Estimation en France des densités d'anguilles jaunes d'Europe (Anguilla anguilla) basée sur leur diffusion dans les bassins versants : Modèle TABASCO (spaTialized Anguilla BASin COlonisation). Rapport final année 3.

J. Domange, Hilaire Drouineau, Cédric Briand, Laurent Beaulaton, Patrick Lambert

Estimation en France des densités d'anguilles jaunes d'Europe *(Anguilla anguilla)* basée sur leur diffusion dans les bassins versants : Modèle TABASCO (spaTialized Anguilla BASin COlonisation)

Rapport final année 3

Thème 11 - Action 23 : Gestion des poissons migrateurs

Jocelyn Domange[(1)], Hilaire Drouineau[(1)], Cedric Briand[(2)], Laurent Beaulaton[(3)], Patrick Lambert[(1)].

(1) **IRSTEA EABX**
(2) **EPTB-Vilaine**
(3) **ONEMA**

6 janvier 2016

Résumé

Estimation des densités d'anguilles jaunes d'Europe (*Anguilla anguilla*) en France métropolitaine, basée sur leur diffusion dans les bassins versants : Modèle TABASCO (spaTialized Anguilla BASin COlonization).

L'approche probabiliste du modèle EDA nous a conduit à expérimenter une démarche alternative de modélisation nommée TABASCO pour « *spaTialized Anguilla BASin COlonization* ». Le modèle TABASCO doit être considéré comme un intermédiaire entre des méthodes strictement statistiques et des modèles mécanistes. Il repose sur la propagation d'une cohorte d'anguilles, formalisée par une distribution gaussienne, au travers d'un graphe orienté associé au réseau hydrographique d'un bassin versant. La résultante de ce processus de diffusion doit permettre d'estimer la répartition et le nombre d'anguilles jaunes à la fin de la colonisation d'un bassin. Un travail de codage en langage C++, avec en parallèle une calibration avec les données de pêches électriques sur un bassin versant de référence (Gironde), a mené à l'établissement d'une version stable du modèle. Dans un second temps, le code a été adapté pour prendre en compte l'ensemble des bassins versants dont l'exutoire est situé en France métropolitaine, ce qui représente un total de 953 bassins pour une surface de 500 979 km^2 (soit 90,81 % de la superficie de la France métropolitaine). La calibration du modèle est exécutée année par année entre 1990 et 2009, et un effet « UGA » (Unités de Gestion Anguille) a été ajouté pour tenir compte de la répartition géographique des anguilles lors de leur arrivée sur les côtes françaises et ainsi affiner les prédictions du modèle. Enfin, les caractéristiques de TABASCO sont prévues pour pouvoir évaluer les mortalités en montaison et les comparer avec les mortalités en dévalaison. Ce rapport technique inclut une discussion des résultats et un diagnostic devant permettre d'améliorer ou de corriger ultérieurement le modèle.

Mots-clefs :

modélisation ; estimation de paramètres ; dynamique de population ; gaussienne ; anguille ; *Anguilla anguilla* ; colonisation ; bassins versants ; réseau hydrographique ; diffusion ; topologie ; France métropolitaine ; rupture de connectivité ; écosystèmes.

Abstract

Assessment of European yellow eel (*Anguilla anguilla*) densities in metropolitan France, based on their diffusion in catchments with the TABASCO model.

The probabilistic approach of the EDA model led us to experiment an alternative modelling method named TABASCO for « spaTialized Anguilla BASin COlonization assessment model ». TABASCO should be considered as an intermediary between strictly statistical approaches and mechanistic models. It is based on the spread of a cohort of eels formalized by a Gaussian distribution through an oriented graph associated to the hydrographic network of a catchment. The result of this diffusion process should allow to estimate the distribution and the number of yellow eels at the end of the colonization of a catchment. A work of C++ coding, together with a calibration with the electrofishing data of a reference catchment (Gironde) led to a stable version of the model. The code has then been adapted to take into account all the catchments whom outlet is located in metropolitan France, which represents a total of 953 catchments and a 500979 km^2 surface (90,81 % of the metropolitan France). The calibration is now carried out year per year between 1990 and 2009, and an « EMU » (Eel Management Units) effect has been added to take into account the geographical distribution of the eels when they arrive at the french coasts, thus refining the model predictions. Moreover, the features of TABASCO have been planned to assess upstream mortalities and to compare them with downstream mortalities. This technical report includes a discussion of the results and a diagnosis in order to improve or rectify the modelling later.

Key words :

modelling; parameter estimation; population dynamics; Gaussian; eel; *Anguilla anguilla*; colonization; catchments; hydrographic network; diffusion; topology; metropolitan France; rupture of connectivity; ecosystems.

Table des figures

Sommaire

Table des matières

Chapitre 1

Synthèse pour l'action opérationnelle

1.1 Contexte général

Le déclin inquiétant de la population d'anguilles européennes observé depuis les années 1980 a amené en 2007 les gestionnaires à décréter des mesures de reconstitution du stock d'anguilles. Ces mesures portent sur les différents types de pêcheries, les obstacles à la circulation des anguilles, le repeuplement, la restauration des habitats et les contaminations. Le plan français de gestion s'est jusqu'à présent principalement appuyé sur la modélisation EDA (Eel Density Analysis), développée depuis 2009 pour calculer l'échappement en anguilles argentées et les différents points de référence (Jouanin *et al.*, 2012b; Briand *et al.*, 2015).
Une approche de modélisation alternative est étudiée depuis 2012 en vue d'essayer d'améliorer les estimations du stock d'anguilles jaunes et de lever certaines limitations du modèle EDA. Il s'agit du modèle TABASCO (pour *spaTialized Anguilla BASin COlonization*), pensé pour être à mi-chemin entre une approche purement statistique et un modèle mécaniste.
Le modèle TABASCO doit permettre d'évaluer les densités d'anguilles jaunes en France métropolitaine à partir de pêches scientifiques en décrivant explicitement le processus de colonisation des bassins versants par les anguilles. Il intègre une prise en compte des caractéristiques topologiques de l'ensemble du réseau hydrographique ainsi qu'un effet spatial correspondant aux Unités de Gestion Anguille (UGA). L'approche calculatoire repose sur la propagation au travers d'un graphe orienté d'une distribution gaussienne correspondant à la résultante d'une diffusion. Une calibration sur vingt années de données (1990-2009) permet d'observer l'évolution du stock au cours du temps. Par ailleurs, le modèle est conçu pour tenir compte de l'effet des obstacles sur la colonisation, et peut être utilisé pour estimer les mortalités à la montaison (par prédation sur les animaux bloqués) et à la dévalaison (lors du passage des turbines hydroélectriques). Enfin, l'utilisation de la méthode du maximum de vraisemblance lors de l'estimation des paramètres de sortie du modèle permet de calculer explicitement l'intervalle de confiance sur les résultats.

1.2 Information sur les méthodes et données utilisées

Le modèle s'appuie sur deux sources d'informations géographiques, qui sont le Réseau Hydrographique Théorique français (RHT, Pella *et al.* (2012)) et le Référentiel des Obstacles à l'Écoulement (ROE). Les données de pêche utilisées proviennent quant à elles de la Base de Données sur les milieux aquatiques et les poissons (BDmap). L'extraction des données pour la calibration du modèle rassemble 19 201 opérations

de pêches électriques sur 11 371 stations, réalisées entre 1966 et 2009. Sur la période d'application retenue (1990-2009), 8 078 observations sont utilisées, dont 5 170 points de données nulles (pas d'observation d'anguilles) et 2 908 points de données positives.
Le RHT, après un redécoupage par le ROE, a été exporté dans le système d'information géographique (SIG) PostGIS, en le structurant en tronçons et nœuds. Un total de 61 961 obstacles du ROE ont été projetés sur le RHT, puis une sélection de 51 769 obstacles a été effectuée pour supprimer certains types d'ouvrages à effet négligeable, comme les ponts et les digues. Les tronçons ont été scindés en deux au niveau de chaque obstacle et les attributs des deux sous-tronçons résultants ont été recalculés. Ainsi, dans le modèle, les obstacles se situent systématiquement au niveau de nœuds. 10 817 stations de pêches électriques sont également projetées sur la topologie.
Le modèle est entièrement codé en C++, et l'export des sorties est compatible avec un traitement sous R ou sous n'importe quel logiciel de calcul scientifique.
Le réseau hydrographique et tous ses attributs ont été formalisés en un graphe orienté grâce à l'utilisation de la bibliothèque logicielle BGL (pour *Boost Graph Library*, Siek *et al.* (2002)). Le programme fonctionne bassin versant par bassin versant. Ainsi, l'utilisateur peut décider de la liste de bassins qu'il souhaite intégrer dans la modélisation.
Les calculs sont ensuite basés sur la détermination de la distribution résultante des anguilles jaunes à la suite du processus de diffusion. Il s'agit en fait d'une adaptation de la loi de Fick. L'onde de diffusion étant une fonction gaussienne dont la formule est explicite, il est possible de calculer la proportion d'anguilles qui se sédentarisent dans chaque tronçon du réseau en tenant compte des obstacles éventuels et de la topologie du réseau.
La calibration du modèle a été réalisée en utilisant la différentiation automatique proposée dans le paquetage logiciel ADMB (*Automatic Differentiation Model Builder*, Fournier *et al.* (2012)).
Nous avons testé deux hypothèses sur l'écart-type de la distribution des anguilles (ci-après dénommées option n°1 et option n°2) : l'une pour laquelle l'écart-type est constant sur tout le réseau hydrographique, et l'autre pour laquelle il dépend de la longueur de chaque drain.
Une approche alternative basée sur des matrices de transition a également été explorée, mais son développement a été stoppé, faute de temps. Toutefois, cette seconde méthode est fonctionnelle même si des problèmes de calibration (optimisation d'une valeur entière) restent à résoudre.

1.3 Principaux acquis transférables obtenus

La version finale du modèle tourne en moins de deux heures, toutes étapes confondues (génération des graphes, imports-exports des données, calibration et optimisation). Les principaux résultats obtenus concernent la répartition des anguilles jaunes en France métropolitaine en terme de densité, et l'évolution du stock en fonction du temps. Les prédictions des densités sont données sur les figures 1.1 et 1.2 pour les deux options de calcul testées, pour l'année 2009. Le patron classique de distribution des anguilles jaunes (fortes densités à proximité des exutoires, et des zones de quasi absence dans les zones les plus amont) est respecté, avec une pénétration plus importante dans les drains principaux pour l'option d'un écart-type de la gaussienne qui dépend de la longueur du drain.
L'évolution du stock d'anguilles jaunes pour la France métropolitaine est présentée ci-dessous sur les figures 1.3 et 1.4 pour les options n° 1 et n° 2 respectivement. On observe dans les deux cas une tendance à la baisse similaire conduisant à des taux de décroissance annuelle de 4,1 et 4,9 % (valeur de 0.041 et de 0.049 pour le coefficient b dans l'expression a.exp(-bt)) respectivement pour l'option n°1 et l'option n°2. Pour autant, les intervalles de confiance importants imposent une prudence dans l'utilisation de ces résultats.
Des problèmes dans l'estimation de certains paramètres (notamment le franchissement

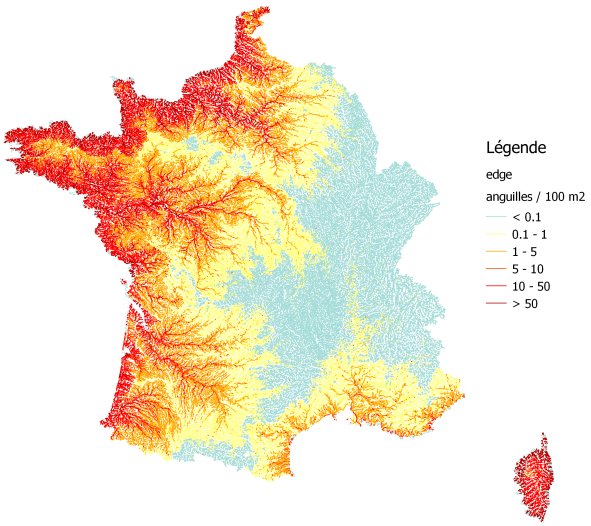

FIGURE 1.1 : Carte des densités d'anguilles jaunes prédites en France métropolitaine pour l'année 2009 avec l'option de calcul n° 1. Les densités sont affichées pour chaque tronçon du réseau hydrographique. La couleur du tronçon correspond à la valeur de la densité prédite selon une échelle allant du gris-vert (pas d'anguille) au rouge foncé (plus de 50 anguilles par dam^2).

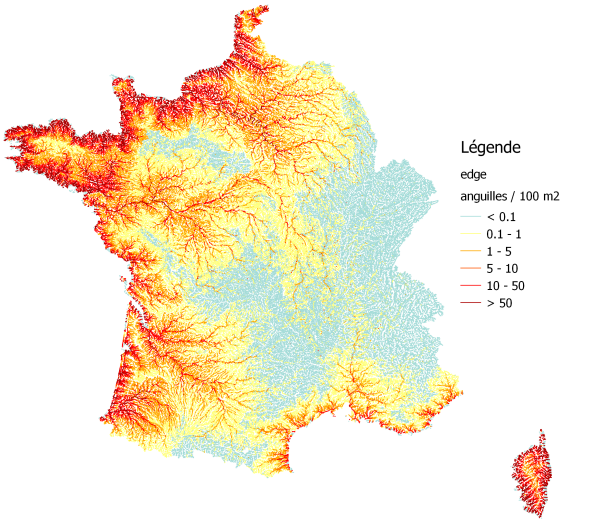

FIGURE 1.2 : Carte des densités d'anguilles jaunes prédites en France métropolitaine pour l'année 2009 avec l'option de calcul n° 2. Les densités sont affichées pour chaque tronçon du réseau hydrographique. La couleur du tronçon correspond à la valeur de la densité prédite selon une échelle allant du gris-vert (pas d'anguille) au rouge foncé (plus de 50 anguilles par dam^2).

des obstacles) nous ont par contre empêché de mener une comparaison des mortalités à la montaison et à la dévalaison. Ce point devrait figurer dans les priorités d'améliorations du modèle à l'avenir. Il conviendrait également de travailler en fonction de la

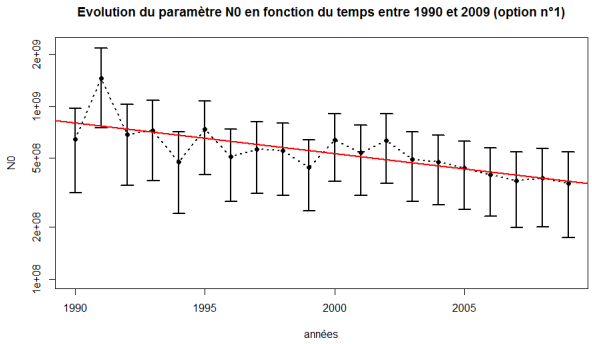

FIGURE 1.3 : Évolution du paramètre N0 en fonction du temps, entre 1990 et 2009, avec intervalle de confiance à 95 % (1,96 σ) sur l'estimation et meilleur ajustement linéaire (en rouge), pour l'option de calcul n° 1. Les axes sont en échelle logarithmique. Le nombre est donné pour la France (métropolitaine) entière.

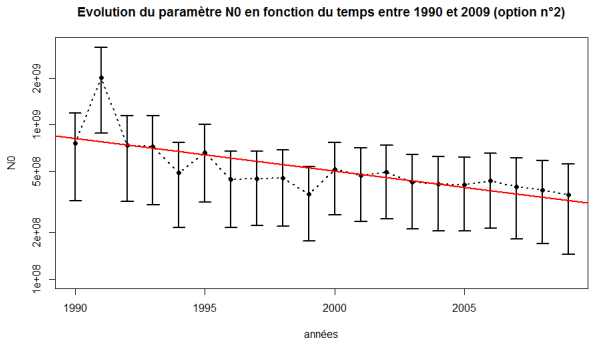

FIGURE 1.4 : Évolution du paramètre N0 en fonction du temps, entre 1990 et 2009, avec intervalle de confiance à 95 % (1,96 σ) sur l'estimation et meilleur ajustement linéaire (en rouge), pour l'option de calcul n° 2. Les axes sont en échelle logarithmique. Le nombre est donné pour la France (métropolitaine) entière.

surface en eau et non plus en fonction du linéaire de berges pour le calcul de la répartition des anguilles dans chaque tronçon (proportion d'anguilles sédentarisées). Cela éviterait probablement un biais sur les densités prédites. Enfin, une implémentation d'un effet « miroir » qui rendrait compte de l'effet des sources sur la distribution des anguilles améliorerait également probablement les prédictions du modèle.

1.4 Travaux scientifiques issus du projet

Ce projet de modélisation a conduit à plusieurs valorisations scientifiques, dont voici la liste :

- Trois rapports techniques : deux rapports intermédiaires en 2013 et 2014, et un rapport final (ci-joint) en 2015.
- Une présentation du concept du modèle à la 144^{ieme} conférence annuelle de l'*American Fisheries Society* (AFS) à Québec au Canada (Lambert *et al.*, 2014).
- Un poster présentant la version finale du modèle en juin 2015 à la conférence internationale « *Fish Passage* » à Groningen aux Pays-Bas (Domange *et al.*, 2015).

Chapitre 2

Introduction

2.1 Contexte général

Afin de réagir au déclin inquiétant de la population d'anguilles européennes observé depuis les années 1980, la commission européenne a institué en septembre 2007 un règlement qui décrète des mesures de reconstitution du stock d'anguilles et impose à chaque État membre de soumettre un plan de gestion de sauvegarde de l'espèce (règlement européen n° 1100/2007 du 18 septembre 2007).
Conformément à ce règlement communautaire qui astreint chaque état membre concerné à mettre en œuvre pour le 1^{er} juillet 2009 un plan par unité de gestion anguille[1], la France a envoyé son plan national le 3 février 2010, et celui-ci a été approuvé par la Commission européenne le 15 février 2010. Son élaboration a été pilotée par les ministères en charge des pêches maritimes et de l'écologie. Nous rappelons sur la figure 2.1 le découpage retenu pour la définition des unités de gestion anguilles.
Les mesures préconisées portent sur les différents types de pêcheries, les obstacles à la circulation des anguilles, le repeuplement, la restauration des habitats et les contaminations.
Programmées sur le court et le moyen terme (horizon 2012-2015), ces mesures sont porteuses d'objectifs ambitieux en matière de réduction des mortalités par la pêche ou liées aux ouvrages.
Une première évaluation de ces plans a été remise en juin 2012. Le plan français, tant dans sa définition que dans sa post-évaluation, s'est jusqu'à présent principalement appuyé sur la modélisation EDA (Eel Density Analysis) pour calculer l'échappement en anguilles argentées et les différents points de référence (Jouanin *et al.*, 2012b).
L'ONEMA, l'Institut d'aménagement de la Vilaine et Irstea développent en effet EDA depuis 2010. Le modèle a été appliqué à plusieurs bassins versants européens et, dans le cadre de la post-évaluation, à l'ensemble des unités de gestion anguille françaises. Un test de fiabilité d'EDA a par ailleurs été réalisé sur une réalité construite à partir du modèle CREPE qui synthétise, sous forme d'une approche individus centrée, la compréhension du fonctionnement d'une fraction de population d'anguilles dans un bassin versant (Lambert, 2012). Au final, il s'avère que dans certains cas, l'extrapolation aux secteurs où la pêche électrique n'est pas possible peut conduire à des estimations d'abondance irréalistes ou des répartitions d'individus dans les réseaux hydrographiques biaisés.
En conséquence, un nouveau type de modélisation est étudiée depuis 2012 en vue d'essayer d'améliorer les estimations du stock d'anguilles jaunes. Il s'agit du modèle TABASCO, qui fait l'objet de ce rapport, et qui a été pensé pour être à mi-chemin entre une approche purement statistique et un modèle mécaniste.

[1] Les « bassins versants anguille » pris en compte comme unités de gestion anguille (UGA) sur les territoires ont été déterminés selon des critères validés par le Comité de Gestion des Poissons Migrateurs compétent sur ces territoires. Les efforts de gestion mis en place sur ces zones doivent

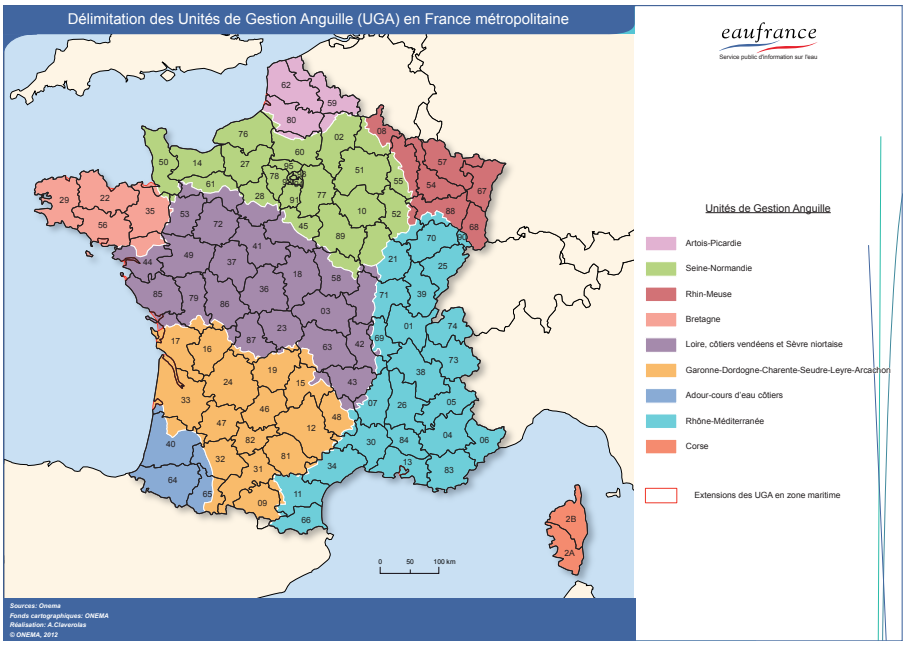

FIGURE 2.1 : Les unités de gestion anguille (UGA) en France.

2.2 Rappel préalable : le modèle EDA

2.2.1 Bases du modèle EDA

Le modèle EDA a été initialement développé pour étudier l'impact des barrages sur la répartition des anguilles dans les cours d'eau bretons (Leprévost, 2007). Il a par la suite été appliqué à l'ensemble du territoire métropolitain dans le cadre de la définition du plan français gestion de l'anguille européenne (2010), puis dans le cadre de l'évaluation de ce même plan (2012). Il a également fait l'objet d'applications dans des unités de gestion en Europe (Walker *et al.*, 2011). Tout au long de ce processus il a fait l'objet d'améliorations successives, en particulier en intégrant des impacts anthropiques au travers de l'occupation du sol (Jouanin *et al.*, 2012b). Son principal atout est la possibilité de son application à large échelle à partir des données de réseau de pêches électriques actuellement disponibles.
Le principe de cette approche (Jouanin *et al.*, 2012b) est :

1. de relier les densités d'anguilles jaunes observées lors de pêches électriques à différents paramètres : méthodes d'échantillonnage, conditions environnementales (distance à la mer, distance relative, température, altitude,...), conditions anthropiques (obstacles, occupation du sol,...) et année d'observation,

2. d'extrapoler les densités d'anguilles jaunes dans chaque tronçon du réseau hydrographique en appliquant un modèle statistique calibré à l'étape 1,

3. de calculer l'abondance totale du stock d'anguilles jaunes en multipliant ces densités par la surface en eau des tronçons et en les additionnant,

4. de calculer un échappement potentiel en convertissant le stock estimé d'anguilles jaunes à l'étape 3 en stock d'anguilles argentées,

5. de calculer un échappement effectif en soustrayant les mortalités d'anguilles argentées (pêcheries, turbines) connues ou estimées,

contribuer à augmenter la part de population d'anguilles dévalantes.

6. de donner une estimation de l'échappement pristine en considérant les conditions anthropiques mises artificiellement à zéro et un jeu temporel de variables avant 1980.

La prise en compte des mortalités par les turbines a fait l'objet d'un développement particulier en 2012 (Jouanin *et al.*, 2012a).
L'approche retenue dans EDA est basée sur le modèle statistique delta-gamma proposé par Stefánsson (Stefánsson et Pálsson, 1996) qui permet de traiter des données présentant une surreprésentation de valeurs nulles. En effet, les effectifs d'anguilles étant faibles et leur distribution hétérogène, l'occurrence d'absences d'observations, et donc de densités mesurées nulles, est forte. L'estimation des densités d'anguilles jaunes est ainsi réalisée au travers de la multiplication d'un modèle de présence-absence (modèle Δ) et d'un modèle de densités non nulles (modèle Γ). Les réponses des variables à modéliser n'étant pas linéaires, des modèles additifs généralisés (GAM) (Hastie et Tibshirani, 1990) ont été utilisés, avec une distribution binomiale et un lien Logit[2] pour le modèle Δ et une distribution Gamma[3] et un lien logarithme pour le modèle Γ (Jouanin *et al.*, 2012b).
Parmi les variables explicatives potentielles, n'ont été gardées que celles qui avaient une répartition comparable dans les jeux d'apprentissage (tronçons avec pêches électriques) et d'extrapolation (l'ensemble des tronçons). Ainsi, la distance à la source, la distance relative (ratio de distance à la mer sur la longueur du drain principal), la surface amont du bassin versant et la pente ont été écartées (au risque de fragiliser les prédictions).
Plusieurs améliorations successives ont été réalisées, notamment la prise en compte des hauteurs d'obstacles, pour obtenir une version finale du modèle en 2015. L'ensemble des caractéristiques de cette version est décrit dans le rapport final du juin 2015 (Briand *et al.*, 2015).

2.2.2 Perspectives d'améliorations du modèle EDA

Dans la version actuelle d'EDA, plusieurs pistes d'améliorations ont été identifiées. Tout d'abord, concernant la distribution naturelle des anguilles dans un réseau hydrographique :

- Le flux de colonisants se répartissant entre tributaires à chaque confluence en plusieurs sous-flux de tailles différentes, les densités de deux tronçons à la même distance de la mer ne devraient pas nécessairement avoir la même densité d'anguilles en fonction du nombre et de la nature des confluences à l'aval (prise en compte de la topologie aval du réseau).

- Deux tronçons à la même distance de la mer ne devraient pas nécessairement avoir la même densité d'anguilles quelle que soit la surface du bassin versant en amont. Cela pose notamment la question de savoir si la relation entre superficie et surface en eau des bassins est linéaire ou pas. On peut supposer que le débit (dont un proxy est la surface de bassin versant amont) influence la colonisation des bassins versants par les anguilles, soit en favorisant cette colonisation avec la notion de

[2]La fonction Logit est une fonction mathématique très utilisée en statistiques et pour la régression logistique. Son expression est :

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

Où p est un nombre compris entre 0 et 1 (typiquement une probabilité). Si le logarithme utilisé est la fonction logarithme népérien, la fonction Logit est la réciproque de la fonction sigmoïde.

[3]La distribution Gamma est une loi de probabilité que l'on peut paramétrer à l'aide d'un paramètre de forme α et d'un paramètre d'intensité β, de telle sorte que sa fonction de densité de probabilité peut se mettre sous la forme :

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\beta\, x}$$

Où α et β sont deux paramètres strictement positifs, et où $\Gamma(x)$ est la fonction Gamma d'Euler.

débit d'attrait, soit comme facteur bloquant la progression des animaux. Une question encore non tranchée est de savoir si cette taille de bassin versant ne joue qu'au niveau de la surface en eau du tronçon (la largeur évoluant en fonction de la racine carrée de la taille de bassin versant) sans influencer la densité.

- Deux tronçons à la même distance de la mer ne devraient pas nécessairement avoir la même densité d'anguilles quelle que soit la distance à la source (non prise en compte de la topologie amont). On peut en effet imaginer qu'une station plus proche de la source, donc moins productive, abrite une densité d'anguilles plus faible.

Concernant maintenant la prise en compte des impacts anthropiques :

- Un barrage ne modifie pas la densité à l'aval. Or, on peut raisonnablement supposer que la densité à l'aval d'un barrage est augmentée (au moins sur une certaine distance) de façon conjointe à une diminution de la densité à l'amont. Dans une simulation avec la version actuelle d'EDA, la suppression d'un barrage conduit à augmenter l'abondance des anguilles en amont sans modifier les densités à l'aval. Cette apparente création d'anguilles suite à l'effacement de l'obstacle revient implicitement, lors de simulations avec le barrage, à considérer un blocage d'individus à l'aval de l'obstacle équivalent en nombre à cette création et à une mortalité totale des mêmes anguilles.

- Les prélèvements d'anguilles jaunes ne sont pas explicitement traités, ils sont supposés se retrouver dans les densités d'anguilles observées lors des pêches électriques (cependant, ce point ne sera pas non plus pris en compte dans TABASCO).

Il est fort vraisemblable que la méthodologie EDA, moyennant l'ajout de nouvelles variables (surface totale du bassin versant, nombre de confluences à l'aval, ratio entre surface de bassin versant amont et surface du bassin versant total, distance à la source, distance au premier barrage en amont, etc...) et d'un jeu d'apprentissage plus représentatif du réseau total, soit en mesure de lever, au moins partiellement, ces difficultés. Concernant ensuite les techniques statistiques :

- La sensibilité des modèles additifs généralisés à la corrélation entre variables explicatives, qui ne garantit pas une causalité avec la variable à expliquer. C'est en particulier gênant pour les extrapolations, notamment si l'on souhaitait prédire ce que serait la distribution des anguilles sans barrages. L'approche retenue dans TABASCO ne garantit cependant pas nécessairement de meilleurs résultats à ce niveau.

Concernant enfin le jeu de données d'apprentissage :

- La sous-représentation des stations d'échantillonnage dans les milieux profonds (drains principaux à l'aval des bassins versants) rendant les extrapolations hasardeuses. Il est néanmoins certain que l'on sera confronté au même problème dans TABASCO, puisqu'il n'existe de toute façon pas de données dans ces zones.

