche 4.2 Etude de faisabilité pour le déploiement opérationnel de la solution (M12→M18)- Orange, Ifsttar, UTT

La cloudification des infrastructures et l'automatisation du monitoring des éléments cœurs des réseaux mobiles sont des enjeux clés pour tous les opérateurs de télécommunications. Cette tâche analysera les contraintes relatives à un déploiement opérationnel de la plateforme d'identification des anomalies, recommandera des solutions techniques en se basant sur les résultats du prototype de plateforme implémenté en Tâche 4.1 et les méthodes sélectionnées dans le Lot 3.

Les grands volumes de données exigent en général des ressources importantes pour la transmission vers le serveur, le stockage et le calcul, soit une centralisation dans une infrastructure cloud cumulant des temps potentiellement incompatibles avec une réactivité temps réel. De plus l'émergence de micro phénomènes au niveau local serait noyée dans la grande masse de données et demanderait plus de temps avant d'être perçu comme pertinents. La plateforme proposée, devra donc reposer sur une approche *Edge Computing*, basée sur l'implémentation d'un système d'apprentissage distribué (*Federative Learning*) pour la détection et la prédiction d'excentricités. Les enjeux de cette tâche résideront notamment dans la compatibilité des approches proposées dans le Lot 3 avec l'architecture du réseau existant, leur scalabilité et leur mise en œuvre dans un dispositif de type *Mobile Edge Computing*.

3.3 Calendrier des tâches, livrables et jalons

La figure 2 présente le calendrier prévisionnel du projet DISCRET et reprend de manière synthétique les dates prévues des livrables et jalons indiquées dans la description des lots. Les noms des responsables de lots et de tâches sont précisés dans le tableau de la partie 4.3.2 ci-dessous.

