# Analyse de l'évolution de la répartition de la population alsacienne de Pie-grièche écorcheur d'après des données non protocolées issues de VisioNature

Cheikh Moustapha Diakhate

**Université de Strasbourg**
**UFR de Mathématique Informatique**

# MASTER
# Mention Mathématiques et Applications
# Spécialité Statistique

## Rapport de Stage

**Cheikh Moustapha DIAKHATE**

# Analyse de l'évolution de la répartition de la population alsacienne de Pie-grièche écorcheur d'après des données non protocolées issues de VisioNature

**Année 2013**

**Stage réalisé à l'Office des Données Naturalistes d'Alsace (ODONAT), Strasbourg**
**Sous la responsabilité de Mme Stéphanie KAEMPF - Chef de Projet.**

**Encadreur à l'UDS : M. Nicolas POULIN.**

## REMERCIEMENTS

Je voudrais adresser mes remerciements à mes deux encadreurs, en l'occurrence Mme Stéphanie KAEMPF et M. Raynald Moratin, qui en plus de n'avoir ménagé aucun effort pour m'aider à la réussite de cette mission, m'ont avant tout accordé leur confiance et donné l'opportunité d'effectuer ce stage au sein de leur structure. Cela a été une belle expérience d'appliquer mes compétences statistiques dans un domaine aussi rare et intéressant qu'est le suivi et la sauvegarde des oiseaux. Je n'oublie pas l'ensemble de l'équipe d'ODONAT qui, sans exception, a œuvré dans la réalisation de ce travail, de par son apport mais surtout de par sa sympathie.

Mes remerciements vont également à l'encontre de mes parents, de mes amis et proches et de ma famille entière sans qui rien ne serait possible. Surtout ma sœur Fanta DIAKHATE, qui n'a cessé de tout mettre en œuvre pour un succès total dans mes études.

## INTRODUCTION

### Les oiseaux : indicateurs de la diversité biologique

Au sommet d'un réseau trophique et pouvant être détectés facilement, les oiseaux ont été choisis comme indicateurs de la biodiversité. En effet, en fonction du contexte local, la composition d'un peuplement d'oiseaux donne dans une certaine mesure, une indication générale sur la qualité biologique d'un site. En dénombrant chaque année les oiseaux présents dans les différents milieux, le suivi permet de connaitre l'évolution des peuplements qui peut alors être mise en relation avec l'état des milieux.

La Pie-grièche écorcheur est une espèce essentiellement caractéristique des milieux campagnards riches, constitués de haies, d'herbages et de milieux semi-ouverts qui lui offrent un large champ de vision pour assurer son évolution et sa prédation. Sa disparition d'un site donné équivaut donc à un appauvrissement de l'écosystème de ces milieux ruraux traditionnels.

Le programme « Suivi Biodiversité en Alsace (SIBA) » a conclu à une régression globale de la population de l'espèce en Alsace entre 2005 et 2012 (ODONAT, Février 2012) à partir de données protocolées relatives à la Pie-grièche écorcheur. C'est-à-dire que les indicateurs d'évolution des effectifs de l'espèce ont été calculés à partir de données recueillies après la mise en place de dispositifs de terrain, de méthodes d'échantillonnage et de comptage bien définies.

### Objectif du stage

Le but de l'étude qui va suivre est, dans un premier temps, de déterminer l'aire de répartition de la Pie-grièche écorcheur en Alsace sur 2011 et 2012 à partir des données non protocolées issues de la base de données faunistiques VisioNature. De telles données sont saisies indépendamment de toute méthode d'échantillonnage ou de dispositifs de collecte préalablement établis. Le calcul de la surface de répartition de l'espèce en Alsace passe par une décomposition de la région en mailles (2*2) km de superficies égales à 4 km² (400ha). Sur chacune des 2282 mailles qui couvrent l'Alsace, est calculée une probabilité de présence en fonction de différents paramètres liés à l'écologie du site et au travail de prospection effectuée par les observateurs. Ainsi, l'ensemble des mailles sur lesquelles sont conclues une présence probable de l'espèce, représente la surface totale d'occupation sur cette année.

Ensuite, il s'agira de mesurer l'influence qu'ont certaines variables, composant ces données non protocolées, sur l'observation de l'espèce au niveau des sites prospectées. Et enfin, de déterminer les sites d'observation sur lesquelles ces interactions entre variables sont plus apparentes.

# I. Structure d'accueil : ODONAT

## I.1. Présentation Générale

L'Office des Données Naturalistes en Alsace (ODONAT) en tant que fédération d'associations naturalistes, œuvre pour la connaissance et la protection des espèces, des espaces naturels et des paysages en Alsace. Elle est donc un réseau associatif au service de l'information sur les espèces et milieux naturels en Alsace.

A ce jour, elle compte en son sein 15 associations adhérentes à ses objectifs ou ayant déjà collaboré à des projets. Et particulièrement celles qui ont été à l'origine sa création en 1995 :

- Alsace Nature ;
- CSA (Conservatoire des Sites Alsaciens) ;
- LPO-Alsace (Ligue pour la Protection des Oiseaux) ;
- Et le GEPMA (Groupe d'Etude et de Protection des Mammifères d'Alsace).

En plus d'être leur point convergent, elle représente une interface entre ces associations affiliées et les partenaires financiers, les bureaux d'études et les collectivités locales.

## I.2. Missions

ODONAT se consacre à la collecte, au traitement, à la diffusion, et la valorisation des données naturalistes recueillies par les spécialistes, les bénévoles et scientifiques au sein des associations membres. ODONAT se charge de diffuser et d'optimiser l'accès aux données et leur utilisation mais elle n'est aucunement une banque de données.

Elle contribue également à la préservation des espèces, des milieux naturels et paysages en Alsace mais aussi à la réalisation d'expertise et d'études en interne ou commandités par des partenaires extérieurs. Elle coordonne et uniformise l'apport des associations membres dans la mise en œuvre de tels projets. Cependant, les contributions des associations affiliées restent toujours volontaires et libres.

En sa qualité de gestionnaire de données naturalistes, ODONAT représente un organisme clé de demande d'information ou de renseignement relatifs à la nature en Alsace.

# II. La Pie-grièche écorcheur

## II.1 Présentation de l'espèce

La Pie-grièche écorcheur est un passereau de taille moyenne, un peu plus grand que le moineau et typiquement migrateur. Elle est présente en Alsace que pendant la belle saison, plus précisément dans la dernière décade d'avril ou en début mai et elle quitte en général son territoire entre la mi-juillet et la mi-août.

C'est un oiseau adepte des milieux semi-ouverts, bien ensoleillés, avec des buissons, des haies et des arbustes bordant des espaces découverts, à végétation rase. On la trouve particulièrement dans les landes sèches colonisées par quelques buissons, les friches agricoles, les vergers, les parcs à bestiaux et les clairières forestières avec des jeunes plantations. En Alsace, l'espèce est bien présente dans les vallées vosgiennes et dans les secteurs dominés par l'élevage, avec des haies, des prairies et des pâturages alors qu'elle est rare dans les zones d'agriculture intensive (cf. Annexe 1-Carte de répartition de la Pie-grièche écorcheur en Alsace).

**Photo prise par Yves Muller –Président de l'association ODONAT**

Elle niche sur des arbustes épineux, des ronciers, à une hauteur généralement comprise entre 0.5 et 2m. Elle se nourrit essentiellement de proies animales très diverses, depuis la petite araignée jusqu'au campagnol, en passant par les petits passereaux, les coléoptères, et d'autres insectes.

### II.2. Pourquoi le choix de la Pie-grièche écorcheur ?

Depuis 2005, la Pie-grièche écorcheur est suivie dans le cadre du programme de Suivi des Indicateurs de la Biodiversité en Alsace (SIBA). Ainsi, chaque année, des observateurs bénévoles, ainsi que des salariés de la Ligue pour la Protection des Oiseaux (LPO), suivent cette espèce selon un protocole de terrain précis et identique chaque année. Ce suivi permet d'analyser annuellement l'évolution de l'espèce sur l'ensemble du territoire alsacien.

Et en 1995, ODONAT et ses associations fédérées ont développé le système de saisie en ligne intitulé VisioNature visant à rassembler de façon volontaire, des données naturalistes de groupes taxinomiques divers, en vue d'en restituer les principaux éléments d'abord aux participants inscrits et ensuite à un public plus large. Grace à cet outil, de nombreux naturalistes bénévoles alsaciens transmettent désormais leurs données tout au long de l'année. Près de 200 000 données naturalistes sont saisies chaque année dont 95% représentent des données d'oiseaux. Des données relatives à la Pie-grièche écorcheur parviennent donc en grande quantité au sein des associations mais ces données sont des données non protocolées, au contraire de celles récoltées dans le cadre du programme SIBA.

ODONAT souhaite aujourd'hui savoir si des données non protocolées, issues des bénévoles et saisies sur l'outil VisioNature, peuvent être utilisées pour suivre et surtout analyser l'évolution des populations de certaines espèces alsaciennes.

En plus de l'existence d'un suivi déjà existant permettant une comparaison des résultats finaux, le choix de l'espèce s'est porté sur la Pie-grièche écorcheur car cette espèce en plus d'être un important indicateur de la biodiversité en Alsace est à la fois commune, détectable (de par son chant), observable et assez suivie pour être notée par les ornithologues.

# III. Présentation synthétique des données

## III.1. Des données non protocolées issues de VisioNature

Les données mises à notre disposition sur la Pie-grièche écorcheur sont issues de la base de données faunistiques en ligne Visio Nature Faune-Alsace. Cette base a été mise en place en collaboration avec la LPO pour la saisie des données faunistiques, de groupes taxinomiques divers, aussi bien les données avifaune, herpétofaune, mammalofaune, entomofaune etc. ODONAT est chargé de la gestion et de l'administration technique du système.

Cette base de données est à libre accès et chaque utilisateur (ornithologue, bénévole, amateur) peut saisir ses observations faunistiques après s'être authentifié. Cependant les données ornithologiques restent toujours les plus abondantes sur VisioNature, et ceci quelle que soit la période de l'année.

Toutefois, la spécificité de ces données saisies est qu'elles sont non protocolées. C'est-à-dire qu'elles sont saisies indépendamment de toute méthode d'échantillonnage de terrain ou de dispositifs scientifiques de recensement d'espèces. La saisie d'une donnée sur VisioNature est donc juste conditionnée par l'observation d'une espèce et une volonté de l'utilisateur de la noter.

Cependant, toutes les données déposées par les observateurs font l'objet d'une vérification, pour détecter les erreurs de saisie, et pour s'assurer de la véracité d'observations peu communes. Les saisies Une donnée non validée n'apparait pas dans les statistiques et les restitutions collectives de la base, y compris dans l'atlas de répartition des espèces.

## III.2. Données disponibles sur la Pie-grièche écorcheur

Les utilisateurs de VisioNature saisissent la donnée pie-grièche écorcheur génériquement selon :

- Le nom scientifique de l'espèce ;
- Un numéro d'identifiant d'espèce ;
- Une date d'observation ;
- Le référencement géographique du lieu d'observation :
- Un niveau taxinomique (Oiseaux, reptiles, amphibiens, etc.) ;
- La famille d'appartenance de l'espèce ;
- Les communes et départements d'observation ;
- Un code INSEE ;
- Un code Atlas ;
- Un indice de nidification simple ;
- Un indice de nidification en toutes lettres ;
- L'altitude à laquelle l'espèce a été observée ;
- Et les noms et prénoms de l'utilisateur.

D'autres variables sur l'espèce peuvent en plus être construites à l'aide du logiciel SIG (Système d'Information Géographique) qui va consolider et croiser ces données issues de VisioNature.

C'est ainsi que les données suivantes sont également disponibles:

- Le nombre de Pie-grièche écorcheur notés sur un site d'observation ;
- Le nombre de passages sur un site d'observation ;
- Et Les indices de nidification de l'espèce spécifiques à chaque maille (2*2) km.

Cependant, toutes ces variables ne seront pas utilisées dans la modélisation de la probabilité d'observation de l'espèce. Ceci pour éviter la colinéarité entre elles et la surparamétrisation des modèles. Nous serons donc par la suite amenés à faire une sélection parmi ces variables et ne garder que celles étant les plus liées à l'observation de l'espèce.

### III.3. La base de données Occupation du sol

Elle a été réalisée en 2008 par la CIGAL (Coopération pour l'Information Géographique en Alsace) sous maitrise d'ouvrage de Région Alsace, un de ses membres fondateurs. Ce projet a pour objectif d'actualiser la connaissance de l'occupation du sol sur le territoire alsacien de façon exhaustive et selon une méthode reconductible.

| **Base de données Occupation du Sol Pie-grièche écorcheur 2008** | | |
|---|---|---|
| **Caté PGE- I** | **Caté PGE- II** | **Caté PGE- III** |
| Forêts | Forêt | Forêt |
| Milieux ouverts | Cultures annuelles | Cultures annuelles |
| | | Cultures spécifiques |
| | | Houblon |
| | Cultures permanentes | Vergers intensifs |
| | | Vignes |
| | Milieux ouverts divers | Bosquets et haies |
| | | Autres espaces libres |
| | | Carrières (Bâtiments) |
| | | Carrières (Zones d'exploitation) |
| | | Chantiers et remblais |
| | | Emprise réseau ferré |
| | | Emprises militaires |
| | | Equipements sportifs et de loisirs |
| | | Gravières et sablières (Bâtiments) |
| | | Gravières et sablières (Zones d'exploitation) |
| | | Fourrés, fructicées et ligneux |
| | | Friches industrielles |
| | | Friches minières (Bâtiments industriels et espaces associés) |
| | | Friches minières (Terrils et anciennes carrières) |
| | | Golfs |
| | | Jardins ouvriers |
| | | Pelouses et zones arborées |
| | | Ripisylves |
| | | Roches nues |
| | Marais, Prairies et landes | Landes |
| | | Pelouses et pâturages de montagne |
| | | Prairies |
| | Vergers | Vergers traditionnels |
| Zones artificialisées | Espaces artificialisés | Espaces urbanisés |
| | | Cimetières |
| | | Emprises aéroportuaires (Autres espaces) |
| Zones humides | Eau | Eau |

Nous allons considérer les superficies de ces milieux occupés par l'espèce comme variables disponibles, en complément des variables listées plus haut et sur lesquelles une sélection va se faire pour constituer un groupe exhaustif de prédicteurs pour les modélisations.

## IV. Sélection des variables pour les modélisations statistiques

Pour modéliser la répartition géographique d'une espèce, il est nécessaire de disposer d'un jeu de données associant les données d'occurrence de l'espèce à des valeurs de paramètres recueillis sur un certain nombre de sites d'observation.

Les données d'occurrence de l'espèce correspondent aux données de présence/absence de la PGE et sont codées suivant la variable binaire « presence » prenant la valeur 1 si la Pie-grièche écorcheur a été observée sur la maille, 0 sinon.

