

VALORISER LES DONNÉES NATURALISTES OPPORTUNISTES EN TENANT COMPTE DES BIAIS

JUILLET 2021

Gestion et dynamisation des bases de données naturalistes du Grand Est



**8 TYPES DE
BIAIS
RÉCURRENTS**

**DES SOLUTIONS
MISES EN ŒUVRE
AU SEIN DES
ASSOCIATIONS
NATURALISTES**

**LA MISE EN
PERSPECTIVE DES
DONNÉES AVEC
L'EXPERTISE
NATURALISTE
RESTE
INDISPENSABLE**

Depuis quelques années, les sciences participatives sont en pleine expansion et permettent l'émergence d'une source de données importantes : les données naturalistes opportunistes. Les observateurs, amateurs ou experts, professionnels mais surtout bénévoles, sont invités à partager directement leurs observations d'espèces dans des bases de données dédiées.

Les techniques ont largement progressé : transmission de tables informatisées, saisie sur des portails en ligne, puis enregistrement via des applications smartphone. Toutes ces évolutions ont permis à plus de naturalistes amateurs d'être contributeurs, avec plus d'observations chacun, dans des formats plus précis (localisation souvent exacte, indices de reproduction...), tout en facilitant les processus de validation.

Dans la région Grand Est, une majorité des données naturalistes opportunistes, en particulier faunistiques, sont collectées et traitées par le réseau associatif ODONAT Grand Est. Une fois centralisées et validées, elles permettent d'accroître significativement le niveau de connaissance de différents pans de la biodiversité régionale.

La force des données naturalistes opportunistes réside dans leur nombre important et leur répartition spatiale. Mais contrairement aux données basées sur des protocoles scientifiques de collecte bien établis en amont, leur caractère opportuniste peut engendrer des biais dans différents types d'analyses. Afin de relativiser certains résultats, il est indispensable d'appréhender ces distorsions. C'est un enjeu du réseau ODONAT Grand Est de développer le maximum de solutions pour en réduire les effets, à appliquer en amont dès la collecte, ou en aval par des modèles correcteurs.

PRINCIPAUX BIAIS DES DONNEES NATURALISTES OPPORTUNISTES

Les biais des données naturalistes opportunistes possèdent différentes origines. Certains biais sont directement liés à l'observateur et d'autres à l'observation.

CHOIX DE L'OBSERVATEUR

Biais spatial - "Où je vais?"

Contrairement à de nombreux protocoles de recherche, les sciences participatives n'imposent aucune localisation des lieux de collecte. Cette liberté introduit un biais spatial (Johnston et al. 2020) : le choix de la part de l'observateur est rarement guidé par le hasard, mais est influencé