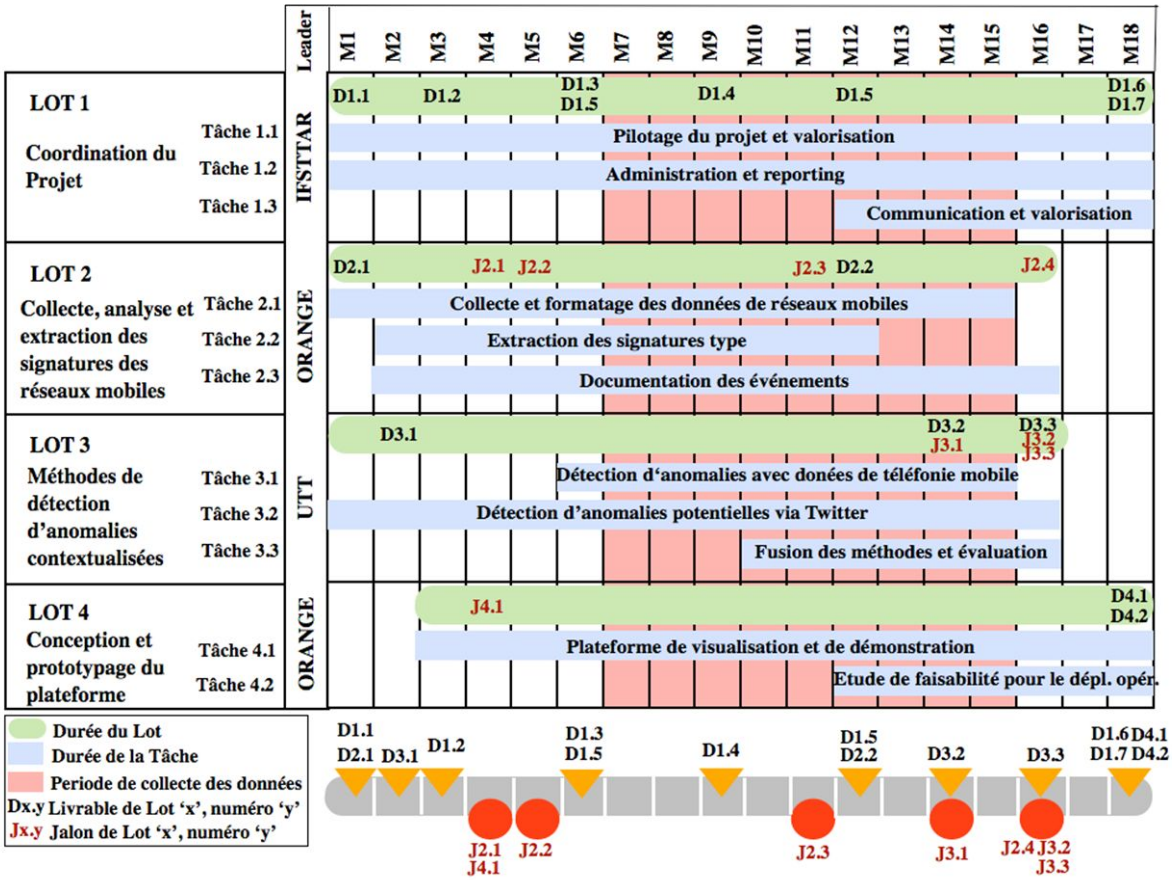

Figure 2 : Calendrier prévisionnel du projet DISCRET, livrables et jalons

3.4 Justifications scientifiques des moyens demandés

Les principales dépenses du projets sont (1) des dépenses de personnel (contribution aux salaires des personnels permanents pour Orange, deux post-docs recrutés pour travailler sur les Lots 2 et 3 et une contribution au financement de salaires d'ingénieurs CDD de l'UTT pour l'extraction des tweets), (2) des frais de mission pour les partenaires (réunions du projet, enquêtes de terrain, dissémination) et (3) des frais d'organisation du séminaire de clôture et d'accès à des API payantes de Twitter. Le détail des coûts et moyens demandés est le suivant.

Ifsttar

Personnel : un post-doc de 18 mois sera recruté par l'Ifsttar pour le projet (T2.2 et T3.1) : 73 862 €. Il sera co-encadré par Orange - le profil recherché est un profil informatique / data science. Il contribuera à la préparation et l'analyse des données mobiles dans un environnement Hadoop/Spark définition des signatures et des méthodes de détection d'anomalies à partir des

données de téléphonie mobile. Trois stages de Master ou de fin d'étude sont aussi prévus, notamment pour contribuer à la tâche 2.3 : environ 5000 € par stage.

Missions : quatre réunions de projet sont prévues à Paris depuis Bron ou Nantes : environ 20 missions à un coût de 200 € soit 4000 €. Dix analyses post-événements sont programmées pour un coût moyen de 500 € de missions. Enfin trois participations à des colloques internationaux sont envisagées (2 000 € en moyenne par mission : frais d'inscription et déplacements)

Autres dépenses : petit matériel dont renouvellement informatique (3 000 €), organisation d'un colloque de clôture pour 50 à 100 personnes (15 000 €). Frais de publication (2 000 €).

ORANGE

Personnel : Les coûts des personnels ont été évalués selon les temps (p.mois) - 12 personnes mois d'ingénieur expert : 119 833,80 € et 54 personnes mois d'ingénieur R&D : 370 321,74 €.

Missions : Participations aux réunions de projet (9 000 €), trois participations à des colloques internationaux (6 000 €).

UTT

Personnel : un post-doc de 12 mois sera recruté par l'UTT pour travailler sur la tâche 3.2 - profil informatique / data science (environ 50 000 €) et 6 mois de salaire d'ingénieur plateforme informatique pour le paramétrage et l'extraction des tweets pour le besoin du projet - plateforme informatique UTT seront financés (23 000 €).

Missions : une vingtaine de missions individuelles à Paris pour les réunions de projet (3 000 €). Trois participations à des colloques internationaux (6 000 €)

Autres dépenses : petit matériel dont renouvellement informatique (4 000 €). Frais de publication (2 000 €). Frais d'accès aux API Twitter payantes (15 000 €).

Le coût complet du projet (salaires des personnels permanents compris) s'élève à près de 1,5 millions d'euros environ. Le taux de financement de l'ANR est donc de l'ordre de 33 %. Le taux de précarité (p.m CDD rapporté aux p.m totales) est de 24 %. Les moyens demandés à l'ANR se répartissent comme suit : 52% pour Orange, 27% pour l'Ifsttar et 21% pour l'UTT.

