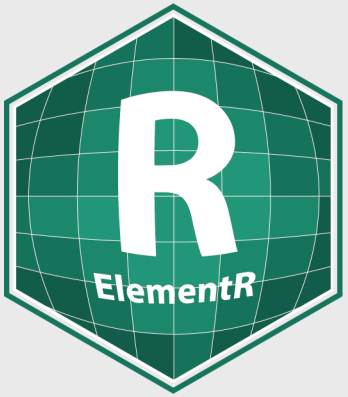


Introduction au Webscraping

Léa Christophe
Robin Cura
Hugues Pecout
Alexandre Cebeillac
Sébastien Rey-Coyrehourcq

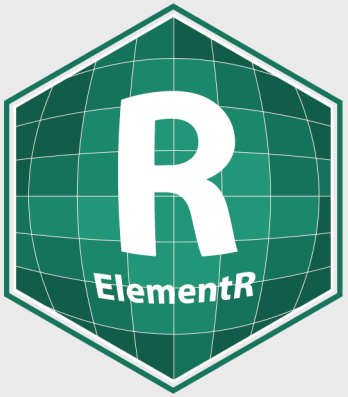
GT Webscraping

21 juin 2024

Géographie-cités
UMR 8504

idees
UMR6266 CNRS Rouen

PRODIG
UMR 8586 (CNRS)



Webscraping

1. Introduction (5 min)
2. Exemples d'application (10-15 min)
3. Mise en pratique

21 juin 2024

Introduction

Définition(s)

Web scraping, web harvesting, or web data extraction *is data scraping used for* ***extracting data from websites****. Web scraping* ***software*** *may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. While web scraping can be done manually by a software user,* ***the term typically refers to automated processes implemented using a bot or web crawler****.*

Wikipedia (en)

“ Le WebScraping est ***une pratique*** *qui nécessite l’articulation plus ou moins complexe de méthodes, de logiciels et d’infrastructures pour* ***naviguer et extraire du contenu*** *(des informations) d’un ou de plusieurs sites Web* ***de façon automatisé****.*

notre GT (wip)

Introduction

De multiples utilisations pour les SHS …

- Constitution de corpus …
  - **en // de corpus existants**
  - **originaux**

études comparatives

sauvegarde / patrimoine

mondialisé

- … faits de de sources et de nature de sources variés (son, images, textes)

- … collectés avec un
  - **flux continu**
  - **échantillonage**

suivi longitudinal

mixte quali/quanti

▶ Intérêt **croissant** au sein des laboratoires en SHS …

Introduction

... mais des verrous à différents niveaux

- **(I) légalité** de la :
  - collecte
  - stockage
  - de la (re)diffusion

  ➡ GT Webscraping (partie éthique & légalité)
- **(II) contrôle qualité**
  - validation
  - harmonisation

  ➡ trop problème dépendant
- **(III) techniques** lié à l'objet à capturer :
  - résilience
  - intégration
  - reproductibilité

  ➡ GT Webscraping (docker, sgbd, forge, etc.)

Exemple : https://journals.openedition.org/cybergeo/36478

Discuté ici (vidéo à venir) : https://atelierdatapaperadn.sciencesconf.org/

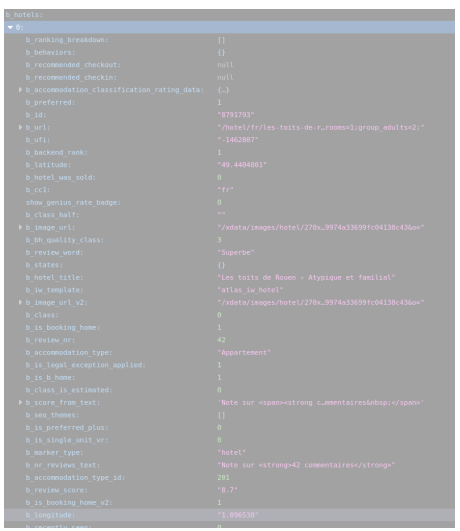
Légalité

une multiplicité de facteurs interdépendants

(... +/- loi du plus fort à l'international entre multinationales ...)

• (1) Pays	→ Jurisprudences & droit d'auteurs
• (2) CGU / Robots.txt	→ Loi Nationale & Européenne
• (3) Nature de l'information	→ Personnelles ? Sensibles ?
• (4) Visibilité de l'information	→ Publiques ? Privés ?
• (5) Qui récolte l'information	→ Institutions ? Chercheurs ? Entreprises ?
• (6) Volumétrie	→ Substansielle ? Non Substantielle ?
• (7) Finalité de la récolte	→ Lucratif / Non lucratif ? Avec ou Sans modifications ? Traitements, Aggrégation ?

Quelques exemples



Quelques exemples

Pratiques touristiques

Airbnb vs Booking.com

Flux multi-échelles

Données Avion FlightRadar24 & Bateaux MarineTraffic

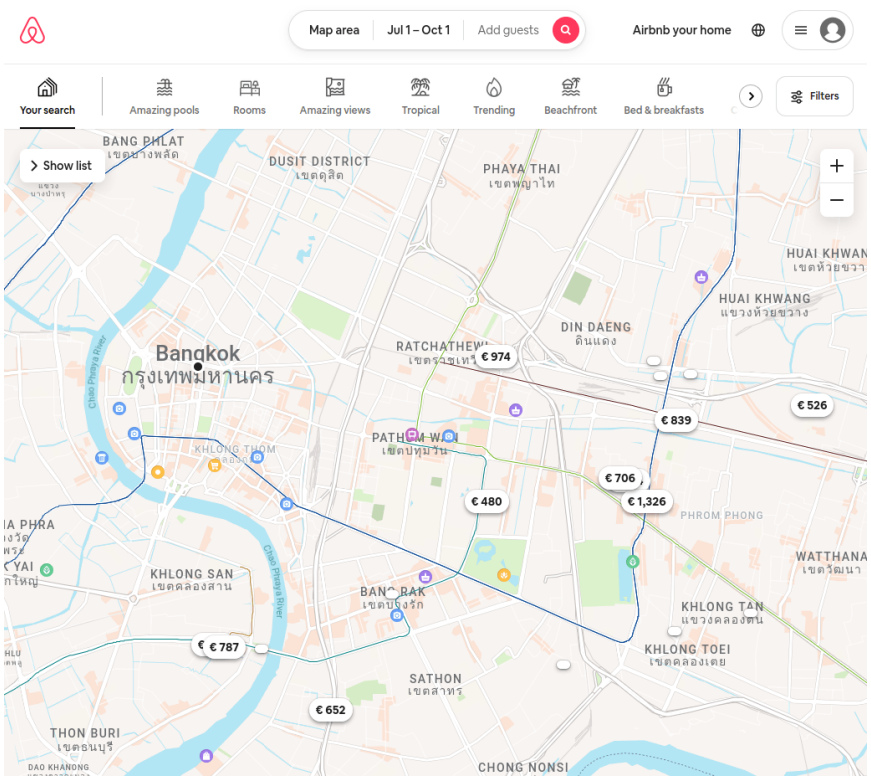
flightradar

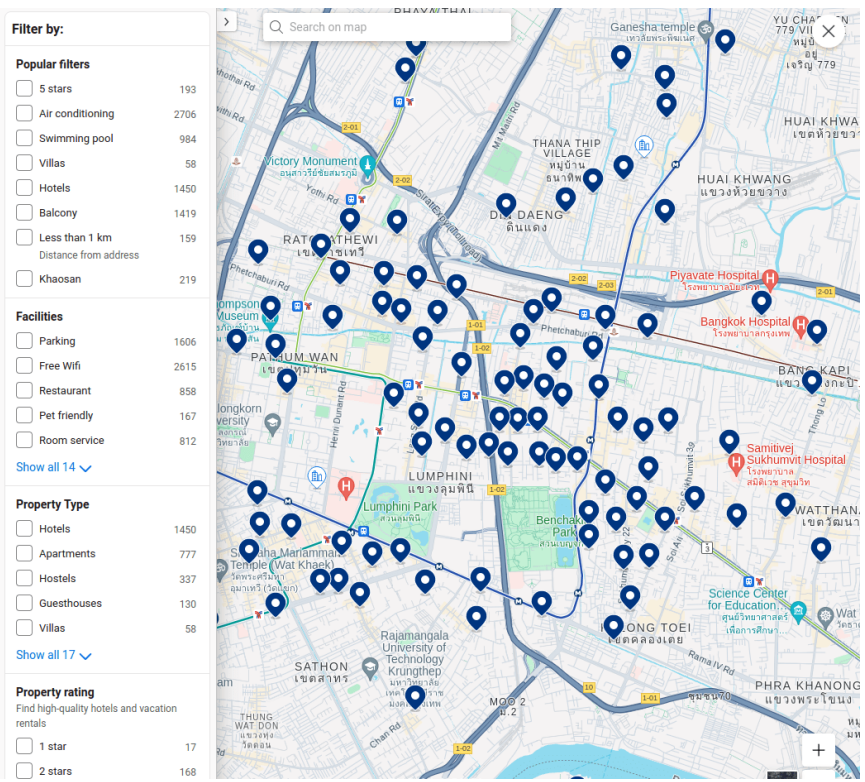
Mobilités quotidiennes & évacuation

Données trafic Bing & TomTom

Pratiques touristiques

Comparer les localisations des types d'offres dans les villes





Arctique & système monde

ANR PUR (2016 - 2020) / Yvette Vaguet

Intégration des zones arctiques dans le système monde

Un des objectifs :

Analyser temporellement les flux affectant les zones arctiques, dont les flux touristiques.

Pourquoi ?

"Tourism is a powerful global force that is turning the north into a Global North—tied into the Global South by facing the same global problem" (Veijola and Strauss-Mazzullo, 2019)

Comment ?

En créant des réseaux à partir des données les plus appropriées et accessibles

Arctique & système monde

PUR Pôles URbains

ANR PUR (2016 - 2020) / Yvette Vaguet

Intégration des zones arctiques dans le système monde

Un des objectifs :

Analyser temporellement les flux affectant les zones arctiques, dont les flux touristiques.

Pourquoi ?

"Tourism is a powerful global force that is turning the north into a Global North—tied into the Global South by facing the same global problem" (Veijola and Strauss-Mazzullo, 2019)

Comment ?

