



# Analyse spatiale des contributeurs de Faune France et leurs contributions

Réalisé par : Axel Fourneyron  
Encadré par : Thierry Joliveau

Mémoire du Master Géographie Numérique en 1<sup>ère</sup> année



Faune  
France





<b>1</b>	<b><u>PROJET SOFT : SOCIOLOGIE DES OBSERVATEURS DE FAUNE-FRANCE SUR LE TERRITOIRE</u></b>	<b>4</b>
1.1	CONTEXTE ET PROBLEMATIQUE DU PROJET	4
1.2	ORGANISATION DU PROJET	5
1.3	MES MISSIONS DANS LE PROJET	6
<b>2</b>	<b><u>LA BASE DE DONNEES FAUNE FRANCE</u></b>	<b>8</b>
2.1	DESCRIPTION GENERALE	8
2.2	QUALITES DES VARIABLES	9
2.3	DEFINITION ET CARACTERISATION DU VOCABULAIRE UTILISE	9
2.4	METHODES DE RESTRUCTURATION	12
2.5	IMPORTATION DES DONNEES	14
<b>3</b>	<b><u>ANALYSES GLOBALES</u></b>	<b>17</b>
3.1	NOMBRE D'OBSERVATIONS	17
3.2	OBSERVATIONS SPATIALES GENERALES	20
3.3	ANALYSE TEMPORELLE	28
3.4	CLASSIFICATIONS DES CONTRIBUTEURS	32
<b>4</b>	<b><u>METHODOLOGIE D'ANALYSE SPATIALE</u></b>	<b>34</b>
4.1	EXPLICATION GENERALE	34
4.2	CONCENTRATION SPATIALE	36
4.3	ETENDUE/PORTEE SPATIALES	39
4.4	STRUCTURATION SPATIALE	44
<b>5</b>	<b><u>ANALYSES SPATIALES</u></b>	<b>45</b>
5.1	CONCENTRATION SPATIALE	45
5.2	ETENDUE/PORTE SPATIALE	50
5.3	STRUCTURATION SPATIALE	59
<b>6</b>	<b><u>CONCLUSION</u></b>	<b>61</b>
6.1	LES DIFFERENTS COMPORTEMENTS DES CONTRIBUTEURS	61
6.2	EVALUATION DES SOLUTIONS MIS EN PLACE	62
6.3	PISTES D'ANALYSES ENVISAGEABLE	62
6.4	REPRODUCTIBILITE ET REUTILISATION DES SCRIPTS	63
<b>7</b>	<b><u>ANNEXES</u></b>	<b>64</b>
7.1	LOGIGRAMMES DES SCRIPTS ET FONCTIONS	64
7.2	BIBLIOGRAPHIE DES FONCTIONS R	68
7.3	REQUETES SQL	69
7.4	GRAPHIQUES STATISTIQUES COMPLEMENTAIRES	70
7.5	REFERENCES BIBLIOGRAPHIQUES	77
7.6	AUTRES DOCUMENTS	78



# 1 Projet SOFT : Sociologie des Observateurs de Faune-France sur le Territoire

## 1.1 Contexte et problématique du projet

Faune-France est un portail naturaliste qui permet la consultation d'une banque de données de plus de 50 millions d'informations collectées par un réseau de naturalistes bénévoles et professionnels, issus d'une cinquantaine d'associations naturalistes. Cette base de données est actuellement animée par la Ligue pour la Protection des Oiseaux (LPO).

Depuis quelques années, ce réseau de naturalistes ne cesse d'augmenter (plus de 70 000 utilisateurs du portail). Ils associent à une pratique de loisir nature des signalements naturalistes sur des bases de données numériques qui permet la géolocalisation de la Faune partout en France.

La LPO souhaite mieux connaître le profil sociologique, les motivations et les pratiques de ces milliers de naturalistes bénévoles. C'est dans cette optique que le projet SOFT a vu le jour. Il a pour principal objectif de mieux comprendre la diversité des publics et leurs motivations, afin d'envisager la façon de faire évoluer l'observatoire Faune-France.org et ses manières de communiquer et de l'adapter au plus près des attentes des contributeurs.

---

### *Identifier les motivations et pratiques des personnes contribuant au recensement de la faune à travers Faune France.org*

---

Faune France met en avant une méthode de Science participative qui vise à faire appel à l'ensemble de la population de manière bénévole pour observer la globalité de la Faune dans son milieux naturel. Cette méthode très utilisée dans le recensement est expliquée par *Florian Charvolin* dans une revue électronique en sciences de l'environnement en 2017 nommée : *Sortie nature, protocole et hybridité cognitive. Note sur les sciences participatives*.

Cette méthode qui présente de réels avantages à récolter des données peut cependant poser des questionnements sur la qualité de la donnée récoltée. Effectivement, un spécialiste fassant un recensement dans la faune pourra indiquer une information bien plus précise, exacte et pertinente de ce qui se passe. Un débutant quant à lui pourra indiquer ce qu'il perçoit. Son information a des risques d'être inexacte, cependant elle présente quand même l'avantage d'être utile.

La difficulté d'interpréter les données de la science participative est qu'il faut avant tout comprendre qui la saisit, comment cette saisie est réalisée et dans quel objectif. En comprenant ces différents facteurs il sera alors possible de traiter les données en cernant leurs avantages et leurs risques ce qui permettra d'avoir une représentation plus proche de la réalité.

Il faut alors prendre en compte que l'accompagnement à la saisie et le traitement de la donnée doit être différent en fonction du type de public. Mais alors comment savoir quels sont les différents publics participant de Faune France ?

## 1.2 Organisation du projet

### 1.2.1 Acteurs

Faune France a donc décidé de faire appel au CNRS notamment aux laboratoires Centre Max Weber (CMW) et Environnement Villes et Sociétés (EVS) pour réaliser différentes analyses de leurs données « opportunistes ».

Les analyses ont pour but de cerner la diversité des publics et leurs motivations, les caractéristiques de leur participation numérique ainsi que leur comportement géographique. Plusieurs outils seront mis en place dans ce projet : analyse statistique et spatiale de la base de données, enquête par questionnaire et par entretien afin de valider certains comportements.

Le projet SOFT doit durer deux ans avec pour acteurs principaux Florian Charvolin (Directeur de recherche en sociologie), Laurent Couzy (Partenaire LPO en charge de la base de données Faune-France), Karine Pietropaoli (Ingénierie Statisticienne au CNRS) et Thierry Joliveau (spécialiste des questions géonumériques).

### 1.2.2 Déroulement

Le déroulement du projet est prévu en quatre grandes étapes qui comprendront enquêtes, analyses, questionnaires et entretiens répartis de la manière suivante :

#### Enquête préliminaire : (Janvier-Avril)

- Recherche documentaire
- Pré-enquête auprès des administrateurs

#### Analyse des caractéristiques des contributeurs de la base de données Faune-France.org : (Avril-Aout)

- *L'enquête se centrera exclusivement sur les contributeurs à la plate-forme numérique Faune-France.org et non sur les usagers qui interrogent la base*
- Connaître l'identité des contributeurs en terme socio-spatial à partir de leurs observations.
- Croiser les lieux de résidence et les lieux d'observation à l'échelle nationale
- Identifier les attentes des contributeurs en termes de fonctionnalités et de restitutions proposées.

**Variables à explorer :** localisation résidence des contributeurs ; distance domicile/lieu d'observation, localisation des lieux d'observation les plus prisés, espèces les plus observées et espèces observées/espèces rares.

#### Enquête par questionnaire auprès des contributeurs Faune-France.org : (début septembre)

- Questionnaire court pour l'ensemble des utilisateurs
- Questionnaire approfondie remis aux contributeurs les plus assidus

#### Enquête par entretiens : (Avril)

- Enquête approfondie
- Analyses des entretiens

### 1.3 Mes missions dans le projet

Mon stage a été principalement dédié à l'analyse géographique des caractéristiques des contributeurs de la base de données Faune-France.org. Pour cela plusieurs missions m'ont été attribuées :

#### 1.3.1 Gestion de la base de données

- Restructuration de la base de données à partir de fichiers .txt et .csv
- Création de scripts pour mettre à jour la base de données

#### 1.3.2 Analyse spatiale des données Faune France

- Visualisation des comportements globaux des variables
- Hypothèses de comportement spatial des contributeurs.
- Méthodes d'analyses
- Analyses spatiale et spatio-temporelle des contributeurs
- Identification du comportement des contributeurs suite aux analyses.
- Création de scripts pour reproduire les analyses.

#### 1.3.3 Documents de communication pour pérenniser le travail

- Recherche bibliographique sur les analyses spatiales liées à des comportements utilisateurs
- Guide méthodologique pour partager, utiliser et actualiser la base de données SQL
- Documents explicatifs sur les différents scripts (fonctions utilisées, paramètres et configuration, objectifs des variables, explications des résultats, hypothèses sur les comportements, ...)
- Répertoire contenant l'ensemble des scripts utilisés pour les analyses

#### 1.3.4 Logiciels utilisés

Pour réaliser mes missions et pouvoir partager mon travail avec le reste de l'équipe, je me suis orienté principalement sur trois logiciels :

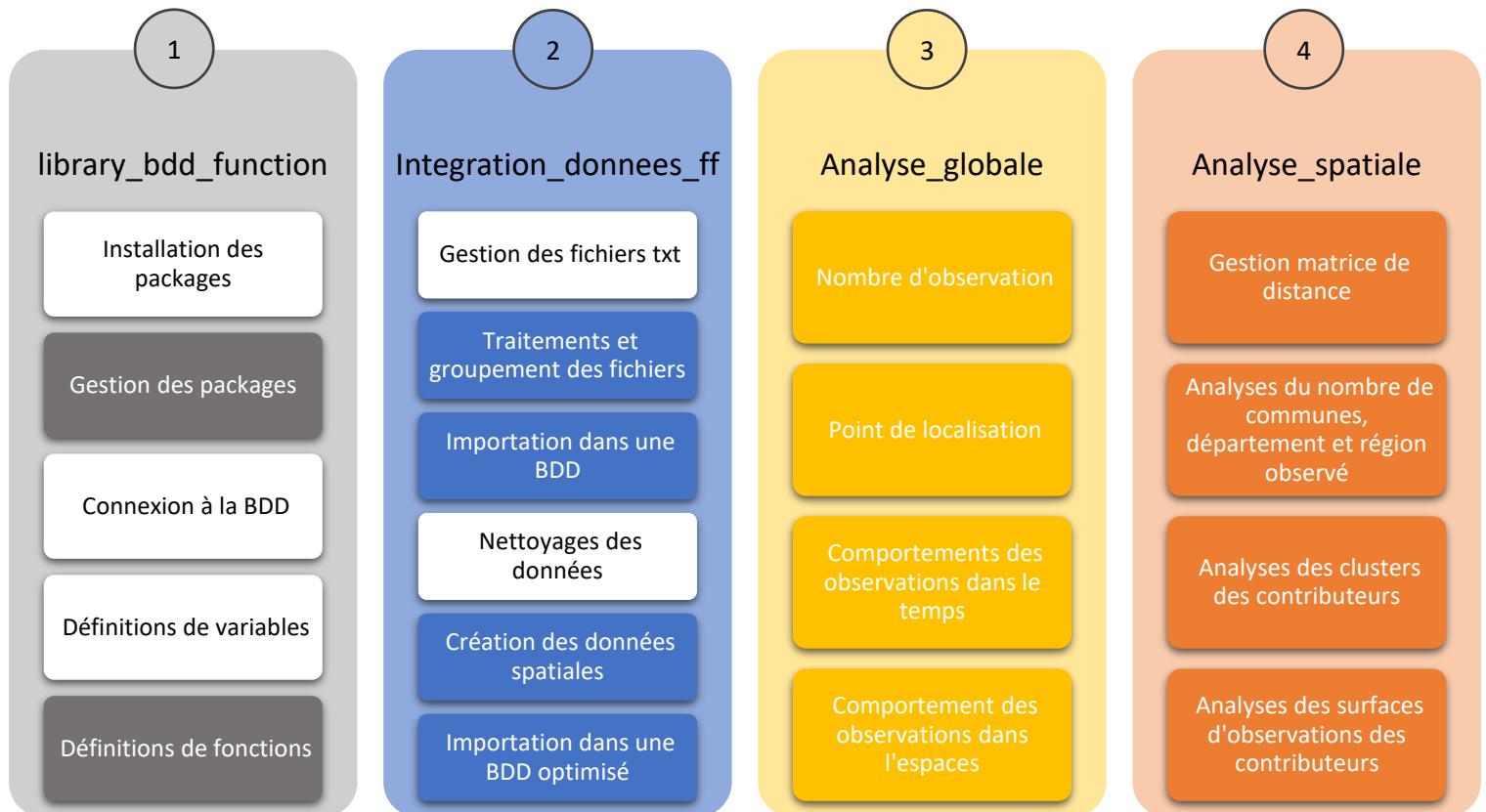
- PostgreSQL/PostGIS : pour la création et gestion de la base de données, mais aussi pour certains calculs, traitements et analyses spatiales.
- R et Rstudio : pour toutes les questions de traitements statistiques, analyses spatiales et traitements de données.
- QGIS : pour certaines visualisations cartographiques de petits jeux de données

Le choix de ces logiciels a porté sur la capacité à traiter des jeux de données importants (plus de 12 millions de données), mais aussi à analyser spatialement des données à travers différentes méthodes. Enfin la coordination avec les autres membres de l'équipe notamment avec Karine afin de mutualiser et de partager un travail réutilisable a été nécessaire pour la suite du projet. De plus ces outils ont l'avantage d'être gratuits, régulièrement mis à jour et documentés si une nouvelle personne intègre le projet.

### 1.3.5 Méthodes de travail

Dans l'objectif de pouvoir fournir un travail réutilisable et reproductible suite aux éventuels ajouts de données (2019 ou 2020), chaque grande partie du projet correspond à un Script R intégralement commenté et organisé en plusieurs sous parties. Pour faciliter leurs utilisations, ses scripts sont séparés en quatre fichiers, qui sont eux même séparés en plusieurs parties de façon à pouvoir relancer seulement certaines parties d'analyse. Mais attention ! Il est impératif que les scripts *library\_bdd\_function* et *Integration\_donnees\_ff* soient réalisés en premier pour pouvoir exécuter les autres.

Ces scripts sont disponibles dans mon répertoire GitHub<sup>1</sup>. On retrouvera alors les scripts suivants :



Chacun de ces scripts est commenté ligne par ligne afin d'être réutilisable, adaptable ou modifiable. Bien sûr, avant de pouvoir réutiliser les scripts certain prérequis sont nécessaire : *la définition des variables de la base de données, l'emplacement des fichiers, le paramétrage de Rstudio pour avoir l'ensemble des packages utilisés, l'intégration de code pour les nouveaux jeux de données ainsi que le nettoyage adapté aux nouveaux jeux de données*. Ces prérequis paramétrables dans les sous-parties sont affichés en blanc.

Les seules parties non reproductibles sont les cartographies, l'interprétation des résultats et des méthodes d'analyse spatiale.

<sup>1</sup> <https://github.com/FourneyronA/Analyse-spatiale-sous-R>

## 2 La base de données FAUNE France

Phillipe Jourde a partagé 26 fichiers texte et un fichier csv qui contiennent l'ensembles des observations de la base de données Faune France de 2017-18 ainsi que l'ensemble des contributeurs inscrits (aperçus des fichiers en annexe page 78). Ces fichiers ont été transférés dans une base de données PostgreSQL afin de faciliter les manipulations et le traitement des données spatiales.

### 2.1 Description générale

La base de données FAUNE France 2017-18 comporte 12 084 804 observations, avec 88 364 contributeurs, sur 4 096 881 lieux différents, pour 3 946 espèces uniques. Pour chaque observation, il y a plus de 75 variables qui sont assimilées. Elles décrivent plusieurs informations sur la temporalité, la spatialité, la personne, la méthode, l'animal qui a été observé.

Parmi l'ensemble de ces 75 champs de variables, nous avons retenu les informations les plus pertinentes qui nous permettent de mieux profiler l'utilisateur de FAUNE France ayant réalisé cette observation. Nous avons donc regardé principalement les informations suivantes :

#### Données temporelles

- La date de l'observation
- La date d'insertion de l'observation
- L'horaire de l'observation
- La date d'inscription de l'utilisateur
- La date de naissance de l'utilisateur
- La date de dernière connexion

#### Données localisations

- Type de localisation
- Latitude et Longitude (WGS84 et L93)
- Altitude

#### Données sur l'animal

- L'espèce précise
- La famille de l'espèce
- La protection nationale de l'espèce
- L'estimation de la sûreté de l'observation
- Le comportement de l'animal
- La contenance de détails se référant à la mortalité de l'animal

#### Données sur l'observation

- Les données de seconde main
- La protection des données
- L'anonymat des données
- La vérification des données
- La contenance des médias
- La participation à une étude
- Les observations sous forme de liste

#### Données supplémentaires ajoutées

- La typologie des paysages de la France (source INSEE)
- La typologie des unités (urbaine/rural) de la France (source INSEE)
- Les polygones des communes
- Les points de localisation des contributeurs
- Les points de localisation de l'observation
- La distance à vol d'oiseau entre l'observation et son observateur.

## 2.2 Qualités des variables

Avant et pendant l'import des données il a fallu vérifier la qualité de ces données pour savoir si elles étaient exploitables. Dans l'ensemble, les champs sont correctement renseignés. Seules les données utilisateurs et la base utilisateurs ont posé des problèmes :

- Le renseignement de la commune, communes hors zone et le code postal car ce sont des champs qui ne sont pas fermés. On retrouve souvent des erreurs tel que le code postal dans le champ commun, ou bien la commune par défaut (Arles).
- L'indication de la date de naissance qui pour certain est indiqué à l'an 0 ou 2019.
- L'indication de la dernière connexion qui est réalisée avant leur première connexion.
- L'indication de la dernière date d'insertion répertoriée avant la date de leur dernière observation.
- Le décalage de certaines colonnes suite à l'utilisation de caractères spéciaux dans la colonne commentaire de l'utilisateur (utilisation des antislash \)
- Et probablement la date d'inscription suite à la mise en commun de plusieurs bases de données, mais sans aucun moyen de le vérifier

Grâce à Karine, certaines données ont pu être nettoyées (communes, communes hors zone, code postal) ce qui aide beaucoup à la localisation du contributeur qui est un point clé de l'analyse spatiale. Cependant pour les dates, cela peut créer un « doute » dans l'analyse temporelle. Il faut donc le garder en tête si des analyses semblent incohérentes ou inhabituelles.

## 2.3 Définition et caractérisation du vocabulaire utilisé

### 2.3.1 Qu'est-ce qu'une observation ?

Faune France a pour but de connaître le comportement de la faune dans son milieu naturel à travers les observations des différents participants (amateurs, passionnés, professionnels).

A travers une observation il est possible de renseigner l'espèce d'un animal, leur nombre, leur comportement et des commentaires/remarques. Cependant, il est important de comprendre que si nous voyons deux animaux d'espèces différentes il faudra alors renseigner deux observations. Dans l'objectif de faciliter cette démarche, Faune France a mis en place un outil de « Liste » qui permet de renseigner à travers une liste plusieurs espèces différentes. Cet outil vient alors créer une multitude d'observations correspondant au nombre d'espèces différentes. Ainsi il sera très facile de renseigner en quelques clics une ou plusieurs observations.

La saisie d'une observation peu se faire sur plusieurs plateformes, le site Faune France ou l'application Naturaliste qui permet aux amateurs, passionnés ou professionnels de pouvoir saisir facilement les observations « naturalistes » réalisées tout en conservant une trace informatique.

L'observation naturaliste venant de professionnels est parfois plus complexe, elle peut parfois prendre en compte une fréquence d'apparition avant une saisie informatique. Même si les méthodes de renseignement sont diverses en fonction des compétences des personnes, la saisie d'une observation sur Faune France tend à garder une démarche similaire pour tous : renseigner ce que l'on peut percevoir. Ainsi certains champs restent facultatifs.

L'objectif dans ce mémoire est de comprendre les contributeurs de Faune France à travers leurs observations. Ce sont donc toutes les variables liées à l'observation, qu'elles relèvent de facteurs temporel, spatial, qualitatif ou encore quantitatif qui permettront d'émettre des hypothèses pour caractériser les contributeurs sur leur volonté d'agir, leurs lieux d'actions, la fréquence, l'impact, la concentration de leurs actions, la diversité de leurs observations et leurs utilisations de l'application.

### 2.3.2 Comment un utilisateur renseigne-t-il une observation ?

**ETAPE 1/3 : CHOIX D'UN LIEU-DIT**

en tapant du texte  **AFFICHER**

par coordonnées géographiques  
Lon  Lat  **AFFICHER**

en choisissant une commune  
 <- Tapez le début d'une commune **AFFICHER** **ZOOMER SUR LA COMMUNE**

en choisissant dans les propositions basées sur votre historique  
mes derniers lieux-dits mes lieux-dits les plus utilisés  
université (campus Tréfilerie)

en cliquant sur la carte  
[aller vers ma dernière donnée] [vue générale de ma région]  
4°20'18" E / 45°31'19" N

\* Date 16.07.2019 **samedi passé** **dimanche passé** **hier** **aujourd'hui**

\* Espèce

\* Nombre total d'individus  
Valeur exacte  1

▼ Les champs ci-dessous sont facultatifs

Nombre	Sexe	Age	Conditions
5	5x mâle	adulte	en main

[ajouter individus supplémentaires]  
5ma

Autres données/informations

Donnée protégée  Donnée de seconde main  
 L'animal est mort ou blessé

Comportement :  Accouplement  
 Se déplace  
 Se nourrit  
 Prédaté  
 Marquage de territoire  
 Rut, parade  
 Sous une plaque

Vous pouvez fournir une image JPEG (max. 450 pixels sur le petit côté) ou un son MP3 de votre animal (max. 1 Mo)

Parcourir... Aucun fichier sélectionné.

Commentaires

Remarque  Remarque protégée

Le contributeur peut renseigner ses observations à travers la plateforme du site, ou l'application *NaturaList*<sup>2</sup>.

Les étapes que l'on peut voir à gauche sont similairement les mêmes sur l'application et sur le site.

Le contributeur vient renseigner une localisation à travers différentes possibilités, choix d'un lieu-dit en cliquant sur le point d'une carte, insertion des coordonnées géographiques Longitude, Latitude, détection de celles-ci par le mobile ou bien par positionnement manuelle sur une carte.

Ensuite il renseigne la date, le groupe taxonomique, l'espèce (si connue), le nombre d'animaux vus, les informations correspondant aux animaux, une photo (si souhaité) et des commentaires visibles ou non (si souhaité).

Ces nombreuses données ainsi renseignées nécessitent un peu de réflexion. Tous les observateurs ne perçoivent par la même rigueur, volonté ou motivation d'observation la nature. Ses différences peuvent alors créer une distorsion de la réalité si nous les traitons toutes de la même façon.

C'est notamment dans la partie des méthodes d'analyses qu'on vient alimenter différentes perspectives et hypothèses pour essayer d'établir des indicateurs qui témoigneraient d'un comportement spatial plus précis.

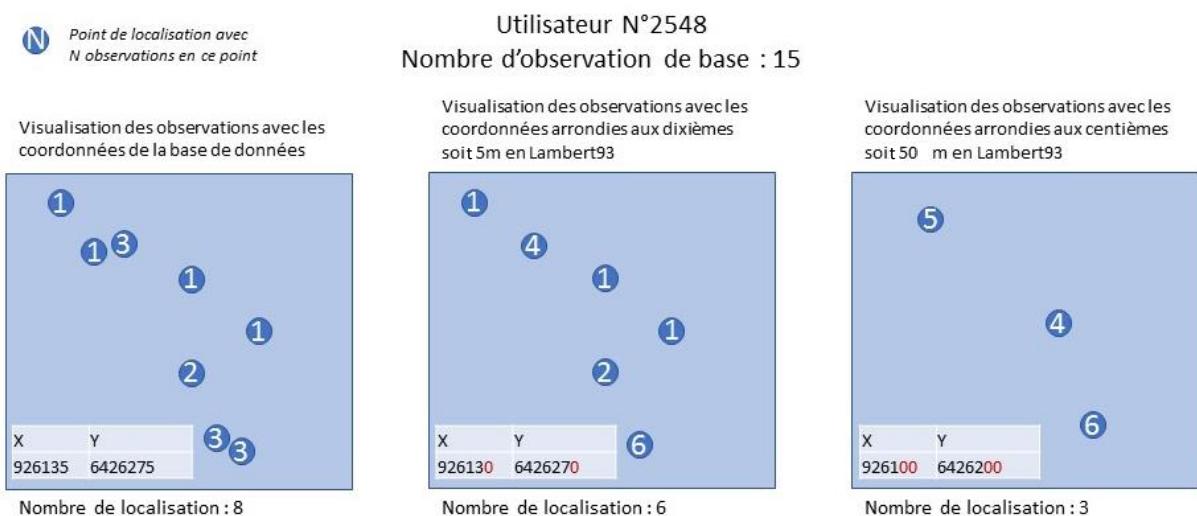
<sup>2</sup> Présentation de l'application : <https://www.youtube.com/watch?v=v8zOwMfUvms>

### 2.3.3 La localisation des observations

Le point de localisation est l'endroit où le contributeur vient renseigner une ou plusieurs observations. Un point de localisation est donc différent d'une observation, car il peut contenir plusieurs observations témoignant de plusieurs espèces animales. Cependant cette information reste relativement influencée avec le nombre d'observations car si un utilisateur ne fait qu'une seule observation, il n'y aura qu'une seule localisation. Il est possible qu'un utilisateur avec plus de 100 observations ait aussi qu'une seule localisation (exemple : un ornithologue qui ne regarde qu'un seul spot).

Les points de localisation peuvent se regarder à différentes échelles, allant du plus précis (X et Y de base dans la base de données) au plus vaste, 5 km aux alentours. En Lambert 93 il est possible de jouer sur cette distance de localisation en arrondissant les données X et Y.

Voici un schéma récapitulatif pour illustrer le propos :



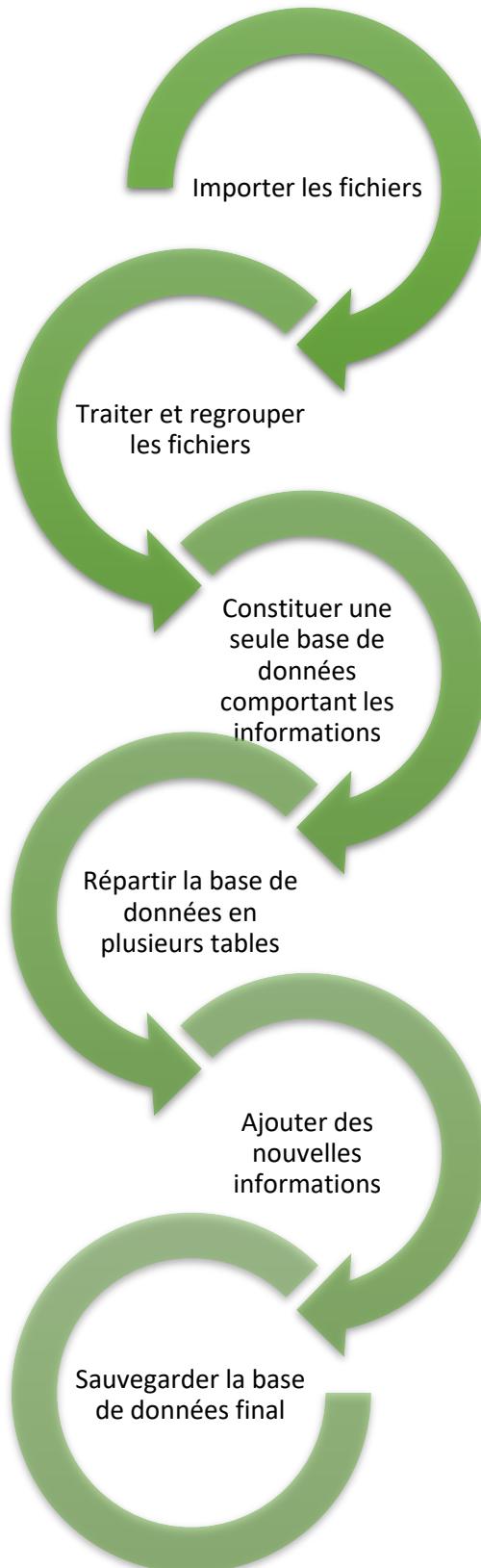
Le nombre de localisations d'un contributeur est très utile pour l'analyse spatiale, car il est parfois nécessaire d'avoir un nombre minimum de point de localisation afin d'effectuer des analyses comme le calcul de surface qu'observe un contributeur (une surface = 3 points minimum).

Cette variable nous servira à la fois d'indicateur sur la concentration spatiale en regardant à différents grains d'observation (1 mètre, 50 mètres, ...) mais aussi un indicateur sur les possibilités d'appliquer des traitements spatiaux (calcul de surface, calcul de périmètre, clustering, ...).

## 2.4 Méthodes de restructuration

### 2.4.1 Etapes et démarches

- Harmonisation des données (les fichiers textes n'ont pas toujours le même nombre de colonnes)
- Regroupement des données
- Suppressions des premiers doublons
- Définition du type de variable
- Création d'une base de données structurée et optimisée sans redondance
- Groupement des données redondantes et transfert vers la table attribuée
- Enregistrement de la base de données pour faciliter le partage et les réutilisations.

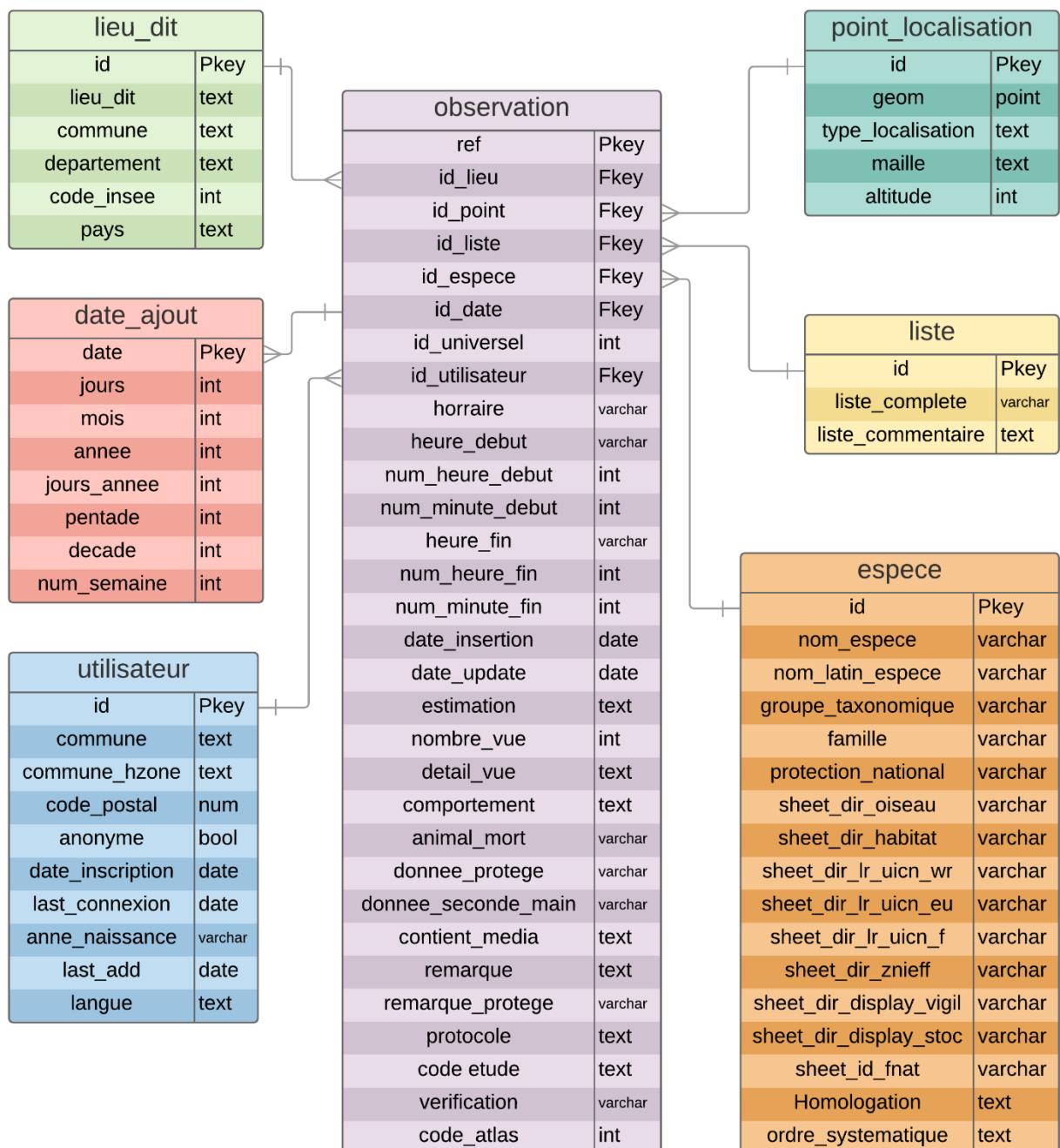


- Transfert des fichiers texte et CSV vers le logiciel R pour permettre une manipulation des données plus facile avec PostgreSQL
- Crédit à l'auteur : [lien vers la source]
- Création d'une table dans PostgreSQL à partir de R
- Transfère de l'intégralité des données dans une seule table
- Suppression des données non désirables (ne faisant pas partie de l'étude)
- Crédit à l'auteur : [lien vers la source]
- Création des variables géométriques (point, ligne) à partir des coordonnées X et Y présentes dans la base de données.
- Importation de la carte de la France (polygone des communes)

#### 2.4.2 Crédation d'un modèle conceptuel pour la base de données

Plusieurs tables n'ont pas été faciles à réorganiser, avec des champs comme le point de localisation et le lieu-dit de l'observation qui font tous les deux référence à une spatialité, mais de façon différente. Ou encore la date d'ajout de l'observation et les horaires d'observations renvoient à une temporalité. Cependant l'horaire étant un champ variable qui n'est pas présent sur toutes les observations, cela devient plus difficile de les grouper.

L'objectif de ce modèle conceptuel n'est pas d'être parfait. Faune France dispose déjà d'un modèle fonctionnel qu'elle ne peut pas nous transmettre. L'objectif était donc de mettre en place un modèle qui diminue la taille de la base de données tout en facilitant son utilisation.



## 2.5 Importation des données

L'importation des données, comme les analyses, est automatisé, cependant elle nécessite quelques prérequis.

1. Le premier prérequis est de créer la base de données sur PostgreSQL en localhost (voir guide en annexe page [80](#))
2. Le deuxième prérequis est de vérifier que tous les packages dans Rstudio soient correctement installés, pour cela il vous suffit de lancer le script `installation_libraries`<sup>3</sup>
3. Le troisième est de changer les paramètres du script `library_bdd_function`<sup>4</sup> qui permet la connexion à la base de données dans la partie « 2 - Connexion BDD FF\_2018 ».
4. Le quatrième prérequis est la gestion de l'emplacement des fichiers texte et csv à importer dans la base de données. Là aussi il faudra changer les paramètres du script `library_bdd_function` dans la partie « 3 - Emplacement des fichiers » afin qu'il corresponde au fichier indiqué.

Enfin une fois ses étapes terminées on peut lancer l'intégralité du script permettant l'intégration de donnée : `integration_donnees_ff`<sup>5</sup>

Il est possible avec un nouveau jeu de données que le script ne marche pas du premier coup. Il peut y avoir des erreurs de doublon qui ne sont pas encore traitées. Plutôt que de supprimer des données automatiquement, il est préférable de comprendre la problématique afin d'appliquer la solution adéquate.

---

<sup>3</sup> [https://github.com/FourneyronA/Analyse-spatiale-sous-R/blob/master/R\\_0\\_installation\\_libraries](https://github.com/FourneyronA/Analyse-spatiale-sous-R/blob/master/R_0_installation_libraries)

<sup>4</sup> [https://github.com/FourneyronA/Analyse-spatiale-sous-R/blob/master/R\\_1\\_Librairie\\_BDD.R](https://github.com/FourneyronA/Analyse-spatiale-sous-R/blob/master/R_1_Librairie_BDD.R)

<sup>5</sup> [https://github.com/FourneyronA/Analyse-spatiale-sous-R/blob/master/R\\_2\\_integration\\_donnees\\_ff.R](https://github.com/FourneyronA/Analyse-spatiale-sous-R/blob/master/R_2_integration_donnees_ff.R)

### 2.5.1 Création / importation de nouvelles données

Afin d'analyser certaines formes de spatialité ou encore de représenter cartographiquement certains comportements, il a été nécessaire d'importer certaines données spatiales. Six nouvelles variables ont été ajoutées à la base de données.

#### 2.5.1.1 *Les points de localisation de l'observation*

Dans l'objectif de réaliser des analyses spatiales comme les matrices de distance, il a été nécessaire d'avoir les points de géométrie de l'observation. La création des points de géométrie s'est faite par requête SQL (voir annexe page 69), via les outils de PostGIS, qui ont permis de créer le point à partir des coordonnées Longitudes et Latitudes en WGS84.

#### 2.5.1.2 *Les polygones des limites administratives (source ADMIN EXPRESS 2019)*

Pour réaliser des géotraitements ou des cartes, les limites administratives allant des communes et des départements jusqu'aux régions de France était indispensable.

#### 2.5.1.3 *La typologie des unités (urbaine/rural) et la typologie des paysages de la France (source INSEE exporté par le biais de Géoclip et d'une jointure sur l'ADMIN EXPRESS)*

Ces données permettent une caractérisation du territoire français à travers deux points de vue, celui-lui de l'opposition urbain/rural qui permet de comprendre si le territoire a été transformé par l'homme et celui des paysages qui permet d'apporter un point de vue sur l'état naturel du paysage. Ces caractérisations du paysage sont réalisées à l'échelle communale.

Travaillant sur des données participatives naturalistes, une visualisation des territoires urbains et ruraux permet d'avoir une fenêtre d'analyse supplémentaire pour comprendre les contributeurs de Faune France.

Les paysages jouant un rôle particulier sur l'habitation de certaines espèces, il était aussi intéressant d'avoir cette information pour visualiser l'impact des spatialités des contributeurs par rapport aux milieux d'habitat des espèces.

Ces données ont été importées dans une table supplémentaire « com\_wgs84 » qu'il est possible de rattacher aux observations par intersection géométrique entre polygone et point d'observation ou par le code postal communal. L'importation s'est faite par l'outil PostGIS 2.0 manager qui permet l'importation de fichier Shapefile dans une base de données.

#### 2.5.1.4 *Les points de localisation des contributeurs*

La encore pour réaliser des mesures entre habitation du contributeurs et observations, il a été nécessaire d'avoir les points de géométries du lieu d'habitation des contributeurs.

La création des points de géométrie s'est faite par requêtes SQL (voir annexe page 69), via les outils de PostGIS, qui ont permis de rattacher par le code communal un contributeur aux polygones d'une commune, puis de prendre le point du centroïde de la commune pour avoir une approximation globale identiques pour chaque contributeur de son lieu d'habitation.

## 2.5.2 Nettoyage des données

### 2.5.2.1 Utilisateurs

Le projet SOFT a pour objectif d'étudier seulement les contributeurs incarnant une personne réelle ayant réalisé une observation en 2017-2018. Dans ce cadre-là, il est nécessaire de supprimer toutes les personnes non concernées. On retrouve alors les personnes observant en-dehors de l'année 2017-2018 ainsi que les utilisateurs inscrit après 2018, les utilisateurs qui utilisent le compte d'une entité collective et qui peuvent être très nombreux, ainsi que les utilisateurs fictifs qui ont été créé lors de tests par Faune France.

Afin de ne pas prendre en compte tous ses utilisateurs, Karine a pu me joindre les informations qu'elle avait obtenu de Phillippe Jourde concernant ses utilisateurs. On retrouve alors les ID :

- 0 à 6546 sont fictifs
- 127861, 128061, 128330, 127826 se sont inscrits après 2018
- Une liste de 75 comptes qui correspondent à des associations

Pour écarter ces utilisateurs des analyses, il a fallu dans un premier temps regrouper tous ces ID dans une liste. Une fois la liste réalisée, l'idée est de créer une table temporaire sur PostgreSQL comportant l'intégralité de ses ID. Puis de lancer une requête de suppression dans la table observation lorsque l'ID de l'utilisateur correspond à l'un des ID à écarter. (Voir script partie IV – Traitement sous partie 1 – suppression dans la BDD)

### 2.5.2.2 ID Lieu-dit

Lors de l'importation de données dans les tables correspondantes, une erreur est survenue fréquemment dans la table lieu\_dit. Cette erreur vient du fait que la base de données Faune France utilise un ID lieu-dit pour définir le lieu d'observation (commune, nom du lieudit, code Insee). Cependant il y a parfois une description variable qui suit le nom du lieudit.

ID lieu dit	Communes	Nom du lieudit	Code Insee	Pays
929172	Pouy-sur-Vannes	Pouy-sur-Vannes (forêt)	10301	France
929172	Pouy-sur-Vannes	Pouy-sur-Vannes (forêt nie)	10301	France

Cette erreur vient ajouter une difficulté dans la migration de données. Effectivement on peut se poser la question de la qualité des données. Est-ce que la différence entre les deux vient nous apporter une information importante ? En regardant de plus près les rares cas d'erreurs, qui se situent souvent dans une information supplémentaires textuelle, le plus simple était de supprimer des lignes en doublons.

### 2.5.2.3 Commentaires

Lors des premières lectures de fichier texte des observations de Faune France, j'utilisais la commande COPY de PostgreSQL afin de lire automatiquement les fichier texte et les transférer dans une base de données. Cependant sur certains fichiers l'importation n'a pas pu aboutir, car le champ *commentaire* ou le champ *remarque* pouvaient contenir le caractère « \ » qui était mal interprété par le logiciel (sur certains champ sa lecture était nécessaire et sur d'autres champ non).

La solution la plus efficace a été de lire ses fichiers avec le logiciel Rstudio. De plus cela a permis la réalisation d'un seul et unique script qui permet la lecture des fichiers, le traitement des données, leurs importations dans une base de données et leur transfert dans des tables adaptées.

### 3 Analyses globales

Une fois la base de données structurée, l'objectif est de décrire quantitativement les variables et leur répartition dans l'espace et le temps. Cela nous permettra de mieux interpréter les analyses axées sur les contributeurs et les phénomènes représentés.