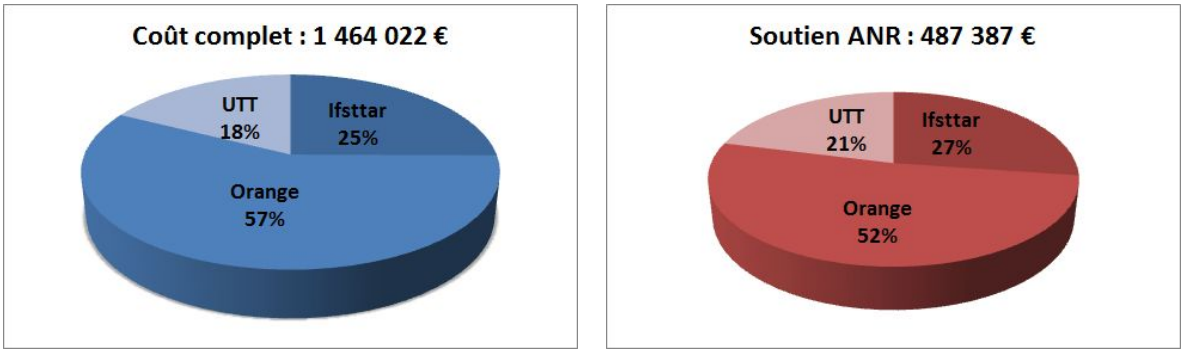

Figure 3 : Répartition du budget total et du soutien ANR entre les partenaires

4 Présentation du partenariat

4.1 Description, adéquation et complémentarité des partenaires

Le consortium du projet DISCRET est resserré et réunit trois établissements partenaires aux compétences très complémentaires, indispensables pour réaliser le projet : (a) l'entreprise Orange, l'un des premiers opérateurs mondiaux de téléphonie mobile et plus particulièrement les équipes de deux

de ses départements de recherche & développement (DI4B, SENSE), (b) l'Ifsttar, établissement public scientifique et technique, référence en matière de recherche sur les transports, l'aménagement et la gestion des risques (équipes LEE, LICIT), et (c) l'équipe Modélisation et Sûreté des Systèmes (LM2S) de l'Institut Charles Delaunay (ICD) de l'UTT, spécialisée dans l'analyse de données de capteurs généralisés, notamment issues des réseaux sociaux. Certains des chercheurs d'Orange et de l'Ifsttar ont déjà travaillé ensemble dans des projets de recherche (ANR ABCD et ANR Cancan). Cette connaissance mutuelle facilitera la collaboration.

L'analyse des données massives (Big Data) offre la possibilité de capturer et d'étudier des évolutions spatio-temporelles à très fine échelle dans les espaces urbains, de mettre en évidence des événements singuliers, des corrélations et des singularités, impossibles à détecter au travers d'un échantillonnage occasionnel. Conduire ce type d'études, nécessite une synergie multidisciplinaire afin de pouvoir 1) échantillonner et traiter de très grandes masses de données; 2) proposer de nouvelles approches, fondées sur le big data et la fusion de données de sources multiples, pour l'interprétation et la modélisation des données; 3) croiser avec d'autres sources d'informations afin de proposer des interprétations aux phénomènes mis en évidence; 4) élaborer des méthodes d'apprentissage et de traitement en temps réel de ce type de données. Le consortium du projet DISCRET rassemble un panel de compétences indispensables et complémentaires : télécommunications (DI4B), programmation informatique et gestion de bases de données (DI4B, LM2S, LICIT, SENSE), ingénierie des systèmes et réseaux (DI4B), fouille de données (DI4B, LM2S, LICIT, SENSE), statistiques, Big data, intelligence artificielle et machine learning (DI4B, LM2S, LICIT, LEE), sciences sociales appliquées aux dynamiques urbaines (SENSE), prévision et gestion des risques (LEE, LM2S).

La répartition des rôles des partenaires dans les lots et tâches vise à maximiser les échanges entre plusieurs communautés de chercheurs pour donner une véritable dimension interdisciplinaire au projet. Les responsables scientifiques sont identifiés dans le tableau 1.

4.2 Qualification du coordinateur du projet

Eric Gaume (Ifsttar), 50 ans, ingénieur général des Ponts des Eaux et des Forêt, docteur, HDR, dirige depuis 2013 le département Géotechnique, Environnement, Risques naturels et Sciences de la Terre (GERS) de l'Ifsttar auquel est rattaché le laboratoire LEE. Chercheur sénior de renommée internationale dans son domaine, il est l'auteur de près de 60 articles scientifiques (h-index du WoS de 26) et co-auteurs de plusieurs ouvrages dont « Hydrologie quantitative » (ed. Springer) [55], prix Roberval du meilleur ouvrage scientifique de langue française en 2013. Eric Gaume est professeur, responsable du cours d'hydrologie de l'Ecole des Ponts ParisTech. Ses activités de recherche portent sur l'amélioration des modèles de prévision et d'alerte pour les phénomènes hydro-météorologiques extrêmes et en particulier les crues soudaines [56] et le développement des méthodes statistiques bayésiennes pour l'étude des extrêmes [57]. Il est co-auteur de la partie bayésienne du package nsRFA du logiciel R. Il promeut actuellement, en concertation avec la Direction générale de la prévention de risques du MTES, l'intégration et la mobilisation de d'informations de sources variées – dont les données de téléphonie mobile font partie, en complément des prévisions pour une meilleure prise de décision lors des événements critiques [58]. Il est à l'initiative du projet DISCRET.