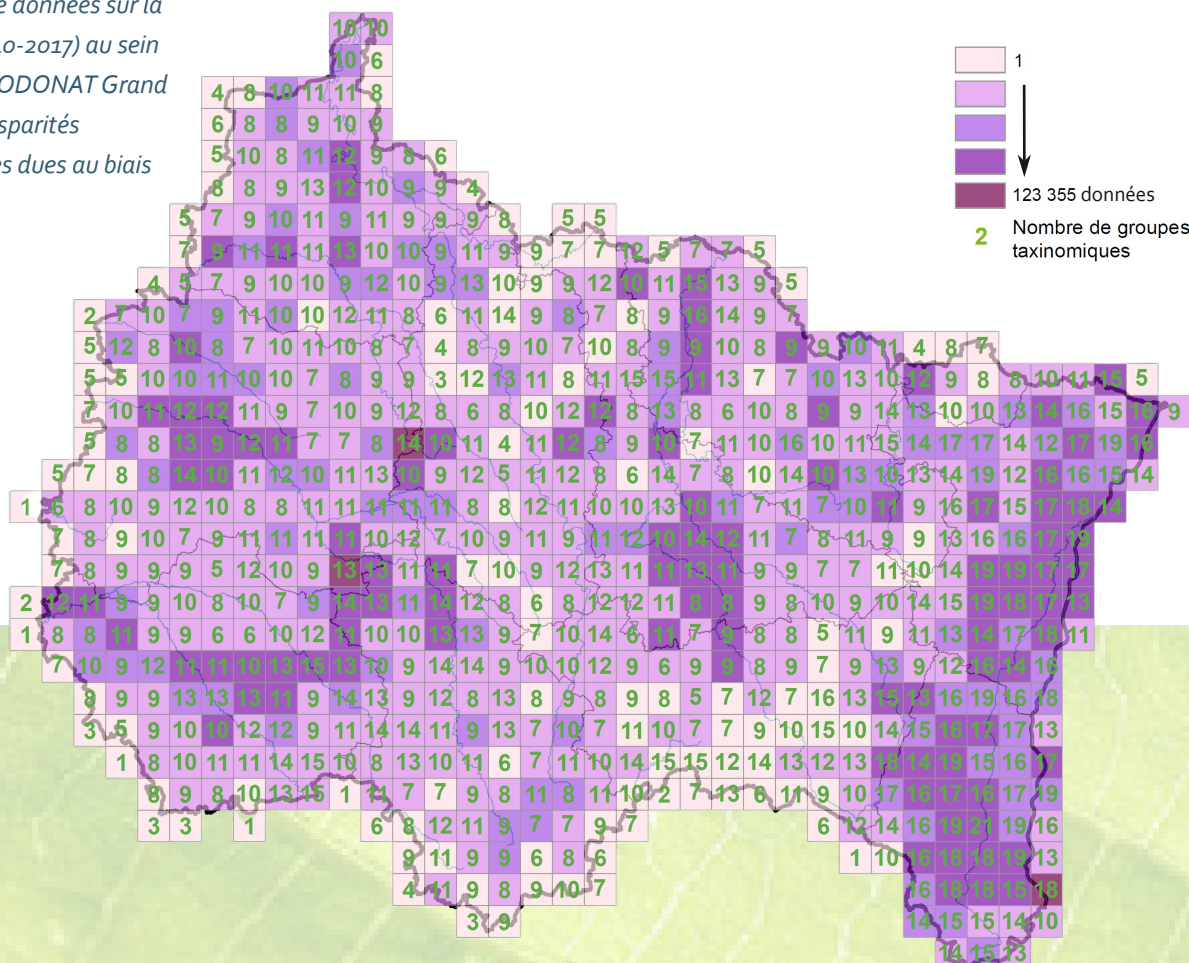
par la proximité et l'accessibilité du site, ou encore la richesse spécifique (parfois supposée) des habitats présents (Mair & Ruete, 2016). Un observateur ira plus rarement dans une zone jugée peu favorable, qui conduirait à un faible nombre de relevés. Malgré cela, dans le cadre d'une vision territoriale assez globale, le nombre important d'observateurs permet généralement de produire une information qualitative étoffée (Callaghan, Rowley, et al. 2019).

Biais temporel - "Quand j'y vais?"