2.2.3 Proposition d'une nouvelle approche : TABASCO

Il a ainsi été choisi de construire un nouveau modèle, parcimonieux en paramètres, analysant les densités d'anguilles mais décrivant explicitement le processus de colonisation sous-jacent à la distribution dans un réseau hydrographique. Il a été nommé TABASCO pour « spaTialised Anguilla BASin COlonisation assessment model ». TABASCO doit donc être considéré comme un intermédiaire entre des approches strictement statistiques comme EDA et des modèles mécanistes comme GlobAng (Lambert et Rochard, 2007) ou SMEP (Aprahamian *et al.*, 2007). Actuellement, il est admis que ce processus

dans un bassin versant est un phénomène diffusif (Ibbotson *et al.*, 2002) avec éventuellement une composante advective[4] vers l'amont sur une courte distance durant les premières années de vie continentale de l'animal (C. Rigaud et al , en préparation). EDA retrouve d'ailleurs une courbe de réponse de l'abondance en fonction de la distance à la mer qui approche une décroissance gaussienne, en accord avec le caractère diffusif de la colonisation[5].
L'approche de modélisation retenue repose sur la propagation au travers d'un graphe orienté d'une distribution gaussienne correspondant à la résultante d'une diffusion (rendant compte de la colonisation des bassins versants par les jeunes anguilles).
TABASCO, par sa représentation explicite de la structure du réseau hydrographique, rend normalement mieux compte de la topologie des bassins. Il permet aussi, à l'instar d'EDA, une évaluation des stocks au cours du temps grâce à une calibration sur plusieurs années et prédit aussi les variations géographiques du stock en fonction des Unités de Gestion Anguille. Par ailleurs, l'impact des obstacles est traité directement en considérant explicitement le processus sous-jacent au niveau du barrage, même si nous utilisons une franchissabilité moyenne sur toute la France qui ne rend pas correctement compte du type d'ouvrage et donc de la capacité de blocage associée.
Nous récapitulons dans le tableau 2.1 ci-dessous les points que le modèle TABASCO a pu améliorer ou résoudre et ceux qui n'ont pas pu être pris en compte ou améliorés. Une discussion des résultats du modèle est par ailleurs effectuée plus loin dans ce rapport.

2.2.4 Zone actuelle d'application du modèle TABASCO

Dans ce rapport, nous devons limiter la zone d'application du modèle à tous les bassins versants de France métropolitaine dont l'exutoire est situé sur le territoire français. Cela revient à omettre environ 9 % de la superficie de la France métropolitaine dans la simulation. On notera qu'en principe, aucune limitation n'empêche l'utilisation du modèle TABASCO en dehors du territoire national. Ce choix est imposé par l'utilisation du Réseau Hydrographique Théorique français, que nous présentons dans la section suivante, et qui n'est pas défini en dehors du territoire français métropolitain. Nous ne pouvons donc pas gérer les calculs relatifs aux cours d'eau dont les bassins versants sont à l'étranger. La surface résultante considérée est affichée sur la carte de la figure 2.2.
Le nombre total de bassins versants concernés est de 953. La surface correspondante est de 500 979 km^2, répartis sur les différentes unités de gestion anguille comme indiqué dans le tableau 2.2. La France métropolitaine ayant une superficie totale de 551 695[6] km^2, le modèle couvre 90,81 % de la France métropolitaine. La surface manquante correspond essentiellement aux unités de gestion anguille Rhin et Meuse, ainsi qu'à une petite partie des unités de gestion Artois et Rhône.

[4] On rappelle que l'advection correspond au transport d'une quantité (scalaire ou vectorielle) par un champ vectoriel. Mathématiquement, l'opérateur advection correspond au produit scalaire du vecteur vitesse par le vecteur gradient (opérateur nabla). En ce qui nous concerne, il s'agit du phénomène de transport orienté des anguilles vers l'amont par leur mouvement propre (*i.e.* par leur nage).

[5] Pour être précis, EDA prédit une décroissance exponentielle de l'abondance. Or, la loi normale est une loi de la famille exponentielle. avec une décroissance dite surexponentielle à droite de la valeur moyenne(Lifschitz, 1995). Il n'est donc pas erroné de comparer les variations de ces deux fonctions sur leur domaine de décroissance.

[6] Surface géodésique de référence selon l'Institut Géographique National (IGN).

Améliorations acquises	**Problèmes non traités**
Prise en compte de l'effet moyen des obstacles sur l'effectif amont (blocage)	Effets des prélèvements d'anguilles jaunes
Prise en compte de l'effet moyen des obstacles sur l'effectif aval (accumulation)	Manque de données dans les drains principaux, notamment dans les 50 km avant l'exutoire
Prise en compte de la surface en eau amont (répartition pondérée par l'espace disponible)	Meilleure extrapolation (sensibilité réduite aux variables corrélées)
Prise en compte du continuum amont-aval	Cartographie par classe d'âge
Prise en compte de la topologie du réseau	Prise en compte de la nature des ouvrages
Calcul des erreurs statistiques sur la détermination des paramètres optimisés	
Prise en compte des variations géographiques du stock d'anguilles (effet "UGA")	
Calcul des mortalités en montaison (prédation) et en dévalaison (mortalités aux turbines)	

TABLE 2.1 : Tableau récapitulatif des améliorations et des problèmes non traités du modèle TABASCO.

UGA	**Surface (km^2)**
Adour	20 007,3
Artois	11 385,7
Loire	126 331,8
Seine	93 810,4
Bretagne	27 992,0
Garonne	96 264,2
Rhône	117 838,9
Corse	7 349,1
Total	500 979,4

TABLE 2.2 : Contributions des différentes unités de gestion anguille à la superficie totale considérée à ce jour dans le modèle TABASCO.

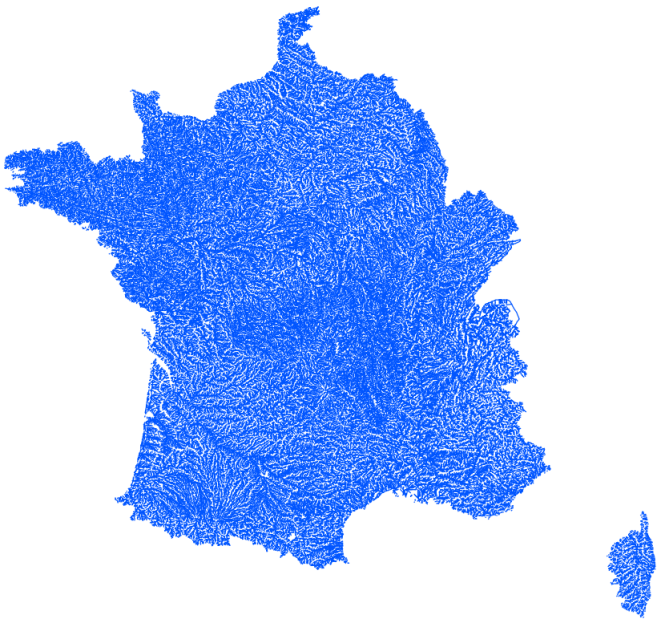

FIGURE 2.2 : Carte de la surface de France métropolitaine actuellement prise en compte dans la simulation. On notera que l'essentiel du territoire omis dans la simulation se situe dans l'UGA Artois (nord de la France) et l'UGA Rhin-Meuse (nord-est de la France).

Chapitre 3

Principes du modèle TABASCO

3.1 Modélisation du réseau hydrographique français

3.1.1 Objets contenant l'information géographique

Le modèle TABASCO s'appuie principalement sur les sources d'information géographique que sont le réseau hydrographique théorique français et le référentiel des obstacles à l'écoulement, ainsi que sur la base de données sur les milieux aquatiques et les poissons. Le modèle utilise les données géoréférencées dans ces différents objets, et permet un affichage spatialisé des résultats de la simulation par projection sur ces mêmes référentiels.

Le RHT

Le réseau hydrographique théorique (RHT) est un nouveau réseau hydrographique dérivé de la BD Alti® de l'Institut Géographique National (IGN) (Pella *et al.*, 2012). La méthode « Agree » utilisée pour construire ce réseau permet de modifier un modèle numérique de terrain à partir d'un réseau d'écoulement pré-défini. Ainsi, le RHT est développé à partir du modèle numérique de terrain de la BD Alti® re-conditionné par le RHE (Réseau Hydrographique Étendu), qui est lui-même une simplification du réseau hydrographique de référence de l'IGN, la BD Carthage®. La compatibilité entre le RHT et la BD Alti® permet d'identifier les bassins versants et de simuler des écoulements avec une meilleure précision. Cette approche permet d'intégrer un ensemble d'attributs spatialisés et de les cumuler le long du réseau. Ainsi, des attributs topographiques, hydrologiques et climatiques sont calculés et intégrés dans un système d'information géographique.

Le ROE

En France métropolitaine, un total de 69 136 obstacles à l'écoulement – barrages, écluses, seuils, etc... - ont été recensés sur les cours d'eau. Ils sont à l'origine de profondes transformations de la morphologie et de l'hydrologie des milieux aquatiques, et perturbent fortement le fonctionnement de ces écosystèmes. Dans le cadre de l'évaluation de l'effet de ces obstacles sur les écosystèmes aquatiques, il était nécessaire de les inventorier et de les rassembler dans une base de données nationale. Le Référentiel des Obstacles à l'Écoulement (ROE) recense ainsi l'ensemble des ouvrages inventoriés sur le territoire national en leur associant des informations (code national unique, localisation, typologie) communes à l'ensemble des acteurs de l'eau et de l'aménagement du territoire (Léonard et Zegel, 2010). Il assure aussi la gestion et la traçabilité des informations en provenance des différents partenaires. Le référentiel actuellement mis en ligne est une version 6.0 figée mettant à disposition les données validées et gelées en date du 7 mai 2014. De lourds développements sont actuellement mis en œuvre pour

assurer l'interopérabilité et la pérennité du Référentiel des Obstacles à l'Écoulement au travers d'une mise à disposition en temps réel, via le format d'échange défini par le Service d'Administration Nationale des Données et Référentiels sur l'Eau (SANDRE). Le ROE, dans sa version de septembre 2013, présente 69 136 ouvrages référencés en France.

TABASCO repose sur l'adaptation de ces deux derniers objets d'information géographique (RHT et ROE) sous forme d'un réseau formalisé exploitable directement dans le code du modèle.

La BDmap

La Base de Données sur les Milieux Aquatiques et les Poissons (BDmap) fait partie du SIE (Système d'Information sur l'Eau), géré par l'Onema. Une extraction de données donne accès aux informations sur les pêches électriques d'anguilles. La calibration des paramètres du modèle s'appuie sur une optimisation basée sur ces données de terrain (voir plus loin dans ce rapport). L'extraction issue de la BDmap sur laquelle nous nous basons actuellement pour le modèle TABASCO regroupe 19 201 opérations de pêches électriques sur 11 371 stations et réalisées entre 1966 et 2009. Des précisions supplémentaires sur la nature de ces pêches sont données dans le paragraphe « calibration du modèle » à la fin de ce chapitre.

Formalisation d'un réseau hydrographique

Un réseau hydrographique peut être représenté sous forme d'un **graphe orienté**. En mathématiques, et plus précisément en théorie des graphes, un graphe orienté est un graphe défini par la donnée d'un ensemble de **nœuds** (ou **sommets**, ou **vertex**) V connectés par un ensemble d'**arcs** A, eux-mêmes définis comme étant un couple de sommets (au sens mathématique du terme, c'est-à-dire la donnée des deux sommets dans un ordre déterminé).
Ainsi, tout graphe orienté G s'écrit $G = (V, A)$, avec :
- Les éléments de l'ensemble V qui sont les nœuds, ou sommets, ou encore vertex.
- Les éléments de l'ensemble A qui sont les arcs orientés, représentés par un couple mathématique de sommets.
Un graphe orienté est qualifié de simple s'il ne comprend pas de boucles ni d'arcs multiples (arcs ayant les mêmes nœuds de départ et d'arrivée). Dans le cas contraire on parle de multigraphe.
Le réseau hydrographique que nous modélisons constitue donc un graphe orienté simple.

Adaptation du RHT et du ROE en vue de leur utilisation dans le modèle TABASCO

Le RHT, découpé par le ROE, a été représenté sous forme d'un objet *topology* tel qu'implémenté par le système d'information géographique (SIG) PostGIS (voir la section concernant les outils informatiques du présent rapport pour de plus amples détails sur ce logiciel), en le structurant en tronçons (*reaches* en anglais) et nœuds (*nodes* en anglais).
Un total de 61 961 obstacles maîtres[1] du ROE ont pu être projetés sur le RHT, puis une sélection de 51 769 obstacles a été effectuée pour supprimer certains types d'ouvrages à effet négligeable, comme les ponts et les digues. Les tronçons ont été scindés en deux au niveau de chaque obstacle et les attributs des deux sous-tronçons ont été recalculés. Ainsi, les obstacles se situent systématiquement au niveau de nœuds. Cette étape tourne sur la France en environ quatre heures.
10 817 stations de pêches électriques ont également pu être projetées sur la topologie

[1] Dans le cas d'un ensemble hydraulique comprenant plusieurs obstacles groupés sur un même cours d'eau et pour lesquels l'effet d'un des éléments prévaut sur tous les autres, on ne considère que ce seul obstacle principal, dit obstacle maître.

créée afin d'identifier les tronçons dans lesquels elles ont été réalisées.
Dans le code, les nœuds et les tronçons du réseau hydrographique ont été récupérés depuis le RHT et le ROE (voir la section outils informatiques pour la méthode utilisée) et un graphe a été créé grâce à l'utilisation d'une bibliothèque logicielle spécifique (BGL, pour Boost Graph Library) en associant les *nodes* et les *reaches* du réseau à des *vertices* et à des *edges* du graphe (Siek *et al.*, 2002).
En outre, le programme a été codé de telle façon qu'il fonctionne bassin versant par bassin versant. Ainsi, tout le réseau hydrographique n'est pas créé d'un seul tenant, mais chaque sous-réseau associé aux différents bassins-versants français est créé successivement. Cela laisse par ailleurs la liberté de ne travailler que sur un bassin versant donné, ou bien sur un ensemble de plusieurs bassins versants, ou encore sur le réseau hydrographique de la France entière.
Pour chaque bassin versant, le graphe associé au réseau hydrographique est obtenu par construction à l'aide d'une requête SQL récursive à partir de l'exutoire du bassin, c'est à dire à partir du tronçon situé le plus en aval.

3.1.2 Quels outils d'analyse du réseau ?

Nous disposons, via le RHT et le ROE, d'un certain nombre de caractéristiques du réseau hydrographique. A chaque tronçon est associé un jeu de données comprenant, entre autres, la largeur du tronçon, sa longueur, sa distance à la source, son rang de Strahler, ou encore la surface du bassin amont, ainsi qu'un identifiant unique permettant de l'indexer. Nous recalculons par ailleurs la distance à la mer de chaque tronçon dans le modèle. Ces différents paramètres permettent de visualiser le réseau en fonction de certaines de ses caractéristiques, mais aussi de prévoir quel comportement global aura le modèle dans ce bassin. La figure 3.1) donne par exemple la répartition des tronçons dans notre bassin versant de référence (Gironde), en faisant apparaître les distributions marginales et le barycentre de la distribution globale.
Ce type de représentation graphique nous a notamment permis de constater que la majorité des tronçons d'un bassin versant sont situés à des distances relativement faibles de leurs sources, et ce, quel que soit leur éloignement de la mer.
En outre, on voit facilement apparaître les différents drains du bassin sur ce type de

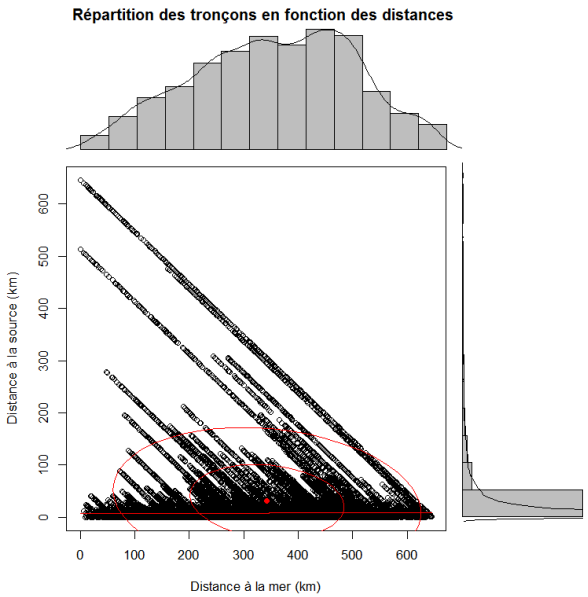

FIGURE 3.1 : Répartition des tronçons du bassin versant de la Gironde en fonction de leur distance à la mer et de leur distance à la source. Les distributions marginales sont également indiquées, ainsi que le barycentre de la distribution globale.

graphe. Tous les tronçons d'un même drain sont sur des droites parallèles d'équation $D_{source} = L_{tot} - D_{mer}$, où L_{tot} représente la longueur totale du drain dans le réseau (depuis sa source jusqu'à l'exutoire). Tous les points du graphe sont contenus dans le triangle formé par les deux axes et la droite correspondant au drain le plus long du réseau (a priori le drain principal du bassin).

3.1.3 Rappels de statistiques utiles pour la suite de ce rapport

Nous rassemblons et explicitons dans cette sous-section les notions de statistiques utiles pour la problématique d'ajustement des données puis pour la discussion des résultats du modèle.

La **fonction de vraisemblance**, notée $\mathscr{L}(x_1, \ldots, x_n|\theta_1, \ldots, \theta_k)$, est une fonction de probabilités conditionnelles qui décrit les valeurs x_i d'une loi statistique en fonction des paramètres θ_j supposés connus. Elle s'exprime à partir de la fonction de densité $f(x|\theta)$ par :

$$\mathscr{L}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{3.1}$$

L'**estimation du maximum de vraisemblance** est une méthode statistique courante, développée par le statisticien Ronald Aylmer Fisher en 1922 (Fisher, 1922) et utilisée pour inférer les paramètres de la distribution de probabilité d'un échantillon donné. Quand θ n'est pas observable, la méthode du maximum de vraisemblance utilise les valeurs de θ qui maximisent $\mathscr{L}(\theta)$ estimateur de θ : c'est l'estimateur du maximum de vraisemblance de θ noté $\widehat{\theta}$. L'estimateur du maximum de vraisemblance peut exister et être unique, ne pas être unique, ou ne pas exister. Le principe est simplement de trouver le maximum de la fonction de vraisemblance pour que les probabilités des réalisations observées soient aussi maximum, ce qui constitue un problème d'optimisation. On maximise souvent le logarithme de la fonction de vraisemblance, qui est plus simple à dériver (étant donné que la fonction de vraisemblance est positive et que la fonction logarithme est une fonction strictement croissante). L'estimateur obtenu par la méthode du maximum de vraisemblance est :

- convergent
- asymptotiquement efficace (il atteint la borne de Cramér-Rao définie ci-dessous)
- asymptotiquement distribué selon une loi normale

En revanche, il peut être biaisé en échantillon fini.

L'**information de Fisher** est une notion introduite par le statisticien du même nom, qui quantifie l'information relative à un paramètre contenue dans une distribution. Dans le cas du modèle TABASCO, la distribution de probabilité considérée dépend de plusieurs paramètres. Ainsi, la recherche du maximum de vraisemblance ne se résume pas à une seule équation mais à un système :

$$E\left[\frac{\partial}{\partial\theta_j} f(X;\theta)\right] = 0, \qquad \forall j \tag{3.2}$$

L'information de Fisher est donc dans ce cas définie comme une **matrice de covariance** :

$$I(\theta_i, \theta_j) = E\left[\left(\frac{\partial}{\partial\theta_i} f(X;\vec{\theta})\right)\left(\frac{\partial}{\partial\theta_j} f(X;\vec{\theta})\right)\right]. \tag{3.3}$$

On notera aussi que l'information de Fisher est une matrice symétrique définie positive. L'inverse de cette matrice permet d'obtenir les **bornes de Cramér-Rao**. Il s'agit en fait des covariances relatives aux estimations conjointes des différents paramètres à

partir des observations. En statistique, la borne Cramér-Rao exprime une borne inférieure sur la variance d'un estimateur sans biais, basée sur l'information de Fisher. Plus formellement, si $\mathcal{I}(\theta)$ est l'information de Fisher d'un paramètre θ, alors son inverse est une borne inférieure de la variance d'un estimateur sans biais de ce paramètre (noté $\widehat{\theta}$). Soit :

$$\text{var}\left(\widehat{\theta}\right) \geq \mathcal{I}(\theta)^{-1} = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}L(X;\theta)\right)^2\right]^{-1} \tag{3.4}$$

Comme notre modèle est régulier[2], la borne de Cramér-Rao peut s'écrire :

$$\mathcal{I}(\theta)^{-1} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}L(X;\theta)\right]^{-1} \tag{3.5}$$

Où $\mathscr{L}(X;\theta)$ est la fonction de vraisemblance. On en déduit que la diagonale de l'opposée (à cause du signe -) de la hessienne de la fonction de vraisemblance donne le vecteur des variances des paramètres estimés dans le modèle (la racine carrée de cette même diagonale donnera accès aux écarts-types). Cela nous permettra d'évaluer l'erreur commise lors de l'estimation et donc de déterminer les intervalles de confiance de chacun des paramètres.

3.1.4 Notions d'algèbre utiles pour la suite de ce rapport

Matrice de corrélation La matrice de corrélation d'un vecteur de p variables aléatoires $\overrightarrow{X}$, tel que :

$$\overrightarrow{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{pmatrix}$$

et dont chacune des variables possède une variance finie, est la matrice carrée dont le terme générique est donné par $r_{i,j} = \text{Cor}(X_i, X_j)$, avec $\text{Cor}(X, Y)$ le coefficient de corrélation des deux variables X et Y, tel que :

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\,\sigma_Y}$$

$\text{Cov}(X, Y)$ est la covariance des variables X et Y, σ_X et σ_Y leurs écarts-types respectifs.
Les termes diagonaux de la matrice de corrélation sont tous égaux à 1, elle est symétrique, semi-définie positive et ses valeurs propres sont positives ou nulles.
Cette matrice, calculée à partir du vecteur des paramètres d'un modèle, indique le niveau de corrélation (et donc d'inter-dépendance) entre chaque paramètre. Cela permet de vérifier, dans le cas où certains paramètres sont fortement corrélés, que le modèle n'est pas sur-paramétré. Il faut néanmoins vérifier le caractère identifiable de chaque paramètre pour confirmer ces conclusions.

Identifiabilité des paramètres du modèle On peut montrer que le nombre de **paramètres identifiables** d'un modèle (paramètres séparément estimables) est égal au **rang du Hessien** (matrice des dérivées secondes de la vraisemblance par rapport aux paramètres, au point de vraisemblance maximale) (Viallefont *et al.*, 1998).

[2] Il y a quatre hypothèses à remplir pour qu'un modèle puisse être considéré comme régulier, mais nous ne les mentionnerons pas. On acceptera donc que le modèle est régulier, et la démonstration peut se faire facilement en vérifiant les quatre hypothèses que l'on trouvera dans tout bon cours de statistiques.

Le **rang** d'une matrice est égal au nombre maximal de ses vecteurs lignes (ou colonnes) linéairement indépendants. Cela équivaut aussi à déterminer la dimension du sous-espace vectoriel engendré par ses vecteurs lignes (ou colonnes).
Mathématiquement, une matrice réelle M de taille $n \times n$ est inversible si et seulement si elle est de rang n. On pourrait donc considérer que si la hessienne est inversible, elle est automatiquement de rang égal au nombre de paramètres optimisés, et donc tous les paramètres du modèle seraient identifiables. Cependant, du point de vue statistique qui est le nôtre, on peut avoir des corrélations très fortes entre vecteurs lignes (colonnes) de la hessienne sans qu'il y ait d'exacte proportionnalité entre eux. Dans le code, la matrice hessienne peut donc être inversible même si certains des paramètres sont fortement corrélés et donc non identifiables. Il faut donc utiliser un critère empirique, tel que celui proposé par Viallefont et al.
La vérification de ce critère suppose au préalable la détermination des **valeurs singulières** et des **valeurs propres** de la matrice hessienne.
Le procédé de **décomposition en valeurs singulières** d'une matrice (ou **SVD**, de l'anglais *Singular Value Decomposition*) peut être considéré comme une généralisation du théorème spectral à des matrices arbitraires, qui ne sont pas nécessairement carrées (même si dans notre cas nous l'utilisons pour la hessienne, qui est une matrice carrée). Cette méthode énonce que si M est une matrice $m \times n$ à coefficients réels ou complexes, alors il existe une factorisation de la forme :

$$M = U \, \Sigma \, V^*$$

avec U une matrice unitaire $m \times m$ sur $\mathbb{R}$ (ou sur $\mathbb{C}$), Σ une matrice $m \times n$ dont les coefficients diagonaux sont des réels positifs ou nuls et tous les autres sont nuls, et V^* est la matrice adjointe à V, matrice unitaire $n \times n$ sur $\mathbb{R}$ (ou sur $\mathbb{C}$). Cette factorisation est la décomposition en valeurs singulières de M. On range souvent, par convention, les valeurs $\Sigma_{i,i}$ par ordre décroissant.
Une fois la décomposition en valeurs singulières effectuée sur la matrice hessienne, Viallefont et al. proposent comme critère pour la nullité des valeurs singulières un seuil s, calculé comme suit :

$$s = q \, \lambda_1 \, \epsilon$$

où q est la dimension de la matrice hessienne, λ_1 est la valeur propre la plus grande, et ϵ est une valeur arbitraire, que les auteurs prennent égale à $1,0.10^{-9}$.
La hessienne étant une matrice semi-définie positive[3] , sa décomposition en valeurs singulières est équivalente à sa décomposition spectrale (ses valeurs singulières sont identiques à ses valeurs propres). Ainsi, λ_1 correspond à la plus grande valeur singulière de la hessienne.
Il ne reste plus qu'à déterminer quelles sont les valeurs singulières qui dépassent le seuil s, et qui sont alors considérées comme non-nulles. Le nombre de ces valeurs dépassant le seuil donne alors le rang de la hessienne, et par conséquent le nombre de paramètres identifiables du modèle.

3.2 Modélisation du processus de diffusion

Initialement, deux approches avaient été retenues pour modéliser le processus de diffusion dans le modèle TABASCO. Le choix a été fait de n'en conserver qu'une, que nous présentons ci-après. Cette décision repose sur des considérations de temps de développement et d'outils informatiques à mettre en place pour les calculs. La solution retenue consiste à étudier la résultante de la propagation d'une distribution gaussienne dans le réseau hydrographique.

[3] Une matrice est semi-définie positive si toutes ses valeurs propres sont positives ou nulles.

3.2.1 Principe général et mise en équations

La première loi que nous utilisons est la loi de conservation locale des anguilles pour un problème unidimensionnel selon l'axe (Ox) :

$$\frac{\partial n}{\partial t} = -\frac{\partial j}{\partial x} \tag{3.6}$$

Où $n(x,t)$ est la concentration d'anguilles, t le temps (qui est lié, mais de façon non triviale, à l'âge des anguilles durant le processus de colonisation du milieu), $\vec{j}(x,t)$ le vecteur densité de courant d'anguilles (c'est-à-dire le produit de la densité d'anguilles par leur vitesse d'ensemble) et x la coordonnée représentative du problème (position dans le bassin versant).
Pour démontrer cette loi de conservation, considérons un tube dans lequel évoluent les anguilles, qui s'appuie sur un cercle d'axe (Ox) et d'aire S. Le tube est un cylindre de génératrices parallèles à (Ox) (figure 3.2).

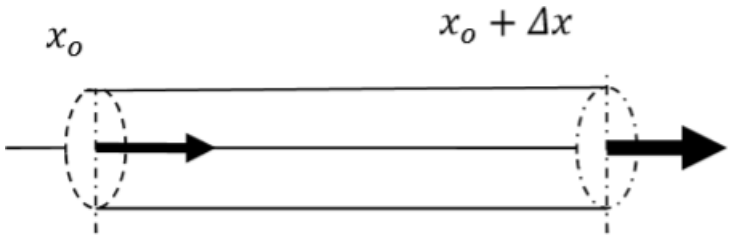

FIGURE 3.2 : Schéma de principe pour la démonstration de la loi de conservation des anguilles selon l'axe (Ox).