#### 3.1 Nombre d'observations

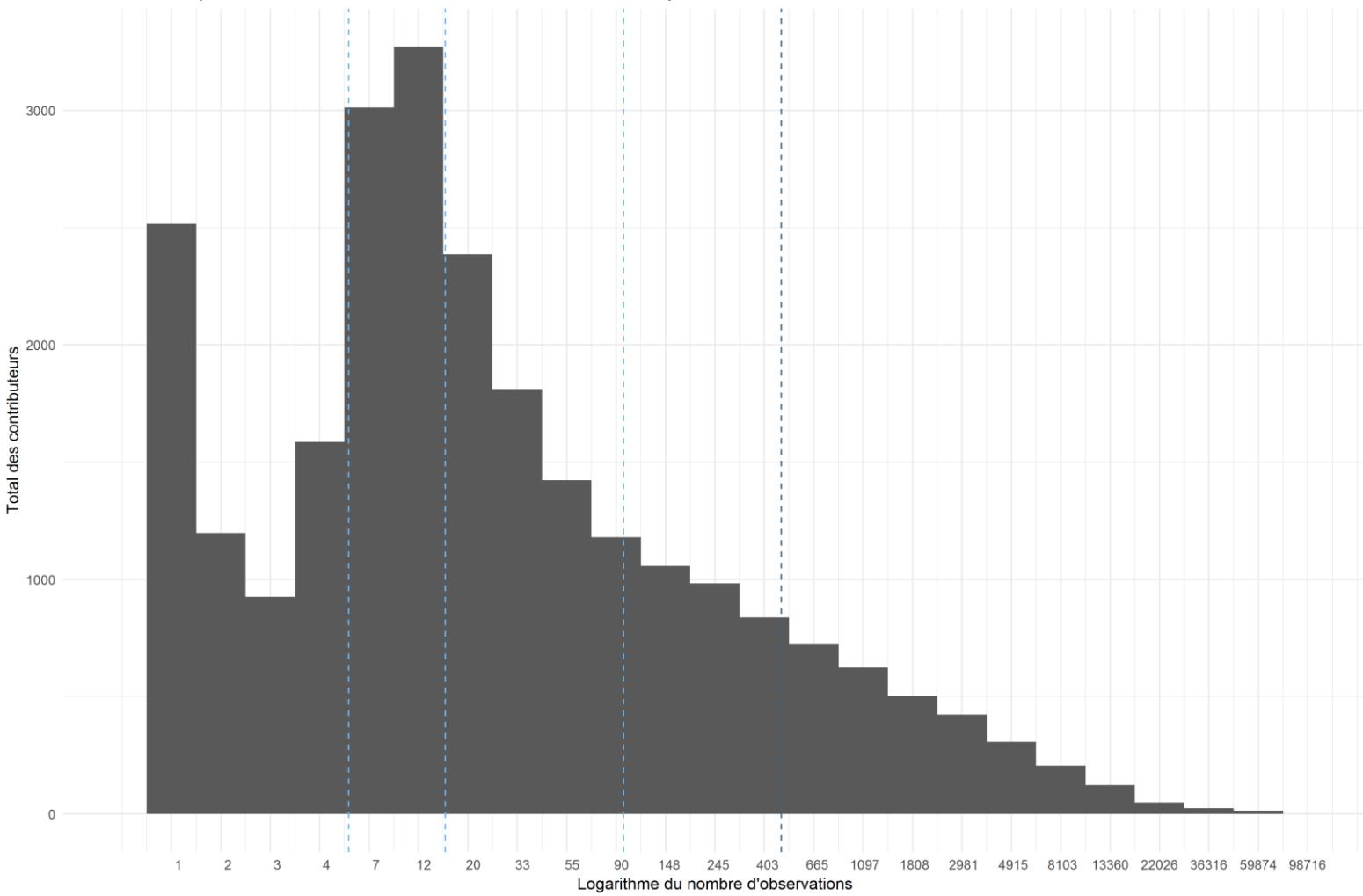
Il est primordial de créer une variable du nombre d'observations par contributeur afin de comprendre les logiques d'implication et de participation des utilisateurs. Pour cela il a fallu grouper l'intégralité des observations réalisées par l'ID des utilisateurs et compter le nombre total d'observations.

Ce qui nous donne l'histogramme si dessous, du total de 25 182 contributeurs ayant réalisé 12 069 291 observations (exprimé en logarithme pour obtenir un graphique plus lisible).

On peut voir que sur ce graphique la répartition des quantiles est proche d'une distribution exponentielle (6,16,97). 10% des personnes ont réalisé seulement une observation. De plus il y a un fort écart entre la médiane et la moyenne ce qui peut montrer une participation inégalitaire. Il est donc judicieux de se pencher vers d'autres outils comme la courbe de Lorenz et l'indicateur de Gini pour mieux percevoir la répartition du nombre d'observations par contributeur.

Participation des contributeurs de FAUNE FRANCE en 2017-18

Indicateurs statistiques sur le nombre d'observations : Q1 = 6 Médiane = 16 Moyenne = 479 Q3 = 97

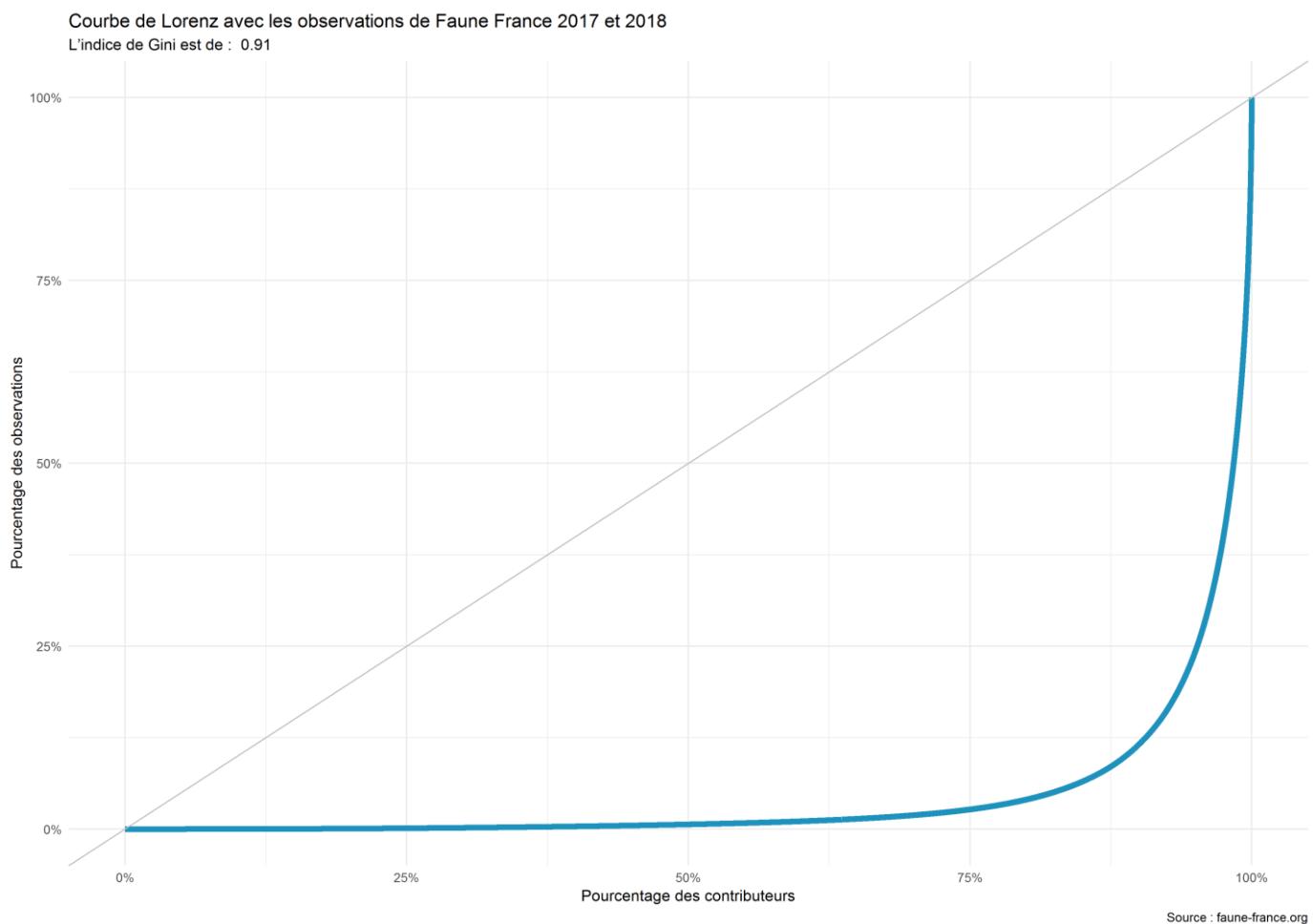


Source : faune-france.org

La courbe de Lorenz nous montre bien qu'une toute petite partie des contributeurs réalise la majeure-partie des observations. C'est un fait habituel, que l'on retrouve dans la plupart des systèmes où les participants sont volontaires (20% des participants réalisent 80% des relevés).

Ici nous retrouvons bien ce phénomène légèrement accentué avec 10% des contributeurs qui réalisent 90% des observations de Faune France.

L'indice de Gini à 91% nous montre bien que la répartition est très concentrée. On en déduit qu'il existe différents profils d'utilisateurs et différents modes de participation dans la communauté de Faune France.

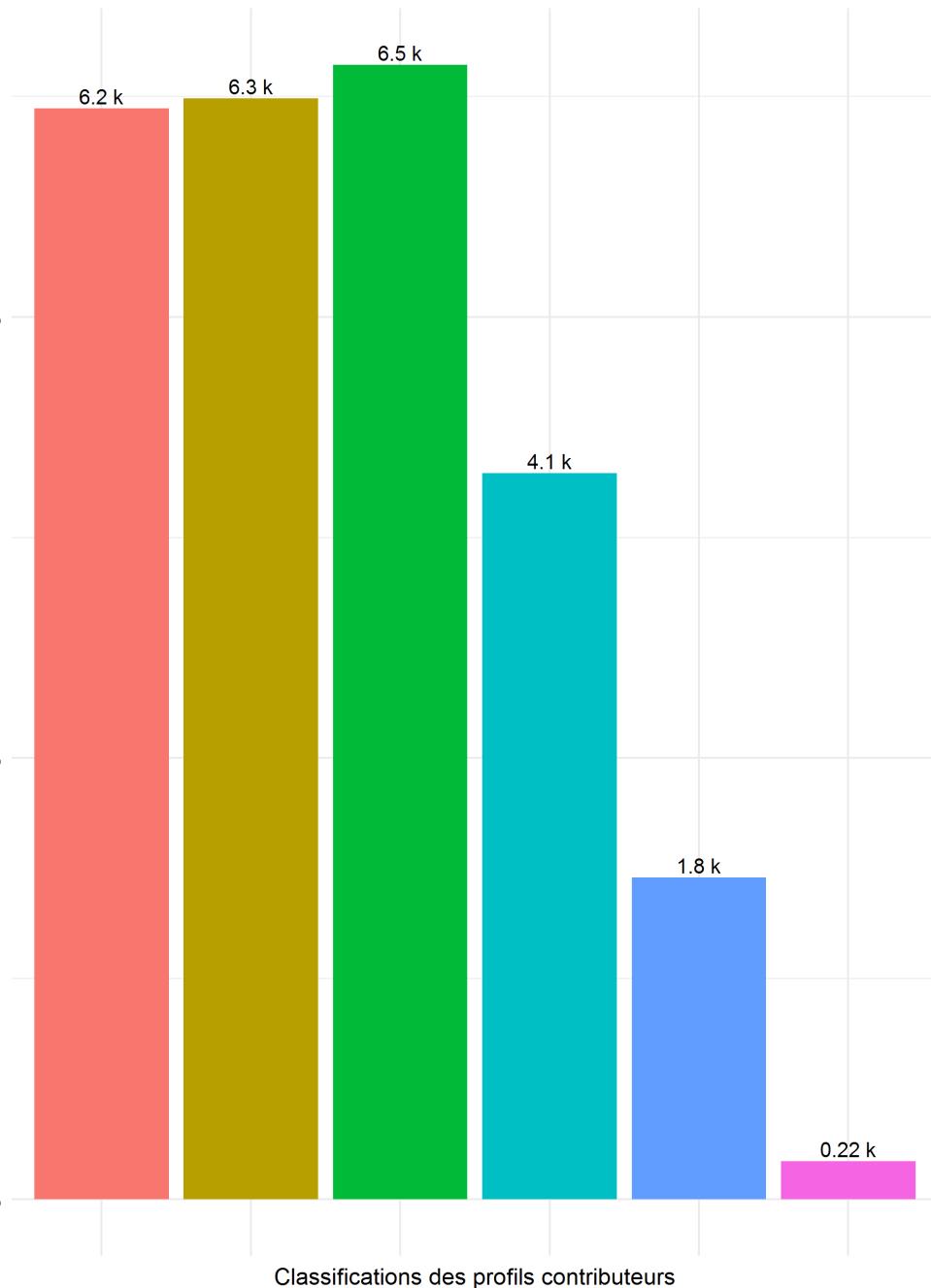


Le nombre d'observations a un effet important pour l'analyse spatiale, car celle-ci dépend directement du nombre de localisations différentes. Il semble alors judicieux de classifier différents profils pour avoir une première fenêtre de visualisation.

Un découpage basé sur les quantiles puis continué de façon exponentielle a été fait, pour respecter cette répartition dissymétrique. On a donc sur la page 19 une première classification basée sur le découpage en classes du nombre d'observations des contributeurs (0, 5, 15, 100, 1 000, 10 000)

### Répartitions des contributeurs dans une classification du nombre d'observation

Pourcentage du nombre de contributeur



Classifications des profils contributeurs

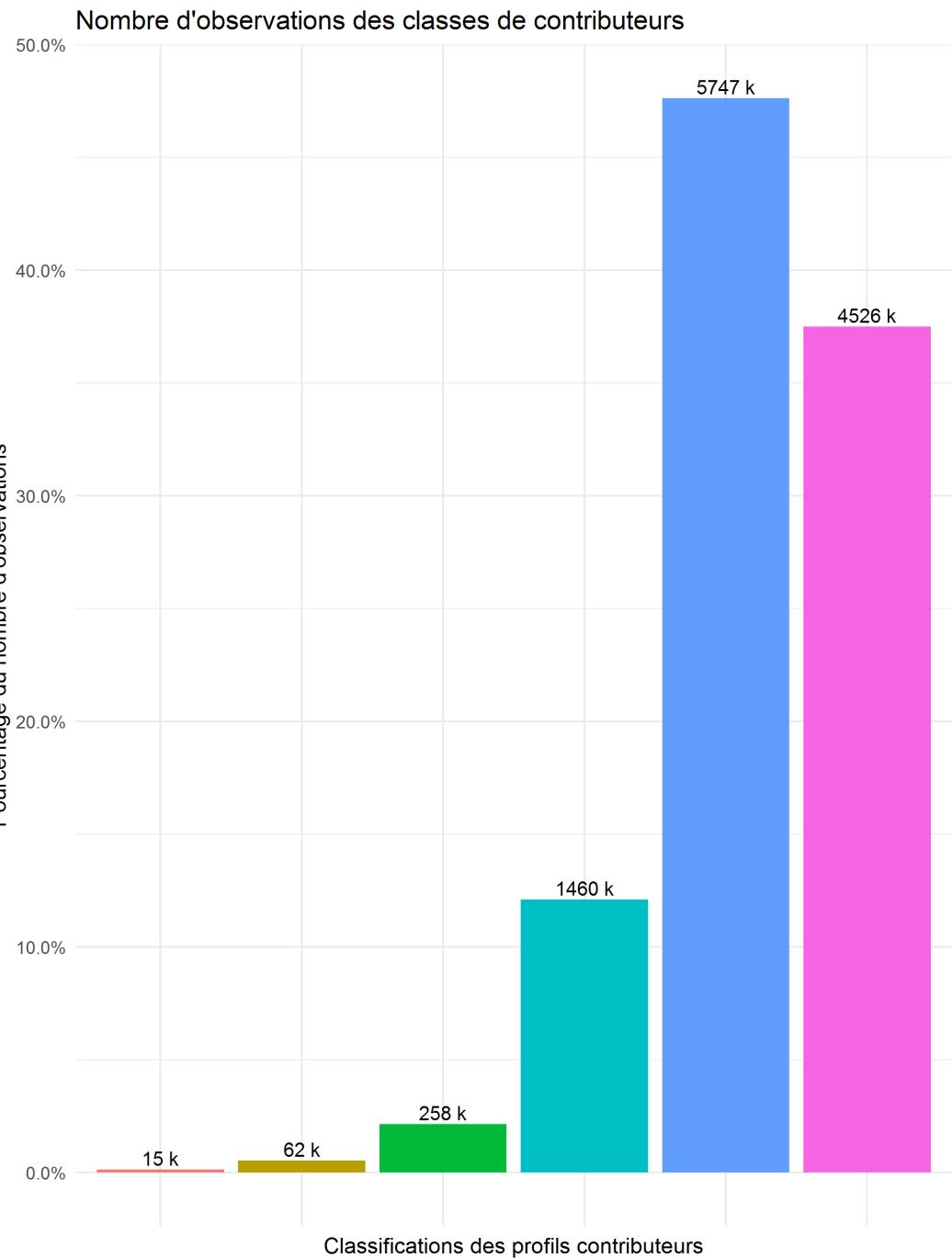
- Contributeurs Occasionnels
- Contributeurs moyens (+5)

- Bon contributeurs (+15)
- Super contributeurs (+100)

- Pro contributeurs (+1000)
- Ultra contributeurs (+10000)

### Nombre d'observations des classes de contributeurs

Pourcentage du nombre d'observations



Source : faune-france.org

## 3.2 Observations spatiales générales

### 3.2.1 Répartition générale des observations et des contributeurs

Dans l'objectif d'avoir un premier aperçu de la répartition spatiale des données d'observation et des contributeurs, j'ai réalisé des cartes à l'échelle départementale accompagnées de graphiques statistiques par régions. Cet aperçu nous permet d'avoir une idée globale de la couverture spatiale sur toute la France, ainsi que les zones les plus observées et ayant le plus de contributeurs, ainsi qu'une première idée sur leur concentration.

#### 3.2.1.1 *Nombre de contributeurs par Région*

On peut voir d'après le document des pages 21-22 plusieurs phénomènes :

1. La région Auvergne-Rhône-Alpes recueille le plus de contributeurs (6000), avec une répartition qui reste relativement homogène.
2. Les régions Île-de-France et Provence-Alpes-Côte-D'azur ont une grande concentration de contributeurs par km<sup>2</sup>, mais un nombre de contributeurs moyen. Pour la région Île-de-France cela s'explique par la concentration de population dans Paris et sa banlieue.
3. L'ouest de la France regroupe une forte zone de contributeurs dans trois régions : Bretagne, Pays de la Loire et Aquitaine-Limousin-Poitou-Charentes. On retrouve ce phénomène dans l'Est avec les régions Alsace-Champagne-Ardenne-Lorraine et Bourgogne-Franche-Comté.

Il est maintenant intéressant de comprendre si cette distribution des contributeurs se traduit en nombre d'observations

#### 3.2.1.2 *Nombre d'observations par Région*

On peut voir d'après le document des pages 23-24 deux phénomènes :

1. Tout d'abord une relation apparente entre le nombre de contributeurs et la présence d'observations sur le territoire notamment avec les concentrations. Une forte présence est notable dans les régions Auvergne-Rhône-Alpes, Île-De-France, Aquitaine-Limousin-Poitou-Charentes et Provence-Alpes-Côte d'Azur.
2. Certaines régions entières n'ont quasiment pas d'observations comme les Hauts-De-France<sup>6</sup>, la Normandie voire le Centre-Val de Loire. Ce qui peut être dû à une non-communication des données de niveau régional à Faune France.
3. Certains départements ont eux aussi très peu d'observations comme les Hautes-Pyrénées ou le Gers dans une région qui pourtant regroupe de nombreux observateurs.

Nous obtenons donc une répartition très inégale des observations au niveau départemental avec notamment un fort impact sur les régions septentrionales. Cela peut être mis en relation soit avec une présence plus faible d'oiseaux, mais peut s'expliquer aussi par le fait que Faune France n'a pas pu collecter toutes les données locales voir régionales.

**Cette information d'un possible manque de données sur certaine région vient augmenter l'importance de la réutilisation et de la reproductivité des scripts R sur l'analyse spatiale qui est un enjeu de tailles pour le projet SOFT.**

---

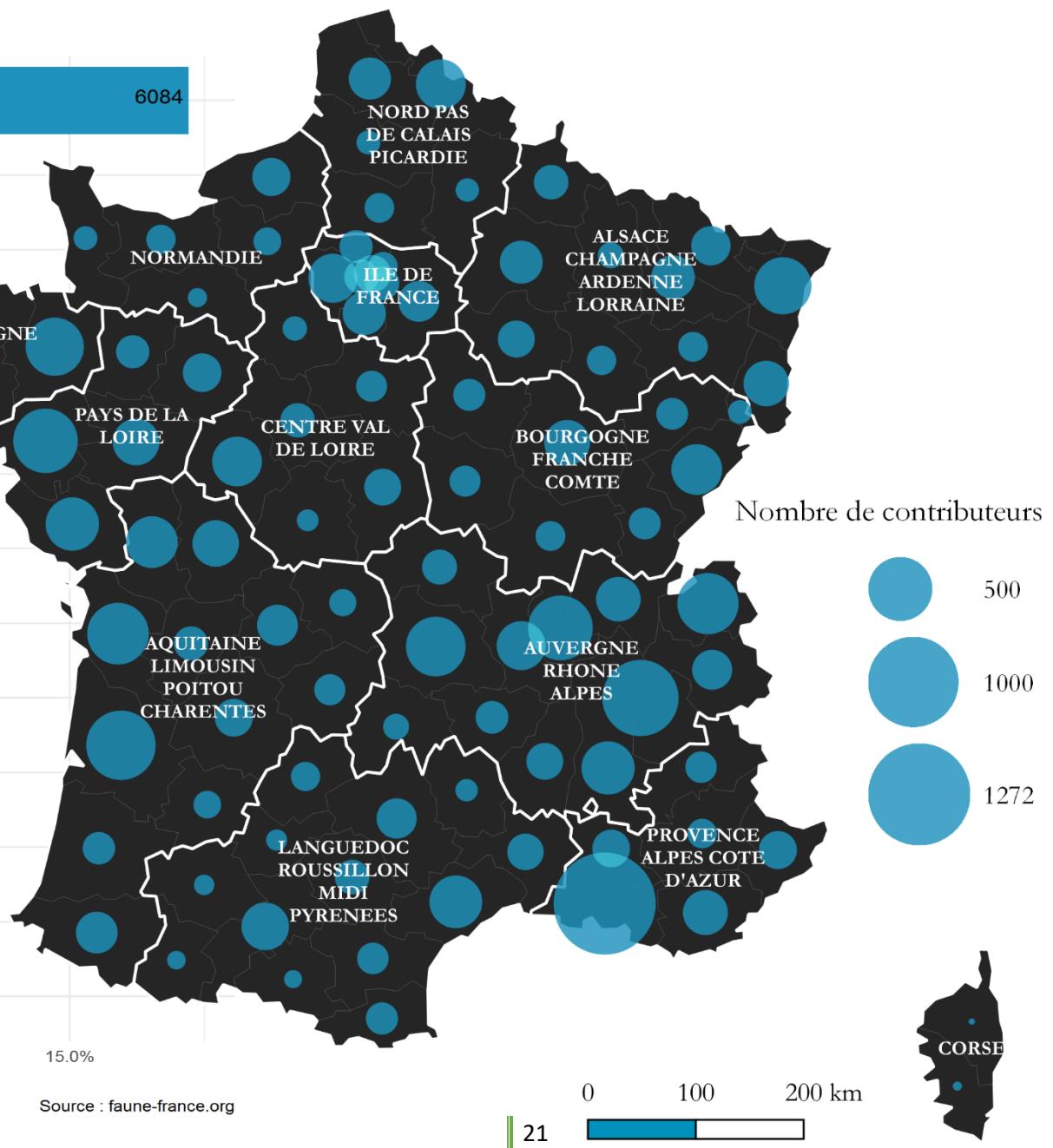
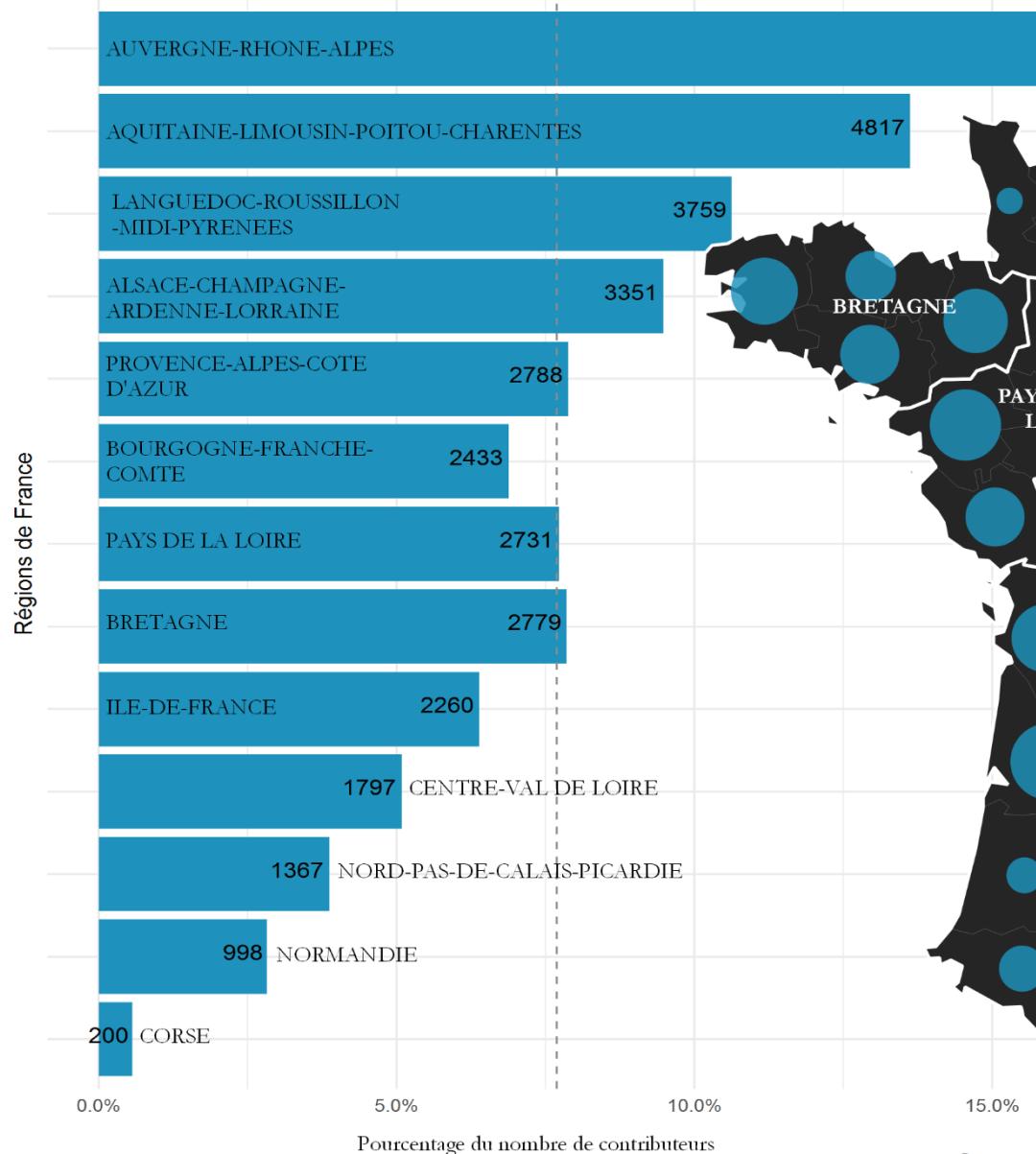
<sup>6</sup> Encore appelée Nord Pas De Calais Picardie dans la plupart des bases de données d'où la différence d'appellation sur les cartes.

# Nombre de contributeurs 2017-2018 à l'échelle départementales et régionales

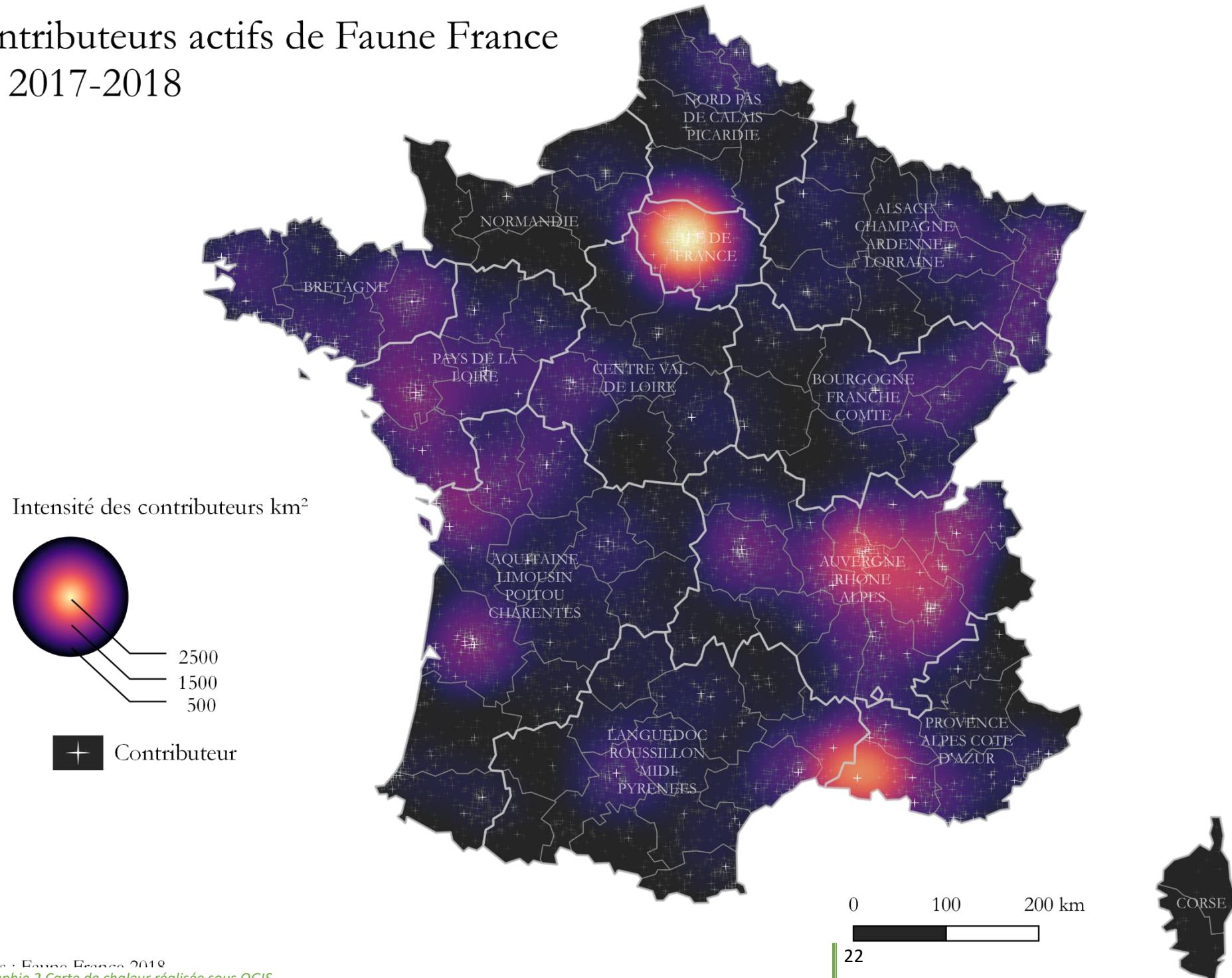
Répartition régionale des contributeurs de Faune France en 2017-18

Moyenne des contributeurs par régions : 2720

Moyenne des contributeurs par départements : 324



# Contributeurs actifs de Faune France en 2017-2018

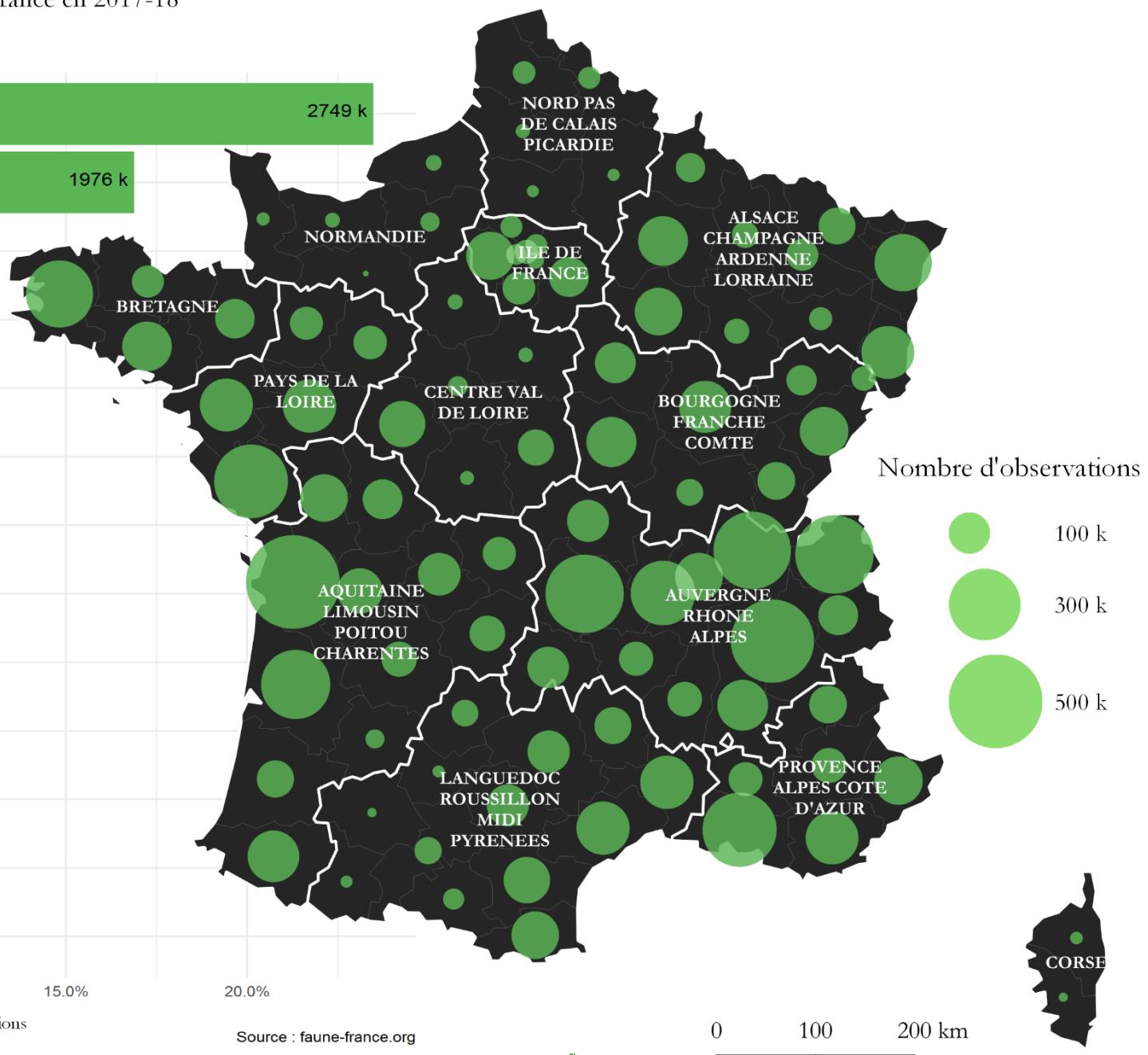
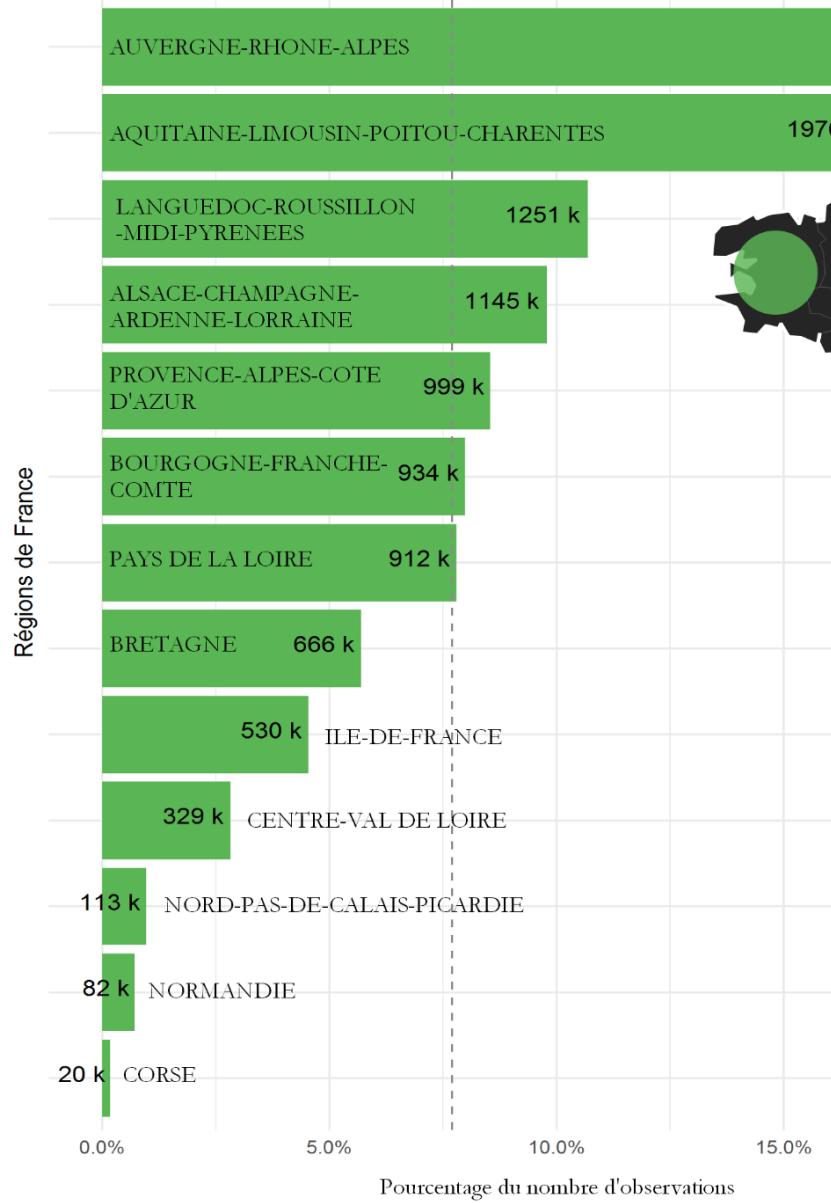


# Nombre d'observations 2017-2018 a l'échelle départementales et régionales

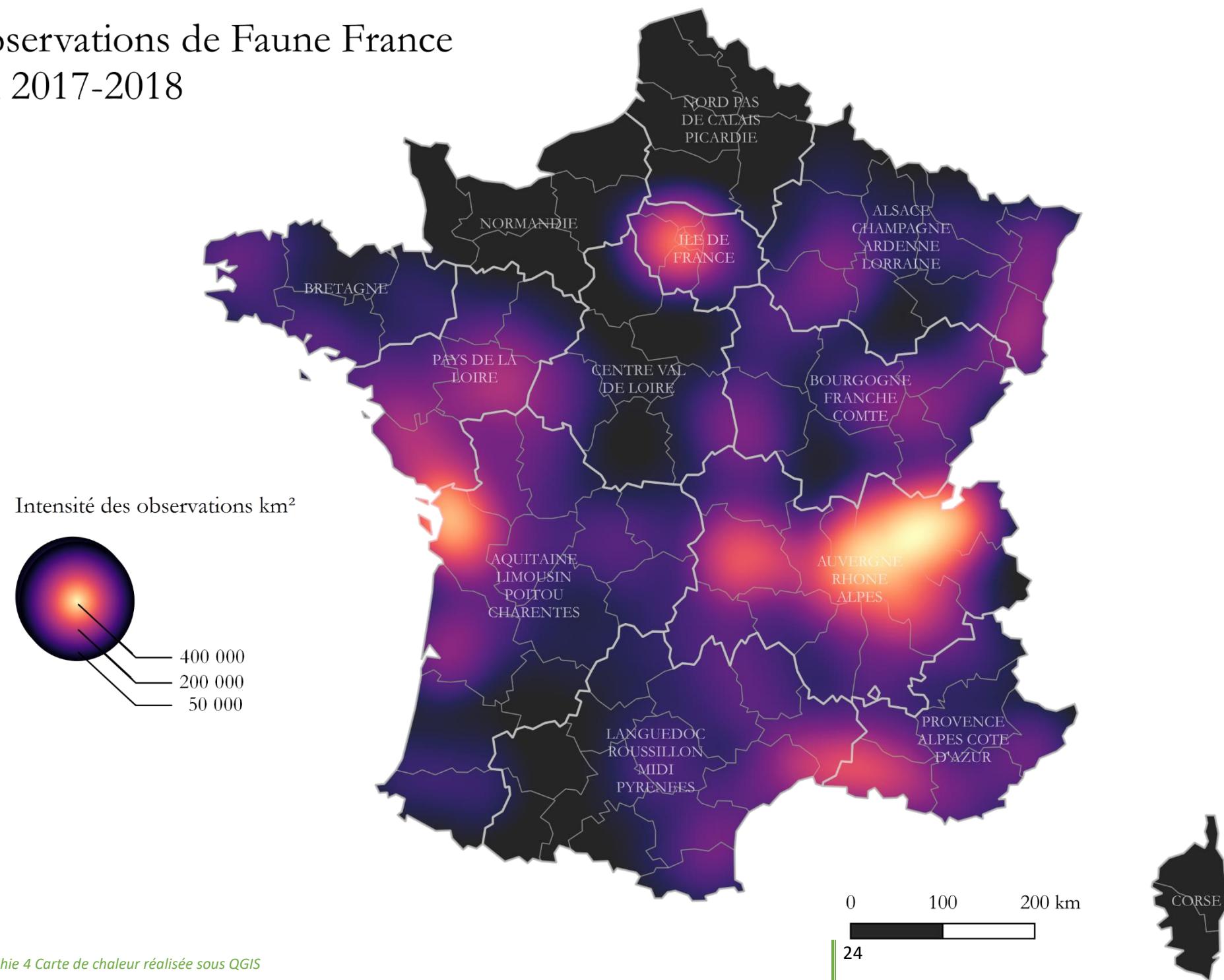
Répartition régionales des observations de Faune France en 2017-18

Moyenne des observations par régions : 900 k

Moyenne des observations par départements : 122 k



# Observations de Faune France en 2017-2018



### 3.2.2 Contributeurs habitants et visiteurs

Grâce à la base de données, nous connaissons la commune d'habitation des contributeurs. Nous connaissons aussi le lieu de l'observation et le contributeur associé à l'observation. Ce qui nous permet de connaître à l'échelle départementale et régionale le nombre d'observations faites par des contributeurs habitant le lieu et des contributeurs non-habitants (visiteurs) ainsi que le nombre de contributeurs habitants et visiteurs par département et région.