La collecte étant participative, les observateurs choisissent quand ils

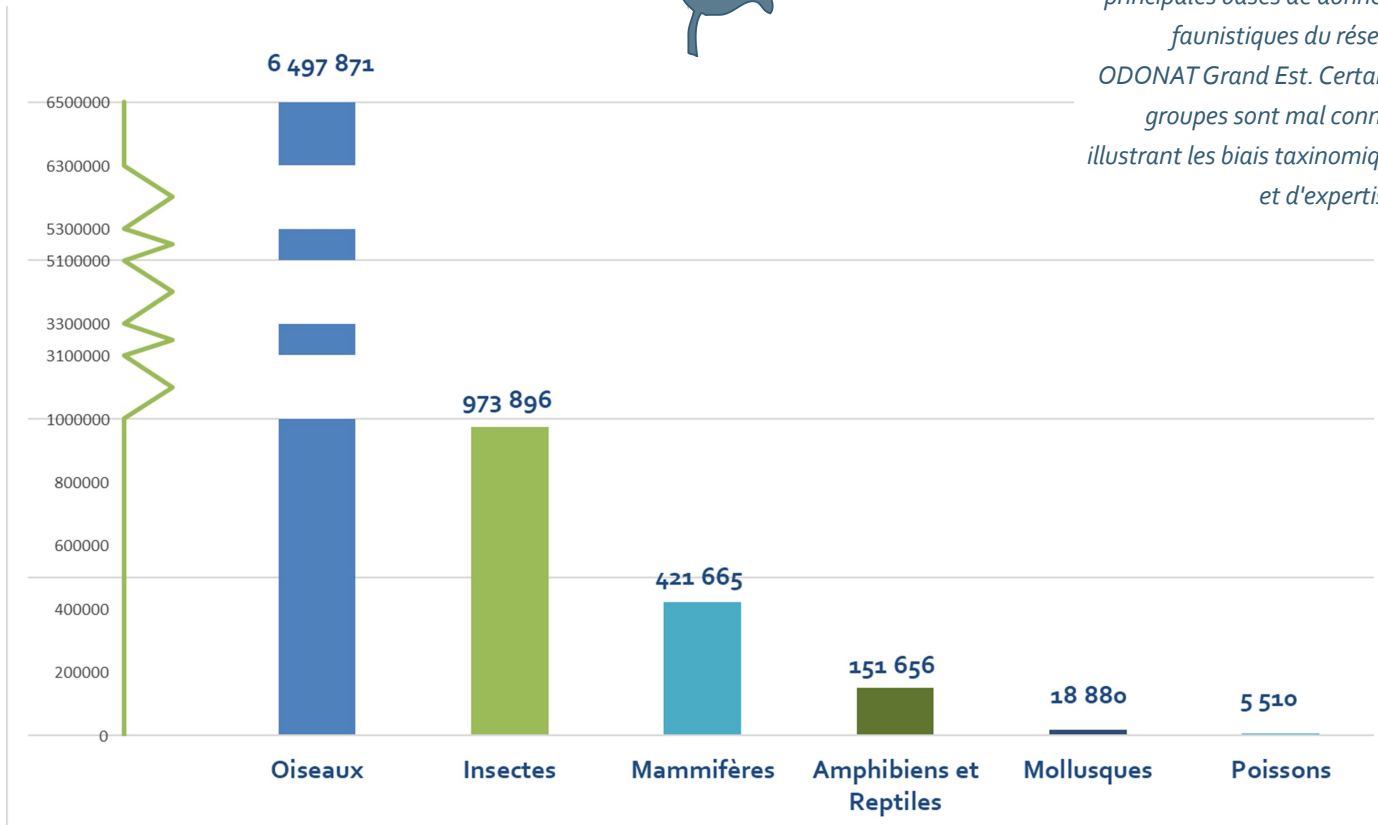
veulent aller sur le terrain. Cela entraîne un biais au niveau de la résolution temporelle des données (Steen et al. 2019). Certaines périodes vont ainsi permettre de recueillir plus de données sans être corrélées à une présence plus importante d'espèces. Par exemple, "le biais du weekend" conduit à plus de signalements d'arrivée d'oiseaux migrants le week-end plutôt que pendant la semaine (Tiedeman et al. 2019). La météo va également avoir une influence sur les jours de sortie des observateurs.

Nombre de données sur la faune (2010-2017) au sein du réseau ODNAT Grand Est, des disparités territoriales dues au biais spatial





Nombre de données par «domaine taxinomique» enregistrées en 2020 dans les principales bases de données faunistiques du réseau ODONAT Grand Est. Certains groupes sont mal connus illustrant les biais taxinomique et d'expertise.



Biais effort de prospection - « Combien de temps j'y passe ? »

Seule la volonté de l'observateur guide un temps de recherche plus ou moins long sur chaque site prospecté. Il en découle une hétérogénéité de l'effort d'échantillonnage, entre observateurs, voire pour un même observateur sur différents sites (Isaac et al., 2014).

COMPETENCES DE L'OBSERVATEUR

Biais expertise - « Ce que je connais »

Les outils de la science participative attirent des personnes aux profils variés,

dépassant le seul cercle d'experts habituellement mobilisable (Robinson et al. 2017). Le biais d'expertise est alors une combinaison entre connaissances (capacité d'identifier peu ou beaucoup d'espèces) et comportements (expérience permettant de rechercher et de détecter les espèces). Les compétences spécifiques réduisent systématiquement le nombre d'observateurs en capacité de participer, et la détection difficile restreint le nombre de données collectées (Du Tien Hat et al., 2018). Les oiseaux (et les vertébrés en général) sont ainsi très représentés dans les bases de

données participatives, alors que certaines familles d'insectes ou de mollusques ne sont presque jamais enregistrées.

Biais taxinomique - « Ce que je veux voir »

Des disparités majeures existent entre les connaissances compilées entre les domaines d'étude taxinomique (Troudet et al. 2017). Celles-ci sont à la fois liées aux préférences manifestées par les collecteurs pour certains domaines d'étude ou à la capacité de détection.