Eric Gaume a coordonné quatre projets nationaux dont le projet ANR Prédiflood (2009-2012) et des tâches des projets Européens Floodsite (2004-2009), HYDRATE (2007-2010), FIMFRAME (2009-2011) et Hymex (2010-2020). Il est l'un des coordonnateurs du projet UrbaRiskLab (2018-2022, projet de l'Isite FUTURE). Il participe aussi au projet ANR PICS (2018-2021). Il a par ailleurs été responsable de la politique de prévention contre les incendies de forêt en Corse du Sud en début de carrière, fonction qui l'a amené à travailler en étroit partenariat avec les services de secours.

4.3 Qualification, rôle et implication des participants

4.3.1. Profil et compétence des participants

- **Orange** SA est un des plus grands opérateurs de télécommunications mondiaux. Orange consacre un budget important à la recherche et l'innovation avec près de 3200 personnes dont 700 personnes travaillent plus spécifiquement sur des objectifs de recherche industrielle. Le département SENSE (Sociology and Economics of Networks and Services) travaille sur la base d'un large spectre des méthodologies, allant des études qualitatives et observations ethnographiques, des enquêtes par questionnaire jusqu'à l'analyse des Big Data des réseaux de télécommunication. SENSE contribuera au projet notamment par la collecte et l'anonymisation des données mobiles ainsi que leur analyse : choix des critères et des approches pertinentes pour l'élaboration des signatures et la détection des anomalies (L2 et L3). L'équipe Data Intelligence for Business (DI4B) d'Orange Labs s'attache à transférer l'innovation en "services de la donnée" vers les unités d'affaires du groupe. DI4B contribuera au projet par l'analyse des données (méthodes d'analyse avancées : L3), la réalisation du prototype de visualisation et l'étude de faisabilité pour une plateforme d'information et d'aide à la prise de décision temps-réel, basée sur l'Intelligence Artificielle (L4).

- **Ifsttar** est un établissement public scientifique et technique, référence pour les recherches portant sur les transports, l'aménagement et les risques, notamment les risques urbains. Il est issu de la fusion en 2011 du Laboratoire central des ponts et chaussées (LCPC) et de l'Institut national de recherche sur les transports et leur sécurité (INRETS). DISCRET implique deux laboratoires de l'Ifsttar. Le LICIT, unité mixte de recherche de l'Ifsttar et de l'ENTPE (UMR T_9401), qui apportera ses compétences à la croisée de l'analyse des données et de la modélisation des comportements urbains (L2). Les chercheurs impliqués apporteront l'expertise sur la qualification de données, la fusion de données et la caractéristisation des activités urbaines et la définition des signatures (L3). Le laboratoire Eau & Environnement (LEE), mène des recherches sur la prévision des phénomènes hydro-météorologiques extrêmes et l'aide à la gestion de crise. Il apportera ses compétences en matière de développement de plateformes d'information en temps réel (projet ANR PICS) et de retours d'expérience post-événements (T2.3).

- **Université Technologique de Troyes (UTT)** est un établissement public à caractère scientifique, culturel et professionnel (EPSCP). Les domaines de recherche du laboratoire ICD de l'UTT couvrent des approches essentiellement statistiques et probabilistes qui s'appuient sur un large ensemble d'hypothèses liées à la nature des informations disponibles a priori dans le système observé, notamment dans le cas des données incomplètes ou manquantes (approche abordée dans la Géo-inférence de Tweets, T3.2). L'équipe LM2S s'intègre pleinement à la thématique transverse « Sciences et Technologies pour la Maîtrise des Risques » de l'UTT pour la reconnaissance des formes et des situations. Elle combine dans le cadre de ces projets les méthodes à noyau, le "time-frequency machine learning" pour la reconnaissance et l'anticipation de situations évolutives. Une partie de ses membres se spécialisent dans l'analyse de données de capteurs technologiques (vidéosurveillance) et de capteur humains via des canaux tels que les réseaux sociaux et leurs applications associées. L'équipe héberge plusieurs plateformes pour la surveillance et la gestion de crise: CapSec- capteurs pour la sécurité, TweetCapt®- collecte large spectre de (méta)données de Twitter, EventAlert®- remontée citoyenne d'alertes pour le monitoring des situations.