Les cartes de la page 26 et 27 représentent le taux d'observations/contributeurs visiteurs par rapport au total. Ils sont calculés de la manière suivante :

$$\text{Taux d'observations visiteurs} = \frac{\text{Nombre d'observations réalisées par des visiteurs}}{\text{Nombres d'observations totales}}$$

$$\text{Taux de contributeurs visiteurs} = \frac{\text{Nombre de contributeurs visiteurs}}{\text{Nombres de contributeurs totales}}$$

Grâce à ses taux situé entre 0 et 1 nous savons que le nombre d'observation/contributeurs est majoritairement visiteur s'il dépasse les 0,5. La sémiologie appliquée est alors basée sur un double dégradé allant du plus clair (0.5) au plus foncé (0 ou 1) avec une représentation rose pour les habitants et bleu pour les visiteurs.

Attention ces cartes ne sont pas aisées de lecture, elles représentent un phénomène particulier lié au choix de l'échelle. Ce phénomène se ressent plus sur la comparaison entre les contributeurs habitants et visiteurs.

Effectivement on peut voir que à l'échelle départementale, il y a une majorité de départements qui ont plus de contributeurs visiteurs (plus de 50% des personnes du département sont des contributeurs visiteurs), alors que à l'échelle régionale ce sont les contributeurs habitants qui sont largement majoritaires.

Cela montre qu'il y a une certaine tendance des contributeurs à se déplacer à l'échelle régionale dans quelques départements. Cependant il reste difficile d'interpréter à travers ces cartes si les déplacements sont plutôt liés aux régions d'appartenance ou aux départements limitrophes. On ne distingue pas les contributeurs qui se déplacent sur l'intégralité de la France de ceux qui ne se déplacent pas.

Pour ce qui est des observations, la plupart des départements ont un très fort taux d'observation réalisé par les habitants du département lui-même, le poids des contributeurs visiteurs est donc très faible sur la totalité des observations. Cela peut être lié à la motivation ou au profil des contributeurs visiteurs qui n'ont pas pour objectif de réaliser beaucoup d'observation ou simplement le temps de réaliser plus d'observations sur le territoire.

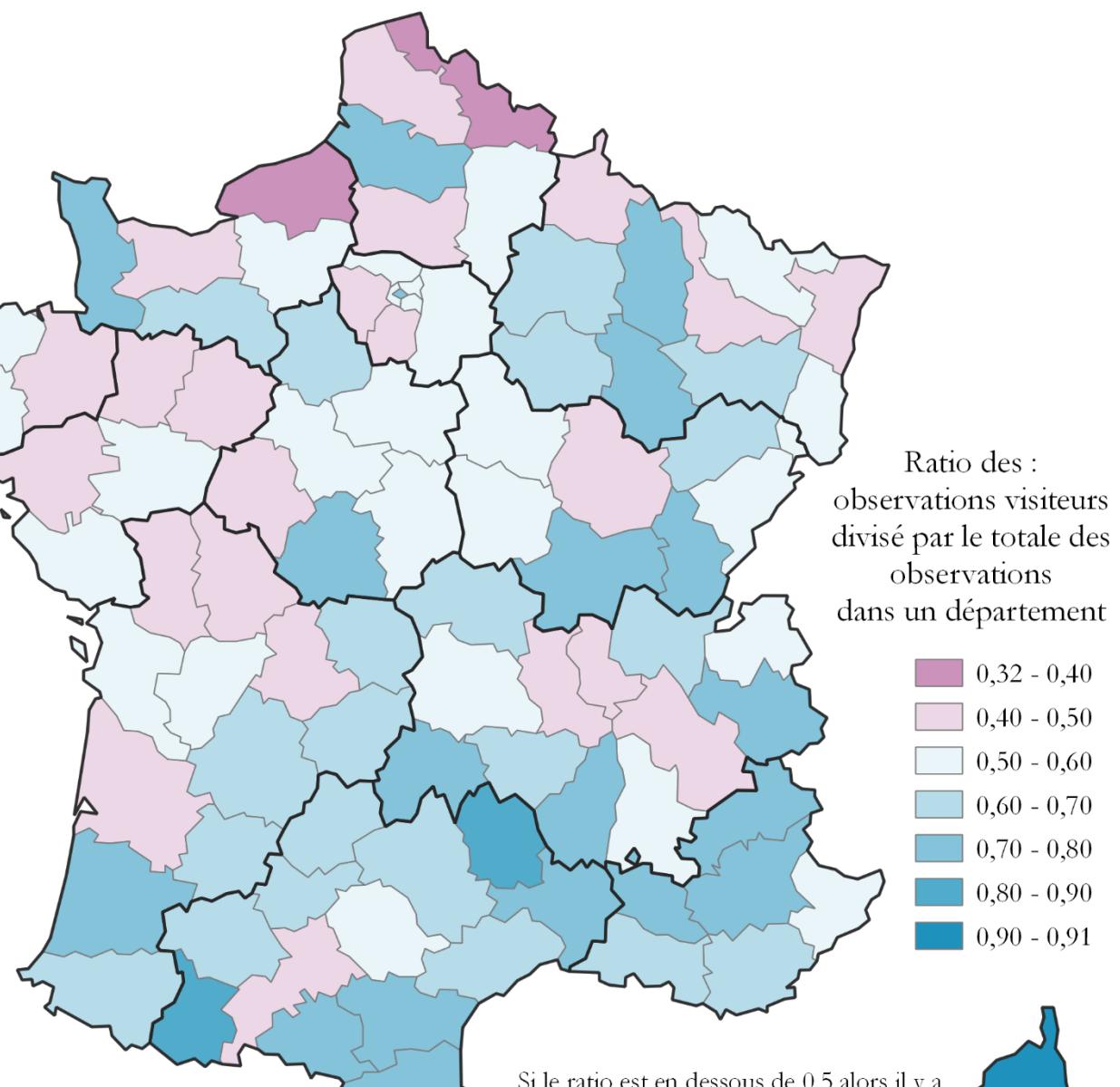
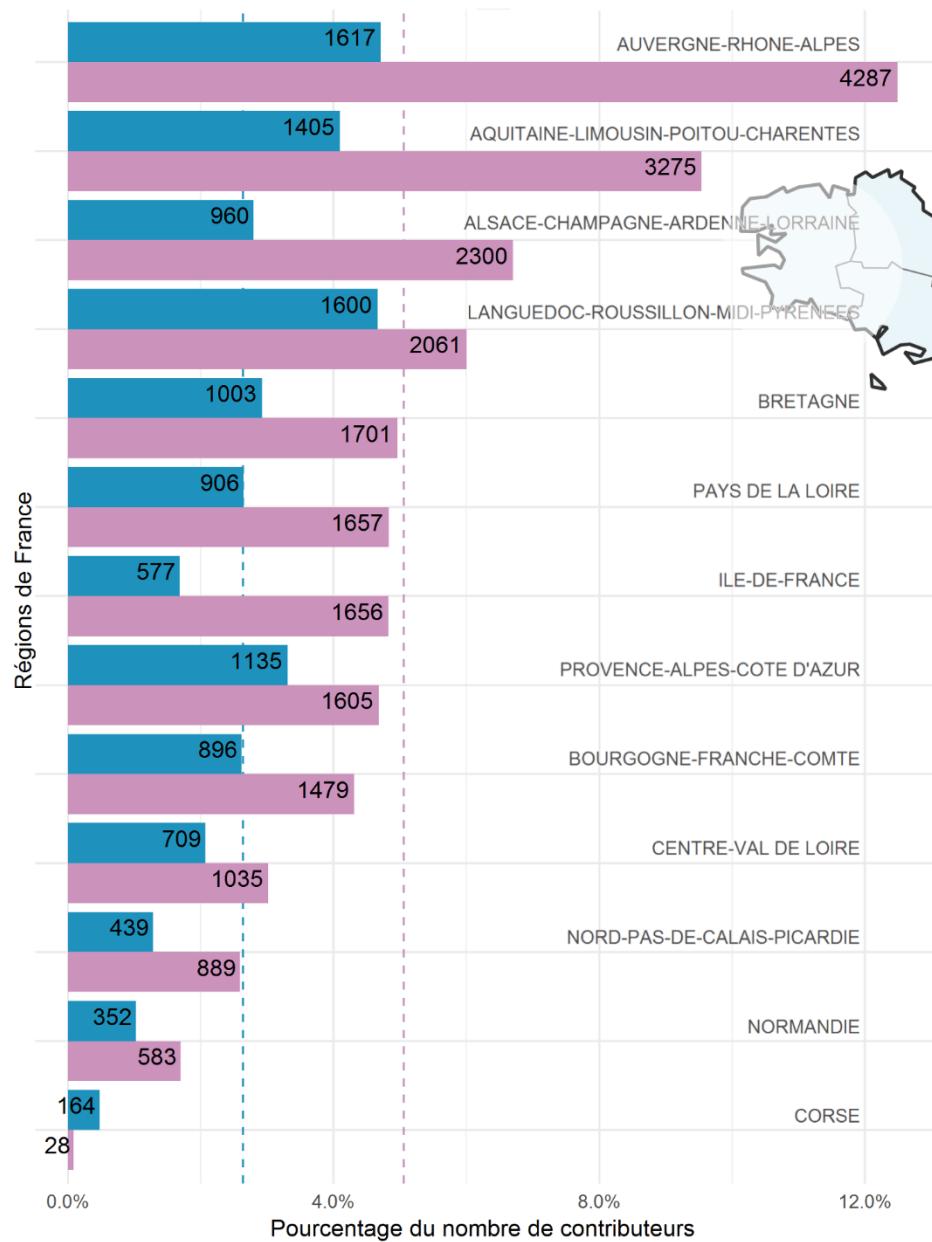
L'enjeux des déplacements des contributeurs est donc une piste à creuser pour mieux interpréter les comportements et motivations des contributeurs à travers une analyse spatiale plus approfondie.

# Taux de contributeur habitants-visiteurs a l'échelle départementales et régionales

Type de contributeur de Faune France en 2017-18

Moyenne des contributeurs visiteurs : 904

Moyenne des contributeurs habitants : 1735

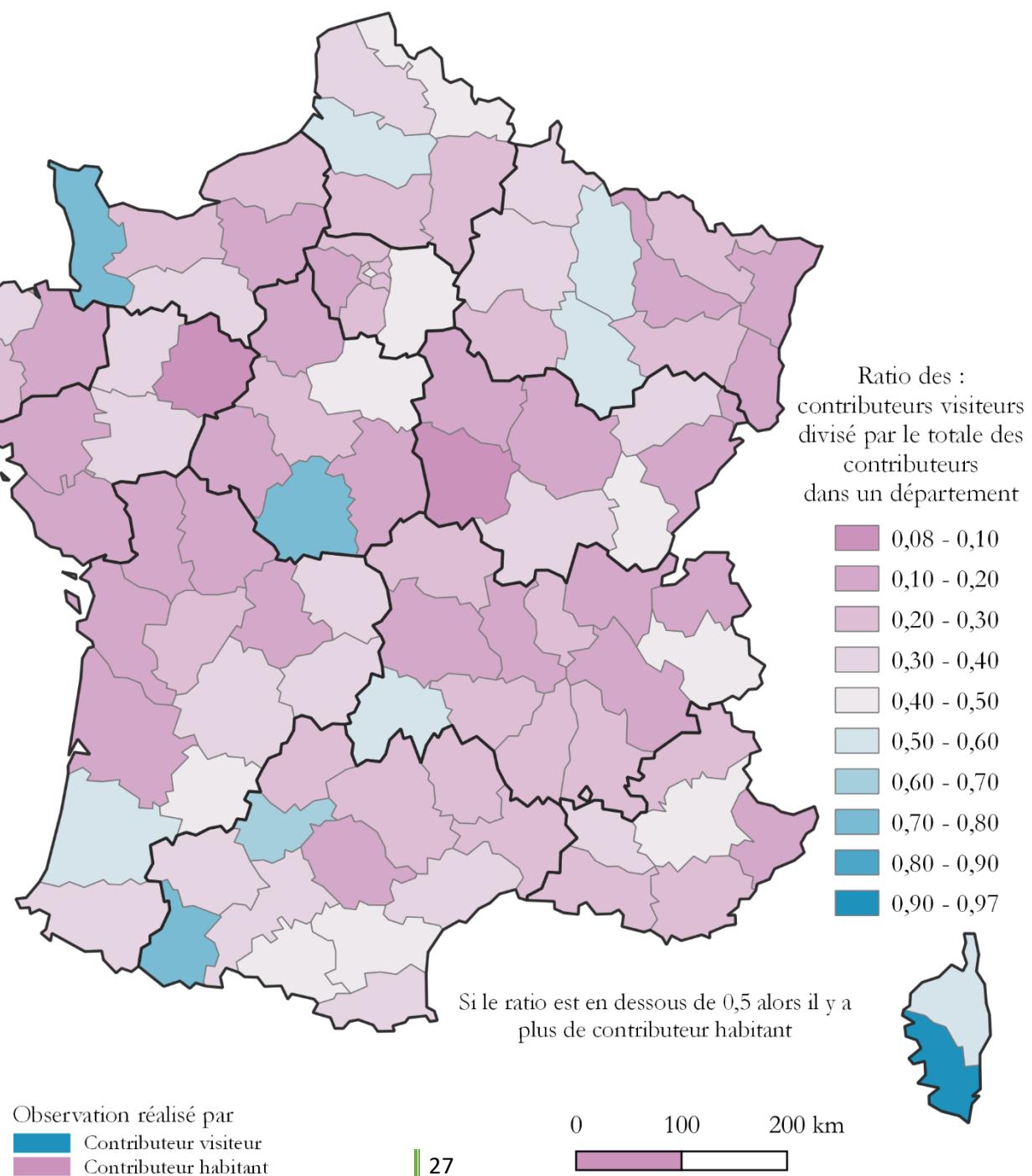
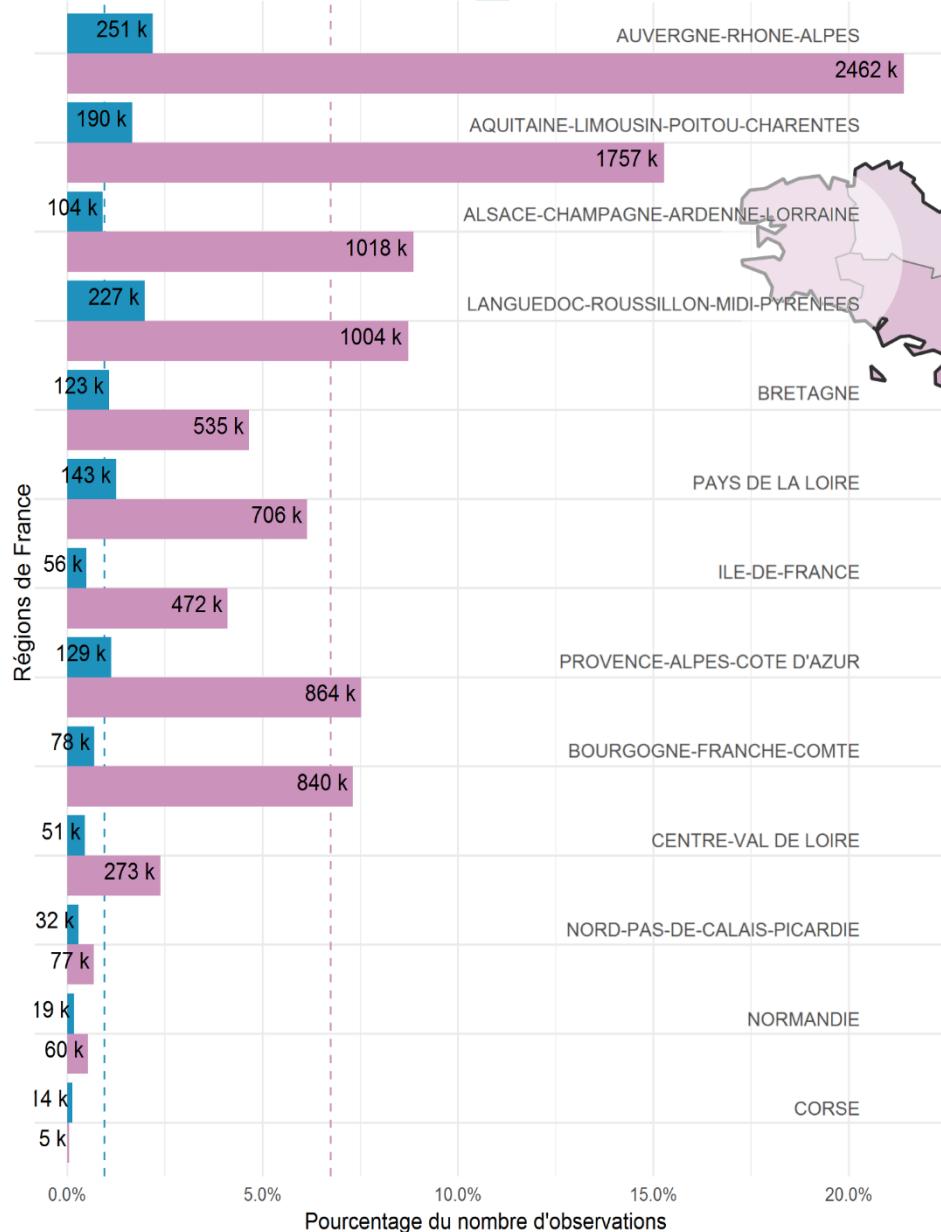


# Taux d'observations habitants-visiteurs a l'échelle départementales et régionales

Type d'observations de Faune France en 2017-18

Moyenne des observations visiteurs : 109 k

Moyenne des observations habitants : 775 k



### 3.3 Analyse temporelle

Dans les analyses suivantes nous avons regardé le nombre d'observations pour chaque jour de l'année, en juxtaposant des regroupements par saisons, mois et semaines, afin de mieux interpréter l'activité des contributeurs. Il faut garder à l'esprit que nous étudions seulement les années 2017-2018 et qu'il faudrait vérifier la répartition chronologique sur plusieurs années successives et vérifier l'évolution entre les années.

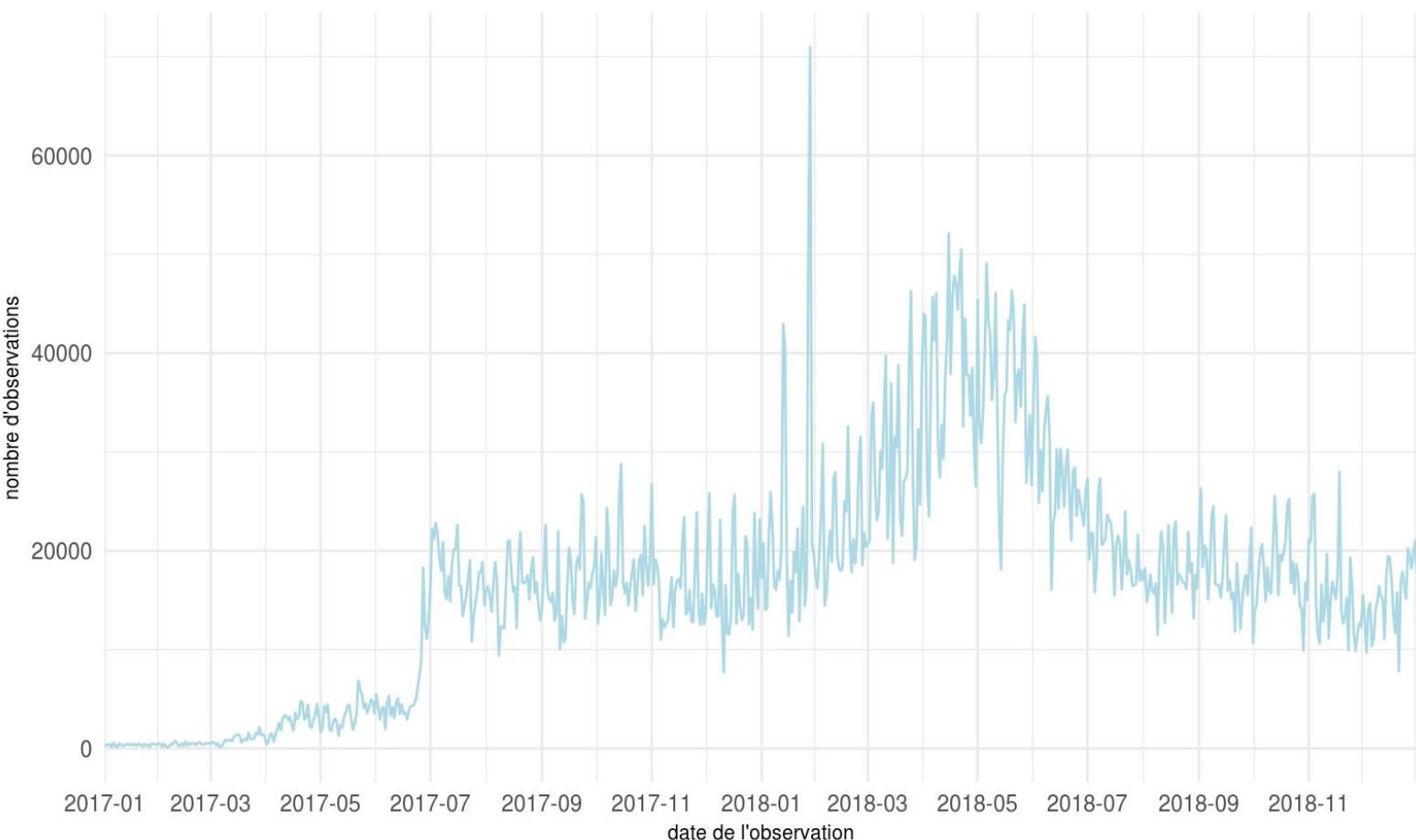
Chaque analyse temporelle (saisons, mois, jours de la semaine) est présentée en graphique Boîte et moustaches afin d'avoir une vision plus large du comportement. Les graphiques en barre de ces analyses sont en annexe dans la partie 70.

#### 3.3.1 Nombre d'observations par jours pour les années 2017 et 2018

A travers ce graphique nous pouvons voir qu'il y a une différence entre l'année 2017 et 2018 assez importante sur le nombre d'observations. Notamment sur la première période de 2017 (janvier-juillet) où le nombre d'observation reste très bas. Nous pouvons aussi voir durant l'année 2018 une période d'observations plus importante entre mars et juin. Ainsi que 2 pics d'observation vers janvier 2018.

Ces différences temporelles peuvent témoigner du lancement de Faune France, il serait donc intéressant d'avoir le début de l'année 2019 (janvier-juin) pour savoir si les pics de janviers sont redondants et si les remontées d'observation entre mars et juin sont annuelles ou caractérisent l'année 2018.

**Nombre d'observations réalisé par jour**



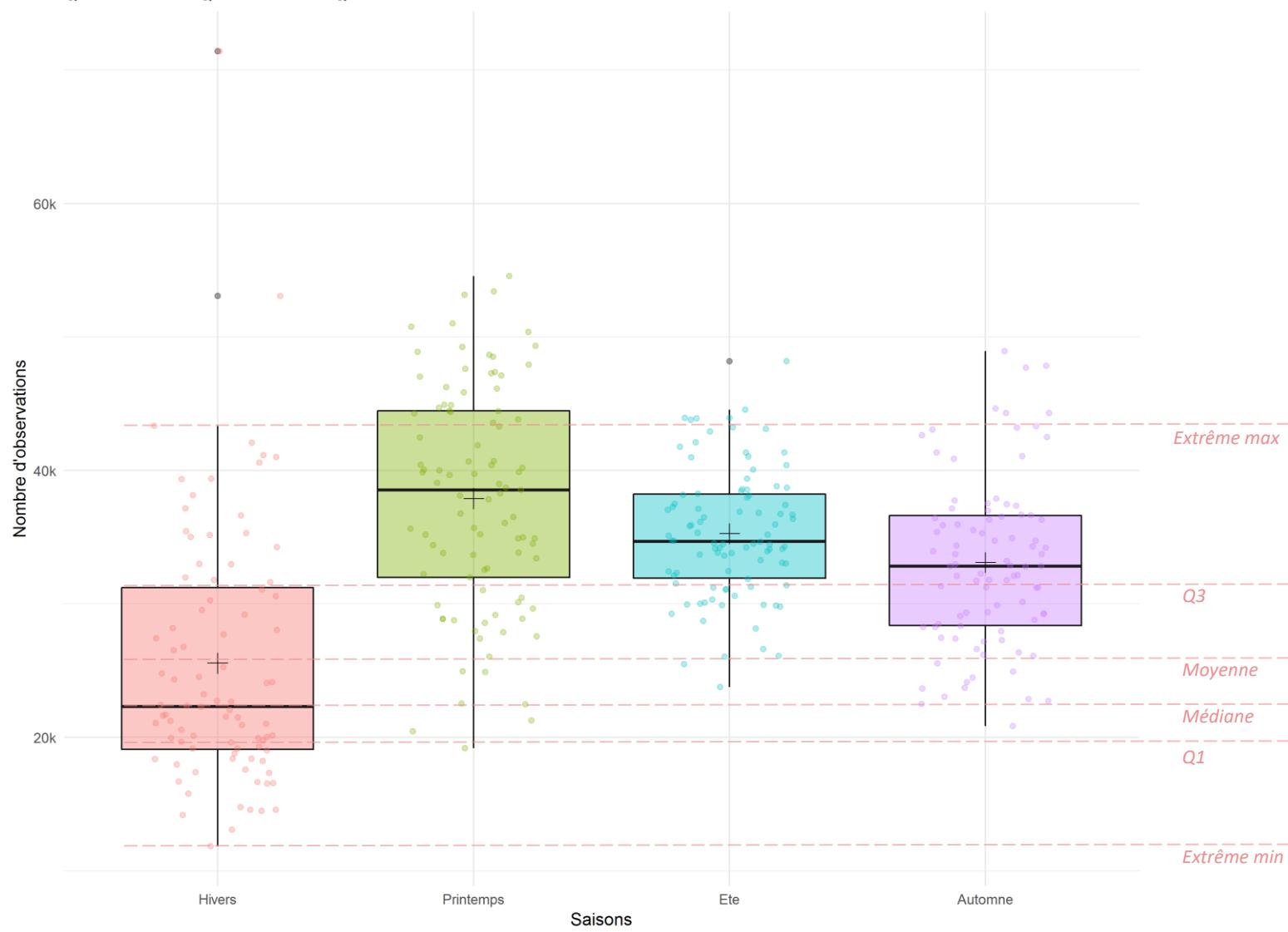
*Graphique 1 Histogramme temporel*

### 3.3.2 Les saisons d'observation

Ce boxplot représente le nombre d'observations réalisées chaque jour (1 point représente 1 jour de l'année) regroupées par saison ce qui permet de visualiser pour chaque saison le Q1, Q3, l'étendue, la médiane, la moyenne et les extrêmes (indiqués dans la légende pour l'hiver). On peut voir que la moyenne journalière des observations par saison se situe entre 32 000 et 38 000 hormis l'hiver qui a une moyenne nettement inférieure située vers les 25 000 observations par jour. Cependant cette saison a aussi deux valeurs extrêmes très importantes qui sont liées à un événement national de la LPO et du Muséum national d'Histoire naturelle (Oiseaux des Jardins<sup>7</sup>)

Nombre d'observations par saison sur 2017-18

La moyenne d'observations est de 3017 k  
 Q1 = 2796 k      Q2 = 3147 k      Q3 = 3368 k



Graphique 2 Boite à moustache des saisons

Source : faune-france.org

<sup>7</sup> Cet événement consiste à inciter toutes les personnes à un niveau national à recenser les oiseaux de leurs jardins, balcons parc.

<https://www.oiseauxdesjardins.fr/index.php>

### 1.1.1 Les mois d'observation

Si l'on regarde le nombre d'observations par jour groupé dans les mois de l'année, on s'aperçoit là aussi que ce sont très nettement les mois de printemps (fin mars, avril, mai, juin) qui ont nettement plus d'observations que les autres mois.

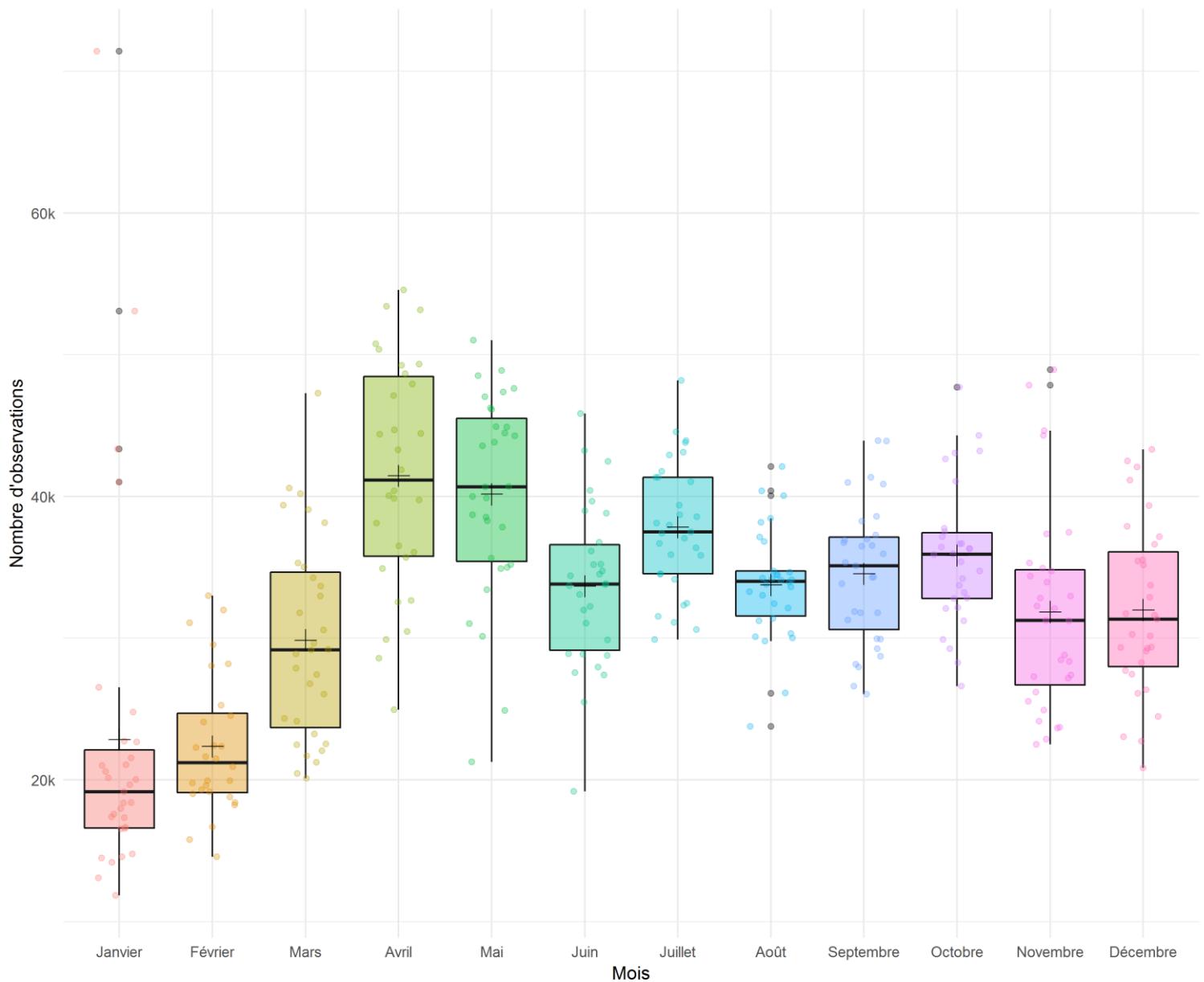
Le mois de janvier montre encore des valeurs extrêmes qui sont liées à l'événement national de Faune France

Dans l'ensemble il est possible que ce pic sur les mois d'avril, mai et juin s'explique surtout par une période plus active de la vie animale (activité, reproduction ...) plus propice à l'observation. Les congés d'été ne correspondent pas à une plus forte intensité d'observations.

Cette distinction pourrait peut-être nous permettre de différencier deux profils d'observateurs, ceux qui utilisent l'application en ayant conscience de la vie de la faune (les « professionnels ») et ceux qui utilisent l'application par divertissement suite à un événement organisé ou au hasard de leurs sorties.

Nombre d'observations par Mois sur 2017-18

La moyenne d'observations est de 1005 k  
 Q1 = 971 k      Q2 = 1011 k      Q3 = 1099 k



Source : faune-france.org

Graphique 3 Boite à moustache mensuelle

### 1.1.2 Les semaines d'observation

Si l'on regarde le nombre d'observations par jour regroupé dans les jours de la semaine, on peut voir que les observations tout au long de la semaine sont plutôt bien réparties. Il s'agit peut-être d'une bonne répartition de profils entre les personnes qui utilise l'application dans un but professionnel (jours de semaines) et ceux qui l'utilisent par passion (temps libre souvent associé au Samedi-Dimanche).

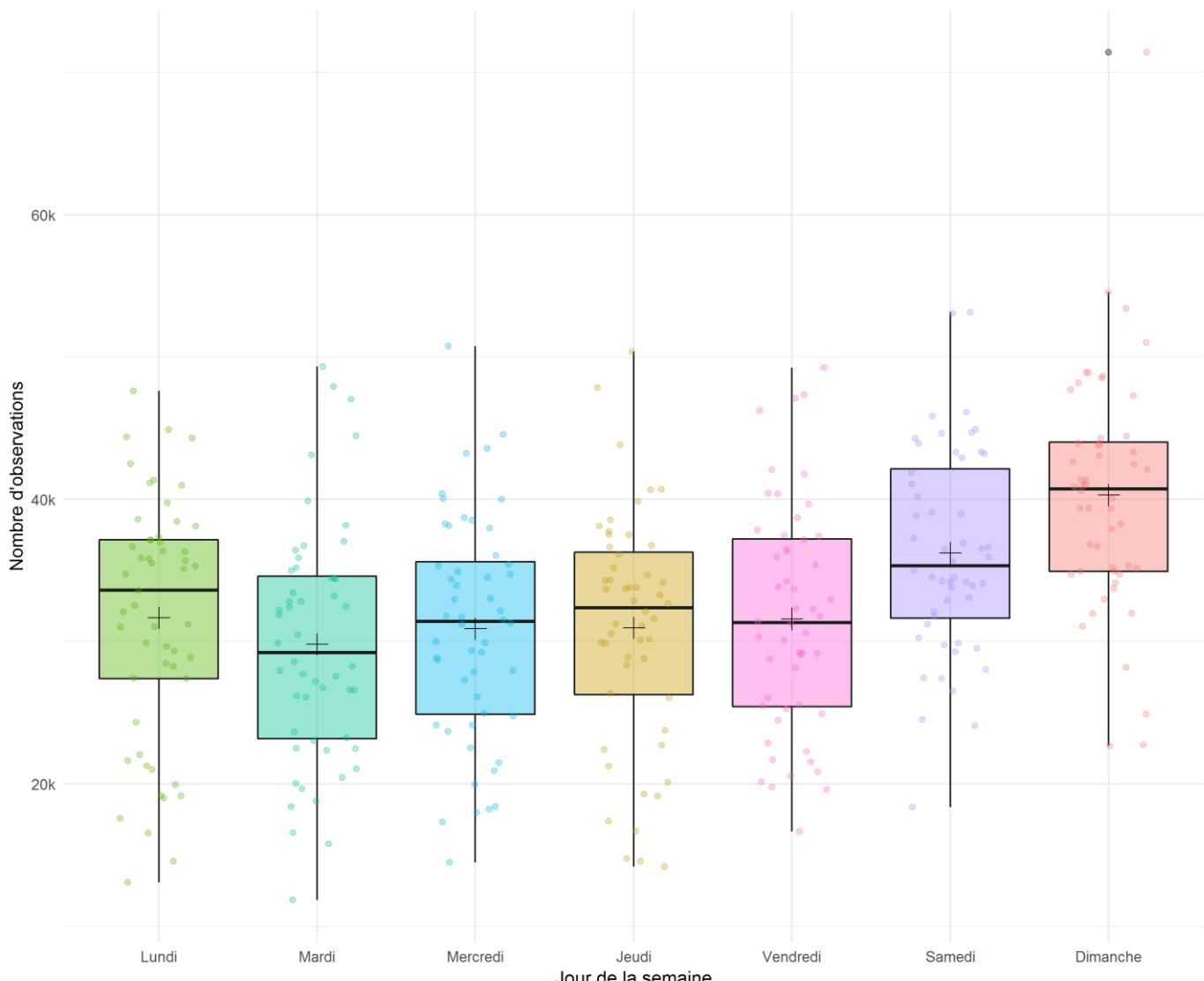
Il y a cependant un peu plus d'observations le week-end, ce qui peut être lié aux personnes très motivées qui apportent beaucoup d'observations et qui auraient plus de temps libre en fin de semaine. On pourrait alors retrouver quatre types de profil des contributeurs :

- Les « professionnels » qui travaillent beaucoup les lundi mardi, mercredi, jeudi et vendredi
- Les passionnés très volontaires qui travaillent principalement les dimanche et samedi et un peu moins le reste de la semaine
- Les passionnés qui travaillent uniquement les samedi et dimanche
- Les amateurs qui travaillent de façon aléatoire en fonction de leurs disponibilités

Cette analyse devrait être refaite avec un traitement or les observations Oiseaux des jardins, ce qui donnerait des résultats peut être différents, car ces observations souvent cumulées sur deux jours peuvent influencer le reste des observations faites tout au long de l'année.

Nombre d'observations dans les jours de la semaine sur 2017-18

La moyenne d'observations est de 1724 k  
 Q1 = 1608 k      Q2 = 1643 k      Q3 = 1781 k



Source : faune-france.org

### 3.4 Classifications des contributeurs

Pietropaoli Karine a réalisé à la mi-juin 2019 une analyse factorielle pour distinguer différents comportements des contributeurs. Cette analyse factorielle est principalement axée sur le nombre d'observations, les différentes variables qualitatives, le type des espèces observés, les jours et temps d'observation dans l'année ainsi que d'autres variables qu'elle a présentées dans une ébauche de rapport<sup>9</sup>. Cette analyse distingue sept typologies de contributeur :

Classes d'observateur	Pratiques d'observation
C1 observateurs de papillons	Activité récente et ponctuelle concentrée en été, diversité taxinomique (notamment observation de papillons très sur-représentée), très localisée, peu contributeurs
C2 observateurs oiseaux du printemps	Activité récente et ponctuelle très concentrée sur l'observation des oiseaux au printemps, très localisée, peu contributeurs, participation à des protocoles de suivi
C3 observateurs réguliers	Activité ancienne et régulière, amateurs "éclairés", diversité taxinomique et des espèces observées, pratique locale (échelle régionale)
C4 observateurs oiseaux d'automne	Activité récente et ponctuelle très concentrée sur l'observation des oiseaux à l'automne, très localisée, peu contributeurs
C5 observateurs "oiseaux des jardins"	Activité récente très ponctuelle concentrée autour du WE "oiseaux des jardins" de janvier, contribue fortement au pic des contributions de ces jours-là par leur nombre, très localisée, néo-observateurs peu contributeurs
C6 consigneurs de rareté <sup>8</sup>	Activité ponctuelle (mais relativement ancienne) de signalements d'espèces rares, relative diversité taxinomique, très localisée, très peu contributeurs
C7 experts	Activité ancienne, régulière et intensive, sur tout le territoire, grande diversité taxinomique et des espèces observées

Source du tableau : [https://analytics.huma-num.fr/Karine.Pietropaoli/Faune%20France/typologie\\_FF.html#utilisation-de-la-base](https://analytics.huma-num.fr/Karine.Pietropaoli/Faune%20France/typologie_FF.html#utilisation-de-la-base)

De mon côté j'ai aussi regardé le comportement global des différentes variables afin de mieux comprendre les pratiques des contributeurs. Ce premier regard permet de mieux interpréter les analyses plus approfondies.

La typologie résultant de l'analyse factorielle de Pietropaoli Karine m'a facilité l'analyse des comportements spatiaux. Il a fallu dans un premier temps comprendre comment se comportent les différents groupes, notamment en terme de nombre d'observations qui joue un rôle très important dans l'analyse des comportements spatiaux. Effectivement il n'est pas possible d'analyser spatialement de la même manière une personne ayant réalisé 500 observations avec une personne qui n'en a fait que 5. Plus d'explications sont données dans la partie Méthodologie d'analyses spatiales 36)

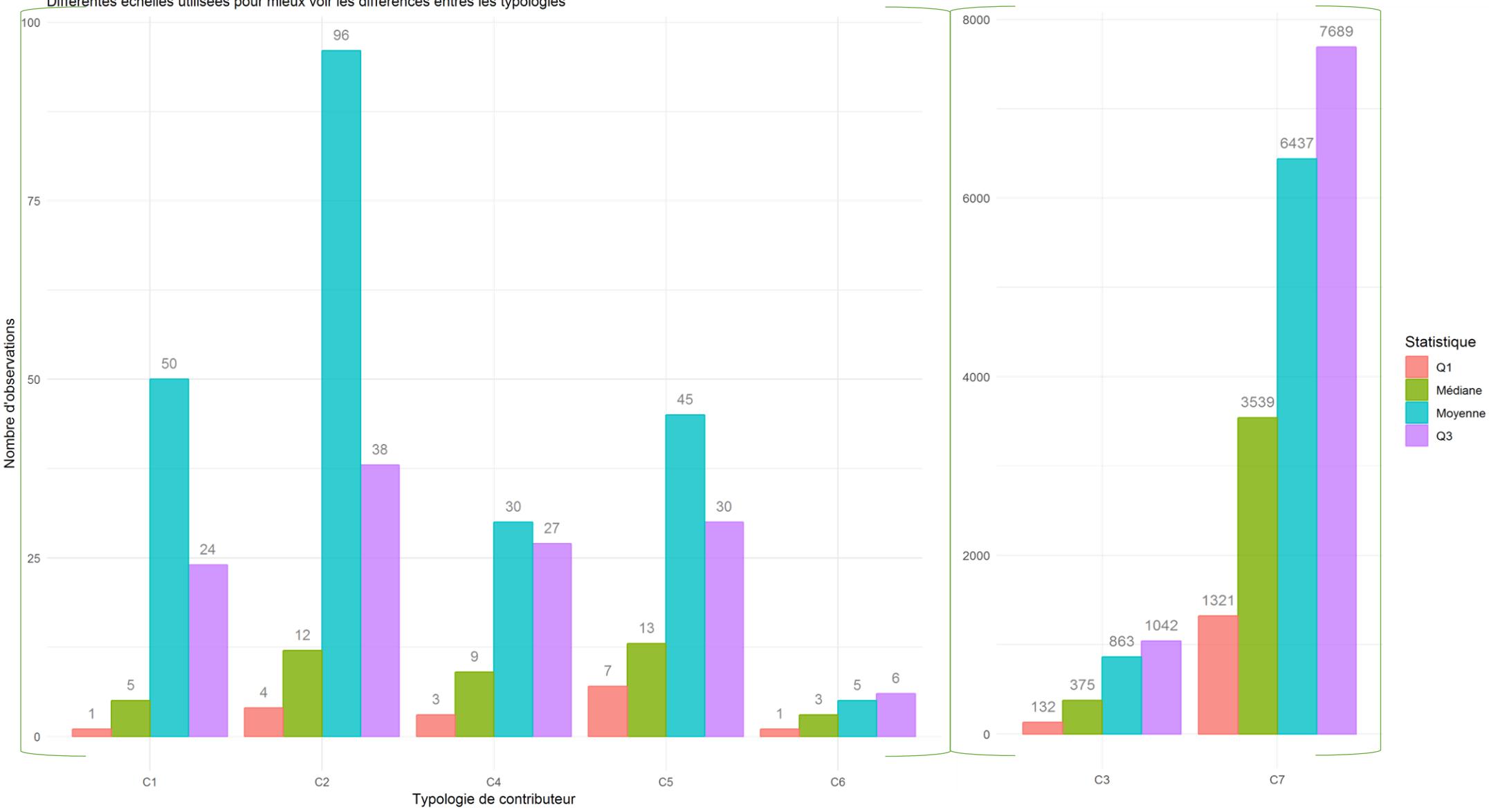
Grâce au graphique de la page 33, on peut visualiser les différents indicateurs statistiques du nombre d'observations pour chaque type de contributeur. On peut voir que les profils C3 et C7 sont difficilement comparables aux autres profils.