En créant des réseaux à partir des données les plus appropriées et accessibles

airbnb

Type	Privately held company
Industry	Lodging
Founded	August 2008; 11 years ago in San Francisco, California
Founders	Brian Chesky Joe Gebbia Nathan Blecharczyk

7M+ Airbnb listings worldwide

100K+ cities with Airbnb listings

191+ countries and regions with Airbnb listings

https://news.airbnb.com/fast-facts/

Airbnb

Collecter des données permettant de former un réseau de flux touristiques vers l'Arctique

Site web : airbnb.com
Langage : R
Framework : rvest / httr
Type de récolte : Campagne
Format : Json
Caractéristiques :

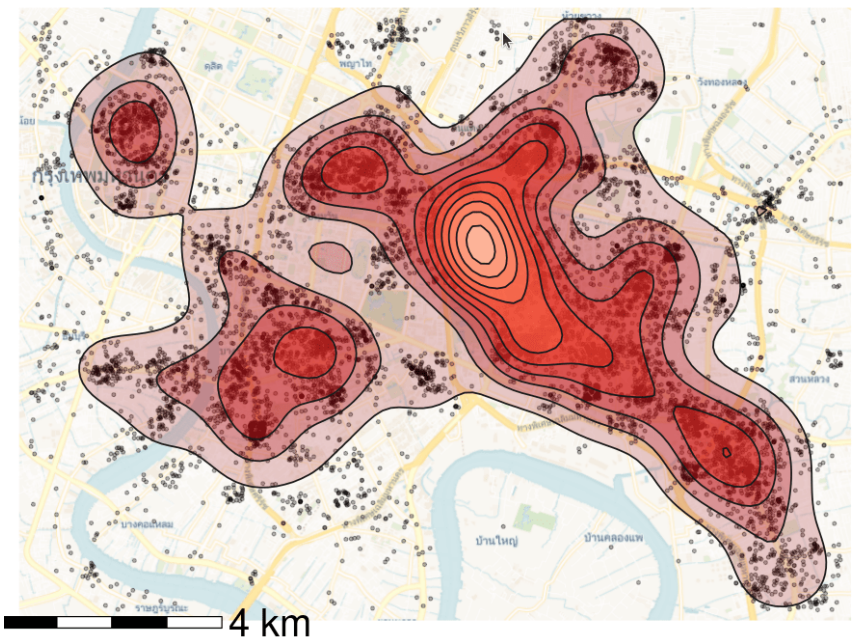
- Accessible via code source
- (+) Plusieurs API officieuses & pas de token privé
- (-) 80 requêtes / min / IP ==> temps de collecte long

- Des extractions de zones (insideairbnb) mais que dans certaines villes
- Des codes sont disponibles (Tom Slee) python, mais pas sous R

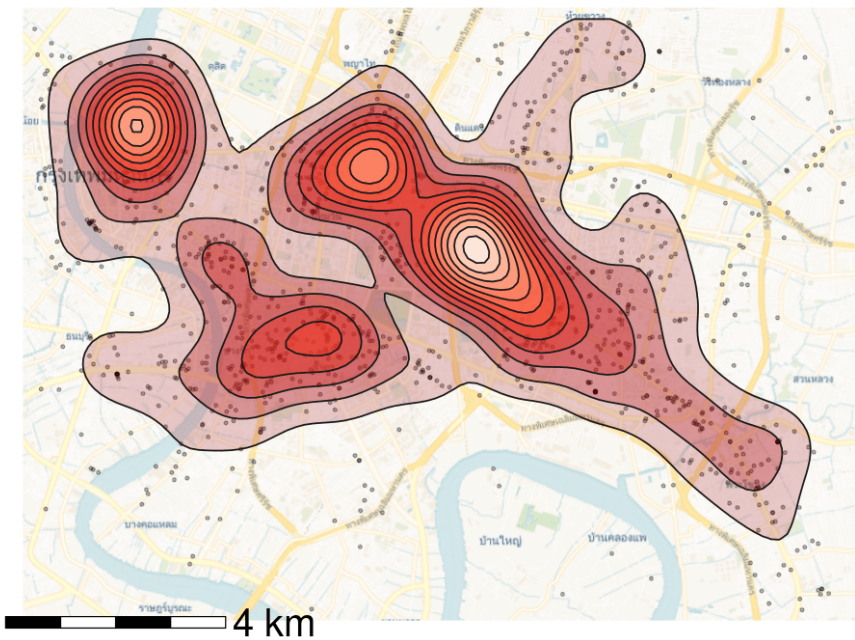
Pratiques touristiques

Comparer les localisations des types d'offres dans les villes

Airbnb



Booking



Localisation des hébergements à Bangkok

Arctique & système monde

PUR
Pôles URbains

ANR PUR (2016 - 2020) / Yvette Vaguet

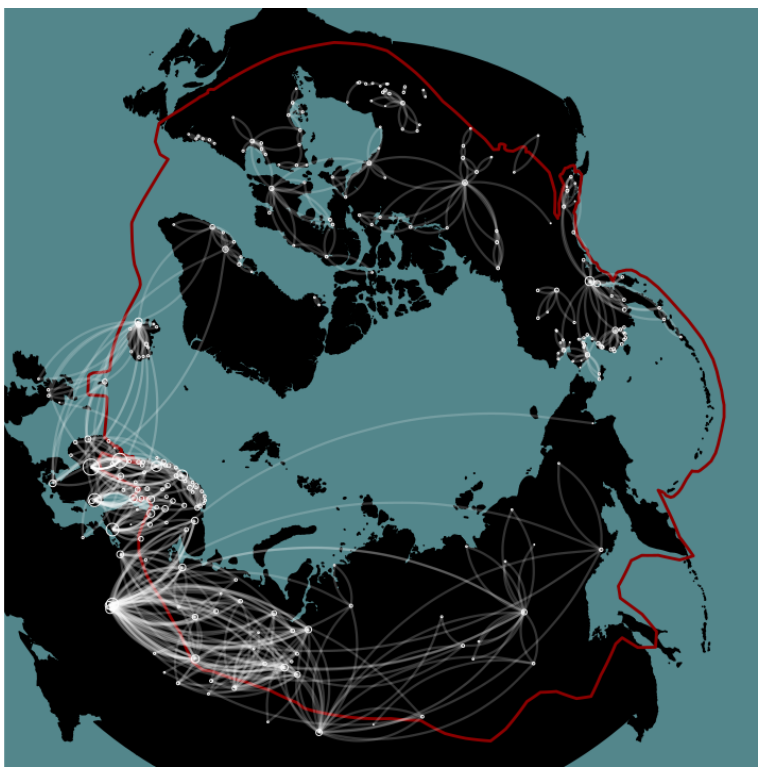
Intégration des zones urbaines arctiques dans le système monde

Un des objectifs :

Analyser temporellement les flux affectant les zones arctiques, notamment les flux aériens.

Comment ?

En créant des réseaux à partir des données les plus appropriées et accessibles

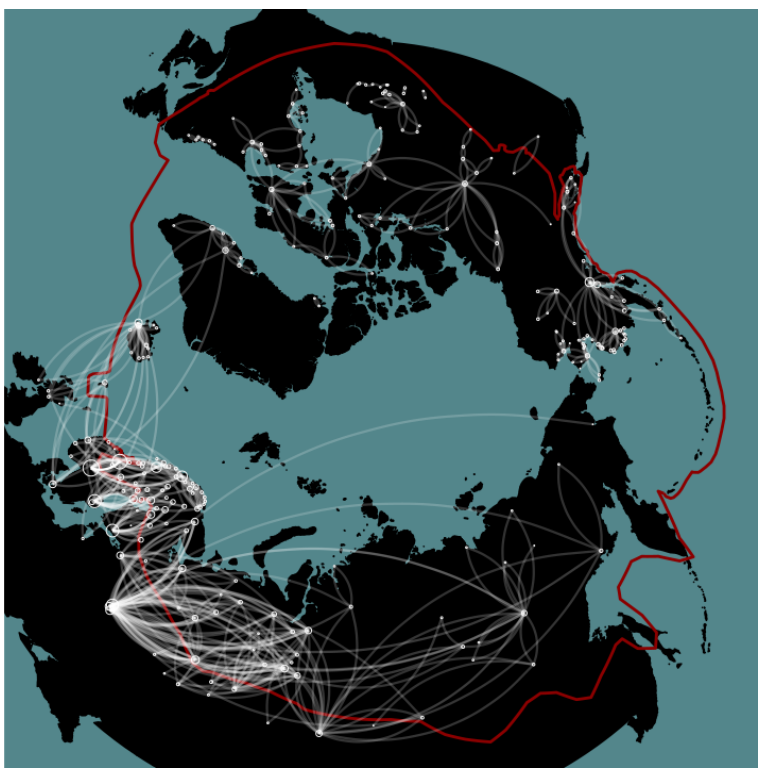


FlightRadar

flightradar 24

ANR PUR (2015 - 2019) / Yvette Vaguet

Récuperer les **départs et arrivées d'avions** pour les **aéroports mondiaux** pendant **plus d'un an**



FlightRadar

ANR PUR (2015 - 2019) / Yvette Vaguet

Récuperer les **départs et arrivées d'avions** pour les **aéroports mondiaux** pendant **plus d'un an**