Les lignes de courant sont des droites, et le flux d'anguilles au travers d'une section d'abscisse x est :

$$\phi(x,t) \equiv \oiint \vec{j}(x,t).d\vec{S}$$

Soit :

$$\phi(x,t) = S\, j(x,t)$$

Faisons un bilan de matière pour les anguilles qui diffusent entre les instants t_0 et $t_0 + \mathrm{d}t$, pour la tranche comprise entre les abscisses x_0 et $x_0 + \Delta x$:

$$N(t_0 + \mathrm{d}t) - N(t_0) = \phi(x_0, t_0)\,\mathrm{d}t - \phi(x_0 + \Delta x, t_0)\,\mathrm{d}t$$

Où N(t) est le nombre d'anguilles à l'instant t entre les deux sections, soit donc :

$$\frac{N(t_0 + \mathrm{d}t) - N(t_0)}{\mathrm{d}t} = \phi(x_0, t_0) - \phi(x_0 + \Delta x, t_0) = S[j(x_0, t_0) - j(x_0 + \Delta x, t_0)]$$

Soit $n(x,t)$ la densité volumique d'anguilles qui diffusent dans le réseau.

$$N(t_0) = \int_{x_0}^{x_0+\Delta x} n(x, t_0) S\, dx$$

$$\frac{N(t_0 + \mathrm{d}t) - N(t_0)}{\mathrm{d}t} = S \int_{x_0}^{x_0+\Delta x} \frac{n(x, t_0 + \mathrm{d}t) - n(x, t_0)}{\mathrm{d}t} = S \int_{x_0}^{x_0+\Delta x} \frac{\partial n}{\partial t}(x_0, t)\mathrm{d}x$$

$$\frac{S}{\Delta x} \int_{x_0}^{x_0+\Delta x} \frac{\partial n}{\partial t}(x_0, t)\mathrm{d}x = -S\frac{j(x_0 + \Delta x) - j(x_0)}{\Delta x}$$

Et en faisant tendre Δx vers 0, il vient :

$$\frac{\partial n}{\partial t}(x_0, t_0) - \frac{\partial j}{\partial x}(x_0, t_0) = 0$$

D'où l'équation de conservation 3.6.
La Loi de Fick à une dimension nous permet ensuite d'exprimer la densité de courant en fonction du gradient de concentration d'anguilles :

$$j = -D \frac{\partial n}{\partial x} \tag{3.7}$$

Où D est le coefficient de diffusion, qui s'exprime en $m^2.\, s^{-1}$ dans les unités SI, et habituellement en $km^2.\, années^{-1}$ dans le cas d'une étude sur la colonisation d'un bassin versant par des anguilles. Ce coefficient quantifie le taux de déplacement des anguilles le long d'un gradient de densité, à travers une surface perpendiculaire à la direction du mouvement (Ibbotson *et al.*, 2002).
En combinant les équations 3.6 et 3.7, on obtient l'équation de diffusion unidimensionnelle :

$$\begin{cases} \dfrac{\partial n}{\partial t} = -\dfrac{\partial j}{\partial x} \\ \\ j = -D \dfrac{\partial n}{\partial x} \end{cases} \Longrightarrow \frac{\partial n}{\partial t} = D \frac{\partial^2 n}{\partial x^2} \tag{3.8}$$

Dans notre problème de diffusion des anguilles dans un bassin versant, nous souhaitons résoudre l'équation 3.8 dans le cas d'un régime non-stationnaire (évolutif) et unidimensionnel (la donnée caractéristique du problème est la distance de pénétration des anguilles dans un bassin versant ramenée sur l'axe (Ox)).
Nous prenons pour l'instant comme hypothèse que le caractère thalassotoque de l'anguille place cette résolution dans le cas simple d'une seule source de diffusion. En effet, les civelles d'anguilles arrivant du milieu océanique, la diffusion dans chaque bassin versant se fait à partir de l'exutoire de ce bassin. Les civelles utilisent les courants de marée montante pour progresser en zone estuarienne. Il serait donc logique de considérer la limite de marée dynamique plutôt que la limite transversale de la mer. Toutefois, la localisation de cette limite de marée dynamique n'est pas une donnée systématiquement renseignée dans les bases de données des réseaux hydrographiques français et européens. Il est donc proposé d'utiliser la distance à la mer comme point de départ du processus de diffusion[4]. On négligera donc, dans la version actuelle du code, les sources secondaires de diffusion des anguilles induites par les obstacles (rétro-diffusion) ainsi que les alevinages. Par ailleurs, la diffusion dans un réseau hydrographique introduit une complexité supplémentaire liée à la topologie du graphe (en fait de l'arbre) que constitue ce réseau hydrographique (Webb et Padgham, 2009, 2013).
Pour résoudre l'équation 3.8, il faut imposer des conditions aux limites. On pose donc qu'à $t = 0$ (instant virtuel où les anguilles sont supposées pénétrer dans le bassin versant par l'exutoire) $n(x, 0) = N_0\, \delta(x)$ où δ désigne la distribution de Dirac et $n(x, t)$ la concentration linéique des anguilles. Autrement dit, on suppose que les N_0 anguilles (la totalité) sont présentes à l'exutoire du bassin versant au départ de la simulation.
On rappelle que la distribution de Dirac possède la propriété suivante :

$$\int_{-\infty}^{+\infty} \delta(x) \mathrm{d}x = 1$$

On retrouve bien un nombre d'anguilles N_0 à t=0 :

[4] La distance à la mer d'un tronçon est définie dans le modèle comme la distance entre l'exutoire du bassin versant et l'extrémité aval du tronçon.

$$\int_{-\infty}^{+\infty} n(x,0)\mathrm{d}x = \int_{-\infty}^{+\infty} N_0\delta(x)\mathrm{d}x = N_0$$

La solution générale de l'équation 3.8 s'écrit alors :

$$n(x,t) = \frac{N_0}{\sqrt{4\,\pi\,D\,t}}\,e^{-\dfrac{x^2}{4\,D\,t}} \tag{3.9}$$

Avec :

$$\int_{-\infty}^{+\infty} n(x,t)\mathrm{d}x = N_0$$

On peut donc aussi définir une densité de probabilité de présence[5], qui est alors :

$$p(x,t) = \frac{1}{N_0}n(x,t) = \frac{1}{\sqrt{4\,\pi\,D\,t}}\,e^{-\dfrac{x^2}{4\,D\,t}} = \frac{1}{\sqrt{2\,\pi}\,\sigma}\,e^{-\dfrac{x^2}{2\sigma^2}} \tag{3.10}$$

Il s'agit bien d'une densité de probabilité, car la fonction p vérifie les deux propriétés suivantes :

$$\begin{cases} p(x,t) \geq 0 \quad \forall x \in \mathbb{R} \\ \displaystyle\int_{-\infty}^{+\infty} p(x,t)\mathrm{d}x = 1 \end{cases}$$

On retrouve bien pour $p(x,t)$ l'expression d'une fonction gaussienne de moyenne $\mu = 0$ et de variance $\sigma^2 = 2\,D\,t$. On notera qu'en présence d'advection due à la nage orientée des anguilles (mouvement de masse), l'expression devient simplement :

$$p(x,t) = \frac{1}{\sqrt{4\,\pi\,D\,t}}\,e^{-\dfrac{(x-\lambda)^2}{4\,D\,t}} = \frac{1}{\sqrt{2\,\pi}\,\sigma}\,e^{-\dfrac{(x-\lambda)^2}{2\sigma^2}} \tag{3.11}$$

La moyenne de la gaussienne est alors égale à λ. Dans la suite du rapport, nous traiterons les calculs sans advection par souci de concision, mais ils sont évidemment facilement adaptables à une situation en présence d'advection. Le taux d'advection n'est d'ailleurs pas inclus dans les paramètres optimisés du modèle (il est considéré nul), mais pourrait l'être sans difficulté si les circonstances l'exigeaient.

3.2.2 Hypothèses retenues pour le paramétrage de la diffusion

Comme nous venons de le déterminer, la variance de la distribution gaussienne est liée au coefficient de diffusion par la relation $\sigma^2 = 2\,D\,t$. Cette variance est donc fonction du temps (ou de façon équivalente de l'âge des anguilles), et l'on conçoit aisément qu'elle est très faible lorsque la cohorte d'anguilles arrive au niveau d'un estuaire (elle est même nulle du point de vue du formalisme mathématique, puisque l'on part de l'hypothèse d'une distribution de Dirac à t=0, c'est-à-dire au point le plus aval d'un bassin versant, ce qui peut être vu comme la cas limite d'une distribution gaussienne de variance nulle) et qu'elle augmente progressivement lors de la colonisation d'un bassin, ce qui traduit l'étalement spatial de la population, jusqu'à une valeur limite qui

[5] Il s'agit là d'une distribution donnant la répartition des anguilles en fonction de la distance caractéristique, et non pas une probabilité de présence telle qu'utilisée dans les modèles de présence-absence.

a priori ne dépend que du bassin considéré. Or, on ne s'intéresse dans le modèle qu'à la distribution résultante de la colonisation. Il faut donc pouvoir estimer la variance de la distribution gaussienne au bout d'un temps inconnu (et qui n'est pas nécessairement identique d'un bassin à un autre) et à partir d'un coefficient de diffusion lui-aussi inconnu, pouvant varier en fonction de nombreux paramètres tels que la capacité d'accueil du milieu, la topologie, le débit, la distance à la mer, la largeur du cours d'eau, etc...
L'analyse de l'évolution du coefficient de diffusion en fonction des différents paramètres écologiques et topologiques constituerait une étude à part entière. Nous avons donc dû établir des hypothèses fortes pour le calcul de la variance de la distribution gaussienne, en essayant de privilégier la simplicité des calculs à partir de considérations basiques. Ainsi, nous avons testé les deux possibilités les plus simples, à savoir :

- une variance estimée constante sur tous les bassins versants intégrés dans la modélisation, ce qui suppose une variance identique en fin de colonisation quel que soit le bassin considéré ;

- une variance estimée constante sur un drain donné, et liée à la longueur totale du drain considéré (on travaille en linéaire de berges, donc sans prendre en compte les variations de la section des drains le long des cours d'eau).

Dans le premier cas, cela suppose que les anguilles se répartissent selon le même schéma spatial quel que soit le bassin versant qu'elles colonisent. Dans le second cas, où l'on ne résonne plus que sur un drain, on s'affranchit en principe des différences de longueur de drain en ayant une variance proportionnelle à la longueur de drain au carré, mais on néglige toujours les autres paramètres, notamment écologiques. Cela reste donc aussi une hypothèse forte.

3.2.3 Probabilités de sédentarisation

La population d'anguilles $N(i)$ dans le $i^{éme}$ tronçon du réseau est donnée par la proportion d'anguilles qui vont se sédentariser dans ce tronçon compris entre $D_{mer}(i)$ et $D_{mer}(i) + l(i)$ parmi toutes celles qui sont issues du tronçon précédent, où $D_{mer}(i)$ est la distance à la mer du $i^{éme}$ tronçon et $l(i)$ sa longueur. On a ainsi, pour une diffusion en milieu infini :

$$N(i) = \overline{N(i-1)} \frac{\int_{D_{mer}(i)}^{D_{mer}(i)+l(i)} p(x,i)\,\mathrm{d}x}{\int_{D_{mer}(i)}^{+\infty} p(x,i)\,\mathrm{d}x} \tag{3.12}$$

Afin d'implémenter ce calcul du nombre d'anguilles qui se sédentarisent dans le tronçon i, nous allons réécrire l'équation à l'aide de la fonction de distribution cumulative (fonction de répartition) de la distribution normale centrée réduite, définie par :

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt \tag{3.13}$$

Exprimons l'équation 3.12 en y mettant l'expression explicite de la distribution des anguilles :

$$
\begin{aligned}
N(i) &= \overline{N(i-1)}\ \frac{\displaystyle\int_{D_{mer}(i)}^{D_{mer}(i)+l(i)} \frac{1}{\sqrt{2\,\pi}\,\sigma}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x}{\displaystyle\int_{D_{mer}(i)}^{+\infty} \frac{1}{\sqrt{2\,\pi}\,\sigma}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x} \\
&= \overline{N(i-1)}\ \frac{\displaystyle\int_{-\infty}^{D_{mer}(i)+l(i)} \frac{1}{\sqrt{2\,\pi}\,\sigma}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x - \int_{-\infty}^{D_{mer}(i)} \frac{1}{\sqrt{2\,\pi}\,\sigma}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x}{\displaystyle\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\,\pi}\,\sigma}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x - \int_{-\infty}^{D_{mer}(i)} \frac{1}{\sqrt{2\,\pi}\,\sigma}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x} \\
&= \overline{N(i-1)}\ \frac{\displaystyle\frac{1}{\sigma}\left(\int_{-\infty}^{D_{mer}(i)+l(i)} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x - \int_{-\infty}^{D_{mer}(i)} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x\right)}{\displaystyle\frac{1}{\sigma}\left(\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x - \int_{-\infty}^{D_{mer}(i)} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x\right)}
\end{aligned}
$$

Et en posant $u = x/\sigma$ ($\mathrm{d}u = \frac{1}{\sigma}\,\mathrm{d}x$) , il vient :

$$
\begin{aligned}
N(i) &= \overline{N(i-1)}\ \frac{\displaystyle\int_{-\infty}^{(D_{mer}(i)+l(i))/\sigma} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{u^2}{2}}\,\mathrm{d}u - \int_{-\infty}^{(D_{mer}(i))/\sigma} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{u^2}{2}}\,\mathrm{d}u}{\displaystyle\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{u^2}{2}}\,\mathrm{d}u - \int_{-\infty}^{(D_{mer}(i))/\sigma} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{u^2}{2}}\,\mathrm{d}u} \\
&= \overline{N(i-1)}\ \frac{\Phi\big((D_{mer}(i)+l(i))/\sigma\big) - \Phi\big((D_{mer}(i))/\sigma\big)}{\Phi(+\infty) - \Phi\big((D_{mer}(i))/\sigma\big)}
\end{aligned}
$$

Or dans la réalité, un bassin versant et tous les drains qui le composent sont de dimension finie. Il faut donc remplacer le signe infini par la longueur totale du drain considéré, que l'on notera $L(i)$ (divisée par σ, car on travaille en coordonnées réduites). D'où finalement :

$$
N(i) = \overline{N(i-1)}\ \frac{\Phi\big((D_{mer}(i)+l(i))/\sigma\big) - \Phi\big((D_{mer}(i))/\sigma\big)}{\Phi\big((L(i))/\sigma\big) - \Phi\big((D_{mer}(i))/\sigma\big)} \tag{3.14}
$$

Cette dernière relation de récurrence peut être utilisée dans le code car la fonction de distribution cumulative de la distribution normale centrée réduite est implémentée dans la bibliothèque de calcul scientifique en C/C++ que nous utilisons (voir dans le chapitre suivant). Une autre possibilité est d'utiliser la fonction d'erreur de Gauss, implémentée dans le fichier d'en-tête <math.h> et définie par :

$$
\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp^{-t^2}\,\mathrm{d}t
$$

En effet, la fonction de distribution cumulative de la loi normale peut s'écrire :

$$
\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]
$$

On en déduit donc une autre expression de N(i) :

$$N(i) = \overline{N(i-1)}\,\frac{\operatorname{erf}\left(\frac{(D_{mer}(i)+l(i))}{\sqrt{2}\,\sigma}\right) - \operatorname{erf}\left(\frac{D_{mer}(i)}{\sqrt{2}\,\sigma}\right)}{1-\operatorname{erf}\left(\frac{D_{mer}(i)}{\sqrt{2}\,\sigma}\right)} \tag{3.15}$$

La méthode de calcul suppose la détermination de la proportion d'anguilles qui se retrouvent dans chaque tronçon du réseau hydrographique suite au processus diffusif. Soit donc P_i la proportion d'anguilles et p_i la probabilité conditionnelle de sédentarisation dans le $i^{ème}$ tronçon du réseau (c'est-à-dire la probabilité pour les anguilles de se trouver dans le $i^{ème}$ tronçon sachant qu'elles ne se sont pas réparties dans les tronçons avals). On suppose pour l'instant une branche unique du graphe orienté (on ignore les confluences et les obstacles pour le moment).
Les équations 3.14 et 3.15 s'écrivent aussi :

$$N_i = \overline{N_{i-1}}\,P_i \tag{3.16}$$

3.2.4 Prise en compte des obstacles

Par construction du modèle, les obstacles sont nécessairement situés à une extrémité de tronçon (par convention au nœud amont), et sont pris en compte de la façon suivante : à chaque obstacle indexé par j est associé un indice de franchissement $\chi(j)$.
Si des obstacle existent entre le tronçon i et le tronçon i+1, alors la probabilité conditionnelle de sédentarisation des anguilles dans le $i{+}1^{ème}$ tronçon sachant qu'elle ne se sont pas sédentarisées dans les tronçons avals est (et toujours en ne considérant qu'un drain unique, sans embranchements) :

$$\begin{aligned} p_{i+1} &= \chi(j)\,\chi(j-1)\,\ldots\,\chi(1)\,P_{i+1}(1-P_i)(1-P_{i-1})...(1-P_0) \\ &= P_{i+1}\prod_{m=1}^{j}\chi(m)\prod_{k=0}^{i}(1-P_k) \end{aligned} \tag{3.17}$$

Dans la version actuelle du modèle, l'indice de franchissement est considéré comme étant identique pour l'ensemble des obstacles : $\chi(j) = \chi(j-1) = \ldots = \chi(1) = \chi$. Il fait néanmoins partie des paramètres optimisés dans le modèle. L'équation précédente devient donc simplement :

$$p_{i+1} = \chi^N\,P_{i+1}(1-P_i)(1-P_{i-1})...(1-P_0) = \chi^N\,P_{i+1}\prod_{k=0}^{i}(1-P_k) \tag{3.18}$$

Où N désigne le nombre d'obstacles présents entre le tronçon 0 et le tronçon i+1.

3.2.5 Prise en compte des confluences

Dans le cas d'une confluence à l'amont (embranchement du réseau), la probabilité de choisir le tronçon amont j est égale à la proportion relative de la surface du bassin versant j par rapport à la somme des surfaces des bassins amonts élevées à une certaine puissance δ :

$$\frac{(S_{BV}(j))^{\delta}}{\sum\limits_{k\in\text{amont}(i)}(S_{BV}(k))^{\delta}}$$

3.2.6 Spécificités de l'ajustement du modèle dans l'approche par propagation d'une gaussienne

L'ajustement du modèle se fait grâce à un algorithme de différentiation automatique (en anglais AD, pour *Automatic Differentiation*). Un tel algorithme permet d'évaluer les dérivées d'une fonction d'intérêt pour un programme, dite fonction-objectif (par exemple une fonction à minimiser, ou dans notre cas, une vraisemblance à maximiser). Pour un développement détaillé des principes et des outils utilisés dans le cadre de la différentiation automatique, on pourra consulter (Rall, 1981).
Nous avons choisi d'utiliser l'outil ADMB (*Automatic Differentiation Model Builder*) (Fournier *et al.*, 2012), utilisable en C++ sous licence BSD (la licence BSD (*Berkeley Software Distribution License*) est une licence libre utilisée pour la distribution de logiciels. Elle permet de réutiliser tout ou une partie du logiciel sans restriction, qu'il soit intégré dans un logiciel libre ou propriétaire.).

3.2.7 Ajustement du modèle

Nous décrivons ici le principe de l'ajustement du modèle TABASCO. Dans un réseau dont la topologie et les caractéristiques des tronçons sont fixées, et connaissant l'effectif total N_{tot}, les composantes diffusives $\sigma(i)$ (qu'elles soient constantes ou qu'elles soient calculées à partir des caractéristiques du tronçon), les probabilités de se sédentariser ou de se déplacer p_i et P_{AM} (calculées au moins à partir de la longueur du tronçon), le coefficient aux confluences β, les probabilités de franchissement vers l'amont $\phi_{AM}(i)$[6] des différents obstacles (constants ou calculés à partir de l'indice de franchissabilité de Steinbach (Steinbach, 2006)), il est possible de calculer de manière récursive tous les effectifs.
Inversement, si l'on dispose d'un jeu d'observation de densités d'anguilles dans une sélection de tronçons, il est possible d'ajuster les paramètres N_{tot}, les $\sigma(i)$ ou p_i et P_{AM}, β, les $\phi_{AM}(i)$ et les $\tau_{AM}(i)$. En particulier, il est intéressant de noter que le paramètre d'intérêt pour la gestion, N_{tot}, est directement exprimé dans les équations, ce qui permet en principe d'en donner un intervalle de confiance. En effet, nous utilisons la méthode du maximum de vraisemblance. Or, l'estimateur du maximum de vraisemblance est asymptotiquement normal[7], on peut donc construire un intervalle de confiance C_n tel qu'il contienne le vrai paramètre avec une probabilité $(1-\alpha)$ (Wasserman, 2004) :

$$C_n = \left(\hat{\theta}_n - \Phi^{-1}(1-\alpha/2)\widehat{\sigma_{\hat{\theta}_n}},\ \hat{\theta}_n + \Phi^{-1}(1-\alpha/2)\widehat{\sigma_{\hat{\theta}_n}}\right)$$

Avec $\Phi^{-1}(1-\alpha/2)$ le quantile d'ordre $(1-\alpha/2)$ de la loi normale centrée réduite et $\widehat{\sigma_{\hat{\theta}_n}}$ l'écart-type estimé de $\hat{\theta}_n$. On a alors :

$$\mathbb{P}(\theta \in C_n) \xrightarrow[n\to+\infty]{} 1-\alpha$$

D'une manière générale, soit $N(i,\theta)$ où θ est le jeu de paramètres qui permet de calculer le nombre d'individus dans le tronçon i. On appellera $\Delta(i,\theta)$ la densité dans le tronçon i, telle que :

$$\Delta(i,\theta) = \frac{N(i,\theta)}{S(i)}$$

Où $S(i)$ est la surface en eau du tronçon i.

[6] Pour le moment, nous considérons dans le modèle que tous les obstacles ont le même indice de franchissement. A terme, cependant, il est prévu d'améliorer cet aspect en tenant compte de franchissabilités variables en fonction de l'ouvrage.

[7] Un estimateur $\hat{\theta}_n$ d'un paramètre d'intérêt θ avec n observations est dit asymptotiquement normal si et seulement si : $n^{p/2}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\Gamma)$, où p est la vitesse de convergence (convergence dite en $1/n^p$) et où Γ est la matrice de covariance (asymptotique). La convergence est une convergence en loi, et on rappelle qu'une suite $(X_n)_{n>=1}$ converge en loi vers X si, pour toute fonction φ continue bornée sur un espace métrique E, à valeurs dans $\mathbb{R}$, $\lim_n \mathbb{E}[\varphi(X_n)] = \mathbb{E}[\varphi(X)]$.

3.2.8 Fonction de vraisemblance

Comme nous l'avons déjà mentionné précédemment, l'ajustement se fait par maximum de vraisemblance. La fonction de vraisemblance pour une observation o s'écrit en considérant soit qu'aucune anguille n'est capturée lors de l'opération (densité observée nulle), soit qu'une densité d'anguilles non nulle est observée. Dans ce deuxième cas, nous postulons que la densité observée suit une loi log-normale (loi de Galton) dont l'espérance correspond à la densité calculée dans le tronçon i(o) où a eu lieu l'observation. Ainsi :

$$\ln(D_{obs}) \sim \mathcal{N}(\mu, \sigma)$$

Avec donc μ l'espérance de $\ln(D_{obs})$ et σ l'écart-type de $\ln(D_{obs})$.
Or, d'après la correction de Laurent (Laurent, 1963) :

$$D(i) = E(D_{obs}) = \mathrm{e}^{(\mu+\sigma^2/2)}$$

D'où :

$$\mu = \ln(D(i)) - \sigma^2/2$$

La fonction de vraisemblance du logarithme népérien de la densité d'anguilles observée a donc pour expression :

$$\frac{1}{\sqrt{2\,\pi}\,\sigma\,D_{obs}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{\ln(D_{obs}-\ln(D(i))+\sigma^2/2}{\sigma}\right)^2}$$

Néanmoins, cette fonction de vraisemblance n'est pas définie lorsque la densité observée est nulle. Afin d'étendre le domaine de définition de la vraisemblance en $D_{obs} = 0$, nous considérons la fonction q, qui correspond à la probabilité de capturer moins d'un individu lors d'une pêche. Nous explicitons le calcul de cette probabilité au paragraphe suivant. La fonction de vraisemblance pour une observation o quelconque s'écrit donc :

$$L(o,\theta) = \begin{cases} q\left(D\left(i(o)\right),\theta\right) & \text{si } D_{obs}(o) = 0 \\ \dfrac{1}{\sqrt{2\,\pi}\,\sigma\,D_{obs}(o)}\,\mathrm{e}^{-\frac{\left(\ln\left(D_{obs}(o)\right)-\ln\left(D\left(i(o)\right)\right)+\sigma^2/2\right)^2}{2\sigma^2}} & \text{si } D_{obs}(o) > 0 \end{cases}$$

La fonction de vraisemblance pour l'ensemble des observations s'écrit alors :

$$L(\theta_D,\theta_{LN},\sigma) = \prod_{o\in(D_{obs}(o)=0)} q\left(D\left(i(o),\theta\right),\theta\right)$$
$$\prod_{o\in(D_{obs}(o)>0)}\left(\frac{1}{\sqrt{2\,\pi}\,\sigma\,D_{obs}(o)}\,\mathrm{e}^{-\frac{\left(\ln\left(D_{obs}(o)\right)-\ln\left(D\left(i(o)\right)\right)+\sigma^2/2\right)^2}{2\sigma^2}}\right)$$

Le modèle est ensuite ajusté sur les observations en minimisant l'opposé du logarithme de la fonction de vraisemblance.

3.2.9 Probabilité de non-capture

Puisque nous avons postulé que la densité observée suivait une loi log-normale, nous pouvons en déduire la distribution de la capture, et en particulier, la probabilité $q(i)$ de capturer moins d'un individu dans un secteur de pêche électrique i donné. Cela correspond à la situation où $D_{obs} = 0$.
On a donc $\ln(D_{obs}) \sim \mathcal{N}(\mu, \sigma)$ d'une part, et $D_{obs}(i) = \frac{nb\_captures(i)}{\Sigma(i)}$ d'autre part, où

$\Sigma(i)$ est la surface prospectée dans le i^{ème} tronçon. Comme $\Sigma(i)$ est un paramètre indépendant défini lors des pêches, il n'influe pas sur la distribution des autres variables aléatoires. Par conséquent, si $\ln(\frac{nb\_captures(i)}{\Sigma(i)}) \sim \mathcal{N}(\mu, \sigma)$, alors $\ln(nb\_captures(i)) \sim \mathcal{N}(\mu, \sigma)$. Les paramètres μ et σ ne changent donc pas, que l'on considère la densité observée ou bien le nombre d'anguilles capturées. L'espérance du nombre d'anguilles pêchées étant le produit de la densité prédite (calculée) $\Delta(i)$ par la surface prospectée $\Sigma(i)$ dans le tronçon i, on peut écrire, en vertu des propriétés de la loi log-normale, que :

$$\mu = \ln(\Delta(i)\,\Sigma(i)) - \frac{\sigma^2}{2} = \ln(\Delta(i)) + \ln(\Sigma(i)) - \frac{\sigma^2}{2}$$

Or de plus, si une variable aléatoire suit une loi log-normale de paramètres μ et σ, sa fonction de répartition Ψ est liée à la fonction de répartition de la loi normale Φ par la formule :

$$\Psi(x) = \Phi\left(\frac{\ln(x) - \mu}{\sigma}\right)$$

On a donc :

$$q(i) = \Phi\left(\frac{\ln(x) - \ln(\Delta(i)) - \ln(\Sigma(i)) + \frac{\sigma^2}{2}}{\sigma}\right)$$

Comme on cherche la probabilité d'absence (de non-capture), c'est-à-dire la probabilité que le nombre d'anguilles pêchées soit inférieur à 1, le calcul final sera :

$$q(i) = \Phi\left(\frac{\ln(1) - \ln(\Delta(i)) - \ln(\Sigma(i)) + \frac{\sigma^2}{2}}{\sigma}\right) = \Phi\left(\frac{-\ln(\Delta(i)) - \ln(\Sigma(i)) + \frac{\sigma^2}{2}}{\sigma}\right)$$

3.2.10 Paramétrage du modèle

Quelle que soit l'expression retenue pour le calcul de la variance de la distribution des anguilles dans les bassins, un total de 31 paramètres sont optimisés dans le modèle. Il se répartissent comme suit :

- 20 paramètres correspondant aux densités surfaciques moyennes en France métropolitaine pour chacune des 20 années prises en compte pour la calibration (variation annuelle des stocks),
- 7 paramètres correspondant aux sept unités de gestion anguille (UGA) en plus de l'unité de gestion que nous prenons comme référence, à savoir l'UGA "Loire, côtiers vendéens et Sèvre niortaise" (variation géographique des stocks),
- 1 paramètre de franchissabilité moyenne des obstacles,
- 1 paramètre de variance de la distribution log-normale des densités observées dans un tronçon,
- 1 paramètre de pondération des flux aux confluences,
- 1 paramètre intervenant dans le calcul de la variance de la distribution gaussienne des anguilles dans les bassins.

Le tableau 3.1 synthétise tous les paramètres utilisés dans le modèle TABASCO (qu'ils soient optimisés ou non).

Définition du paramètre	**Dénomination dans le modèle**	**Symbole usuel**	**Remarques**
Logarithme de la densité surfacique moyenne dans le bassin pour l'année xxxx	logMeanDensityxxxx	$\log(D_0)$	optimisé dans le modèle pour chacune des 20 années de données
Taux d'advection	lambdaAdvection	λ	en km, non optimisé (pris égal à 0)
Variation géographique du stock d'anguilles en fonction de l'unité de gestion xxxx	EMUeffect_FR_xxxx	-	optimisé dans le modèle pour les 7 unités (en plus de l'unité de référence)
écart-type de la distribution des anguilles dans un drain	SigmaDiffusive	σ	optimisé dans le modèle
Logit de la franchissabilité des obstacles	logitPassabilities0	$\text{logit}(\chi)$	optimisé dans le modèle
Logarithme de la variance de la distribution des anguilles capturées dans un tronçon	logsigma	$\text{Log}(\sigma)$	optimisé dans le modèle
Paramètre de pondération des flux aux confluences	d	-	optimisé dans le modèle
Coefficient de diffusion	-	D	en $\text{km}^2.\text{années}^{-1}$
Distance du tronçon à sa source	Dsource	D_{source}	en km
Distance du tronçon à la mer	Doutlet	D_{mer}	en km
Longueur du tronçon	length	l	en km
Longueur totale du drain	-	L_{tot}	en km

TABLE 3.1 : Tableau récapitulatif des paramètres utilisés dans le modèle TABASCO.