PRECISION ET COMPLETUDE DES DONNEES

Biais déclaration- « Ce que je dis »

Les observateurs restent libres dans la complétude de signalement des espèces identifiées. Toutes ne sont pas enregistrées (van Strien et al., 2013). Les espèces communes sont ainsi souvent sous-représentées dans les bases alors que les espèces rares sont plus systématiquement déclarées, parfois même avec une précision allant jusqu'à la localisation de chaque individu (ou station) (Steen et al. 2019). Ce facteur peut entraîner une surestimation ou une sous-estimation de l'abondance des espèces. D'autre part, la saisie des autres champs descriptifs importants n'est pas

systématique, ce qui réduit la précision des informations transmises : pas de dénombrement de l'effectif, code de reproduction non spécifié, géolocalisation approximative, etc.

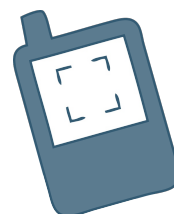
Biais d'absence d'une espèce - « Ce que je n'ai pas vu »

La très grande majorité des données enregistrées correspondent à des données de présence. Les données précisant l'absence de contact alors qu'un protocole de recherche dédié a été suivi restent très minoritaires (Lobo et al. 2018).

Biais "erreur"

Lors de la saisie d'une donnée, différentes sources d'erreurs

peuvent affecter les valeurs des champs descriptifs. La principale erreur reste liée à la qualité de l'identification du taxon. (Du Tien Hat et al., 2018).



DES SOLUTIONS POUR MIEUX MOBILISER LES DONNEES

Un enjeu essentiel du réseau associatif ODONAT Grand Est est de pouvoir mobiliser les données opportunistes au service de la préservation de la biodiversité. Cela implique de mettre en place différentes démarches pour pallier aux biais décrits précédemment.

DES SOLUTIONS DEJA EN PLACE AU SEIN DU RESEAU

Améliorer les outils des bases de données en ligne

La science participative reposant sur l'apport des observateurs, il faut mettre en place des outils de collecte qui restent faciles d'utilisation pour le plus grand nombre, tout en motivant à la précision et la complétude des informations associées à chaque

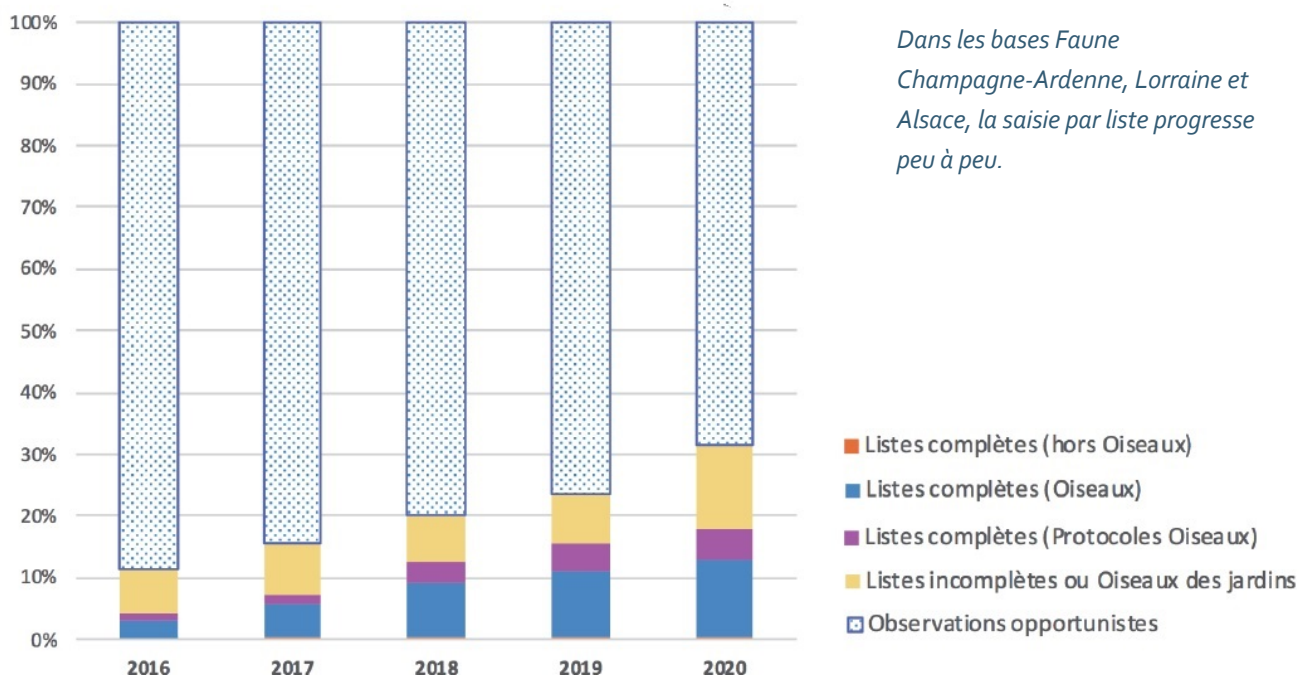
donnée. Les interfaces web et smartphones permettant la saisie des informations sont en constante évolution pour inciter au remplissage d'un maximum d'informations.