Site web : flightradar24.com
Langage : Python
Framework : Scrapy

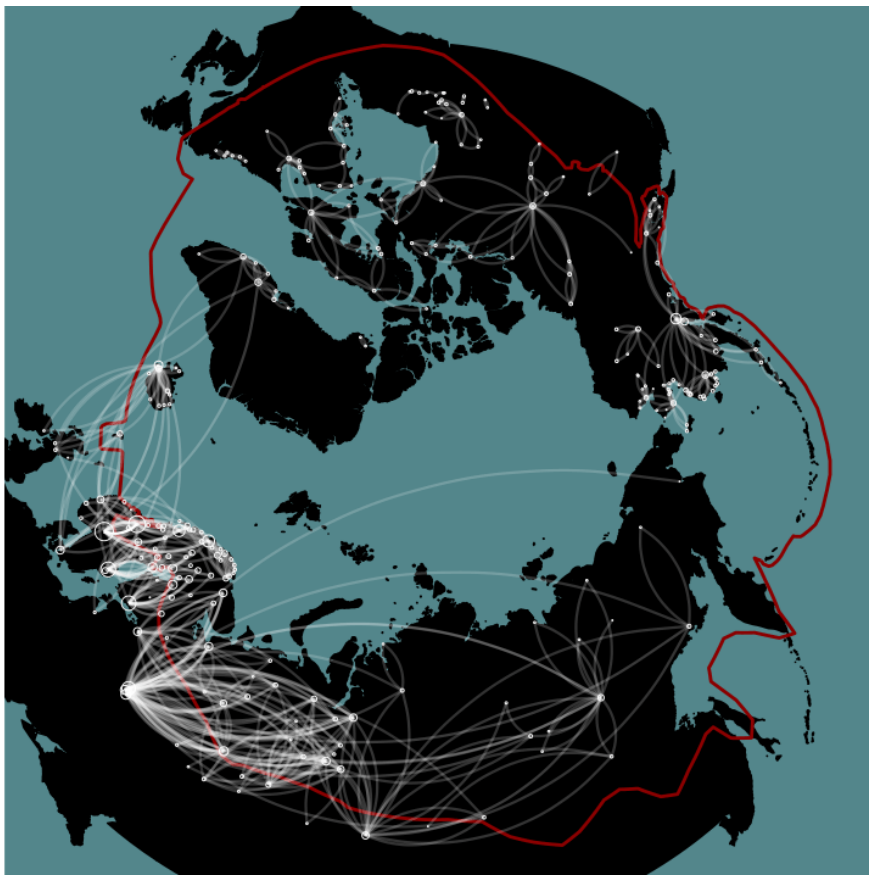
Free as in Freedom

https://github.com/IDEES-Rouen/Flight-Scrapping

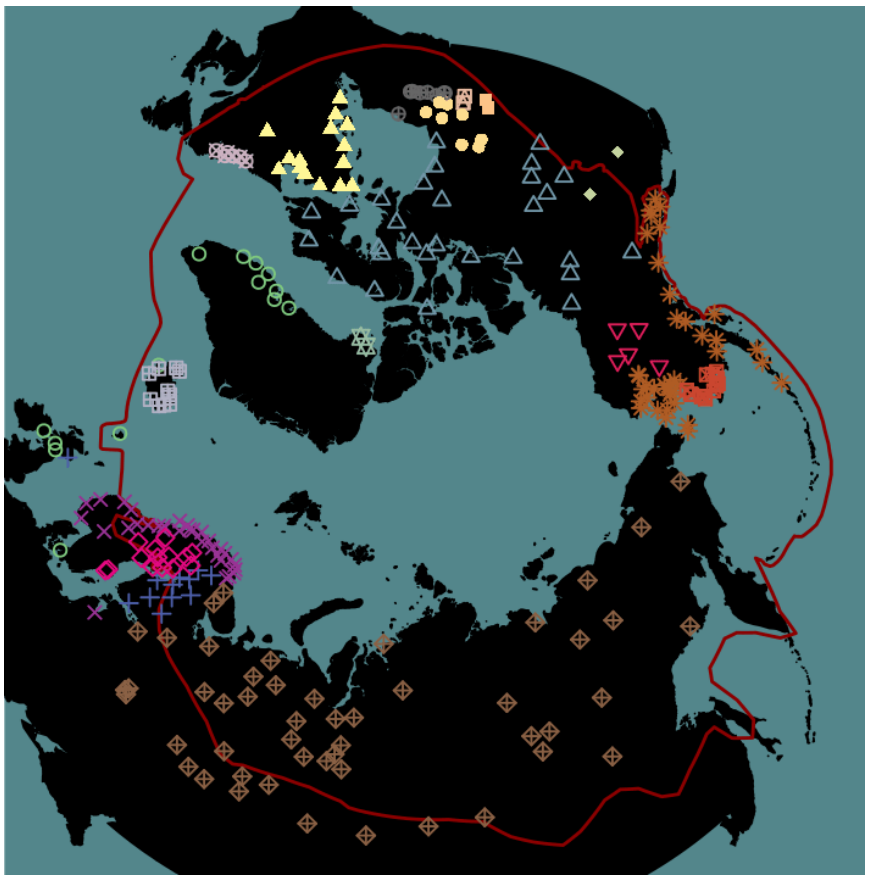
Type de récolte : Continue (depuis le **2018/06/10**)
Format : mongoDB / Json
Difficultés :

- architecture technique pour une récolte continue
- gestion des timezones
- intégration/consolidation des données SGBD (work in progress)

FlightRadar



Vol vers la zone arctique

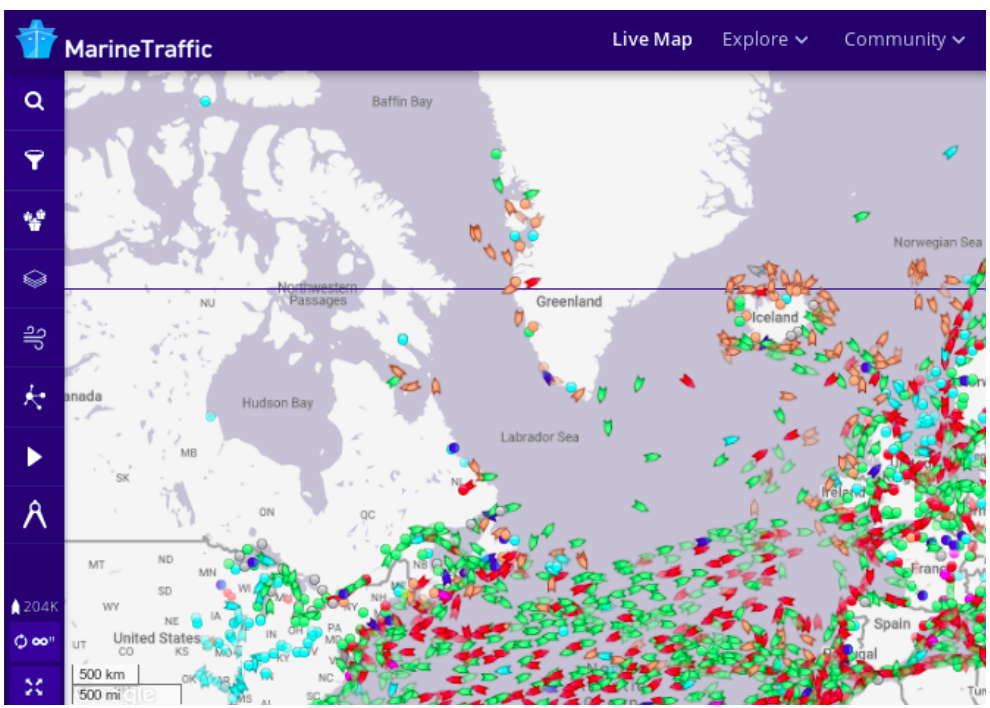


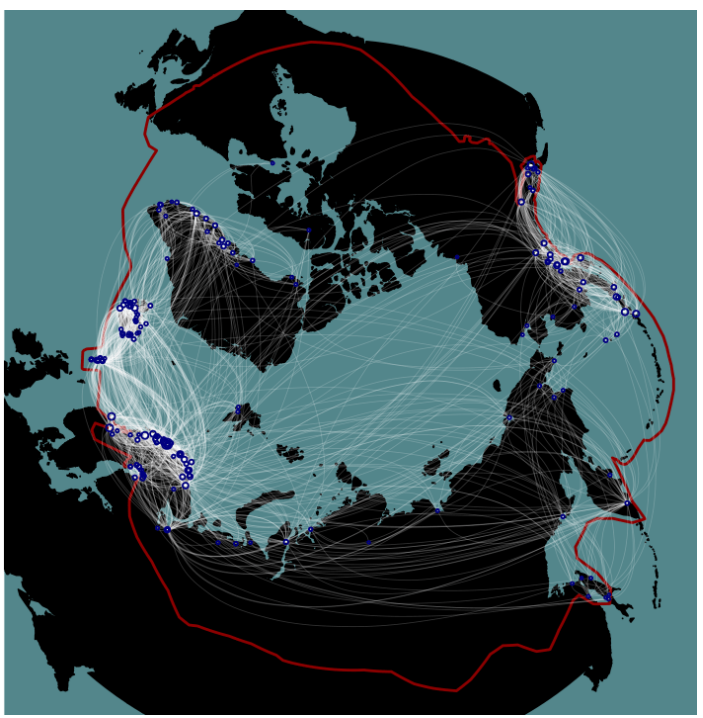
Détection de communauté

Marine Traffic

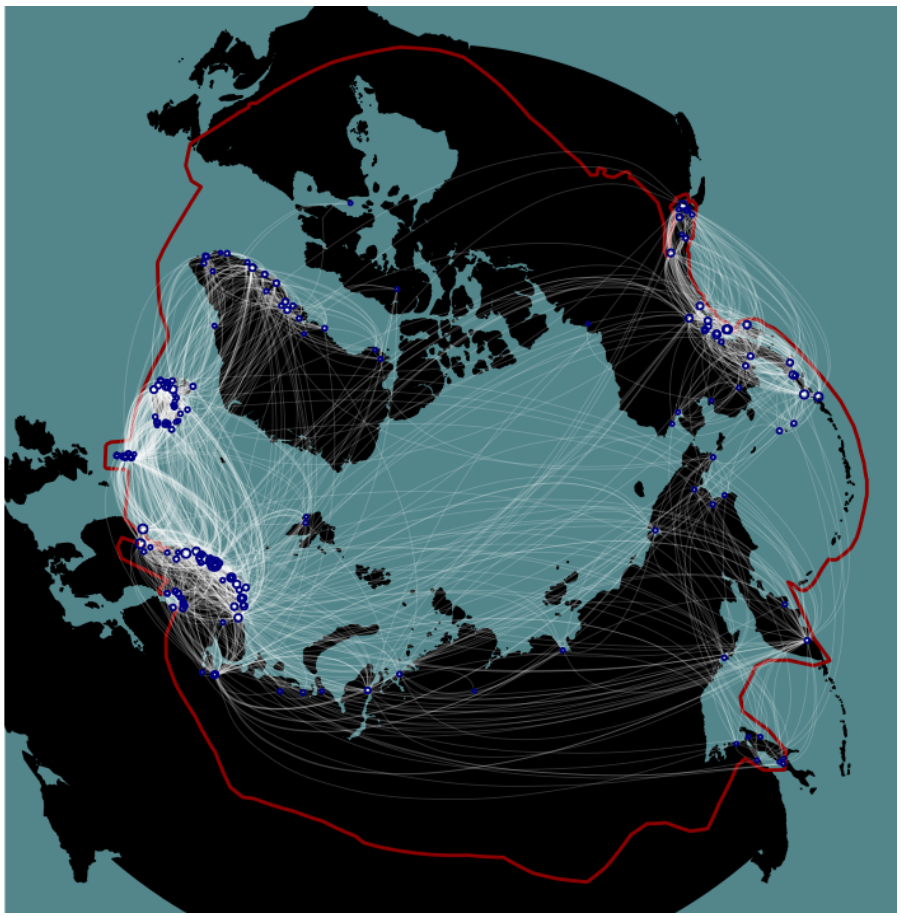
ANR PUR (2015 - 2019) / Yvette Vaguet

Récuperer les **départs et arrivées de bâteaux** partant où allant vers la zone arctique