3.2.11 Calibration du modèle

Nous disposons des données de pêches électriques complètes réalisées entre les années 1990 et 2009 incluses (soit 20 ans de données). Dans le détail, l'extraction de la BD-map dont nous disposons regroupe un total de 19201 opérations de pêches électriques effectuées entre 1966 et 2009, dont 9837 pêches complètes. Ce total se réduit à 16510 opérations pour la période que nous avons retenue, à savoir entre 1990 et 2009. Sur

cette période, seules 8819 opérations sont des pêches complètes. C'est ce dernier échantillon de données sur lequel nous basons la calibration de TABASCO. Il faut garder à l'esprit que ce nombre de données utilisables est très faible. Si l'on considère par exemple le bassin de la Gironde (Garonne-Dordogne), qui est pourtant l'un de ceux pour lesquels les pêches sont les plus nombreuses, on ne dispose que de 1293 mesures sur 20 ans, ce qui fait une moyenne de 65 données par an seulement (présence et absence confondues).
Nous calibrons le modèle sur les bassins versants dans lesquels ont été menées des opérations de pêches électriques, soit une grosse centaine parmi l'ensemble des 953 bassins versants pris en compte dans le modèle. Le nombre total de bassins (953) correspond à la totalité des bassins versants dont les exutoires sont situés en France métropolitaine (le long des côtes de la Manche, en Bretagne, sur tout l'arc Atlantique et sur le littoral méditerranéen).
Nous disposons d'indicateurs numériques pour juger de l'efficacité de l'optimisation et de la rapidité de la convergence : la valeur de la fonction objectif lors de la minimisation (qui est l'opposée de la fonction de vraisemblance), la valeur du gradient de chaque paramètre fournie par l'algorithme d'optimisation, le nombre d'itérations de l'algorithme, et le nombre d'évaluations de la fonction-objectif. Une valeur trop élevée du gradient indique par exemple que l'algorithme ne parvient pas à trouver un minimum local pour le paramètre considéré et ne converge donc pas. La meilleure estimation des paramètres se produit donc lorsque la fonction objectif est minimale et que les coefficients de la matrice des dérivées des paramètres sont minimaux en valeur absolue.
Il est aussi important de rappeler que comme tous les algorithmes qui utilisent la méthode de Quasi-Newton, ADMB ne trouve qu'un minimum local, qui n'est pas nécessairement le minimum absolu de la fonction objectif (mais qui peut l'être). Ainsi, il est nécessaire de vérifier la sensibilité du point de départ de l'optimisation afin de garantir que l'ensemble des paramètres optimisés soit effectivement le meilleur possible.

3.2.12 Approche de calcul alternative : méthode matricielle

Une approche de modélisation parallèle a été étudiée en 2013 et 2014, mais il a été choisi de ne pas poursuivre davantage son développement, principalement par manque de temps. Néanmoins, nous exposerons ici le travail qui a été effectué sur cette méthode alternative et qui constitue une possible piste de poursuite du modèle à l'avenir. Cette approche traite le phénomène de diffusion en utilisant des matrices de transition (ou de transfert) pour modéliser la propagation des effectifs d'anguilles d'un tronçon à l'autre. Cela suppose de modéliser la probabilité pour une anguille de passer d'un tronçon à un tronçon adjacent au cours d'un pas de temps unitaire. La probabilité est une fonction des caractéristiques du tronçon (surface du tronçon, surface du bassin-versant amont, longueur du tronçon, etc...) et des nœuds entre les tronçons (barrage ou confluence). Les probabilités de mouvement vers un tronçon adjacent pour un pas de temps unitaire sont alors réunies dans une matrice, dite matrice de transition.

Principe général et concept

Il s'agit en fait d'une amélioration du modèle OMMER (Obstacle Mitigation Model for Eel in Rivers) (Lambert *et al.*, 2011).
Initialement, l'idée d'utiliser des matrices de transition dans la dynamique de population revient à P. H. Leslie en 1948 (Leslie, 1948). L'approche retenue était un modèle déterministe basé sur les distributions en âge des populations, qui prédisait la distribution résultante par classe d'âge à intervalles successifs. En 1965, L. P. Lefkovitch proposa une alternative en groupant les populations non plus par classe d'âge mais par stades du développement (Lefkovitch, 1965). Cela évitait notamment d'avoir besoin de connaître l'âge des individus et autorisait des différences dans les taux de développement puisqu'il n'y a pas de relation invariante entre taille et âge. La méthode matricielle a ensuite été appliquée à de nombreux objets d'étude en écologie, jusqu'à

la dynamique de population des arbres.
De façon plus formelle, le concept de base est celui de matrice de transition[8]. Il s'agit d'une matrice utilisée pour décrire les transitions d'une chaîne de Markov[9]. Une telle matrice est carrée et chacun de ses coefficients est un nombre réel compris entre 0 et 1, et représentant une probabilité.
Reprenons le cas des anguilles qui colonisent un bassin. Soit $P_{i,j}$ la probabilité pour les anguilles de se déplacer d'un tronçon i vers un tronçon quelconque j du réseau au cours d'un pas de temps. La matrice de transition T est définie par la donnée des $i \times j$ coefficients tels que $T_{i,j}$ est le coefficient de la i^{e} ligne et de la j^{e} colonne de T. Soit :

$$T = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,j} & \cdots \\ p_{2,1} & p_{2,2} & \cdots & p_{2,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ p_{i,1} & p_{i,2} & \cdots & p_{i,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \end{pmatrix} \tag{3.19}$$

Étant donné que les anguilles ne peuvent se déplacer que dans un tronçon adjacent à celui dans lequel elles se trouvent à un instant donné, les coefficients $T_{i,j}$ sont nuls pour tous les tronçons j non adjacents au tronçon i. C'est la raison pour laquelle les matrices que nous utilisons dans le modèle de diffusion par matrices sont dites "creuses" : elles contiennent en effet beaucoup de zéros.
Puisque la probabilité de déplacement des anguilles d'un tronçon i vers un autre tronçon adjacent (incluant aussi la probabilité qu'elles restent dans le même tronçon au cours du pas de temps) vaut 1, on a alors :

$$\sum_{j} T_{i,j} = 1 \quad \forall\, j \text{ adjacent à i} \tag{3.20}$$

Formellement, on dit que la matrice est une matrice stochastique droite.

Probabilités de déplacement et sédentarisation

Le bassin versant est représenté sous forme d'un vecteur regroupant les effectifs par tronçon sur les M tronçons :
$N(t) = (N(t,1), N(t,2), \ldots, N(t,M)$
Les individus rentrant initialement à l'aval, on a :
$N(0) = (N_{tot},\ 0\ , \ldots,\ 0)$
On note T la matrice de transition décrivant les probabilités p$_{i,j}$ de passer du tronçon de départ i à un tronçon d'arrivée j au cours d'un pas de temps unitaire. Cette matrice est creuse (remplie de zéro) puisque les individus ne peuvent se déplacer qu'entre tronçons adjacents.
Le problème revient donc essentiellement à déterminer le modèle d'état. Cela consiste à calculer la matrice T en déterminant les probabilités de passer d'un tronçon à un autre au cours d'un pas de temps unitaire. Les probabilités de passage sont décomposées en étapes successives :

- une probabilité de rester dans le tronçon ou d'essayer de bouger,
- une probabilité de bouger vers l'aval ou vers l'amont sachant que l'anguille change de tronçon,
- une probabilité de franchir le nœud dans la direction choisie,

[8] On trouve aussi d'autres dénominations : matrice de probabilité, matrice de substitution, matrice stochastique, ou encore matrice de Markov

[9] Une chaîne de Markov désigne de façon générale un processus de Markov à temps discret. En mathématiques, un processus de Markov est un processus stochastique possédant la propriété de Markov, c'est-à-dire que toute l'information utile pour la prédiction du futur est contenue dans l'état présent du processus.

- une probabilité de choisir un des deux tronçons possibles si j'ai franchi le nœud amont.

Ainsi, la probabilité de ne pas essayer de quitter le tronçon i est :

$$p_i = \frac{1}{1 + \exp^{-(\text{un terme de position relative dans le réseau})}}$$

La probabilité de choisir de partir vers l'amont est P_{AM}, celle de partir vers l'aval $(1 - P_{AM})$.
La probabilité de franchir un nœud vers l'amont est :

$$P_{n,AM}(i) = \begin{cases} 1 \text{ si confluence} \\ \phi_{n,AM}(i) \text{ si barrage} \end{cases}$$

La probabilité de franchir un nœud vers l'aval est :

$$P_{n,AV}(i) = 1$$

Dans le cas d'une confluence à l'amont (embranchement du réseau), la probabilité de choisir le tronçon amont j est égale à la proportion relative des surfaces de bassins versants amonts élevées à la puissance β :

$$\frac{(S_{BV}(j))^{\beta}}{\sum\limits_{k \in \text{amont}(i)} (S_{BV}(k))^{\beta}}$$

En résumé, les probabilités $P_{i \to j}$ sont alors :

$$P_{i \to j} = \begin{cases} 0 \text{ si } (j \notin \text{amont}(i) \mid j \notin \text{aval}(i)) \\ p_i + (1 - p_i)[P_{AM}\,(1 - P_{n,AM}(i)) + (1 - P_{AM})] \text{ si } j = i \\ (1 - p_i)(1 - P_{AM}) \text{ si } j \in \text{aval}(i) \\ (1 - p_i) P_{AM} P_{n,AM}(i) \dfrac{(S_{BV}(j))^{\beta}}{\sum\limits_{k \in \text{amont}(i)} (S_{BV}(k))^{\beta}} \text{ si } j \in \text{amont}(i) \end{cases}$$

Spécificités de l'ajustement du modèle pour l'approche par matrices de transition

L'ajustement du modèle pour l'approche par matrices de transition se fait grâce à l'algorithme d'optimisation numérique BOBYQA (Bound Optimization BY Quadratic Approximation) développé par Michael J. D. Powell pour résoudre des problèmes d'optimisation avec contraintes aux limites sans utiliser les dérivées de la fonction-objectif (fonction à minimiser), ce qui le classe dans la catégorie des algorithmes sans dérivées (Powell, 2009). C'est aussi un algorithme itératif à région de confiance, c'est-à-dire qu'il minimise une fonction en procédant par améliorations successives. Au point courant, BOBYQA effectue un déplacement obtenu en minimisant un modèle simple de la fonction (en l'occurrence un modèle quadratique généré par interpolation) sur une région de confiance. Le rayon de confiance (caractérisant la région du même nom) est ajusté de manière à faire décroître suffisamment la fonction à chaque itération.

Chapitre 4

Outils informatiques

4.1 Langage et normes de programmation

Le modèle TABASCO a été intégralement codé en C++. Le C++ est un langage de programmation compilé, libre d'utilisation, permettant la programmation sous de multiples paradigmes comme la programmation procédurale, la programmation orientée objet et la programmation générique. Ses principaux atouts sont ses fonctionnalités multiples (il dérive d'une amélioration du langage C), et sa portabilité, puisqu'il est compatible avec une grande variété de plateformes matérielles et de systèmes d'exploitation.

4.2 Logiciels et interfaces

4.2.1 Environnement de développement

L'environnement de développement (dont nous utiliserons par la suite l'acronyme en anglais : IDE, pour *Integrated Development Environment*) choisi est le logiciel gratuit, open-source et multi-plateformes Code::Blocks, compatible avec les langages de programmation C, C++, et Fortran.

4.2.2 Base de données

La gestion de la base de données est effectuée grâce à l'outil PostgreSQL. Il s'agit d'un système de gestion de base de données relationnelle et objet (SGBDRO). C'est un outil libre disponible selon les termes d'une licence de type BSD (voir la définition des licences BSD dans la section traitant de l'ajustement).

Interface avec l'utilisateur

Nous utilisons également deux interfaces utilisateur, détaillées ci-dessous.

- psql, qui est une interface en ligne de commande permettant la saisie de requêtes SQL[1], directement ou par l'utilisation de procédures stockées.

- pgAdmin, qui est un outil d'administration graphique pour PostgreSQL, distribué selon les termes de la licence PostgreSQL.

[1]SQL (Structured Query Language, ou langage de requête structurée en français) est un langage informatique normalisé servant à exploiter des bases de données relationnelles. Créé en 1974, et normalisé depuis 1986, ce langage est reconnu par la grande majorité des systèmes de gestion de bases de données relationnelles du marché.

Interface avec le SIG

Le module spatial PostGIS nous permet de travailler en lien avec un système d'information géographique (SIG) sur des données spatialisées ou géoréférencées. C'est ce qui confère à PostgreSQL le statut de SGDBRO spatial.
La version de PostgreSQL que nous utilisons est la version 9.3, celle de PostGIS est la version 2.1.

Système d'Information Géographique

Le SIG que nous utilisons est le logiciel libre et multi-plateformes QGIS. Il gère, via la bibliothèque GDAL3, les formats d'images matricielles (*raster*) et vectorielles, ainsi que les bases de données. QGIS fait partie des projets de la Fondation Open Source Geospatial.
Ses caractéristiques principales sont :

1. La gestion de PostGIS, l'extension spatiale de PostgreSQL.
2. La prise en charge d'un grand nombre de formats de données vectorielles (Shapefile, les couvertures ArcInfo, Mapinfo, GRASS GIS, etc.)
3. La prise en charge d'un nombre important de formats de couches matricielles (GRASS GIS, GeoTIFF, TIFF, JPG, etc.)

4.2.3 Interface du langage avec la base de données

La bibliothèque libpqxx est l'outil nous permettant de faire le lien entre le langage C++ et le logiciel PostgreSQL. Il s'agit de l'interface de programmation (dont nous utiliserons par la suite l'acronyme anglais : API, pour *Application Programming Interface*) standard entre des programmes codés en C++ et des bases de données exploitées avec PostgreSQL. Le code source de libpqxx est lui-aussi disponible sous licence BSD. Un tutoriel détaillant la procédure - assez complexe - d'installation de libpqxx sous Windows est par ailleurs fourni en annexe du présent rapport. L'installation de cette bibliothèque sous Windows nous a en effet posé quelques difficultés lors de la préparation de nos machines en vue de la programmation du modèle TABASCO (notamment car cette bibliothèque est initialement prévue pour fonctionner sous systèmes UNIX).

4.2.4 Bibliothèques pour le calcul et l'algorithmique

Implémentation de graphes

Comme nous l'avons vu précédemment dans ce rapport, la formalisation du réseau hydrographique est réalisée à l'aide de la bibliothèque graphique Boost (en anglais BGL, Boost Graph Library) qui nous permet d'utiliser des outils pour la gestion de graphes. Boost est en fait un ensemble de bibliothèques logicielles libres écrites en C++, qui vise à remplacer la Bibliothèque Standard du C++. L'écriture des modules au sein de cet ensemble est soumise à un comité de lecture. La plupart du code est distribué selon les termes de la licence logicielle Boost, laquelle autorise autant son intégration dans les logiciels libres que propriétaires. La plupart des fondateurs de Boost se trouvent dans le comité du standard C++ et plusieurs de ses bibliothèques ont été acceptées pour faire partie de la base de travail de la norme C++11. Un tutoriel d'installation des bibliothèques non pré-compilées de Boost est fourni en annexe du présent rapport.

Optimisation

Dans ce cas, l'optimisation est faite grâce à ADMB (AD Model Builder). C'est un paquetage logiciel, très utilisé notamment pour le développement de modèles statistiques

non-linéaires. ADMB est construit autour de la bibliothèque AUTODIF, une extension du langage C++ qui implémente la différentiation automatique en mode inverse[2].

Optimisation pour l'approche alternative de calcul (matricielle)

Nous recourons aussi à la bibliothèque gratuite et open-source NLopt pour l'optimisation non linéaire. Contrairement à l'approche par propagation d'une gaussienne, son utilisation a été rendue nécessaire pour l'approche matricielle car dans ce modèle l'un des paramètres à optimiser est à valeurs discrètes (l'exposant de la matrice stochastique), ce qui rend impossible l'utilisation d'ADMB. D'autre part, la capacité de mémoire qui était requise pour les calculs avec ADMB était de toute façon trop importante. NLopt fournit une interface commune pour différentes routines d'optimisation disponibles en ligne, ainsi que des implémentations inédites de plusieurs autres algorithmes. Ses fonctionnalités incluent :

- Une compatibilité avec les langages C, C++, Fortran, Matlab ou GNU Octave, Python, GNU Guile, Julia, GNU R, Lua, et OCaml.
- Une interface commune pour de nombreux algorithmes. Il est possible d'essayer un algorithme différent en changeant simplement un paramètre.
- Un fonctionnement pour l'optimisation large échelle (algorithme avec un très grand nombre de paramètres et de contraintes).
- Une compatibilité avec des algorithmes d'optimisation locale ou globale.
- Une compatibilité avec des algorithmes n'utilisant que les valeurs de la fonction-objectif (sans dérivées) et des algorithmes appliquant des gradients fournis par l'utilisateur.
- Une compatibilité avec des algorithmes destinés à l'optimisation sans contraintes, à l'optimisation avec contraintes aux bornes, et avec contraintes d'égalité/inégalité non linéaires.

L'algorithme BOBYQA, que nous avons déjà mentionné dans ce rapport, et qui assure l'optimisation des paramètres de l'approche matricielle, est implémenté dans NLopt.

Calcul numérique additionnel

Nous nous servons aussi d'Eigen, une bibliothèque C++ de haut niveau contenant des modèles de déclarations (template headers) pour l'algèbre linéaire, le calcul vectoriel et matriciel, ou la résolution numérique de problèmes. C'est notamment cette bibliothèque qui nous permet de gérer efficacement le calcul avec des matrices creuses dans l'approche matricielle.

Parallélisation du code

Par ailleurs, en prévision d'une réduction du temps de calcul si celui-ci s'avérait trop important lors d'une modélisation de la diffusion des anguilles à l'échelle de la France entière, l'interface de programmation OpenMP (Open Multi-Processing) a été intégrée dans le programme pour paralléliser certaines portions du code. OpenMP est une interface de programmation pour le calcul parallèle sur architecture à mémoire partagée. Cette API est supportée sur de nombreuses plateformes, incluant Unix et Windows, pour les langages de programmation C, C++ et Fortran. Il se présente sous la forme

[2]Le mode inverse de différentiation, ou méthode de l'état adjoint, est une approche dans laquelle il n'y a pas besoin de calculer la dérivée de la fonction implicite (qui permet de réécrire le problème initial d'optimisation avec contraintes en un problème d'optimisation sans contraintes grâce notamment au théorème des fonctions implicites).

d'un ensemble de directives, d'une bibliothèque logicielle et de variables d'environnement. OpenMP est portable et dimensionnable. Il permet de développer rapidement des applications parallèles à petite granularité en restant proche du code séquentiel. Une programmation parallèle hybride peut être réalisée par exemple en utilisant à la fois OpenMP et MPI. OpenMP est une implémentation du multithreading (multi-tâches), une méthode de parallélisation dans laquelle un thread[3] maître (une série d'instructions exécutées de façon séquentielle) se divise en un nombre donné de threads esclaves avec lesquels le système partage les tâches. Les threads fonctionnent donc en parallèle, et l'environnement d'exécution alloue les threads aux différents processeurs en fonction de leur disponibilité.

[3]Un thread, ou fil (d'exécution), ou tâche (terme et définition normalisés par ISO/CEI 2382-7 :2000, mais d'autres appellations sont connues : processus léger, unité de traitement, unité d'exécution, fil d'instruction, processus allégé, exétron), est similaire à un processus car tous deux représentent l'exécution d'un ensemble d'instructions du langage machine d'un processeur. Du point de vue de l'utilisateur, ces exécutions semblent se dérouler en parallèle. Toutefois, là où chaque processus possède sa propre mémoire virtuelle, les threads d'un même processus se partagent sa mémoire virtuelle. Par contre, tous les threads possèdent leur propre pile d'appel.

Chapitre 5

Sorties du modèle

5.1 Présentation des sorties du modèle

5.1.1 Liste et nature des sorties

Nous listons ci-après les variables d'intérêt pour le modèle, en indiquant pour chacune d'elles sa signification et l'unité dans laquelle elle est déterminée.

1. Nombre et densité surfacique d'anguilles jaunes par tronçon du réseau hydrographique (respectivement en nombre d'individus, et en nombre d'individus pour 100 m^2).
2. Nombre et densité surfacique d'anguilles jaunes, en l'absence d'obstacles, par tronçon du réseau hydrographique (respectivement en nombre d'individus, et en nombre d'individus pour 100 m^2).
3. Nombre d'anguilles bloquées à l'aval direct de chaque obstacle (pour le calcul de la mortalité en montaison).
4. Nombre de turbines hydroélectriques en aval de chaque tronçon (pour le calcul de la mortalité en dévalaison).

D'autres variables ne sont utilisées que pour l'exploration et la calibration du modèle, et ne présentent pas d'intérêt particulier en tant que résultats de la simulation :

1. Probabilité absolue de sédentarisation dans chaque tronçon (il s'agit d'une probabilité, donc un nombre sans unité et compris entre 0 et 1).
2. Vraisemblance locale (par opération de pêche).

On rappelle également que nous disposons de l'erreur statistique sur chacun des paramètres optimisés. Celle-ci est disponible dans un fichier texte qui regroupe tous les paramètres optimisés.

5.1.2 Sorties graphiques

Comme nous l'avons déjà vu, la totalité des sorties peut être affichée sur le RHT, ce qui permet d'avoir une vision globale (France entière), à l'échelle d'un bassin versant ou même un zoom sur une région ou un ensemble de tronçons particuliers. Un script utilisable par le logiciel R a été mis au point pour récupérer les sorties intéressantes depuis la base de données gérée par PostgreSQL, les afficher graphiquement puis les sauvegarder sous forme de fichier png (ou jpeg, ou pdf, au choix de l'utilisateur). Le script est disponible en annexe à la page 73. A titre d'information, cet affichage est réalisé en quelques dizaines de secondes sur les 953 bassins versants actuellement pris en compte dans la simulation.

Une autre possibilité est d'utiliser l'éditeur de cartes de QGIS, qui fournit tous les outils utiles pour l'affichage et la sauvegarde des sorties(placement libre des cartes et des légendes notamment), si ce n'est que le processus n'est pas automatisé (il faut relancer l'éditeur à chaque fois que l'on souhaite sauvegarder des résultats).

5.1.3 Machine de référence

Dans tout ce rapport, les performances données pour le modèle sont obtenues sur une machine de référence, équipée d'un processeur Intel Core i7-4900MQ cadencé à 2,7 GHz et de 16 Go de RAM. On rappelle ici que le modèle est fait pour être multi-plateformes, et peut donc fonctionner sous n'importe quel système d'exploitation, moyennant une procédure d'installation des outils informatiques spécifique à chaque système.

5.2 Données de pêches électriques

Les données de pêches électriques que nous utilisons présentent une hétérogénéité à la fois temporelle et spatiale, ce qui peut constituer une source de biais dans les prédictions du modèle. Néanmoins, étant sont les seules à disposition pour la calibration de TABASCO, il est bon d'observer leur répartition dans le temps et dans l'espace afin de prévenir de potentielles mauvaises interprétations des résultats. C'est le but de ce paragraphe.
Dans l'ensemble, ce sont 8 078 observations que nous utilisons sur 20 ans, dont 5170 points de données nulles et 2908 points de données positives.
Intéressons-nous tout d'abord à l'évolution de la quantité de données disponibles au cours des 20 années considérées dans la modélisation. Nous présentons sur la figure 5.1 la quantité de données disponibles pour chacune des années entre 1990 et 2009. Sur la même figure, sont indiquées le nombre de données nulles (absence d'anguilles) et le nombre de données positives (présence d'anguilles), ce qui permet aussi d'évaluer la contribution relative de ces données.
On constate que le taux de données positives reste stable sur toute la période consi-

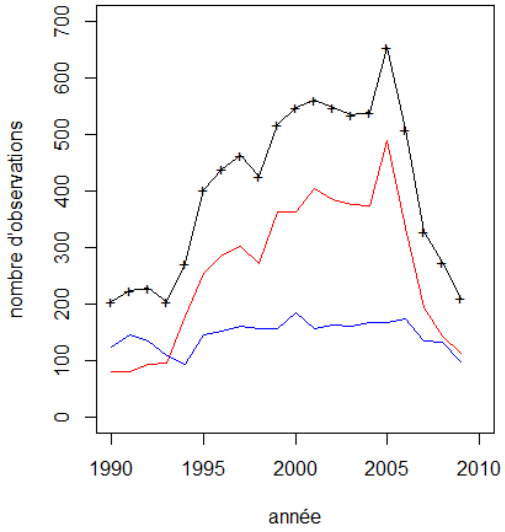

FIGURE 5.1 : Quantité et nature des données utilisées dans le modèle. Le nombre total de données annuelles est indiqué en noir, les données nulles (absence d'anguilles) sont en rouge, et les données positives (présence d'anguilles) sont en bleu.

dérée, autour de 130 observations positives par an. En revanche, le taux de données nulles varie sensiblement au cours du temps. Il passe ainsi de 80 observations nulles en 1990 à 488 en 2005, avant de redescendre rapidement en 2009 pour retrouver son niveau initial. La variation du nombre total de données annuelles est donc la conséquence directe de la variation du nombre d'observations nulles au cours du temps.
Si l'on s'intéresse ensuite à la répartition spatiale des données, on note là encore de fortes disparités en fonction des années. Nous avons regroupé, sur la figure 5.2, la répartition des pêches électriques complètes utilisées pour la calibration du modèle pour quatre années qui jalonnent les 20 années prises en compte avec un intervalle de temps de cinq ans, soit les années 1990, 1995, 2000 et 2005. On visualise bien qu'en 1990 par

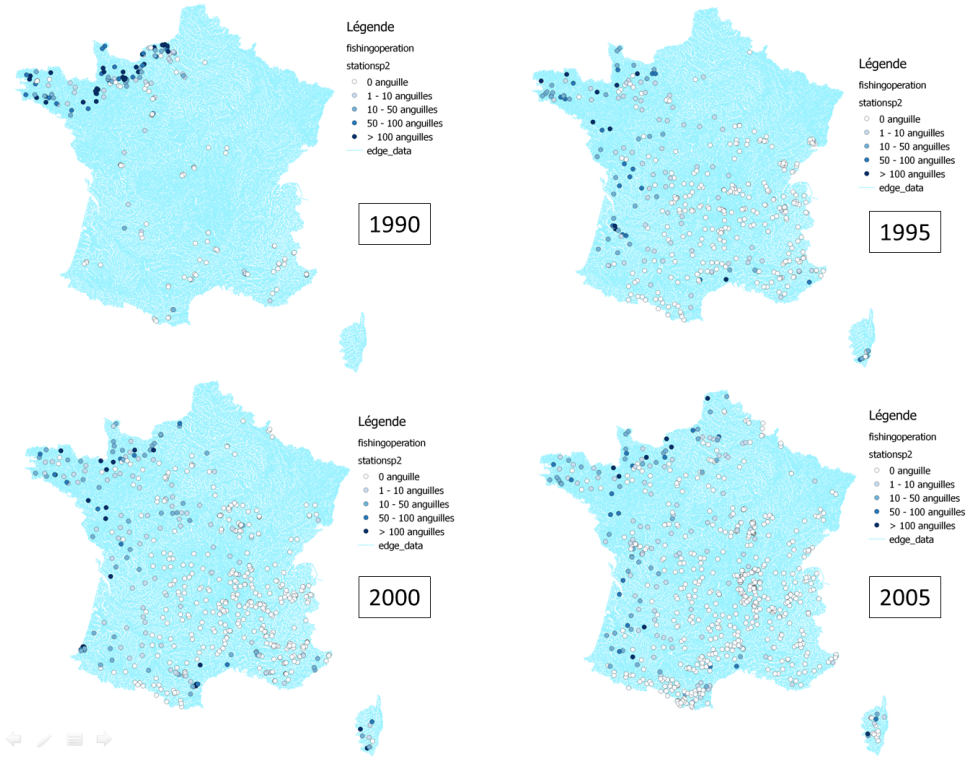

FIGURE 5.2 : Répartition en France métropolitaine des pêches électriques complètes utilisées pour la calibration du modèle pour les années 1990, 1995, 2000 et 2005. Chaque point correspond à une opération de pêche électrique complète. La couleur du point indique l'effectif pêché selon les catégories spécifiées en légende des cartes (de blanc = aucune anguille pêchée à bleu foncé = plus de 100 anguilles pêchées).

exemple, les pêches ont été effectuées presque essentiellement dans les UGA Bretagne et Seine, tandis que le reste du territoire métropolitain n'est pas ou très peu couvert. Il n'y a par exemple aucune opération de pêche en Corse cette année-là, et quasiment aucune dans l'UGA Loire, alors qu'il s'agit de la plus grande unité de gestion. On remarque aussi, en 1990, qu'en dehors des pêches dans les UGA Bretagne et Seine, toutes les autres opérations ont abouti à des effectifs observés nuls (dans leur grande majorité) ou faibles (pour quelques-une d'entre-eux). A partir de 1995, le maillage du territoire par les stations de pêches semble être nettement plus homogène. Seules les UGA Rhin/Meuse et Artois sont très peu représentées, mais cela n'est pas préoccupant pour TABASCO, puisque nous ne réalisons pas de calculs dans les bassins versants dont l'exutoire est situé en dehors du territoire métropolitain (ce qui est le cas de la Meuse et du Rhin, entre autres). Du point de vue des effectifs, on retrouve à partir de 1995 un schéma de répartition spatiale relativement constant, dont nous rediscuterons un

peu plus loin en le comparant avec les prédictions du modèle.
Traçons maintenant les histogrammes de répartition des données nulles (respectivement des données positives) en fonction de la distance à la mer et de la distance à la source tel que cela est fait sur les figures 5.3 et 5.4 (respectivement figures 5.5 et 5.6).
On peut constater que :

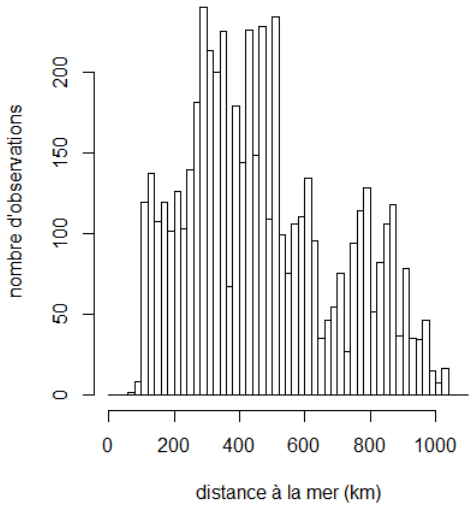

FIGURE 5.3 : Distribution des données nulles en fonction de la distance à la mer.