<sup>8</sup> Les « consigneurs de rareté » ou plus couramment appelé « les cocheurs » sont historiquement les personnes qui cochent une page du guides Delachaux pour connaître toutes les espèces d'oiseaux et ainsi pouvoir cocher la page lorsqu'ils ont perçu le spécimen. Aujourd'hui de multiples sites permettent cette pratique sur toute la Faune (<http://cocheurs.fr>)

<sup>9</sup> [https://analytics.huma-num.fr/Karine.Pietropaoli/Faune%20France/typologie\\_FF.html#utilisation-de-la-base](https://analytics.huma-num.fr/Karine.Pietropaoli/Faune%20France/typologie_FF.html#utilisation-de-la-base).

## Indicateur statistique des observations des contributeurs par leur typologie sur 2017-18

Differentes échelles utilisées pour mieux voir les différences entre les typologies



Graphique 5 Diagramme en bâton des typologies

## 4 Méthodologie d'analyse spatiale

### 4.1 Explication générale

Pour analyser spatialement le comportement des contributeurs, il a fallu réfléchir aux différentes attitudes possibles de ceux-ci et trouver une méthode qui permette de représenter ce comportement à partir des données disponibles.

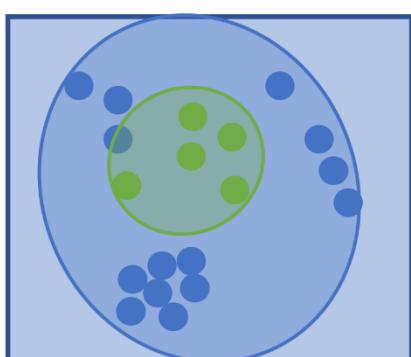
En parallèle, les recherches bibliographiques sur l'analyse spatiale (exposées en annexe page 77) m'ont permis d'avoir une idée plus vaste sur les comportements possibles ainsi que sur les méthodes utilisées. On retrouve dans l'analyse écologiques notamment avec le document *Spatial point patterns methodology and application with R* des méthodes d'analyses spatiales très poussées pour analyser certains comportements. Toutes ces méthodes et comportements n'étaient pas forcément adaptés à nos objectifs.

Grâce à ces différentes connaissances il a été possible de mettre en place différents indicateurs qui peuvent expliquer les comportements des contributeurs. Cependant cette approche reste très expérimentale et ne permet pas d'assurer à coup sûr une pratique d'observation. De plus la taille du jeu de données et son hétérogénéité viennent ajouter une difficulté supplémentaire pour l'interprétation des indicateurs.

L'objectif était d'arriver à percevoir trois critères spatiaux :

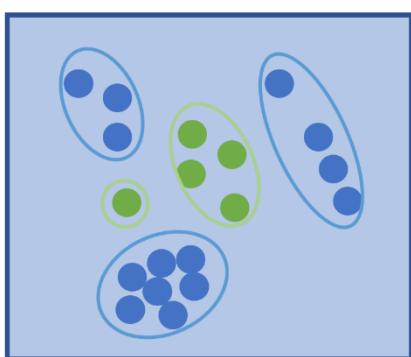
- La portée spatiale permet de voir l'étendue d'observation d'un contributeur. On peut alors avoir une idée de la mobilité du contributeur et de sa tendance à aller dans des « lieux » d'observations éloignés mais aussi une indication de la taille de sa zone d'observation.
- La structuration spatiale permet d'avoir une représentation des lieux d'observations. On peut essayer de restructurer des zones spatiales à partir des observations pour identifier des lieux précis d'observations.
- La concentration spatiale permet de voir la répartition des observations dans l'espace. Cela permet de savoir si les personnes ont tendance à plus observer un endroit spécifique.

Portée spatiale



- Points d'observations du contributeur A
- Points d'observations du contributeur B

Structuration spatiale



Concentration spatiale

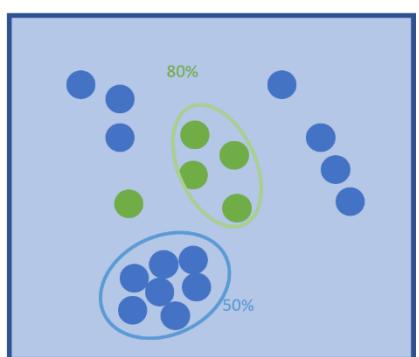


Schéma 1 représentation des critères spatiaux

Les différents critères peuvent ainsi nous permettre d'avoir une première représentation des comportements des contributeurs et de leurs tendances. Plusieurs méthodes sont possibles pour représenter ses critères. C'est pourquoi j'ai choisi différents indicateurs pour représenter ces critères.

# Concentration spatiale

# Porté spatiale

# Structuration spatiale

## 4.1.1 Les différents indicateurs spatiaux utilisés

Les indicateurs sont mis en place pour comprendre les comportements spatiaux. Ils s'appuient principalement sur les variables du lieu d'habitation et la localisation des observations du contributeur. Seulement quelques critères sont mis en place à partir de la temporalité qui nous permettent d'avoir une visualisation de l'espace dans le temps. Cependant cette démarche reste très difficile à appliquer à cause de l'hétérogénéité des contributeurs qui n'ont pas la même répartition temporelle.

Chacun de ses indicateurs est calculé pour un contributeur, ce qui nous permet d'avoir une visualisation de l'ensemble des contributeurs, ce qui prend parfois beaucoup de temps. Bien sûr une visualisation par typologie est aussi primordiale car elle nous aide confirmer des comportements spatiaux différents et donc à mesurer la logique spatiale de la science participative.

---

Le nombre de point de localisation différents

---

Taux entre le nombre de points de localisation différents et le nombre d'observations

---

Le nombre de communes, départements, régions où le contributeur a réalisé une observation

---

Le taux d'observation dans la commune, le département, la région d'habitation

---

Les distances (min, max, moyenne) entre un contributeur et ses observations

---

La distance moyenne entre un contributeur et ses points de localisation

---

La surface d'observation sur l'intégralité des observations  
La surface d'observation journalières (totale et moyenne des surfaces)

---

Le périmètre d'observation sur l'intégralité des observations  
Le périmètre d'observation journalière (total et moyenne des périmètres)

---

La distance totale et la moyenne journalières

---

Le nombre de cluster optimal

---

Le nombre d'observations par cluster

---

La surface moyenne et maximum de chaque cluster

---

Le taux d'observations dans le cluster principal

## 4.2 Concentration spatiale

### 4.2.1 Nombre de localisations

Comme expliqué dans la partie [2.3.3 Qu'est-ce qu'un point de localisation ?](#), le nombre de localisations est un critère spatial primordial, car il permet d'avoir une représentation du nombre de « lieux » différents visités par un contributeur. C'est aussi un critère qui nous permettra de savoir si l'on peut calculer une surface, ou réaliser un cluster.

Grâce aux coordonnées de base enregistrés en Lambert 93, nous connaissons la précision de l'observation au mètre près. Cependant nous souhaitons principalement savoir si le contributeur réalise toutes ses observations dans le même lieu. Or la taille d'un lieu est très variable :

- 1 m<sup>2</sup> : Observation de base
- 5 m<sup>2</sup> : Un ou deux pas de côté
- 50 m<sup>2</sup> : un petit jardin
- 500 m<sup>2</sup> : un parc ou un bosquet
- 5 km<sup>2</sup> : une forêt

Il est donc judicieux de regarder la concentration du nombre de localisations à différentes échelles (1 m<sup>2</sup>, 5m<sup>2</sup>, 50 m<sup>2</sup>, ...).

Cette information est donc pertinente et intéressante notamment pour les contributeurs ayant réalisé un nombre minimum d'observations. Si par exemple un utilisateur avec plus de 4 000 observations (gros contributeur) n'a que 2 ou 3 localisations différentes à 50 mètres, la concentration spatiale est très forte. Cet indicateur peut nous aider à interpréter une méthode d'observation qui peut témoigner d'une habitude ou d'un rituel du contributeur, ce qui peut impacter les différentes analyses de Faune France.

Cependant le nombre de localisations étant directement liée au nombre d'observations, celui-ci reste très corrélé et impacté. Il sera alors difficile de comparer deux contributeurs qui n'ont pas réalisé le même nombre d'observations.

### 4.2.2 Taux d'observation à différentes localisations

Avec l'objectif de mieux interpréter le résultat précédent, sans être influencés par le nombre d'observations réalisé par le contributeur, nous pouvons mettre en place un taux d'observations qui se calcule de la manière suivante :

$$\text{Taux d'observations} = \frac{\text{Nombre de localisations différentes}}{\text{Nombre d'observations totales}}$$

Ce taux est un indicateur qui nous permet de voir si un contributeur a :

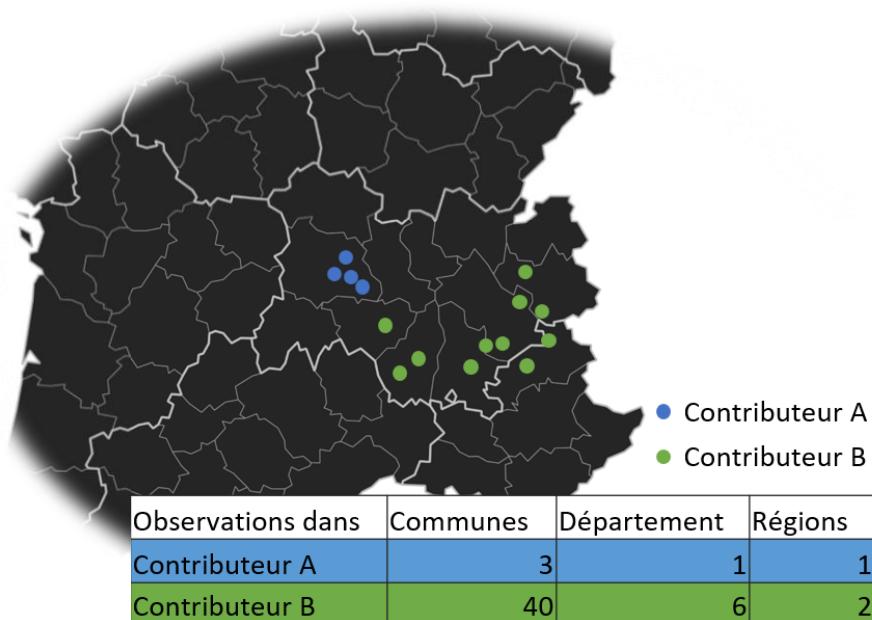
- Ses observations concentrées aux mêmes endroits/lieux (taux proches de 0)
- Ses observations très dispersées (taux proches de 1)

Attention ! Avec cette méthode, pour toutes les personnes ayant très peu d'observation (entre 1 et 4) le taux sera moins fiable. Car effectivement la personne ayant réalisé une seule observation, son taux sera de 100% donc aura une dispersion totale, alors que au contraire sa dispersion est nulle car elle est située aux mêmes endroits.

#### 4.2.3 Nombre de communes/Départements/Régions/paysages différents observés

Toujours avec l'objectif de comprendre la concentration spatiale dans un « lieu », nous pouvons caractériser le lieu par les limites administratives habituelles. Nous avons aussi ajouté la typologie des paysages qui permet d'indiquer pour chaque commune quel type de paysages est présent (13 paysages différents au total). Nous pouvons ainsi compter pour tous les contributeurs le nombre de communes, départements et régions différents où ils ont réalisé une observation.

Ses indices pourront alors nous aider à comprendre l'échelle d'observation d'un contributeur, ce qui peut nous donner un premier indice sur la mobilité du contributeur. Il faut cependant garder en tête que les limites administratives ne sont pas forcément homogènes (surface de l'Île-de-France par rapport à Auvergne-Rhône Alpes), mais aussi l'emplacement de certain département qui est adjacent à trois régions. Ce sont éléments qui viennent ajouter une certaine difficulté d'interprétation.



*Schéma 2 Représentation des observations de deux contributeurs*

Exemple : le contributeur A reste très sédentaire sur une zone avec une concentration à l'échelle communal alors que le contributeur B est plutôt dispersé avec une mobilité à l'échelle régionale.

Cependant là encore il est difficile de comparer deux contributeurs n'ayant pas le même nombre d'observations. La comparaison ne doit donc pas se faire entre typologie d'utilisateurs mais plutôt au sein même d'une typologie pour savoir si un profil peut contenir deux comportement spatiaux différents.

#### 4.2.4 Taux d'observations locale (commune, département, région)

Il peut être intéressant de se servir du lieu d'habitation indiqué par le contributeur pour voir si celui-ci a un impact sur ses observations. Pour cela nous avons regardé le nombre d'observations dans son lieu d'habitation par rapport au nombre d'observations total ce qui nous permet d'avoir un taux d'observations par rapport à son lieu d'habitation.

$$Taux\ d'observation\ locale = \frac{Nombre\ d'observations\ dans\ la\ commune}{Nombre\ d'observations\ totales}$$

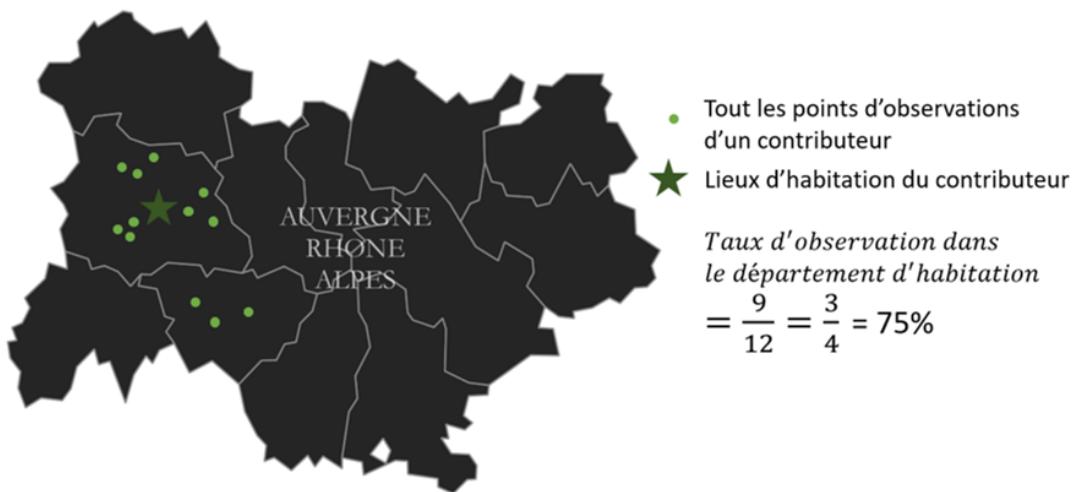


Schéma 3 Représentation des observations d'un contributeur et de son lieu d'habitation

Ce taux réalisé sur les différentes échelles nous permet d'avoir une idée si le contributeur à une observation dépendante de son lieu d'habitation. De plus cette représentation n'est pas biaisée par le nombre d'observations, on peut alors interpréter deux profils :

- Un taux est élevé (1) : le contributeur se limite à une zone d'observation limitée (communes, département, régions), le contributeur n'a pas de motivation ou d'occasion d'aller en dehors de sa zone d'observation. Sa concentration d'observations dans une zone est alors importante. Cela peut aussi lui permettre d'avoir une certaine maîtrise d'observation sur son milieu (connaissance de spot et de la diversité des espèces présentes, ...).
- Un taux est faible (0), le contributeur ne démontre pas d'attache particulière à son lieu d'habitation. Il aura donc tendance à être attiré par d'autres enjeux (recherche de diversité faunistique, envie de déplacement dans de nouveaux paysages, participation avec d'autres acteurs, recherche d'une espèce en particulier ou bien participation lors des déplacements professionnels ou de vacances personnelles). Sa concentration d'observation peut alors soit être faibles, soit concentrée sur un autre lieu.

## 4.3 Etendue/Portée spatiales

### 4.3.1 Matrice de distance :

La matrice de distance entre les points d'observations du contributeur et son lieu d'habitation nous permet de connaitre pour chaque utilisateur la distance minimale, maximale, moyenne et totale.

Ces distances peuvent nous apporter une idée sur les déplacements réalisés par les contributeurs pour observer. Bien sûr pour chacune des distances il faut prendre du recul sur leur signification.

Cette matrice de distance est créée à partir d'une requête (présente dans l'annexe page 69) sur une table de la base de données. Cette table est ensuite traitée dans le script R donnée spatiale pour créer les min, max, moyenne avant de permettre une visualisation.

#### 4.3.1.1 *Distance minimum*

La distance minimum entre les points d'observations du contributeur et son lieu d'habitation permet d'avoir une première vision sur le motif d'utilisation de l'application.

On peut alors retrouver différents profils :

- Les contributeurs qui utilisent très régulièrement l'application auront tendance à avoir une distance minimum très petite.
- Pour un contributeur voulant essayer l'application on retrouvera alors une distance minimum proche de 0.

Ce sont sur ces derniers qu'il est intéressant de se pencher car il est possible d'avoir là aussi différents profils :

- Ceux qui ont découvert l'application pendant des vacances, ou séjours (cas exceptionnel)
- Ceux qui souhaitent renseigner leurs observations lors de rares événements ou lorsqu'il rencontre une espèce exceptionnelle.

#### 4.3.1.2 *Distance moyenne*

La distance moyenne entre les points d'observations du contributeur et son lieu d'habitation permet d'avoir une première approche sur la distance moyenne des observations de l'utilisateur.

Cette première approche ne représente cependant pas du tout la distance parcourue par le contributeur, car si celui-ci se déplace une seule fois à 500 km pour réaliser une majorité d'observations alors qu'il réalise l'autre partie des observations à côté de chez lui, nous avons l'impression qu'il observe à 250km aux alentour en moyenne alors qu'en réalité une majorité de ses observations sont très proches de chez lui et seulement quelques-unes sont éloignées.

De plus, une vigilance supplémentaire doit être apportée aux contributeurs ayant peu d'observations, car l'indicateur peut vite être tronqué par des valeurs extrêmes, qui peuvent témoigner d'un comportement spatial particulier (observation uniquement sur les lieux des vacances, ...)

Une deuxième approche est de regarder la distance moyenne entre les points de localisation du contributeur et son lieu d'habitation, ce qui permet de supprimer le facteur nombre d'observations qui vient fausser l'information.

Mais là encore nous ne pouvons pas représenter la distance parcourue par l'utilisateur mais seulement à quelle distance se situent en moyenne ses lieux d'observations, ce qui peut nous permettre d'avoir une deuxième représentation de la mobilité spatiale du contributeurs (en plus du nombre de communes, département, régions).

#### 4.3.2 Fonction calcul des surfaces, périmètres, distances moyennes journalières

Grâce à la matrice de distance nous arrivons à percevoir une certaine mobilité des contributeurs. Cependant l'étendue spatiale n'est pas qu'une mobilité vers un lieu d'observation, c'est une mobilité lors de l'observation dans le lieu lui-même. Comment les contributeurs réalisent leurs observations sur place ?

On peut alors penser à plusieurs profils de contributeurs :

- Ceux qui ne regardent qu'un espace en restant à un endroit fixe
- Ceux qui renseignent toutes les observations lors d'une balade
- Ceux qui renseignent toutes les observations lors d'un voyage
- Ceux qui notent leurs observations à différents spots, mais les renseignent seulement au lieu-dit.

Afin d'essayer d'entrevoir ces comportements, il a fallu mêler données spatiales et données temporelles pour percevoir plusieurs indicateurs qui pourraient nous aider à comprendre les différentes démarches et logiques :

- La surface totale
- La surface moyenne journalière/mensuelle
- Le périmètre total
- Le périmètre moyen journalier/mensuel
- La distance moyenne aux barycentres des surfaces journalières et mensuelles.

Là encore il est difficile d'interpréter et de connaître exactement le mode d'observation des experts, passionnés ou amateurs. Il y a une grosse distorsion entre la spatialité des données présentes dans la base de données et ce qui se passe dans la réalité.

Par exemple, il est possible qu'un contributeur ait observé toute une forêt avant de d'apercevoir réellement seulement une espèce. Notre indicateur nous indiquera alors que le contributeur n'a observé qu'un lieu et non l'intégralité de la forêt. La représentation de la réalité n'est alors pas fidèle. Ce sont donc des critères qui reste très exploratoires, qui peuvent nous donner des indices sur des comportements mais qui ne représentent pas exactement la réalité.

Cette fonction qui permet de calculer l'ensemble de ces indicateurs pour un contributeur s'avère chronophage en temps de calcul pour l'ensemble des contributeurs. Malgré les dernières améliorations, il faut compter une nuit de calcul (au lieu de 4 jours précédemment).

La fonction a besoin de plusieurs paramètres en entrée, une standardisation des données ainsi que plusieurs packages dans R. De plus, comme expliqué depuis un moment, le calcul de surface et de périmètres nécessite un nombre de points de localisation minimum, ce qui ne permet pas de représenter l'ensemble des contributeurs

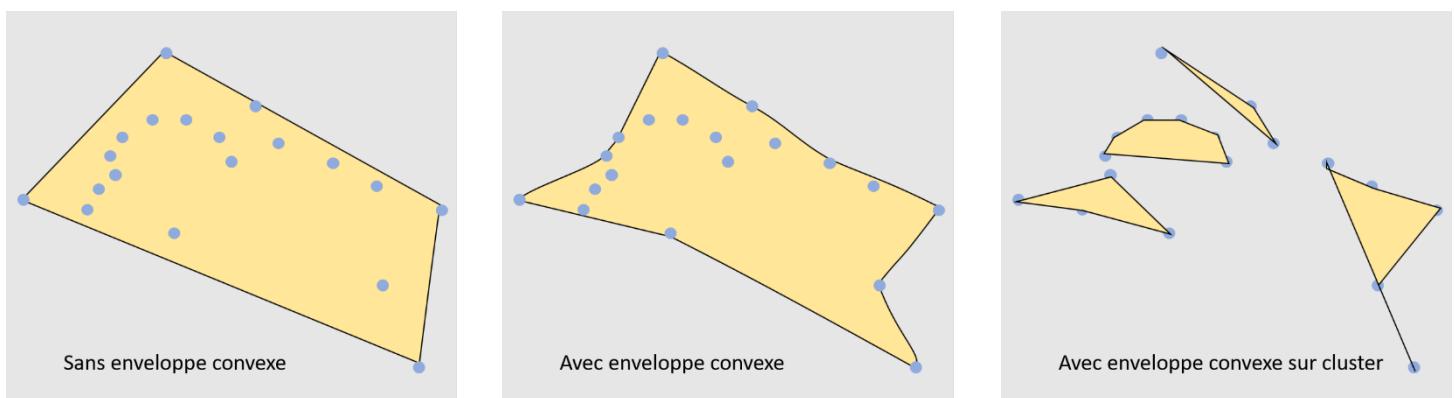
Cette fonction fait appel à plusieurs fonctions du package notamment SF et concaveman qui permettent de réaliser des géotraitements tel que (st\_as\_sf, st\_centroid, st\_cast, st\_union, st\_area, st\_length, st\_distance, concaveman) (voir bibliographie des fonctions R en annexe page 68)

#### 4.3.2.1 La surface totale

La surface globale du contributeur calculée à partir de l'ensemble de ses observations nous permet de donner un indicateur sur la couverture spatiale réalisée par celui-ci. Cet indice nous permet d'avoir une superficie en km<sup>2</sup>, cependant elle ne représente pas la superficie observée, mais donne plutôt une perception de l'étendue des lieux observés.

Le calcul de la surface se fait à partir de l'ensemble des points géométriques de l'utilisateur, puis l'utilisation de la fonction concaveman permet de trouver l'enveloppe convexe de l'ensemble des points afin d'obtenir un polygone. Une fois le polygone obtenu la fonction st\_area m'a permis d'avoir la surface du polygone en mètres carrés (géométrie fournie en Lambert 93). L'enveloppe convexe peut aussi intégrer des paramètres tel que le clustering (voir partie [5.4.1 Analyses clustering](#)), cependant la difficulté et le temps de traitement étant plus lourd j'ai préféré me centrer pour cet indicateur sur une simple enveloppe convexe qui nous permet déjà d'obtenir une bonne représentation.

Schéma récapitulatif des différents traitements de calcul de surface :

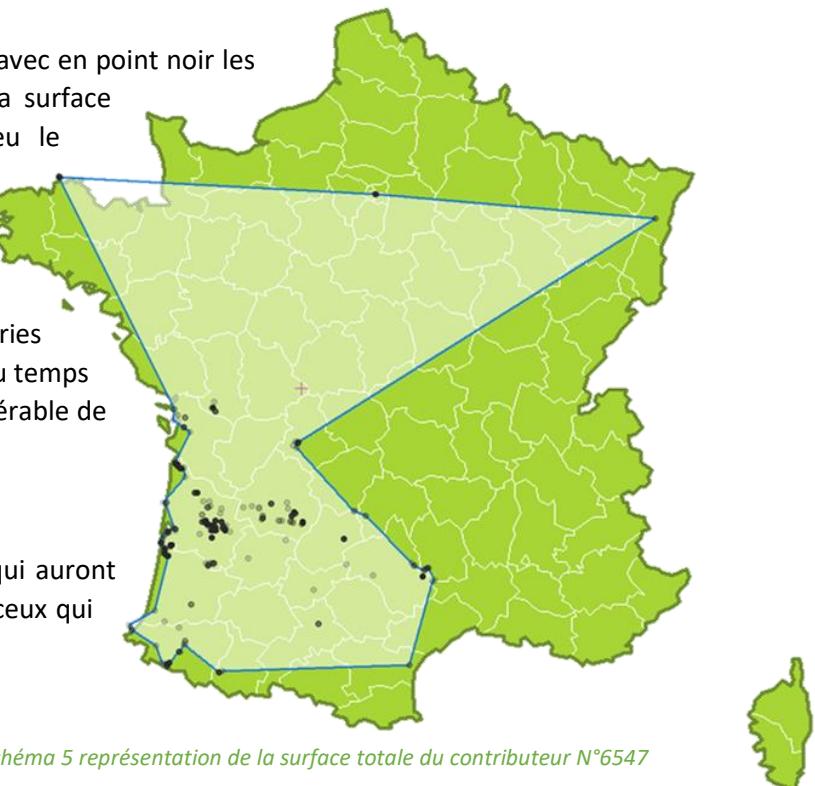


*Schéma 4 Représentation des différentes méthodes de calcul de surface*

Voici un exemple de résultat de la fonction avec en point noir les observations d'un contributeur, en blanc la surface réalisée par l'enveloppe convexe, en bleu le périmètre de cette surface et la croix rose le centroïde de surface.

La fonction permet de visualiser les géométries créées obtenues. Cependant, pour gagner du temps en traitement et calcul de surface il est préférable de ne pas les afficher.

On peut ainsi distinguer les contributeurs qui auront tendance à observer partout en France, et ceux qui sont restreints à une certaine zone.



*Schéma 5 représentation de la surface totale du contributeur N°6547*

#### 4.3.2.2 La surface moyenne et totale journalières/mensuelles

La surface moyenne journalières/mensuelles n'est pas calculée de la même façon. Cette fois-ci pour chaque contributeur on récupère l'ensemble de ses points de géométries que l'on vient ensuite scinder par jours/mois. Ensuite pour chaque élément scindé, on réalise l'enveloppe convexe pour calculer la surface. Toutes les surfaces calculées sont ensuite enregistrées dans un tableau temporaire qui va permettre le calcul de la moyenne de toutes les surfaces ainsi que le total des surfaces.

Cette fois-ci la prise en compte de la temporalité permet d'avoir une vision de la surface observée par jour du contributeur. Cette vision nous permet de cerner les méthodes spatiales d'observations du contributeur. Ainsi on peut commencer à voir émerger plusieurs profils :

- Ceux qui observent seulement sur un spot (surface moyenne très petite)
- Ceux qui observent sur tous lieux (surface moyenne grande)
- Ceux qui observent lors des trajets en voiture (surface moyenne très grandes)

*Attention, il arrive parfois que nous percevions ce phénomène lorsque deux personnes utilisent un seul compte alors qu'ils sont à des endroits différents, mais ceci vient d'un problème d'approche du contributeur qui doit être fait avant et non de son analyse.*

Cependant ce niveau de précision d'analyse est très élevé, car nos restrictions sont très grandes. On demande d'avoir au minimum quatre points d'observations par jour d'observation. Ce qui est beaucoup, même pour les plus gros contributeurs. L'indice de surface moyenne journalière est rarement calculée pour la totalité des jours mais souvent pour les quelques jours où il y a eu le plus d'observations. Ce qui permet déjà d'avoir un échantillonnage.

Sur cet exemple (même contributeur que le précédent) on peut voir ses observations en point noir, ses surfaces journalières en blanc avec le périmètre en bleu.

On peut tout de suite voir, contrairement à la carte précédente qu'il y a très peu de surface et qu'une grande surface est liée à des observations durant un trajet (voiture ou train).



#### 4.3.2.3 Le périmètre total

Tout comme la surface totale, le périmètre total peu lui aussi nous permettre d'avoir un indice sur l'étendue globale des contributeurs. Cependant le périmètre ne présente pas les mêmes contraintes de création. Il lui suffit de deux points contrairement à quatre pour la surface. Ce qui peut permettre de représenter toutes les personnes ayant entre deux et quatre points de localisation différents qui ne sont pas représentés par la surface.

Pour calculer le périmètre total, la méthode est la suivante :

- Plus de trois points de localisation : prise de la géométrie du polygone de la surface totale, transformation en géométrie de type ligne grâce à la fonction **st\_cast**, puis calcul de la longueur des lignes par la fonction **st\_length** (applicable seulement sur des lignes)
- Entre deux et trois points : création de la géométrie ligne à partir des points de localisation (utilisation de la fonction **st\_union** pour regrouper tous les points en un seul objet et utilisation de la fonction **st\_cast** pour transformer du multipoint en multiligne), puis calcul de la longueur avec **st\_length**

#### 4.3.2.4 Le périmètre moyen journalier/mensuelles

J'ai appliqué la même logique que le périmètre total, pour le périmètre moyen et total journalier/mensuelle. Cette fois-ci, les calculs se font à partir de la surface moyenne et totale journalière/mensuelle mais la même méthode est appliquée. L'objectif est vraiment de pouvoir visualiser plus de contributeur si la surface est trop exigeante.

#### 4.3.2.5 La distance moyenne et totale aux barycentres des surfaces journalières et mensuelles.

Pour affiner et représenter une réelle distance parcourue par les utilisateurs entre leurs lieux d'observations et leur lieu d'habitation, il a fallu se pencher sur la distance entre le lieu d'habitation du contributeur et le point barycentre journalier/mensuel (le point à équidistance de toutes les observations journalière/mensuelles).

Pour calculer ce barycentre, il suffit de prendre en fonction des points de localisation :

- Le centroïde du polygone de la surface journalière
- Le centroïde de la ligne du périmètre journalier
- Le point d'observation

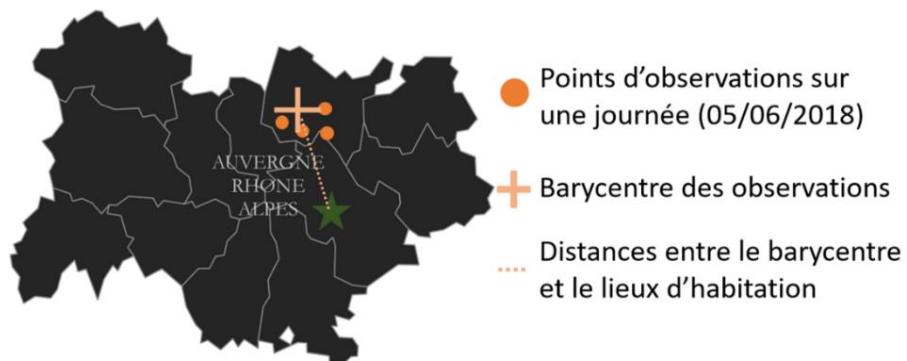


Schéma 7 Représentation de la distance barycentre habitation d'un contributeur

Une fois le barycentre trouvé il suffit alors d'utiliser la fonction **st\_distance** qui permet de trouver la distance en mètre entre deux points (en Lambert 93)

Cependant, même si nous nous rapprochons d'une réalité de la distance parcourue journalière, celle-ci n'est pas entièrement juste. Si la personne change d'habitation, que ce soit sur une courte période ou pendant un long moment, cela impactera directement la distance moyenne et totale journalière/mensuelle. Ce qui montre que la méthode reste indicative pour des pistes de réflexion sur les comportements en général.

## 4.4 Structuration spatiale

### 4.4.1 Calcul des clusters

Dans l'objectif de comprendre les différents lieux d'observations des contributeurs sur les années 2017 et 2018, on a utilisé une démarche de clustering. Il permet de rassembler des groupes de variable qui ont les mêmes tendances. Ici le clustering sera basé sur les variables spatiales pour permettre un regroupement spatial des variables.

Le clustering permet une visualisation du nombre des groupements (donc un nombre de lieux d'observation), de connaître le nombre d'observations par groupement, la taille des groupements (un grand groupement ayant quelques observations ou bien un petit groupement ayant beaucoup d'observations), ainsi que l'isolement des groupement (est-ce que les clusters sont éloignés l'un de l'autre ou très proches ?). Tout cela nous permet de comprendre comment le contributeur observe son territoire avec différents points de vue :

- Est-ce que le contributeur a plusieurs lieux d'observations ? Identifier le nombre de zones distinctes observées par les contributeurs.
- Est-ce que ses lieux d'observations sont vastes ? Calculer des indicateurs sur la taille des clusters pour savoir si ses zones d'observations des contributeurs sont grandes ou concentrées.
- Est-ce que ses lieux d'observations sont éloignés ? Calculer des indicateurs sur la distance entre les clusters.
- Est-ce que le contributeur observe particulièrement un lieu précis ? Calculer son taux d'observation afin de savoir si l'utilisateur observe en particulier une zone.

#### *Méthode de calcul :*

Bien-sûr cette analyse n'est pas possible sur tous les contributeurs, que ce soient les personnes n'ayant pas suffisamment d'observation, ou à l'inverse celles en ayant trop. Pour comprendre tous les cas et arriver à réaliser des clusters sur un maximum de contributeurs, une série de tests est mis en place. Elle permet de voir dans l'ordre si :

- Le contributeur a moins de 10 observations (nombre insuffisant pour représenter significativement le contributeur)
- Le contributeur à moins de 3 points de localisations (s'il y a que deux points, le cluster se fait forcément sur les deux points à 100%, donc une information tronquée)
- Le contributeur a plus de 8 000 points de localisation (nombre important d'information, au-delà il est long et difficile de réaliser les clusters)
- Le contributeur a plus de 8 000 points de localisation agrégé à 5 m<sup>2</sup> (permet d'échantillonner un peu les points de localisation)
- Le contributeur à plus de 8 000 points de localisation agrégé à 50 m<sup>2</sup> (deuxième échantillonnage un peu plus grand)
- Le contributeur à plus de 8 000 points de localisation agrégé à 50 m<sup>2</sup> qui contiennent au minimum 3 observations (troisième échantillonnage pour prendre seulement les points de localisation les plus significatifs)

Grâce à cette première série de tests sur le nombre d'observations et de localisations, nous pouvons attribuer au contributeur l'échantillon du nombre d'observation adaptés pour réaliser le cluster. Il faut bien-sûr prendre en compte que cette série de test est adapté à notre base de données actuelle, il est possible que l'on soit obligé d'ajouter des étapes d'échantillonnage si un contributeur a trop d'information.

## 5 Analyses spatiales

Toutes les méthodes expliquées ont été mis en place, les résultats ont permis d'obtenir de nombreux graphiques, qu'il a fallu représenter en s'assurant de garder l'intégralité de l'information mais aussi en comprenant bien les différents phénomènes représentés.

Ayant énormément de graphiques en sortie, tous ne sont pas représentés, seuls les plus pertinents ayant une information à apporter sont expliqués. Il est possible d'en consulter d'autres en annexe page 70-76, cependant pour certains non documentés, ils ne sont pas présents dans ce document mais sont conservés dans les scripts.

### 5.1 Concentration spatiale

#### 5.1.1 Nombre de localisations

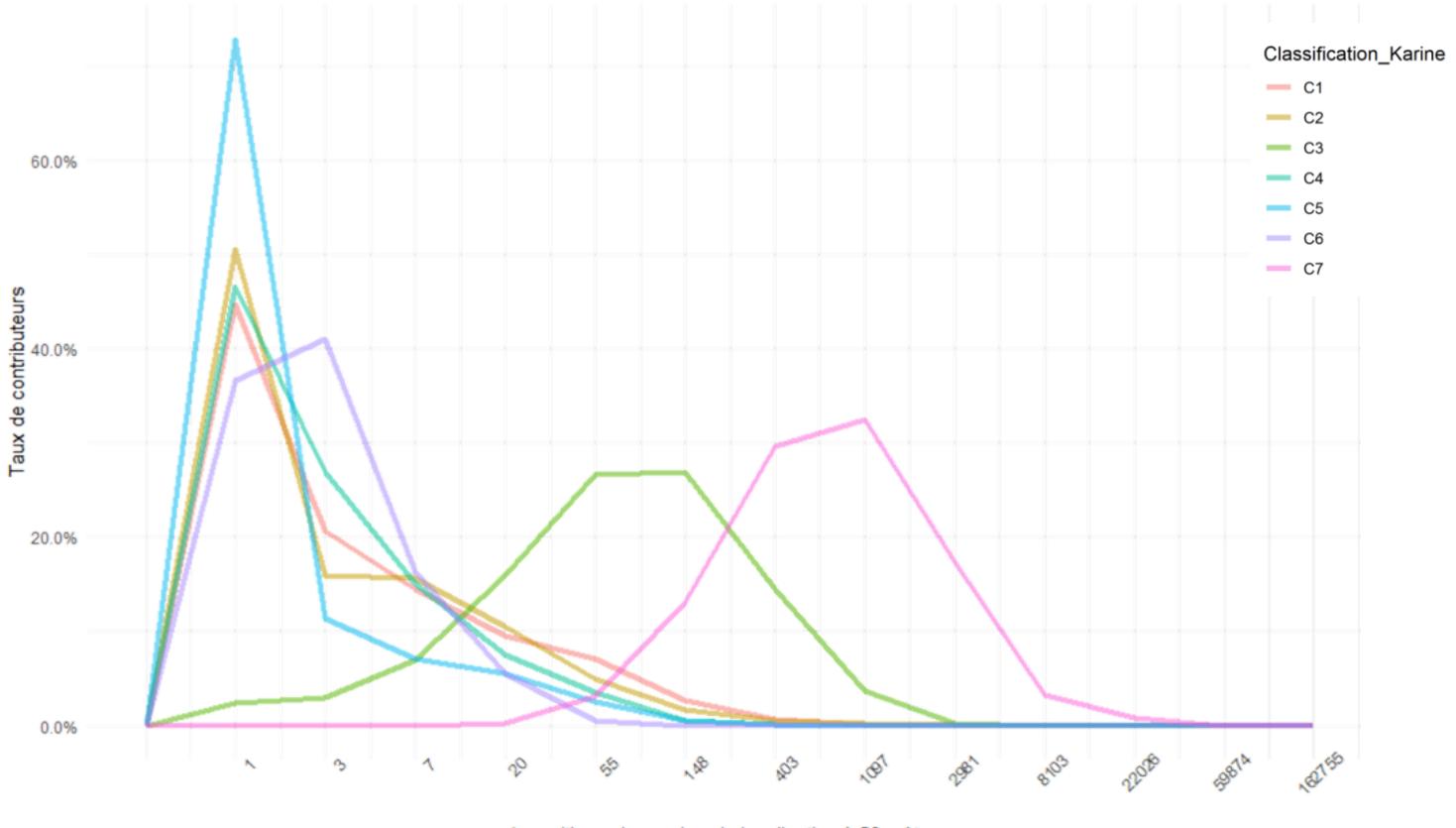
Ce graphique représente le pourcentage de contributeurs pour chaque classification en fonction du nombre de localisations différentes à 50 mètres. On peut alors lire que 72% des contributeurs du profil C5 « oiseaux des jardins » ont réalisé des observations sur un seul point de localisation.

Dans l'ensemble, le nombre de points de localisation est très différente pour chaque typologie d'utilisateurs, mais chaque typologie reste homogène dans sa distribution. On peut voir 5 comportements différents :

- Le profil C5 où les observateurs n'ont qu'un seul lieu d'observation
- Les profils C1, C2 et C4 avec un peu plus d'observations mais très peu de localisations.
- Le profil C6 qui a beaucoup de localisations différentes (2 à 4) alors qu'ils ont en moyenne seulement 5 observations ce qui montre une réelle volonté d'observer différents lieux.
- Enfin les profils C3 et C7, qui par leur grand nombre d'observations ont beaucoup plus de lieux de localisations.

Globalement ce résultat est très influencé par le nombre d'observations et rend l'interprétation difficile. Mais nous pouvons déjà distinguer différents comportements spatiaux.

Nombre de point de localisation différent à 50 mètres



### 5.1.2 Taux d'observation à différentes localisations

La lecture des graphiques N°7 et 8 de la pages 46 reste la même que celle du graphique précédent hormis l'abscisse qui est remplacé par le taux d'observation pour chaque lieu différent à 1 mètre et 5 mètres. On peut alors lire sur le graphique N°7 que 40% des contributeurs de la typologie C5 ont un taux d'observation à différente localisations a 10%.