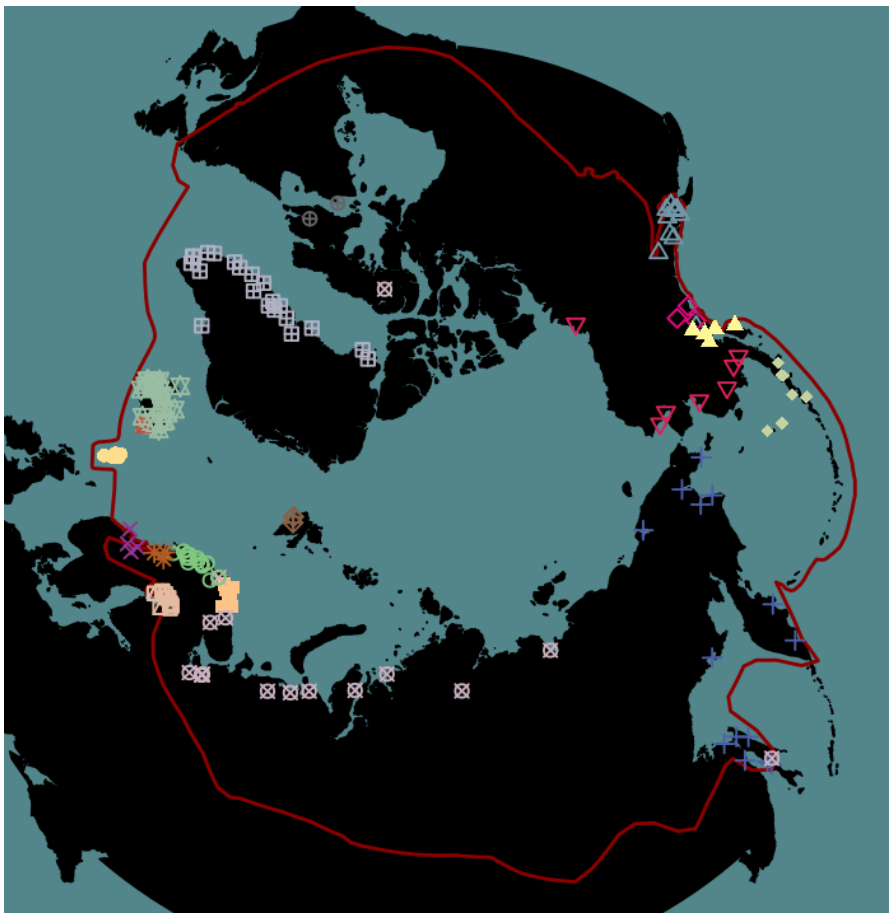




Marine Traffic



Flux maritime vers la zone arctique



Détection de communauté

Mobilités urbaines

FP7 DENFREE & ANR MO3
Mobilité quotidiennes & propagation des épidémies de dengue

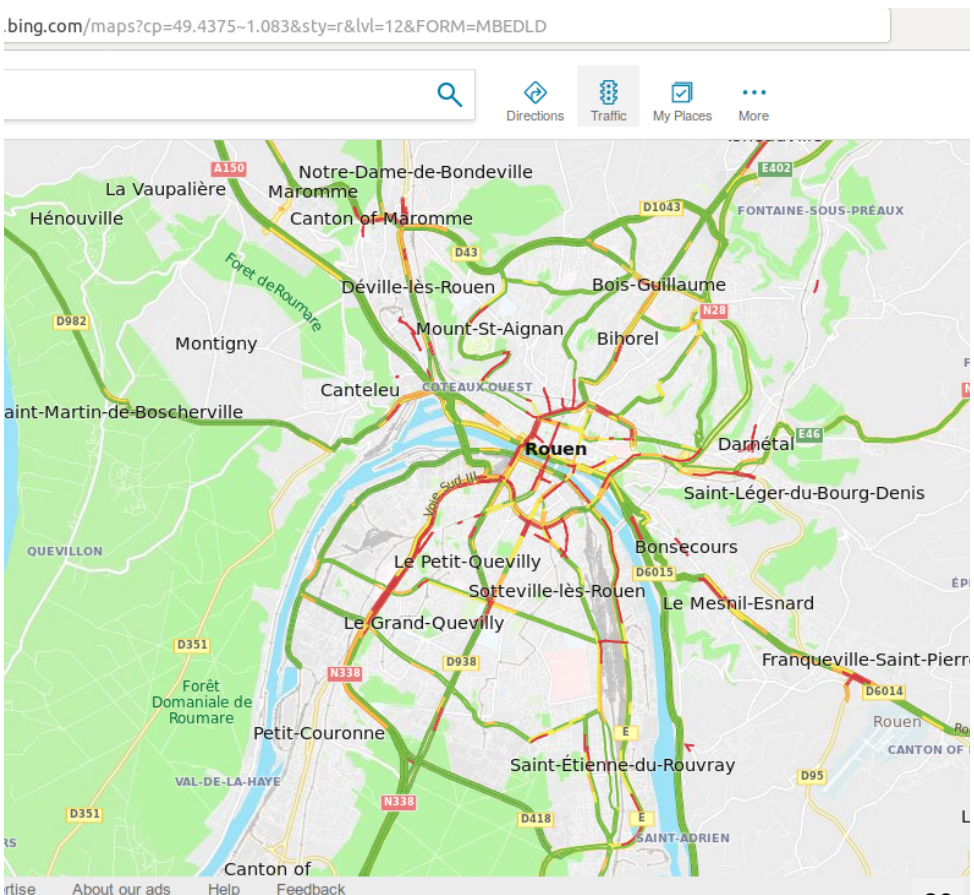
ANR Escape (2016 - 2024)
Modéliser les mobilités lors d'évacuation en cas d'urgence

Besoin de données sur les mobilités comme les conditions de trafic pour nourrir et/ou valider les modèles

Condition de trafic

Bing Map (Microsoft)

Collecte en temps réel des conditions de circulation



Condition de trafic

Bing Map (Microsoft)

Collecte en temps réel des conditions de circulation

Site web : /www.bing.com/maps

Langage : R

Type de récolte : Campagne

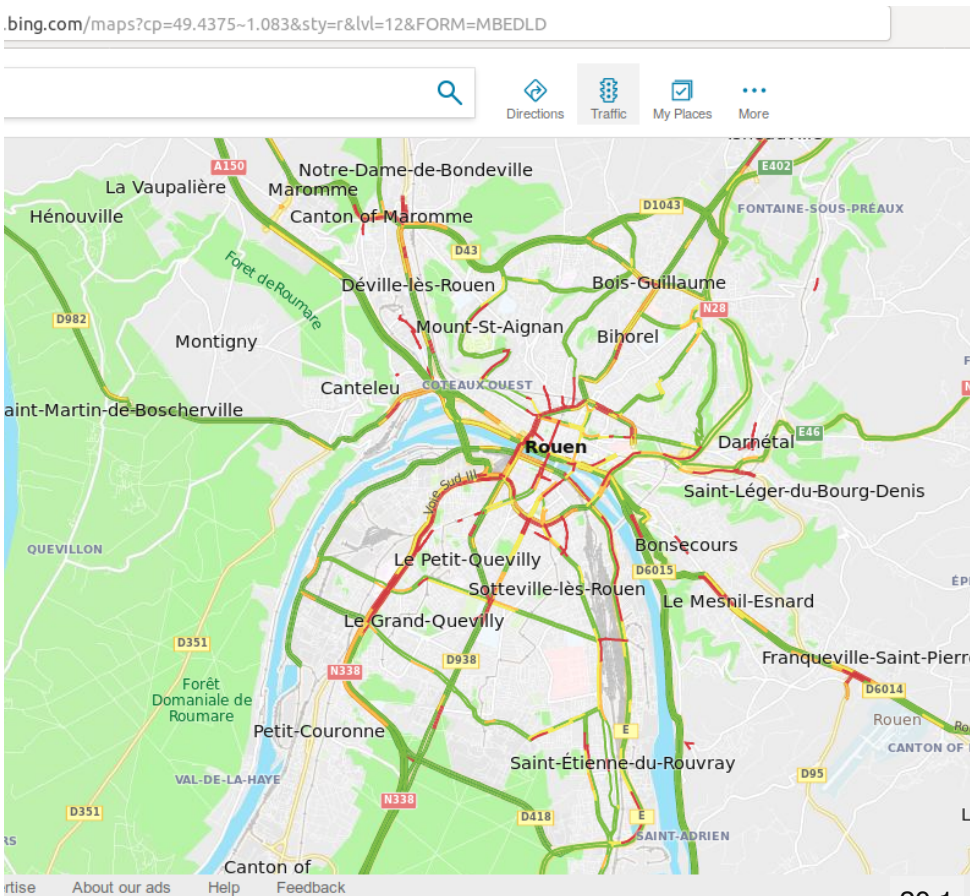
Format : Raster (image)

Avantages & limites :

- Trafic en temps réel
- Disponible dans 55 pays
- Méthodologie de Bing inconnue

Principe :

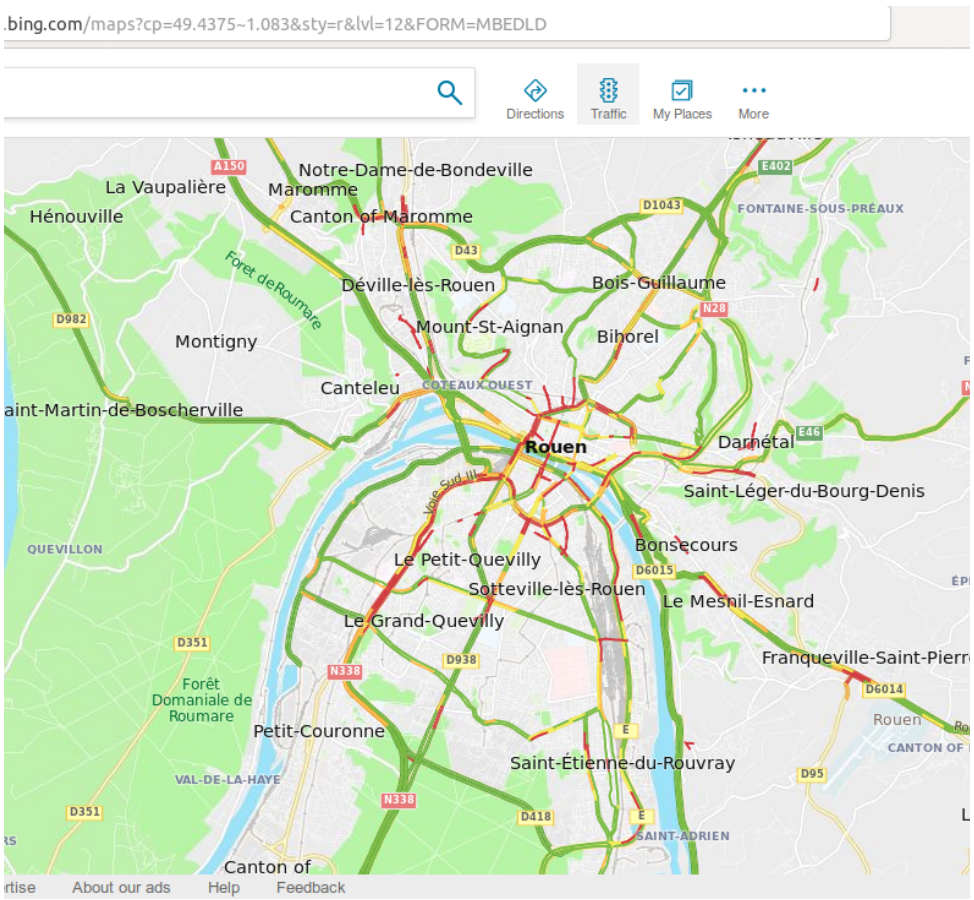
- Collecter et reprojecter des images
- Extraire les conditions de circulation