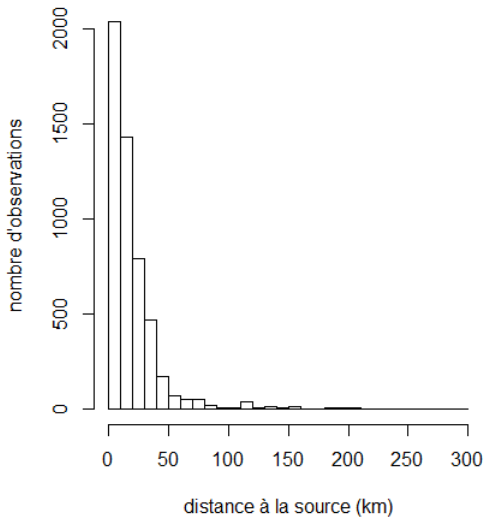

FIGURE 5.4 : Distribution des données nulles en fonction de la distance à la source.

- Les données nulles sont présentes à des distances à la mer très variables, avec cependant une absence de données dans les zones situées à moins de 100 km de la mer, ce qui s'explique par l'absence de pêches électriques dans ces zones[1]. Il semble y en avoir environ deux fois plus dans l'intervalle $[0; 500[$ km que dans l'intervalle $[500; 1000[$ km, ce qui s'expliquerait a priori par le fait que la majorité

[1] Ce qui est d'ailleurs un gros problème pour calibrer les modèles, et qui devrait être une priorité pour les futures campagnes de pêches scientifiques.

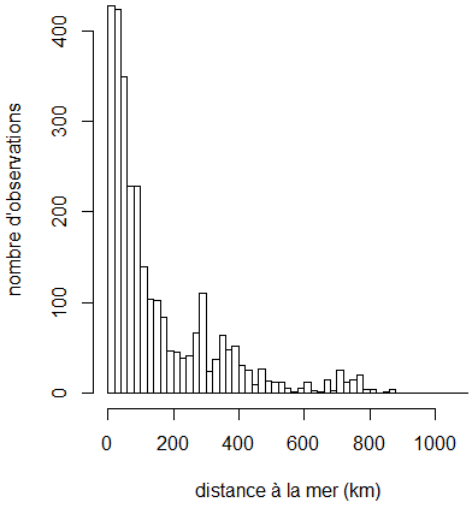

FIGURE 5.5 : Distribution des données positives en fonction de la distance à la mer.

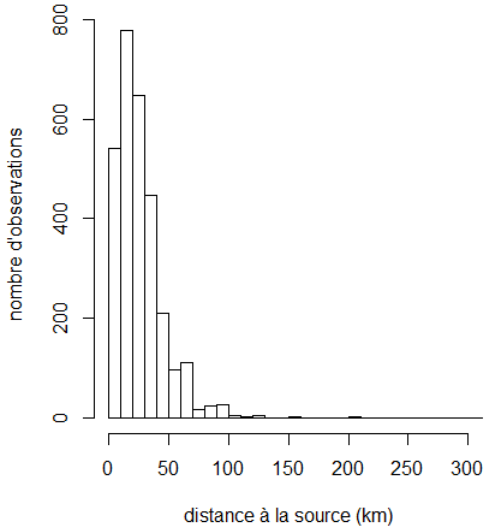

FIGURE 5.6 : Distribution des données positives en fonction de la distance à la source.

des bassins versants n'ont pas de drains dont la longueur est supérieure à 500 km.

- Ces mêmes données nulles suivent par contre une distribution exponentielle décroissante lorsqu'on s'éloignent des sources des drains. Cela n'est pas étonnant, car les anguilles ne peuvent en général par remonter jusqu'en haut des drains, et il est donc logique que la probabilité de ne pas en observer lors d'une pêche électrique augmente lorsqu'on se rapproche d'une source.
- Les données positives suivent une distribution exponentielle décroissante en fonction de la distance à la mer, ce qui là encore est logique car les densités observées les plus fortes sont regroupées à proximité des côtes et elles diminuent rapidement avec la distance à la mer. Lorsqu'on atteint les sources, il est probable ne plus détecter d'anguilles.
- Ce qui est plus étonnant, c'est que les densités positives sont les plus nombreuses

pour de petites distances à la source, ce qui sous-entendrait qu'elles sont plus fréquentes sur les drains de taille modeste. Néanmoins, il s'agit probablement d'un biais dû au fait que les pêches électriques à proximité de la mer sont réalisées sur des petits drains, et non pas sur les drains principaux où il n'est pas possible de mener de pêches électriques à cause de la hauteur d'eau et du débit. Ainsi, il y a peu de drains conséquents pour lesquels on possède des données proches de l'estuaire, ce qui biaise probablement l'histogramme des données positives en fonction de la distance à la source.

Enfin, nous donnons dans le tableau 5.1 les 10 principaux bassins versants contributeurs aux données de pêches électriques entre 1990 et 2009 afin d'avoir un aperçu de leur provenance géographique. On retrouve bien les quatre plus gros bassins (Loire, Rhône, Garonne/Dordogne et Seine) en tant que plus gros contributeurs, même s'il est étonnant de n'avoir que 592 observations sur 20 ans dans le bassin de la Seine, ce qui peut paraître un peu faible en comparaison des trois autres « grands » bassins. On constate également qu'en dehors de ces quatre bassins principaux, les données issues de tous les autres bassins sont très peu nombreuses (moins de 10 données par an et par bassin en moyenne, à l'exception de l'étang de Berre).

Bassin versant	**Nombre d'opérations**
Loire	1911
Rhône	1880
Gironde	1291
Seine	592
étang de Berre	246
Adour	170
Vilaine	156
Charente	112
Orne	104
Touques	73

TABLE 5.1 : Tableau listant les 10 bassins versants qui fournissent le plus de données de pêches électriques entre 1990 et 2009 et le nombre de pêches correspondantes.

5.3 Résultats du modèle

Sauf mention contraire, tous les résultats présentés ici sont issus de simulations en mode multi-années (calibration sur 20 années de données), multi-bassins (953 bassins) et avec prise en compte des variations du stock entre unités de gestion anguilles. Ces résultats sont ceux obtenus au bout des trois années de développement du projet TABASCO.
Comme nous l'avons expliqué précédemment dans ce rapport, deux options de calcul de la variance ont été testées. Dans toute la suite du rapport, elles seront désignées en tant qu'option n° 1 et option n° 2, avec les correspondances suivantes :
Une **variance constante sur tous les bassins versants** pour l'option **n° 1**.
Une **variance constante sur un drain** (proportionnelle au carré de la longueur du drain sur chaque drain du réseau) pour l'option **n° 2**.

5.3.1 Résultats numériques de l'optimisation des paramètres pour les deux options de calcul de la diffusion

L'algorithme de minimisation converge correctement dans les deux cas, mais pas sous n'importe quelles conditions. La convergence dépend des plages de variation allouées à chacun des paramètres optimisés. Il est impossible de savoir à ce stade si ces variations sont uniquement dues à des restrictions mathématiques dans l'optimisation, mais c'est une hypothèse probable. Une autre considération concerne le comportement du paramètre de franchissabilité des obstacles. Dans les deux options de calcul, et comme nous l'avons déjà mentionné, ce dernier tend toujours vers la borne maximale (et donc vers un indice de franchissement des obstacles de 100 %), et ce, quelles que soient les plages de variation des paramètres, du moment que la convergence a lieu.
En ce qui concerne l'option n°1, l'algorithme converge en 830 itérations et 1 085 évaluations de la fonction objectif (on ne compte pas la dernière itération pour laquelle l'algorithme ne progresse plus malgré les 200 appels supplémentaires de la fonction objectif, ce qui correspond à l'un de nos critères d'arrêt de l'algorithme). La valeur absolue finale de la logvraisemblance est de 9 651,47, avec une valeur absolue maximale de gradient de $9,62.10^{-6}$. Aucun des paramètres n'a atteint sa borne minimale ou maximale, sauf l'indice de franchissabilité des obstacles, qui atteint sa borne maximale (il vaut donc 100 %). Néanmoins, comme la dérivée de ce paramètre a une valeur finie à la borne même à la fin de l'optimisation, le hessien peut être calculé car tous ses coefficients sont à valeurs finies. Le modèle a tourné en 1 h 34 min 50 s (toutes étapes confondues).
En ce qui concerne l'option n°2, l'algorithme converge en 767 itérations et 1 026 évaluations de la fonction objectif. La valeur absolue finale de la logvraisemblance est de 9 785,18, avec une valeur absolue maximale de gradient de $6,26.10^{-4}$.
Dans les deux cas, nous considérons que la convergence de l'algorithme de minimisation s'est déroulée correctement, c'est-à-dire que l'optimum de la logvraisemblance a été trouvé sous les contraintes que nous avons imposées.
Nous récapitulons dans les tableaux 5.2 et 5.3 les résultats obtenus pour l'optimisation des 31 paramètres du modèle pour l'option n°1 et l'option n°2 respectivement. On y observera notamment que l'indice de franchissabilité atteint ou tend vers sa borne maximale. La valeur après optimisation de l'écart-type de la distribution log-normale est très élevée, comme nous l'avons déjà mentionné. L'optimisation de ce paramètre engendre un certain nombre problèmes dans les prédictions des densités (voir plus haut dans ce chapitre), mais nous ne sommes pas en mesure de savoir quelles en sont les causes. On notera enfin que le paramètre de gestion des flux aux confluences à une valeur optimisée inférieure à 1 dans l'option n°1, ce qui implique dans ce cas que les anguilles privilégient le drain le plus petit lorsqu'elle parviennent à une confluence, tandis qu'il prend une valeur supérieure à 1 dans l'option n°2, ce qui implique dans cet autre cas que les anguilels privilégient le plus gros drain lorsqu'elle parviennent à une confluence (ce qui serait plutôt la situation à laquelle on s'attendrait dans la réalité).

5.3.2 Évaluation de la qualité du modèle

Afin de juger de la qualité du modèle TABASCO, nous utiliserons le critère d'évaluation le plus courant, à savoir le critère d'information d'Akaike (Akaike, 1973, 1974), plus connu sous son acronyme anglais d'AIC (pour *Akaike Information Criterion*). Il se calcule comme suit :

$$AIC = 2k - 2\ln(\mathcal{L})$$

Où k est le nombre de paramètres à estimer du modèle, et $\mathcal{L}$ est le maximum de la fonction de vraisemblance du modèle.
Dans notre cas, nous estimons 31 paramètres, ce qui donne un AIC de 19 364,94 pour l'option de calcul n° 1 et de 19 632,36 pour l'option n° 2. Ces deux valeurs sont très proches, ce qui n'est évidemment pas étonnant car nous utilisons le même nombre de

paramètres dans les deux cas, et les valeurs des maxima des vraisemblances sont également très proches. Néanmoins, à nombre de paramètres égal, l'option de calcul n° 1 apparaît légèrement meilleure.
Nous avons réalisé ce calcul afin de pouvoir comparer le modèle TABASCO avec d'autres modélisations. Par contre, son intérêt est nul pour la comparaison de nos deux options de calcul, car le nombre de paramètres est identique, et la différence était déjà quantifiable par la comparaison des vraisemblances. Cela s'explique par le fait que le critère d'information d'Akaike n'est qu'un prolongement de la méthode du maximum de vraisemblance.

5.3.3 Indépendance et caractère identifiable des paramètres du modèle

Un contrôle rapide de la matrice de corrélation du vecteur des paramètres du modèle permet d'avoir une idée de la corrélation, et donc de l'inter-dépendance des paramètres du modèle. Du fait de la grande taille de cette matrice (31x31), nous l'avons déplacée en annexe du présent rapport (à la page 77), et n'avons gardé, par souci de clarté, que la partie sous la diagonale, car il s'agit d'une matrice symétrique. Les coefficients de la partie supérieure sont donc les symétriques de ceux de la partie inférieure par rapport à la diagonale.
On vérifie bien que les coefficients sur la diagonale sont tous égaux à 1. On note aussi qu'à une exception près (coefficient en rouge), tous les coefficients sont inférieurs à 0,47 (pour les deux options de calcul), ce qui correspond a priori à une faible corrélation (en pratique, on considère tout coefficient de corrélation inférieur à 0,5 comme représentant un faible niveau de corrélation). Le seul coefficient dont la valeur est remarquable est celui entre le paramètre de gestion des flux aux confluences et le paramètre σ de la gaussienne (écart-type). Sa valeur de +0,70 environ (pour l'option 2) et de -0.55 environ (pour l'option 1) tendrait à démontrer d'une part qu'il existe une corrélation entre ces deux paramètres, et d'autre part qu'elle est fortement dépendante des hypothèses prises pour le calcul de la diffusivité (la corrélation étant positive dans un cas, négative dans l'autre). Il faut néanmoins rester prudent car ces valeur ne sont pas très élevées (en tout cas elles ne sont pas proches de 1) et elles ne traduisent donc pas une forte corrélation.
Afin de vérifier l'indépendance statistique de nos 31 paramètres (leur identifiabilité), nous procéderons comme cela a été expliqué à la page 17 pour contrôler que tous les paramètres du modèle sont bien identifiables, et que le modèle n'est pas sur-paramétré. Nous avons procédé pour cela à une décomposition en valeurs singulières de la matrice hessienne des paramètres au point optimum. Le résultat est présenté sur la figure 5.7. La valeur maximale des valeurs singulières[2] est de 21 320,64. En appliquant la méthode

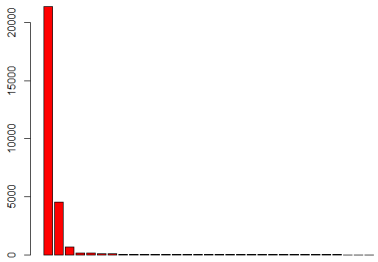

FIGURE 5.7 : Décomposition en valeurs singulières de la hessienne au point optimum, avec l'option de calcul n°2. les valeurs singulières sont classées par valeur décroissante.

[2]Qui sont identiques aux valeurs propres car la matrice hessienne est semi-définie positive.

du seuil donnée par Viallefont *et al.* (1998), on obtient un seuil S qui vaut :

$$S = 31 \times 21320,64 \times 1,0 \times 10^{-9} = 6,6 \times 10^{-4} \tag{5.1}$$

Ce seuil étant inférieur à la plus petite valeur des valeurs singulières, on en déduit que tous les paramètres du modèle sont identifiables.

5.3.4 Étude des écarts entre prédictions et observations

Le but de ce paragraphe est d'une part de comparer quantitativement les écarts obtenus entre les prédictions du modèle et les données issues des observations, et d'autre part d'analyser les variations de ces écarts en fonction des paramètres géographiques pour identifier les zones problématiques dans les prédictions. Cela permet d'opérer un premier diagnostic du modèle.
La façon la plus simple de juger de la qualité des prédictions du modèle est de tracer les densités observées en fonction des densités prédites, ainsi que nous l'avons fait sur la figure 5.8. On constate immédiatement que le meilleur ajustement linéaire des don-

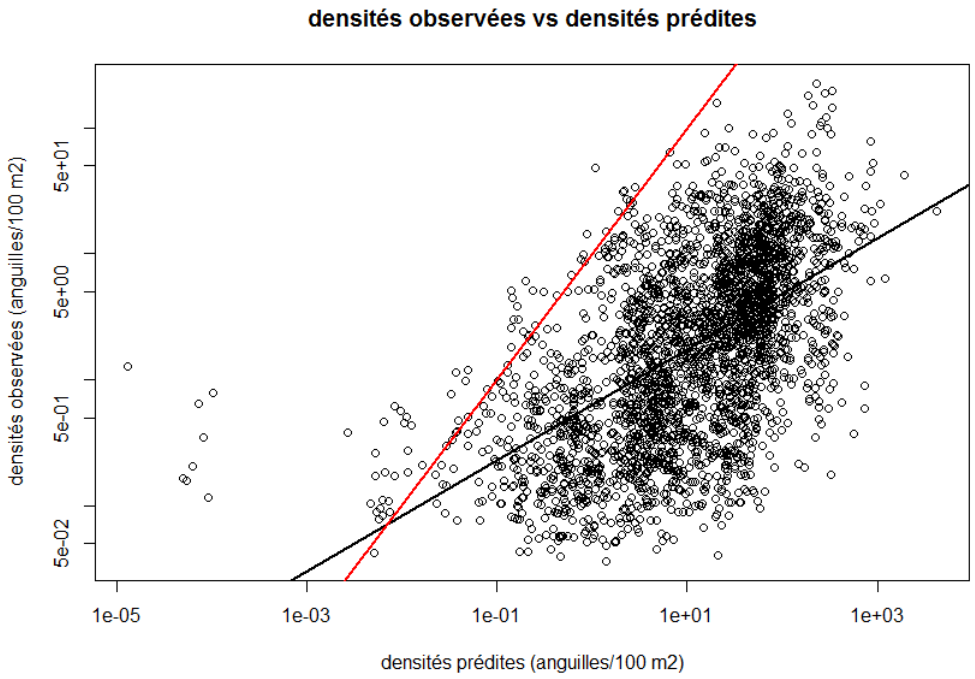

FIGURE 5.8 : Graphique sur lequel sont tracées les densités observées en fonction des densités prédites pour chaque opération de pêche électrique complète entre 1990 et 2009, mais seulement pour les densités observées positives (car on ne peut pas représenter les données nulles en échelle logarithmique). Les deux axes sont en échelle logarithmique. Le meilleur ajustement linéaire des données est tracé en rouge, tandis que la droite d'équation y=x est tracée en noir. Dans l'idéal (c'est-à-dire si on prédisait dans le modèle les valeurs exactes des densités observées), la ligne rouge devrait coïncider avec la ligne noire. Le fait que les deux droites ne soient pas superposées indique un écart entre prédictions et observations (en l'occurrence une surestimation des densités puisque la barycentre des points est en-dessous de la droite y=x). L'estimation du modèle est réalisée avec l'option de calcul n° 1, mais les résultats sont assez similaires (même si la surestimation est un peu moindre) avec l'option n° 2. Les densités sont exprimées en anguilles/dam^2.

nées (tracé en rouge) ne coïncide pas avec la droite d'équation y=x (tracée en noir), ce qui implique un biais dans la prédiction des densités d'anguilles. En l'occurrence, la quasi totalité des points sont situés sous la droite y=x, ce qui suppose un biais quasi systématique, qui se trouve être une surestimation des densités. Il semble donc assez clair que TABASCO surestime de façon générale les densités d'anguilles jaunes.
En reprenant le même graphe, nous avons également tracé les quantiles à 5 et à 95 % pour la valeur optimisée dans le modèle de l'écart-type de la loi log-normale, et pour une valeur théorique "raisonnable" de 0,3 (figure 5.9). On voit facilement sur cette

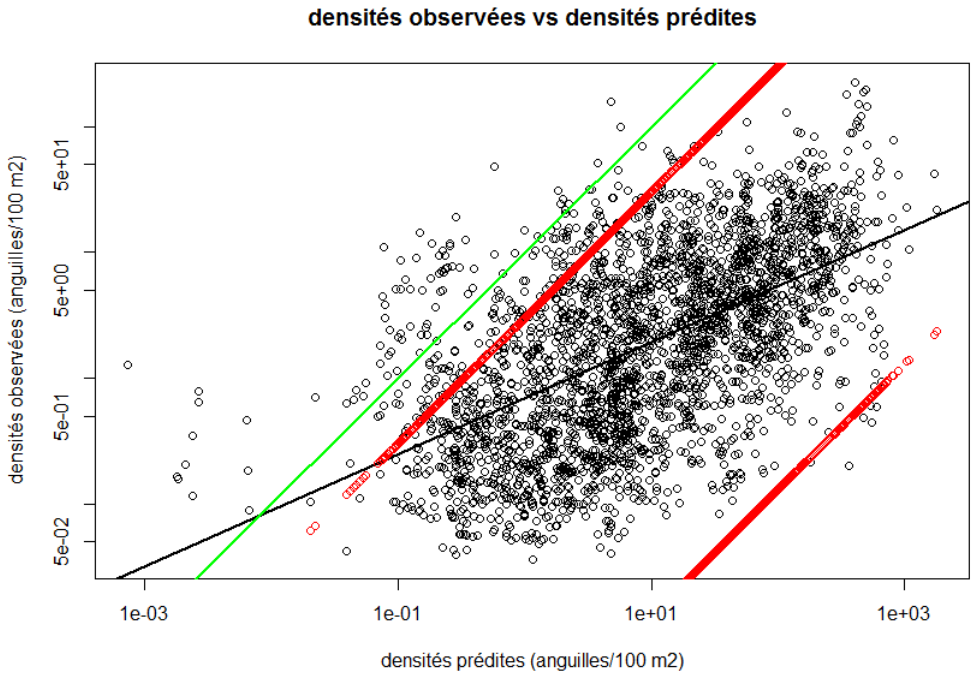

FIGURE 5.9 : Graphique sur lequel sont tracées les densités observées en fonction des densités prédites pour chaque opération de pêche électrique complète entre 1990 et 2009, mais seulement pour les densités observées positives (car on ne peut pas représenter les données nulles en échelle logarithmique). Les deux axes sont en échelle logarithmique. Les densités sont exprimées en anguilles/dam^2. Les quantiles à 5 et à 95 % sont également tracés en rouge (ligne du bas pour le quantile à 5 % et ligne du haut pour celui à 95 %), et la droite y=x est tracée en vert.

figure que l'essentiel des points (quantiles) sont situés entre les deux quantiles à 5 et 95 % pour la valeur optimisée de sigma, ce qui est normal, mais que l'ensemble de la distribution est décentrée par rapport à la droite d'équation y=x. Cela traduit en fait la dissymétrie de la distribution log-normale qui augmente avec la valeur de sigma. Avec de fortes valeurs de sigma (ce qui est notre cas lors de l'optimisation), la loi log-normale devient très fortement dissymétrique avec une longue queue de distribution, et les densités observées peuvent prendre des valeurs très éloignées des densités prédites, ce qui explique le décalage du nuage de points vers la droite et vers le bas.
Afin d'essayer d'identifier les zones les plus problématiques pour l'estimation des densités, nous avons étudié les variations géographiques des écarts de densités. Pour cela, nous avons tout d'abord tracé la carte de France sur laquelle nous avons positionné les tronçons dans lesquels nous pouvions calculer l'écart absolu |densité prédite - densité observée| (figure 5.10). Cette carte semble indiquer un écart d'autant plus fort que l'on se rapproche des côtes, et d'autant plus faible que l'on s'en éloigne. Cela

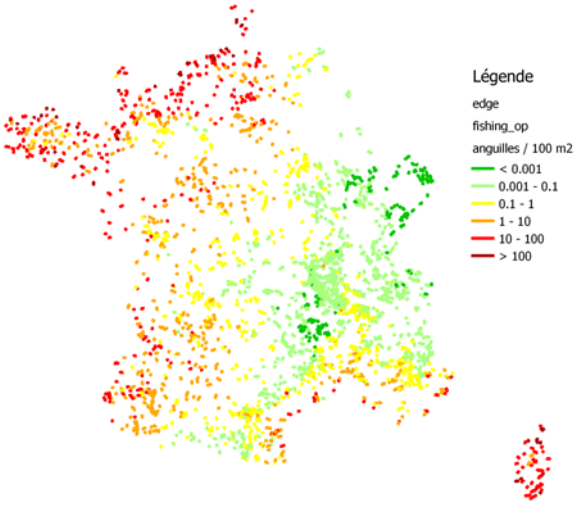

FIGURE 5.10 : Carte de France sur laquelle sont localisés les tronçons pour lesquels on peut calculer un écart absolu |densité prédite - densité observée|. La couleur indique la valeur de l'écart (en valeur absolue), selon la légende fournie à droite de la carte, et allant du vert foncé (écart nul) jusqu'au rouge foncé (écart supérieur à 100 anguilles/dam^2). Cette carte est réalisée sur les 20 années de données dont nous disposons (de 1990 à 2009), ce qui implique qu'il peut y avoir des recouvrements sur certains tronçons. En effet, certaines stations ont fourni des opérations de pêche plusieurs fois entre 1990 et 2009, ce qui mène sur la carte à la superposition des couleurs au même endroit.

est cohérent avec la figure 5.16 sur laquelle il apparaissait qu'on prédisait des nombres d'anguilles un peu trop forts par rapport aux observations dans la bande littorale, tandis que les observations à plus grande distance de la mer semblaient relativement correctes. Il faut néanmoins être prudent sur ces conclusions, car il s'agit d'un écart absolu, qui n'est donc pas forcément très représentatif de la qualité de la prédiction. Par exemple, un écart absolu d'une anguille/dam^2 avec une densité observée de 0 anguille est très problématique, tandis que le même écart pour une densité observée de 50 anguilles/dam^2 est négligeable.
Nous avons donc aussi regardé les écarts relatifs (densité prédite - densité observée) / densité observée, exprimés en pourcentages. Nous avons choisi l'année 2005, car c'est l'année durant laquelle nous disposons du maximum de données. Nous excluons évidemment les densités observées nulles, car l'écart relatif n'est pas défini dans ce cas. Le résultat est présenté sur la figure 5.11. On peut vérifier sur cette nouvelle carte que le modèle sur-estime globalement les densités (peu d'écarts négatifs), en particulier le long des côtes, en Corse, et sur les grands fleuves (Loire et Garonne notamment). Dans une zone centrale allant de la Vendée jusqu'à la Bourgogne, il semble par contre que l'on ait tendance à sous-estimer les densités, mais cela semble secondaire comparé à la sur-estimation globale. Il est assez difficile d'interpréter cet effet.
Afin d'essayer de déterminer s'il existe un autre effet géographique dans la distribution spatiale des écarts en densités, nous avons aussi tracé sur la figure 5.12 une carte du réseau hydrographique théorique avec des nuances de bleu en fonction de l'altitude de chaque tronçon, et sur laquelle sont superposés des points correspondant à chaque opération de pêche électrique complète menée entre 1990 et 2009, et dont la couleur varie en fonction de la valeur du résidu logarithmique $\ln(\text{densité prédite} + 1) - \ln \text{densité observée} + 1)$ (sans unité). Cette quantité à la particularité d'être négative si la densité prédite est inférieure à la densité observée (sous-estimation du modèle), d'être nulle en cas de prédiction parfaite, et d'être positive en cas de densité prédite

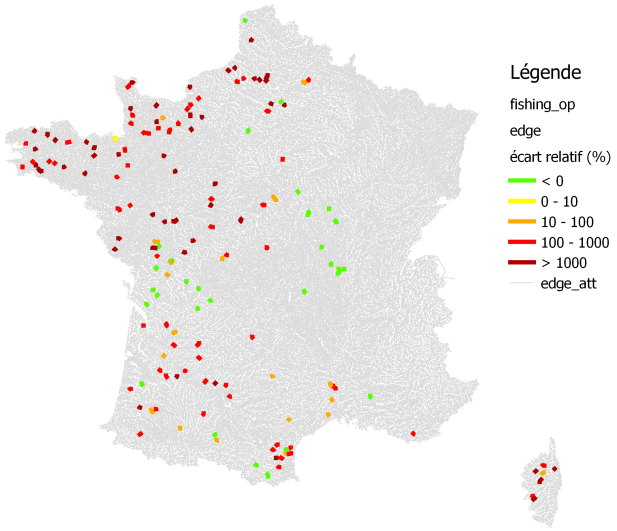

FIGURE 5.11 : Carte de France sur laquelle sont localisés les tronçons pour lesquels on peut calculer un écart relatif (densité prédite - densité observée)/densité observée, exprimé en pourcentage. La couleur indique la valeur de l'écart (qui est négative en cas de sous-estimation du modèle, positive en cas de sur-estimation), selon la légende fournie à droite de la carte, et allant du vert (écart négatif) jusqu'au rouge foncé (écart positif supérieur à 1000 %). Cette carte est réalisée sur les données de l'année 2005 et en excluant les données nulles pour lesquelles il est impossible de calculer un écart relatif.

supérieure à la densité observée (sur-estimation du modèle). Le but est de rechercher des patrons spatiaux qui pourraient expliquer, au moins partiellement, la distribution des écarts en densités que nous observons.

Le premier commentaire que l'on peut tirer de cette figure est que les écarts observés ne semblent pas corrélés à l'altitude des tronçons, même si globalement, les écarts sont plus faibles dans les zones montagneuses, en particulier les Alpes, le Jura, le Massif Central et les Pyrénées Orientales. Cependant, nous ne pouvons pas réellement savoir s'il s'agit d'un effet de l'altitude, de la distance à la mer, de la distance à la source ou d'un autre paramètre topologique.

Par contre, il semble y avoir un effet "UGA" un peu plus net. En effet, les écarts semblent être globalement plus élevés dans les UGA dont les exutoires débouchent sur les côtes de la Manche (Artois-Picardie, Seine-Normandie, Bretagne) et sur le littoral atlantique (Bretagne, Loire et côtiers vendéens, Adour) ainsi que l'UGA Corse. A l'inverse, l'UGA Rhône -Méditerranée et - plus étonnamment - l'UGA Garonne-Dordogne semblent présenter des écarts plus faibles, voire même des sous-estimations des densités d'anguilles jaunes. Cela semblerait indiquer un problème dans l'estimation de l'effet "UGA" tel que nous l'avons intégré au modèle. Il faudra donc travailler sur ce point à l'avenir.