Les paramètres auxquels nous allons associer les données d'occurrence de l'espèce sont obtenus au terme d'une sélection dans les données issues de VisioNature et présentées plus haut. Cette sélection est faite pour ne considérer que les variables directement liées à la présence de l'espèce sur un site donné.

Les variables retenues et disponibles annuellement sur les 2282 mailles (2*2) km en Alsace sont :

- Le nombre d'oiseaux notés sur la maille : Correspondant au nombre d'oiseaux vus et comptés sur le site, toutes espèces confondues;

- Le nombre de passages des observateurs sur la maille : correspondant au nombre de fois que le site a été prospecté par les observateurs sur l'année;

- L'année de prospection : Variable qualitative à deux modalités (2011/2012). Son intégration dans le modèle permettra de mesurer l'impact des années d'étude passées et/ou futures à chaque fois que des données sur ces années seront rajoutées aux tables.

- L'indice de nidification de l'espèce : Cet indice est laissé à l'appréciation de l'observateur.

    - 1 si la nidification de l'espèce est jugée certain sur la maille ;
    - 2. Si elle est probable.
    - 3. Si elle est possible ;
    - 4. Si elle est inconnue.

- Et une variable « Occupation du sol » : Constituée en réalité des superficies des variables présentées dans la base de données OCS 2008 par la Pie-grièche écorcheur-Section III.3.

Cette variable « Occupation du sol » est présentée en trois sous-ensembles dont nous ne considérerons que les deux premières (Caté PGE-I et Caté PGE-II). Ce qui nous amène à travailler sur deux tables de données aux champs similaires sauf pour les superficies de milieux écologiques occupés par la Pie-grièche écorcheur.

### IV.1. Construction de la première table de données

- Le nombre d'individus ;
- Le nombre de passages sur la maille;
- L'année de prospection ;
- Et l'indice de nidification de l'espèce.

Et la variable « Occupation du sol » constituée des superficies de :

- Milieux forestiers ;
- Milieux ouverts ;
- Zones artificialisées ;
- Zones humides.

### IV.2. Construction de la deuxième table de données

- Le nombre d'individus ;
- Le nombre de passages sur la maille;
- L'année de prospection ;
- Et l'indice de nidification de l'espèce.

Et la variable « Occupation du sol », ensemble des superficies en :

- Cultures annuelles;
- Cultures permanentes;
- Zones d'eau;
- Zones artificialisées;
- Forêt;
- Milieux ouverts;
- Marais, prairies et landes;
- Vergers.

## V. Planification des modélisations statistiques pour l'atteinte des objectifs

- **Principe de la modélisation statistique**

La modélisation de la présence / absence de l'espèce sur une maille donnée revient à modéliser la niche écologique de présence. Elle consiste à construire une fonction de paramètres environnementaux qui prédit la probabilité de présence de l'espèce à partir d'un jeu de données de calibration comprenant des données de présence/absence (variable « presence ») et d'abondance (Nombre d'oiseaux relevés) de l'espèce ou d'autres espèces, et des valeurs de paramètres environnementaux relatives aux sites d'observation (variable « Occupation du Sol »).

En fait, de tels modèles prédisent plutôt des combinaisons de facteurs environnementaux qui sont attendues favorables à la présence de l'espèce étudiée ; il est donc plus approprié de considérer qu'ils prédisent des habitats potentiels.

- **Outils statistiques utilisés**

Nous cherchons à modéliser une variable qualitative binaire : Observation (Présence/Absence) de la Pie Grièche Ecorcheur sur une maille à l'aide d'un ensemble déterministe de variables explicatives qualitatives et quantitatives et à déterminer les relations d'interactions possibles entre ces variables explicatives.

Pour ce faire, il existe de nombreuses méthodes dites de « statistique supervisée » parmi lesquelles la régression logistique, les arbres de décision, l'analyse discriminante, l'analyse en composantes principales… etc.

Nous allons présenter deux méthodes appropriées dans le cas de données de présence/absence d'une espèce : L'analyse en Composantes Principales (ACP) et la Régression Logistique.

La régression logistique, qui suppose une loi de distribution binomiale de la variable « presence » à prédire et une fonction de lien logistique, convient pour la modélisation d'événements de présence/absence d'une espèce ; dans ce cas, la variable à prédire est l'indice binaire décrivant la présence ou l'absence de l'espèce et les prédicteurs entrant dans le modèle sont les variables constituant les différents jeux de données.

L'Analyse en Composantes Principales (ACP) est une méthode de l'analyse multidimensionnelle de données, permettant de déterminer les relations entre les variables ainsi que les spécificités individuelles des mailles (2*2) km, et ceci en réduisant l'information de façon à ne pas en perdre.

## VI. Mise en œuvre de la régression logistique pour l'étude de l'évolution de la répartition des populations de Pie-grièche écorcheur

### VI.1 Présentation de la régression logistique

- **Définitions**

La régression logistique est un type de modèle statistique qui appartient à la famille des Modèles Linéaires Généralisés. Elle s'utilise lorsque la variable à expliquer est qualitative, le plus souvent binaire. Les variables explicatives peuvent être par contre soit qualitatives ou soit quantitatives.

La variable dépendante est habituellement la survenue ou non d'un évènement (présence d'une maladie, présence d'une espèce) et les variables explicatives sont celles susceptibles d'expliquer la survenue de cet évènement, c'est-à-dire les variables mesurant l'exposition à un facteur d'absence ou à un facteur de présence de l'espèce.

Contrairement à la régression linéaire multiple et l'analyse discriminante, la régression logistique n'exige pas une distribution normale des prédicteurs, ni l'homogénéité des variances. Par ses nombreuses qualités donc, cette technique est de plus en plus préférée à l'analyse discriminante par les statisticiens et les spécialistes du scoring.

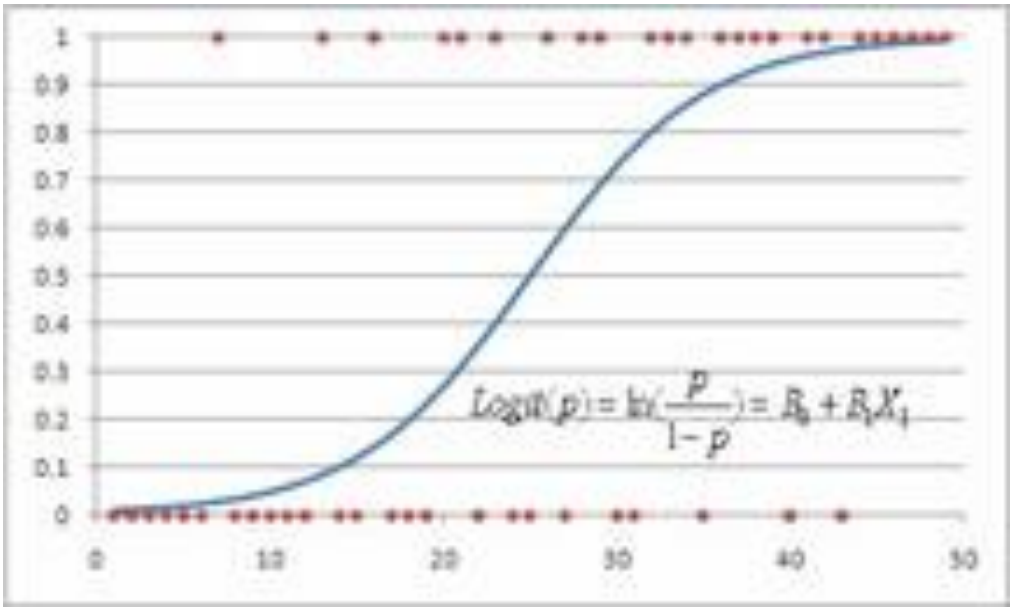
- **Principes mathématiques et conditions d'applications**

Lorsque nous voulons modéliser une variable à réponse binaire (**Absence/Présence de la PGE**), la forme de la relation est souvent non linéaire. On recourt alors à une fonction non-linéaire, de type logistique par exemple, en pareils cas.

Le principe de la régression logistique binaire est de considérer une variable à prévoir binaire (variable cible admettant uniquement deux modalités possibles) Y = {0,1} d'une part, et *p* variables explicatives notées X = (X1, X2, ..., Xp), continues, binaires ou qualitatives. L'objectif de la régression logistique est de modéliser l'espérance conditionnelle E (Y/X=x), par l'estimation d'une valeur moyenne de Y pour toute valeur de X. Pour une valeur Y valant 0 ou 1 (loi de Bernoulli), cette valeur moyenne est la probabilité qu'Y=1. On a donc :

$$E(Y/X=x) = prob(Y=1/X=x).$$

Une propriété essentielle de la régression logistique est qu'elle n'exige pas que les prédicteurs (variables explicatives) suivent une loi normale, ou soient distribués de façon linéaire, ou encore qu'ils possèdent une variance égale entre chaque groupe. La forme de courbe (en « s ») que nous remarquons par ailleurs est appelé sigmoïde, ou courbe logistique.




**Graphe 1. <u>Fonction logit de la régression logistique.</u>**

Si l'on suit l'expression de cette courbe, on peut écrire la fonction logistique **E(Y) = $p_i$ = probabilité (Y=1/X=x) (probabilité de présence)** sous la forme:

$$E(Yi)=pi=\frac{\exp(\beta 0+\beta iXi)}{1+\exp(\beta 0+\beta iXi)}$$

Encore, la probabilité d'occurrence selon la formule logistique s'écrit :

$$\text{Log}\left(\frac{pi}{1-pi}\right)=\beta 0+\beta iXi.$$

En fait, en cherchant à expliquer la probabilité de réalisation de l'évènement de présence **prob (Y=1/X=x)**, il nous faudrait une transformation de **E(Y)** qui étende l'intervalle de définition **[0,1]**. C'est le calcul des ratios de chance « odds ratio » qui permet d'envisager cette transformation. Ainsi le quotient **pi / (1-pi)** est appelé « odds », et la fonction **f(p)=ln (pi/1-pi)** est appelée « logit ».

### VI.2. <u>Explication du choix de la méthode</u>

La régression logistique convient pour la modélisation d'événements de présence/absence d'une espèce. On cherche à prédire la probabilité de présence/absence de la Pie-grièche écorcheur en expliquant la variable binaire « presence » qui suit une loi binomiale.

Dans pareil cas, la nécessité d'utiliser des modèles particuliers se justifie par le fait que :

- L'utilisation d'un modèle de régression linéaire classique n'est plus adéquat ;
- La mise en œuvre d'une régression linéaire va produire des valeurs continues qui ne s'interprètent pas comme des probabilités, or on veut uniquement des probabilités sur [0,1] ;
- La variable de réponse « presence » suit une loi binomiale.

Ainsi, la régression logistique peut alors être utilisée pour prédire cette variable en fonction de variables quantitatives – continues -, binaires et qualitatives.

### VI.3. Régression logistique sur la première table de données

Notre but est de modéliser la variable « presence », désignant l'observation ou non de l'espèce sur une maille, en fonction des 8 variables explicatives constituant cette 1e table de données.

Autrement dit, on va chercher à estimer la probabilité de présence de l'espèce sur chacune des 2282 mailles (2*2) km en fonction de huit (08) prédicteurs considérés.
La probabilité d'absence en est déduite avec :

**P (absence)=1- P (presence).**

**1e étape : Régression logistique posé sur les variables et interactions d'ordre 2**

Il s'agit d'écrire le modèle de régression sur la variable réponse «presence » en fonction des 8 variables explicatives de base et de leurs interactions d'ordre 2. Ce qui fait un groupe de 8 variables principales et 28 interactions entre elles. (cf. Annexe 2).

La variable « presence » sera donc expliquée à l'aide de 36 prédicteurs et à partir de l'estimation des coefficients associés aux variables influentes, seront calculées les probabilités de presence/absence de l'espèce sur les 4564 mailles prospectées en 2011/2012.

Notons qu'on a choisi de négliger toutes les interactions d'ordre strictement supérieur à 2.

**2e étape : Analyse descendante ou SELECTION BACKWARD sur le modèle**

Le modèle de base postulé nécessite l'estimation de 36 paramètres dont il faut tester la significativité sur le modèle posé. Vu le nombre élevé de variables et donc de tests individuels de significativité, des méthodes de sélection de modèle peuvent être envisagées.

Nous pouvons appliquer une analyse descendante avec la SELECTION BACKWARD de R. Cette sélection est basée sur le critère AIC. La procédure de sélection part du modèle de base où toutes les variables sont intégrées dans le modèle. L'opération s'arrête lorsque toutes les variables considérées sont retirées du modèle ou lorsque le retrait de l'une d'entre elles conduit à une augmentation ou à une stagnation de la dernière valeur prise par le score AIC. Au final, cette analyse ne garde que les variables influentes parmi toutes celles utilisées pour écrire le modèle.

Dans notre cas, la procédure de sélection de modèle s'arrête à une étape où les 8 variables principales et 13 interactions sont retenues (cf. Annexe 2).

Ainsi, du modèle de départ à 36 variables, on aboutit à un modèle de 21 variables dont 13 interactions. Ce dernier modèle obtenu est jugé être le « meilleur » modèle au sens de l'AIC.

**3e étape : Calcul des probabilités de présence/absence de l'espèce sur les mailles**

Les probabilités d'occurrence de la Pie-grièche écorcheur sont obtenues à partir des estimations des coefficients associés à chacune des variables retenues par la sélection BACKWARD.

Elles sont données par l'espérance de la variable à prédire conditionnellement aux valeurs prises par les prédicteurs, et calculées à l'aide de la fonction logistique (logit) :

$$\mathbf{P\ (Presence) = \frac{exp(\ a + bX)}{1+ exp\ (a + bX)}}\ \mathbf{, où:}$$

**a :** la valeur de l'ordonnée à l'origine ;

**b :** Vecteur colonne constitué des estimations des 21 coefficients associés aux variables retenues par l'analyse BACKWARD ;

**X :** Matrice constituée des 21 valeurs de variables retenues par l'analyse BACKWARD. Les interactions sont égales aux produits des valeurs de variables les composant.

**4e étape: Capacité de prédiction du modèle choisi**

Il est essentiel d'apprécier le pouvoir de prédiction de notre modèle final. Il peut être estimé à l'aide d'un critère qu'est le pouvoir de discrimination.

Le pouvoir discriminant du modèle est sa capacité à discriminer les événements de présence et d'absence de l'espèce. Il peut être évalué en comparant les événements de présence/absence prédits par le modèle avec les valeurs observées dans les données. (cf. Annexe 2).