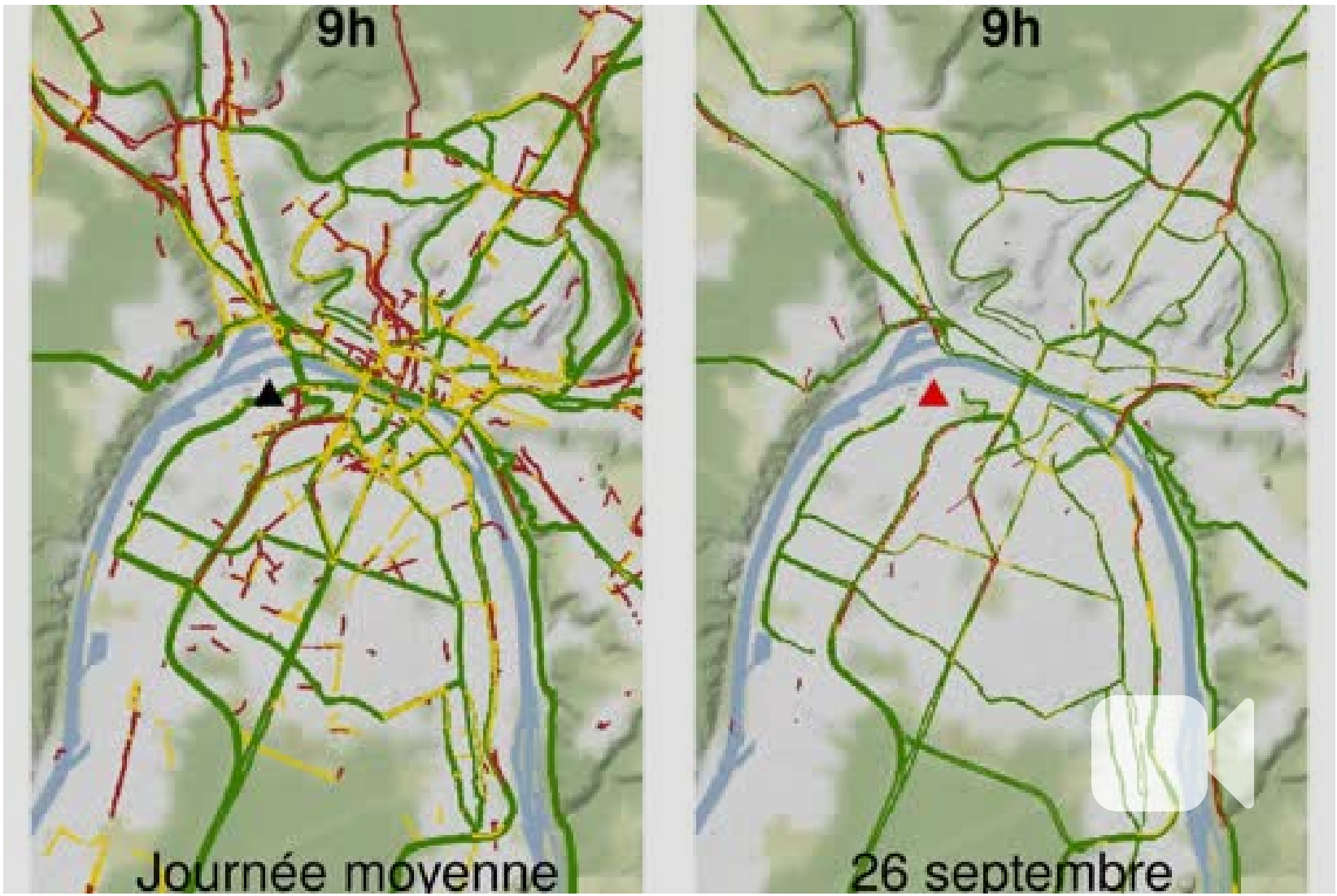
Condition de trafic

Bing Map (Microsoft)