Nous avons enfin tracé la quantité ln(densité observée) - ln(densité prédite) + $\sigma^2/2$, sous forme de boîtes à moustaches. S'il n'y a pas d'erreur de calcul, cette quantité doit être centrée sur 0 puisque ln(densité prédite) + $\sigma^2/2$ correspond à l'espérance de la densité observée. Nous avons tracé les boxplots correspondants en fonction de quatre variables : distance à la mer, distance à la source, UGA et altitude. Le résultat est présenté sur la figure 5.13. Il semble, à la lecture de cette figure, que la qualité des prédictions est plutôt bonne pour les UGA Artois, Bretagne, Corse et Seine. Elle est moyenne pour les UGA Adour, Garonne/Dordogne et Loire, et plutôt mauvaise pour le Rhône. Cela ne correspond pas forcément aux sur-estimations ou aux sous-estimations en densités que nous avons pu observer de façon plus globale un peu plus haut, mais

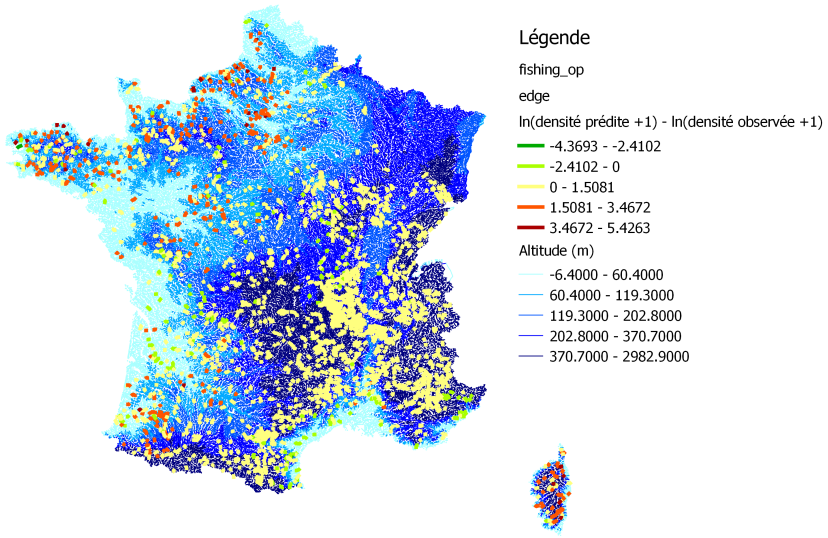

FIGURE 5.12 : Carte de France sur laquelle sont placés les sites de pêches électriques pour lesquels on a calculé le résidu logarithmique $\ln(\text{densité prédite} + 1) - \ln \text{densité observée} + 1)$ (sans unité) sur les 20 années de données (de 1990 à 2009). Cette quantité est négative en cas de sous-estimation des densités, positive dans le cas contraire. L'échelle de couleur correspondante va du vert foncé pour les plus fortes sous-estimations au rouge foncé pour les plus fortes sur-estimations. Sur la même carte, nous avons superposé le réseau hydrographique théorique en indiquant l'altitude de chaque tronçon par un deuxième code couleur, qui va du bleu le plus clair pour les plus faibles altitudes au bleu le plus foncé pour les plus élevées.

donne une indication sur le déroulement de l'optimisation et la qualité des prédictions UGA par UGA. Lorsque l'on regarde en fonction des distances à la mer et des distances à la source, il y a assez peu de commentaires à faire : l'ensemble des boîtes sont centrées sur 0, à l'exception de la boîte associé aux distances à la mer les plus grandes. Cela sous-entendrait qu'avec l'option de calcul n°1, on ne prédit pas forcément bien les densités au bout des drains principaux. Enfin, il semble que nous ayons un léger effet de l'altitude sur la qualité des prédictions, même si cela ne se voyait pas forcément sur la carte 5.12. Plus l'altitude est élevée, moins la prédiction semble être bonne, mais l'effet paraît assez faible.

5.3.5 Prédictions du modèle pour les densités d'anguilles jaunes en France.

Nous présentons dans ce paragraphe les résultats de TABASCO en termes de répartition des anguilles jaunes sur le territoire métropolitain, à travers des cartes de densités par tronçon du réseau hydrographique.
La carte 5.14 est le résultat de la simulation avec l'option de calcul n° 1, tandis que la carte 5.15 est le résultat du calcul avec l'option n° 2. Le schéma de répartition général à l'échelle de la France est sensiblement le même dans les deux cas. En particulier, la zone qui semble inatteignable aux anguilles, et qui apparaît en gris-vert sur les cartes, ne change pas entre les deux options de calcul. A l'échelle macroscopique, on observe aussi en comparant les deux cartes que les zones de fortes densités sont plus étendues à l'issue du calcul avec l'option n° 1 (surtout dans la bande situé à moins de 100 km des côtes), ce qui laisse à penser que le nombre global d'anguilles jaunes prédit pour l'option n° 1 est plus élevé que celui prédit pour l'option n° 2, ce qui sera vérifié un peu plus loin dans le rapport lorsque l'on calculera le paramètre d'intérêt N_0 (nombre d'an-

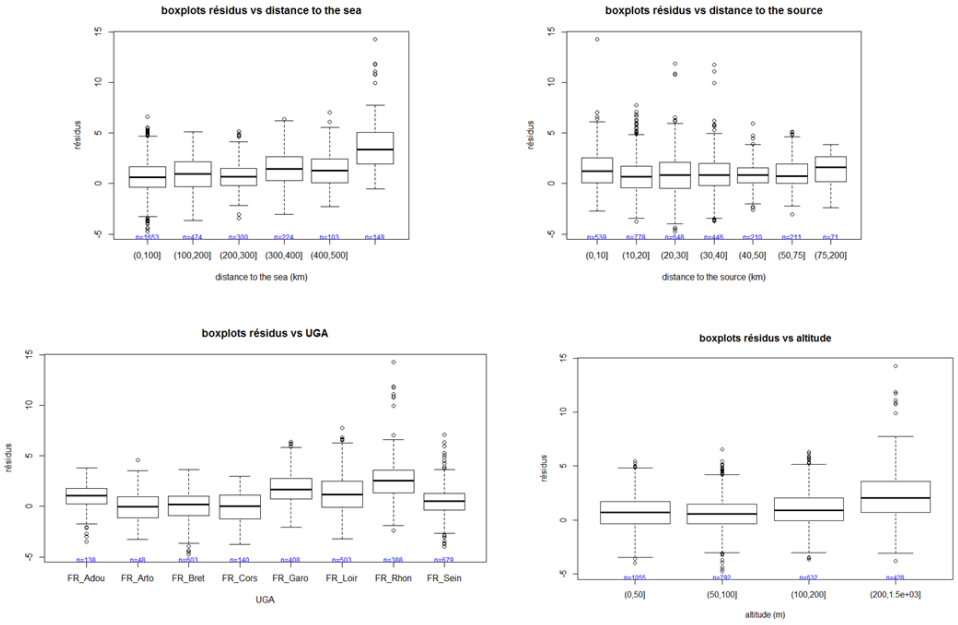

FIGURE 5.13 : Boxplots des résidus de l'espérance de la densité observée, avec prise en compte de la correction de Laurent. Résultats pour l'option de calcul n°1. De haut en bas : en fonction de la distance à la mer, en fonction de la distance à la source, en fonction de l'UGA et en fonction de l'altitude.

guilles jaunes prédit pour la France entière). Les différences interviennent à une échelle plus petite, dans le détail des zones globales que nous venons d'évoquer. Les zones de répartition obtenues avec l'option n° 2 sont plus découpées, et suivent beaucoup plus la géométrie du réseau hydrographique que pour l'option n° 1. Cette observation est logique étant donné les hypothèses sous-jacentes : dans le premier cas, la diffusivité est constante quel que soit le drain, alors que dans le second cas, on prend en compte la longueur de chaque drain, ce qui affine en quelque sorte le schéma de répartition des anguilles, notamment au niveau des petits drains du réseau.

Afin de comparer le schéma global de répartition des anguilles prédit par TABASCO avec les effectifs observés lors des pêches, nous avons tracé sur la figure 5.16 une superposition d'une carte donnant les prédictions du modèle en terme de nombre d'anguilles avec une carte des effectifs d'anguilles estimés d'après les pêches menées entre 1990 et 2009. Les résultats sont ceux de l'année 2009, avec l'option de calcul n° 2. Nous ne présentons pas ceux obtenus avec l'option n° 1 car le maillage des pêches ne nous permet de toute façon pas d'opérer une comparaison à une échelle autre que l'échelle de la France entière. Or, comme nous l'avons déjà mentionné, les différences de résultats entre les deux options de calcul interviennent à une échelle plus petite. Il est surtout intéressant de regarder sur cette figure si la répartition globale des anguilles coïncide avec celle observée sur le terrain, au moins dans ses grandes lignes. Il semble que cela soit le cas, en tout cas dans la bande littorale, même si on peut soupçonner une surestimation au niveau des côtes. Lorsque l'on s'éloigne de la mer, il est par contre clair que l'on a tendance à faire remonter les anguilles trop loin à l'intérieur des terres, surtout sur les plus gros drains (Seine, Loire, Garonne/Dordogne et Rhône). Cela pourrait s'expliquer par la non prise en compte de l'altitude dans le modèle ou par une hypothèse fausse dans le calcul de la diffusivité en fonction des paramètres topologiques. Les zones vides d'anguilles sont par contre assez bien prédites. On ne prend pas en considération les bassins versants situés dans le nord-est de la France (Rhin et Meuse notamment) puisqu'ils sortent du cadre de notre modèle.

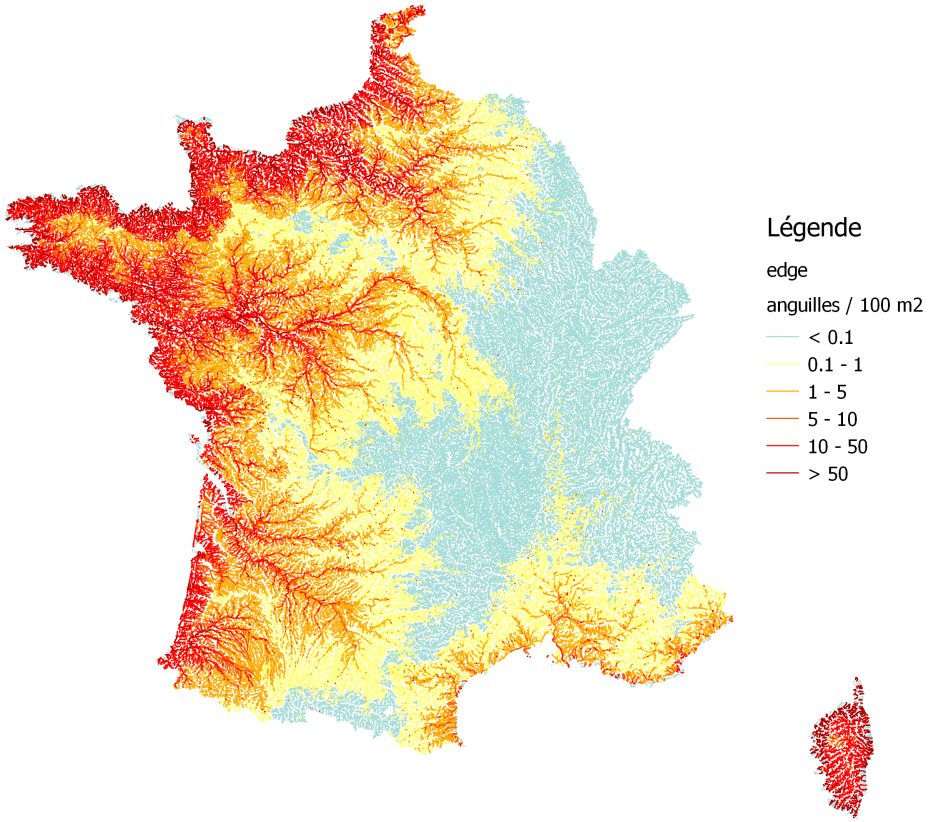

FIGURE 5.14 : Carte des densités d'anguilles jaunes prédites en France métropolitaine pour l'année 2009 avec l'option de calcul n° 1. Les densités sont affichées pour chaque tronçon du réseau hydrographique. La couleur du tronçon correspond à la valeur de la densité prédite selon l'échelle indiquée en légende sur la droite, qui va du gris-vert (pas d'anguille) au rouge foncé (plus de 50 anguilles par dam^2).

5.3.6 Paramètre d'intérêt pour la gestion

Il faut bien noter que le paramètre directement optimisé n'est pas le nombre d'anguilles jaunes, mais le logarithme népérien de la densité moyenne surfacique d'anguilles dans la zone couverte par la simulation ($D0$, qui s'exprime en nombre d'anguilles par km^2). Il est très important de noter que cette densité surfacique est à entendre comme le nombre moyen d'anguilles par unité de surface sur toute la superficie prise en compte dans la simulation (donc dans ce cas sur la superficie totale des 953 bassins versants). Cette précision est importante, car cette quantité n'a pas la même signification que la densité surfacique d'anguilles présentées dans chaque tronçon (notamment sur les cartes de répartition), et qui, bien que s'exprimant dans la même unité, correspond au nombre d'anguilles par unité de surface sur la seule surface en eau du tronçon.
Afin de rendre le paramètre d'intérêt $N0$ plus explicite, nous avons calculé sa valeur dans les tableaux 5.4 et 5.5 (pour les options n° 1 et n° 2 respectivement) pour chaque année entre 1990 et 2009, sur les 953 bassins versants intégrés dans la simulation, et en extrapolant sur l'ensemble de la France métropolitaine à partir des superficies respectives. Les valeurs obtenues sont plus élevées d'au moins un ordre de grandeur

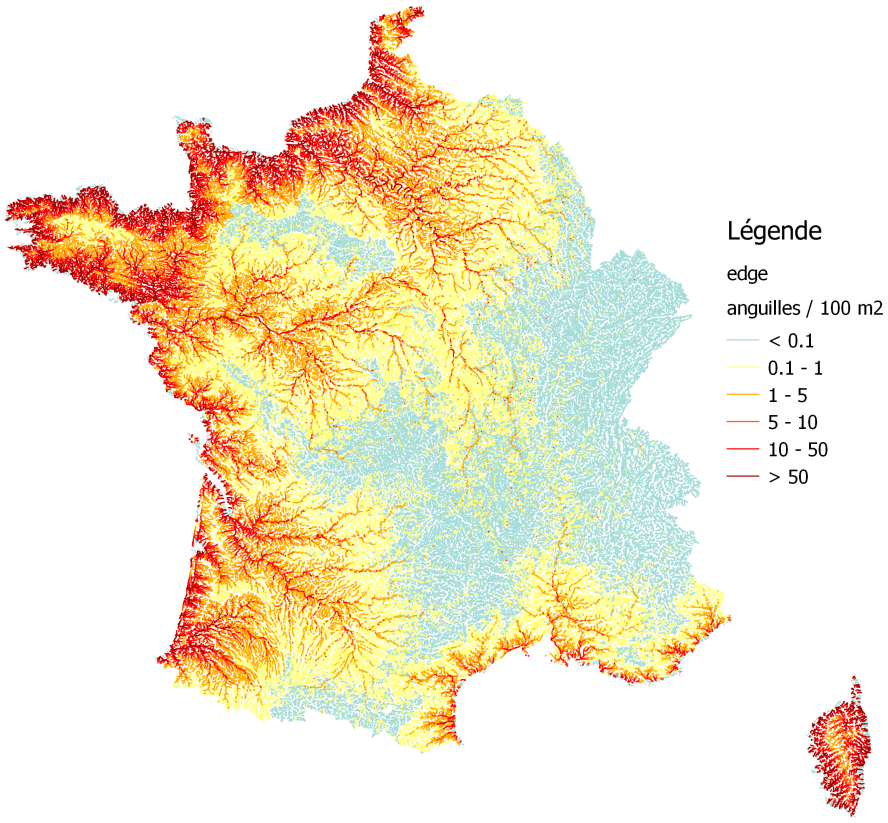

FIGURE 5.15 : Carte des densités d'anguilles jaunes prédites en France métropolitaine pour l'année 2009 avec l'option de calcul n° 2. Les densités sont affichées pour chaque tronçon du réseau hydrographique. La couleur du tronçon correspond à la valeur de la densité prédite selon l'échelle indiquée en légende sur la droite, qui va du gris-vert (pas d'anguille) au rouge foncé (plus de 50 anguilles par dam^2).

que celles indiquées dans le plan de gestion anguille de 2012 (Anonyme, 2012), en l'occurrence 1,40.10^7 anguilles jaunes, dans le rapport intermédiaire du modèle EDA (Jouanin *et al.*, 2012b), à savoir 4,54.10^7 anguilles jaunes et dans le dernier rapport du modèle EDA, paru en juin 2015 (Briand *et al.*, 2015), dans lequel le nombre d'anguilles jaunes pour l'année 2009 est estimé à 4,63.10^7. Il est actuellement impossible de vérifier ces nombres et de savoir quelle est l'estimation la plus proche de la réalité, mais il est néanmoins possible que nous surestimions les effectifs d'anguilles jaunes étant donné les écarts en densité que nous avons évoqués précédemment dans ce rapport.

Nous avons aussi tracé l'évolution du paramètre N0 en fonction du temps entre 1990 et 2009, afin d'observer la tendance globale du nombre d'anguilles jaunes recrutées en France durant cette période. Nous avons tracé, pour chacune des deux options de calcul, un graphique en échelle naturelle et un graphique en échelle logarithmique (figures 5.17 et 5.18 pour l'option n° 1, 5.19 et 5.20 pour l'option n° 2). Quelle que soit l'option de calcul considérée, on note clairement une tendance décroissante, très proche d'une décroissance linéaire au cours du temps lorsque l'on est en échelle logarithmique. Cela suggère donc fortement une décroissance exponentielle du paramètre

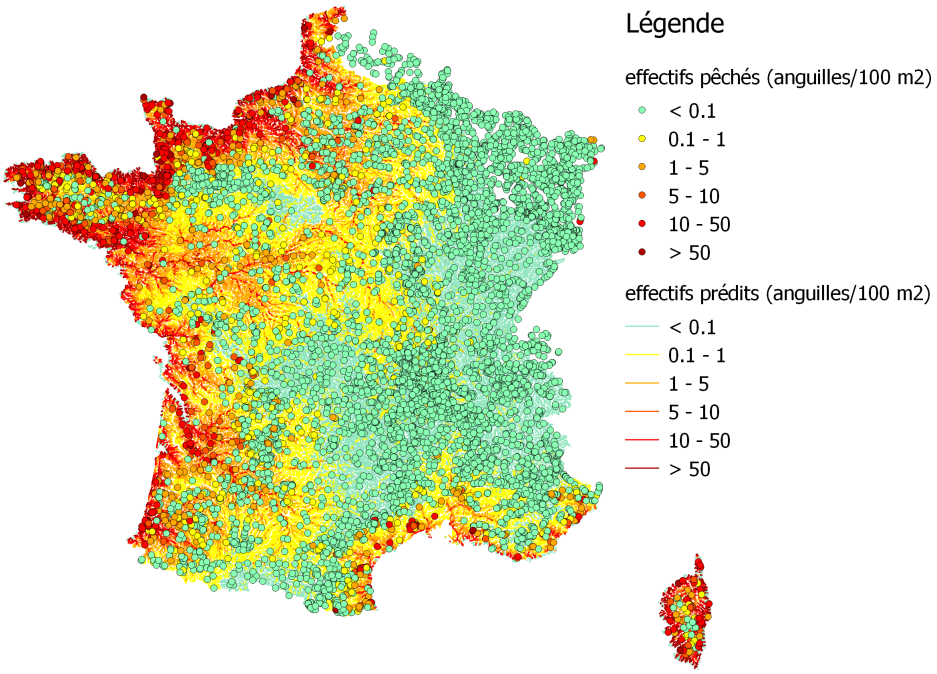

FIGURE 5.16 : Carte des effectifs prédits par le modèle TABASCO pour l'année 2009 avec l'option de calcul n° 2, en arrière-plan, et localisation des opérations de pêche en premier plan. Chaque opération de pêche est représentée par un point dont la couleur indique la valeur de l'effectif observé selon l'échelle indiquée à droite de la figure, qui est identique à l'échelle de couleur pour les effectifs prédits dans chaque tronçon du réseau hydrographique.

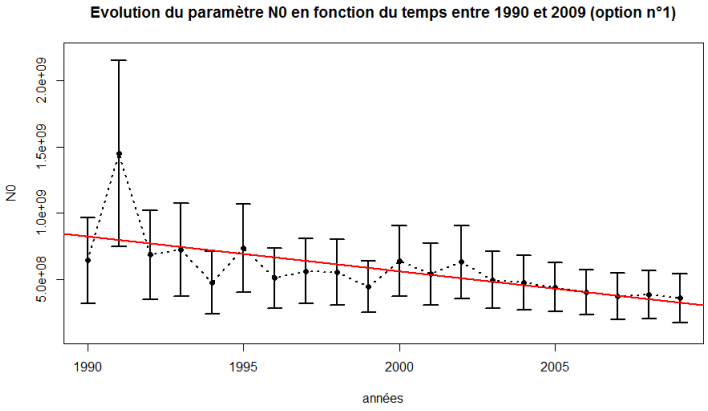

FIGURE 5.17 : Évolution du paramètre N0 en fonction du temps, entre 1990 et 2009, avec intervalle de confiance à 95 % (1,96 σ) sur l'estimation et meilleur ajustement linéaire (en rouge), pour l'option de calcul n° 1. Les axes sont en échelle linéaire. Le nombre est donné pour la France (métropolitaine) entière.

d'intérêt N_0. Un paramétrage de N_0 sous la forme a.exp(-b t) nous permet d'évaluer le taux de décroissance annuelle à 4,1 et à 4,9 % respectivement pour l'option n°1 et l'option n°2 (soit une valeur de 0,041 et 0,049 pour le coefficient b) sous l'hypothèse d'une décroissance exponentielle. On passe ainsi d'environ 10^9 anguilles en 1990 à 3.5×10^8 anguilles en 2009, soit une diminution de 65 % du stock d'anguilles jaunes en 20 ans.

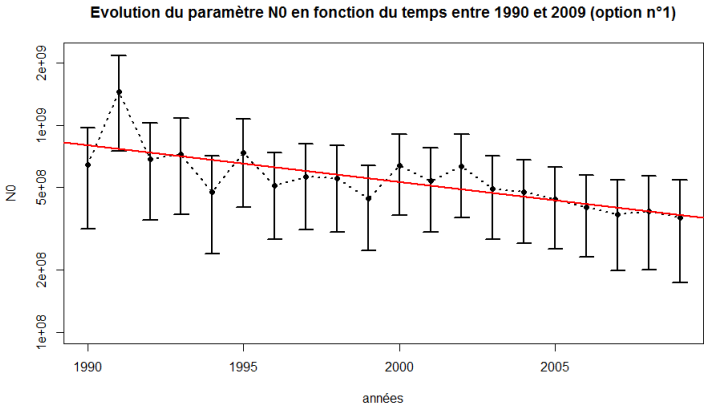

FIGURE 5.18 : Évolution du paramètre N0 en fonction du temps, entre 1990 et 2009, avec intervalle de confiance à 95 % (1,96 σ) sur l'estimation et meilleur ajustement linéaire (en rouge), pour l'option de calcul n° 1. Les axes sont en échelle logarithmique. Le nombre est donné pour la France (métropolitaine) entière.

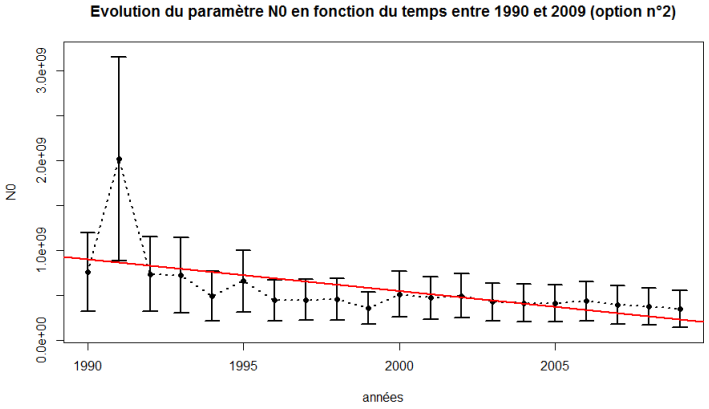

FIGURE 5.19 : Évolution du paramètre N0 en fonction du temps, entre 1990 et 2009, avec intervalle de confiance à 95 % (1,96 σ) sur l'estimation et meilleur ajustement linéaire (en rouge), pour l'option de calcul n° 2. Les axes sont en échelle linéaire. Le nombre est donné pour la France (métropolitaine) entière.

Cela est plutôt cohérent avec l'évolution de l'indice de recrutement des civelles sur la même période, qui passe de 33.3 à 4.3 (base 100 = indice 1980), soit une diminution de 87 %, d'autant plus que dans le même temps, le taux d'exploitation des civelles a lui-aussi diminué (Anonyme, 2015).
Seule l'année 1991 semble présenter une remontée du nombre d'anguilles, sans que nous puissions dire s'il s'agit d'une prédiction fiable du modèle TABASCO qui reproduirait une fluctuation observée, ou d'un artefact de calcul dû, par exemple, à un manque de données ou à des données lacunaires en France cette année-là. On remarquera néanmoins que 1991 est l'année pour laquelle le modèle GEREM (Drouineau *et al.*, 2016) prédit le taux d'exploitation de civelles le plus bas de toute la période 1990-2009 (à environ 14 %), ce qui pourrait conforter la hausse observée dans TABASCO, même si dans le même temps, l'indice de recrutement était à son plus bas de la période 1990-2000 (Anonyme, 2015).

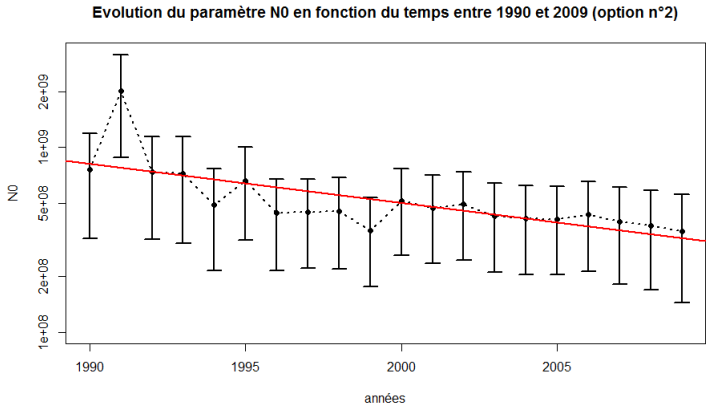

FIGURE 5.20 : Évolution du paramètre N0 en fonction du temps, entre 1990 et 2009, avec intervalle de confiance à 95 % (1,96 σ) sur l'estimation et meilleur ajustement linéaire (en rouge), pour l'option de calcul n° 2. Les axes sont en échelle logarithmique. Le nombre est donné pour la France (métropolitaine) entière.

5.4 Discussion générale des résultats

5.4.1 Commentaires généraux sur l'optimisation

Quelle que soit l'option de calcul utilisée pour la variance, nous allons voir un peu plus loin dans ce chapitre que nous sommes confrontés à deux problèmes à l'issue de la phase d'optimisation, dont nous ne savons pas pour l'heure s'ils sont liés.
Le premier problème, et probablement le principal, concerne le paramètre de franchissement moyen des obstacles, qui tend systématiquement vers 100 % lors de l'optimisation. Cela signifierait dans la réalité que toutes les anguilles franchissent l'ensemble des obstacles sans qu'aucune ne reste bloquée, ce qui n'est pas possible.
Le second problème concerne le paramètre donnant la dispersion de la distribution log-normale des densités observées dans les tronçons. L'optimisation aboutit dans les 2 cas à une valeur beaucoup trop élevée, de l'ordre de 2,4, alors qu'on attendrait plutôt une valeur de l'ordre de 0,3. A titre d'illustration, nous avons tracé sur la figure 5.21 l'allure d'une distribution log-normale pour des densités prédites d'1 anguille/100 m^2, de 10 anguilles/100 m2 et de 50 anguilles/100 m2 (valeurs raisonnables de densités d'anguilles dans un tronçon du réseau hydrographique). Pour les trois courbes tracées, l'écart-type σ de la distribution vaut 2,4, c'est-à-dire la valeur obtenue après optimisation. On voit immédiatement qu'une valeur de 2,4 pour l'écart-type de la distribution est beaucoup trop grande. Les queues de distribution sont très importantes comparativement à l'aire globale sous les courbes, et ce d'autant plus que la densité prédite augmente. Ainsi, on observe qu'une densité prédite d'1 anguille/100 m^2 peut correspondre à des densités observées allant jusqu'à environ 20 anguilles/100 m^2, et pour une densité prédite de 50 anguilles/100 m^2, on peut observer entre 0 et plus de 100 anguilles/100 m^2, ce qui suppose des résidus (densités prédites - densités observées) importants et avec une large variabilité. Nous vérifierons d'ailleurs ce point un peu plus loin dans ce rapport.

5.4.2 Discussion sur l'ordre de grandeur et les valeurs possibles de la variance

Nous discutons ici de la valeur de la variance de la distribution gaussienne lors du phénomène de diffusion, notamment en comparant l'ordre de grandeur théorique attendu et la valeur du paramètre optimisé dans les deux options de calcul.