4.3.2. Activités des participants

Les coordinations de lots et tâches sont indiquées en gras.

Partenaire	Nom	Prénom	Emploi actuel	p.mois	Rôle/responsabilité dans le projet (4 lignes max)
Ifsttar-LEE	GAUME	Eric	Chercheur senior	7	**Coordonnateur du projet, Lot 1 (T1.1, T1.2, T1.3)** Contributeur Lot 1, 2 et 4 (T2.3, T4.1)
Ifsttar-LICIT	EL FAOUZI	Nour-Eddin	Directeur de recherche	6	Contributeur Lots 2 et 3 (T2.2, **T3.1**, T3.3) **Coordination T3.1**
Ifsttar-LICIT	FURNO	Angelo	Chargé de recherche	3	Contributeur Lots 2 et 3 (T2.1, **T2.2**, T3.1) **Coordination T2.2**
Ifsttar-LICIT	FIORE	Marco	Chercheur associé	**	Contributeur Lots 2 et 3 (T2.1, T2.2, T3.1)
Ifsttar-LEE	LEBOUC	Laurent	Technicien supérieur	4	Contributeur Lot 2 (**T2.3**) **Coordination T2.3**
UTT-ICD	BIRREGAH	Babiga	Enseignant - chercheur	12	**Coordonnateur du Lot 3 (T3.2)** Contributeur Lot 1, 3, 4 .(T1.1.,**T3.2**, T3.3, T4.2)
UTT-ICD	GUEPIE	Blaise Kevin	Enseignant - chercheur	4	Contributeur Lot 3 (T3.2, T3.3)
UTT-ICD	DUHAMEL	Andrea	Enseignant-chercheur	4	Contributeur Lot 3 (T3.2, T3.3)
UTT-ICD	SNOUSSI	Hichem	Professeur	4	Contributeur Lot 3 (T3.2, T3.3)
UTT-ICD	OUARET	Rachid	Ingénieur de recherche post-doctoral	1	Contributeur Lot 3 (T3.2, T3.3)
Orange SA	TOSIC	Tamara	Data Scientist	13	**Coordonnatrice du Lot 4 (T3.3, T4.2)** Contributrice de Lots 1, 3, 4 (T1.1, T1.3, **T3.3**, **T4.2**)
Orange SA	HARO	Herve	Ingénieur R&D	8,5	Contributeur de Lot 3 (T3.1, T3.3)
Orange SA	SIBILEAU	Natalie	Ingénieur R&D	8,5	Contributrice de Lots 3 et 4 (T3.1, T4.1)
Orange SA	BERNHARD	Raphael	Ingénieur R&D	6	Contributeur de Lot 4 (**T4.1**, T4.2) **Coordination T4.1**
Orange SA	SMOREDA	Zbigniew	Chercheur senior	6	**Coordonnateur du Lot 2 (T2.1)** Contributeur Lots 2 et 3 (**T2.1**, T2.2, T3.1)
Orange SA	RUBRICHI	Stefania	Chercheuse	8	Contributrice Lots 2, 3, 4 (T2.1, 2.2, 3.1, 4.1)
Orange SA	BOUCHOIR	Bruno	Ingénieur R&D	8	Contributeur Lot 2, 3, 4 (T2.1, 3.3, 4.1)
Orange SA	ZIEMLICKI	Cezary	Ingénieur R&D	8	Contributeur Lots 2 et 4 (**2.1**, 2.2, 4.1) **Coordination T2.1**
Ifsttar- LICIT	à recruter		Post-doc	18	Développement des méthodologies de détection des anomalies - analyse des données (Lots 2 et 3, T2.2,T3.1)
UTT-ICD	à recruter		Post-doc	12	Traitement des données des réseaux sociaux, extraction des informations contextuelles (Lot 3, T 3.2)
UTT-ICD	service interne de l'UTT		Ingénieur de recherche	6	Paramétrage- Admin. Bases de données et plateforme (Lot 3, T3.2)