Collecte en temps réel des conditions de circulation



Condition de trafic

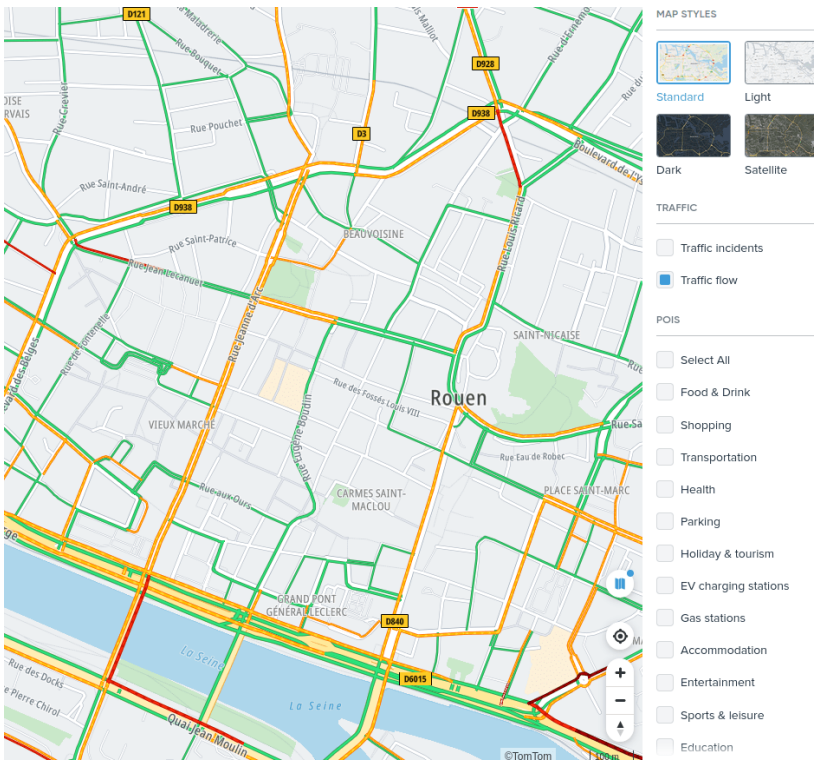


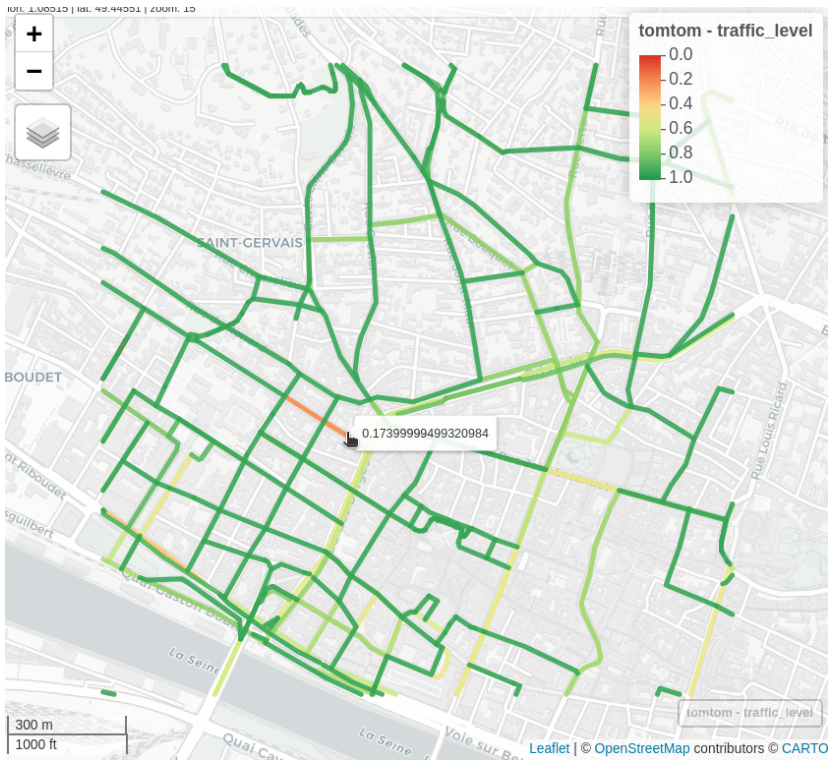
Trafic après l'explosion de Lubrizol

Condition de trafic

TomTom

Collecte en temps réel des conditions de circulation



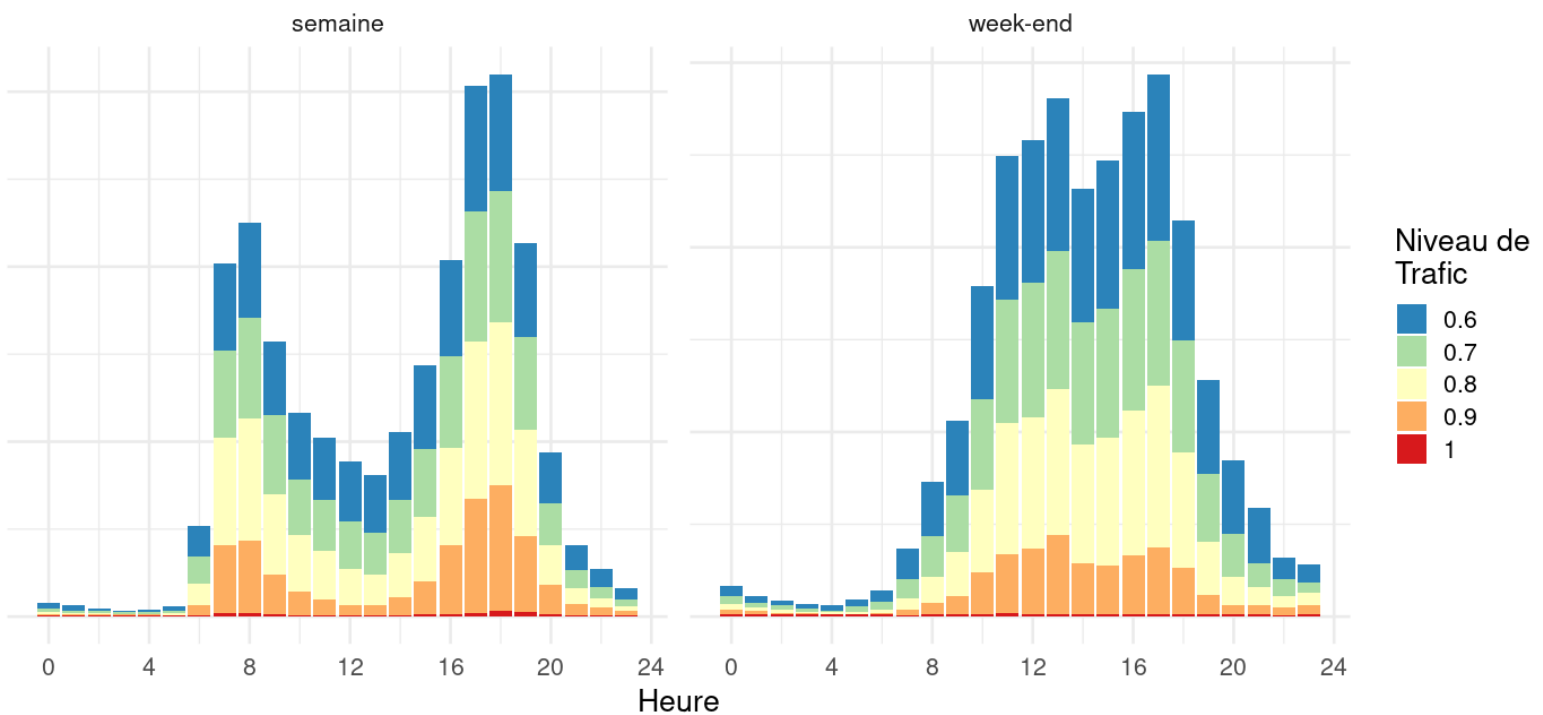


Données vectorielles scrappées plus complètes (niveau de trafic)

Condition de trafic

TomTom

Collecte en temps réel des conditions de circulation



Permet la validation de modèles de mobilités

Écart entre mobilité simulées et trafic observé

GeoCaching

GRR Trenum (2015 - 2019) / Philippe Vidal

Récupérer **toutes les informations** sur les **~ 13000 géocaches** (1/page) de Normandie à un instant **t**

Site web : geocaching.com
Langage : Python
Framework : Scrapy
Type de récolte : Campagne
Format : Json

Free as in Freedom

DOI 10.5281/zenodo.3542261

Difficultés :

- Pas d'API & limitation à 1000 résultats par recherche
- Mon premier scraping
- Login/Pwd pour accès à la donnée complète
- ASPX avec gestion des états précédents

Téléchargement automatique

הספרייה הלאומית
المكتبة الوطنية الإسرائيلية
The National Library of Israel

Projet RachelNet / Denis Eckert
Récupération d'ouvrages numérisés en ligne

Téléchargement semi-automatisé des pages d'ouvrages numérisés accessibles uniquement en navigation.

- Permettre une consultation et un archivage hors-ligne et plus rapide des documents


```r
library(tidyverse)
library(curl)
library(glue)
library(magick)

## Livre Bibliothèque Nationale d'Israel
premier <- "https://rosetta.nli.org.il/delivery/DeliveryManagerServlet?dps_func=stream&dps_pid=FL25784474"
dernier <- "https://rosetta.nli.org.il/delivery/DeliveryManagerServlet?dps_func=stream&dps_pid=FL25785206"

tous <- 25784474:25785206
errors <- numeric()
for (i in tous) {
  thisAdress <- glue("https://rosetta.nli.org.il/delivery/DeliveryManagerServlet?dps_func=stream&dps_pid=FL{i}")
  tryCatch({
    curl_download(url = thisAdress, destfile = glue("scraping_denis/bouquin_bibli_israel/{i}.jpg"))
  },
  error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}

system('convert "*.{jpg}" -quality 100 ouvrage.pdf')
```

Marché immobilier mexicain

PrimoMex / A. Ribardière, J.-F. Valette & L. Salinas

Caractériser le marché immobilier et son évolution dans l'agglomération de Mexico (ZMVM)

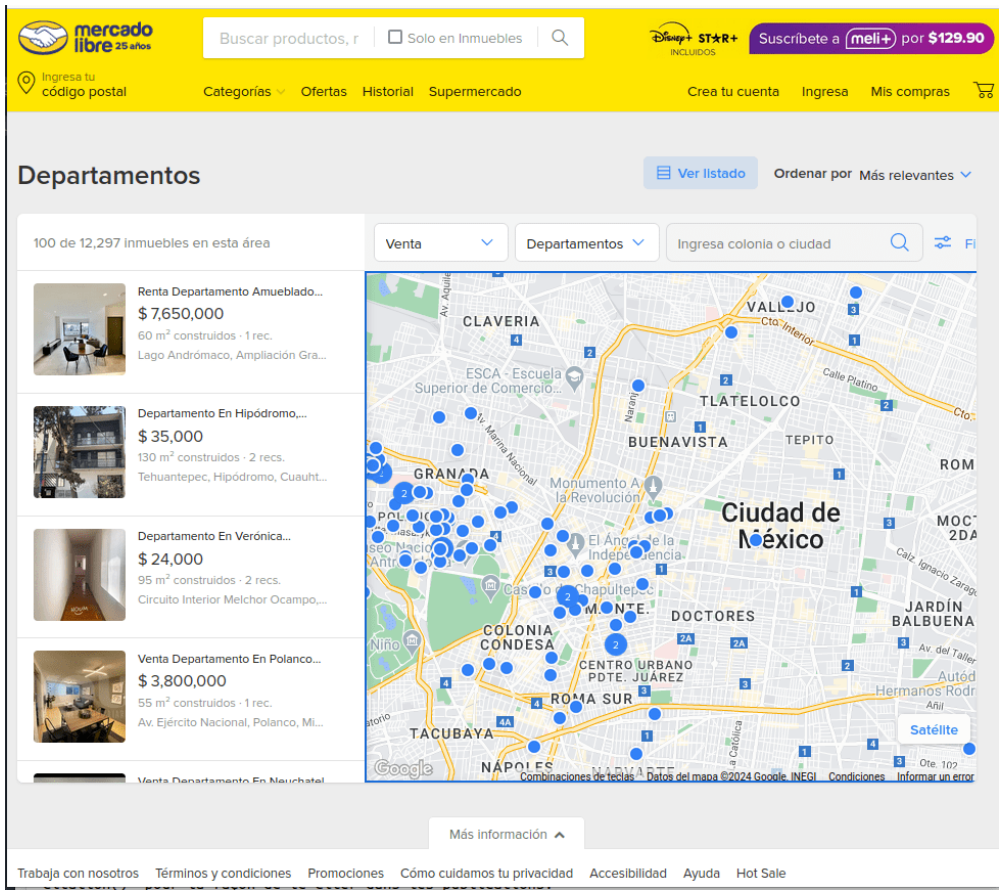
Un des objectifs :

Étudier l'évolution des prix à différentes échelles depuis 2015

Moyen : Base 2015 exhaustive, scraping de l'ensemble des annonces en 06/2024

Un des enjeux :

Détecter les annonces incomplètes, mal renseignées, ou trompeuses.



MGB : tarif des bars parisiens

Projet pédagogique : Interpolations et lissages

Détection de hot-spots et lissages spatiaux
des prix de la pinte de bière

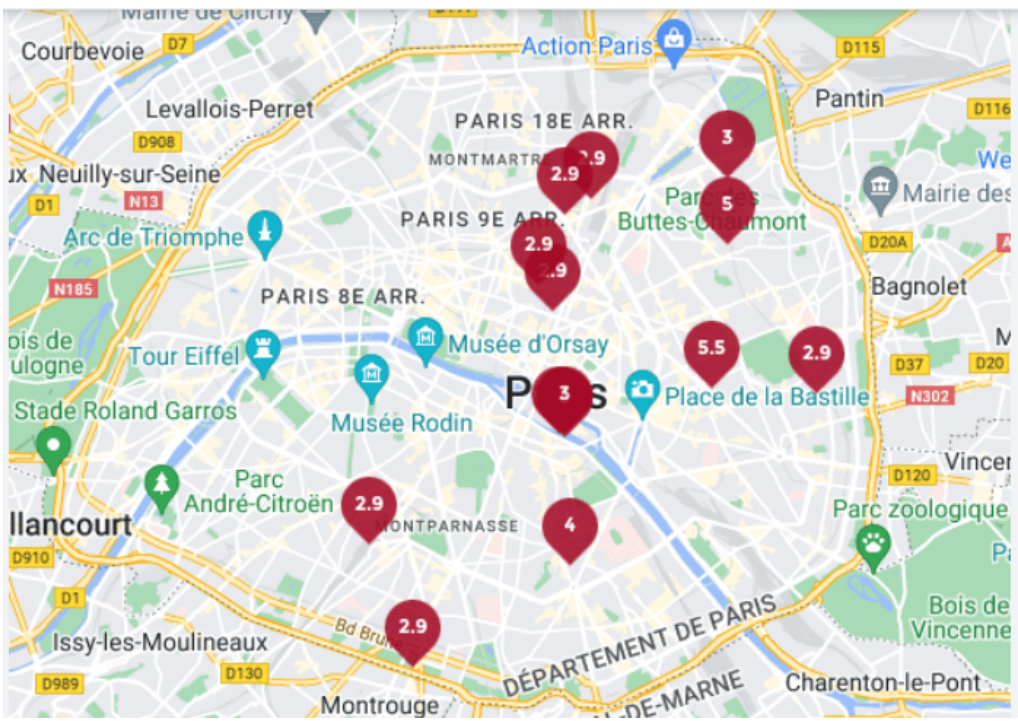
Un des objectifs :

Récupérer un jeu de données ponctuel très bruité mais présentant une nette structure spatiale

Moyen : Scraping "géomatique" de la carte du site MGB via fenêtres spatiales mouvants

Un des enjeux :

Passer outre la limite du site qui n'affichait que 20 annonces à la fois



Mobilités urbaines

FP7 DENFREE & ANR MO3 / Eric Daudé
Modéliser et simuler les épidémies de dengue

Un des objectifs :