Pour prédire une variable binaire, on doit se munir d'une règle de décision et d'un seuil θ :

$$\mathbf{Y=} \begin{cases} \mathbf{1 \text{ si } P\ (Y=1) > \theta,} \\ \mathbf{0 \text{ sinon.}} \end{cases}$$

En règle générale, on prend **θ=0.5** et **Y=1** correspond à une présence de l'espèce.
Autrement dit, une probabilité prédite p supérieur à 0.5 (50%) représente une observation probable de l'espèce sur la maille (2*2) km concernée.

Les comptages de comparaison prédiction/données sont reportés dans une matrice de confusion.

- Sur 2011 :

| Variable Observation | Observations prédites | | |
|---|---|---|---|
| | Absence | Présence | Total |
| Non observé | 1993 | 53 | **2046** |
| Observé | 16 | 220 | **236** |
| Total | **2009** | **273** | **2282** |

- Sur 2012 :

| Variable Observation | Observations prédites | | |
|---|---|---|---|
| | Absence | Présence | Total |
| Non observé | 1950 | 28 | **1978** |
| Observé | 253 | 51 | **304** |
| Total | **2203** | **79** | **2282** |

Le tableau de classification consolidé sur les deux années donne :

| | Absence | Présence | Total |
|---|---|---|---|
| Non observé | 3943 | 81 | **4024** |
| Observé | 269 | 271 | **540** |
| Total | **4212** | **352** | **4564** |

La capacité prédictive ou score d'exactitude du modèle est donné par la formule:

**ACC = (VP+VN) / N**, où :

**VP** : Nombre de mailles où une prédiction de présence est égale à une observation réelle;
**VN** : Nombre de mailles où une prédiction d'absence correspond à une non observation.

Cette formule est l'estimation du pourcentage de vrais positifs (présence prédite égale à une observation réelle) et de vrais négatifs (absence prédite égale à une absence notée) sur l'ensemble des prospections.

Numériquement,

**ACC= [(3943 + 271) / 4564]= 92,33%.**

Le pouvoir discriminant du modèle sur cette 1$^{e}$ table de données est de 92.33%. Ce modèle illustre donc bien les observations réelles de l'espèce sur les sites.

**5$^{e}$ étape : Variables influentes sur l'observation de l'espèce.**

Il s'agit de tester la significativité à un seuil de 5%, des variables finales de la sélection BACKWARD. Cela revient à tester les hypothèses : (cf. Annexe 2)

**H0 : {βj =0} contre H1 : {βj≠0}, pour j= {1,…,21},**

où les **βj** sont les coefficients associés aux variables retenues par la sélection BACKWARD.

L'élimination descendante renvoie en sortie les paramètres (variables et interactions) qui semblent être les plus significatives dans la prédiction des probabilités d'observation de l'espèce. Le modèle obtenu est considéré comme étant le « meilleur » au sens de l'AIC.

Cependant, tous les coefficients de ce modèle ne sont pas significatifs au seuil de 5%. Nous allons donc faire appel au test du rapport de vraisemblance pour déterminer lesquels sont significativement pertinents sur l'observation de l'espèce avec la commande « summary »de R.

On a une significativité d'une variable (décision de H1) si la p-value associée au coefficient de cette variable (colonne **Pr (>|z|))** est inférieure à 0.05 (seuil des tests de 5%), sinon cette variable n'est pas réellement influente sur la variable « presence » (décision de H0).

Les résultats des tests sur R sont relevés dans le tableau ci-après :

| Coefficients | Estimate | Std. Error | z value | Pr (>\|z\|) | |
|---|---|---|---|---|---|
| **(Intercept)** | -6.950e+00 | 5.936e-01 | -11.708 | < 2e-16 | *** |
| **nois** | -1.162e-02 | 4.027e-03 | -2.886 | 0.003900 | ** |
| **npass** | 3.269e-02 | 2.577e-02 | 1.269 | 0.204542 | |
| **nidif** | 1.380e+01 | 1.637e+00 | 8.430 | < 2e-16 | *** |
| **foret** | 8.041e-04 | 1.315e-03 | 0.611 | 0.541007 | |
| **milouv** | 3.031e-03 | 1.224e-03 | 2.475 | 0.013306 | * |
| **zonart** | -1.095e-02 | 3.169e-03 | -3.456 | 0.000548 | *** |
| **zonhum** | 7.522e-03 | 8.169e-03 | 0.921 | 0.357131 | |
| **annee2012** | 3.449e+00 | 4.184e-01 | 8.244 | < 2e-16 | *** |
| **nois:npass** | -2.379e-05 | 8.265e-06 | -2.879 | 0.003991 | ** |
| **nois:nidif** | 1.233e-02 | 4.389e-03 | 2.809 | 0.004963 | ** |
| **nois:foret** | 8.313e-05 | 1.371e-05 | 6.064 | 1.33e-09 | *** |
| **nois:annee2012** | 5.082e-03 | 3.800e-03 | 1.337 | 0.181060 | |
| **npass:nidif** | -1.062e-01 | 2.408e-02 | -4.412 | 1.02e-05 | *** |
| **npass:milouv** | 2.611e-04 | 6.316e-05 | 4.134 | 3.57e-05 | *** |
| **npass:annee2012** | 7.427e-02 | 2.538e-02 | 2.926 | 0.003434 | ** |
| **nidif:foret** | -1.777e-02 | 4.176e-03 | -4.255 | 2.10e-05 | *** |
| **nidif:milouv** | -1.938e-02 | 4.449e-03 | -4.357 | 1.32e-05 | *** |
| **foret:milouv** | 2.354e-05 | 6.518e-06 | 3.611 | 0.000305 | *** |
| **foret:zonart** | 3.796e-05 | 2.603e-05 | 1.458 | 0.144714 | |
| **foret:zonhum** | -1.897e-04 | 8.774e-05 | -2.162 | 0.030584 | * |
| **zonart:zonhum** | -3.448e-04 | 1.981e-04 | -1.741 | 0.081715 | . |

Les variables principales les moins influentes dans l'observation de l'espèce sont le nombre de passages sur la maille, les territoires forestiers et les zones humides.
Les p-values associées à leurs coefficients sont supérieures à 0.05 donc l'hypothèse nulle (non significativité des variables) n'est pas totalement rejetée.

Certaines interactions notamment entre la forêt et les zones artificialisées, entre la forêt et les zones humides, entre les zones artificialisées et les zones humides n'ont pas d'intérêt en un sens écologique, mais ils apportent plus de robustesse au modèle postulé et de précision aux estimations des coefficients associés aux variables. Par ailleurs, aucune d'elles n'est pertinente dans l'écriture du modèle.

Les variables de base influant dans l'observation de l'espèce sont la nidification (résultat prévisible), le nombre d'oiseaux notés sur la maille, les milieux ouverts, les zones artificialisées et l'année 2012. L'influence du nombre d'oiseaux sur l'observation de l'espèce est confirmée par la significativité de son interaction avec l'indice de nidification.

L'année 2012 a été plus favorable à l'observation de l'espèce qu'en 2011.

L'interaction entre le nombre de passages et les milieux ouverts est influente dans ce modèle. Plus on fait de prospections dans les milieux ouverts, plus l'observation de l'espèce est possible.

L'interaction entre la forêt et le nombre d'oiseaux est aussi significative. L'observation est possible sur des milieux forestiers à la condition que ces milieux soient riches en avifaune.

Le terrain commun entre la forêt et les milieux ouverts est également propice à l'observation de la PGE. Cette interaction est représentée par les lisières de forêt et les clairières, autrement dit les anciens espaces forestiers défraichis et devenus moins denses dont la constitution se rapproche plus à celle de milieux ouverts.

**6e étape : Calcul de l'aire de répartition de la Pie-grièche écorcheur**

Cette aire de répartition est calculée à partir des probabilités de présence de l'espèce calculées sur chaque maille (2*2) km en 2011 et 2012.

L'aire totale de présence prédite par le modèle statistique est égale à la somme des superficies des mailles sur lesquelles on a une probabilité de présence supérieure ou égale à 0.5, autrement dit la somme des aires des mailles sur lesquelles on a conclu une présence probable de l'espèce. Ce nombre de mailles correspond aux colonnes « Presence » des matrices de confusion.

En 2011, le modèle prédit une présence sur 273 mailles et en 2012 sur 79 mailles.
L'espèce se répartit donc sur 1092 km² en Alsace sur 2011 et sur 316 km² en 2012.

**7e étape : Conclusions**

Les données de cette 1e table ont été obtenues sur 8322 km² de territoires alsaciens sur les deux années 2011 et 2012. Pour rappel, la superficie théorique de l'Alsace est de 8280 km². Cette différence de 42,45 km² résulte du dépassement sur certaines mailles (2*2) km des superficies de 4km² imposées. Ce dépassement est néanmoins négligeable car ne dépassant pas une surface de 0.1 km²/maille en général.

Et à partir des résultats obtenus précédemment, la Pie-grièche écorcheur s'est raréfiée sur 776km² entre 2011 et 2012, soit une disparition sur 9,32% du territoire alsacien.

### VI.4. Régression logistique sur la deuxième table de données

**1e étape : Régression logistique sur les variables et interactions d'ordre 2**

Ce modèle est posé sur le 2e jeu de données composé de la variable de réponse « presence » et de 12 variables explicatives relatives à la Pie-grièche écorcheur. (cf. Annexe 2).

Elle est similaire au 1e jeu de données, à l'exception de la variable «Occupation du sol » qui a été décomposée en des milieux écologiques plus précis. Elle est organisée selon une table de données de 13 colonnes représentant les valeurs des variables sur 4564 lignes (mailles 2*2) km.

Avec 12 prédicteurs de base et 66 interactions d'ordre 2, le modèle postulé est constitué de 78 variables de base dont nous allons mesurer l'impact sur l'observation de la Pie-grièche écorcheur. Les probabilités de présence/absence de l'espèce sont calculées à l'aide de la fonction logistique et des estimations des coefficients associés aux variables influentes du modèle.

On supposera comme précédemment que toutes les interactions d'ordre strictement supérieur à 2 sont négligeables.

**2e étape : Analyse descendante ou SELECTION BACKWARD sur le modèle**

Nous allons, dans cette partie, effectuer une élimination descendante basée sur le score AIC. Cette procédure permet de supprimer, pas à pas, du modèle de départ les variables peu influentes sur l'observation de la Pie-grièche écorcheur.

La procédure stoppe lorsque toutes les variables sont retirées du modèle ou lorsque la valeur de l'AIC du dernier modèle choisi, croit ou stagne.

A la fin de la procédure, sur un ensemble de 78 variables, le modèle obtenu se compose des 12 variables de base et de 29 interactions. Ce nouveau modèle ainsi obtenu est le « meilleur » modèle postulé au sens de l'AIC. (cf. Annexe 2).

**3e étape : Calcul des probabilités de présence/absence de l'espèce sur les mailles**

On se fixe la règle de décision selon laquelle on conclut une présence de l'espèce sur une maille avec une probabilité supérieure ou égale à 0.5. Sinon, on considère que l'espèce est absente de ce site d'observation.

Les probabilités d'occurrence de l'espèce sont calculées à l'aide de la fonction logistique :

$$P(Presence) = \frac{\exp(a + bX)}{1 + \exp(a + bX)}$$ , où:

**a :** la valeur de l'ordonnée à l'origine ;

**b :** le vecteur colonne (41*1) constitué des 41 coefficients estimés associés aux variables retenues par l'analyse BACKWARD ;

**X :** la matrice (4564*41) constituée des 41 valeurs des variables retenues par l'analyse BACKWARD sur chacune des 4564 mailles (2*2) km prospectées en 2011 et 2012. Les interactions sont égales aux produits des valeurs de variables les composant.

**4e étape: Capacité de prédiction du modèle**

Après le calcul des probabilités de présence/absence de l'espèce sur les mailles, on se propose de mesurer le pouvoir discriminant du modèle retenu. Ceci en comparant la présence/absence prédite par le modèle, aux données réelles recueillies par les observateurs. (cf. Annexe 2).

Les comptages sur 2011, 2012 et consolidés sur les deux années d'observation sont reportés dans des matrices de confusions présentées ci-après.

- En 2011 :

| Variable Observation | Observations prédites | | |
|---|---|---|---|
| | Absence | Présence | Total |
| Non observé | 1997 | 49 | **2046** |
| Observé | 16 | 220 | **236** |
| Total | **2013** | **269** | **2282** |

- En 2012 :

| Variable Observation | Observations prédites | | |
|---|---|---|---|
| | Absence | Présence | Total |
| Non observé | 1984 | 25 | **2009** |
| Observé | 203 | 70 | **273** |
| Total | **2187** | **95** | **2282** |

Globalement, sur les deux années d'étude, les prédictions comparées avec les observations réelles donnent la matrice suivante :

| Variable Observation | Observations prédites | | |
|---|---|---|---|
| | Absence | Présence | Total |
| Non observé | 3981 | 74 | **4055** |
| Observé | 219 | 290 | **509** |
| Total | **4200** | **364** | **4564** |

Et la capacité de prédiction du modèle est :

**ACC = [(3981+290)] / 4564 = 93,58%.**

Toutefois, notons que les valeurs de la matrice de confusion changent en fonction du seuil d'affectation choisi. On admet toutefois que ce seuil de 0.5 est choisi en règle générale et donne le meilleur score d'exactitude pour un modèle donné.

**5e étape : Variables influentes sur l'observation de l'espèce**

La sélection BACKWARD renvoie le meilleur modèle au sens de l'AIC. Cependant, toutes les variables retenues ne sont pas significatives à un seuil de 5%. On se propose d'appliquer le test du rapport de vraisemblance sur les variables retenues pour déterminer leur significativité.

Nous allons tester l'hypothèse nulle H0 contre l'hypothèse alternative H1, ci-dessous :

**H0 : {βj =0} contre H1 : {βj≠0}, pour j= {1,…,41}, où** les **βj** sont les coefficients associés à chacune des 41 variables obtenues par la sélection BACKWARD.