Cette fois-ci l'indicateur n'est pas lourdement influencé par le nombre d'observation, ce qui nous permet de voir différents comportements au sein des classes.

Pour la classe C5 on peut voir une majorité des contributeurs ave un taux d'observations à des localisations différentes très faible, ce qui nous montre une tendance chez ses contributeurs à réaliser des observations au même endroit.

Pour les classes C1 et C6 cette fois-ci c'est une majorité de contributeurs à avoir un taux élevé, on peut donc penser que ce sont des classes qui ont tendance à ne jamais refaire des observations au même endroit. Cette tendance est intéressante, pour C6 les consigneurs de rareté (cocheurs), nous avons un premier indicateur qui vient confirmer cette volonté d'aller dans des lieux différents afin d'apercevoir une espèce. Pour C1 les personnes qui observent les papillons cela peut confirmer une méthode d'observation différente des ornithologues, avec pour objectifs d'observer différents lieux.

Pour les classes C2 et C4 qui représentent les observateurs des oiseaux en printemps (C2) et en automne (C4), on perçoit une certaine similitude, avec pour les deux classes deux comportements bien distincts, soit les contributeurs observent au même endroit, soit les contributeurs ne font que des lieux différents. Il n'y a que très peu de contributeurs avec une répartition intermédiaire.

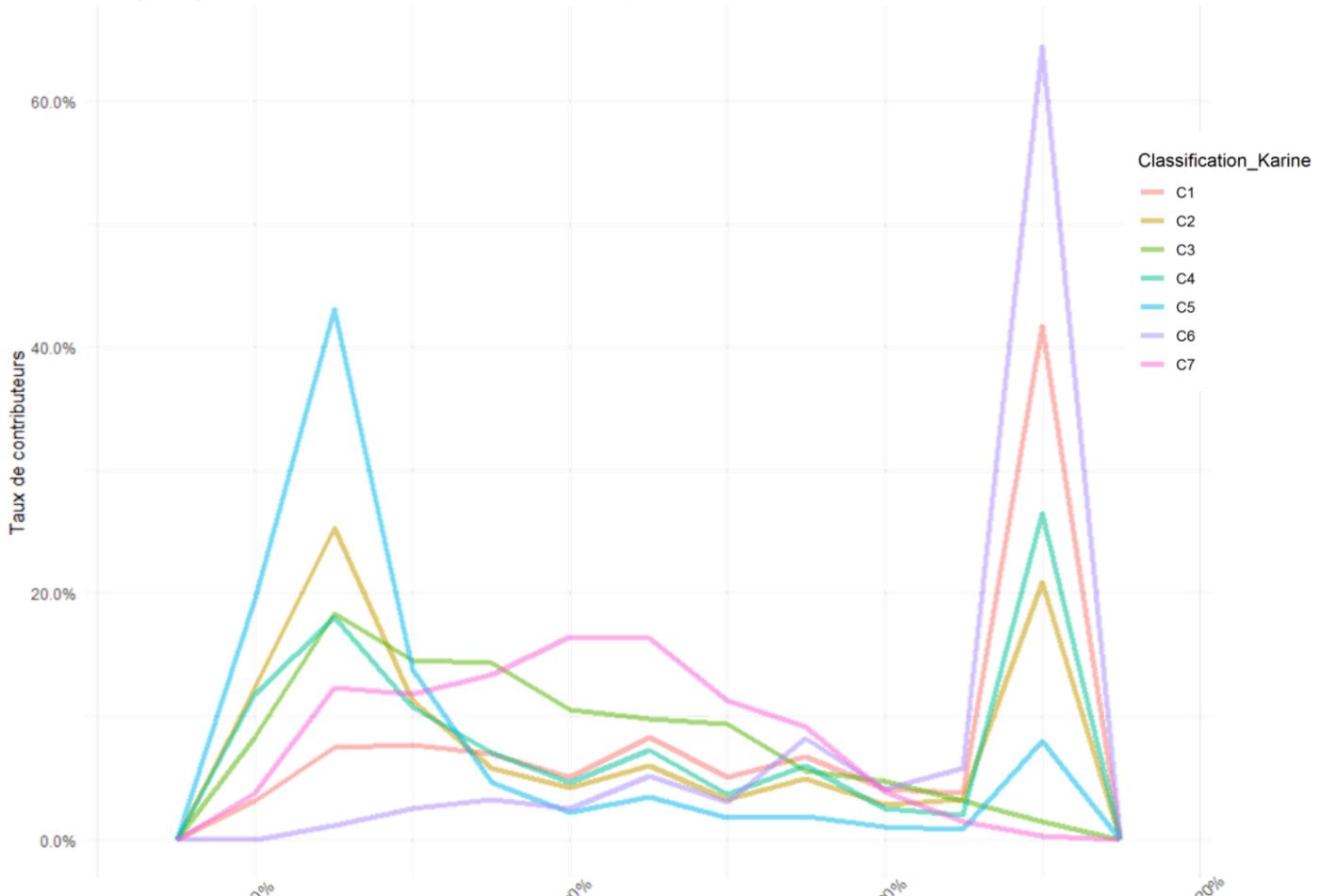
A l'inverse la classe C7 qui représente les experts/professionnel est la mieux répartie, il n'y a pas d'extrême. Il est possible que les contributeurs aient une méthode différente des précédentes classes, qui visent à la fois à réaliser beaucoup d'observation sur un lieu mais aussi sur beaucoup de lieux différents.

Enfin la classe C3 qui représente les observateurs réguliers présente une bonne répartition là aussi des contributeurs, même si on peut voir une certaine tendance d'observation sur le même lieu. On peut donc penser que leur méthode se rapproche de celle des experts mais qu'ils ne peuvent pas déployer les mêmes moyens.

Cette différence de méthode et de comportement (ceux qui observent sur les mêmes lieux, ceux qui un lieu différent à chaque fois et ceux qui font un peu des deux) peut avoir des répercussions non négligeables pour Faune France, car les données collectées n'ont pas les même objectif et représentation de la réalité. Les données doivent être traitées différemment. Si les ornithologues visent à multiplier l'information, il est possible que chez d'autres naturaliste la pratique ne soit pas similaire.

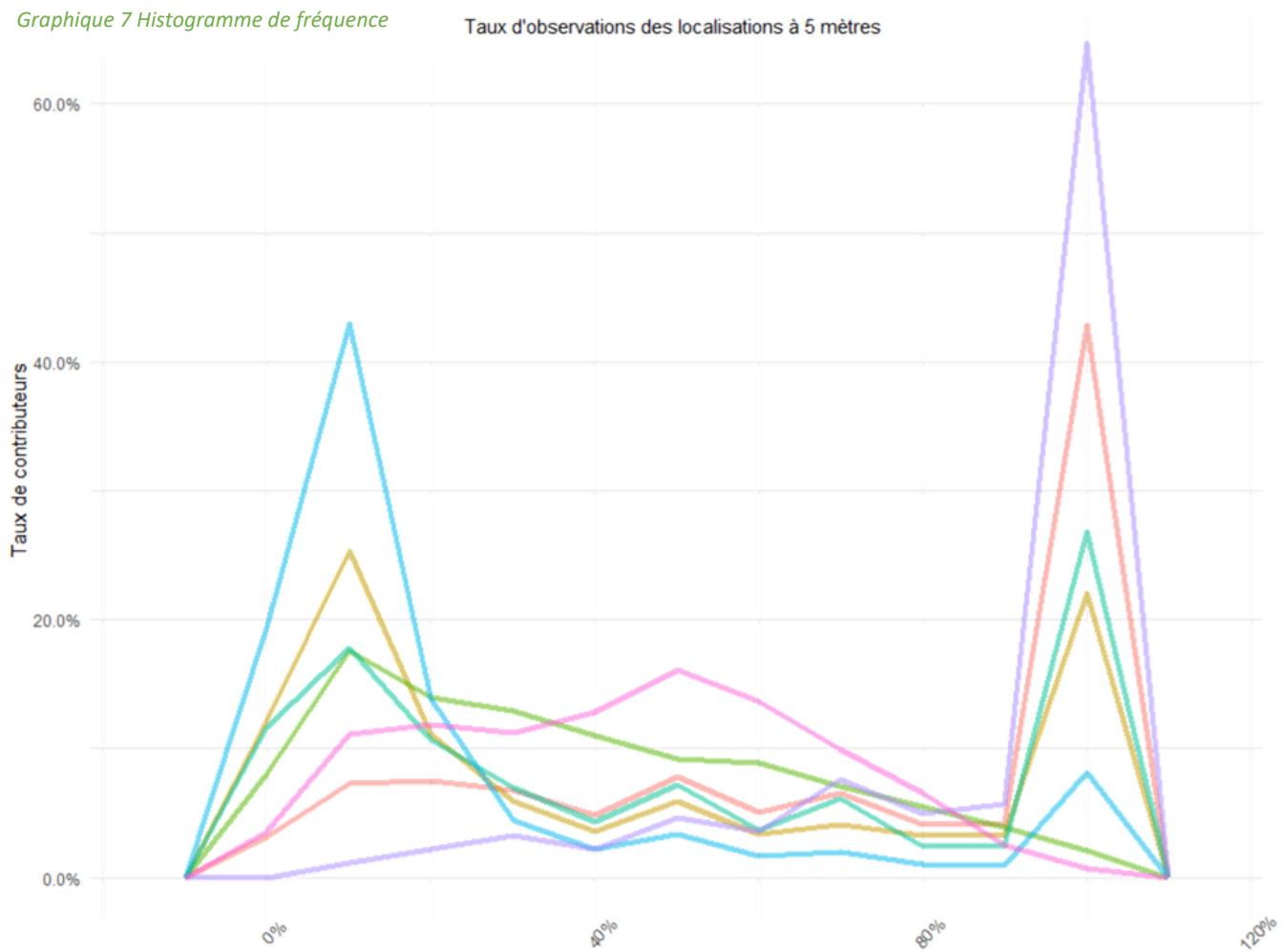
Il est possible de regarder les graphiques du taux d'observation à différents mètres : 1 mètres, 5 mètres, 50 mètres, 500 mètres, 5 km, graphiques disponibles dans l'annexe à la page 72. On peut alors voir que si on change le grain d'observation c'est principalement les profils C3 et C7 qui ont une variation d'histogramme. Cette variation nous permet de voir que ces profils réalisent des observations à différents grain contrairement aux autres profils.

### Les typologies des contributeurs ont-il une vision dispersé ou concentré ?



Graphique 7 Histogramme de fréquence

Taux d'observations des localisations à 5 mètres



Graphique 8 Histogramme de fréquence

Taux d'observations des localisations à 1 mètres

### 5.1.3 Nombre de communes/départements/régions/paysages différents observés

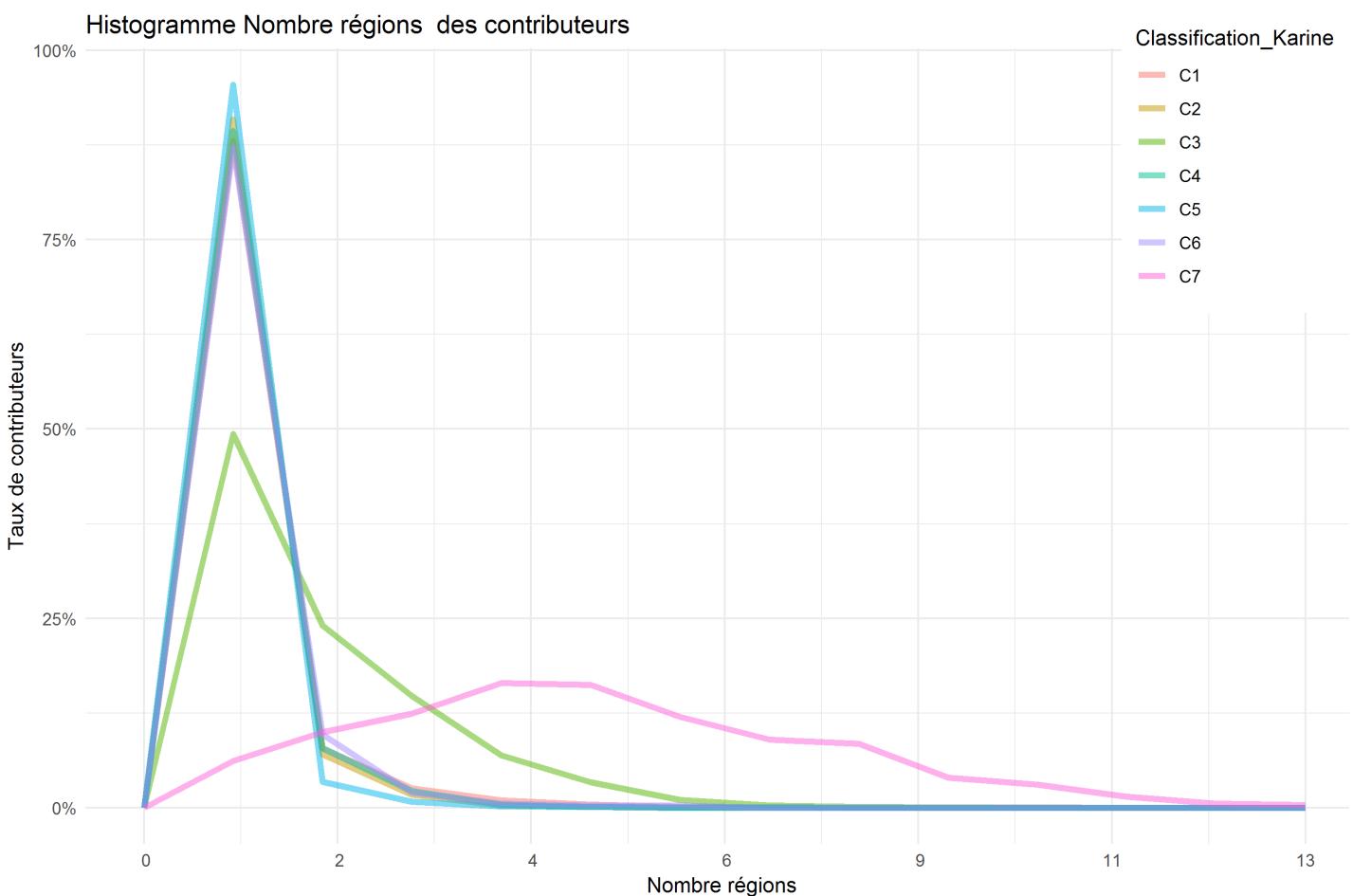
Sur cet histogramme nous pouvons lire le pourcentage de contributeurs au sein de chaque classe de façons indépendante, par rapport au nombre de régions parcourues. On peut alors lire que 50 % des contributeurs de la classe C3 ont observé seulement dans une région.

Cependant on s'aperçoit que cet indicateur ne différencie pas beaucoup de comportements différents. Il y a trois grands comportements : C1, C2, C4, C5 et C6 qui ont un comportement très sédentaire fixé pour la majorité sur une seule région.

C3 qui a deux profils dans la même classe, d'une part 50 % des contributeurs qui restent sédentaires tout comme le profil précédent, l'échelle d'observation est dispersée au minimum sur quelques départements, mais l'autre partie des contributeurs de cette classification a des déplacements plus importants à l'échelle régionale (quelques régions et plusieurs départements).

Enfin C7 a une mobilité très importante. On peut voir que la plupart des contributeurs de C7 parcourent la moitié de la France avec un nombre de départements et de régions qui restent très importants. Les contributeurs de cette classification observent au minimum à l'échelle régionale mais pour beaucoup d'entre eux c'est presque une échelle nationale.

(Les autres graphiques avec le nombre de départements, communes et les histogrammes généraux sont disponible en Annexe page [73](#))



Graphique 9 Histogramme de fréquence

#### 5.1.4 Taux d'observations dans sa propre commune/département/région

L'histogramme le pourcentage de contributeurs de chaque typologie en fonction de leurs taux d'observation dans la commune d'habitation des contributeurs. On peut alors lire que 11% des contributeurs du profile C6 réalisent 50% de leurs observations en dehors de leur commune d'habitation.

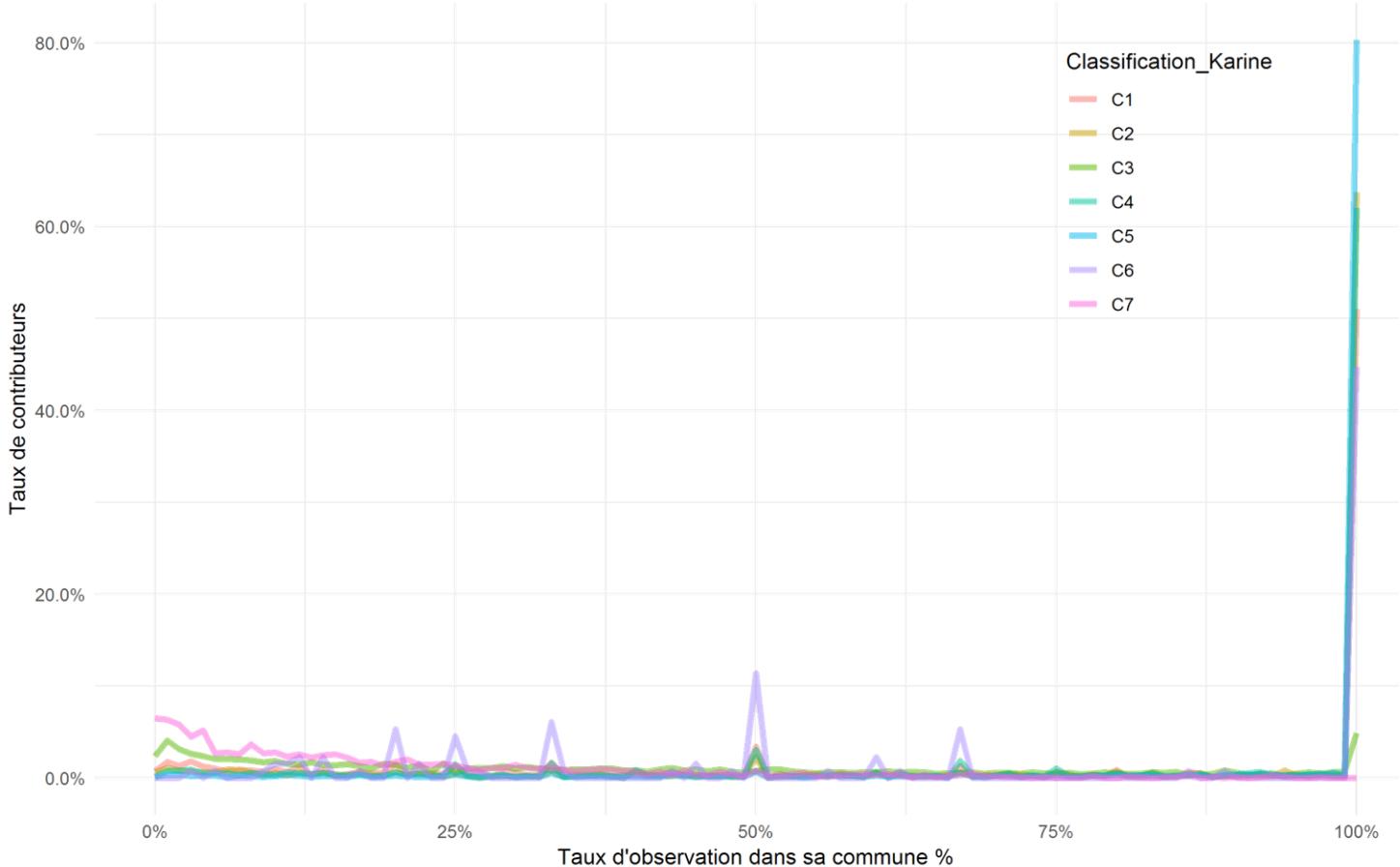
La lecture de l'histogramme n'est pas évidente, car les profiles sont très variés mais aussi parce que le nombre d'observations parfois faibles créer rapidement des petits pics à 25%, 33%, 50%, 66%, 75%, ...

Cependant malgré ces difficultés on peut voir que pour la plupart des profiles l'observation se fait majoritairement voire uniquement sur la commune d'habitation.

Seule les typologies C7 et C3 ont un taux d'observation faible qui tend vers 0% ce qui montre un certain détachement de la commune d'habitation et une volonté d'observé dans différent milieux.

Pour la typologie C6 on peut aussi retrouver ce comportement chez certaine personne, on peut voir comme des petits pics ce qui est lié aux faibles nombres d'observations réalisé, cependant cela montre aussi qu'il y a peu d'attachement aux lieux d'habitation. (Voir autres graphiques sur la thématique avec un grain départemental et régional en annexe page [74](#))

Histogramme Taux d'observation dans sa commune % des contributeurs



Graphique 10 Histogramme de fréquence

## 5.2 Etendue/Porté spatiale

### 5.2.1 Distance minimum

Le graphique est toujours un histogramme de fréquences par typologie de contributeur qui est cette fois-ci cumulé en fonction de la plus petite distance entre leurs observations et leur lieu d'habitation. Cette distance en kilomètre est visualisée en logarithme pour mieux étaler et comprendre les histogrammes. De cette façon nous pouvons lire que 20% des contributeurs de la classe C6 ont réalisé leurs observations les plus proches entre 24 et 48 km.

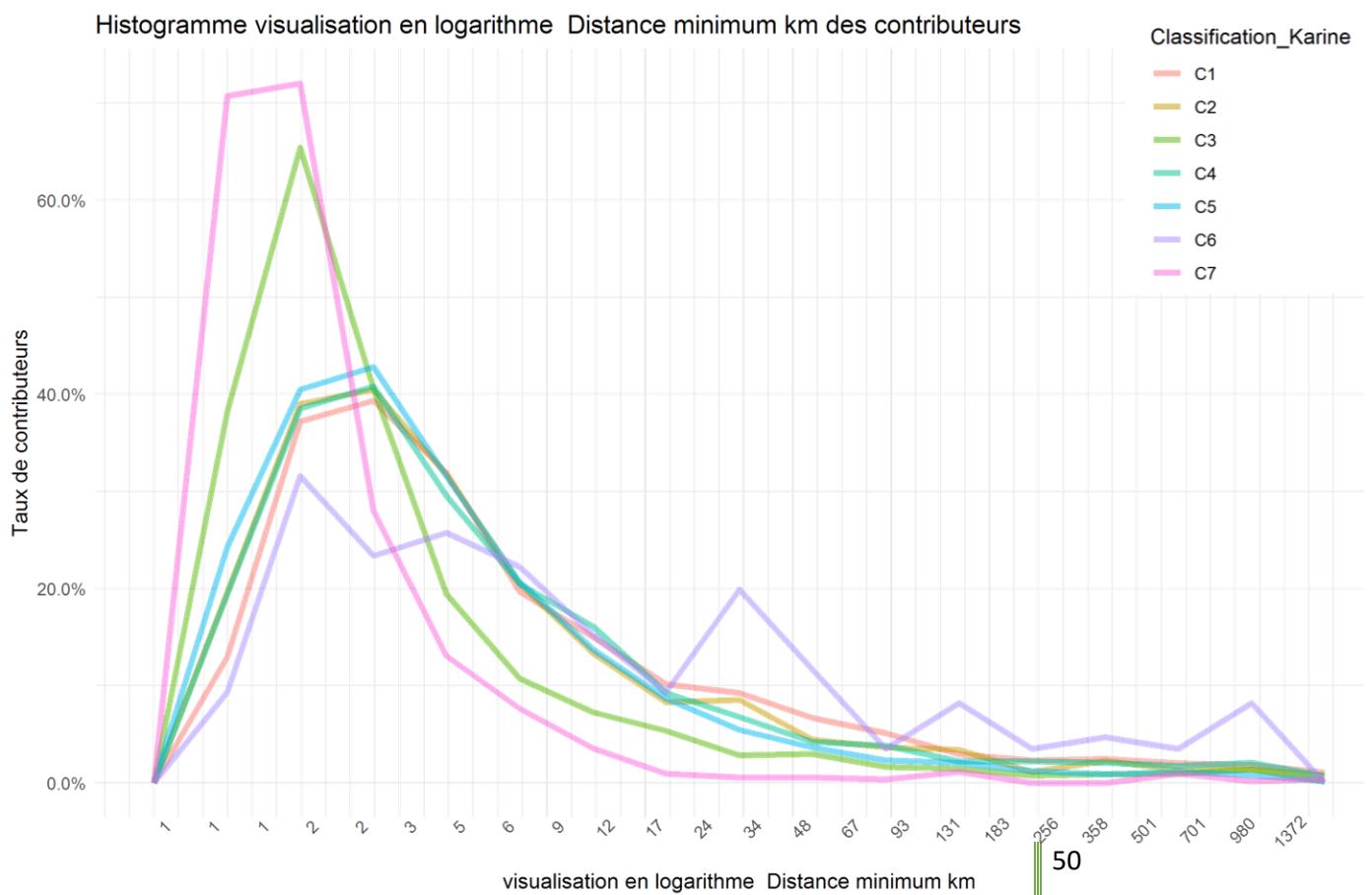
Ce graphique nous permet alors de visualiser trois grands comportements :

Les utilisateurs des typologies C3 et C7 qui réalisent des observations très proches de chez eux avec 75% contributeurs qui ont réalisé une observation à moins de deux kilomètres de leurs lieux d'habitation.

Pour les profils C1, C2, C4 et C5 qui sont pratiquement similaires, 80% des contributeurs réalisent leurs observations la plus proche à entre 0 et 10 km.

Enfin la typologie C6 qui sort du lot, on peut distinguer trois comportements au sein de la typologie. Il y a tout d'abord ceux qui réalisent des observations proches de chez eux moins de 20km comme les autres profils ( $\pm 50\%$  des contributeurs). Ceux qui réalisent leur observation la plus proches à des distances plus éloignées entre 20 et 100 km ( $\pm 30\%$  des contributeurs). Puis ceux qui réalisent leurs observations la plus proche à des distances très éloignés entre 100 et 1000 km ( $\pm 20\%$  des contributeurs).

De façon générale, on peut voir que les contributeurs de Faune France ont tendance à réaliser au minimum une observation proches de leur lieu d'habitation. Seuls certains contributeurs du profil C6 « les cocheurs » n'ont pas de motivation à observer aux alentours de chez eux, et certains sont prêts à parcourir plus de 500km pour réaliser une observation. La encore la consistance de la donnée n'est pas encore la même, car l'effort fourni par le contributeur montre une certaine motivation voire une certaine expertise.



### *Graphique 11 Histogramme de fréquence*

### 5.2.2 Distance moyennes

La lecture de cet histogramme est la même que celle du précédent, la seule différence est l'axe des abscisses qui représente la distance moyenne parcourue par les contributeurs. On peut alors lire que dans la catégorie C7 42% des contributeurs réalisent leurs observations à une distance moyenne entre 40 et 80 km. D'après le graphique on arrive à regrouper trois profils :

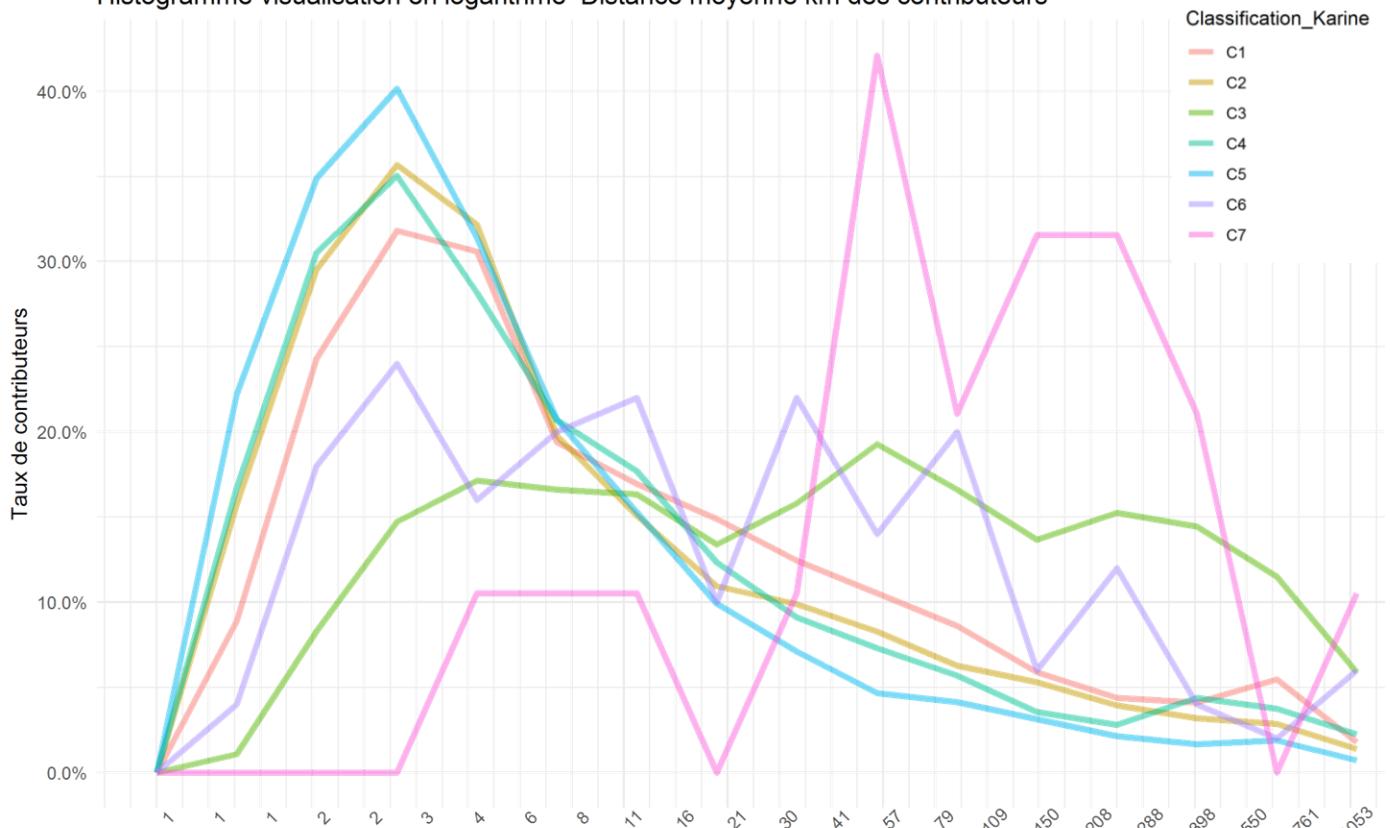
Les profils C1, C2, C4 et C5 ont une distance moyenne très faible, on peut voir que la majorité des contributeurs de ses classes ne dépassent pas les 20km en moyenne pour réaliser leurs observations. Nous sommes donc sur un profil d'utilisateurs ayant une portée spatiale très petite, ce qui montre une certaine limite dans les déplacements de ses contributeurs. Ce qui vient renforcer l'observation locale.

Les profils C3 et C6 quant à eux n'ont pas une distance particulière, la répartition des contributeurs est relativement stable pour toutes les distances (malgré les variations de la classe C6 qui manque peut-être de contributeurs pour rendre le phénomène plus stable). Cette homogénéité témoigne d'une grande variabilité de la portée spatiale. Cette variabilité nous permet d'avoir des comportements de contributeurs différents avec des motivations différentes.

Enfin le profil C7 présente quant à lui une majorité de contributeurs réalisant des observations plutôt éloignées de leurs lieux d'habitation. Pour eux les déplacements sont habituels et ne représentent pas un frein.

Dans l'ensemble on peut visualiser trois grands comportements sur tous les contributeurs confondus. Il y a ceux qui réalisent des observations très proches de chez eux (20km max) qui peut être n'ont pas de motivation à observer au-delà. Ceux qui ont tendance à observer plus loin (20 à 100km) qui ont une volonté d'observer plus loin mais qui sont peut-être limités par des obligations (travail, transport,...). Puis ceux qui observent très loin (100 à 1000 km), qui ont la volonté ou l'obligation d'observer dans des endroits éloignés du lieu d'habitation avec la possibilité de le faire. Attention tout de même, cette distance moyenne, comme expliquée dans la partie 4.3.1.2, n'est qu'une approche des comportements pas une distance réellement parcourue.

Histogramme visualisation en logarithme Distance moyenne km des contributeurs



### 5.2.3 Distance moyenne mensuelle

Sur le graphique N°13 page 53, nous pouvons voir l'histogramme de la distance moyenne mensuelle en kilomètre de tous les contributeurs de Faune France. Nous pouvons lire sur cet histogramme que  $\pm 2750$  contributeurs parcourent entre 1,5 et 2 km pour réaliser leurs observations.

Globalement nous pouvons voir que au moins 25% des contributeurs de Faune France ne parcourent pas plus de 1 kilomètre en moyenne par mois pour réaliser leurs observations. 50 % des contributeurs parcourent jusqu'à 5 kilomètres. On peut voir que l'écart interquartile montre une dispersion pouvant aller jusqu'à 35 km. Enfin la distance moyenne mensuelle d'observation de 54 km montrent que certains contributeurs ont des tendances à observer très loin par rapport à une grande majorité des contributeurs.

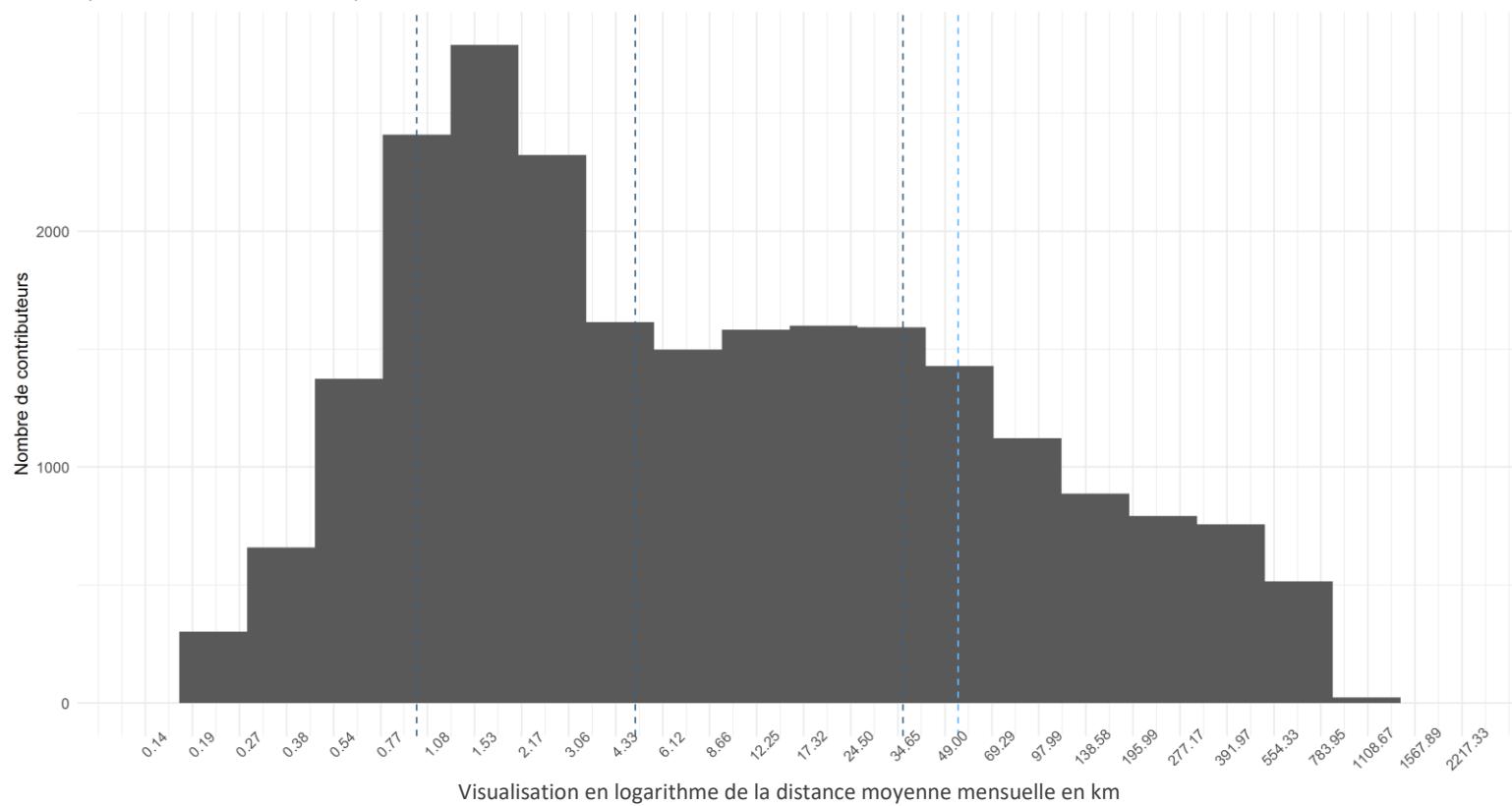
Pour mieux comprendre ces tendances il faut regarder l'histogramme de la distance moyenne mensuelle en fonction des typologies (graphique N°14 page 53). Cet histogramme représente un taux de contributeurs et non un nombre de contributeurs afin de pouvoir comparer plus facilement des comportements entre typologies différentes. On peut alors lire que 45% des contributeurs de la classe C7 réalisent entre 69 et 140 km pour faire leurs observations. Dans l'ensemble on peut distinguer quatre comportements spatiaux différents :

- 1) Tout d'abord les contributeurs qui ne se déplacent pas et observent sur place, avec une distance moyenne mensuelle ne dépassant pas les 4 kilomètres. Ce comportement est très représenté par la classe C5 (oiseaux des jardin) qui a une très grande majorité de contributeurs dans ce cas. On retrouve aussi les classes C2 (oiseaux printemps) et C4 (oiseaux automne) qui ont une majorité de contributeurs dans ce cas. Ainsi que la classe C1 (papillons) avec un tiers des contributeurs.
- 2) Ensuite on peut voir les contributeurs qui ont une certaine mobilité sur le territoire avec des déplacements qui peuvent se rapprocher de la distance domicile travail ( $\pm 30$ km) leurs distances moyennes mensuelles situées entre 5 à 35 kilomètres montrent une certaine volonté de la part des contributeurs à vouloir observer. On trouve majoritairement la classe C3 (observateur régulier) et C6 (cocheurs) dans ce profil. Une petite partie ( $\pm 15\%$ ) des profils C2, C4 et C1 font également partie de ce comportement.
- 3) Il y a aussi des contributeurs avec cette fois-ci une grande mobilité qui montrent une grande motivation pour des déplacements importants, leurs distances moyennes mensuelles sont variables entre 35 et 200 km et leurs distances moyennes journalières entre 35 et 130 km. Ces distances sont relativement importantes et témoignent aussi d'une certaine mobilité. Effectivement tout le monde ne peut pas se permettre de réaliser régulièrement des observations à plus de 100 km. On retrouve principalement voire uniquement dans ce profil la classe C7 (expert).
- 4) Enfin on trouve une petite partie de contributeurs qui ont des mobilités extrêmement grandes, qui témoignent de motivation particulière pour visiter ou découvrir quelques choses en particulier. Leurs distances moyennes mensuelles dépassent les 200 km allant jusqu'à 800 km. Parmi ce comportement on retrouve un tiers des contributeurs du profil C6 ainsi qu'un tiers du profil C1, ainsi qu'une petite partie des profils C2, C4 et C7.

### Distance moyenne mensuelle des contributeurs en 2017 - 2018

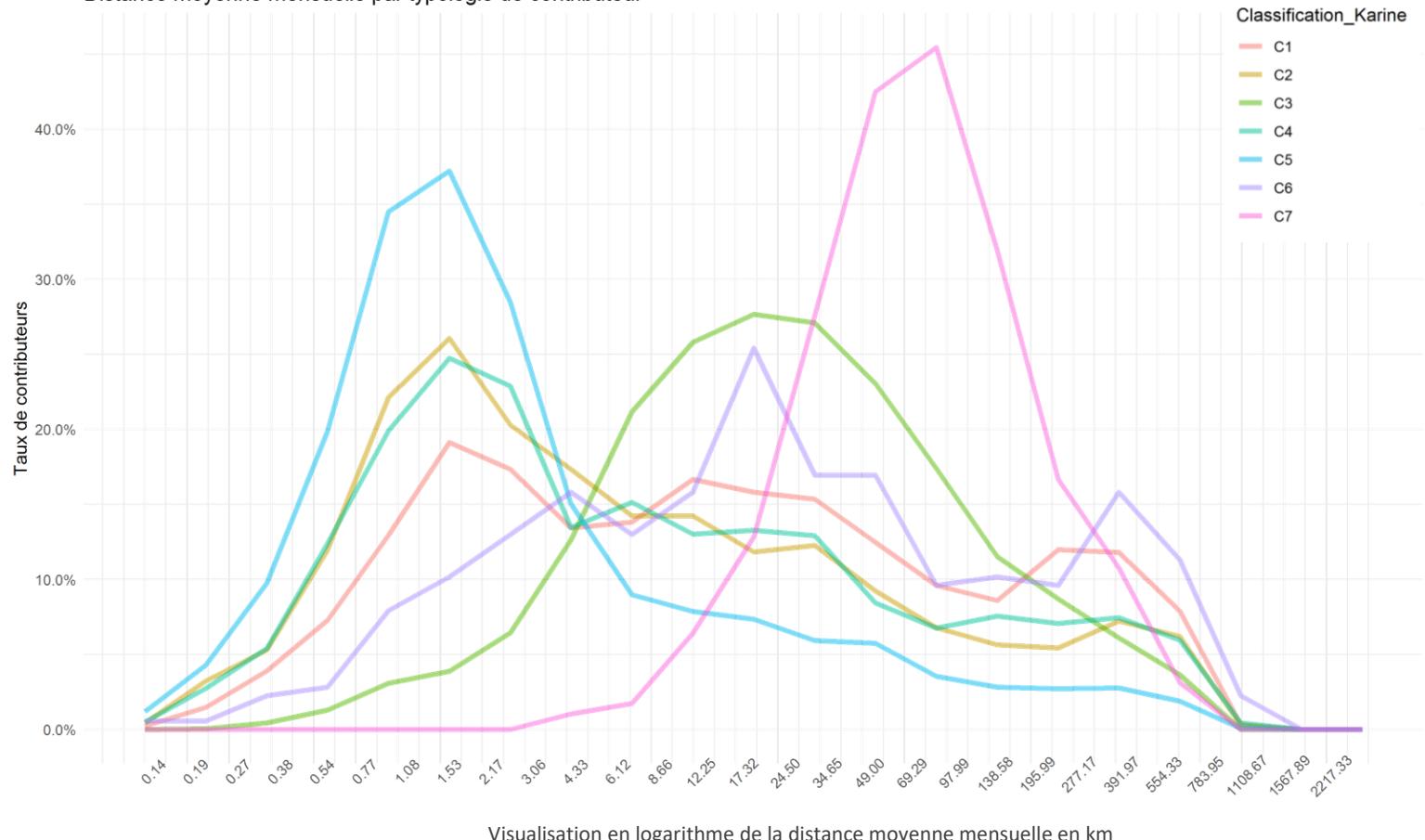
La moyenne est de 54 km  
Q1 est de 1 km

La mediane est de 5 km  
Q3 est de 36 km



Graphique 14 Histogramme

### Distance moyenne mensuelle par typologie de contributeur



Graphique 13 Histogramme de fréquence

#### 5.2.4 Surface totale

La surface totale des observations des contributeurs nécessite au minimum trois points de localisation différents. De part cette contrainte, il n'est pas possible d'avoir cette information pour l'ensemble des contributeurs.