Former des observateurs

En s'appuyant sur ces constats, les associations naturalistes organisent des formations, créent des clés d'identifications (...) pour étoffer le nombre d'observateurs en capacité d'étudier un grand nombre de domaines taxinomiques. Des événements comme les 24h de la biodiversité permettent aux naturalistes de développer leurs compétences en permettant l'échange entre observateurs.

Favoriser la saisie des données par liste

Systématiser la saisie des données sous forme de listes complètes (l'observateur s'engage à enregistrer l'ensemble des espèces contactées sur un site dans une période de temps déterminée) est un enjeu essentiel. Il permet d'attribuer une valeur d'absence aux espèces non relevées qui peut être intégrée dans des modèles d'analyse statistiques. Ainsi, on limite les biais expertise, déclaration, absence et effort de prospection.



Valider les données

Les données renseignées subissent une étape de validation par des naturalistes confirmés. Ils apportent ainsi la « principale garantie de fiabilité des informations » (Bogéa Soares et al. 2013). Ce processus de validation a lieu en trois grandes étapes. Des filtres automatisés vérifient la pertinence de la donnée aux regards des données habituelles connues pour l'espèce (répartition, phénologie). Les filtres permettent également d'avertir les observateurs d'erreurs de saisie. La deuxième étape est la vérification systématique par

des naturalistes confirmés. Les observateurs sont invités à participer à la validation en répondant le plus finement possible aux questions de ces experts afin de faciliter la validation ou la correction de leur donnée. La dernière étape est l'identification d'informations aberrantes passées inaperçues lors d'analyses globales des jeux de données a posteriori (Faune France 2020). Les biais d'expertise et d'erreur sont, par cette méthode, en grande partie résolus, la plupart des erreurs étant corrigées à la source. Certains groupes taxinomiques ne