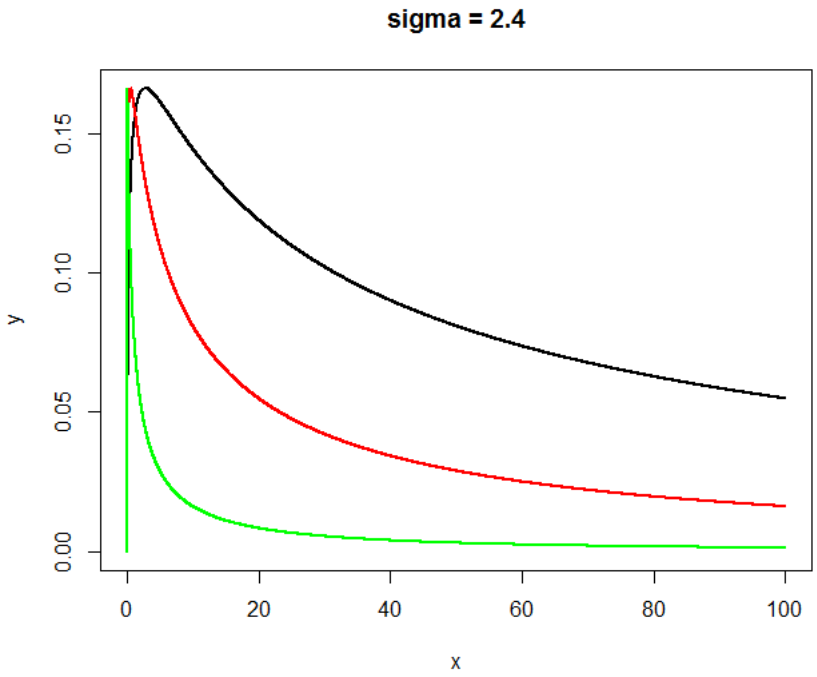

FIGURE 5.21 : Allure des distributions log-normales (densités de probabilité) pour un écart-type $\sigma = 2,4$ et des densités prédites d'1 anguille/100 m^2 (en vert), de 10 anguilles/100 m2 (en rouge) et de 50 anguilles/100 m2 (en noir). Dans un souci de clarté, les courbes sont normalisées de telle sorte que les maxima des fonctions sont tous égaux à la valeur du maximum de la courbe verte (soit 0,166 environ), dont l'aire sous la courbe vaut 1.

- Tout d'abord, s'agissant d'un carré, la variance est nécessairement un nombre positif ou nul.

- On exclut des calculs la situation limite (qui est une abstraction théorique) dans laquelle toutes les anguilles sont placées simultanément à l'exutoire d'un bassin, ce qui impliquerait une variance nulle. On exclut ainsi la borne inférieure égale à zéro.

- La variance représente une distance au carré (que nous calculons en kilomètres carrés) correspondant à l'étalement spatial des anguilles en fin de colonisation, dans la direction moyenne de leur mouvement (qui est donnée par la droite passant par la source du drain colonisé et l'exutoire du bassin). De plus, à la fin de la colonisation, tout le stock d'anguilles qui a pénétré dans un drain donné doit toujours s'y trouver. Cela se traduit par une contrainte mathématique sur l'écart-type (racine carrée de la variance). En effet, la table statistique de la loi normale centrée réduite nous indique que pour que 100 % du stock soit dans le drain, ce dernier doit avoir une longueur au moins égale à 4 σ.

En conséquence, la variance doit être telle que :

$$0 < \sigma^2 \leqslant \frac{L_{drain}^2}{16} \tag{5.2}$$

Soit un écart-type (racine carrée de la variance) tel que :

$$0 < \sigma \leqslant \frac{L_{drain}}{4} \tag{5.3}$$

On vérifie au passage que comme l'écart-type mesure une distance d'étalement, les membres de l'inégalité 5.3 sont homogènes à des longueurs.
Enfin, on peut déduire de ces considérations une valeur maximale pour l'écart-type en France métropolitaine. En effet, le drain le plus long en France est la Loire, avec une longueur de 1012 km depuis sa source jusqu'à son embouchure. L'inégalité 5.2 nous permet donc de dire qu'en France métropolitaine, la valeur maximale de σ est de 253 km. On notera que puisque $\sigma^2 = 2\,D\,t$, cela permet aussi de donner une borne maximale pour la valeur du coefficient de diffusion des anguilles en fonction de leur âge. Par exemple, pour des anguilles âgées de 2 ans, le coefficient de diffusion sera au maximum de 16 002 km^2 par an (en France) :

$$2\,D\,t \leqslant 64009\,\mathrm{km}^2 \Rightarrow D \leqslant 16002,25\,\mathrm{km}^2.\mathrm{an}^{-1}$$

Néanmoins, ces valeurs maximales sont à considérer avec précaution car elles reposent sur les hypothèses du modèle, et non sur des observations de terrain. Elles permettent cependant d'avoir un ordre de grandeur (sinon exact, du moins pas irréaliste) du mouvement des anguilles durant leur phase de colonisation des bassins.
Un autre point important est que lorsque nous considérons une variance constante quel que soit le bassin considéré, nous ne pouvons plus appliquer le raisonnement précédent car la variance optimisée est dans ce cas unique pour toutes les configurations de bassins et les longueurs de drain, et l'on ne peut plus vraiment définir de "borne maximale" par bassin.

Paramètre	**Borne min**	**Borne max**	**Valeur initiale**	**Valeur finale**	**Erreur stat. (+/-)**
logD0_1990	3.0	10.0	4.5	**7.57242**	0.442122
logD0_1991	3.0	10.0	4.5	**8.38671**	0.419264
logD0_1992	3.0	10.0	4.5	**7.63572**	0.426072
logD0_1993	3.0	10.0	4.5	**7.63846**	0.451798
logD0_1994	3.0	10.0	4.5	**7.26933**	0.428653
logD0_1995	3.0	10.0	4.5	**7.70989**	0.383472
logD0_1996	3.0	10.0	4.5	**7.33972**	0.378426
logD0_1997	3.0	10.0	4.5	**7.43925**	0.368326
logD0_1998	3.0	10.0	4.5	**7.42216**	0.376365
logD0_1999	3.0	10.0	4.5	**7.19998**	0.365114
logD0_2000	3.0	10.0	4.5	**7.56459**	0.346106
logD0_2001	3.0	10.0	4.5	**7.39594**	0.363484
logD0_2002	3.0	10.0	4.5	**7.55617**	0.363607
logD0_2003	3.0	10.0	4.5	**7.31063**	0.361318
logD0_2004	3.0	10.0	4.5	**7.26951**	0.362225
logD0_2005	3.0	10.0	4.5	**7.19399**	0.351312
logD0_2006	3.0	10.0	4.5	**7.10254**	0.352818
logD0_2007	3.0	10.0	4.5	**7.02295**	0.397691
logD0_2008	3.0	10.0	4.5	**7.05755**	0.411407
logD0_2009	3.0	10.0	4.5	**6.98882**	0.451569
effect_Sein	-3.0	3.0	0	**0.391667**	0.235175
effect_Garo	-3.0	3.0	0	**-1.07949**	0.234025
effect_Adou	-3.0	3.0	0	**-0.122936**	0.391818
effect_Rhon	-3.0	3.0	0	**-2.76997**	0.229943
effect_Bret	-3.0	3.0	0	**0.483197**	0.26191
effect_Arto	-3.0	3.0	0	**-0.046129**	0.689041
effect_Cors	-3.0	3.0	0	**0.0808307**	0.438032
σDiffusive	0.01	10000	500.0	**276.312**	14.982
logitPassab	-5	5	0	**5**	$8{,}64426.10^{-8}$
logsigma	-2.0	1.0	-0.5	**0.878676**	0.0249953
d	0.01	10.0	1.0	**0.750234**	0.0200862

TABLE 5.2 : Tableau récapitulatif des résultats de l'optimisation des paramètres du modèle TABASCO pour l'option de calcul n° 1 de la variance. L'intervalle de confiance pour le calcul de l'erreur est à 1,96 σ (95 %).

Paramètre	**Borne min**	**Borne max**	**Valeur initiale**	**Valeur finale**	**Erreur stat. (+/-)**
logD0_1990	3.0	10.0	4.5	**7.15150**	0.453646
logD0_1991	3.0	10.0	4.5	**8.13185**	0.438494
logD0_1992	3.0	10.0	4.5	**7.11669**	0.444530
logD0_1993	3.0	10.0	4.5	**7.10139**	0.458604
logD0_1994	3.0	10.0	4.5	**6.71305**	0.439247
logD0_1995	3.0	10.0	4.5	**7.00772**	0.394406
logD0_1996	3.0	10.0	4.5	**6.61484**	0.387640
logD0_1997	3.0	10.0	4.5	**6.62430**	0.375543
logD0_1998	3.0	10.0	4.5	**6.63189**	0.384966
logD0_1999	3.0	10.0	4.5	**6.39280**	0.376242
logD0_2000	3.0	10.0	4.5	**6.75869**	0.360904
logD0_2001	3.0	10.0	4.5	**6.67471**	0.373916
logD0_2002	3.0	10.0	4.5	**6.71916**	0.371256
logD0_2003	3.0	10.0	4.5	**6.57091**	0.372378
logD0_2004	3.0	10.0	4.5	**6.54303**	0.375667
logD0_2005	3.0	10.0	4.5	**6.53104**	0.369802
logD0_2006	3.0	10.0	4.5	**6.59026**	0.379149
logD0_2007	3.0	10.0	4.5	**6.49903**	0.418126
logD0_2008	3.0	10.0	4.5	**6.44843**	0.425946
logD0_2009	3.0	10.0	4.5	**6.37781**	0.468308
effect_Sein	-3.0	3.0	0	**1.28681**	0.218351
effect_Garo	-3.0	3.0	0	**-0.291026**	0.216253
effect_Adou	-3.0	3.0	0	**0.975502**	0.381283
effect_Rhon	-3.0	3.0	0	**-1.72981**	0.210322
effect_Bret	-3.0	3.0	0	**1.70169**	0.254327
effect_Arto	-3.0	3.0	0	**1.02537**	0.675071
effect_Cors	-3.0	3.0	0	**1.27328**	0.43636
σDiffusive	0.01	100	1.0	**2.94797**	0.150054
logitPassab	-10	10	0	**4.14831**	0.202038
logsigma	-2.0	1.0	-0.5	**0.856453**	0.0251586
d	0.1	10.0	1.0	**1.1282**	0.0506957

TABLE 5.3 : Tableau récapitulatif des résultats de l'optimisation des paramètres du modèle TABASCO pour l'option de calcul n° 2 de la variance. L'intervalle de confiance pour le calcul de l'erreur est à 1,96 σ (95 %).

Année	N0 (953 BV)	N0 (France)
1990	$5{,}82.10^8$	$6{,}41.10^8$
1991	$1{,}31.10^9$	$1{,}45.10^9$
1992	$6{,}20.10^8$	$6{,}83.10^8$
1993	$6{,}56.10^8$	$7{,}22.10^8$
1994	$4{,}30.10^8$	$4{,}74.10^8$
1995	$6{,}68.10^8$	$7{,}36.10^8$
1996	$4{,}62.10^8$	$5{,}08.10^8$
1997	$5{,}10.10^8$	$5{,}61.10^8$
1998	$5{,}01.10^8$	$5{,}52.10^8$
1999	$4{,}02.10^8$	$4{,}42.10^8$
2000	$5{,}78.10^8$	$6{,}36.10^8$
2001	$4{,}88.10^8$	$5{,}38.10^8$
2002	$5{,}73.10^8$	$6{,}31.10^8$
2003	$4{,}48.10^8$	$4{,}94.10^8$
2004	$4{,}30.10^8$	$4{,}74.10^8$
2005	$3{,}99.10^8$	$4{,}40.10^8$
2006	$3{,}64.10^8$	$4{,}01.10^8$
2007	$3{,}37.10^8$	$3{,}71.10^8$
2008	$3{,}48.10^8$	$3{,}83.10^8$
2009	$3{,}25.10^8$	$3{,}58.10^8$

TABLE 5.4 : Tableau récapitulatif des résultats de l'optimisation du paramètre d'intérêt N0 (stock d'anguilles) année par année, pour l'option de calcul n° 1.

Année	N0 (953 BV)	N0 (France)
1990	$6{,}89.10^8$	$7{,}59.10^8$
1991	$1{,}84.10^9$	$2{,}02.10^9$
1992	$6{,}66.10^8$	$7{,}33.10^8$
1993	$6{,}56.10^8$	$7{,}22.10^8$
1994	$4{,}45.10^8$	$4{,}90.10^8$
1995	$5{,}97.10^8$	$6{,}58.10^8$
1996	$4{,}03.10^8$	$4{,}44.10^8$
1997	$4{,}07.10^8$	$4{,}48.10^8$
1998	$4{,}10.10^8$	$4{,}52.10^8$
1999	$3{,}23.10^8$	$3{,}56.10^8$
2000	$4{,}66.10^8$	$5{,}13.10^8$
2001	$4{,}28.10^8$	$4{,}72.10^8$
2002	$4{,}48.10^8$	$4{,}93.10^8$
2003	$3{,}86.10^8$	$4{,}25.10^8$
2004	$3{,}76.10^8$	$4{,}14.10^8$
2005	$3{,}71.10^8$	$4{,}09.10^8$
2006	$3{,}94.10^8$	$4{,}33.10^8$
2007	$3{,}59.10^8$	$3{,}96.10^8$
2008	$3{,}42.10^8$	$3{,}76.10^8$
2009	$3{,}18.10^8$	$3{,}51.10^8$

TABLE 5.5 : Tableau récapitulatif des résultats de l'optimisation du paramètre d'intérêt N0 (stock d'anguilles) année par année, pour l'option de calcul n° 2.

Chapitre 6

Conclusion et perspectives

6.1 Historique du projet TABASCO

Durant la première phase du projet (2013-2014), un important travail de choix des outils informatiques (bibliothèques de calcul, algorithme de minimisation par différentiation automatique, interfaces avec la base de données et avec le système d'information géographique), puis d'intégration des données au sein de ces outils a été réalisé. L'implémentation de deux approches de calcul (approche par propagation d'une gaussienne et approche par matrice de transition) a permis de produire une première version du code, qui a été testée sur un bassin versant de référence, en l'occurrence celui de la Gironde (Garonne/Dordogne). Le concept du modèle a été présenté à la 144^{ieme} conférence annuelle de l'*American Fisheries Society* (AFS) à Québec au Canada (Lambert *et al.*, 2014).
En 2014, après une phase de correction des problèmes apparus suite à la première version du code, le modèle a été adapté pour pouvoir fonctionner sur une sélection de soixante bassins de taille importante. Une seconde période de test a débuté en juillet 2014, en particulier sur des calculs liés au paramètre sigma (écart-type) associé à la diffusion pour l'approche par propagation d'une onde gaussienne. Au vu du temps nécessaire pour chacune de ces étapes, et en particulier pour les phases de test qui sont coûteuses en temps, il a été décidé de ne continuer à travailler que sur l'approche de calcul par propagation d'une gaussienne dans un graphe orienté. Le développement de l'option par matrice de transition a été ajourné, mais il est important de noter que cette possibilité fonctionne correctement. Elle pourrait donc être exploitée de nouveau si à l'avenir cela était nécessaire.
A la fin de l'année 2014 et au début de l'année 2015, nous avons travaillé à l'adaptation du code TABASCO afin qu'il puisse prendre en compte la totalité des bassins versants de France métropolitaine dont l'exutoire est situé à l'intérieur des frontières nationales. La modèle prend maintenant en compte 953 bassins, couvrant 500 979 km^2 (soit 90,81 % de la superficie de la France métropolitaine). L'application aux cours d'eaux transfrontaliers sera possible dès que les informations nécessaires (graphes du réseau hydrographique et recensement des obstacles) seront disponibles. Nous avons également intégré une calibration multi-années (sur les 20 années de données entre 1990 et 2009) dans la modélisation, et une prise en compte d'un effet « UGA » qui permet de mieux ajuster les prédictions de densités entre unités de gestion. Toutes ces améliorations ont mené à une version stable du modèle au printemps 2015. Cette version aboutie du modèle a été présentée en juin 2015 à la conférence internationale « *Fish Passage* » à Groningen aux Pays-Bas (Domange *et al.*, 2015).
La dernière phase de développement du projet (deuxième moitié de 2015) a consisté à comparer deux options de calcul du paramètre sigma associé à la diffusion et à essayer de mener une étude sur la comparaison des mortalités par prédation des individus bloqués à la montaison et des mortalités dues aux turbines hydroélectriques à la dé-

valaison. Toutefois, l'estimation du paramètre de franchissement des obstacles à une valeur proche de 100 % n'a pas rendu possible cette comparaison des mortalités. Enfin, un travail de préparation et de normalisation des outils de diagnostic a été réalisé pour la présentation des résultats actuels et futurs.

6.2 Bilan des résultats du modèle

Nous pouvons dégager un certain nombre de conclusions sur le modèle TABASCO à l'issue de ce rapport. Nous resterons les plus factuels possible dans ce paragraphe, et nous garderons les perspectives possibles pour le dernier paragraphe. Nous distinguerons ici entre les points positifs et négatifs que l'on peut faire ressortir de notre analyse.

6.2.1 Apports du modèle

Schéma de répartition globale Le premier apport du modèle TABASCO est la prédiction des densités d'anguilles jaunes en France métropolitaine, avec une bonne cohérence vis-à-vis des données de pêche. L'étude des résidus montre en effet que les plus gros écarts sont localisés aux zones littorales de l'arc atlantique et de la Manche, ou plus ponctuellement aux sources de certains drains ou dans certaines zones montagneuses. Cependant, le patron classique de la répartition spatiale des anguilles est plutôt bien reproduit à l'échelle de la France, avec en particulier les zones à très faible densité d'anguilles qui coïncident avec celles d'absence de l'anguille dans les pêches électriques.

Détermination de l'évolution du paramètre d'intérêt $\mathbf{N_0}$ Nous avons pu estimer l'évolution temporelle du nombre total d'anguilles jaunes en France métropolitaine (paramètre N_0) pour nos deux options de calcul. Dans les deux cas, nous estimons le taux de décroissance annuelle entre 4 et 5 %, sous l'hypothèse d'une décroissance exponentielle, par ailleurs fortement suggérée par les données.

Prise en compte de la topologie Le modèle a été créé pour tenir compte de la topologie des bassins versants, et les outils informatiques mis en œuvre remplissent parfaitement cette tâche. Le réseau hydrographique est formalisé sous forme de graphe orienté, ce qui permet de réaliser les calculs en tenant compte de la forme et des dimensions de chaque bassin versant.

Détermination des intervalles de confiance sur l'estimation des paramètres Un point positif du modèle est qu'il permet de présenter les résultats de la simulation accompagnés de leur intervalle de confiance. Cela facilite considérablement la comparaison avec d'autres modèles, donne une meilleure idée de la qualité de l'estimation en fournissant des indications sur l'erreur statistique commise. En particulier, les valeurs pour le paramètre d'intérêt pour la gestion de l'anguille, à savoir le stock total d'anguilles jaunes par année, peuvent être considérées avec toutes les précautions qu'impose leur marge d'erreur.

Détermination des mortalités Nous tenons ici à souligner que, même si cette étude n'est pas possible pour le moment compte tenu des problèmes sur le paramètre de franchissement des obstacles, le modèle inclut tous les outils nécessaires au calcul des mortalités en montaison (prédation sur les animaux bloqués à l'aval des barrages) et en dévalaison (passage dans les turbines hydroélectriques). Si le problème de l'évaluation de l'indice de franchissabilité était résolu, une étude de ce type pourrait être immédiatement menée avec TABASCO.

Temps de calcul Enfin, l'une des forces de TABASCO repose sur le temps de simulation qui est très court en comparaison du nombre d'opérations effectuées. Ainsi, le modèle tourne complètement (c'est-à-dire en incluant la formalisation du réseau hydrographique, la calibration, les calculs et les imports/exports de données dans la base et vers les sorties graphiques) en moins de deux heures (typiquement 1h30 min) pour les 953 bassins actuellement considérés (donc quasiment sur la France entière). Cela rend le modèle facilement utilisable en pratique, y compris pour des phases de test pour lesquelles il est nécessaire d'enchaîner plusieurs simulations à la suite.

6.2.2 Problèmes identifiés

Paramètre σ d'observation Le principal problème qui est apparu dans notre étude est l'optimisation du paramètre σ quantifiant la dispersion de la distribution log-normale des densités observées dans un tronçon donné par rapport à la valeur théorique. Il est clairement ressorti que la valeur optimisée de ce paramètre était élevée (de l'ordre de 2,4 dans les deux options de calcul), alors que l'on attendrait plutôt une valeur proche de 0,3, ou tout du moins inférieure à 0,5. En d'autres termes, cela revient à dire que la calibration du modèle s'appuie peu sur les informations contenues dans les observations. Nous n'avons pas réussi à savoir si cela était dû à l'algorithme d'optimisation lui-même, ou bien à la façon dont nous avons implémenté le processus de diffusion dans TABASCO. De très nombreux tests ont pourtant été menés, que ce soit en modifiant la plage de variation des paramètres du modèle, ou bien leurs valeurs de départ pour l'optimisation, ou encore en fixant a priori ce paramètre σ d'observation (et donc en le retirant du processus d'optimisation). Dans tous ces essais, le processus d'optimisation menait à une valeur élevée pour ce paramètre, ou bien à une valeur plus faible mais au prix d'une log-vraisemblance très élevée en valeur absolue et/ou d'une mauvaise convergence de l'algorithme d'optimisation.

Paramètre de franchissement des obstacles Le second problème est du même type que le précédent, mais il concerne cette fois le paramètre indiquant la franchissabilité moyenne des ouvrages en France. En effet, dans tous les cas de figure testés, ce paramètre s'approche de sa borne maximale (i.e. de 100 % de franchissement). Cela ne reflète évidemment pas la réalité. Nous proposons trois explications. La première est qu'il s'agit d'un paramètre moyen, dans le sens où il ne différencie pas les ouvrages selon leur type. En effet, cette valeur moyenne peut « effacer » les effets des ouvrages les plus bloquants, qui seraient moyennés avec les effets nuls ou faibles des petits obstacles, les plus nombreux. La seconde explication est que, par construction, TABASCO ne modélise que la résultante du processus de franchissement. Ainsi, le modèle ne tient pas compte du nombre de tentatives de passage par une anguille qui, d'une part, en cohérence avec un processus diffusif, varie en fonction de la distance à la mer, et d'autre part varie également en fonction de l'âge des individus et donc indirectement de la localisation de l'obstacle dans le réseau hydrographique. Comme pour l'hypothèse précédente, cela revient à remettre en cause une franchissabilité constante dans le modèle. Enfin, une troisième hypothèse est que le nombre d'individus bloqués en aval des barrages est déjà lissé dans les données de pêches. Nous manquons effectivement de données spatialement (sur quelques dizaines de kilomètres en aval direct de chaque obstacle) et temporellement (afin d'avoir une idée de la surmortalité par prédation induite par le blocage aux obstacles par exemple), pour mesurer l'effet des barrages sur les densités d'anguille.

Effet UGA Il apparaît que nous avons un patron spatial dans la carte des résidus (sous-estimation dans les UGA Gironde et Rhône, et sur-estimation dans les autres). Actuellement, l'effet UGA se limite aux densités moyennes (le N_0). On peut imaginer l'existence de structures spatiales à des échelles plus petites (bassins versants) ou plus grandes (arc atlantique, Manche, ...). Une autre possibilité est le paramètre σ de la

diffusion (diffusivité ou durée de colonisation) soit différent d'une UGA à l'autre. Il ne nous est en tout cas pas possible de trancher ce point à ce stade.

Surestimation générale des densités Nous avons aussi observé une surestimation possible des densités prédites, notamment sur la partie du territoire la plus proche des côtes. La comparaison avec les seules autres estimations connues (modèle EDA) indiquent une surestimation du stock total d'anguilles jaunes d'un facteur 10. Cela peut paraître important, mais il suffit d'une surestimation dans les zones aval (dans lesquelles on dispose de peu d'information) pour engendrer des différences d'un facteur 10 à l'échelle nationale.

Paramètre de gestion des flux aux confluences Le dernier problème que nous pouvons faire ressortir de notre étude est la valeur du paramètre de répartition des flux aux confluences, qui est d'une part dépendante des hypothèses retenues pour la modélisation de la diffusion, et d'autre part corrélée au paramètre σ de la diffusion. Ainsi, lorsque nous avons testé l'option de calcul avec un sigma diffusif constant, le paramètre de bifurcation a pris une valeur de 0,75, donc inférieure à 1. Cela implique dans ce cas que les anguilles privilégient le drain le plus petit lorsqu'elles parviennent à une confluence. Par contre, avec un sigma diffusif qui diminue avec la distance à la source, la paramètre a été optimisé à une valeur de 1,13, supérieure à 1. Dans ce cas, les anguilles privilégieraient le plus gros drain à une confluence. On voit donc que pour deux options de calcul finalement assez proches, on obtient une interprétation opposée du comportement des anguilles aux confluences.

6.3 Perspectives pour le modèle

Nous avons identifié de nombreuses pistes d'améliorations possibles pour TABASCO que nous allons énumérer dans cette section. Nous les avons classées par ordre d'importance et de faisabilité.

- Tout d'abord et simplement, le calcul de la répartition des anguilles dans chaque tronçon (proportion d'anguilles sédentarisées) pourrait se faire non plus en fonction du linéaire de berges mais en fonction de la surface en eau, ce qui éviterait probablement un biais sur les densités prédites.

- Afin d'éviter tout problème d'accumulation anormale des anguilles au niveau des tronçons sources, et de ne pas fausser les prédictions en redistribuant des individus sur la longueur totale d'un drain, il serait très avantageux de pouvoir implémenter un effet miroir à chaque extrémité amont des drains pour rendre mieux compte des conditions aux limites amont. Ce type de calcul pourrait aussi être utilisé pour répercuter l'effet des obstacles sur la diffusion. Dans un premier temps, des simulations simples avec une marche aléatoire dans un petit drain permettraient d'explorer les différences obtenues dans le cas d'un calcul avec redistribution (tel qu'il est effectué actuellement) et avec effet miroir.

- Une modification de l'implémentation du paramètre de franchissement semble nécessaire. Une amélioration importante serait de prendre dorénavant en compte la diversité des obstacles, afin que le paramètre de franchissement soit maintenant dépendant de leur position dans le réseau hydrographique, de leur nature, de leur dimension et de leur équipement (par exemple en tenant compte du fait qu'ils sont associés ou non à un équipement de type passe à poissons).

- Le paramètre σ d'observation (lié à la distribution des densités observées dans les tronçons) pourrait être fixé a priori à une ou plusieurs valeurs prédéfinies, et non plus optimisé dans le modèle. Cela pourrait sensiblement améliorer les difficultés durant la phase d'optimisation car la valeur de ce paramètre est très

sensible pour la recherche de l'optimum de la fonction de vraisemblance. Enfin, dans ce premier lot d'améliorations, il serait pertinent d'utiliser un algorithme de recherche globale avant l'algorithme de minimisation pour améliorer la recherche de l'optimum, et guider les plages de variation et les valeurs initiales des paramètres optimisés.

- Dans un deuxième temps, le calcul de l'écart-type de la distribution gaussienne étant un point critique de la modélisation, une étude complémentaire sur les hypothèses à choisir et la formule de calcul à utiliser serait très profitable. C'est en effet ce paramètre qui contrôle l'essentiel de la façon dont la diffusion s'effectue dans les bassins versants. Toutefois, compte tenu de notre expérience, ce travail ne sera pas simple.

- Afin de compléter la prise en compte de la topologie des bassins versants, il serait avantageux de tenir compte de l'altitude. Dans un esprit de « dépense énergétique », on pourrait par exemple faire diminuer le σ de la gaussienne avec l'altitude.

- Par ailleurs, une évolution qui ne pose pas de problème technique particulier consiste à appliquer le modèle sur les abondances d'anguilles par classes de taille (par exemple en cm :]15,30],]30,45] et]45,+∞[). Les prédictions du modèle permettraient alors de calculer des stocks en termes de biomasses et non plus seulement en termes de nombres ou de densités. Cela permettrait également d'aligner les fonctionnalités du modèle sur les derniers développement d'EDA.

- A moyen terme, une évaluation de la fiabilité et de la robustesse de TABASCO pourrait être effectuée de plusieurs manières : en utilisant des données issues du modèle opératoire CREPE (Lambert, 2012), et en définissant une procédure systématique de comparaison des modèles (basée par exemple sur une comparaison de l'AIC). En particulier, la sensibilité de la réponse de TABASCO à différents niveaux d'abondance totale pourrait être testée pour mesurer la capacité du modèle à détecter une amélioration ou une dégradation du stock d'anguilles dans un hydrosystème.

- Enfin, si, à l'avenir, l'approche matricielle (par matrices de transition) était de nouveau développée, il faudrait travailler, si cela s'avérait possible (complexité non négligeable), sur l'utilisation d'exposants réels pour la matrice stochastique dans l'approche matricielle (actuellement, on rappelle que les exposants sont des entiers). Ce serait en effet la seule façon de pouvoir définir des intervalles de confiance pour les sorties du modèle avec ce type d'approche de calcul.

Chapitre 7

Annexes

7.1 Liste par surface décroissante des principaux bassins versants intégrés dans la modélisation

Chaque nom de bassin versant est suivi de la mer ou de l'océan dans lequel se jette le fleuve qui lui est associé, ou dans lequel il se vide au niveau de son exutoire (dans le cas d'étangs par exemple), ainsi que sa superficie totale. Seuls les bassins dont la superficie dépasse 1500 km^2 sont mentionnés.