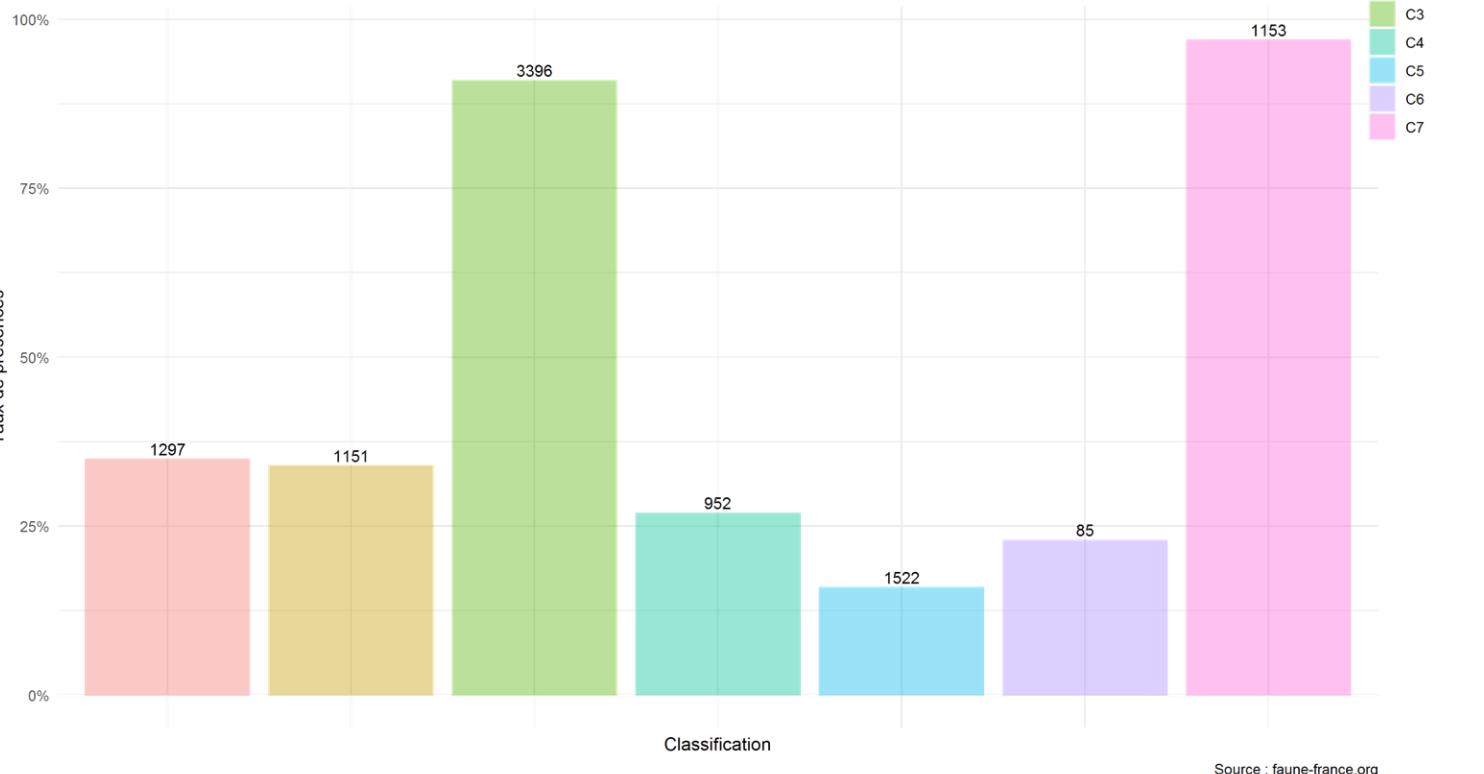
On peut voir que seulement 37,95 % des contributeurs sont représentés par cet indicateur, on s'aperçoit aussi seules les classes C3 et C7 sont correctement représentées par cet indicateur.

Pour les autres classes cela représente seulement entre 25% à 30% des contributeurs, ce qui est peu. Il faudra donc prendre en compte dans l'analyse de la surface totale le fait que certaines classes ne sont pas parfaitement représentées et que les phénomènes s'appliquent parfois seulement à une partie de la classe.

Cependant malgré cette représentation imparfaite des contributeurs, l'échantillonnage de quasiment 38% des contributeurs est suffisant pour être signifiant de comportements spatiaux différents.

#### Représentation des contributeurs

La moyenne est de 23025 km<sup>2</sup>      La mediane est de 1485 km<sup>2</sup>  
 Q1 est de 134 km<sup>2</sup>      Q3 est de 13099 km<sup>2</sup>  
 Représentation de 9556 Soit 37.95 % des contributeurs sont représenté au total



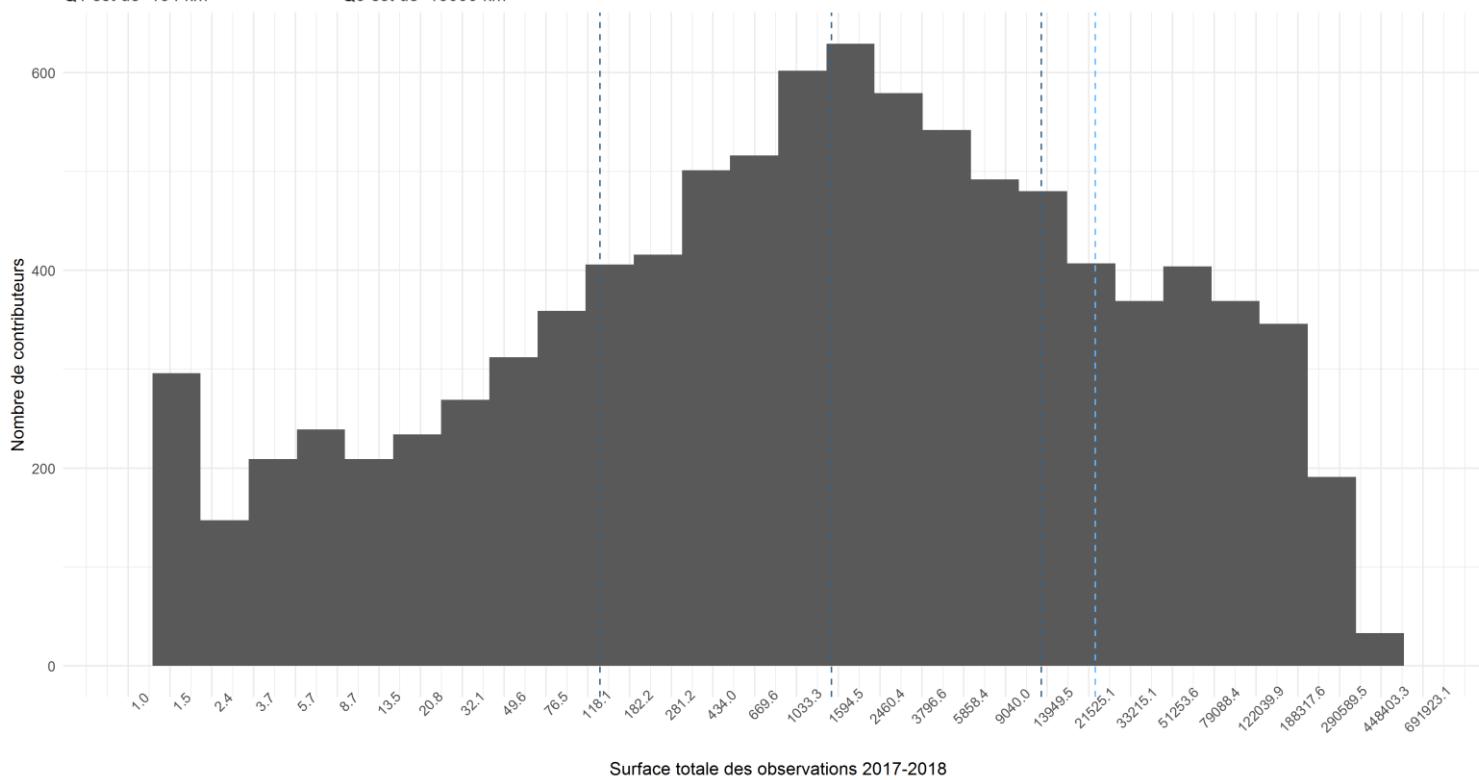
Graphique 15 Diagramme en bâton des typologies

Le graphique N°16 nous permet de visualiser la répartition des contributeurs en fonction de leurs surface totale parcourue sur l'année 2017 et 2018. Cette surface en km<sup>2</sup> et très variable en fonction des contributeurs. Son étendue maximale de 690 000 km<sup>2</sup> montre que certains contributeurs ont parcouru la globalité de la France. La encore, nous pouvons aussi penser que si la surface totale d'un contributeur est trop importante cela peut provenir d'un usage de plusieurs personnes sur un compte. Cependant il reste parfois difficile de distinguer les personnes très motivées parcourant l'intégralité de la France de celle qui sont plusieurs.

Surface totale des contributeur en 2017-2018

La moyenne est de 23025 km<sup>2</sup>  
Q1 est de 134 km<sup>2</sup>

La mediane est de 1485 km<sup>2</sup>  
Q3 est de 13099 km<sup>2</sup>



Visualisation en logarithme de la surface totale en km<sup>2</sup>

Graphique 16 Histogramme

Cet histogramme représente les fréquences des typologies de contributeurs par rapport à leur surface d'observation totale en km<sup>2</sup> visualisé en logarithme. On peut donc lire que 35% des contributeurs du profil C7 ont une surface d'observation totale entre 120 000 et 180 000 km<sup>2</sup>.

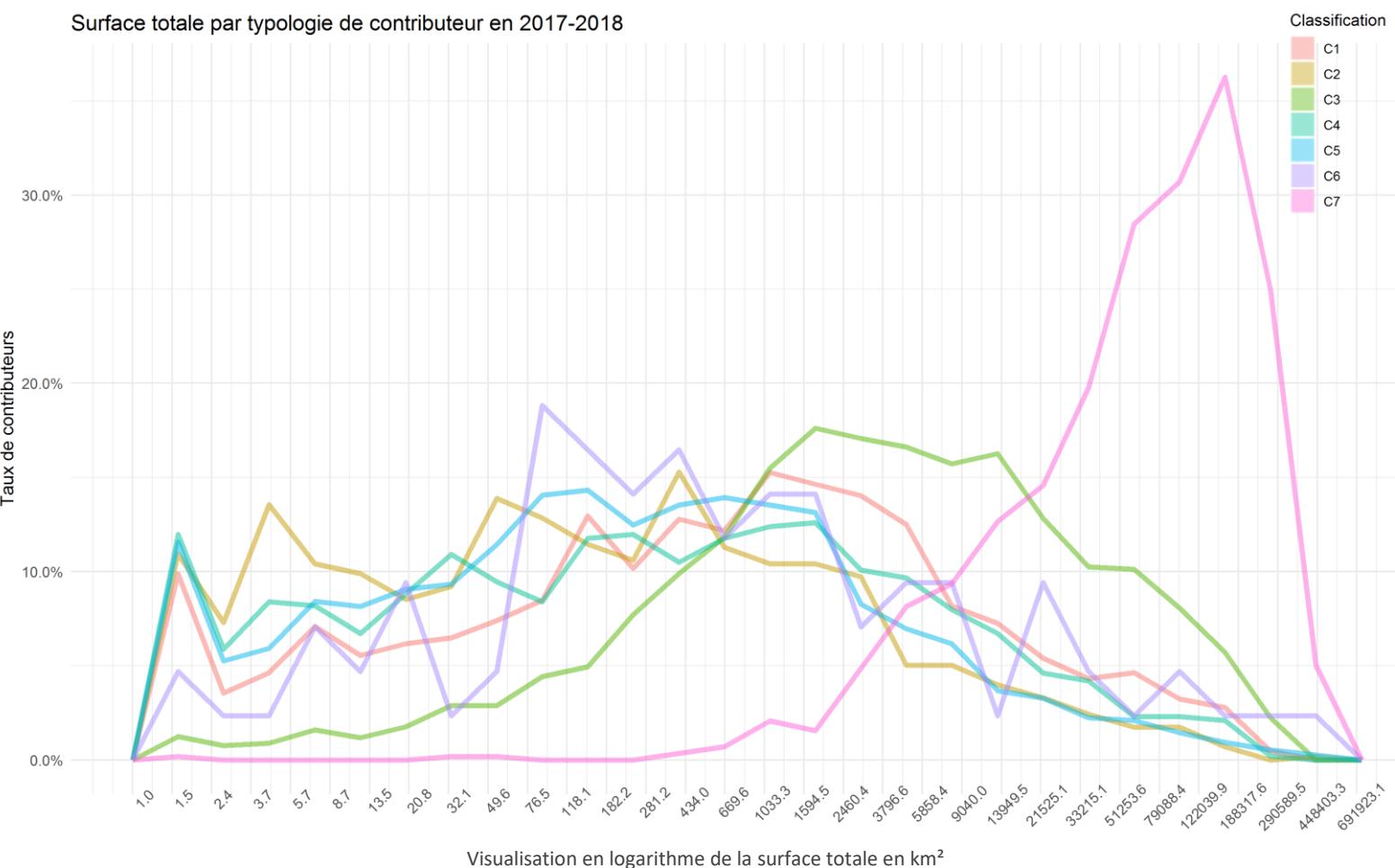
Dans l'ensemble toutes les classes qui sont représentées à seulement 25-30% (C1, C2, C4, C5, C6) ont un comportement similaire avec une surface d'observation qui dépasse rarement les 2 500 km<sup>2</sup>. Leur couverture spatiale est alors relativement réduite, il est alors possible de penser que ces contributeurs réalisent leurs observations souvent aux mêmes endroits dans des lieux bien précis.

Pour la classe C3 qui est plutôt correctement représentée on peut voir une surface d'observation qui a tendance à être plus élevée entre 1 000 km<sup>2</sup> et 30 000 km<sup>2</sup>. Leur couverture spatiale peut parfois couvrir l'échelle d'un ou deux départements, ce qui peut montrer une volonté d'observer dans différents lieux.

Enfin pour la classe C7, on voit un comportement spatial différents avec une surface d'observation entre 30 000 à 400 000 km<sup>2</sup>. On perçoit que l'enjeu n'est pas le même, avec cette fois-ci une couverture spatiale d'envergure nationale. On perçoit une réelle volonté d'observer l'ensemble du territoire.

On perçoit tout de même un petit pic d'observation pour le profil C6 avec une surface totale relativement élevée, ce qui peut témoigner d'un certain éloignement entre leurs observations et donc une certaine volonté d'observer dans différents milieux.

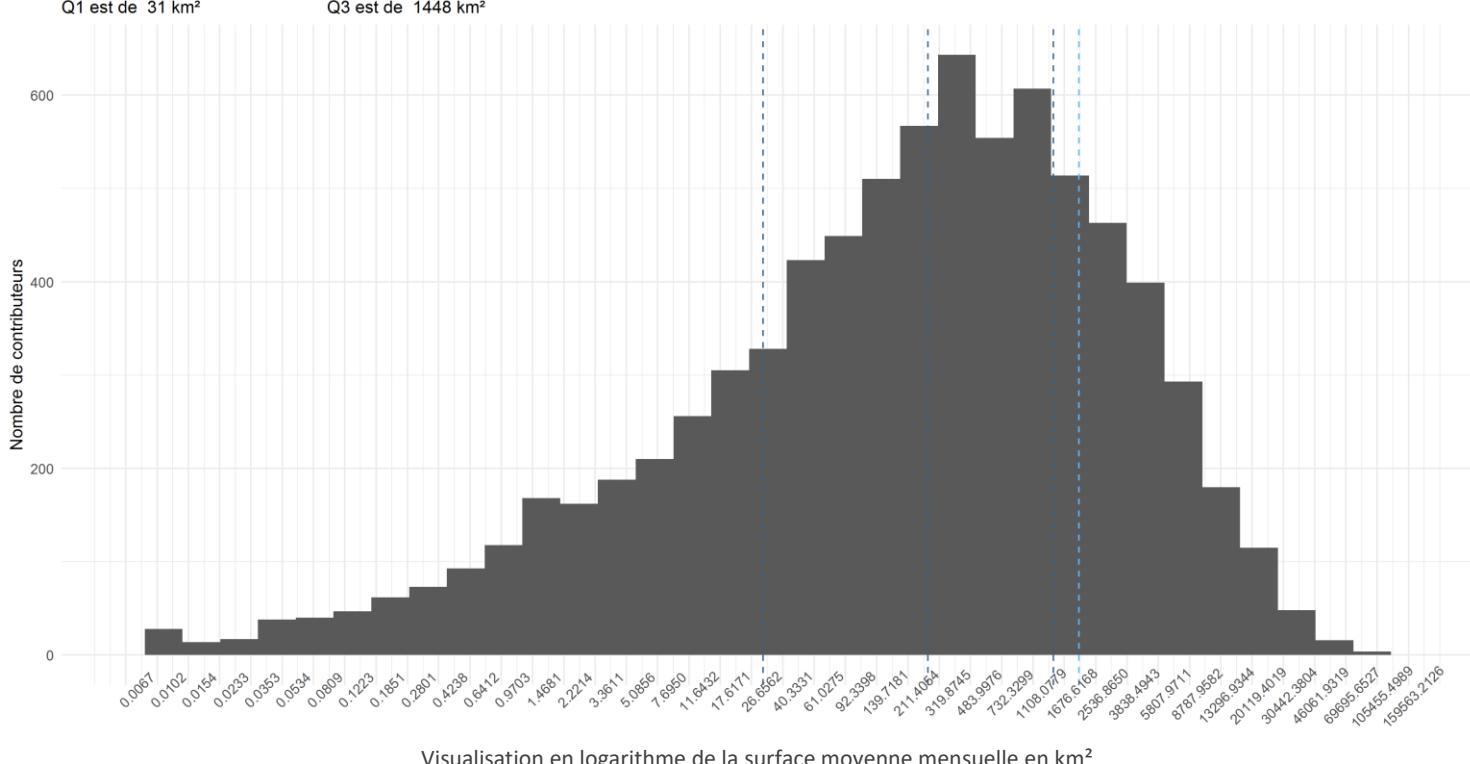
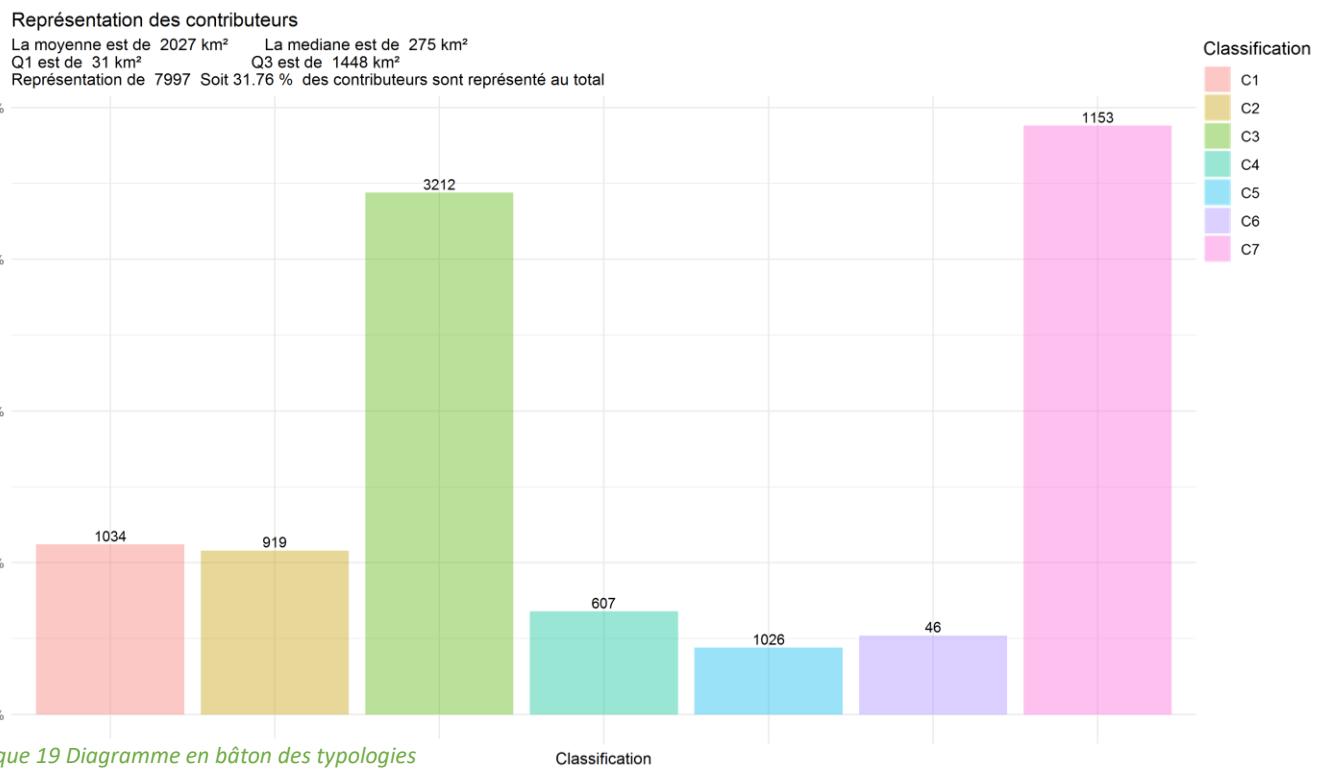
Surface totale par typologie de contributeur en 2017-2018



Graphique 17 Histogramme de fréquence

## 5.2.5 Surface moyenne mensuelle

Tout comme la surface totale, la surface moyenne mensuelle ne permet pas de représenter tous les contributeurs, l'ajout du critères spatiotemporels de manière mensuelle et journalières rend la représentation encore plus restreinte. Effectivement en plus d'avoir au minimum trois point de localisation différents, on exige que ce soit trois points de localisation dans le même mois (voir dans le même jours). Cependant l'échantillons de 30% reste là encore relativement représentatif et peut nous permettre de visualiser certaine tendance.



Graphique 18 Histogramme

Cet histogramme nous permettent de voir la surface moyenne d'observation mensuelle en km<sup>2</sup> pour chaque typologie de contributeurs. La visualisation en logarithme nous permet de mieux visualiser les différentes courbes des typologies. De cette façon nous pouvons voir que 36% des contributeurs de C7 observent mensuellement une surface entre 2 500 et 4 000 km<sup>2</sup>. Grâce à lui nous pouvons supposer trois comportement spatial différents.

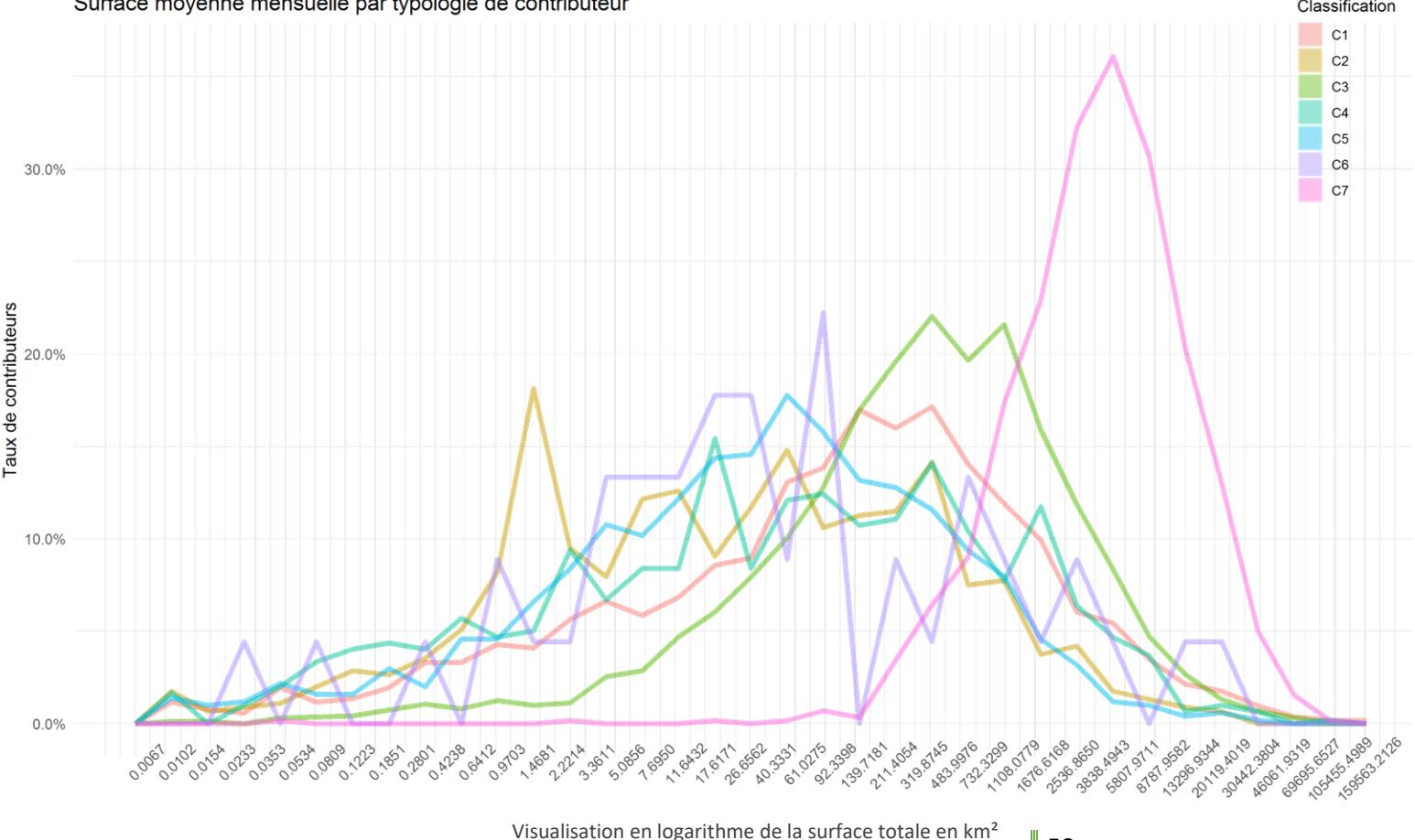
Tout d'abord la classe C7 qui a une surface d'observation moyenne mensuelle immense située entre 1 000 à 13 000 km<sup>2</sup>. Ce qui représente la surface d'observation totale (analyse précédente) de plus 75% des contributeurs tous profils confondus. Ce comportement montre que les méthodes d'observations mensuelles ont pour objectifs de visualiser une grande partie d'un territoire.

Ensuite il y a la classe C3 et C1 qui ont une surface d'observation moyenne mensuelle assez élevée, entre 100 et 1 000 km<sup>2</sup>. Ce comportement se rapprochant de celui des experts reste assez éloigné on peut alors penser différents scénarios, les contributeurs n'ont pas le temps libre nécessaire pour réaliser toutes les observations qu'ils souhaiteraient, ou bien leurs méthodes d'observation sont moins étendues sur le territoire ou n'ont pas volonté d'être plus entendue sur le territoire.

Puis les classes C2, C4 et C5 qui ont globalement une surface d'observation moyenne mensuelle faible situé entre 1 et 100 km<sup>2</sup>. Une petite distinction entre le profile C2 (oiseau printemps) qui tend plus vers une surface de 1 km<sup>2</sup> et C4 (oiseau automne) qui tend plus vers une surface de 100 km<sup>2</sup>. On peut alors voir ici une volonté d'observer seulement dans certains lieux bien définis.

Enfin la classe C6 qui est décomposée entre deux comportements, à la fois une surface moyenne mensuelle faible, donc une partie des contributeurs qui souhaitent aller dans des milieux bien précis, mais à la fois des surfaces élevées donc une étendue de territoire très grande. Sachant que ce profil a tendance à réaliser peu d'observations mais souvent à une grande distance, cela peut créer avec peu d'observations dans des lieux précis et éloignés une grande surface.

Surface moyenne mensuelle par typologie de contributeur



Graphique 20 Histogramme de fréquence

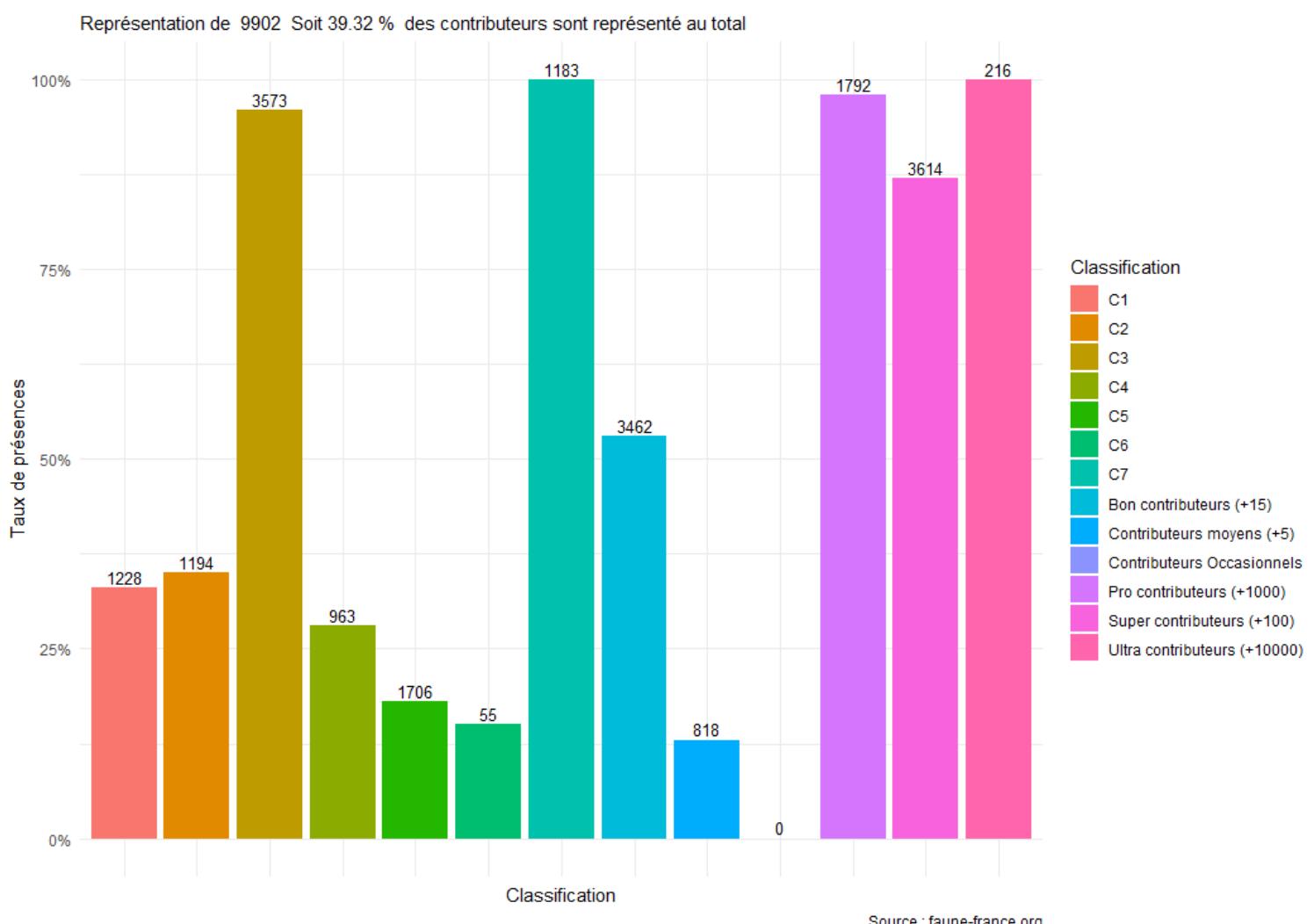
### 5.3 Structuration spatiale

La structuration spatiale réalisée par le calcul de clustering n'est pas réalisable pour tous les contributeurs (car elle nécessite un certain nombre d'observations et de localisations).

On peut voir sur le graphe ci-dessous que seulement 40% de l'intégralité des contributeurs Faune France ont une possibilité de clustering. Dans ces 40% on peut voir que certaines typologies sont très impactées. Seules C3 et C7 sont représentées en totalité alors que les autres profils représentent moins de 30%. Cependant sur ces 30% il y a peut-être des comportements intéressants qui peuvent se distinguer des comportements généraux.

En parallèle, on peut voir à travers la classification sur le nombre d'observations que les possibilités de clustering est directement liée au nombre d'observations. Au-dessus de 100 observations, les clusters sont très majoritairement faits, entre 100 et 15 observations il y a une chance sur deux de pouvoir réaliser un cluster et en dessous de 15 observations la possibilité de cluster est quasi nulle.

Les résultats étant parfois très hétérogènes les différentes analyses sur les clusters sont à revoir avec un paramétrage plus adapté. Il est quand même possible d'observer les résultats en annexe pages [75](#)-[76](#)



Graphique 21 Diagramme en bâton des différentes classifications

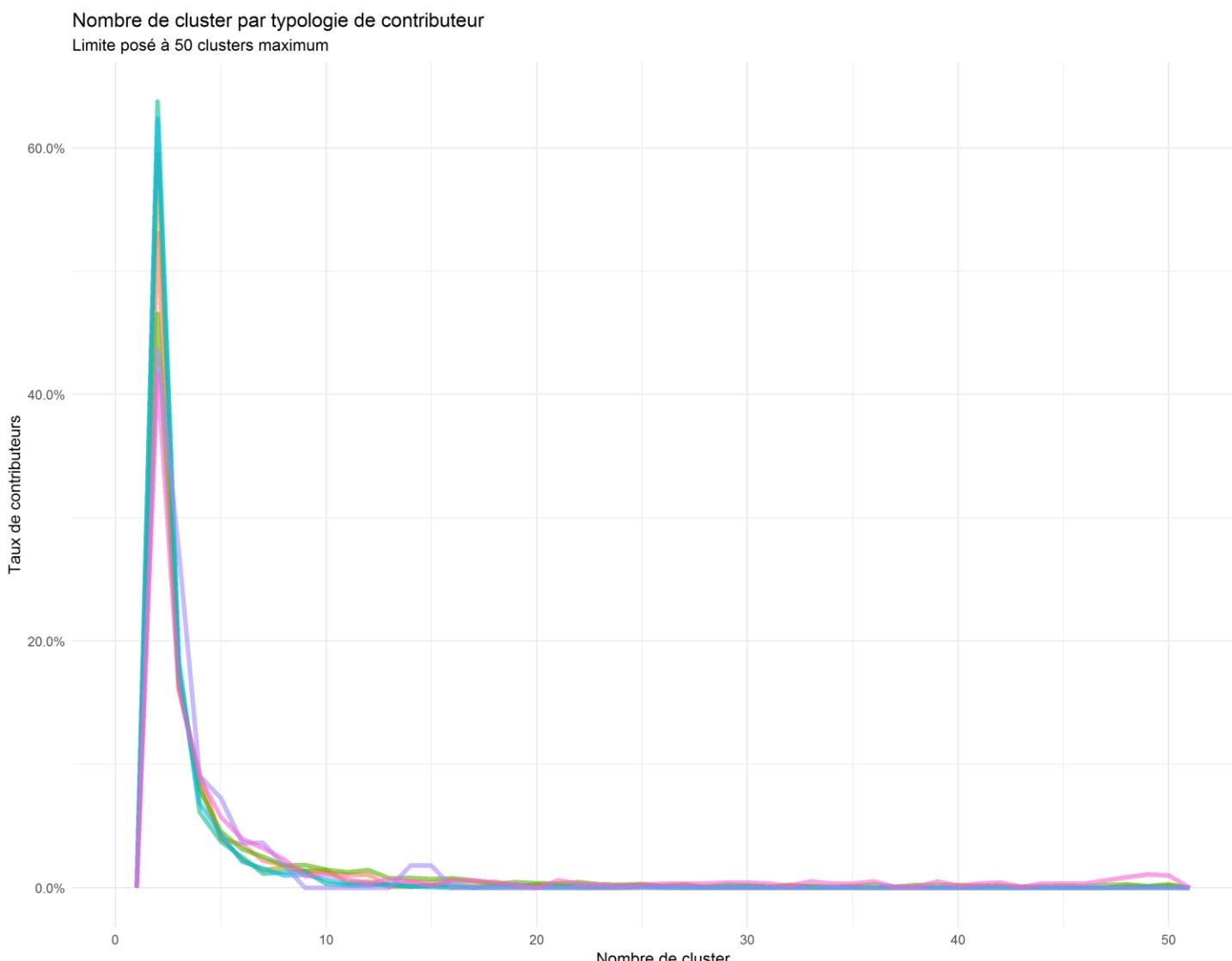
### 5.3.1 Nombre de clusters

Nous pouvons voir que le nombre de cluster reste majoritairement très faible entre deux et trois clusters pour l'ensemble des typologies. Cet indice nous indiquerait que les contributeurs de Faune France ont tendance à observer seulement dans quelques lieux différents ou éloignés.

Les typologies C7 et C6 ont une légère part de contributeur ayant plus de clusters que la majorité, on retrouve alors presque 5% des contributeurs C6 qui ont entre 12 et 15 clusters et 6% des contributeurs C7 qui ont entre 45 et 50 clusters. Pour ces contributeurs, nous pouvons penser que le nombre de lieux observés est plus important que les autres, ce qui peut être lié à leur méthode d'observation qui consiste à observer dans de nouvelle zone.

Ce nombre de clusters qui nous permet d'avoir un indice sur le nombre de zones d'observations reste cependant assez loin de la réalité. La taille des clusters peut beaucoup varier, être parfois très précis avec une délimitation très représentative, parfois très flous avec des clusters couvrant la moitié de la France.

Les différents indices des clusters ne sont donc pas exacts pour la représentation de la structuration spatiale. Cependant peut-être que ses valeurs peuvent nous apporter une indication utile délimitant différents comportements à travers des analyses plus poussées telles que l'analyses factorielle.



Graphique 22 Histogramme de fréquence

## 6 Conclusion

### 6.1 Les différents comportements des contributeurs

Grâce aux différents indicateurs mis en place pour les analyses spatiales nous arrivons à percevoir différents comportements spatiaux des utilisateurs. Ceux-ci sont parfois liés à une certaine typologie de contributeurs déterminés auparavant à travers d'autres données. D'autres fois nous pouvons remarquer une distinction de comportements au sein d'une typologie, ce qui permet d'affiner les typologies. Parmi ses comportements spatiaux on retrouve trois profils globaux :

1. Les contributeurs sédentaires, observant souvent les mêmes lieux relativement proches de leur commune d'habitation, voire uniquement celle-ci. Ils ne présentent pas de volontés d'observer plus loin et leur zone d'observation reste relativement petite et locale. On a alors un effet de concentration de leurs observations sur une zone déterminée. On retrouve quasiment en permanence parmi ces profils les typologies C5, C2 et C4.
2. Les contributeurs visiteurs, essayant d'observer des lieux différents, leurs observations sont faites sur des zones petites. Cependant pour certains d'entre eux, ces zones très proches de leur domicile alors que pour d'autres elles sont très éloignées. On a alors cette fois-ci un effet de dispersion des observations. C'est très marqué pour la typologie C6 ainsi qu'une partie de C1.
3. Les contributeurs réguliers, vont observer dans beaucoup de lieux, mais repassent aussi par les lieux déjà visités. Ils n'ont pas de distance d'observation privilégiée, certains vont assez loin d'autres restent très proches de leur lieu d'habitation. De manière générale leur échelle d'observation reste souvent à un niveau départemental. Leurs surfaces d'observation commencent à être vastes, avec des étendues de concentration couvrant une superficie importante. On peut penser que ce profil est limité par certaines contraintes qui l'empêche d'aller plus loin. Ce profil concerne uniquement la typologie C3.
4. Les contributeurs à très grande mobilité parcourent de grandes distances pour aller observer sur toute une région. Ils vont parfois observer la moitié voire l'intégralité de la France. Leur surface d'observation est très grande en multipliant les lieux et en repassant par ceux-ci. Ils n'ont aucune contrainte et montrent une motivation sans limite pour observer. Cela concerne uniquement le profil C7.

Ces comportements peuvent nous permettre de comprendre certaines motivations et profils des utilisateurs mais peuvent être interprétés comme une limite de la science participative. Effectivement, les différentes méthodes d'observations dans l'espace peuvent impacter la représentation de la réalité. Cependant chaque méthode est correcte et peut apporter un avantage.

Pour exemple, les contributeurs visiteurs, qui observent plus rarement les mêmes lieux vont permettre d'éviter une redondance d'information. Les contributeurs réguliers qui vont repasser par les mêmes lieux vont permettre une certaine connaissance du terrain donc de la faune déjà présente et de son évolution. Les contributeurs à très grande mobilité qui observent de grandes surfaces fournissent une sûreté de l'information.

La science participative présente l'avantage d'une récolte de données importantes par des acteurs aux profils divers, mais aussi la difficulté à cerner la qualité de la donnée. L'analyse des typologies des contributeurs et de leur comportement spatial permet mieux comprendre comment Faune France doit traiter ses données.

## 6.2 Evaluation des solutions mis en place

Cette méthode d'analyse spatiale reste exploratoire ; certains indicateurs ne sont pas représentatifs de la réalité, d'autres sont difficiles à interpréter. Cette partie du mémoire est consacrée à un regard critique sur le travail fourni pour l'améliorer si d'autres études doivent être réalisées.

1. La différence du nombre d'observations des utilisateurs impacts énormément les indicateurs spatiaux mis en place. Même si la technique de représentation par taux permet d'éviter cette impact la lecture, la lecture par taux tronque aussi l'information.
2. La visualisation des indicateurs spatiaux en fonction des typologies nous permet de mieux percevoir certain comportement mais la comparaison ou association entre typologie n'est pas à réaliser à chaque fois. Notamment lorsque les indicateurs nous montrent un comportement similaire entre typologie, il est possible que la réalité et les motivations des contributeurs soi différentes.
3. La fonction de clustering dans R est très complexe, je l'ai mis en place avec des paramétrages et une méthode qui me semble être correct cependant il est possible que d'autres paramètres ou d'autres indicateurs permettent de réaliser des analyses plus représentatives des contributeurs.