Les résultats des tests sont notés ci –après :

| Coefficients | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| **(Intercept)** | -6.963e+00 | 6.312e-01 | -11.030 | < 2e-16 | *** |
| **nois** | -1.429e-02 | 4.954e-03 | -2.885 | 0.003917 | ** |
| **npass** | 3.624e-02 | 2.849e-02 | 1.272 | 0.203379 | |
| **nidif** | 1.417e+01 | 1.867e+00 | 7.592 | 3.14e-14 | *** |
| **culann** | 1.621e-03 | 1.513e-03 | 1.071 | 0.283986 | |
| **culper** | 5.620e-03 | 3.685e-03 | 1.525 | 0.127236 | |
| **eau** | 7.913e-03 | 7.219e-03 | 1.096 | 0.272990 | |
| **zonart** | -8.631e-03 | 3.175e-03 | -2.718 | 0.006566 | ** |
| **foret** | 7.919e-04 | 1.358e-03 | 0.583 | 0.559714 | |
| **milouv** | 4.167e-04 | 4.775e-03 | 0.087 | 0.930472 | |
| **prailan** | 6.520e-04 | 3.950e-03 | 0.165 | 0.868895 | |
| **verg** | -5.836e-03 | 2.397e-02 | -0.243 | 0.807619 | |
| **annee2012** | 3.465e+00 | 4.688e-01 | 7.391 | 1.46e-13 | *** |
| **nois:npass** | -2.788e-05 | 8.121e-06 | -3.433 | 0.000596 | *** |
| **nois:nidif** | 1.438e-02 | 4.907e-03 | 2.930 | 0.003392 | ** |
| **nois:culper** | -1.388e-04 | 4.327e-05 | -3.209 | 0.001332 | ** |
| **nois:zonart** | 1.850e-05 | 9.781e-06 | 1.891 | 0.058615 | . |

| | | | | | |
|---|---|---|---|---|---|
| **nois:foret** | 8.014e-05 | 1.521e-05 | 5.270 | 1.37e-07 | *** |
| **nois:verg** | 7.476e-04 | 2.326e-04 | 3.214 | 0.001309 | ** |
| **nois:annee2012** | 5.898e-03 | 4.422e-03 | 1.334 | 0.182314 | |
| **npass:nidif** | -1.133e-01 | 2.804e-02 | -4.042 | 5.29e-05 | *** |
| **npass:culann** | 1.464e-04 | 7.162e-05 | 2.044 | 0.040962 | * |
| **npass:culper** | 1.146e-03 | 2.841e-04 | 4.033 | 5.51e-05 | *** |
| **npass:prailan** | 7.442e-04 | 1.476e-04 | 5.041 | 4.63e-07 | *** |
| **npass:verg** | -2.101e-03 | 1.014e-03 | -2.073 | 0.038181 | * |
| **npass:annee201** | 2 7.044e-02 | 2.940e-02 | 2.396 | 0.016554 | * |
| **nidif:culann** | -1.648e-02 | 4.978e-03 | -3.311 | 0.000931 | *** |
| **nidif:culper** | -1.930e-02 | 8.347e-03 | -2.312 | 0.020794 | * |
| **nidif:foret** | -1.755e-02 | 4.622e-03 | -3.797 | 0.000146 | *** |
| **nidif:milouv** | -2.199e-02 | 1.223e-02 | -1.799 | 0.072034 | . |
| **nidif:prailan** | -2.663e-02 | 5.784e-03 | -4.604 | 4.14e-06 | *** |
| **culann:zonart** | -4.839e-05 | 2.471e-05 | -1.959 | 0.050139 | . |
| **culann:prailan** | 4.243e-05 | 2.132e-05 | 1.990 | 0.046581 | * |
| **culper:eau** | 1.540e-03 | 1.083e-03 | 1.422 | 0.154941 | |
| **culper:foret** | 3.796e-05 | 2.490e-05 | 1.525 | 0.127273 | |
| **culper:milouv** | -6.435e-04 | 2.574e-04 | -2.499 | 0.012442 | * |
| **culper:prailan** | 1.315e-04 | 5.266e-05 | 2.497 | 0.012525 | * |
| **culper:verg** | -3.823e-04 | 1.625e-04 | -2.353 | 0.018641 | * |
| **eau:foret** | -1.526e-04 | 8.404e-05 | -1.816 | 0.069361 | . |
| **foret:prailan** | 5.041e-05 | 1.833e-05 | 2.750 | 0.005961 | ** |
| **prailan:verg** | -2.894e-04 | 1.282e-04 | -2.258 | 0.023949 | * |
| **verg:annee2012** | 6.792e-02 | 2.024e-02 | 3.356 | 0.000792 | *** |

Les variables de base significatives dans l'observation de l'espèce à un seuil de 5% sont :

- Le nombre d'oiseaux relevés sur la maille;
- L'indice de nidification ;
- Les zones artificialisées ;
- Et l'année 2012.

On peut donc dire que, plus le nombre d'oiseaux relevés est important, plus la présence de l'espèce est probable sur cette maille. Il est évident que l'observation de l'espèce reste fortement corrélée à l'indice de nidification. Les zones artificialisées représentent les milieux écologiques les plus favorables à la Pie-grièche écorcheur.

Les interactions influentes dans le modèle sont :

- Le nombre d'oiseaux avec le nombre de passages ;
- Le nombre d'oiseaux avec la nidification ;
- Le nombre d'oiseaux avec les cultures permanentes ;
- Le nombre d'oiseaux avec la forêt ;
- Le nombre d'oiseaux avec les vergers ;
- Le nombre de passages et la nidification ;
- Le nombre de passages avec les cultures annuelles ;
- Le nombre de passages avec les cultures permanentes ;
- Le nombre de passages avec les prairies et landes ;

- Le nombre de passages avec les vergers ;
- Le nombre de passages avec l'année 2012 ;
- La nidification avec les cultures annuelles ;
- La nidification avec les cultures permanentes ;
- La nidification avec la forêt ;
- La nidification avec les prairies et landes ;
- Les cultures annuelles avec les prairies et landes.

Le nombre d'oiseaux avec les cultures permanentes, le nombre d'oiseaux avec la forêt et le nombre d'oiseaux avec les vergers sont des interactions significatives. Plus haut, on a vu que ce nombre d'oiseaux est explicatif de l'observation de l'espèce mais aussi son interaction avec la nidification. C'est donc dire que l'observation de l'espèce est fort possible au niveau des cultures permanentes, des vergers et des territoires forestiers sous réserve que de telles zones soient riches en avifaune.

Le nombre de passages évolue dans le sens de la nidification et ses interactions avec les cultures annuelles, les cultures permanentes, les vergers, les prairies et les landes influent sur l'observation de l'espèce. Plus les prospections sont importantes dans de telles zones, plus l'observation de l'espèce est probable.

D'ailleurs, la significativité des interactions de l'indice de nidification avec des milieux comme les cultures annuelles, les cultures permanentes, la forêt, les prairies et landes confirment que l'observation de la Pie-grièche écorcheur est probable dans ces milieux sous condition qu'elles soient riches en oiseaux et que le nombre de passages y soit important.

Les domaines communs aux cultures annuelles et aux prairies et landes sont aussi favorables à l'observation de l'espèce.

**6e étape : Calcul de l'aire de répartition de la Pie-grièche écorcheur**

Ce calcul est basé sur les probabilités de présence prédites par le modèle.

**En 2011**, on a une **présence de l'espèce sur 269 mailles** contre **95 mailles en 2012**.
L'espèce est donc présente sur **1076 km²** en **2011** et sur **380 km²** en **2012.**

**7e étape : Conclusions**

Les résultats obtenus à l'issue de l'analyse nous font voir que l'espèce n'est plus présente sur 696 km² en 2012, autrement dit elle est passée d'une répartition en Alsace de 12.93% en 2011 à 4.57% en 2012, marquant ainsi sa disparition sur 696 km² soit 8,35% du territoire alsacien.

### VI.5. Comparaison des résultats de la régression logistique sur les deux tables de données

#### VI.5.1. Score d'exactitude des 2 modèles

Le modèle posé sur la 1e table de données présente une capacité prédictive de 92.33% contre une capacité prédictive de 93.58% du modèle sur la 2e table de données.

Le 2e modèle semble plus refléter la réalité –les données collectées – au niveau des prédictions de probabilité de présence de la Pie-grièche écorcheur.

### V.5.2.Comparaison des aires de répartition

Les modèles posés sur les deux tables de données montrent une baisse importante de l'aire de répartition de la Pie-grièche écorcheur en Alsace.

Cette baisse est de 9.32% (1e table de données) contre 8.35% de l'Alsace (2e table de données). Une différence de 0.97%, soit sur 265 km² (64 mailles) où les probabilités prédites divergent entre les 2 modèles.

### V.5.3. Variables influentes sur l'observation de la Pie-grièche écorcheur selon les deux modèles

Dans cette partie, on cherche à comparer les variables jugées pertinentes dans l'explication de l'observation de l'espèce, obtenues sur chacun des deux modèles.

- **Pour les variables de base des modèles**

Sur les deux modèles, les variables de base influentes dans l'observation de la Pie-grièche écorcheur sont l'indice de nidification, l'année 2012, le nombre d'oiseaux relevés sur le site, les milieux ouverts et les zones artificialisées.

- **Pour les interactions entre les variables de base**

Les interactions entre variables de base significatives et communes aux deux modèles sont le nombre d'oiseaux avec le nombre de passages sur le site, le nombre d'oiseaux et la nidification, le nombre d'oiseaux avec la forêt, le nombre de passages avec l'indice de nidification, le nombre de passages avec l'année 2012, l'indice de nidification avec la forêt.

### VI.6. Conclusions générales sur la régression logistique

La régression logistique a été posée à partir de deux tables de données.
Des tests de significativité ont été faits ainsi que des calculs de probabilités d'occurrence pour déterminer l'aire de répartition de l'espèce en Alsace.

L'exactitude des probabilités calculées sur le 2e jeu de données étant supérieur à celles obtenues à partir du 1e, nous pouvons penser que la 2e table de données décrit mieux les variables explicatives de l'observation de l'espèce. Et sur la base du modèle posé sur cette table, on peut donc dire que :

- L'étendue de présence de la Pie-grièche écorcheur a baissé de 8.35% en Alsace entre 2011 et 2012, soit une raréfaction de l'espèce sur 696 km² ;

- Les zones artificialisées et les milieux ouverts sont influentes dans l'observation de l'espèce;

- De façon générale, sur un site d'observation riche en avifaune est probable d'observer l'espèce notamment dans les cultures permanentes, la forêt et les vergers ;

- La Pie-grièche écorcheur peut être aussi observée dans les zones de cultures annuelles et les prairies et landes, sous condition qu'elles soient bien prospectées.

# VII. Mise en œuvre de l'Analyse en Composantes Principales (ACP)

## VII.1. Présentation de la méthode

L'Analyse en Composantes Principales (ACP) est une méthode de la famille de l'analyse de données et plus généralement de la statistique multi variée, qui permet de synthétiser un ensemble de données en identifiant la redondance entre celles-ci. Autrement dit, elle résume et identifie les structures et les tendances des données et fournit également une synthèse graphique pertinente des résultats.

L'ACP cherche à mettre en évidence les relations globales existant entre les variables dès que p>3, où p le nombre de variables explicatives, car ils sont impossibles à visualiser dans ce cas. La solution de l'ACP est alors de condenser l'information de manière à retirer les relations vraiment caractéristiques (proximités entre variables et individus) ceci en limitant la perte d'information.

Il s'agit d'une approche à la fois géométrique (les variables étant représentées dans un nouvel espace, selon des directions d'inertie maximale) et statistique (la recherche portant sur des axes indépendants expliquant au mieux la variabilité – la variance – des données). Lorsqu'on veut compresser un ensemble de N variables aléatoires, les n premiers axes de l'analyse en composantes principales sont un meilleur choix, du point de vue de l'inertie ou de la variance.

Plusieurs packages fournissent des outils permettant de réaliser une analyse en composantes principales. On peut citer :

- Anciennement le package NVA et désormais aussi dans le package stats : prcomp, princomp ;
- Le package ade4 : dudi.pca ;
- Le package FactoMineR : PCA.

## VII.2 Explication du choix de la méthode

L'ACP est simple à mettre en œuvre car en réalité les seuls outils mathématiques utilisés sont le calcul des valeurs/vecteurs propres d'une matrice, et les changements de base. Grâce aux graphiques qu'elle produit, elle permet facilement d'appréhender une grande partie de ses résultats. C'est aussi une méthode puissante, car en quelques opérations, elle offre un résumé et une vue complète des relations existant entre les variables quantitatives d'une population d'étude, résultats qui ne seraient obtenues qu'au prix de manipulations fastidieuses.

Les résultats les plus complets semblent toutefois être fournis par la procédure PCA. En effet, l'objet PCA fournit des informations très complètes sur les résultats de l'ACP. Entre autres :

- Les valeurs propres ;
- Les corrélations des variables avec les axes factoriels ;
- Les cos2 des variables (carré des corrélations) avec les axes ;
- Les contributions des variables aux axes ;
- Les coordonnées des individus ;
- Les cosinus carrés des individus ou encore contribution relatives des individus;
- Les contributions des individus ;
- Et les résultats pour individus et/ou variables supplémentaires illustratives.

### VII.3. Sélection des variables

Contrairement à la régression logistique, une sélection de variables n'est pas nécessaire dans une analyse de données à composantes principales. Toutes les variables retenues pour l'explication de l'observation de l'espèce sont traités simultanément et indépendamment d'une sélection de variables influentes.

L'ACP ne se fait que sur des variables à quantitatives et semi-quantitatives qui vont constituer les variables actives de l'ACP. En partant de ce principe, les variables **« annee** », variable catégorielle à deux modalités : 2011 et 2012, et « **presence** », variable qualitative dichotomique/variable binaire marquant l'observation ou non de l'espèce, ne seront pas prises en compte dans l'analyse mais seront plutôt considérées comme des **variables illustratives**. C'est-à-dire qu'elles ne vont pas intervenir dans la construction des composantes principales mais vont nous aider à l'interprétation des dimensions de variabilité.

### VII.4 Mise en œuvre de l'ACP sur la première table de données

#### VII.4.1. Répartition des variables de cette table de données

Cette première table de données est la même utilisée dans la régression logistique et décrite dans la section IV.1 de ce rapport. Elle est constituée de 9 colonnes correspondant aux 9 variables mesurées sur les 4564 mailles (2*2) km² (lignes) en 2011 et en 2012.