Comprendre le rôle des mobilités quotidiennes dans la propagation des épidémies à Bangkok & Delhi

Moyen :

Créer un modèle de mobilité, individu centré, à base d'agents (couplé à d'autres modèles)

Concept d'espace d'activité

Un des enjeux :

Collecte de données de mobilité

Mobilités urbaines

FP7 DENFREE & ANR MO3 / Eric Daudé
Modéliser et simuler les épidémies de dengue

Un des objectifs :

Comprendre le rôle des mobilités quotidiennes dans la propagation des épidémies à Bangkok & Delhi

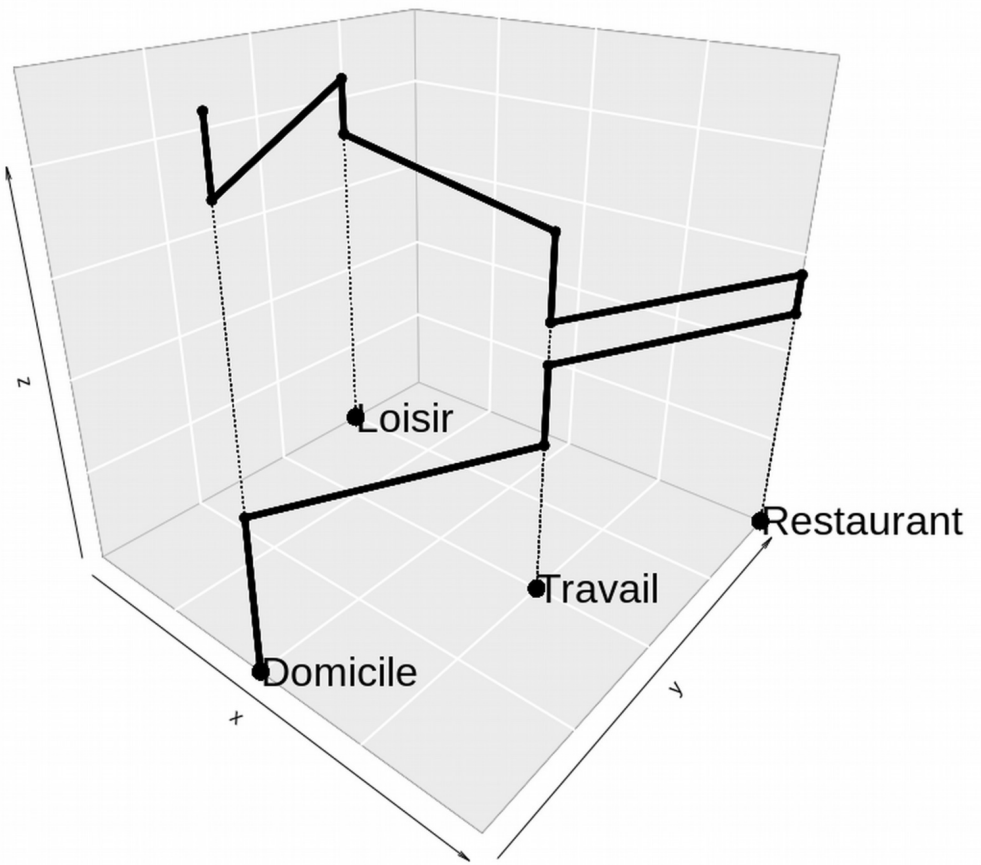
Moyen :

Créer un modèle de mobilité, individu centré, à base d'agents (couplé à d'autres modèles)

Concept d'espace d'activité

Un des enjeux :

Collecte de données de mobilité



Mobilités urbaines

FP7 DENFREE & ANR MO3 / Eric Daudé
Modéliser et simuler les épidémies de dengue

Un des objectifs :

Comprendre le rôle des mobilités quotidiennes dans la propagation des épidémies à Bangkok & Delhi

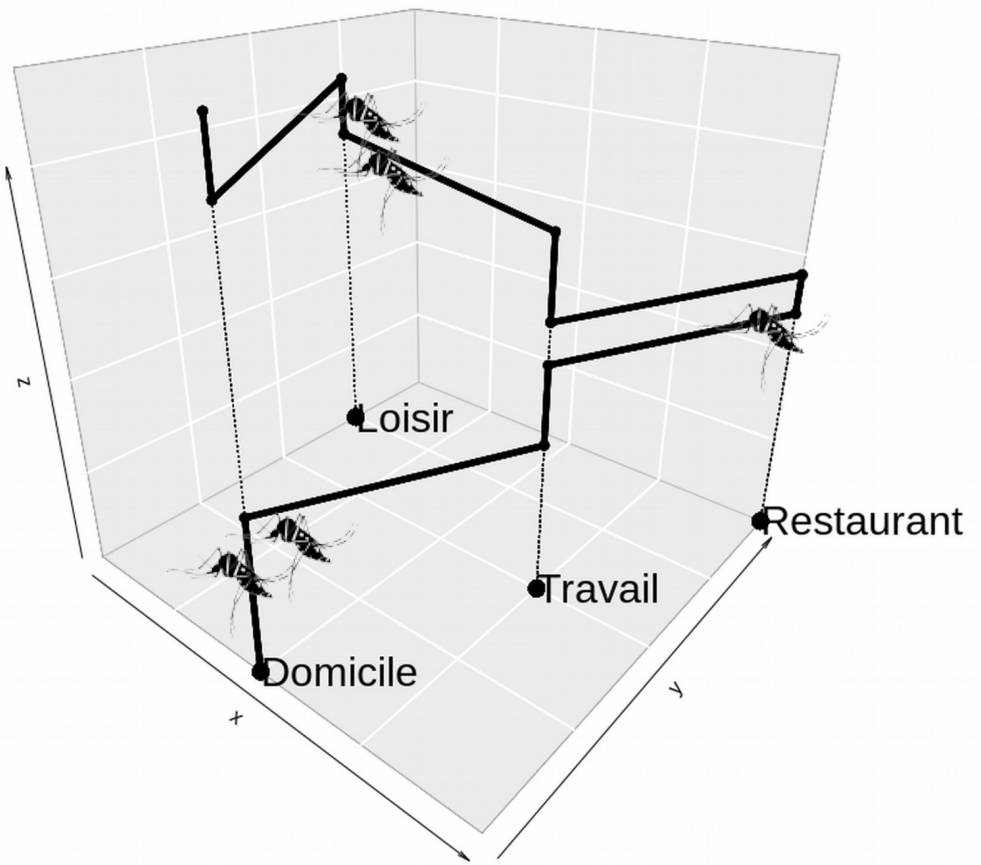
Moyen :

Créer un modèle de mobilité, individu centré, à base d'agents (couplé à d'autres modèles)

Concept d'espace d'activité

Un des enjeux :

Collecte de données de mobilité



Mobilités urbaines

FP7 DENFREE & ANR MO3 / Eric Daudé
Modéliser et simuler les épidémies de dengue

Un des objectifs :

Comprendre le rôle des mobilités quotidiennes dans la propagation des épidémies à Bangkok & Delhi

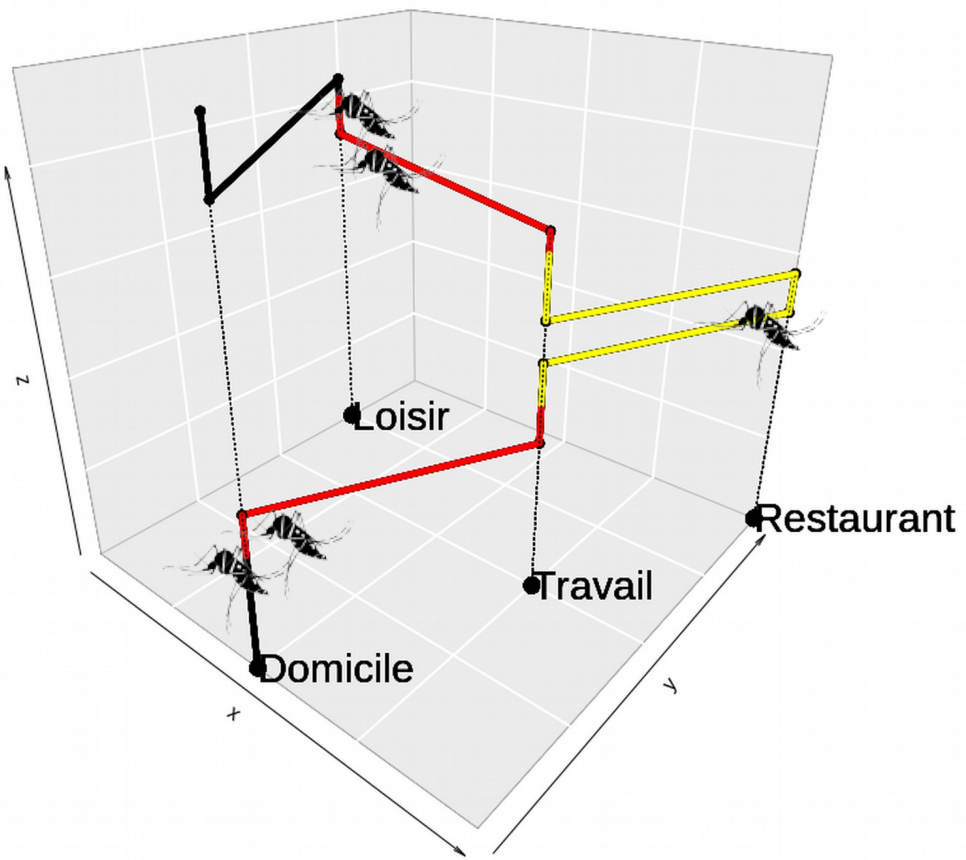
Moyen :

Créer un modèle de mobilité, individu centré, à base d'agents (couplé à d'autres modèles)

Concept d'espace d'activité

Un des enjeux :

Collecte de données de mobilité



Évacuations massives

ANR Escape (2016 - 2024) / Eric Daudé

"Contribuer à la conception de systèmes d'aide à la décision dans le cas d'évacuations massives"

Un des objectifs :

Simuler les mobilités en cas de catastrophes (industrielles ou naturelles). Multisites

Comment ?