1. Loire (Océan Atlantique, 117 032 km^2).
2. Rhône (Mer Méditerranée, 89 521 km^2).
3. Gironde (Océan Atlantique, 80 080 km^2).
4. Seine (Manche, 76 119 km^2).
5. Adour (Océan Atlantique, 16 825 km^2).
6. Vilaine (Océan Atlantique, 10 485 km^2).
7. Charente (Océan Atlantique, 9 831 km^2).
8. Somme (Manche, 6 150 km^2).
9. Étang de Berre (Mer Méditerranée, 5 776 km^2).
10. Sèvre Niortaise (Océan Atlantique, 3 075 km^2).
11. Orne (Manche, 2 858 km^2).
12. Lez (Mer Méditerranée, 2 851 km^2).
13. Var (Mer Méditerranée, 2 828 km^2).
14. Argens (Mer Méditerranée, 2 756 km^2).
15. Hérault (Mer Méditerranée, 2 682 km^2).
16. Blavet (Océan Atlantique, 2 605 km^2).
17. Leyre (Océan Atlantique, 2 133 km^2).
18. Vire (Manche, 1 972 km^2).
19. Lay (Océan Atlantique, 1 931 km^2).

20. Canal de Nantes à Brest (Océan Atlantique, 1 880 km^2).)
21. Dives (Manche, 1 770 km^2).
22. Chenal de Caronte (Mer Méditerranée, 1 633 km^2).
23. Orb (Mer Méditerranée, 1 559 km^2).

7.2 Tutoriel d'installation des fichiers non pré-compilés de la bibliothèque graphique Boost

Prérequis :

1. Télécharger (par exemple en allant sur http://sourceforge.net/projects/mingw/files/) et installer MinGW, de préférence à la racine d'un disque local (ce qui donne par exemple le chemin d'accès "C:\MinGW" si on choisit de l'installer dans le répertoire "MinGW" à la racine du disque local C:).
2. Télécharger la dernière version stable de Boost sur http://www.boost.org/users/download/.
3. Télécharger la dernière version de PostgreSQL sur http://www.postgresql.org/download/.

Installation des bibliothèques non pré-compilées de Boost :

Ouvrir l'invite de commande de Windows (Sous Windows 7, aller dans Tous les programmes→Accessoires→Invite de Commande). Si vous n'êtes pas administrateur de votre machine, ouvrez l'invite en mode administrateur (clic droit, et "Exécuter en tant qu'administrateur"). L'utilisateur doit maintenant choisir parmi les méthodes décrites ci-après celle qui fonctionnera sur sa machine et avec sa configuration logicielle.

Méthode 1

1. Dans l'invite de commande, déplacez-vous dans le répertoire de téléchargement de Boost, puis dans "tools\build\v2\engine". Exemple de commande à taper :
   cd boost_1_55_0\tools\build\v2\engine
2. Garder l'invite de commande ouverte. A l'aide d'un éditeur de texte (idéalement Notepad++), ouvrir le fichier "build.bat". Dans la routine intitulée " :Guess_Toolset", trouver la structure contenant le mot-clef "mingw". Éditer le fichier en conséquence (avec le bon chemin d'accès vers votre répertoire d'installation de MinGW, et effacer toutes les autres structures similaires. Au final, votre routine " :Guess_Toolset" doit contenir quelque-chose comme :
   :Guess_Toolset REM Try and guess the toolset to bootstrap the build with...
   REM Sets BOOST_JAM_TOOLSET to the first found toolset.
   REM May also set BOOST_JAM_TOOLSET_ROOT to the REM location of the found toolset.

   call :Clear_Error
   call :Test_Empty %ProgramFiles%
   if not errorlevel 1 set ProgramFiles=C :\Program Files

   if EXIST "C :\MinGW\bin\gcc.exe" (
   set "BOOST_JAM_TOOLSET=mingw"

```
set "BOOST_JAM_TOOLSET_ROOT=C :\MinGW\"
goto :eof)
call :Clear_Error
if NOT "_%CWFolder%_" == "__" (
set "BOOST_JAM_TOOLSET=metrowerks"
set "BOOST_JAM_TOOLSET_ROOT=%CWFolder%¨
goto :eof )
call :Clear_Error
call :Test_Path mwcc.exe
if not errorlevel 1 (
set "BOOST_JAM_TOOLSET=metrowerks"
set "BOOST_JAM_TOOLSET_ROOT=%FOUND_PATH%..\..\"
goto :eof)
call :Clear_Error
call :Error_Print "Could not find a suitable toolset."
goto :eof
```

3. Dans l'invite de commande, vérifiez que vous êtes toujours dans le sous-répertoire "\tools\build\v2\engine", puis tapez la commande "build.bat".Cela va normalement créer le fichier "bjam.exe" pour MinGW/gcc.

4. Ajouter le chemin d'accès vers ce fichier dans la variable d'environnement PATH de Windows (au début de la variable, et séparé des autres chemins par un point-virgule). En principe, cela doit ressembler à :
"C :\boost_1_55_0\tools\build\v2\engine\bin.ntx86 ;"

5. A la racine de votre répertoire d'installation de Boost, taper alors dans l'invite de commande :
bjam –toolset=gcc –layout=system –with-thread install

En cas de difficultés, quelques explications additionnelles sont disponibles sur ce forum : http://www.developpez.net/forums/d1096176/c-cpp/cpp/bibliotheques/boost/aide-installation-boost-version-1-46-1-a/.

Méthode 2

1. Allez dans le répertoire "tools\build\v2\"

2. Tapez la commande : bootstrap.bat gcc

3. Tapez la commande : b2 install –prefix=PREFIX (sans espace entre les deux signes moins, et où PREFIX est le répertoire dans lequel vous souhaitez que Boost.Build soit installé).

4. Ajoutez "PREFIX\bin" dans votre variable d'environnement PATH.

Méthode 2 bis

1. Choisir un répertoire d'installation. Boost.Build placera tous les fichiers intermédiaires qu'il génère lors de l'installation dans le répertoire d'installation. Si votre répertoire racine de Boost est accessible en écriture, cette étape n'est pas strictement nécessaire : par défaut, Boost.Build créera pour cela un sous-répertoire "bin.v2\" au sein de votre répertoire de travail courant.

2. Invoquer b2. Changer votre répertoire courant en vous plaçant dans le répertoire racine de Boost, et invoquer b2 en tapant dans votre invite de commandes :

b2 –build-dir=build-directory toolset=toolset-name –build-type=complete stage
Pour une description complète de cette utilisation et des différentes options, consultez la documentation Boost.Build.
Votre session devrait ressembler à cela :
%C :\WINDOWS> cd C :\Program Files\boost\boost_1_55_0
%C :\Program Files\boost\boost_1_55_0> b2 ^
%More ? –build-dir="C :\Documents and Settings\dave\build-boost" ^
%More ? –build-type=complete msvc stage
L'option "–build-type=complete" fera que Boost.Build installera toutes les versions supportées des différentes bibliothèques. Pour savoir comment n'installer que certaines versions spécifiques, référez-vous à l'aide en ligne de Boost.Build.

Infos supplémentaires (en anglais) :
Building the special stage target places Boost library binaries in the stage\lib\subdirectory of the Boost tree. To use a different directory pass the –stagedir=directory option to b2.
Note : b2 is case-sensitive ; it is important that all the parts shown in bold type above be entirely lower-case.
For a description of other options you can pass when invoking b2, type :
b2 –help
In particular, to limit the amount of time spent building, you may be interested in :

- reviewing the list of library names with –show-libraries
- limiting which libraries get built with the –with-library-name or –without-library-name options
- choosing a specific build variant by adding release or debug to the command line.

Note : Boost.Build can produce a great deal of output, which can make it easy to miss problems. If you want to make sure everything is went well, you might redirect the output into a file by appending ">build.log 2>&1" to your command line.
Expected Build Output
During the process of building Boost libraries, you can expect to see some messages printed on the console. These may include :

- Notices about Boost library configuration—for example, the Regex library outputs a mes-sage about ICU when built without Unicode support, and the Python library may be skipped without error (but with a notice) if you don't have Python installed.
- Messages from the build tool that report the number of targets that were built or skipped. Don't be surprised if those numbers don't make any sense to you ; there are many targets per library.
- Build action messages describing what the tool is doing, which look something like :
- toolset-name.c++ long/path/to/file/being/built
- Compiler warnings.

7.3 Tutoriel d'installation de la bibliothèque libpqxx

Le présent tutoriel indique comment installer et compiler la bibliothèque libpqxx sur une machine Windows.

1. Installer la dernière version de PostgreSQL sur votre ordinateur. Cela ne pose aucun problème particulier en utilisant l'exécutable pgInstaller.

2. Après l'installation, vous devez pouvoir trouver dans votre répertoire d'installation le sous-répertoire *lib* qui contient la bibliothèque libpq. Exemple de chemin d'accès : "C :\Program Files\PostgreSQL\Version\lib".

3. Télécharger la dernière version de lipqxx, et décompresser les fichiers dans le répertoire de votre choix. Ensuite, il faut préparer la compilation de libpqxx (ce qui suppose d'avoir déjà sur votre machine la bibliothèque libpq sous la forme de ses fichiers binaires et de ses fichiers d'en-tête). Pour cela, copiez le fichier "win32\common-sample", renommez-le en " win32\common", et éditez-le pour qu'il tienne compte des chemins d'accès à vos fichiers « include » et « lib » de PostgreSQL. Vérifiez pour cela que la variable PGSQLSRC pointe bien vers votre installation de PostgreSQL, puis suivez les instructions du fichier "win32\common" pour commenter ou décommenter certaines lignes. Ensuite, vous devez créer les fichiers d'en-tête de configuration, de la forme "include\pqxx\config-*-*.h". Le plus simple pour cela est de copier les fichiers-types dans "config\
sample-headers". Dans ce dernier répertoire, trouvez les sous-répertoires correspondant le mieux à votre compilateur et à votre version de la bibliothèque libpq respectivement. Prenez les fichiers config-*-*.h dans ces répertoires, et copiez-les dans le répertoire "include\pqxx\".

4. Il reste à compiler et installer la bibliothèque lipqxx. Dans cette optique, installer (si ce n'est pas déjà fait) MinGW sur votre machine (directement à la racine du disque, par exemple dans "C :\MinGW"). Ensuite, installer MSYS (mais attention, pas dans l'arborescence de MinGW, par exemple dans "C :\MSYS"). Puis, lancer MSYS (aller dans Démarrer → Tous les Programmes → MinGW → MSYS → msys). Dans l'invite de commande de MSYS, aller dans le répertoire principal d'installation de lipqxx (celui dans lequel vous avez décompressé les fichiers), puis taper :
export LDFLAGS=-lws2_32.

5. Vérifiez que le chemin vers le répertoire "\bin" de MinGW est bien dans votre variable d'environnement PATH (le rajouter au besoin).

6. Tapez enfin :
./configure –prefix="C :/MinGW/local" –enable-static && make &&
make install

Il y a alors de grandes chances pour que la compilation ne fonctionne pas. Pas d'inquiétude, cela est normal sous Windows, et vient du fait qu'il faut éditer manuellement les fichiers d'en-tête, ce qui conduit inexorablement à des erreurs lors de la compilation de la bibliothèque (en général, il s'agit d'erreurs spécifiant des fichiers d'en-tête introuvables ou manquants). Une erreur typique indique ainsi que le fichier d'en-tête <sys/select.h> est manquant. Vous trouverez alors une variable de configuration nommée PQXX_HAVE_SYS_SELECT_H, qui ne doit pas être définie si votre système d'exploitation ne possède pas de fichier sys/select.h (ce qui est le cas sous Windows). Dans ce cas, supprimez-la ou supprimez les lignes définissant cette variable dans le fichier d'en-tête de configuration (a priori le fichier config-internal-compiler.h), et retentez une compilation. L'erreur aura normalement disparu, et procédez de la même façon pour d'autres nouvelles erreurs qui apparaîtraient. Il n'est pas anormal de supprimer plusieurs lignes dans le fichier de configuration pour que tout fonctionne au final !

7.4 Script R pour l'affichage graphique des sorties

SCRIPT DE GESTION DES SORTIES GRAPHIQUES DU MODELE TABASCO
(© Jocelyn Domange, Hilaire Drouineau and Patrick Lambert / IRSTEA Bordeaux / Janvier 2015)

```
    # Nettoyage de la mémoire
rm(list=ls())

    # Suppression des graphiques déjà ouverts
graphics.off()

    # Chargement des bibliothèques potentiellement utiles
library(RPostgreSQL)
library(rgeos)
library(sp)
library(squash)
library(RColorBrewer)
library(classInt)
library(maptools)
library(graphics)
library(Rlab)
library(gplots)
library(robustbase)
library(outliers)

    # Fonction qui prend le mot de passe de connexion à la base de données
getPass=function(){
print("password : ")
pass=scan(n=1,what=character(),quiet=TRUE)
cat(" 014 ")
return(pass)
}

    # Connexion à la base de données
m <- dbDriver("PostgreSQL")
con <-dbConnect(m,host="xx.xx.fr",port=XXXX,dbname="xxxx",
user="xxxx.xxxx",password=getPass())

    # Récupération des résultats de la modélisation
print(Sys.time())
res=dbGetQuery(con,"SELECT *, ROW_NUMBER() OVER () AS position, st_astext(geom)
wkt FROM tabasco_hilaire_Jo.edge LEFT JOIN public.edge_att ON edge.edge_id=edge_att.edge_id
LEFT JOIN tabasco_hilaire_jo.fishing_op ON edge.edge_id=fishing_op.edge_id WHERE
year=2009 AND edge.edge_id <> 41850 AND edge.edge_id <> 111913")
print(Sys.time())

    # Création des lignes géoréférencées avec attributs
row.names(res)=res$edge_ id
tot=do.call("rbind",mapply(readWKT,res$wkt,res$edge_ id))
res_ final=SpatialLinesDataFrame(tot, res[-length(row.names)])
print(Sys.time())

    # Tracé et sauvegarde de la carte
#x11(height=16/2.54,width=16/2.54)
graphics.off()
jpeg("C :/Mes_ programmes/carte_ densite_ 60BV.jpeg",height=16/2.54,
width=16/2.54,units="in",res=300)
par(mar=c(2.5,2.5,.5,13),mgp=c(1.5,.5,0))
```

```
    nombre_ classe_ couleur=5
bornes=pretty(c(0,max(log(res$density))),n=5, min.n=5)
couleurs=rev(heat.colors(length(bornes)-1))

    plot(res_ final,col=(sapply(res_ final$density,getcolor)))

    getcolor=function(x) {
if (log(x)<=bornes[1]) {
return(couleurs[1])
} else{
return(couleurs[max(which(log(x)>bornes))])
}
}

    box()
par(xpd=NA)

    text(grconvertX(1,"npc"),grconvertY(.75,"npc"),"densité",adj=c(0,0))
legend(grconvertX(1,"npc"),grconvertY(.75,"npc"), c(paste("<",
round(exp(bornes[2])),sep=""),paste(round(exp(bornes[2 :(length(bornes)-1)])),
round(exp(bornes[3 :(length(bornes))])),sep="-")), lty=rep(1,length(couleurs)),
col=couleurs,bty="n",xjust=0,yjust=1)
dev.off()
print(Sys.time())
```

AUTRES OPTIONS D'AFFICHAGE

```
    # color = function(x)rev(heat.colors(x))
# color = function(x)heat.colors(x)
mycol <- colorRampPalette(c("lightcyan", "blue4"))
plot(res_ final)
plot(res_ final, col=res_ final$number)
plot(res_ final, col=res_ final$number, lwd=res_ final$strahler)
map<-makecmap(res_ final$number, breaks = c(0,100,10000,1000000,
100000000,10000000000), colFn = mycol)
pl.color<-cmap(res_ final$number, map = map)
plot(res_ final, col=pl.color, lwd=res_ final$strahler, main="Nombre d'anguilles par
tronçon")
mycol <- colorRampPalette(c("yellow", "red4"))
map<-makecmap(res_ final$concentration, breaks = c(0,100,10000,1000000,
100000000,1000000000000000), colFn = mycol)
pl.color<-cmap(res_ final$concentration, map = map)
plot(res_ final, col=pl.color, lwd=res_ final$strahler/2)
print(Sys.time())

    # discrétisation en 7 classes (quantiles)
distr <- classIntervals(res$concentration,7,style="quantile",dataPrecision=5)$brks

    # choix d'une gamme de couleurs
# pour voir les palettes disponibles : display.brewer.all()
colours <- brewer.pal(7,"YlOrRd")

    # optionnel - codes des couleurs utilisées
colours
```

```
    # attribution des couleurs aux régions
colMap <- colours[(findInterval(res$concentration,distr,all.inside=TRUE))]

    # Affichage de la carte
par(xpd=T, mar=par()$mar+c(6,0,0,0))
plot(res_ final, col=colMap, lwd=res_ final$strahler/2)

    # affichage de la légende
# legend("bottom", inset=c(-0.05,0),legend=leglabs(distr,under="Inférieur à",
over="Supérieur à", between="-"), fill=colours, bty="n",
title="Densité surfacique d'anguilles par tronçon", cex=0.5,
pt.cex=0.5)
legend(x=c(2.0, 2.0), y=c(2.0, 2.0), legend=leglabs(distr),
fill=colours, bty="n", title="Densité surfacique d'anguilles par tronçon")
legend(locator(1), legend=leglabs(distr), fill=colours, bty="n",
title="Densité surfacique d'anguilles par tronçon", cex=0.7,
pt.cex=0.7)
# l'introduction de la chaine de caractère «  n » entraine un saut de ligne dans
# le texte à afficher

    # titre et sous titres
title(main="Prédiction par le modèle TABASCO de la densité surfacique d'anguilles
par tronçon ",
sub="auteurs : Jocelyn Domange, Hilaire Drouineau, Patrick Lambert, IRSTEA Bor-
deaux (équipe EABX/PMA)"
```

7.5 Programme de calcul du nombre total d'anguilles jaunes

Ce petit programme, codé en C++, permet d'obtenir la valeur du paramètre d'intérêt N0 sur l'ensemble des 953 bassins pris en compte dans le modèle, et par extrapolation sur la France entière, à partir d'un fichier texte d'entrée contenant les 7 paramètres rendant compte de la variabilité des densités en fonction de l'unité de gestion anguille, les 20 paramètres donnant la variabilité des densités au cours du temps entre 1990 et 2009, et les 27 erreurs statistiques associées (20 + 7). Il calcule donc aussi l'erreur statistique (intervalle de confiance) sur la détermination de ce paramètre d'intérêt pour la gestion, pour chacune des années.

```
# include <iostream>
# include <fstream>
# include <math.h>

using namespace std;

int main()
{

fstream fichier("N0_ input_ option2.txt", ios::in | ios::out | ios::ate);
if(fichier)
{
int i,j,k=0;
double effect[7];
double N0perYear[20];
```

```
double erreur[27];
double erreur_fin[8];
for(i=0;i<7;i++){
fichier>>effect[i];
}
for(j=0;j<20;j++){
fichier»N0perYear[j];
for(k=0;k<27;k++){
fichier»erreur[k];
}
}
for(int j=0;j<20;j++){double N0_ total = exp(N0perYear[j])*127544
+exp(N0perYear[j]+effect[0])*93810.362+exp(N0perYear[j]+effect[1])*96264.188
+(N0perYear[j]+effect[2])*20007.265+(N0perYear[j]+effect[3])*117838.883
+(N0perYear[j]+effect[4])*27991.965+(N0perYear[j]+effect[5])*11385.668+(N0perYear[j]+effect[6])*7349.106;
cout«"Le N0 total pour l'annee "«1990+j«" est de : "
«N0_ total«endl;
cout«"Le N0 total pour l'annee "«1990+j«" sur toute la France est de : "
«N0_ total*100/90.8073«endl;
fichier«"Le N0 total pour l'annee "«1990+j«" est de : "«N0_ total«endl;
erreur_fin[0] = 127544*exp(N0perYear[j])*(exp(erreur[j])-exp(-erreur[j]));
erreur_fin[1] = 93810.362*exp(N0perYear[j]+effect[0])*(erreur[j]+erreur[20]);
erreur_fin[2] = 96264.188*exp(N0perYear[j]+effect[1])*(erreur[j]+erreur[21]);
erreur_fin[3] = 20007.265*exp(N0perYear[j]+effect[2])*(erreur[j]+erreur[22]);
erreur_fin[4] = 117838.883*exp(N0perYear[j]+effect[3])*(erreur[j]+erreur[23]);
erreur_fin[5] = 27991.965*exp(N0perYear[j]+effect[4])*(erreur[j]+erreur[24]);
erreur_fin[6] = 11385.668*exp(N0perYear[j]+effect[5])*(erreur[j]+erreur[25]);
erreur_fin[7] = 7349.106*exp(N0perYear[j]+effect[6])*(erreur[j]+erreur[26]);
double erreur_N0 = (erreur_fin[0]+erreur_fin[1]+erreur_fin[2]+erreur_fin[3]+erreur_fin[4]+erreur_fin[5]+erreu
cout«"L'erreur sur le N0 total pour l'annee "«1990+j«" est de : "«erreur_N0 «endl;
fichier«"L'erreur sur le N0 total pour l'annee"«1990+j«" est de : "«erreur_N0 «endl;
N0_ total=0;erreur_N0=0;
erreur_fin[0] = erreur_fin[1] = erreur_fin[2] = erreur_fin[3] = erreur_fin[4] = er-
reur_fin[5] = erreur_fin[6] = erreur_fin[7] = 0;
}
fichier.close();
}

else cerr « "Erreur a l'ouverture!" « endl;

return 0;

}
```

7.6 Matrice de corrélation

Nous affichons dans cette section la matrice de corrélation des 31 paramètres du modèle pour l'option de calcul n° 2 (la matrice de corrélation pour l'option de calcul n° 1 n'apportant rien de plus, nous ne la présentons pas).

FIGURE 7.1 : Matrice de corrélation des 31 paramètres du modèle pour l'option de calcul n° 2. La matrice étant symétrique, nous ne présentons que sa moitié inférieure par souci de clarté.

Bibliographie

Hirotogu Akaike : Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Inter. Symposium on Information Theory*, pages 267 – 281, 1973.

Hirotogu Akaike : A new look at the statistical model identification. *EEE Transaction on Automatic Control*, 19(6):716–723, 1974.

Anonyme : Plan de gestion anguille de la france, rapport de mise en oeuvre - juin 2012. Rapport technique, Ministère de l'écologie, de l'énergie, du développement durable et de la mer / Ministère de l'alimentation, de l'agriculture et de la pêche / Onema, 2012.

Anonyme : Plan de gestion anguille de la France. Rapport de mise en oeuvre - juin 2015. Rapport technique, Ministère de l'écologie, de l'énergie, du développement durable et de la mer / Ministère de l'alimentation, de l'agriculture et de la pêche / Onema, 2015.

M. W. Aprahamian, A. M. Walker, B. Williams, A. Bark et B. Knights : On the application of models of European eel (*Anguilla anguilla*) production and escapement to the development of Eel Management Plans : the River Severn. *ICES Journal of Marine Science*, 64:1472–1482, 2007.

C. Briand, L. Beaulaton, P. M. Chapon, H. Drouineau et P. Lambert : Eel density analysis (EDA 2.2), Estimation de l'échappemement en anguilles argentées (Anguilla anguilla) en France, Rapport 2015. Rapport technique, Institut d'Aménagement de la Vilaine, ONEMA, Irstea Bordeaux, June 2015.

J. Domange, H. Drouineau, C. Briand, L. Beaulaton et P. Lambert : Tabasco (spatialized anguilla basin colonisation assessment model) : A diffusion model within an hydrographic network to estimate yellow eels densities. *In International conference on river connectivity best practices and innovations*, 2015.

H. Drouineau, C. Briand, P. Lambert et L. Beaulaton : GEREM (Glass Eel Recruitment Estimation Model) : A model to estimate glass eel recruitment at different spatial scales. *Fisheries Research*, 174:68–80, February 2016.

R. A. Fisher : On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, 222:309–368, 1922.

D. A. Fournier, J. R. Sibert, J. Majkowski et J. Hampton : Multifan a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (thunnus maccoyii). *Can*, 47:301–317, 1990.

David A. Fournier, Hans J. Skaug, Johnoel Ancheta, James Ianelli, Arni Magnusson, Mark N. Maunder, Anders Nielsen et John Sibert : AD Model Builder : using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 2012.

T. J. Hastie et R. J. Tibshirani : *Generalized Additive Models*, volume 43 de *Monogrpahs on Statistics and applied Probability*. CRC Press, 1990.

Anton Ibbotson, Jim Smith, Peter Scarlett et Miran Aprahamian : Colonisation of freshwater habitats by the European eel *Anguilla anguilla*. *Freshwater Biology*, 47(9):1696–1706, September 2002.

C. Jouanin, P. Gomes, C. Briand, V. Berger, F. Bau, H. Drouineau, P. Baran, P. Lambert et L. Beaulaton : Évaluation des mortalités d'anguilles induites par les ouvrages hydroélectriques en France. Projet SEA-HOPE (Silver Eels escApment From HydrOPowEr). Convention ONEMA-Irstea. Rapport final., 2012a.

Céline Jouanin, Cédric Briand, Laurent Beaulaton et Patrick Lambert : Eel Density Analysis (EDA 2.x), Un modèle statistique pour estimer l'échappement des anguilles argentées (Anguilla anguilla) dans un réseau hydrographique, Rapport Final. Rapport technique, Irstea Bordeaux, Institut d'Aménagement de la Vilaine, ONEMA, 2012b.

P. Lambert, H. Drouineau, C. Briand et L. Beaulaton : Yellow eel stock assessment oriented by the colonisation process to evaluate impacts of obstacles. *In International Eel Symposium 2014 : Are Eels Climbing Back up the Slippery Slope ?, Part 6*, 2014.

Patrick Lambert : Développement d'outils de modélisation de la population d'anguille européenne prenant en compte la diversité des paramètres de dynamique par grande fraction d'aire de répartition continentale de l'espèce. Rapport technique, Irstea Bordeaux, ONEMA, 2012.

Patrick Lambert et Eric Rochard : Identification of the inland population dynamics of the european eel using pattern-oriented modelling. *Ecological Modelling*, 206:166–178, 2007.

Patrick Lambert, Guy Verreault, Brigitte Lévesque, Valérie Tremblay, Jean-Denis Dutil et Pierre Dumont : Détermination de l'impact des barrages sur l'accès de l'anguille d'Amérique (Anguilla rostrata) aux habitats d'eau douce et établissement de priorités pour des gains en habitat. Rapport technique, Institut Maurice-Lamontagne, 2011.

Andre G. Laurent : The log-normal distribution and the translation method : description and estimation problems. *Journal of the American Statistical Association*, 58:231–235, 1963.

L. P. Lefkovitch : The study of population growth in organisms grouped by stages. *Biometrics*, 21(1):1–18, March 1965.

Aurélie Léonard et Pascale Zegel : Référentiel des obstacles à l'écoulement, Version 1, Descriptif de contenu. Rapport technique, ONEMA, 2010.

Gaëlle Leprévost : Développement d'un indicateur pour caractériser l'impact migratoire sur le stock d'anguille européenne à l'échelle des bassins. Rapport technique, ONEMA / IAV, 2007.

P. H. Leslie : Some further notes on the use of matrices in population mathematics. *Biometrika*, 35(3 and 4):213–245, December 1948.

Simone Libralato, Villy Christensen et Daniel Pauly : A method for identifying keystone species in food web models. *Ecological Modelling*, 195:153–171, 2006.

M. Lifschitz : *Gaussian Random Functions*. Kluver Academic Publishers, 1995.

P. T. Manders : A transition matrix model of the population dynamics of the clanwilliam cedar (widdringtonia cedarbergensis) in natural stands subject to fire. *Forest Ecology and Management*, 20:171–186, 1987.

Hervé Pella, Jérôme Lejot, Nicolas Lamouroux et Ton Snelder : Le réseau hydrographique théorique (RHT) français et ses attributs environnementaux. *Géomorphologie : relief, processus, environnement*, 2012.

Michael J. D. Powell : The bobyqa algorithm for bound constrained optimization without derivatives. Rapport technique, Department of Applied Mathematics and Theoretical Physics, Cambridge University, 2009.

J. Radinger et C. Wolter : Disentangling the effects of habitat suitability, dispersal, and fragmentation on the distribution of river fishes. *Ecological Applications*, 25 (4):914–927, 2015.

Louis B. Rall : *Automatic Differentiation : Techniques and Applications*, volume 120 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 1981.

Jeremy G. Siek, Lie-Quan Lee et Andrew Lumsdaine : *The Boost Graph Library, User Guide and Reference Manual*. C++ In-Depth. Addison-Wesley / Pearson Education, 2002.

Gunnar Stefánsson et Ólafur K. Pálsson : Statistical evaluation and modelling of the stomach contents of icelandic cod (gadus morhua). *Can. J. Fish. Aquat. Sci.*, 54 (1):169–181, 1996.

P. Steinbach : Expertise de la franchissabilité des ouvrages hydrauliques transversaux par l'anguille dans le sens de la montaison. Rapport technique, Conseil Supérieur de la Pêche, 2006.

A. Viallefont, J.-D. Lebreton, A.-M. Reboulet et G. Gory : Parameter identificability and model selection in capture-recapture models. *Biometrical Journal*, 40 (3):313–325, 1998.

A.M. Walker, E. Andonegi, P. Apostolaki, M. Aprahamian, L. Beaulaton, P. Bevacqua, C. Briand, A. Cannas, E. De Eyto, W. Dekker, G. De Leo, E. Diaz, P. Doering-Arjes, E. Fladung, C. Jouanin, P. Lambert, R. Poole, R Oeberst et M Schiavina : Pilot projects to estimate potential and actual escapement of silver eel. Rapport technique, The European Commission Directorate-General for Maritime Affairs and Fisheries, 2011.

Larry Wasserman : *All of Statistics : A Concise Course in Statistical Inference*. Springer-Verlag, september 2004.

J. Angus Webb et Mark Padgham : How does network structure and complexity in river systems affect population abundance and persistence ? *Limnologica*, 43:399–403, 2013.

J.A. Webb et M. Padgham : Patterns of dispersal through stream networks respond simply to multiple structural modifications. *In 18th World IMACS / MODSIM Congress, Cairns, Australia*, pages 1809–1815, 2009.