*** Marco Fiore, chercheur du CNR (Turin, Italie), apportera son expertise au projet en tant que membre associé du partenaire Ifsttar (T2.2 et T3.1). Son temps n'est pas comptabilisé pour des raisons d'éligibilité et de simplicité administrative.*

4.4 Éléments financiers des partenaires entreprises

Année	CA (€)	Subventions d'exploitation (€)	EBE (€)	Capitaux propres (€)	Disponibilités (VMP + disponibilités) (€)	Dettes auprès des établissemen ts de crédit (€)
n*	41 381 000 000	4 829 000 000		30 669 000 000		25 441 000 000
n-1	40 859 000 000	4 778 000 000		30 975 000 000		23 843 000 000
n-2	40 708 000 000	3 917 000 000		31 241 000 000		24 444 000 000

Plan de financement :

Année	Apport en capital (€)	Apport en compte courant (€)	Emprunt (€)	Autofinancement (€)	Subvention (dont aide ANR) (€)	Autres (préciser) (€)
n+1*	0	0	0	391 771,94	167 902,26	0
n+2*	0	0	0	195 885, 97	83 951,13	0

*prévisions

Annexes

Références bibliographiques

Les noms des participants au projet apparaissent **en gras**

[1] P. Gendre and G. Ostyn, "Système de recueil d'information trafic via les réseaux téléphoniques cellulaires: opportunité et faisabilité," Note analytique, *CETE Méditerranée*, 2006.

[2] J.-L. Ygnace et al., "Travel time/speed estimates on the french rhone corridor network using cellular phones as prob Final report of the SERTI V program, INRETS, Lyon, France, 2001.

[3] J.-L. Ygnace and C. Drane, "Cellular telecommunication and transportation convergence: a case study of a research conducted in California and in france on cellular positioning techniques and transportation issues," in ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 01TH8585), pp. 16–22, IEEE, 2001.

[4] P. Bonnel, É. Hombourger, A.-M. Olteanu-Raimond, and **Z. Smoreda**, "Apports et limites des données passives de la téléphonie mobile pour la construction de matrices origine-destination," *Revue d'Economie Régionale Urbaine*, no. 4, pp. 647–672, 2017.

[5] J. P. Bagrow, D. Wang, and A.-L. Barabasi, "Collective response of human populations to large-scale emergencies," *PloS one*, vol. 6, no. 3, e17680, 2011.

[6] A. Dobra, N. E. Williams, and N. Eagle, "Spatiotemporal detection of unusual human population behavior using mobile phone data," *PloS one*, vol. 10, no. 3, e0120449, 2015.

[7] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, p. 779, 2008.

[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[9] C. M. Schneider, V. Belik, T. Couronné, **Z. Smoreda**, and M. C. González, "Unravelling daily human mobility motifs," *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20130246, 2013.

[10] **A. Furno**, D. Naboulsi, R. Stanica, and **M. Fiore**, "Mobile demand profiling for cellular cognitive networking," *IEEE Transactions on Mobile Computing*, vol. 16, no. 3, pp. 772–786, 2016.

[11] **A. Furno**, **M. Fiore**, R. Stanica, **C. Ziemlicki**, and **Z. Smoreda**, "A tale of ten cities: Characterizing signatures of mobile traffic in urban areas," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2682– 2696, 2016.

[12] **A. Furno**, **M. Fiore**, and R. Stanica, "Joint spatial and temporal classification of mobile traffic demands," in IEEE INFOCOM 2017-IEEE Conference on Computer Communications, pp. 1–9, IEEE, 2017.

[13] **A. Furno**, **N.-E. El Faouzi**, **M. Fiore**, and R. Stanica, "Fusing gps probe and mobile phone data for enhanced land-use detection," in 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 693–698, IEEE, 2017.

[14] A. Vespignani, "Predicting the behavior of techno-social systems," Science, vol. 325, no. 5939, pp. 425–428, 2009.

[15] S. Eubank, H. Guclu, V. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, no. 6988, p. 180, 2004.

[16] M. Irfan, L. Marcenaro, and L. Tokarchuk, "Crowd analysis using visual and non-visual sensors, a survey," in 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1249–1254, IEEE, 2016.

[17] Z. Xu, L. Mei, K.-K. R. Choo, Z. Lv, C. Hu, X. Luo, and Y. Liu, "Mobile crowd sensing of human-like intelligence using social sensors: a survey," Neurocomputing, vol. 279, pp. 3–10, 2018.