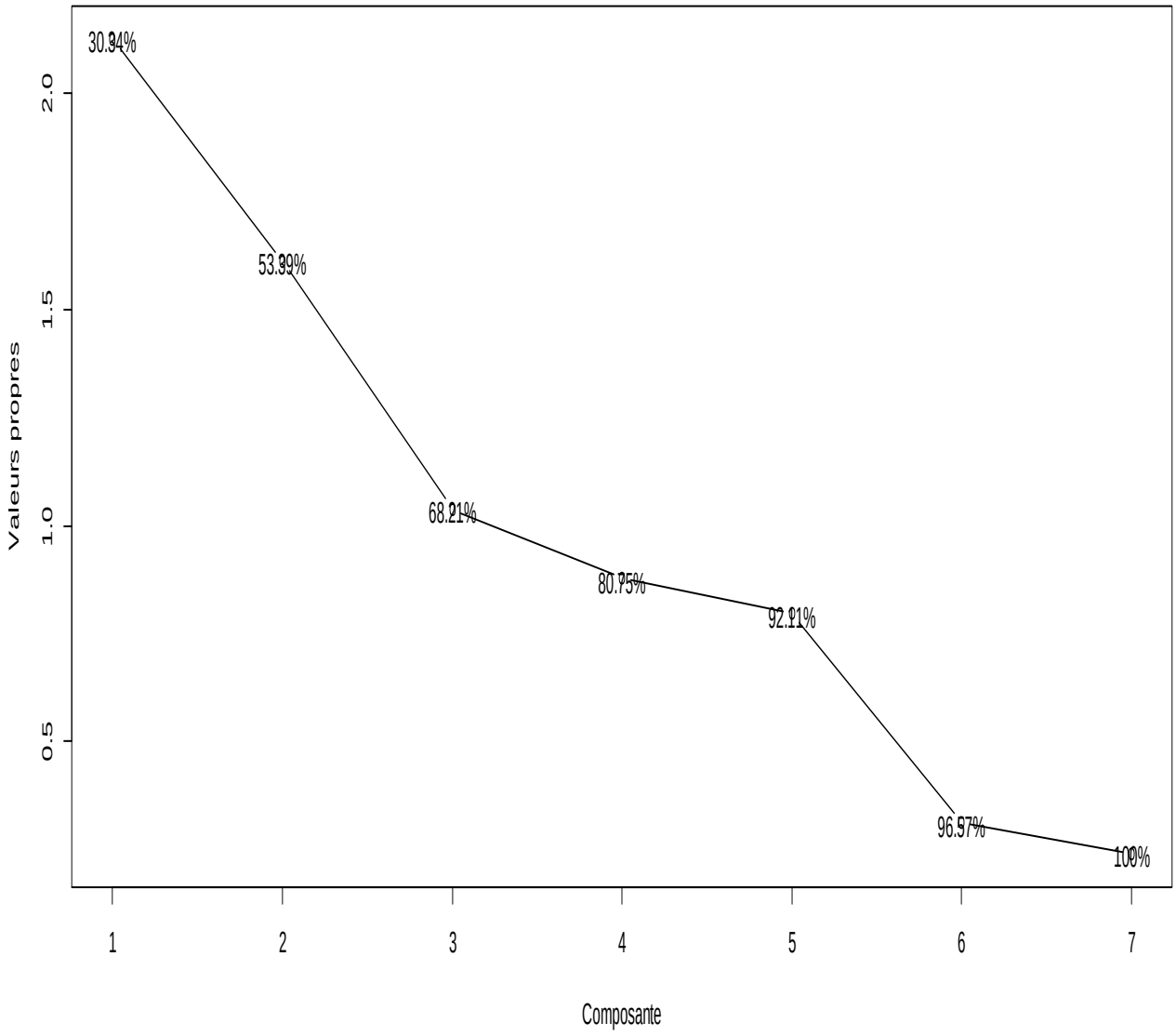
**Variables actives de l'ACP** : Colonnes de 1 à 7 : Nombre d'oiseaux relevés par les observateurs, nombre de passages sur les mailles, l'indice de nidification et les superficies respectives de forêt, de milieux ouverts, de zones artificialisées et de zones humides.

**Variables illustratives** : Colonne 8, la variable binaire-qualitative dichotomique « presence » ;
Colonne 9, la variable qualitative « annee » à 2 niveaux.

#### VII.4.2. Eboulis de valeurs propres

Ce graphique permet de sélectionner le nombre de composantes principales à conserver dans l'analyse des données. Pour $j \in \{1,\dots,7\}$, le poids de la j-ème composante principale est mesuré par le rapport de la j-ème valeur propre de la matrice des composantes principales empiriques C et la somme de toutes les valeurs propres de cette matrice C. (cf. Annexe 3).

Autrement dit, chaque point j du graphique estime le pourcentage de la variance expliquée par la j-ème composante principale.




**Graphe 2. Eboulis des valeurs propres**

En pratique, on cherche un « coude » sr l'éboulis des valeurs propres et on admet qu'il est suffisant de conserver uniquement comme nombre d'axes, le nombre de points avant ce coude. On entend par coude, le décrochement du graphique suivi d'une décroissance régulière.

Sur notre graphique, le coude se situe au niveau du 3$^{e}$ point. En effet, au-delà de ce point, la pente de la droite semble décroître moins rapidement.

Sous ce principe, la procédure PCA a retenu les deux premières composantes principales pour résumer les données variables-individus.

A elles deux, elles expliquent 53.39% de la variance totale des individus.

## VII.4.3. Résultats de l'analyse sur les variables.

### o Corrélations entre les variables et les axes retenus.

| | **Dim.1** | **Dim.2** | Dim.3 | Dim.4 | Dim.5 |
|---|---|---|---|---|---|
| nois | **0.5673777** | **0.5853252** | 0.056315325 | 0.25550449 | -0.39330158 |
| npass | **0.7394334** | **0.5059378** | -0.007110162 | 0.16309611 | -0.03882961 |
| nidif | **0.2659144** | **0.1369404** | 0.827761544 | -0.05452742 | 0.46523905 |
| foret | **-0.6971414** | **0.5884697** | 0.054306679 | 0.18335564 | 0.07161659 |
| milouv | **0.5490088** | **-0.6674786** | 0.237417813 | -0.08306312 | -0.29830431 |
| zonart | **0.5184516** | **-0.2109132** | -0.457522603 | 0.38586427 | 0.55217361 |
| zonhum | **0.3604409** | **0.3979133** | -0.283236399 | -0.77042138 | 0.15417783 |
| presence | **0.229553** | **0.1575305** | 0.5805996 | -0.002758422 | 0.2648155 |

Ce tableau fait état des corrélations existant entre les variables actives de l'analyse et les composantes principales retenues à partir de l'éboulis de valeurs propres. Les deux colonnes Dim1. Et Dim. 2 sont respectivement les coefficients de corrélation des variables actives avec la 1$^{e}$ et la 2$^{e}$ composante principale.

Le nombre de passages est la variable la plus expliquée par la 1$^{e}$ composante principale avec une corrélation de 74%. Mais également, le nombre d'oiseaux relevés sur les mailles (corrélation de 57%), les milieux ouverts (corrélation de 55%) et les zones artificialisées (corrélation de 52%). A l'opposé, la forêt est négativement corrélée à cette composante de 70%.

La 2$^{e}$ composante est explicative de la forêt (corrélation de 59%), du nombre d'oiseaux (corrélation de 58%), du nombre de passages des observateurs (corrélation de 50%). Les milieux ouverts sont négativement corrélées à cette composante (corrélation de 67%).

La variable illustrative « presence » n'est pas bien expliquée par les deux composantes principales. En effet, elle est faiblement corrélée aux 1$^{e}$ et 2$^{e}$ axes respectivement de 22% et 15%. Cependant, plus corrélée avec la première dimension, elle semble dépendante du nombre d'oiseaux relevés par les observateurs, le nombre de passages sur les mailles, les milieux ouverts et les zones artificialisées.

### o Contributions relatives des variables aux axes

| | **Dim.1** | **Dim.2** | Dim.3 | Dim.4 | Dim.5 |
|---|---|---|---|---|---|
| nois | **0.32191747** | **0.34260562** | 0.0031714159 | 0.065282545 | 0.154686130 |
| npass | **0.54676180** | **0.25597304** | 0.0000505544 | 0.026600343 | 0.001507739 |
| nidif | **0.07071046** | **0.01875268** | 0.6851891738 | 0.002973240 | 0.216447378 |
| foret | **0.48600608** | **0.34629655** | 0.0029492154 | 0.033619289 | 0.005128936 |
| milouv | **0.30141071** | **0.44552773** | 0.0563672177 | 0.006899482 | 0.088985459 |
| zonart | **0.26879206** | **0.04448436** | 0.2093269319 | 0.148891236 | 0.304895696 |
| zonhum | **0.12991768** | **0.15833502** | 0.0802228579 | 0.593549100 | 0.023770803 |

Nous nous intéressons ici à la contribution relative des variables actives aux deux composantes principales retenues. Par contribution relative, on entend la proximité de la variable donnée à la composante principale. Une valeur de contribution relative proche de 1 (100%) signifie que la variable donnée est proche de la composante principale, autrement dit elle est essentiellement caractérisée par cette composante.

Ainsi, les variables caractérisées par le 1e axe sont le nombre de passages sur les mailles et la forêt. Les milieux ouverts et le nombre d'oiseaux sont caractérisés par le 2e axe.

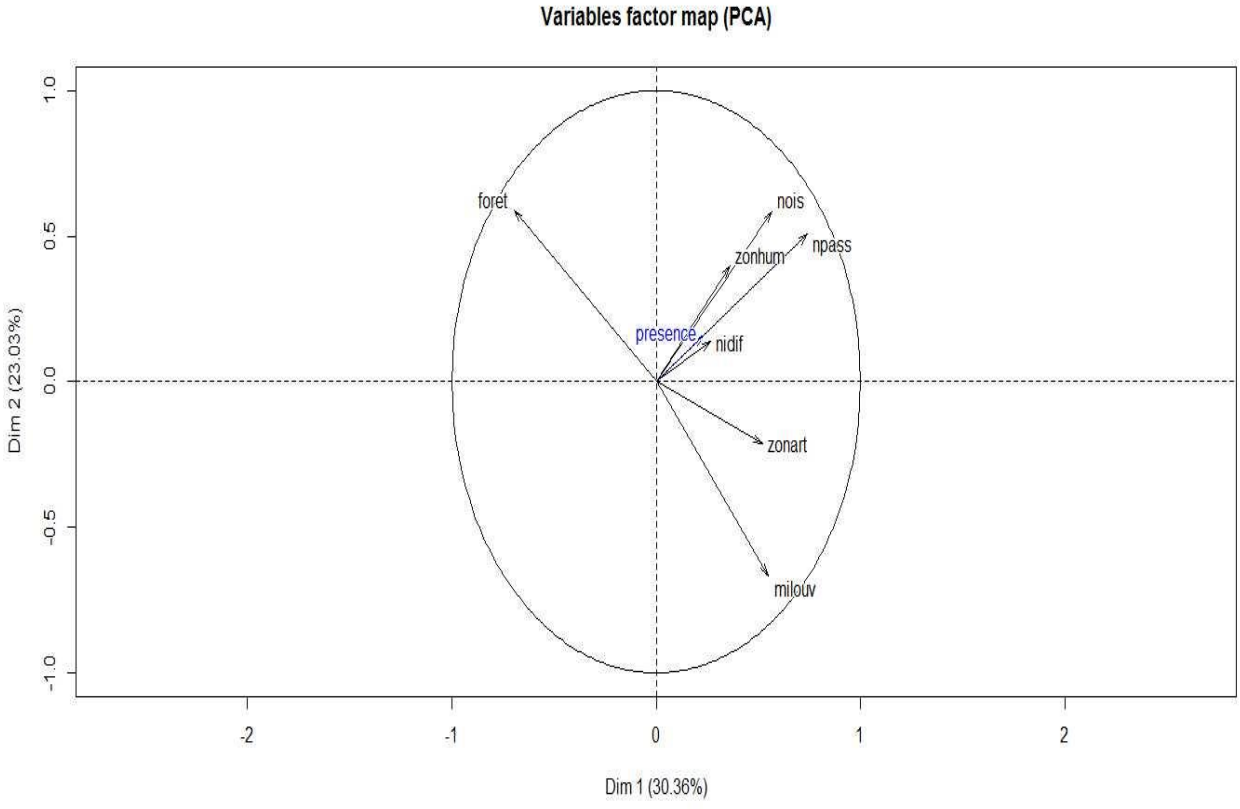
- **Contributions absolues des variables aux axes retenus**

Ce tableau marque les contributions des variables à la construction des composantes principales. Une valeur de contribution absolue proche de 100 signifie que la variable donnée contribue essentiellement seul à la création de la composante principale. En d'autres termes, cette composante principale a été créée principalement pour modéliser cette variable.

| | **Dim.1** | **Dim.2** | Dim.3 | Dim.4 | Dim.5 |
|---|---|---|---|---|---|
| Nois | **15.145378** | **21.253780** | 0.30574425 | 7.4369346 | 19.4470485 |
| npass | **25.723717** | **15.879467** | 0.00487376 | 3.0302895 | 0.1895520 |
| nidif | **3.326743** | **1.163335** | 66.05650481 | 0.3387090 | 27.2116361 |
| foret | **22.865320** | **21.482749** | 0.28432274 | 3.8298822 | 0.6448068 |
| milouv | **14.180588** | **27.638625** | 5.43415093 | 0.7859834 | 11.1871991 |
| zonart | **12.645966** | **2.759619** | 20.18042026 | 16.9615689 | 38.3313062 |
| zonhum | **6.112288** | **9.822424** | 7.73398326 | 67.6166324 | 2.9884513 |

Le nombre de passages est la variable ayant le plus grand apport sur la 1e composante principale, suivi de la forêt et des milieux ouverts. Les milieux ouverts, la forêt et le nombre d'oiseaux sont ceux qui ont le plus contribué à la construction de la 2e composante principale.

- **Cercle de corrélation des variables**



Comme son nom l'indique, ce graphique fait état des corrélations existant entre les variables actives de l'ACP. Plus généralement :

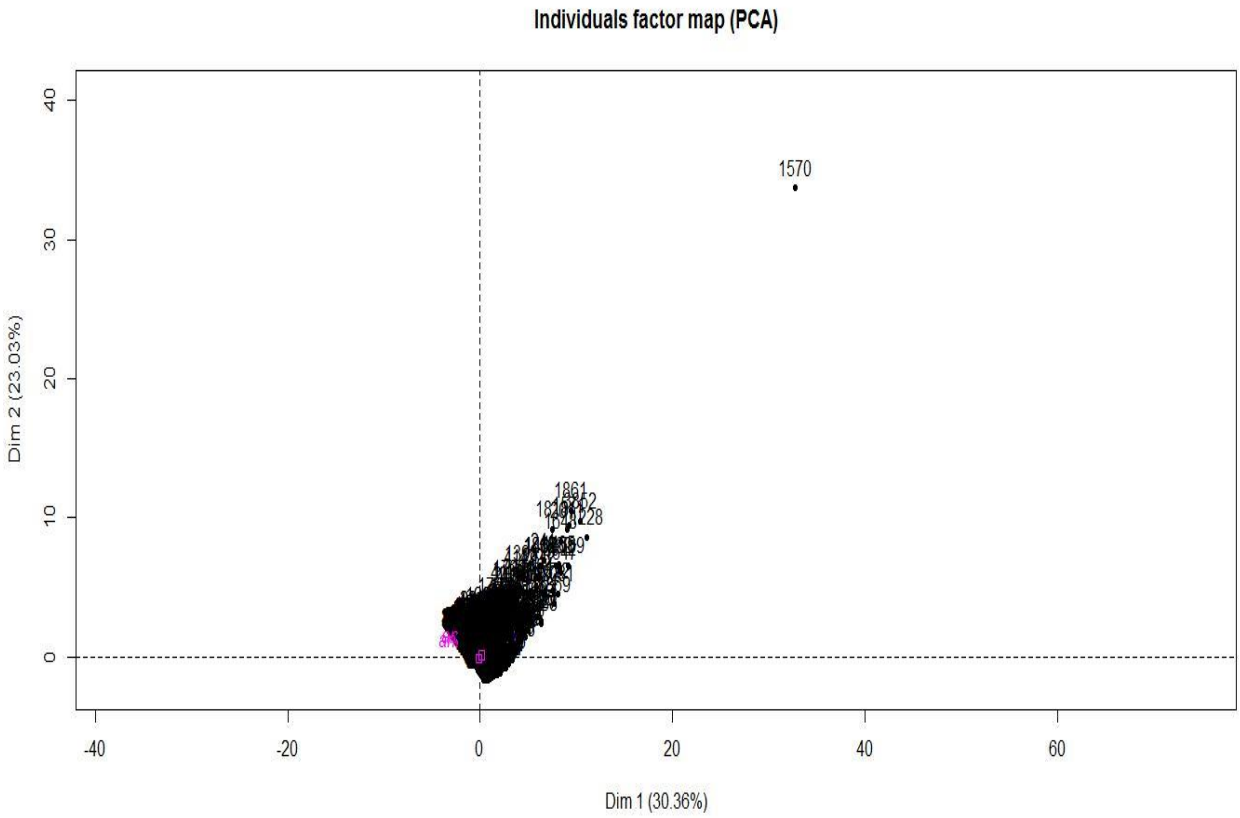
- Une variable qui se trouve en projection sur ce cercle est parfaitement déterminée sur le plan ;
- Une variable proche de la bordure du cercle est dite « bien représentée » dans le plan ;
- Une variable proche de l'origine du cercle est par contre une variable mal représentée.