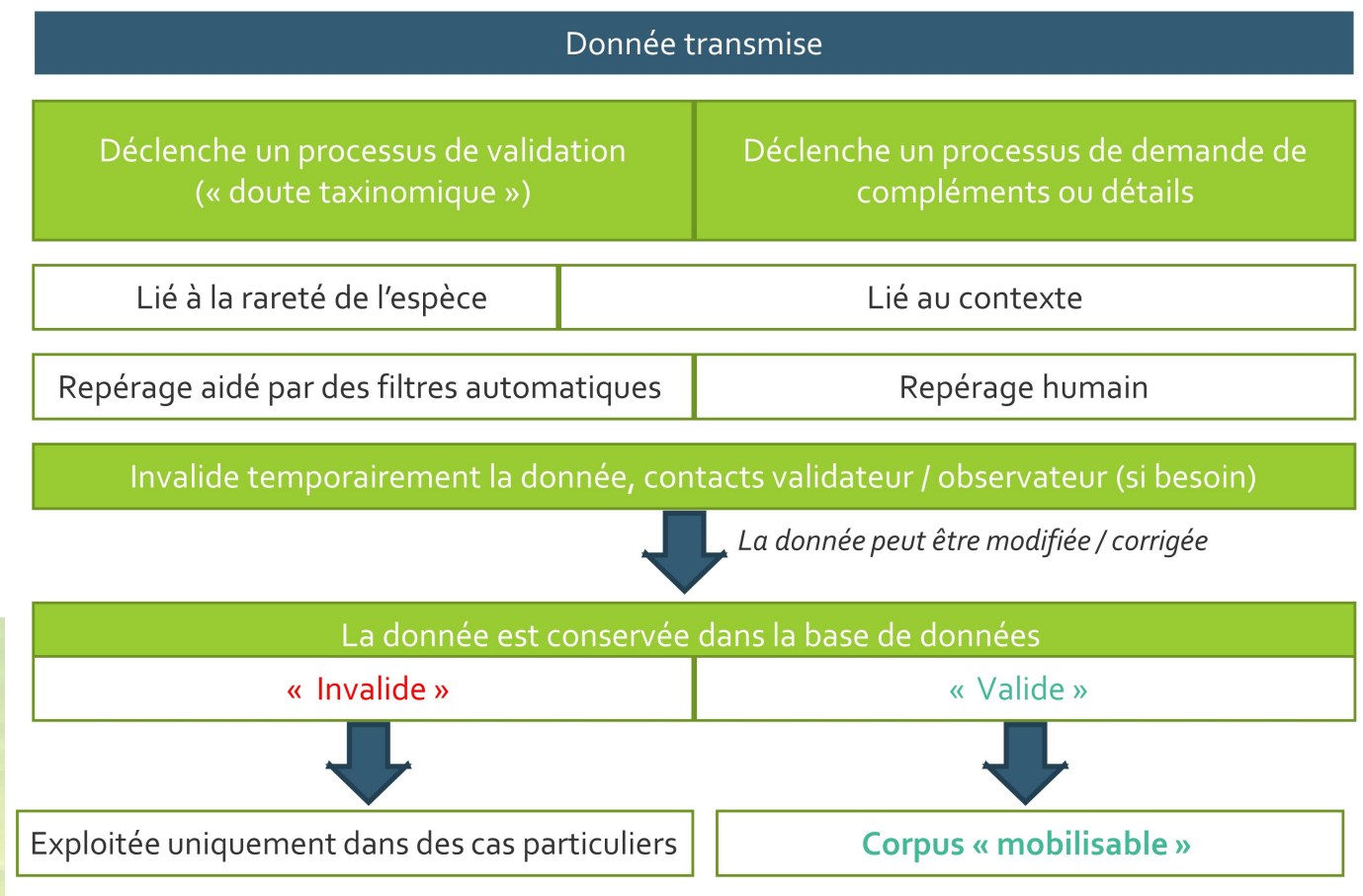
sont pas associés à un circuit de validation, mais les données sont exclusivement faites par un noyau d'observateurs confirmés.

DES SOLUTIONS A DEVELOPPER

Vérifier ponctuellement la fiabilité des données

Ponctuellement, il est intéressant de comparer les données opportunistes avec des données protocolées pour mesurer le niveau de connaissance apportés par les deux types de jeux de données. Il existe, par exemple, des méthodes visant à améliorer

Circuit de validation d'une donnée déclenchant un processus de vérification





la collecte des données en comparant l'acquisition par des observateurs bénévoles avec l'analyse faites par des observateurs professionnels (McDonough MacKenzie et al. 2017). Ces méthodes permettent de corriger les biais d'effort et spatial. Une autre possibilité est de regarder le taux de détection sur les deux types de jeux de données. Par exemple, la comparaison entre un jeu de données opportunistes, les données de l'INPN et un jeu de données protocolées, le dispositif de surveillance des papillons (STERF) a montré que les valeurs brutes peuvent fonctionner aussi bien que les mesures corrigées (correction des biais). Il reste désormais à déterminer si les fréquences brutes peuvent être utilisées pour d'autres groupes avec moins de données ou présentant des variations de phénologie (Cima et al. 2020).

Mettre au point des modèles correcteurs

De nombreux articles scientifiques ont été écrits pour identifier des méthodes permettant de passer outre les biais résiduels. En général, il faut sélectionner un (ou deux maximum) biais sur lequel un

modèle correctif est développé. Il est possible de traiter les biais par une méthode de type statistique. Ces modèles sont définis par des observations indépendantes pour lesquelles le modèle va chercher à décrire le mode de collecte des données (Cadre 2016). Les modèles linéaires généralisés (GLM) en sont un exemple. D'un autre côté, les modèles de distribution d'espèces (SDM) sont des modèles simulant la distribution des espèces au niveau spatial et temporel (Razafimahefa 2017). Les SDMs peuvent être « corrélatifs ou mécanistes ». Les modèles corrélatifs relient les données naturalistes aux données environnementales alors que les SDMs mécanistes sont basés sur « des traits fonctionnels et des contraintes physiologiques » (Kearney et al. 2010). L'utilisation des SDMs peut permettre de compléter et normaliser simultanément les biais géographiques et le taux de détection (Kelling et al. 2015). Il est néanmoins nécessaire de rester prudent avec ces modèles correctifs car ils peuvent être plus ou moins fiables (Steen et al. 2019).