[18] B. Guo, Z. Yu, X. Zhou, and D. Zhang, "From participatory sensing to mobile crowd sensing," in 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS), pp. 593–598, IEEE, 2014.

[19] H. Ma, D. Zhao, and P. Yuan, "Opportunities in mobile crowd sensing," IEEE Communications Magazine, vol. 52, no. 8, pp. 29–35, 2014.

[20] B. Guo, C. Chen, D. Zhang, Z. Yu, and A. Chin, "Mobile crowd sensing and computing: when participatory sensing meets participatory social media," IEEE Communications Magazine, vol. 54, no. 2, pp. 131–137, 2016.

[21] X. Lu, L. Bengtsson, and P. Holme, "Predictability of population displacement after the 2010 haiti earthquake," Proceedings of the National Academy of Sciences, vol. 109, no. 29, pp. 11576–11581, 2012.

[22] T. Yabe and S. V. Ukkusuri, "Integrating information from heterogeneous networks on social media to predict post-disaster returning behavior," Journal of Computational Science, vol. 32, pp. 12–20, 2019.

[23] T. Yabe., Y. Sekimoto, K. Tsubouchi, and S. Ikemoto, "Cross-comparative analysis of evacuation behavior after earthquakes using mobile phone data," PLoS One vol. 14 no. 2, e0211375, 2019.

[24] L. Hong, M. Lee, A. Mashhadi, and V. Frias-Martinez, "Towards understanding communication behavior changes during floods using cell phone data," in International Conference on Social Informatics, pp. 97– 107, Springer, 2018.

[25] O. Oh, M. Agrawal, and H. R. Rao, "Information control and terrorism: Tracking the mumbai terrorist attack through twitter," Information Systems Frontiers, vol. 13, no. 1, pp. 33–43, 2011.

[26] D. Naboulsi, **M. Fiore**, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: a survey," IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 124–161, 2015.

[27] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. Von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti," PLoS medicine, vol. 8, no. 8, p. e1001083, 2011.

[28] D. Pastor-Escuredo, A. Morales-Guzmán, Y. Torres-Fernández, J.-M. Bauer, A. Wadhwa, C. Castro-Correa, L. Romanoff, J. G. Lee, A. Rutherford, V. Frias-Martinez, et al., "Flooding through the lens of mobile phone activity," in IEEE Global Humanitarian Technology Conference (GHTC 2014), pp. 279–286, IEEE, 2014.

[29] H. T. Marques-Neto, F. H. Xavier, W. Z. Xavier, C. H. S. Malab, A. Ziviani, L. M. Silveira, and J. M. Almeida, "Understanding human mobility and workload dynamics due to different large-scale events using mobile phone data," Journal of Network and Systems Management, pp. 1–22, 2018.

[30] Y. Dong, F. Pinelli, Y. Gkoufas, Z. Nabi, F. Calabrese, and N. V. Chawla, "Inferring unusual crowd events from mobile phone call detail records," in Joint European conference on machine learning and knowledge discovery in databases, pp. 474–492, Springer, 2015.

[31] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," IEEE Transactions on Knowledge and data Engineering, vol. 26, no. 9, pp. 2250–2267, 2013.

[32] Y.-A. de Montjoye, S. Gambs, V. Blondel, G. Canright, N. de Cordes, S. Deletaille, K. Engø-Monsen, M. Garcia-Herranz, J. Kendall, C. Kerry, **Z. Smoreda**, et al., "On the privacy-conscientious use of mobile phone data," Scientific data, vol. 5, 2018.

[33] P. Earle, M. Guy, R. Buckmaster, C. Ostrum, S. Horvath, and A. Vaughan, "Omg earthquake! can twitter improve earthquake response?," Seismological Research Letters, vol. 81, no. 2, pp. 246–251, 2010.

[34] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, "# earthquake: Twitter as a distributed sensor system," Transactions in GIS, vol. 17, no. 1, pp. 124–147, 2013.

[35] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in Proceedings of the 19th international conference on World wide web, pp. 851–860, ACM, 2010.

[36] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," IEEE Transactions on Knowledge and data engineering, vol. 25, no. 4, pp. 919–931, 2012.

[37] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1079–1088, ACM, 2010.

[38] X. Lu and C. Brelsford, "Network structure and community evolution on twitter: human behavior change in response to the 2011 japanese earthquake and tsunami," Scientific reports, vol. 4, p. 6773, 2014.