En définitive, pour interpréter ce graphique, on sélectionne les variables bien expliquées et on analyse les proximités de ces variables sélectionnées en termes de corrélations.

Ainsi, les variables les mieux représentées sur les deux axes sont le nombre d'oiseaux, le nombre de passages, la forêt, les milieux ouverts et les zones artificialisées. L'observation de la Pie-grièche écorcheur (Variable « presence ») n'est pas bien représentée sur les plans factoriels.

Les milieux ouverts sont négativement corrélés aux territoires forestiers. Et tous deux sont faiblement corrélés au nombre d'oiseaux et au nombre de passages, qui sont positivement corrélés entre eux. C'est dire donc que le nombre d'oiseaux notés par les observateurs croît en fonction du nombre de passages sur les mailles. Aussi, la constitution des mailles (2*2) en terme de forêts et de milieux ouverts est opposée, c'est-à-dire plus les espaces forestiers occupent une grande partie de la maille, moins il y'aura de milieux ouverts, et vice versa.

### VII.4.4. Interprétation de l'analyse sur les individus




Ce graphe représente la distribution des individus de cette 1e table de données suivant les deux composantes principales retenues.

On observe un point levier, le point N° 1570 correspondant à la maille E1040N6734 présentant de fortes valeurs sur chacun des deux axes.

Or ces deux axes sont explicatifs essentiellement du nombre de passages, du nombre d'oiseaux, des milieux ouverts et des territoires forestiers. Et en effet, cette maille présente 263 passages avec 12744 oiseaux sur 217 ha en milieux ouverts et 72 ha forêt. Ce qui confirme nos interprétations par rapport aux contributions relatives, aux contributions absolues et à la corrélation des variables actives bien représentées.

A l'exception de ce point, d'autres interprétations seront difficiles à faire sur ce graphique car les mailles (2*2) km sont nombreuses (4564 sur 2011/2012) et forment un nuage de points condensés autour des deux composantes principales. Il est quasi impossible de les discerner les unes des autres pour apprécier leur distribution suivant les deux axes.

On va donc s'intéresser à deux résultats plus précis, fournis par le package FactoMineR : le tableau des contributions relatives (CTR) et le tableau des contributions absolues (CTA), relatives aux individus. Comme pour les variables, ces deux sorties correspondent respectivement aux individus caractérisés par les deux composantes principales et les individus ayant le plus contribué à la construction des deux composantes principales.

On peut également s'intéresser à la représentation des individus selon les modalités de la variable illustrative « annee ». Un coloriage différent sera appliqué aux individus de chaque année afin d'apprécier leur représentation.

### o **Tableau des contributions relatives des individus**

La contribution relative d'un individu (CTR) mesure sa proximité avec une composante principale. Elle est égale au carré de la corrélation de l'individu avec un axe donné. Une valeur de la CTR proche de 1 signifie que l'individu donné est très proche de cette composante principale. Autrement, cet individu est essentiellement caractérisé par les variables ayant la plus forte CTR par rapport à cette composante principale.

Rappelons que les variables caractérisées par le 1$^{e}$ axe sont le nombre de passages, le nombre d'oiseaux, la forêt et les milieux ouverts. Mais qu'aussi la forêt est négativement corrélée aux milieux ouverts et ils sont tous deux faiblement corrélée au nombre d'oiseaux et au nombre de passages des observateurs. On peut donc se douter que ces mailles bien situés sur le 1$^{e}$ axe sont les mailles avec de grandes superficies en espaces forestiers et/ou de milieux ouverts mais non prospectées et par conséquent, aucune donnée recensée dessus.

En effet, c'est le cas des mailles E1026N6816 (0 passages, 0 oiseaux sur 234 ha de forêt et 150 ha de milieux ouverts), E1006N6798 (0 passages, 0 oiseaux sur 250 ha de forêt et 100 ha de milieux ouverts), E1026N6816 (0 passages, 0 oiseaux sur 234 ha de forêt et 150 ha de milieux ouverts), E1016N6756 (0 passages, 0 oiseaux sur 234 ha de forêt et 143 ha de milieux ouverts), E998N6758 (1 passage, 1 oiseau noté sur 228 ha de forêt et 151 ha de milieux ouverts), E998N6752 (0 passages, 0 oiseaux sur 368 ha de forêt et 32 ha de milieux ouverts). Ces mailles ont toutes des valeurs de CTR, par rapport à la 1$^{e}$ composante, supérieures à 0.80%.

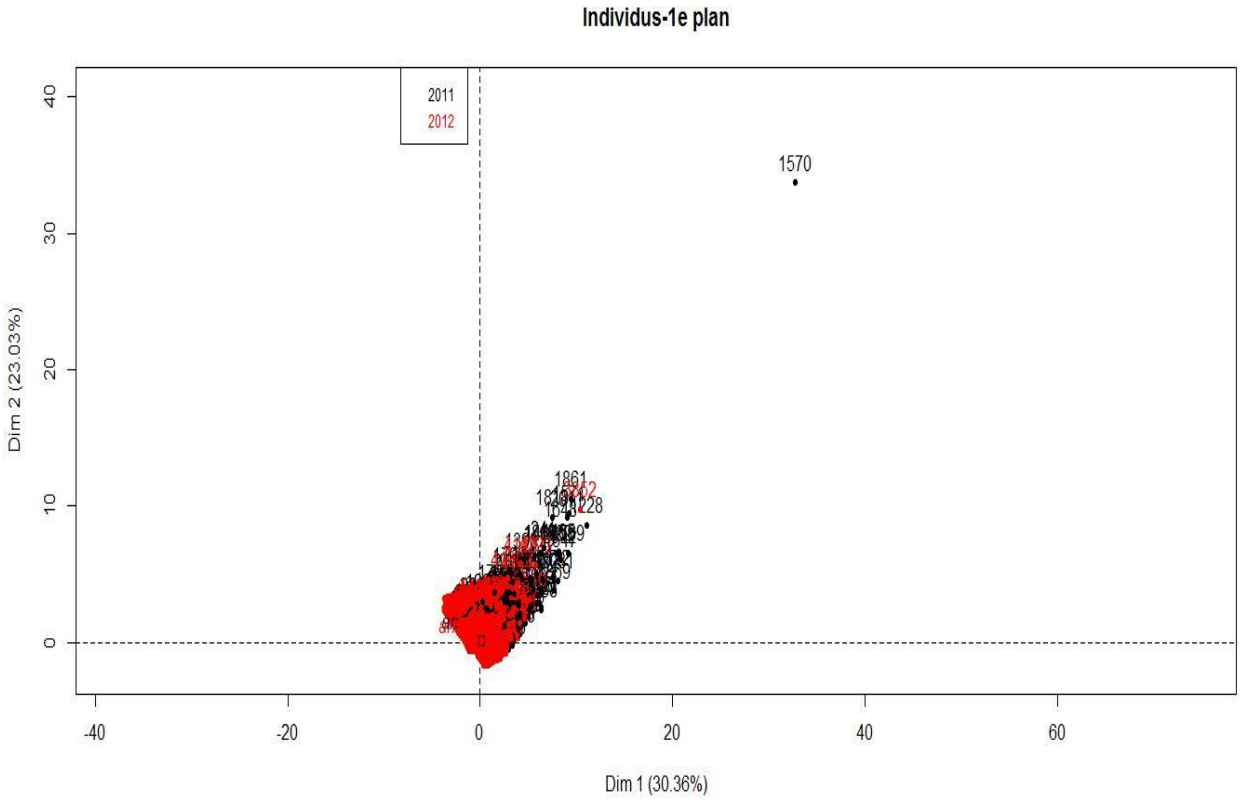
Il en est de même pour la 2$^{e}$ composante qui caractérise le même groupe de variables. Les mailles à forte CTR par rapport à cette composante, c'est-à-dire les mailles constituées en majorité de forêt et de milieux ouverts mais avec aucune observation d'oiseaux et peu de prospections. Il s'agit des mailles E1052N6858, E1052N6858, E1052N6882, E1028N6768, E1040N6806.

- **Tableau des contributions absolues des individus**

Les valeurs de ce tableau correspondant aux contributions de chaque individu à la construction des deux composantes principales. En d'autres termes, une valeur de CTA proche de 100 sur un axe signifie que cet axe a été créé essentiellement pour modéliser l'individu en question.

Les mailles (2*2) km ayant le plus contribué à la construction de la 1$^{e}$ composante sont les mailles E1040N6734 (12744 oiseaux pour 263 passages dans une zone constitué de 73 ha de forêt où l'espèce a été vue), E1032N6732, E1040N6734, E1050N6826. Toutes ces mailles présentent de fortes valeurs en nombre de passages-corrélée au nombre d'oiseaux-et en superficie de forêt, variables étant caractéristiques des deux composantes principales.

- **Représentation des individus suivant l'année d'observation**




Ce graphique montre la distribution des individus suivant les deux composantes principales. Il a été choisi de le faire, vu la qualité représentative très condensée obtenue précédemment mais surtout pour faciliter l'interprétation des résultats sur les mailles (2*2) km.

De façon générale, on constate que les individus sont bien représentés dans le repère formé par les deux composantes principales. Mais toutefois, on peut remarquer que le plus gros du nuage d'individus sont de coordonnées positives sur les deux axes. Il s'agit des mailles les plus prospectées, ayant donc un grand nombre d'oiseaux recensés dessus, et à majorité constituée de forêt et de milieux ouverts et de zones artificialisées. Ce constat est surtout fait en 2012.

Notamment, on peut distinguer la maille E1040N6734 qui en 2012 a fait l'objet de 92 passages sur 217 ha de milieux ouverts, 73 ha de forêt et 72 ha de zones artificialisées. Aussi, la maille E1048N6824 qui a été prospecté sur la même année 45 fois sur 222 ha de forêt et 85 ha de milieux ouverts.

Or on a montré plus haut que l'observation de l'espèce semble évoluer dans le sens de ces variables à l'exception de la forêt (négativement corrélée à la 1e composante principale), ce qui se confirme sur ces deux mailles car l'espèce y a été notée.

En outre, les mailles (2*2) km sont également nombreuses sur les coordonnées positives du 1e axe et négatives sur celles du 2e axe. C'est-à-dire que les prospections ont été plus fréquentes dans les espaces forestiers et moins dans les milieux ouverts, surtout en 2012.

## VII.5. Mise en œuvre de l'ACP sur la deuxième table de données

### VII.5.1. Répartition des variables de la table de données

Cette 2e table de données est la même, utilisée dans la régression logistique et également présentée dans la section IV.2 de ce document.

**Variables actives - Colonnes de 1 à 11** : Nombre d'oiseaux relevés, nombre de passages, l'indice de nidification et les superficies respectives en cultures annuelles, cultures permanentes, eau, zones artificialisées, forêt, milieux ouverts, prairies et landes et en vergers.

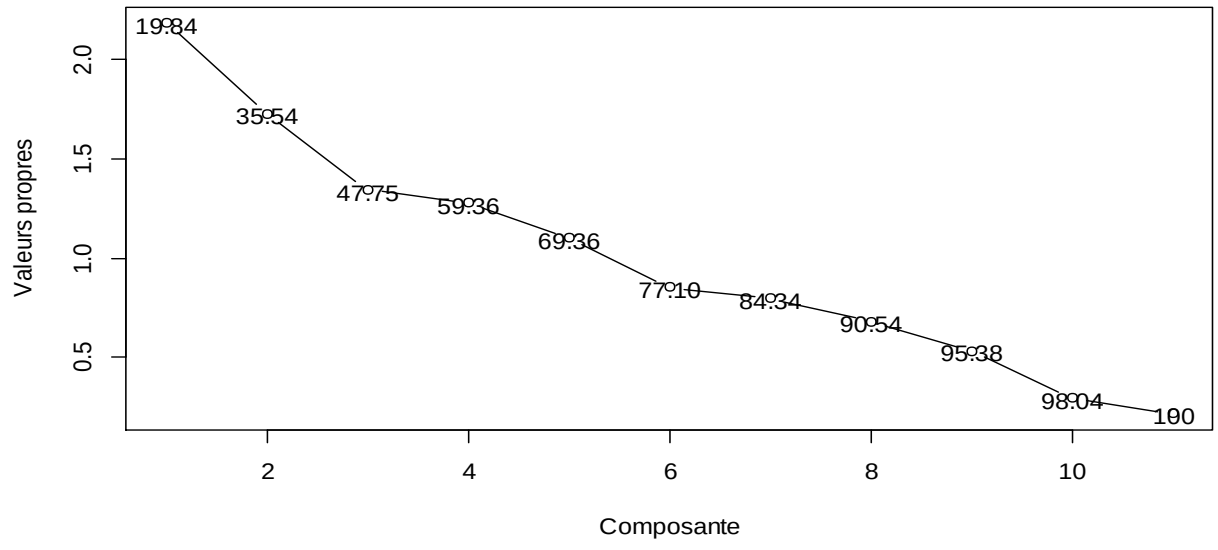
**Variables illustratives/supplémentaires** :

Colonne 12, la variable binaire-qualitative dichotomique « presence » ;
Colonne 13, la variable qualitative « annee » à 2 niveaux.