Un Modèle GLM pour tester les biais spatiaux et temporels (Callaghan, Poore, et al. 2019)

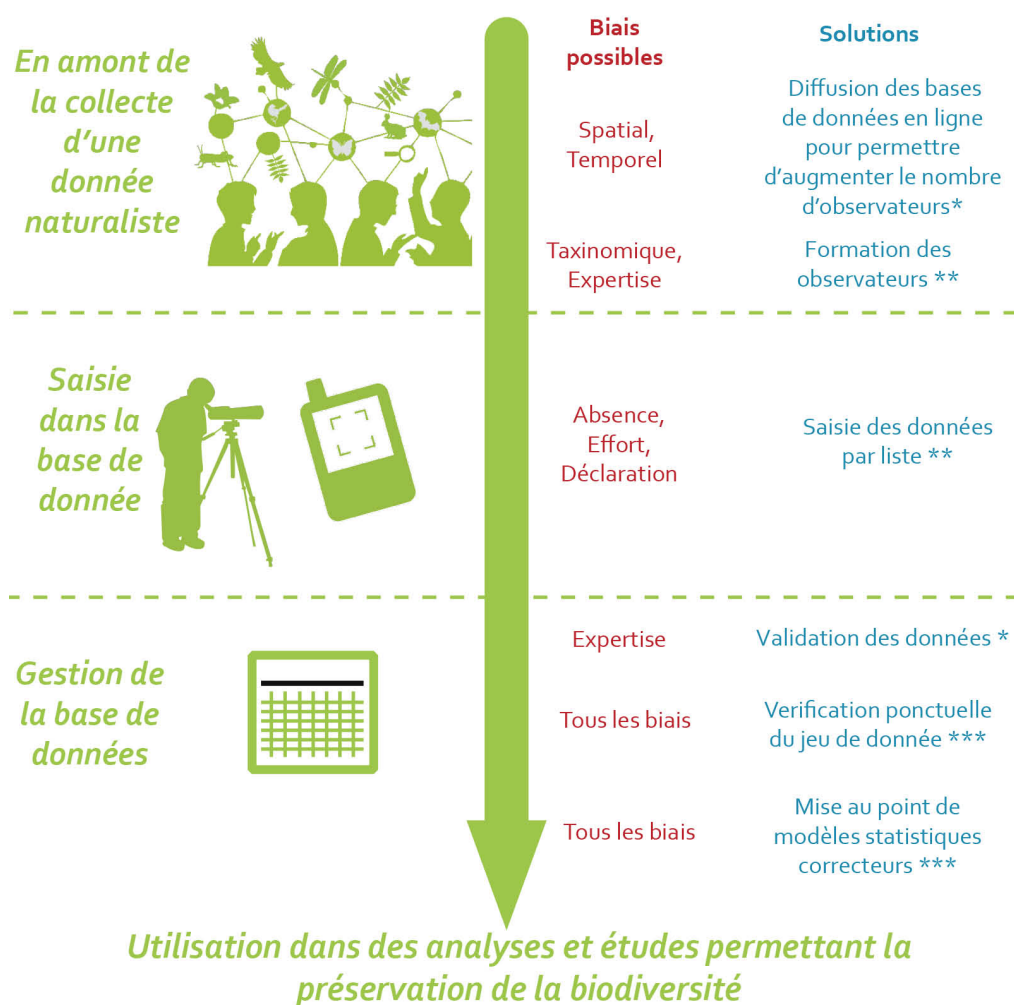
La question initiale était de savoir si l'échantillonnage spatial et temporel des données opportunistes avait une influence sur les tendances de biodiversité. Ces chercheurs ont donc développé un modèle linéaire généralisé basé sur la présence/absence. Dans ce modèle, l'intérêt a été porté à la fois au jour de collecte mais aussi à la taille de la maille sur laquelle faire l'analyse.

Un effet de levier a été produit pour chaque maille lié à des paramètres temporels et spatiaux (la maille avait-elle déjà été visitée, distance à la maille la plus proche échantillonnée...). La conclusion de cet article est qu'un échantillonnage de toutes les mailles n'est pas forcément utile, les citoyens identifient et échantillonnent de manière appropriée les points chauds qui méritent le plus d'attention. Mais, des point chauds régionaux auraient pu échapper aux citoyens et sans échantillonnage ces points chauds ne seront jamais découverts.