## 6.3 Pistes d'analyses envisageable

Toujours dans l'objectif d'améliorer cette étude, pour pallier les différents problèmes évoqués précédemment mais aussi d'aller plus loin dans la visualisation de certain comportement voici différentes pistes d'analyses et de méthode envisageable :

1. La mise en place d'un échantillonnage des données pour réaliser des analyses pointues qui sont parfois chronophages, mais aussi pour limiter les effets de dépendance entre variable.
2. Une analyse de clustering prenant en compte plus de paramètres tel que la date de l'observation, les espèces d'animaux ou le type de paysage.
3. Une analyse factorielle avec les indicateurs spatiaux les plus pertinents. Pour améliorer la typologie des contributeurs mais aussi réaliser des groupements de comportement de manière plus méthodologique.
4. Une analyse des comportements spatiaux des contributeurs à travers une visualisation animé de leurs déplacements en fonction du temps. Pour représenter, illustré les différents comportements des contributeurs et ainsi mieux percevoir leurs méthodes d'observations.  
*(Début de cette analyse commencer, voir dans la fonction : R\_Visu\_carto\_intensity.R )*
5. Une matrice de distance entre les observations et la route la plus proche, pour comprendre les motivations et facilité de déplacements des contributeurs.
6. Calculer la concentration d'observation (Kernel estimation) des contributeurs sans le rendre dépendant du nombre d'observation  
*(Début de cette visualisation commencer, voir dossier fonction le script : R\_Recuperation\_info\_surface\_perimetre\_journalier.R avec certaine ligne de visualisation mise en commentaire)*

#### 6.4 Reproductibilité et réutilisation des scripts

Les scripts permettant l'analyse statistique générale ainsi que l'analyse spatiale sont conçus pour pouvoir être réutilisés par la suite. Si la base de données vient à être changée avec de nouvelles données (2019, 2020) ; il sera alors possible de relancer les mêmes analyses avec les scripts afin de comparer et voir les différentes évolutions.

Chaque ligne des scripts sont commentées afin qu'ils soient faciles de comprendre leurs fonctionnements, et de pouvoir les réutiliser ou les modifier très facilement. De plus les logigrammes en annexe pages 64-67 permettent de voir les différentes étapes mise en place dans le traitement des données pour faciliter leur compréhension.

Enfin si ses scripts venaient à être réutilisés sur différentes années, il est possible que de nouveaux comportement voient le jour tout comme une typologie de contributeurs peu changer de comportement ou se diviser en adoptant différents comportements. Ainsi il serait alors possible de percevoir des changements de typologie pour certains contributeurs ce qui pourrait nous aider à identifier des comportements de prédiction (exemple : 30% des contributeurs de C2 observateurs des oiseaux en printemps pourraient passer sur un profil C3 observateurs réguliers)

A propos des contributeurs ayant fait quelques observations car leur inscription est récente, il est aujourd'hui difficile de savoir si ce contributeur va arrêter de réaliser des observations, s'il participera seulement à des événements de Faune France ou s'il va tendre à devenir un ornithologue amateur. C'est peut-être le comportement spatial couplé à d'autres informations sur plusieurs années qui peut nous indiquer que ses motivations vont le faire évoluer vers un profil différent.

Ainsi il sera possible pour Faune France de mieux comprendre et cerner les différents profils de leurs contributeurs. Ce profil pourra alors permettre une meilleure prédiction de leurs comportements à venir, mais aussi de mieux analyser les données d'un contributeur selon qu'il s'agit d'un amateur ou d'un expert.

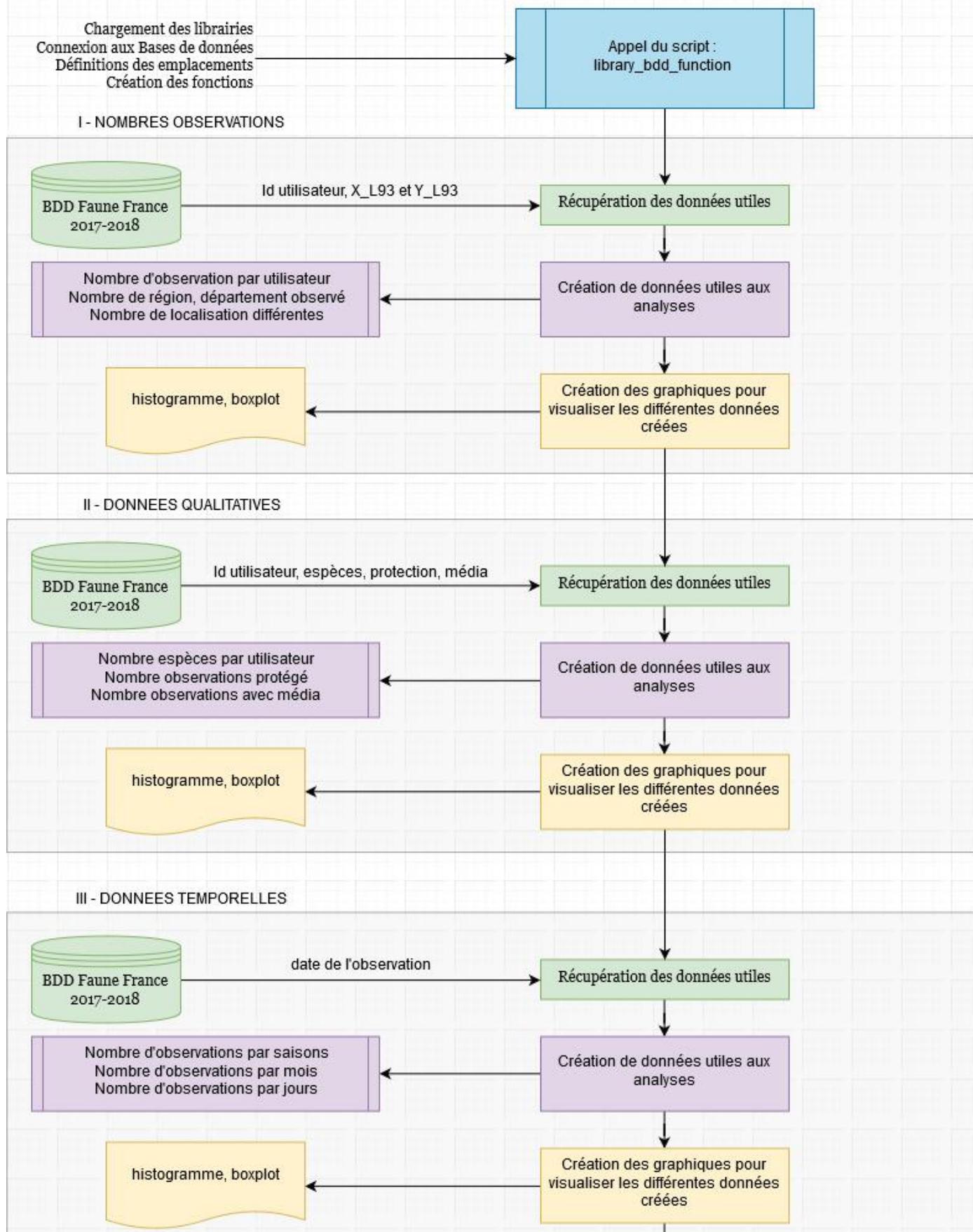
Cette différence de traitement de la donnée entre différents profils permettra alors d'améliorer les résultats des méthodes de science participative.

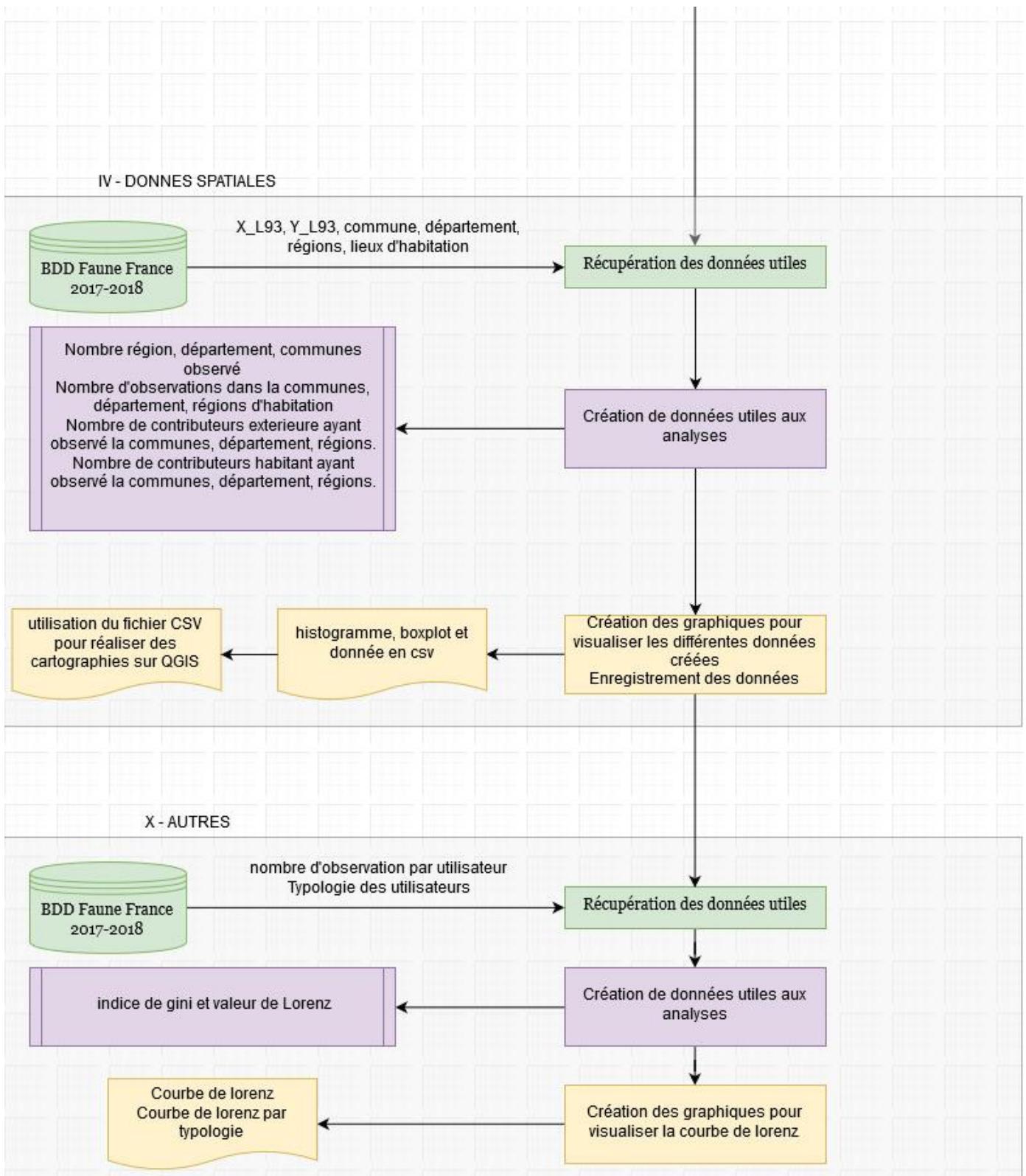
## 7 Annexes

### 7.1 Logigrammes des scripts et fonctions

#### 7.1.1 Données statistiques

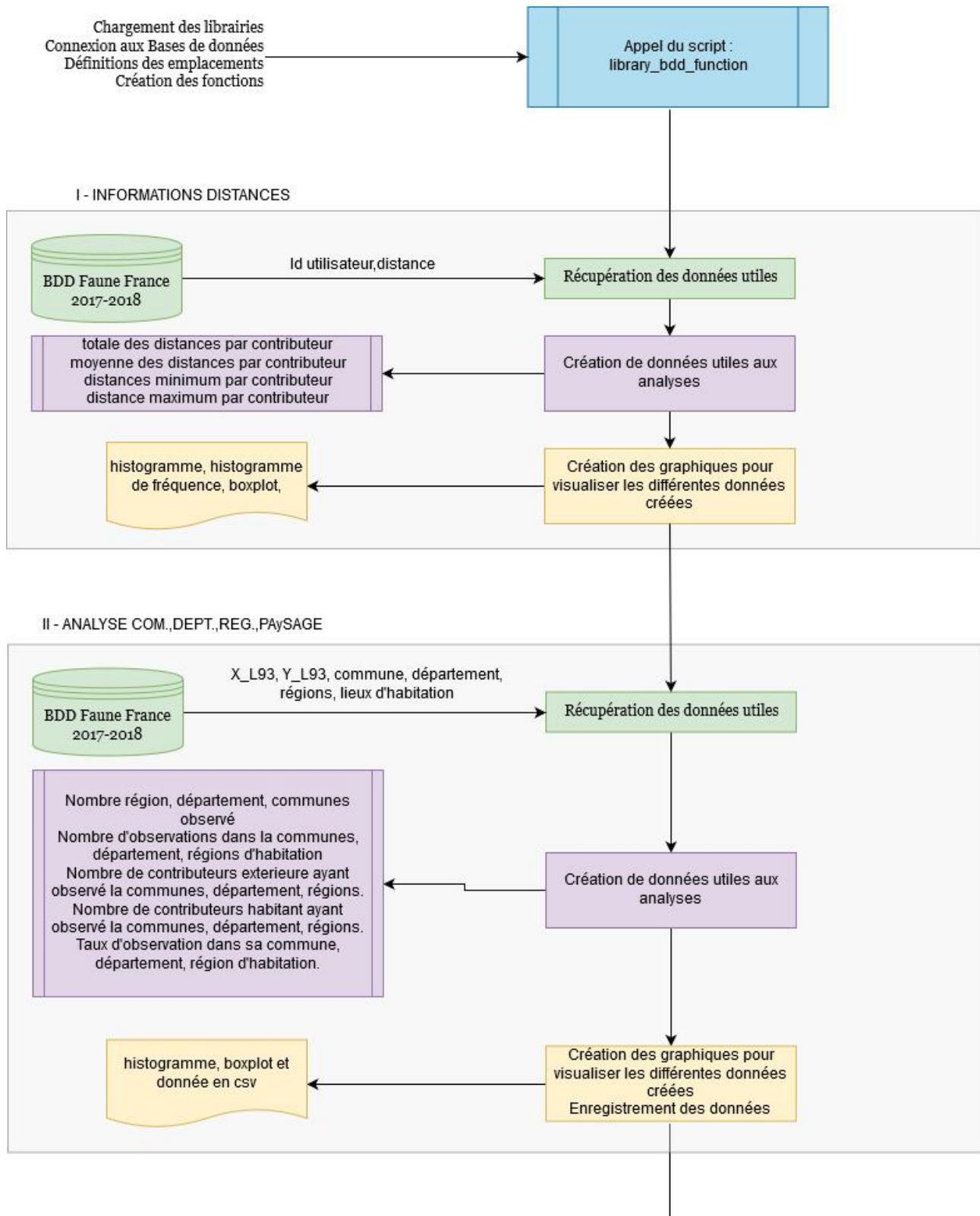
#### Script : R\_3\_Analyse\_globale.R

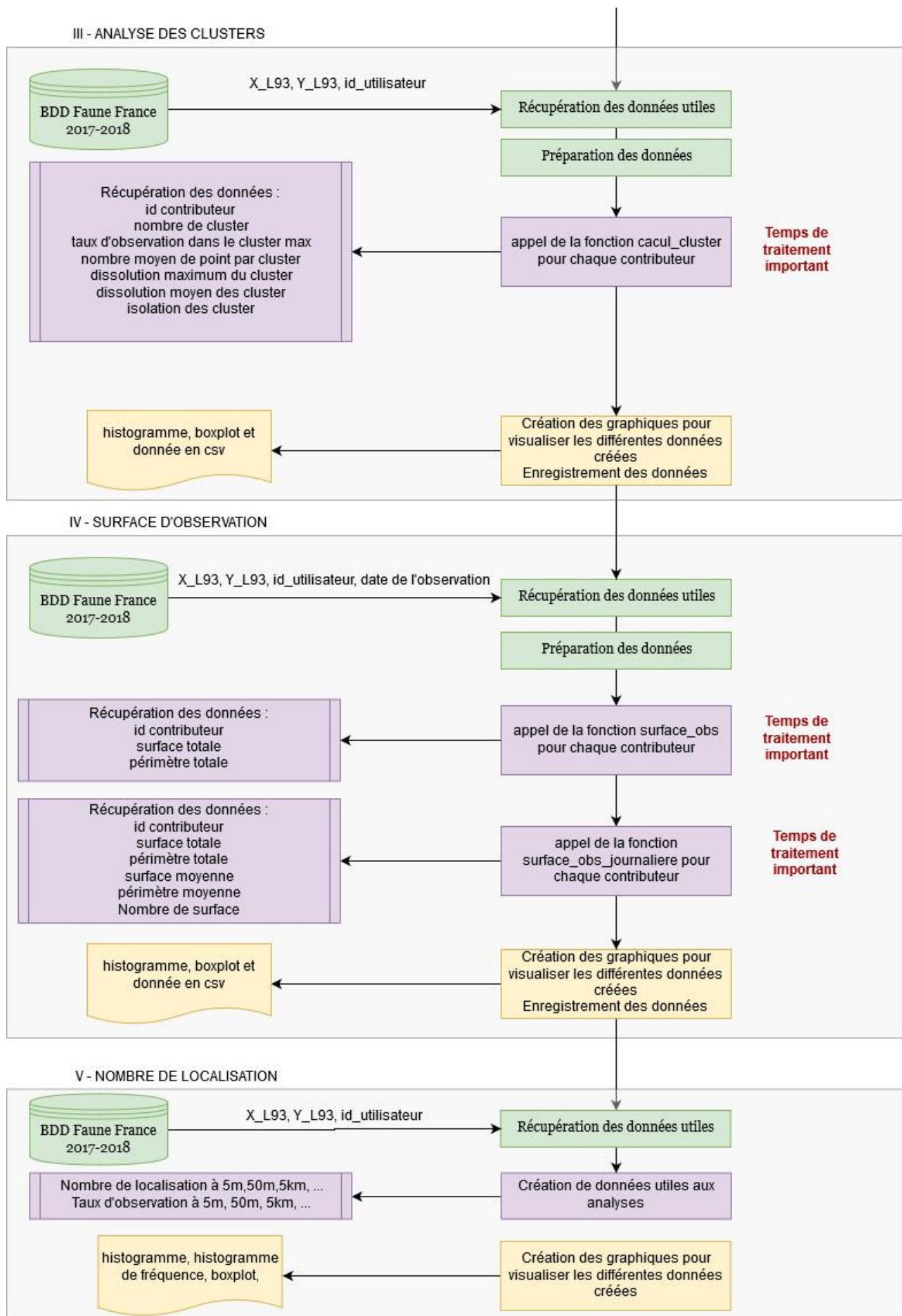




### 7.1.2 Données spatiales

#### Script : R\_4\_Analyse\_spatiale.R





## 7.2 Bibliographie des fonctions R

### 7.2.1 Géotraitements

- `st_as_sf` : transf
- `st_centroid` : récupère le centroïde d'un segment ou d'un polygone
- `st_cast` : groupe différentes géométries (couches) en une géométrie (ligne en multiligne)
- `st_union` : permet d'assembler une géométrie (couche) (multipolygone en polygone)
- `st_area` : retourne la superficie en m<sup>2</sup> (si l'objet géométrique est créé à partir de coordonnée L93)
- `st_lenght` : retourne la longueur d'une géométrie en m (si l'objet géométrique est créé à partir de coordonnée L93)
- `st_distance` : retourne la distance entre deux points en m (si les objets géométriques sont créés à partir de coordonnée L93)
- `concaveman` : permet de transformer des multipoints en un polygone

### 7.2.2 Clusterings

- `Clara` : permet de créer les clusters à partir de différents paramètres avec la méthode clara.
- `fviz_nbclust` : permet de visualiser le nombre de cluster optimal

### 7.2.3 R to SQL

- `dbDriver` : permet de définir le type de communication de la base de donnée
- `dbConnect` : permet d'établir une communication entre R et une base de donnée
- `dbExistsTable` : permet de vérifier si une table existe dans une base de donnée
- `dbExecute` : permet d'exécuter une requête SQL dans une base de donnée
- `dbCreateTable` : permet de créer un table dans une base de donnée
- `dbWriteTable` : permet d'écrire dans une table dans une base de donnée

### 7.2.4 Lecture de fichier

- `read.delim` : permet de lire un fichier de donnée en indiquant une délimitation
- `load` : permet de chargé un fichier R
- `read_excel` : permet de lire un fichier excel
- `read.csv` : permet de lire un fichier CSV
- `cbind` : permet de combiner deux tableaux (l'un à côté de l'autre ajout de colonne)
- `rbind` : permet de combiner deux tableaux (l'un en dessous de l'autre ajout de ligne)

## 7.3 Requêtes SQL

### 7.3.1 Les point de localisation des observations

```
UPDATE FICHIER_RECEPTION
SET point_geom84 = ST_SetSRID(ST_MakePoint(fichier_reception.Lon_WGS84,
fichier_reception.Lat_WGS84),4326);
```

### 7.3.2 Les point de localisation des contributeurs

```
UPDATE utilisateur as ut1 SET geom =
  (SELECT ST_Centroid(com_wgs84.geom)
   FROM utilisateur as ut2, com_wgs84
   WHERE (ut2.code_insee_fin = com_wgs84.insee_com)
   AND ut1.id = ut2.id
   LIMIT 1) ;
```

### 7.3.3 Matrice de distances

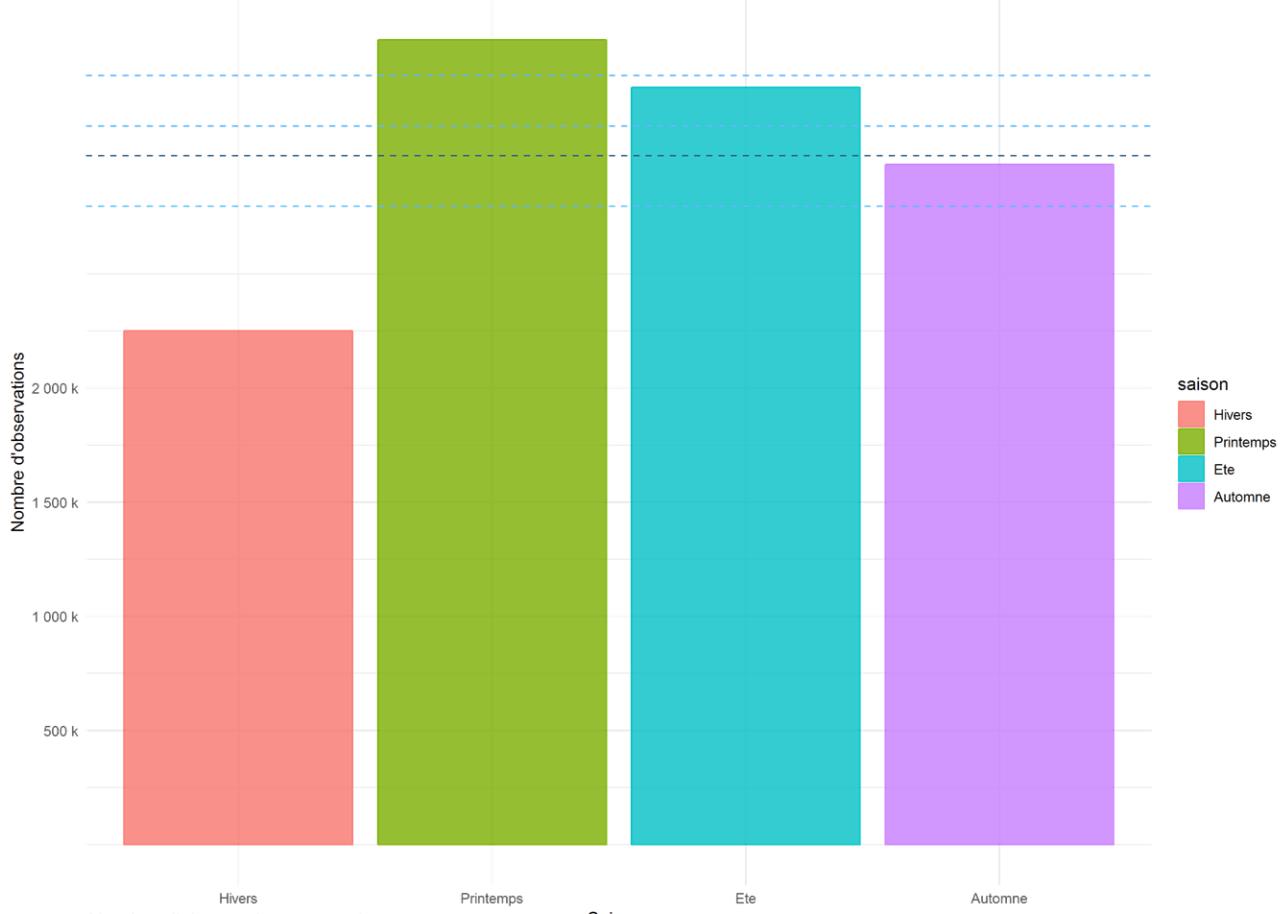
```
INSERT INTO matrice_distance(id_utilisateur, id_localisation, distance,
geom)
  SELECT utilisateur.id, point_localisation.id,
ST_Distance(ST_Transform(point_localisation.geom,2154),
ST_Transform(utilisateur.geom,2154)),
ST_SetSRID(ST_MakeLine(ST_Transform(point_localisation.geom,2154),
ST_Transform(utilisateur.geom,2154)),2154)
  FROM utilisateur, observations, point_localisation
  WHERE observations.id_point = point_localisation.id
  AND observations.id_utilisateur = utilisateur.id;
```

## 7.4 Graphiques statistiques complémentaires

### 7.4.1 Autres visualisations des analyses temporelles

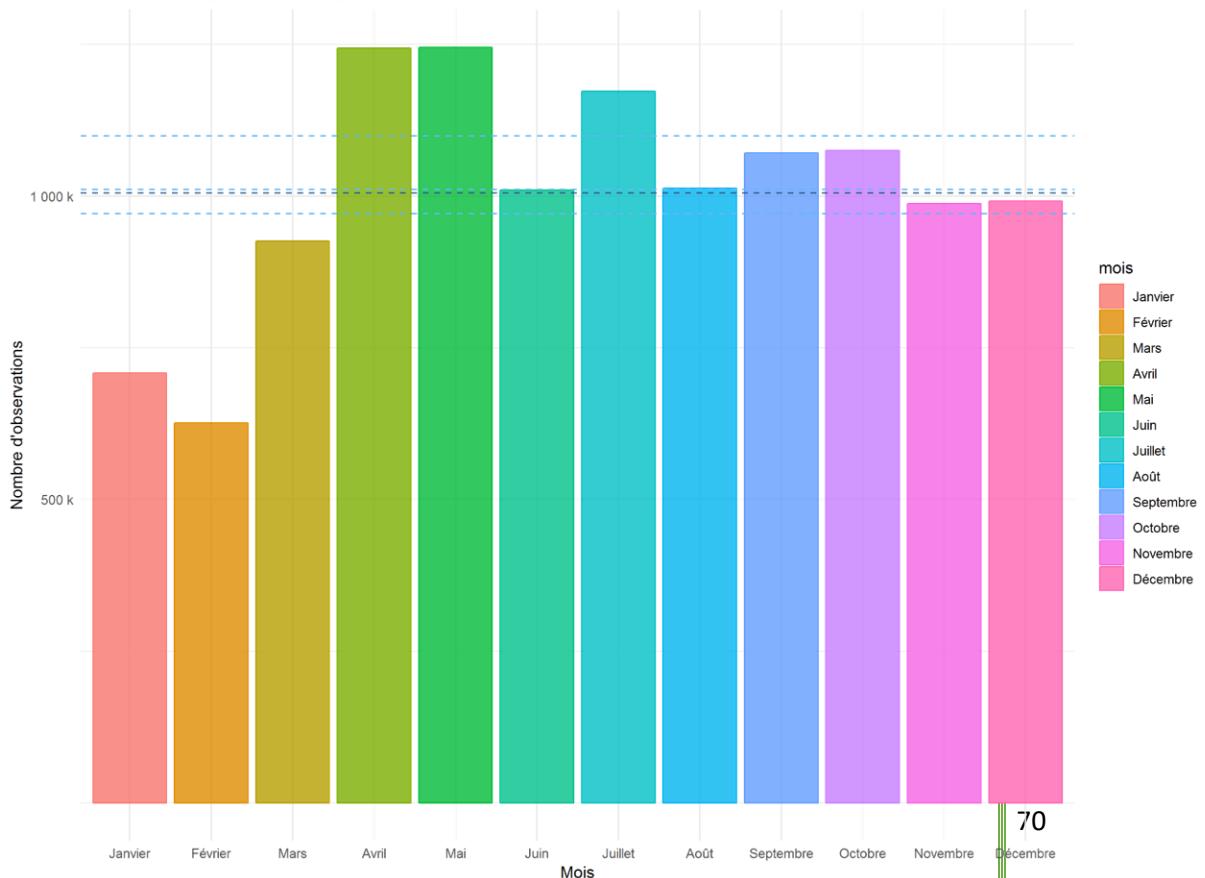
#### Nombre d'observations par saison sur 2017-18

La moyenne d'observations est de 3017 k  
 Q1 = 2796 k    Q2 = 3147 k    Q3 = 3368 k



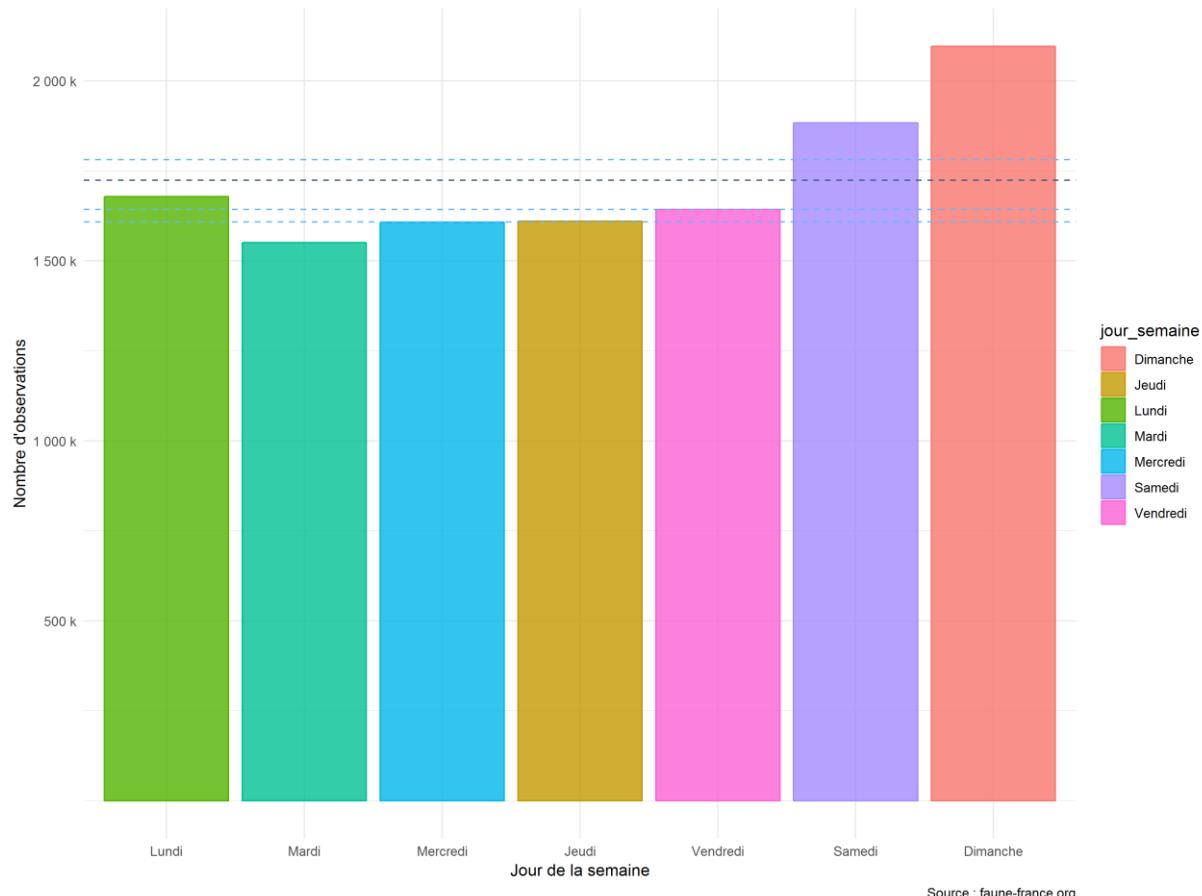
#### Nombre d'observations par mois sur 2017-18

La moyenne d'observations est de 1005 k  
 Q1 = 971 k    Q2 = 1011 k    Q3 = 1099 k



### Nombre d'observations dans les jours de la semaine sur 2017-18

La moyenne d'observations est de 1724 k  
 Q1 = 1608 k    Q2 = 1643 k    Q3 = 1781 k

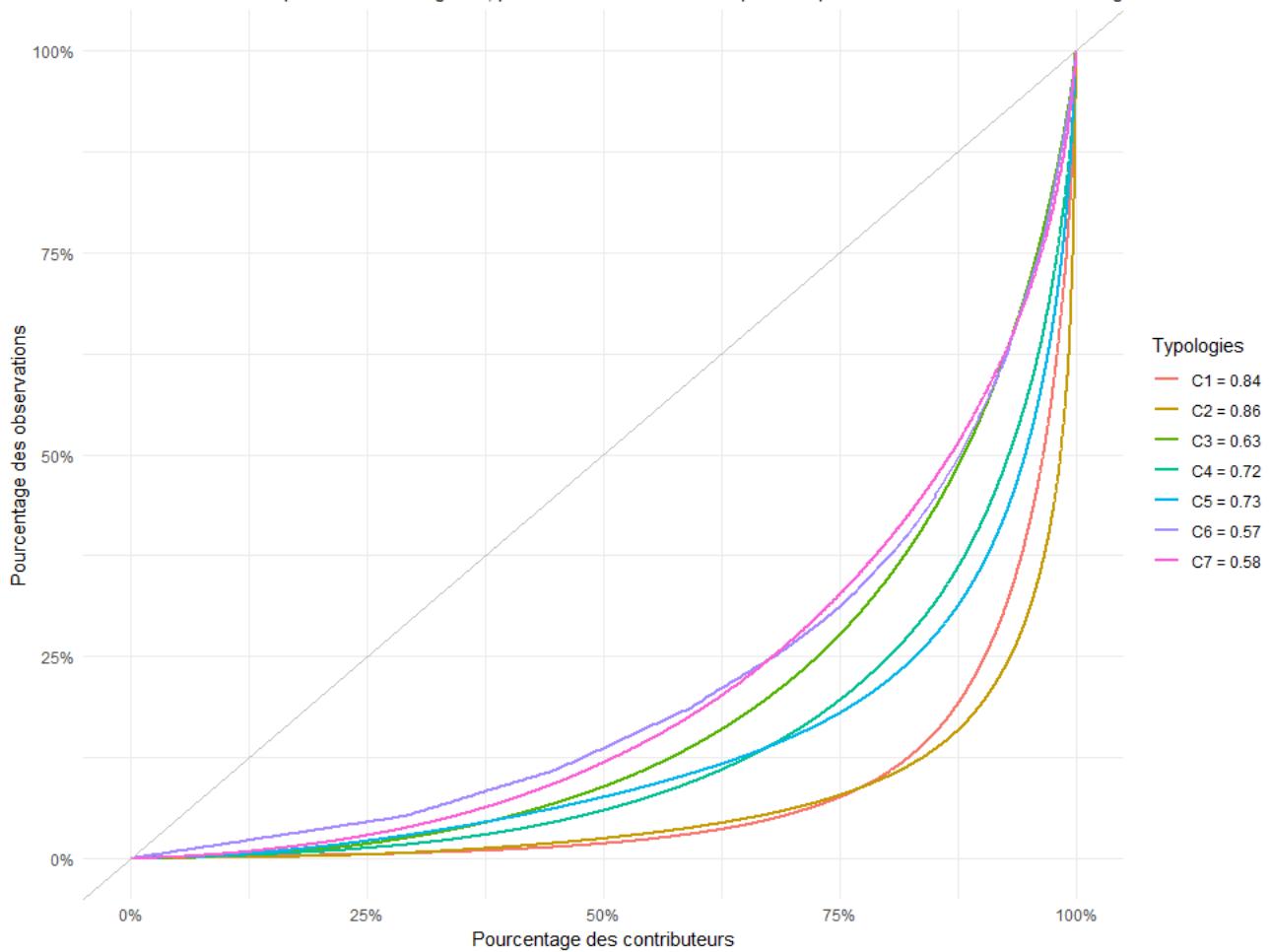


Source : faune-france.org

### 7.4.2 Visualisation de la répartition d'observation dans chaque typologies

#### Courbe de Lorenz avec les observations de Faune France 2017 et 2018

L'indice de Gini est indiquer sur dans la légende, plus cette indice tend vers 1 plus la répartitions entre contributeur est inégalitaire

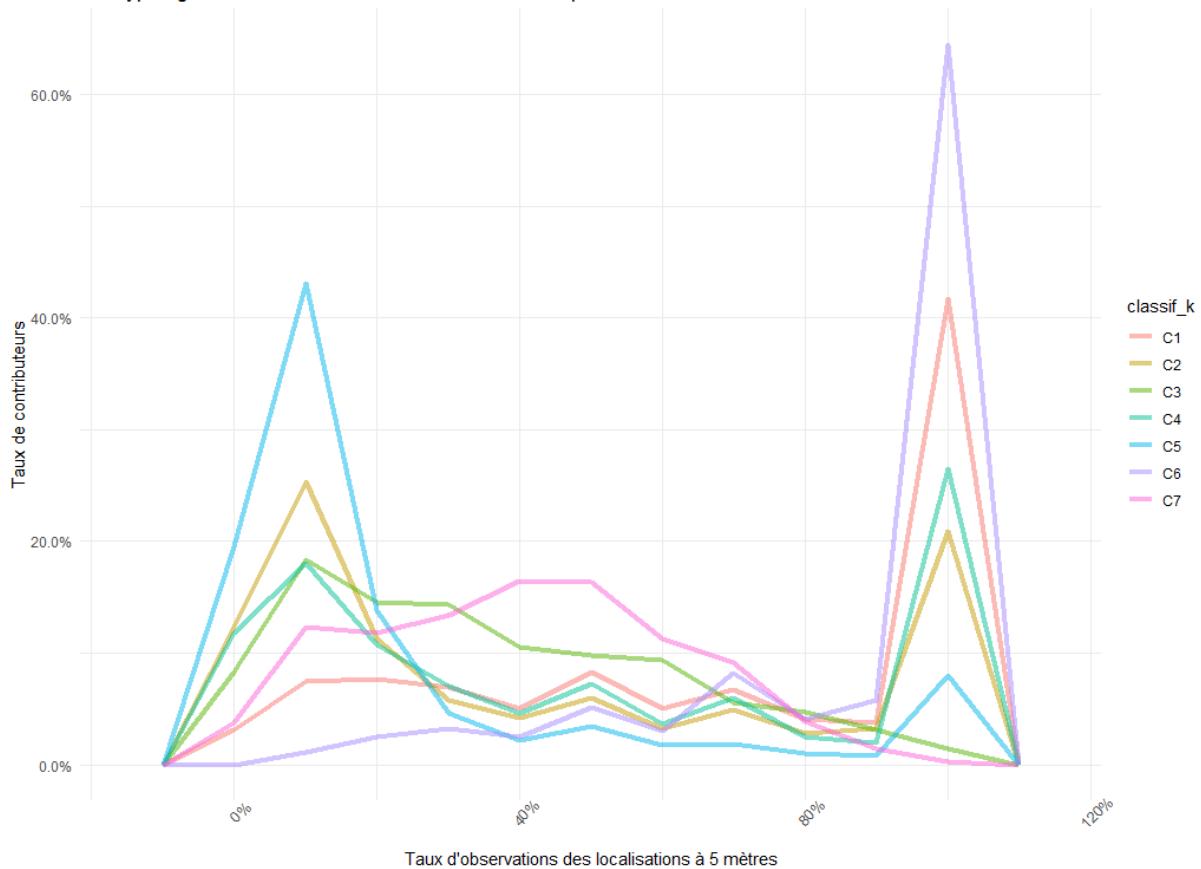


Source : faune-france.org

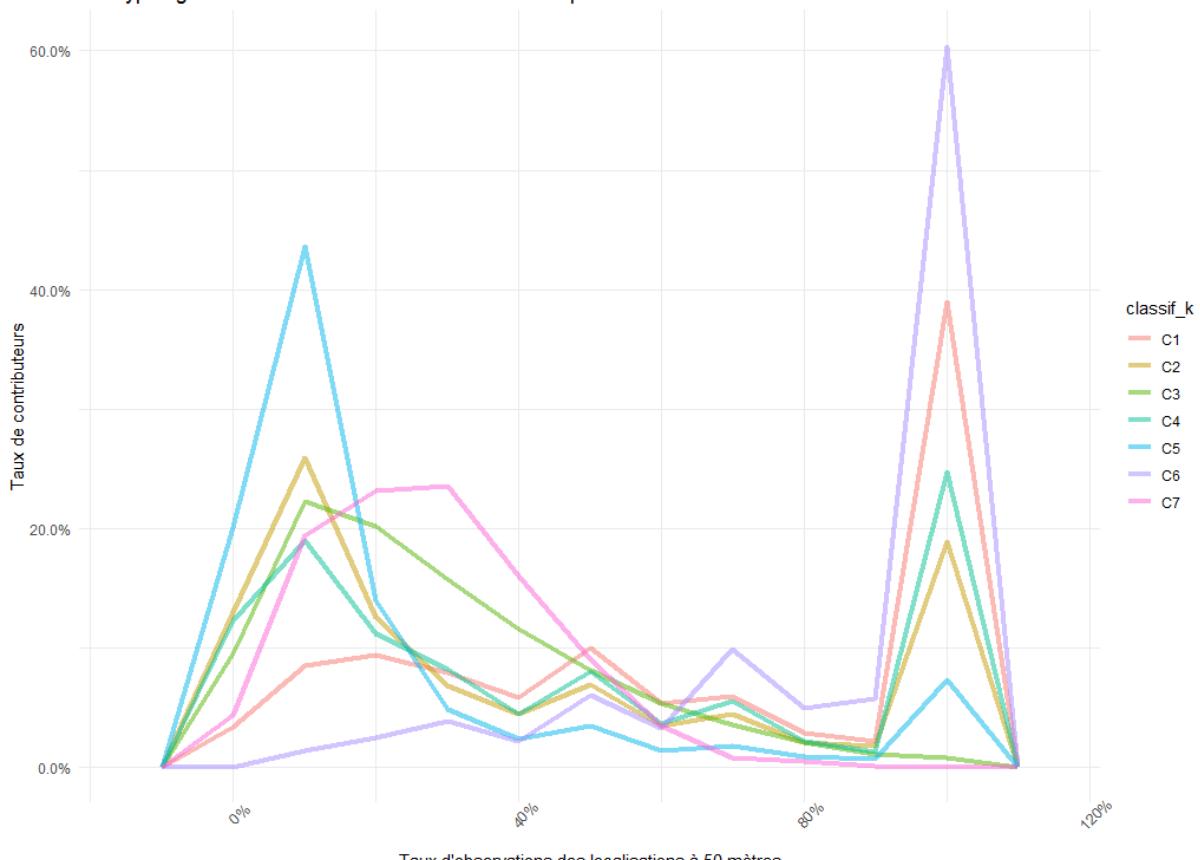
#### 7.4.3 Nombre de localisation à différentes échelles