### VII.5.2. Eboulis de valeurs propres

Comme vu précédemment, ce graphique est construit pour nous permettre de définir le nombre de composantes principales à retenir pour l'ACP. (cf. Annexe 3)

Le nombre d'axes correspond au nombre de points placés avant le « coude ». On entend par « coude », le décrochement abrupt de la courbe à partir duquel la pente semble devenir régulière. Sur notre graphique, ce coude se situe à partir du 3e point. L'ACP va se faire sur deux composantes principales qui à elles deux, représentent 35.54% de la dispersion des individus.



## VII.5.3. Résultats de l'analyse sur les variables

### o Corrélations entre les variables et les axes retenus

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 |
|---|---|---|---|---|---|
| nois | **0.62187333** | **0.3675067** | 0.0209861924 | -0.38273109 | 0.09722833 |
| npass | **0.81268788** | **0.3261999** | 0.0009316096 | -0.19730609 | 0.08981134 |
| nidif | **0.38560574** | **0.1649352** | 0.4603869757 | -0.28329608 | 0.01030772 |
| culann | **0.26047893** | **-0.7143393** | -0.3451889592 | -0.30979664 | -0.20572941 |
| culper | **0.06965664** | **-0.1100814** | -0.0108043574 | 0.11253052 | 0.90757155 |
| eau | **0.39378427** | **0.3163800** | -0.3531440463 | -0.04358804 | -0.18775198 |
| zonart | **0.54078419** | **-0.1701671** | -0.0907229892 | 0.62658365 | 0.07126036 |
| foret | **-0.55396126** | **0.7018698** | 0.2265775802 | -0.01739262 | 0.02702404 |
| milouv | **0.31430663** | **0.3152878** | -0.0364356179 | 0.70315250 | -0.19790644 |
| prailan | **0.18272199** | **-0.2019158** | **0.6995795721** | 0.09461608 | -0.33339113 |
| verg | **0.16660412** | **-0.4169091** | **0.5797818552** | 0.06509504 | 0.15864888 |
| presence | **0.2568418** | **0.09071464** | **0.3541903** | -0.1848708 | 0.01911229 |

Les variables les plus corrélées au 1e axe sont le nombre de passages des observateurs, le nombre d'oiseaux relevés et les zones artificialisées. La forêt est négativement corrélée à la 1e dimension (corrélation de -55%).

Les cultures annuelles sont négativement corrélées avec la 2e composante qui elle, est explicative des espaces forestiers. En anticipation du cercle de corrélations, on se doute que les cultures annuelles seront négativement corrélées aux espaces forestiers.

L'observation de la Pie-grièche écorcheur est expliquée à 25% par le 1e axe et 9% par le 2e axe. Elle n'est donc pas bien représentée sur les plans factoriels. Cependant, de par sa corrélation positive avec ces axes, on peut dire qu'elle est dépendante du nombre de passages des observateurs, du nombre d'oiseaux relevés, des zones artificialisées et des cultures annuelles.

### o Contributions relatives des variables aux axes

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 |
|---|---|---|---|---|---|
| nois | **0.386726432** | **0.13506119** | 4.404203e-04 | 0.1464830836 | 0.0094533475 |
| npass | **0.660461595** | **0.10640639** | 8.678964e-07 | 0.0389296936 | 0.0080660774 |
| nidif | **0.148691787** | **0.02720362** | 2.119562e-01 | 0.0802566684 | 0.0001062490 |
| culann | **0.067849275** | **0.51028065** | 1.191554e-01 | 0.0959739586 | 0.0423245911 |
| culper | **0.004852047** | **0.01211791** | 1.167341e-04 | 0.0126631174 | 0.8236861274 |
| eau | **0.155066050** | **0.10009628** | 1.247107e-01 | 0.0018999176 | 0.0352508071 |
| zonart | **0.292447537** | **0.02895683** | 8.230661e-03 | 0.3926070756 | 0.0050780387 |
| foret | **0.306873082** | **0.49262125** | 5.133740e-02 | 0.0003025034 | 0.0007302989 |
| milouv | **0.098788656** | **0.09940641** | 1.327554e-03 | 0.4944234361 | 0.0391669580 |
| prailan | **0.033387324** | **0.04076998** | 4.894116e-01 | 0.0089522029 | 0.1111496473 |
| verg | **0.027756932** | **0.17381324** | 3.361470e-01 | 0.0042373648 | 0.0251694680 |
| presence | **0.06596769** | **0.008229145** | 0.1254508 | 0.03417721 | 0.0003652796 |

Le nombre d'oiseaux, le nombre de passages et les zones artificialisées sont caractérisés par la 1e composante principale. Les cultures annuelles et la forêt par la 2e composante principale.

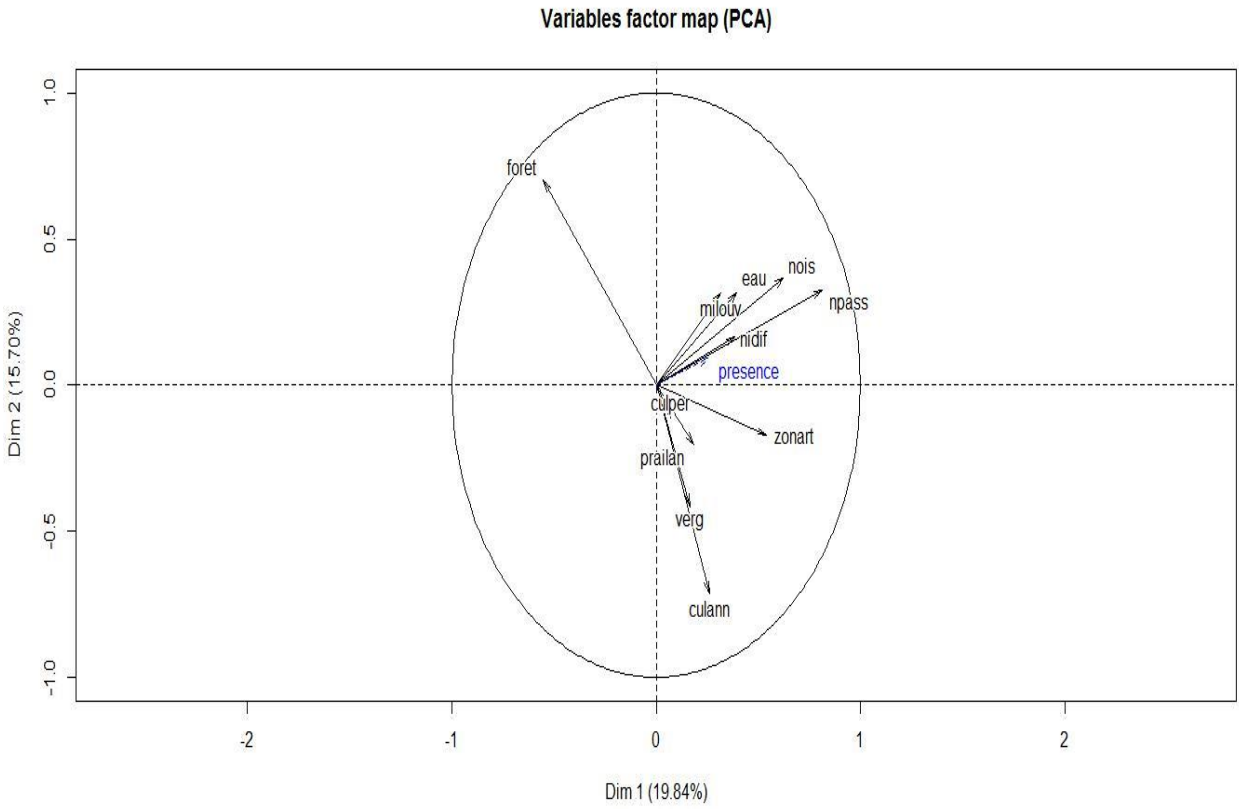
On remarque encore que la présence de l'espèce n'est expliquée par aucun des axes retenus avec respectivement des corrélations de 0.06 avec le 1e axe et 0.008 avec le 2e axe.

- **Contributions absolues des variables aux axes**

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 |
|---|---|---|---|---|---|
| nois | 17.7161714 | 7.8217728 | 3.279781e-02 | 11.47331039 | 0.859253361 |
| npass | 30.2561445 | 6.1622930 | 6.463167e-05 | 3.04917433 | 0.733158723 |
| nidif | 6.8116605 | 1.5754382 | 1.578424e+01 | 6.28611608 | 0.009657408 |
| culann | 3.1082163 | 29.5517849 | 8.873425e+00 | 7.51717529 | 3.847054952 |
| culper | 0.2222752 | 0.7017821 | 8.693114e-03 | 0.99184065 | 74.868196260 |
| eau | 7.1036694 | 5.7968567 | 9.287125e+00 | 0.14881134 | 3.204089834 |
| zonart | 13.3971983 | 1.6769715 | 6.129319e-01 | 30.75101050 | 0.461563677 |
| foret | 14.0580412 | 28.5290798 | 3.823062e+00 | 0.02369362 | 0.066379850 |
| milouv | 4.5255680 | 5.7569045 | 9.886209e-02 | 38.72579283 | 3.560044782 |
| prailan | 1.5294935 | 2.3611039 | 3.644616e+01 | 0.70118269 | 10.102845409 |
| verg | 1.2715618 | 10.0660125 | 2.503265e+01 | 0.33189226 | 2.287755743 |

Comme on le constate toujours, le nombre de passages et le nombre d'oiseaux sont les variables ayant le plus contribué au 1e axe. Les cultures annuelles et les espaces forestiers au 2e axe. En d'autres termes, les deux composantes principales modélisent essentiellement ces 4 variables.

- **Cercle de corrélation entre les variables**.

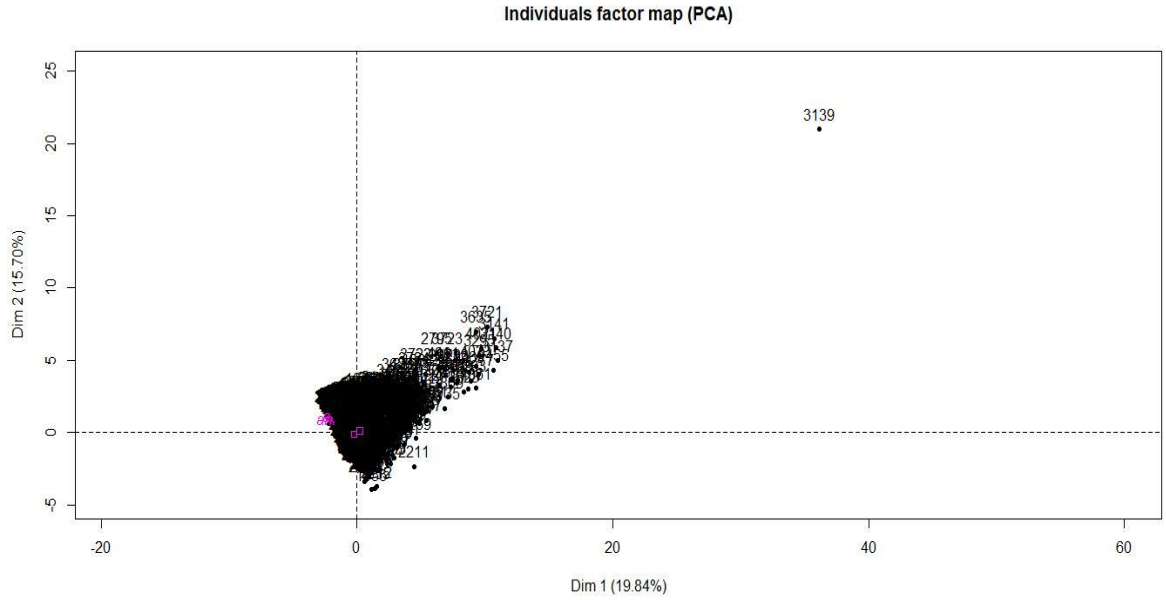


Les deux composantes principales retenues représentent 35,54% de la variance totale des données. Les variables les plus expliquées par ces composantes principales sont le nombre d'oiseaux, le nombre de passages, la forêt, les zones artificialisées et les cultures annuelles.

Les cultures annuelles sont négativement corrélées à la forêt (résultat prévisible avec le tableau des corrélations entre variables et composantes principales), et toutes deux sont faiblement corrélées au nombre d'oiseaux relevés et au nombre de passages. Ces deux dernières variables restent positivement corrélées entre elles.

Les prairies et landes et les vergers ne sont pas bien représentés sur les plans factoriels, mais cependant, ils sont positivement corrélés aux zones artificialisées et aux cultures annuelles.

### VII.5.3. Résultats de l'analyse sur les individus




Le point remarquable de la distribution des individus est la maille N°3140-E1040N6734 qui se détache du nuage des points et qui présente de fortes valeurs aussi bien pour la 1$^{e}$, que pour la 2$^{e}$ composante principale. En effet, cette maille est caractérisée par le nombre d'oiseaux élevés qui a été noté dessus (12744) au terme de 263 passages. Elle est constituée de 105 ha de cultures annuelles, 73 ha de forêt et 72 ha de zones artificialisées. Cette maille est donc essentiellement caractérisée par les variables que l'on a trouvés étant plus expliquées par les deux composantes principales. Ce qui confirme les résultats de l'ACP.

Le nuage de points des individus étant fortement concentré autour de l'origine du repère, nous nous intéressons aux tableaux de contributions relatives et de contributions absolues des individus par rapport aux deux composantes principales.

- **Tableau des contributions relatives des individus (CTA)**

Ce tableau montre l'ensemble des individus caractérisés par les composantes principales retenues, autrement dit les individus caractérisés par les variables notées comme les mieux expliquées par les composantes principales de l'analyse.

Ainsi, les mailles E1044N6822, E1024N6720, E990N6874, E1044N6822, E1026N6752, E1024N6720 avec des CTR de plus de 0.7 par rapport à la 1e composante, sont caractérisés par le nombre d'oiseaux relevés dessus, les passages des observateurs et les superficies de zones artificialisées. Les cultures annuelles et la forêt sont caractéristiques des mailles E1040N6868, E1030N6726, E1062N6878, E1024N6852, E1064N6884 avec des CTR de plus de 0.8 par rapport à la 2e composante.