Simulation à base d'agents

Préalable :

Connaître les comportements de mobilités en condition "normale"

==> Intérêt des conditions de trafic

(en plus des EMD)

Évacuations massives

ANR Escape (2016 - 2024) / Eric Daudé

"Contribuer à la conception de systèmes d'aide à la décision dans le cas d'évacuations massives"

Un des objectifs :

Simuler les mobilités en cas de catastrophes (industrielles ou naturelles). Multisites

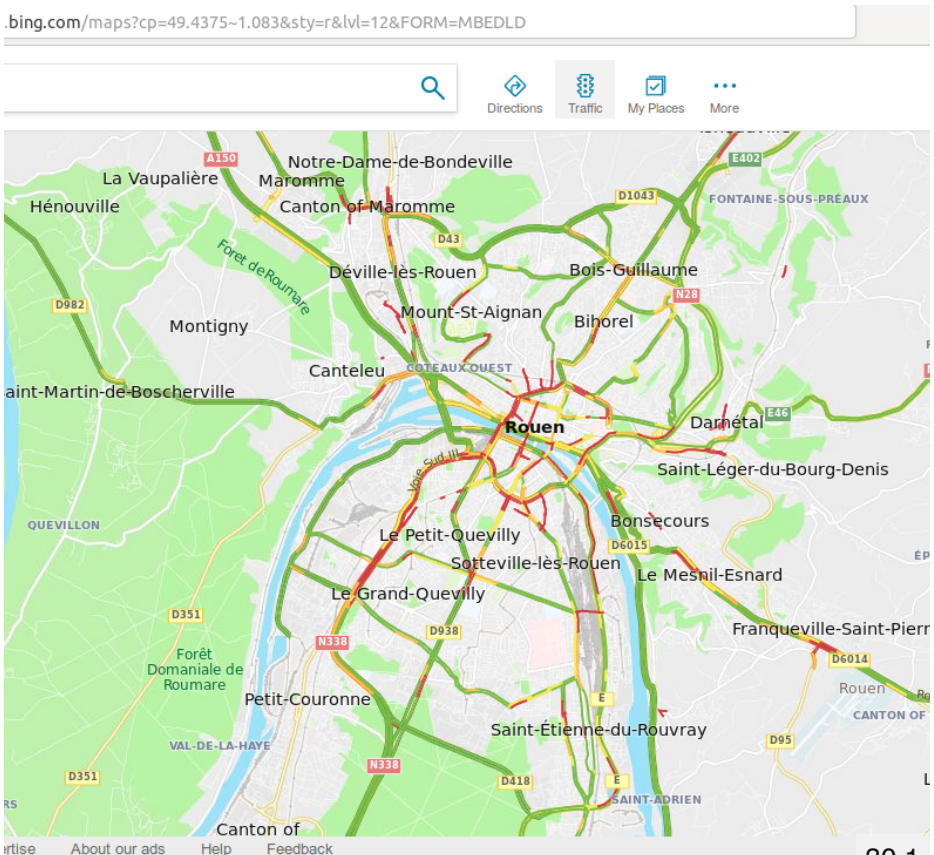
Comment ?

Simulation à base d'agents

Préalable :

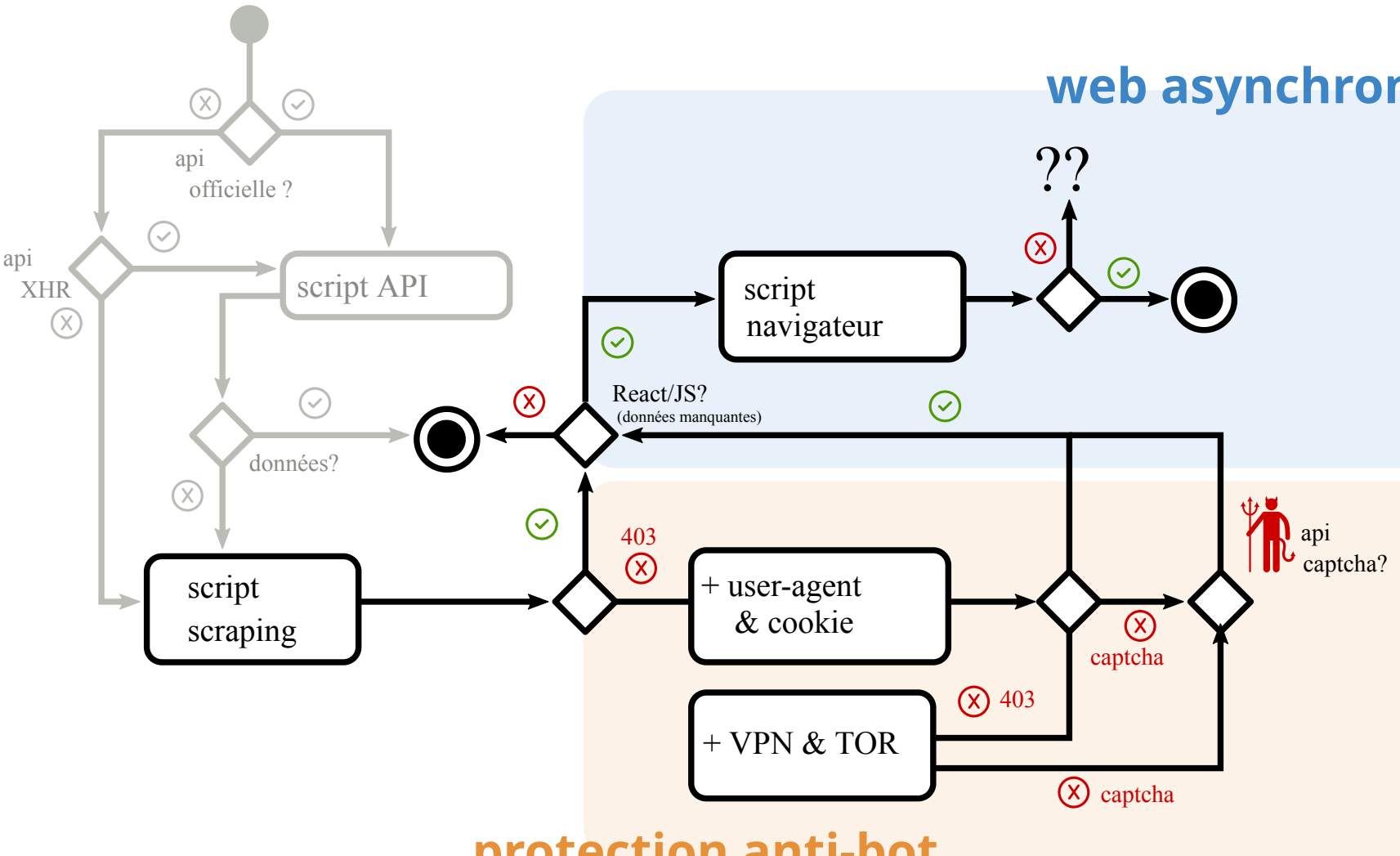
Connaître les comportements de mobilités en condition "normale"

==> Intérêt des conditions de trafic (en plus des EMD)



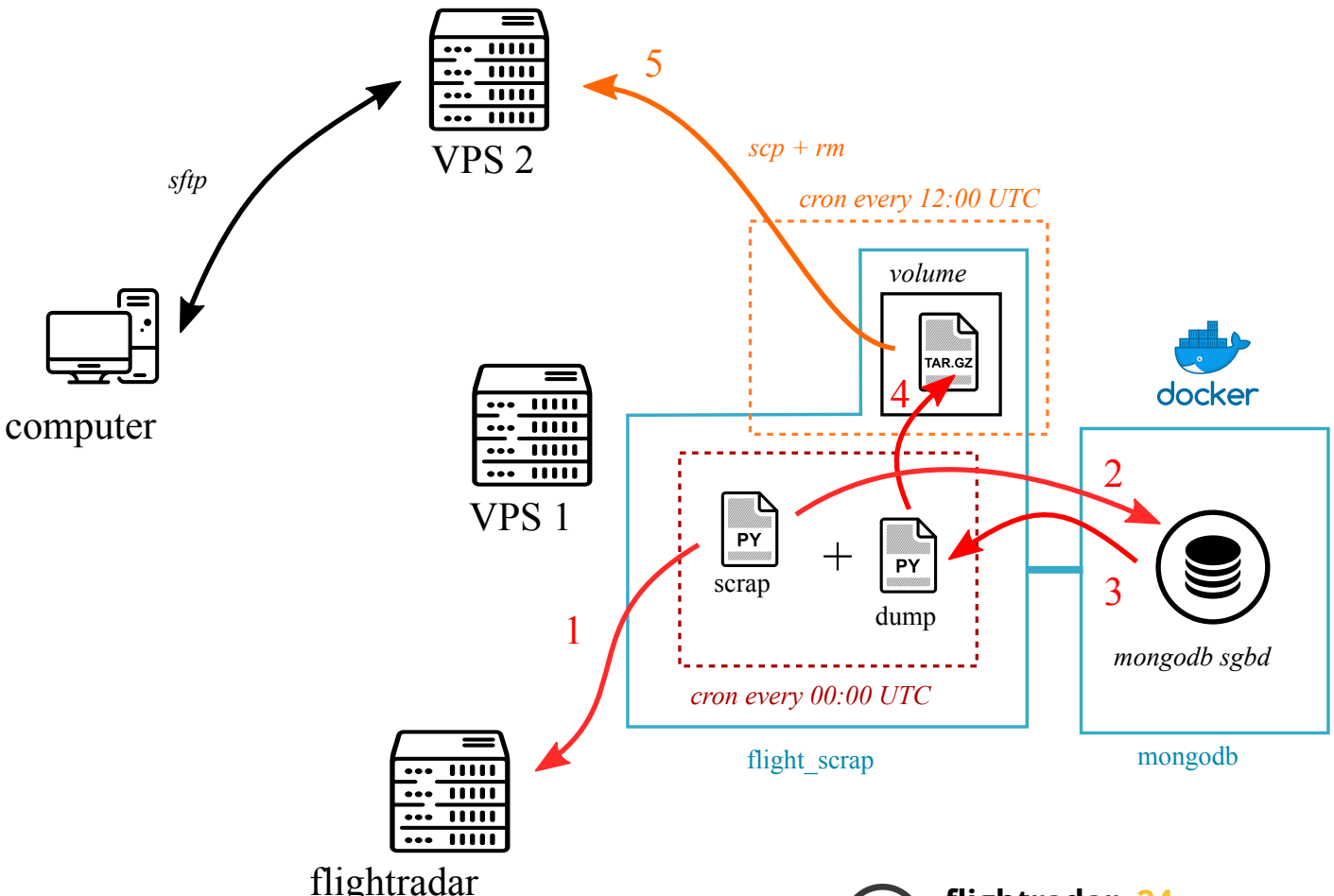
(III) Verrous techniques

(a) définir un logigramme pour mieux s'y retrouver



(III) Verrous techniques

(b) mettre une campagne de collecte h24/7J en place, une problématique complexe rapidement ...



flightradar 24
https://www.flightradar24.com
https://github.com/IDEES-Rouen/Flight-Scrapping