UNE AMELIORATION EN CONTINU

Le réseau ODONAT Grand Est met en place des dispositifs permettant de limiter les biais. Néanmoins dans chaque étude développée par le réseau, les jeux de données sont vérifiés pour

limiter les biais au maximum et les experts-naturalistes mettent en perspectives leurs savoirs pour pouvoir identifier les enjeux au plus juste.



* : Bon fonctionnement actuel
** : nécessité de continuer à être développé
*** : A mettre en place



POUR ALLER PLUS LOIN

- BOGÉA SOARES et al., 2013. Réflexions sur la validation des données naturalistes : le cas des erreurs d'occurrence dans la distribution des espèces. Rapport SPN-2014-38. S.I.
- CADRE, Benoit, 2016. Chapitre 1 : Modélisation statistique [en ligne]. S.I. ENS Rennes.
- CALLAGHAN, POORE et al., 2019. Optimizing future biodiversity sampling by citizen scientists. In : Proceedings of the Royal Society B: Biological Sciences. 9 octobre 2019. Vol. 286, n° 1912, pp. 20191487.
- CALLAGHAN, ROWLEY et al., 2019. Improving big citizen science data: Moving beyond haphazard sampling. In : PLOS Biology. 27 juin 2019. Vol. 17, n° 6, pp. e3000357.
- CIMA et al., 2020. A test of six simple indices to display the phenology of butterflies using a large multi-source database. In : Ecological Indicators. 1 mars 2020. Vol. 110, pp. 105885.
- DU TIEN HAT et al., 2018. Utilisation des bases de données opportunistes pour mettre à jour la répartition des oiseaux nicheurs [en ligne]. rapport PatriNat 2018 - 119. S.I.
- FAUNE FRANCE, 2020. Processus de validation - www.faune-france.org. In : Faune France [en ligne]. 28 août 2020.
- ISAAC et al., 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. In : Methods in Ecology and Evolution. 2014. Vol. 5, n° 10, pp. 1052-1060.
- JOHNSTON et al., 2020. Estimating species distributions from spatially biased citizen science data. In : Ecological Modelling. 15 avril 2020. Vol. 422, pp. 108927.
- KEARNEY et al., 2010. Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. In : Conservation Letters. 2010. Vol. 3, n° 3, pp. 203-213.
- KELLING et al., 2015. Taking a 'Big Data' approach to data quality in a citizen science project. In : Ambio. 1 novembre 2015. Vol. 44, n° 4, pp. 601-611.
- LOBO et al., Castor, 2018. KnowBR: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. In : Ecological Indicators. 1 août 2018. Vol. 91, pp. 241-248.
- MAIR et al., 2016. Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa. In : PLoS ONE. 28 janvier 2016. Vol. 11, pp. e0147796.
- MCDONOUGH MACKENZIE et al., 2017. Lessons from citizen science: Assessing volunteer-collected plant phenology data with Mountain Watch. In : Biological Conservation. 1 avril 2017. Vol. 208, pp. 121-126.
- RAZAFIMAHAFI, Ntsoa Rakoto, 2017. Modélisation de la distribution d'espèces benthiques par l'utilisation d'une plateforme multiagents. Mémoire de fin d'études. S.I. Agrocampus Ouest.
- ROBINSON et al., 2017. Correcting for bias in distribution modelling for rare species using citizen science data. In : Diversity and Distributions. 27 décembre 2017. Vol. 24.
- STEEN et al., 2019. An evaluation of stringent filtering to improve species distribution models from citizen science data. In : Diversity and Distributions. 2019. Vol. 25, n° 12, pp. 1857-1869.
- TIEDEMAN et al., 2019. The quality and contribution of volunteer collected animal vehicle collision data in ecological research. In : Ecological Indicators. 1 novembre 2019. Vol. 106, pp. 105431.
- TROUDET et al., 2017. Taxonomic bias in biodiversity data and societal preferences. In : Scientific Reports. 22 août 2017. Vol. 7, n° 1, pp. 9132.
- VAN STRIEN et al., 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. In : Journal of Applied Ecology. 2013. Vol. 50, n° 6, pp. 1450-1458.



Equipe projet : Guillaume Morante, Anaïs Gsell-Epailly
Photos et cartographies : ODONAT Grand Est
Mise en page : Anaïs Gsell-Epailly
Validation et relecture : Yves Muller, Raynald Moratin, Aurore Sindt

Avec le soutien de :