[39] M. F. Goodchild and J. A. Glennon, "Crowdsourcing geographic information for disaster response: a research frontier," International Journal of Digital Earth, vol. 3, no. 3, pp. 231–241, 2010.

[40] B. Birregah, T. Top T., Perez C. et al. "Multi-layer crisis mapping: a social media-based approach", 21st IEEE International Workshop on Enabling Technologies - Infrastructure for Collaborative Enterprises (WETICE) Toulouse, France, 2012.

[41] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, **Z. Smoreda**, and **C. Ziemlicki**, "Data for development: the d4d challenge on mobile phone data," arXiv preprint arXiv:1210.0137, 2012.

[42] K. O. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, and R. Jenssen, "Time series cluster kernel for learning similarities between multivariate time series with missing data," Pattern Recognition, vol. 76, pp. 569–581, 2018.

[43] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," IEEE Transactions on Signal Processing, vol. 65, no. 13, pp. 3551–3582, 2017.

[44] **E. Gaume** and M. Borga, "Post-flood field investigations in upland catchments after major flash floods: proposal of a methodology and illustrations," Journal of flood risk management, vol. 1, no. 4, pp. 175– 189, 2008.

[45] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning," arXiv preprint arXiv:1809.07857, 2018.

[46] N. di Pietro, M. Merluzzi, E. C. Strinati, and S. Barbarossa, "Resilient design of 5g mobile-edge computing over intermittent mmwave links," arXiv preprint arXiv:1901.01894, 2019.

[47] S. Kekki, W. Featherstone, Y. Fang, P. Kuure, A. Li, A. Ranjan, D. Purkayastha, F. Jiangping, D. Frydman, G. Verin, et al., "Mec in 5g networks," ETSI White Paper, ISBN No. 979-10-92620-22-1, vol. 28, 2018.

[48] **R. Ouaret**, **B. Birregah**, and O. Jaafor, "Spatial patterns of the french rail strikes from social networks using weighted k-nearest neighbour," International Journal of Social Network Mining, (in press), 2019.

[49] A. Ben-Aissa, **N.-E. El Faouzi**, and E. Lefevre, "Classification multisource via la théorie des fonctions de croyance: application à l'estimation du temps de parcours," Revue de Statistique Appliquée, p. 17p, 2009.

[50] **N.-E. El Faouzi**, "Fusion numérique d'informations multi-sources et extraction de connaissances: application à l'ingénierie du trafic," Revue des Nouvelles Technologies de l'Information (numéro spécial: Entreposage et fusion de données),, pp. 61–74, 2003.

[51] **N.-E. El Faouzi**, L. A. Klein, and O. De Mouzon, "Improving travel time estimates from inductive loop and toll collection data with dempster– shafer data fusion," Transportation research record, vol. 2129, no. 1, pp. 73–80, 2009.

[52] **N.-E. El Faouzi** and L. A. Klein, "Data fusion in intelligent traffic and transportation engineering: Recent advances and challenges," in Multisensor Data Fusion, pp. 586–617, CRC Press, 2015.

[53] **N.-E. El Faouzi** and L. A. Klein, "Data fusion for its: techniques and research needs," Transportation Research Procedia, vol. 15, pp. 495– 512, 2016.

[54] **N.-E. El Faouzi**, "Combining predictive schemes in short-term traffic forecasting," in 14th International Symposium on Transportation and Traffic Theory - Transportation Research Institute, 1999.

[55] P.A. Roche, J. Miquel., and **E. Gaume**, Hydrologie quantitative: Processus, modèles et aide à la décision. Springer ed., Amsterdam, Netherlands, 2012.

[56] W. Amponsah, P.-A. Ayral, B. Boudevillain, C. Bouvier, I. Braud, P. Brunet, G. Delrieu, J.-F. Didon-Lescot, **E. Gaume**, L. Lebouc, et al., "Integrated high-resolution dataset of high-intensity european and mediterranean flash floods," Earth System Science Data, 10(4), 1783–1794, 2018.

[57] **E. Gaume**, "Flood frequency analysis: The bayesian choice," Wiley Interdisciplinary Reviews: Water, 5(4),. e1290, 2018.

[58] G. L. Bihan, O. Payrastre, **E. Gaume**, D. Moncoulon, and F. Pons, "The challenge of forecasting impacts of flash floods: test of a simplified hydraulic approach and validation based on insurance claim data," Hydrology and Earth System Sciences, vol. 21, no. 11, pp. 5911–5928, 2017.