#### 7.4.4 Taux de localisation à différentes échelles

**Les typologies des contributeurs ont-il une vision dispersé ou concentré ?**

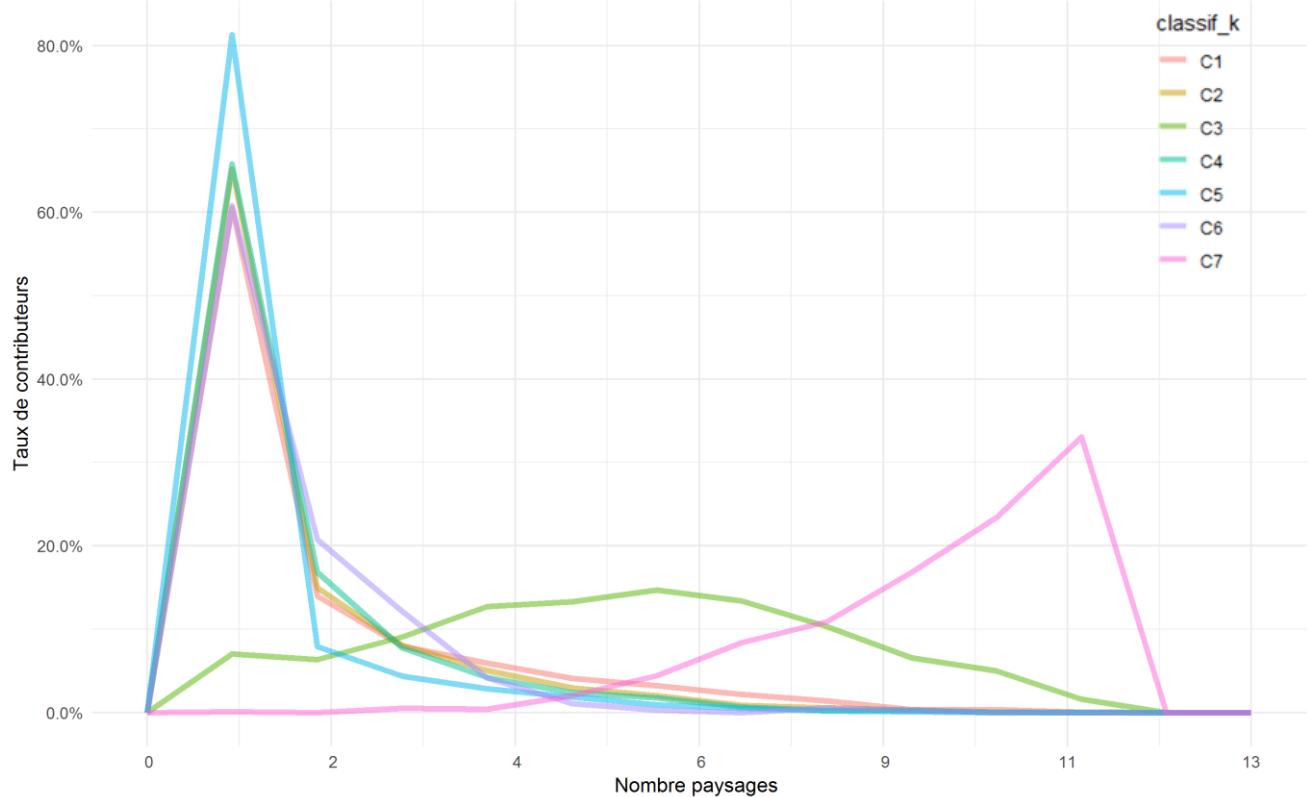


**Les typologies des contributeurs ont-il une vision dispersé ou concentré ?**

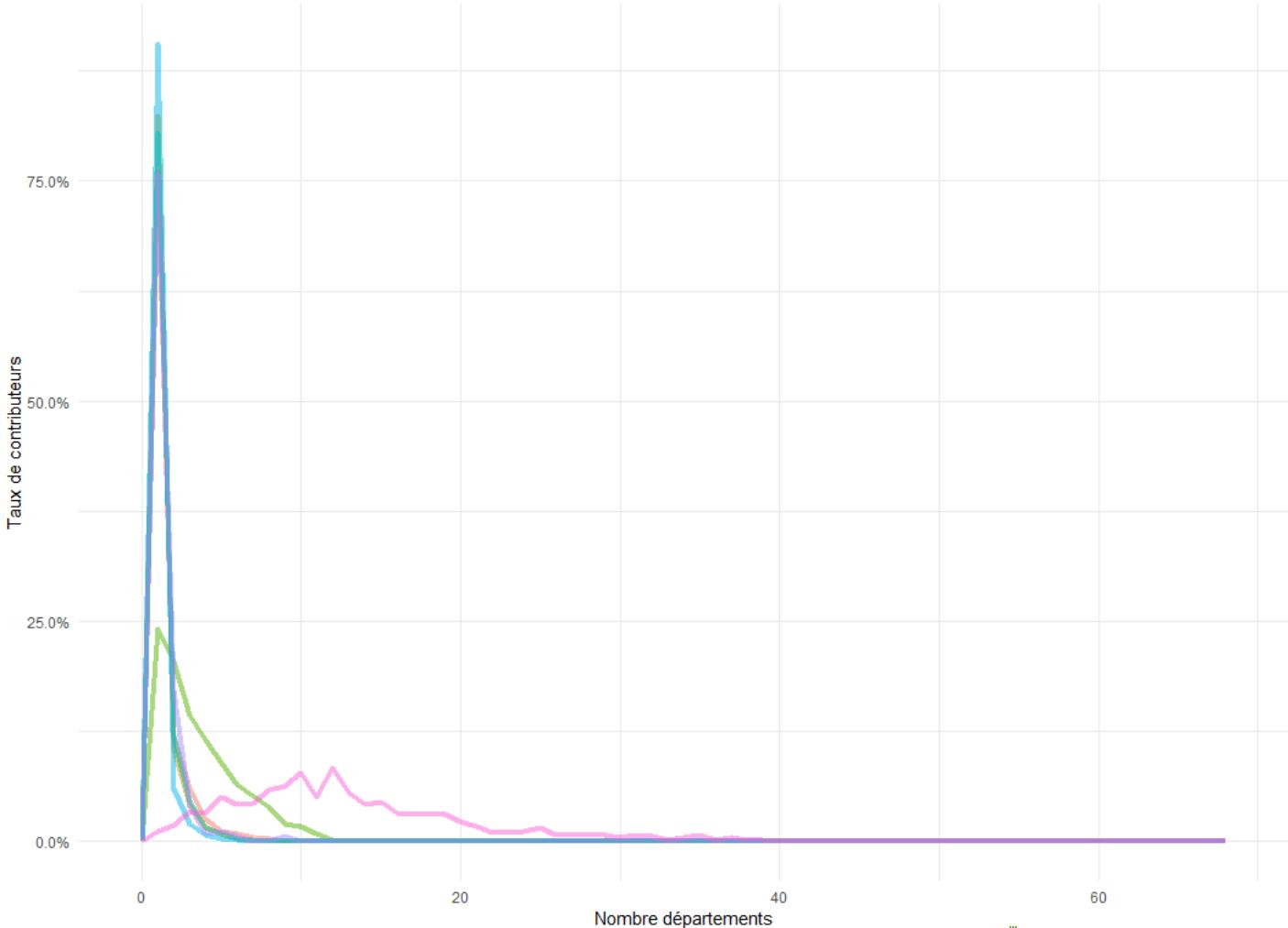


#### 7.4.5 Nombre de communes, départements, régions et paysages différents visités

Histogramme Nombre paysages des contributeurs

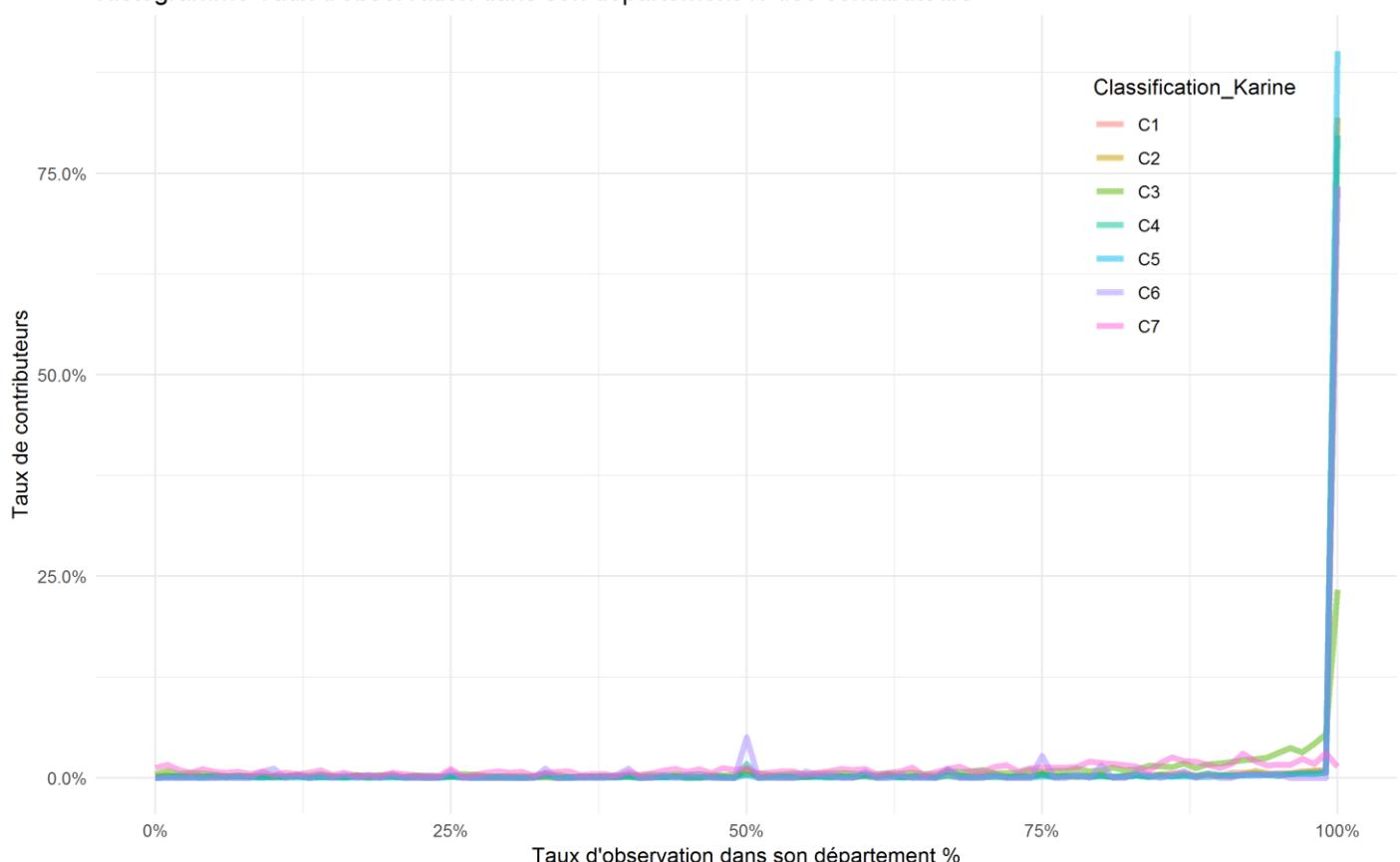


Nombre de département observé par un contributeur de Faune France

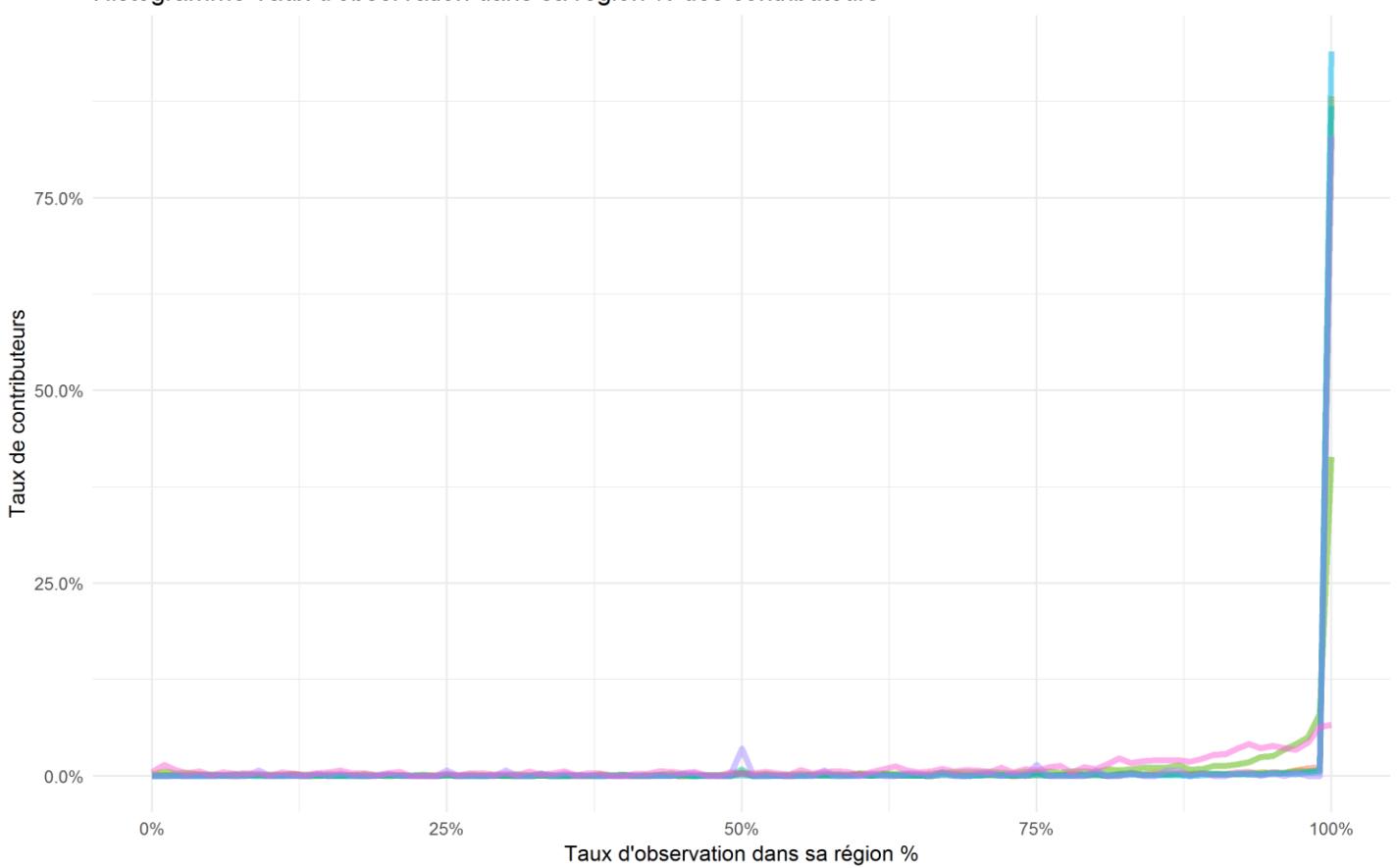


#### 7.4.6 Taux d'observation dans sa commune, son département, sa région d'habitation

Histogramme Taux d'observation dans son département % des contributeurs

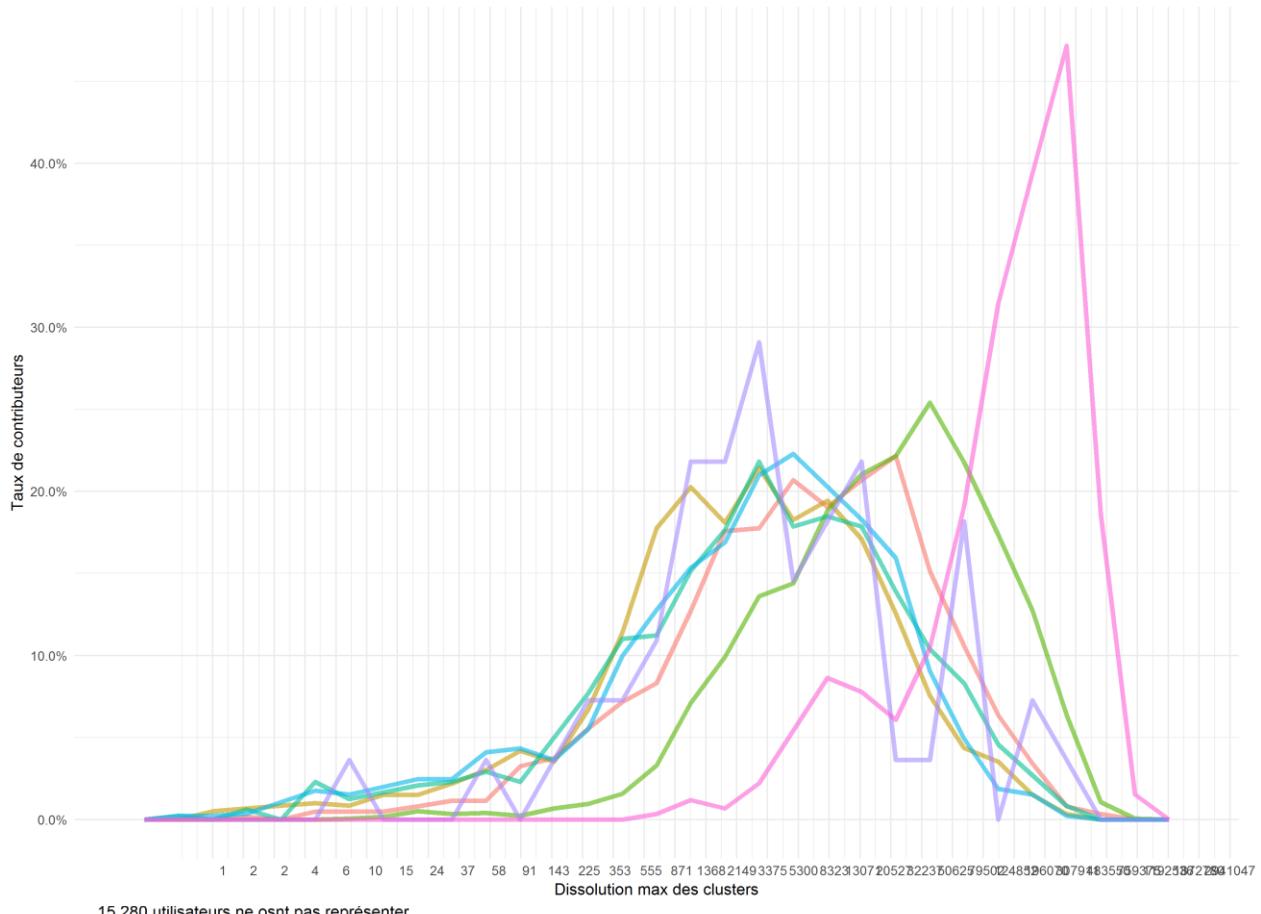


Histogramme Taux d'observation dans sa région % des contributeurs

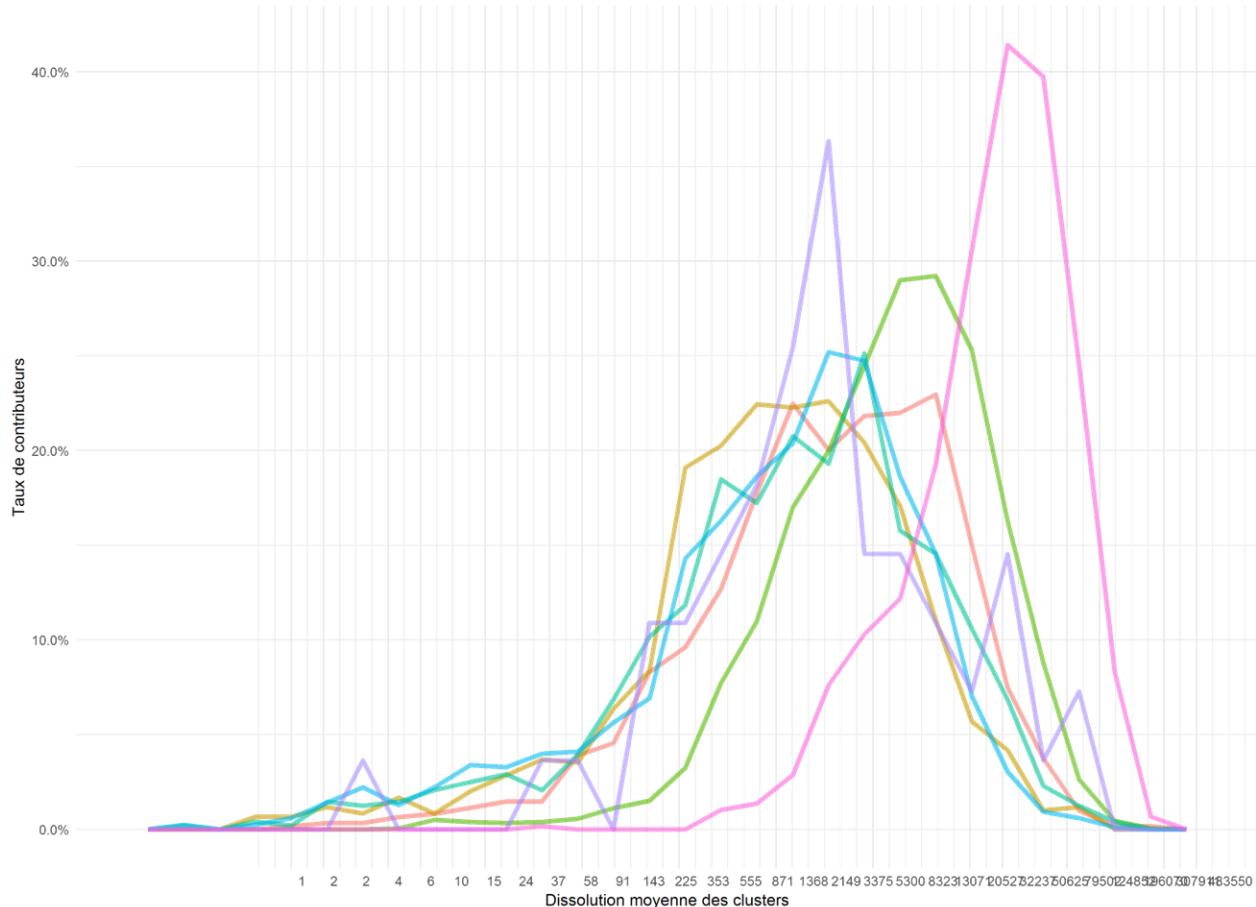


#### 7.4.7 Dissolution des clusters

Dissolution max des clusters par typologie de contributeur  
15 280 utilisateurs ne sont pas représentés



15 280 utilisateurs ne sont pas représentés

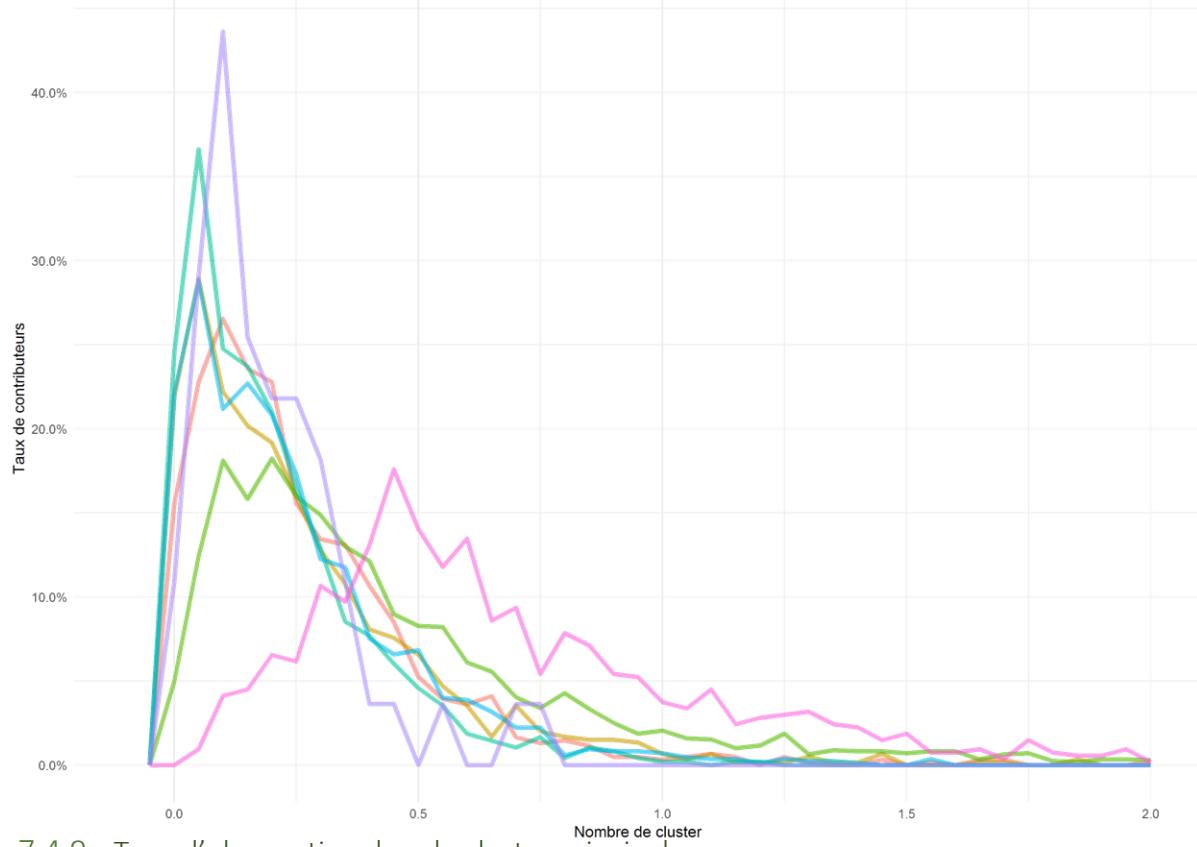


#### 7.4.8 Isolation des clusters

Nombre de cluster par typologie de contributeur

15 280 utilisateurs ne sont pas représentés

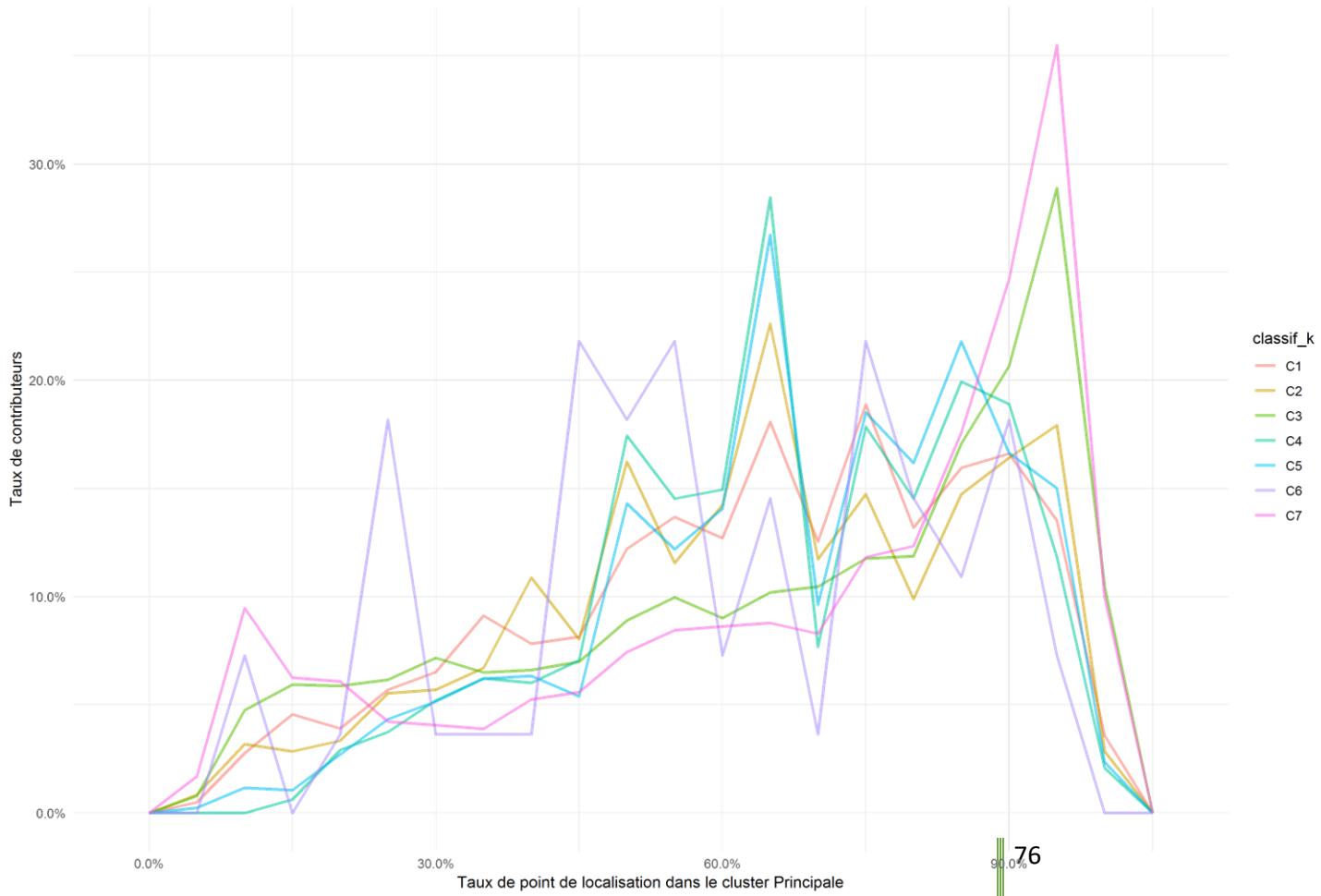
L'isolation est calculé par la fonction clara qui permet d'indiquer si les clusters ont tendance à être éloigné (proche de 0) ou très rapproché voire chevauché (plus de 1)



#### 7.4.9 Taux d'observation dans le cluster principal

Concentration de localisation des contributeurs par typologie

15 280 utilisateurs ne sont pas représentés



## 7.5 Références bibliographiques

### 7.5.1 Livres

*Alboukadel Kassambara - Practical Guide to Cluster Analysis in R. Unsupervised Machine Learning-STHDA (2017)*

*Alboukadel Kassambara - Multivariate Analysis II Practical Guide to Principal Component Methods in R*

*Baddeley, Adrian Rubak, Ege Turner, Rolf - Spatial point patterns methodology and applications with r - CHAPMAN & HALL CRC (2016)*

*Robin Lovelace, Jakub Nowosad, Jannes Muenchow - Geocomputation with R (2019)*

*Hadley Wickham - ggplot2 Elegant Graphics for Data Analysis - Springer-Verlag New York (2009)*

*Daniel Borcard, Francois Gillet, Pierre Legendre - Numerical Ecology with R - Springer-Verlag New York (2011)*

*Regina O. Obe and Leo S. Hsu, Foreword by Paul Ramsey, PostGIS in Action, Second Edition (2015)*

### 7.5.2 Site

[HTTPS://JUBA.GITHUB.IO/TIDYVERSE/08-GGPLOT2.HTML](https://juba.github.io/tidyverse/08-ggplot2.html)

[HTTPS://GITHUB.COM/RMCELREATH/STATRETHINKING\\_WINTER2019](https://github.com/rmcelreath/statthinking_winter2019)

[HTTPS://SOCVIZ.CO/GROUPFACETX.HTML#STATFUNCTIONS](https://socviz.co/groupfacetx.html#statfunctions)

[HTTPS://CRAN.R-PROJECT.ORG/WEB/PACKAGES/GGMAP/GGMAP.PDF](https://cran.r-project.org/web/packages/ggmap/ggmap.pdf)

[HTTP://PERSO.ENS-LYON.FR/LISE.VAUDOR/](http://perso.ens-lyon.fr/lise.vaudor/)

[HTTPS://BOOKDOWN.ORG/ROBINLOVELACE/GEOCOMPR/SPATIAL-CLASS.HTML](https://bookdown.org/robinlovelace/geocompr/spatial-class.html)

[HTTPS://FRANCE-DECOUVERTE.GEOCLIP.FR/#C=INDICATOR&I=STA.LOC&I2=TYPO\\_RUR.TYPO\\_CH3&T=A01&T2=A01&VIEW=MAP12](https://france-decouverte.geoclip.fr/#c=INDICATOR&i=STA.LOC&i2=TYPO_RUR.TYPO_CH3&t=A01&t2=A01&view=map12)

[HTTPS://WWW.FAUNE-FRANCE.ORG/INDEX.PHP](https://www.faune-france.org/index.php)

[HTTP://WWW.POSTGIS.FR/CHROME/SITE/DOCS/WORKSHOP-FOSS4G/DOC/PostGISINTRO.PDF](http://www.postgis.fr/chrome/site/docs/workshop-foss4g/doc/PostGISIntro.pdf)

## 7.6 Autres documents

### 7.6.1 Visualisation des jeux de données

Ref	ID Espèce Biolovisio	Nom espèce	Nom latin	Groupe taxonomique	Famille	Ordre systématique	SEARCH_EXPORT	Protection nationale
63673796	8990	Salamandre tachetée	<i>Salamandra salama</i>	Amphibiens	Salamandridae	30	-	Non
63673797	8950	Triton alpestre	<i>Ichthyosaura alpestris</i>	Amphibiens	Salamandridae	690	-	Non
63677146	8680	Rainette méridionale	<i>Hyla meridionalis</i>	Amphibiens	Hylidae	2010	-	Non
63681057	8990	Salamandre tachetée	<i>Salamandra salama</i>	Amphibiens	Salamandridae	30	-	Non
63682337	15295	Crapaud épineux	<i>Bufo spinosus</i>	Amphibiens	Bufonidae	1710	-	Non

SEARCH_EXPORT	SEARCH_EXPORT_S	SEARCH_EXPORT	SEARCH_EXPORT_S	SEARCH_EXPORT	SEARCH_EXPORT_S	SEARCH_EXPORT	SEARCH_EXPORT	SEARCH_EXPORT	Date
-	LC	LC	LC	Non	Oui	Non			01.01.2018
-	LC	LC	LC	Non	Oui	Non			01.01.2018
-	LC	LC	LC	Non	Oui	Non			01.01.2018
-	LC	LC	LC	Non	Oui	Non			01.01.2018
-	LC	LC	LC	Non	Oui	Non			01.01.2018

Mois	Année	Jour de l'année	Pentade	Décade	numéro de la liste	Horaire	ID liste	Heure début
1	2018	1	1	1	1	1	0	0
1	2018	1	1	1	1	1	0	0
1	2018	1	1	1	1	1	0	0
1	2018	1	1	1	1	1	0	0
1	2018	1	1	1	1	1	0	0

Heure de début	Minute de début	Heure fin	Heure de fin	Minute de fin	Liste complète	Commentaire de la liste	ID Lieu-dit	Lieu-dit
0	0	0	0	0	0		99471	Etang de la Tournière
0	0	0	0	0	0		99471	Etang de la Tournière
0	0	0	0	0	0		962888	Maisons du bois de l'A
0	0	0	0	0	0		550962	Sauronnet
0	0	0	0	0	0		1107342	Saint-Vincent-Rive-d'

Commune	Département	Code INSEE	Pays	X Lambert Ile [m]	Y Lambert Ile [m]	X Lambert93 [m]	Y Lambert93 [m]	Lat (WGS84)
Lent	1	1211	France	823729	2128070	872321	6559762	46.1157485
Lent	1	1211	France	823729	2128070	872321	6559762	46.1157485
Saint-Laurent	17	17353	France	336890	2115128	385821	6550908	45.9851891885
Antignac	15	15008	France	617139	2038727	665169	6472255	45.348809074987

(Tableau suite)

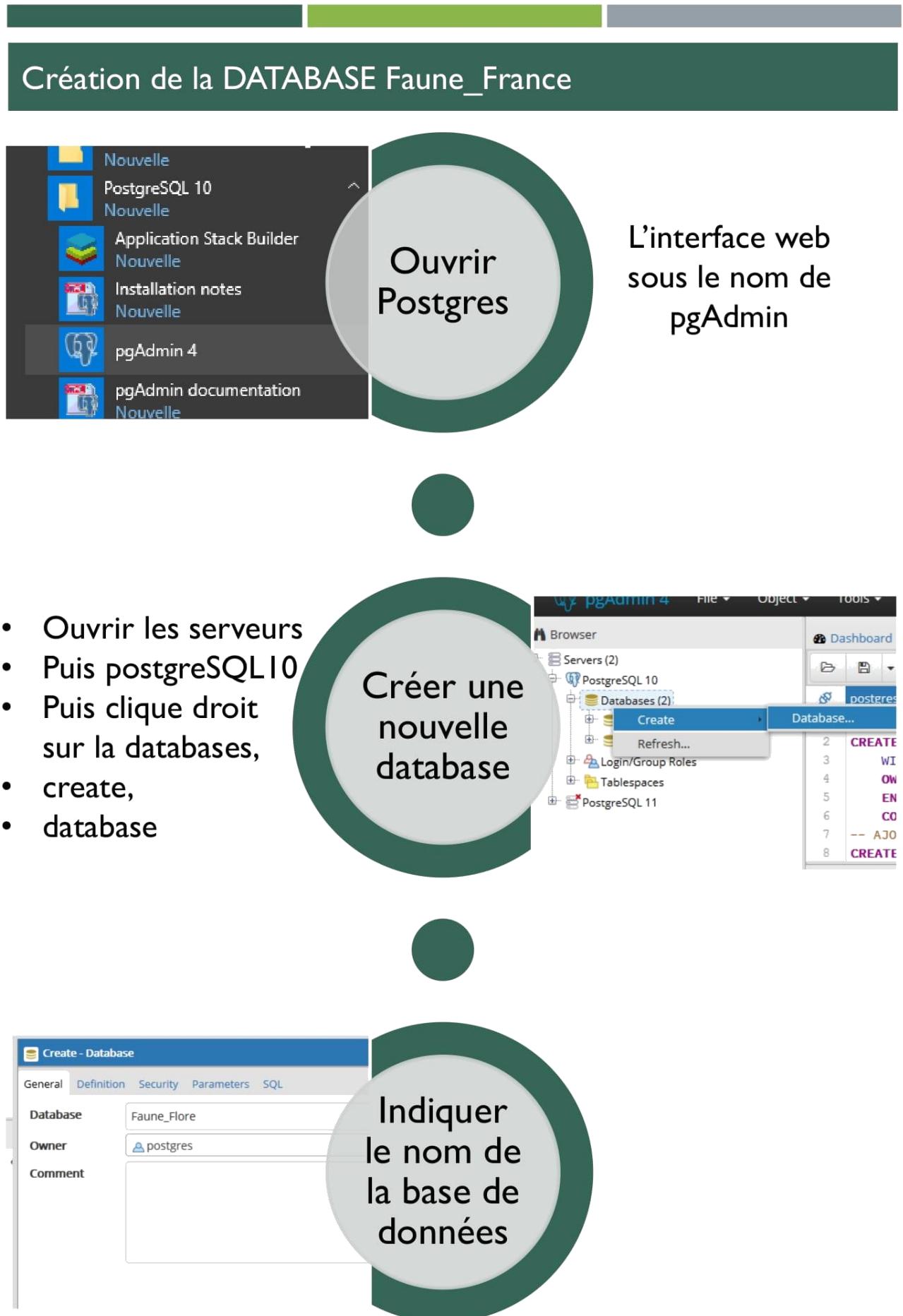
Lon (WGS84)	latitude (DMS)	longitude (DMS)	fuseau UTM Nord	UTM X [m]	UTM Y [m]	Type de localisat	Maille	Estimation
5.23157446	46°06'56.69"N	5°13'53.67"E	31	672433	5109328	Lieu-dit	E087N655	
5.23157446	46°06'56.69"N	5°13'53.67"E	31	672433	5109328	Lieu-dit	E087N655	
-1.060223633	45°59'6.68"N	1°03'36.81"W	30	650240	5094231	Lieu-dit	E038N655	
2.555212659	45°20'55.71"N	2°33'18.77"E	31	465157	5021796	Localisation préc	E066N647	
1.301088826	44°27'50.06"N	1°18'3.92"E	31	364856	4924803	Localisation préc	E056N637	

Altitude	Nombre	Détails	Comportement	Donnée de seconde	Protégée	Vérification	Remarque	Remarque privée
259	1			Non	Non			
259	1	1x femelle		Non	Non			
7	1	1x (vu)		Non	Non		Dans les pierres autour du bassin	
557	2			Non	Non			
117	1	1x adulte (vu)		Non	Non			

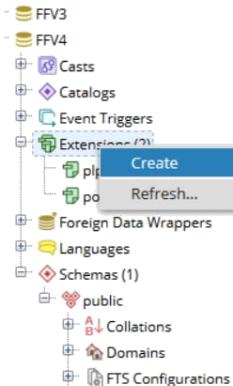
Code étude	Contient des media	ID de l'observati	ID universel observ	Anonyme	Transmetteur	Protocole	Contient des dé	Date d'insertion
	Oui	28859	45950	Non	Non		Non	01.01.2018 03:14
	Oui	28859	45950	Non	Non		Non	01.01.2018 03:16
	Non	36633	23821	Non	Non		Non	01.01.2018 11:58
	Non	8381	18418	Non	Non		Non	01.01.2018 14:37
	Oui	64022	82521	Non	Non		Non	01.01.2018 15:53

Date de dernière modification								
01.01.2018 09:23								
01.01.2018 09:22								
01.01.2018 11:58								
01.01.2018 14:37								
01.01.2018 15:53								

## 7.6.2 Guides créations d'une base de données dans PostgreSQL



## Création de la DATABASE Faune\_France



Ajouter  
l'extension  
postgis

- Clique droit sur extension,
- create,
- extension

- postgis

Indiquer le  
nom de la  
base de  
données

Create - Extension

General Definition SQL

Name: post

Comment:

Select from the list

- pointcloud\_postgis
- postgis
- postgis\_sfsgal
- postgis\_tiger\_geocoder
- postgis\_topology