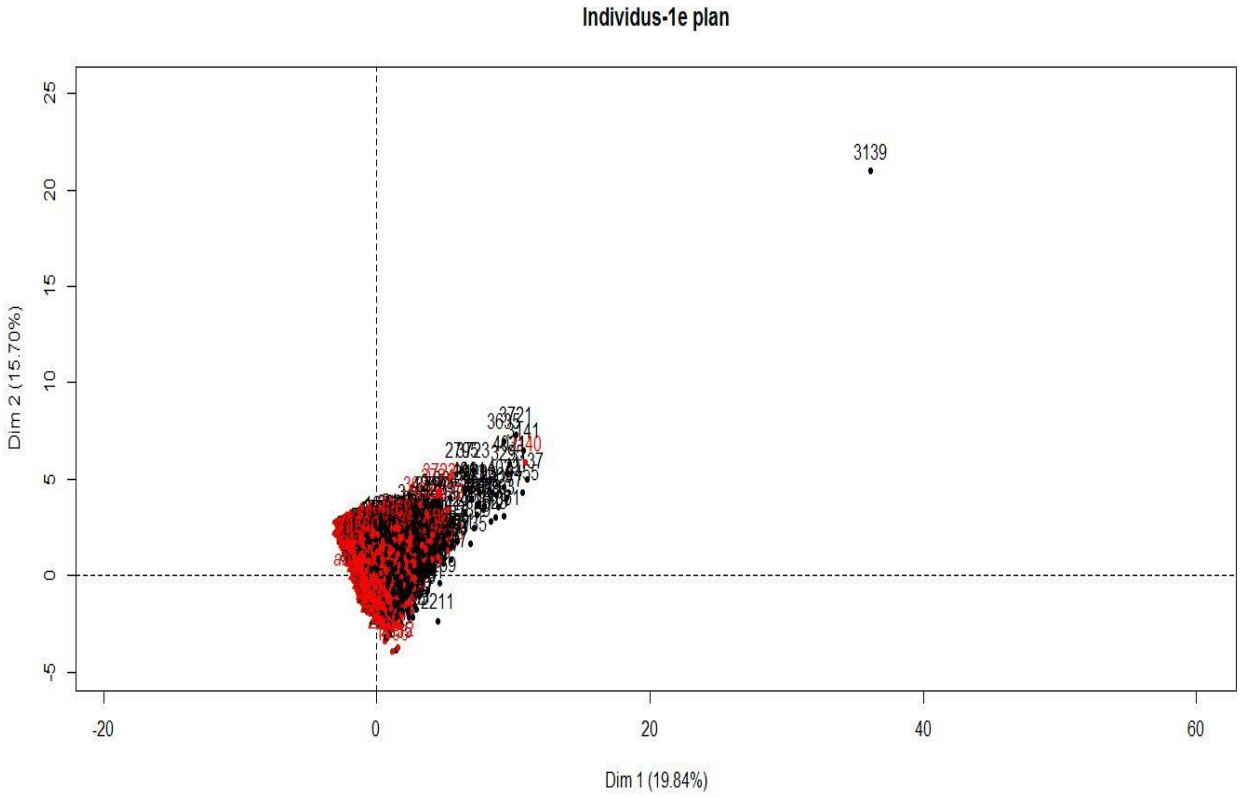
- **Tableau des contributions absolues des individus (CTA)**

Ce tableau détermine les individus ayant le plus contribué à la construction des composantes principales de l'ACP. Autrement dit, elle permet de déterminer les mailles (2*2) km essentiellement modélisées par les composantes principales.
De tels individus ont une valeur de CTA très proche de 100.

Cependant, aucune des mailles ne semblent avoir grandement contribué aux axes retenus. A l'exception de la maille E1040N6734, représentant le point levier sur la graphique de représentation des individus, toutes les autres mailles ont des CTA inférieures à 6.

Toutefois, on peut citer les mailles E1040N6734, E1040N6732, E1040N6736, E1050N6826 qui ont le plus contribué à la construction de la 1e composante et les mailles E1048N6824, E1040N6736, E1040N6734 à la construction de la 2e composante principale.

- **Représentation des individus suivant l'année d'observation.**



Les mailles (2*2) km sont réparties en majorité sur les coordonnées positives des deux composantes principales surtout en 2011 (abondance de points en noir). Il s'agit des mailles qui ont été les plus prospectées, avec un important nombre d'oiseaux relevés sur des grandes superficies de cultures annuelles, de forêt et de zones artificialisées.

## VII.6. Comparaisons des sorties des ACP sur les deux tables de données

### ➢ Variance expliquée par les composantes principales

Les deux dimensions de l'ACP sur le 1e jeu de données représentent 53,39% de la variance totale. Sur le 2e jeu de données, la variance totale expliquée par les deux composantes principales retenues est de 30,36%.

Le but de l'ACP étant de maximiser la dispersion des individus projetés sur les composantes principales, l'analyse sur le 1e jeu de données semble mieux réduire les données et représenter les relations entre variables et entre individus.

### ➢ Variables bien représentées sur les dimensions

Les variables du 1e jeu de données qui sont bien représentées sur les dimensions sont le nombre de passages des observateurs, le nombre d'oiseaux relevés sur les mailles, les milieux ouverts, les zones artificialisées et la forêt. Pour le 2e jeu de données, il s'agit également, en plus des cultures annuelles, du nombre de passages des observateurs, du nombre d'oiseaux, des zones artificialisées et des espaces forestiers.

### ➢ Corrélations entre les variables bien représentées

Sur les deux jeux de données, le nombre d'oiseaux relevés sur les mailles est positivement corrélé au nombre de passages des observateurs sur les mailles. Les milieux ouverts sont négativement corrélés à la forêt sur le 1e jeu de données, et sur le 2e, c'est les cultures annuelles qui sont opposés à la forêt. Les zones artificialisées sont plus proches des milieux ouverts et des cultures annuelles que des autres variables.

### ➢ Explication de l'observation de la Pie-grièche écorcheur par les composantes principales

La variable « presence » relative à l'observation de l'espèce n'est expliquée par aucune des composantes principales des deux ACP. Elle est cependant plus corrélée aux premières dimensions des deux analyses notamment de 25% pour le 1e jeu de données et de 23% avec le 2e jeu de donnés. Ces axes étant explicatifs du nombre de passages, du nombre d'oiseaux, des milieux ouverts, des cultures annuelles et des zones artificialisées, on peut donc dire que l'observation de l'espèce est plus liée à ce groupe de variables. A l'opposé, elle est négativement corrélée aux espaces forestiers.

### ➢ Comparaisons des mailles expliquées par les deux composantes principales

Les individus sont représentés de façon très condensée sur les axes des deux ACP. Cependant, la maille N°3140-E1040N6734 est le point commun des deux analyses qui se présente comme un point levier avec la plus grande valeur sur les deux axes retenus. Sur cette maille a été relevé 12744 oiseaux au terme de 263 passages sur 105 ha de cultures annuelles, 72 ha de forêt et 73 ha de zones artificialisées.

Le plus grand nombre d'individus est réparti sur les coordonnées positives des deux composantes principales retenues surtout en 2012. C'est donc dire que la Pie-grièche écorcheur a été plus recensée en 2012 sur les cultures annuelles, les milieux ouverts et les zones artificialisées.

### VII.7 Conclusions générales sur les ACP

Nous avons choisi de réaliser une ACP sur chacune des tables de données afin de visualiser les relations existant entre les variables relatives à la Pie-grièche écorcheur mais aussi la dispersion des mailles (2*2) km suivant les axes factoriels construits dans les analyses.

Il en résulte que les variables les plus expliquées par l'ACP sont le nombre d'oiseaux relevés sur les mailles d'observation, le nombre de passages des observateurs, les zones artificialisées, les milieux ouverts, les cultures annuelles et les espaces forestiers.

La présence de l'espèce n'est pas bien représentée sur les cercles de corrélations. Cependant, elle est toujours corrélée positivement aux premières dimensions qu'aux deuxièmes. Or d'après les tableaux de contributions relatives et absolues des variables, on a montré que ces premiers axes sont caractéristiques du nombre d'oiseaux, du nombre de passages, des superficies de milieux ouverts, de zones artificialisées et de cultures annuelles.

A partir de ces résultats, on a pu en déduire que l'observation de l'espèce évolue selon ces variables. Et ceci est notamment vérifié en 2012 avec les coloriages appliqués sur les mailles (2*2) km suivant l'année d'observation. En effet, cette représentation des individus (mailles 2*2) nous a fait voir que la majorité des points sur 2012 sont répartis sur les coordonnées positives des deux composantes principales caractéristiques des variables citées précédemment.

On peut envisager une possibilité d'observer l'espèce également sur les vergers, les prairies et landes car de tels milieux écologiques sont positivement corrélés aux cultures annuelles. Mais cependant, elles ne sont pas bien représentées sur les plans factoriels.

## VIII. Comparaisons des sorties de la régression logistique et de l'ACP

La régression logistique a été appliquée sur les tables de données pour prédire une probabilité de présence de la Pie-grièche écorcheur, sur chacune des 2282 mailles (2*2) km prospectées par année en Alsace. Les probabilités ont été calculées à partir des coefficients associés aux prédicteurs retenus à l'issue d'une sélection BACKWARD faite sur l'ensemble des variables. Et par la suite, des tests de significativité à un niveau de 5% ont été faits sur ces prédicteurs sélectionnés pour mesurer leurs influences sur la variable réponse « presence ».

L'ACP a été mis en œuvre pour visualiser la structure des variables dans leur totalité, c'est-à-dire pour voir les corrélations existant entre elles et identifier les mailles (2*2) km caractérisées par les variables les mieux représentées sur les axes factoriels des analyses.

D'un point de vue résultats, les seuls pouvant être mis en relief et comparés entre ces deux méthodes sont les variables et interactions significatives sur l'observation probable de l'espèce.

Toutes ces méthodes ont permis de conclure à une observation plus probable de l'espèce sur les milieux ouverts, les lisières de foret représentées par l'interaction entre la forêt et les milieux ouverts, les zones artificialisées, les cultures annuelles, les prairies et landes mais également dans les espaces forestiers riches en avifaune.

Toutefois, les ACP montrent que la variable « presence » que l'on cherche à expliquer, n'est pas bien représentée sur les axes factoriels des deux analyses. En raison de son caractère binaire, elle a été considérée comme variable illustrative et n'intervient donc pas directement dans la construction des plans factoriels mais elle a servi à interpréter les dimensions de variabilité.

Par conséquent, nous pouvons dire que les tests de rapports de vraisemblance appliqués sur les variables semblent être les mieux adaptés pour mesurer l'impact des prédicteurs sur l'observation de la Pie-grièche écorcheur. Ces tests sont d'autant plus réalisés sur des modèles écrits avec des interactions de variables que l'ACP ne prend pas en compte.

# IX. Conclusions générales et Discussion

## IX.1. Résumé des résultats

Dans le but d'étudier l'évolution de la Pie-grièche écorcheur en Alsace sur 2011/2012 et des facteurs influant son observation sur un site, nous avons choisi d'appliquer une régression logistique et une analyse en composantes principales (ACP) aux données mises à notre disposition et organisées en deux tables.

Un choix de deux méthodes pour espérer obtenir de meilleurs résultats et pouvoir comparer les sorties. Ce qui permettrait de définir la meilleure modélisation qui puisse se faire sur des données relatives aux oiseaux.

Le pouvoir de prédiction du modèle 2 (posé sur la 2$^{e}$ table de données) étant supérieur au pouvoir de prédiction du modèle 1 (posé sur la 1$^{e}$ table de données), nous décidons de plus nous intéresser aux résultats obtenus sur cette 2$^{e}$ table.

Ainsi, on a pu voir que l'espèce s'est raréfiée sur 696 km² en Alsace entre 2011 et 2012. Son observation est plus probable sur les milieux ouverts, les clairières (anciens espaces forestiers défraichis), les zones artificialisées, les cultures annuelles, les prairies et landes et les espaces forestiers riches en avifaune. Elle est également conditionnelle à un important nombre de passages et par conséquent à un grand nombre d'oiseaux notés sur les mailles (2*2) km.

La variable réponse « presence » étant binaire, elle a été considérée comme variable illustrative dans les analyses en composantes principales. Cependant, elle n'est pas bien représentée sur les axes factoriels construits. C'est sur cette base que nous accordons plus de fiabilité aux tests du rapport de vraisemblance effectués sur les modèles de régression logistique pour mesurer l'influence des prédicteurs sur la réponse « presence ».

## IX.2. Perspectives

A l'issue de cette étude, d'autres zones de recherche peuvent être définies notamment sur la cause de la disparition de l'espèce sur les mailles ou elle a été notée sur une année et pas sur l'autre. Est-ce dû à un changement au niveau du milieu écologique (intensification de l'agriculture, installation de structures urbains, braconnage) ou aux migrations tardives de l'espèce dues aux changements climatiques.

Mais également, on peut voir comment enrichir le modèle de régression logistique et l'automatiser à toutes les données non protocolées disponibles sur d'autres espèces d'oiseaux saisies sur VisioNature par les utilisateurs.

### IX.3. Bilan du stage

Ce stage au sein d'ODONAT en plus d'être enrichissante, a été une énorme découverte professionnelle. Découverte, parce qu'appliquer des modélisations statistiques sur un domaine aussi hors du commun et intéressant qu'est le suivi et la sauvegarde des oiseaux, était inattendu. Cela m'a vraiment encore une fois prouvé l'étendue du domaine d'application des statistiques.

Au cours de ce stage, j'ai pu me rendre compte que les phases préliminaires de compréhension et de traitement des données sont cruciales pour préparer au mieux la modélisation avec les outils statistiques choisis. Ce processus m'a pris plus de la moitié des deux mois en entreprise. D'autre part, je me suis aperçu que les connaissances acquises en Master 1 n'étaient pas suffisantes pour mener à bien ce projet. Il a fallu énormément me documenter sur les méthodes statistiques utilisées et leur application sur les données mises à ma disposition.

Les résultats obtenus semblent toutefois refléter la réalité quant aux habitats potentiels de la Pie-grièche écorcheur et à la possibilité de pouvoir mener un suivi de l'évolution des espèces à partir de données non protocolées recueillies par les utilisateurs dans VisioNature.

## X. Bibliographie

- **Livres**

- Gilbert Saporta (2006), *« Probabilités, Analyse de données et Statistique »*, Editions Dunod.

- Michel Lejeune (2004), «*Statistiques – La théorie et ses applications* », Chapitre 6 Modèles à réponse dichotomique.

- **Articles.**

- Ricco Rakotomalala-Université de Lyon 2 « *Analyse en composantes principales avec les packages FactoMineR et dynGraph du logiciel R »,* pages 1-13.

- Renaud Lancelot et Matthieu LESNOFF (2005), « *Sélection de modèles avec l'AIC et critères d'information dérivés »*, Version 3.

- ODONAT (2012), « *Bilan des saisies naturalistes-première année(2011)-et perspectives 2012* », Région Alsace, Union européenne, Conseil général du Bas Rhin, 50 pages.

- ODONAT (2012), « *Suivi des Indicateurs de la Biodiversité en Alsace* », Rapport année 2012, Région Alsace, Département 67 et département 68, 140 pages.

- Jean-Paul Sampoux (2009) « *Innovations agronomiques-Modélisation de la niche écologique et gestion des ressources génétiques* », pages 79-91.

- **Sites Internet**

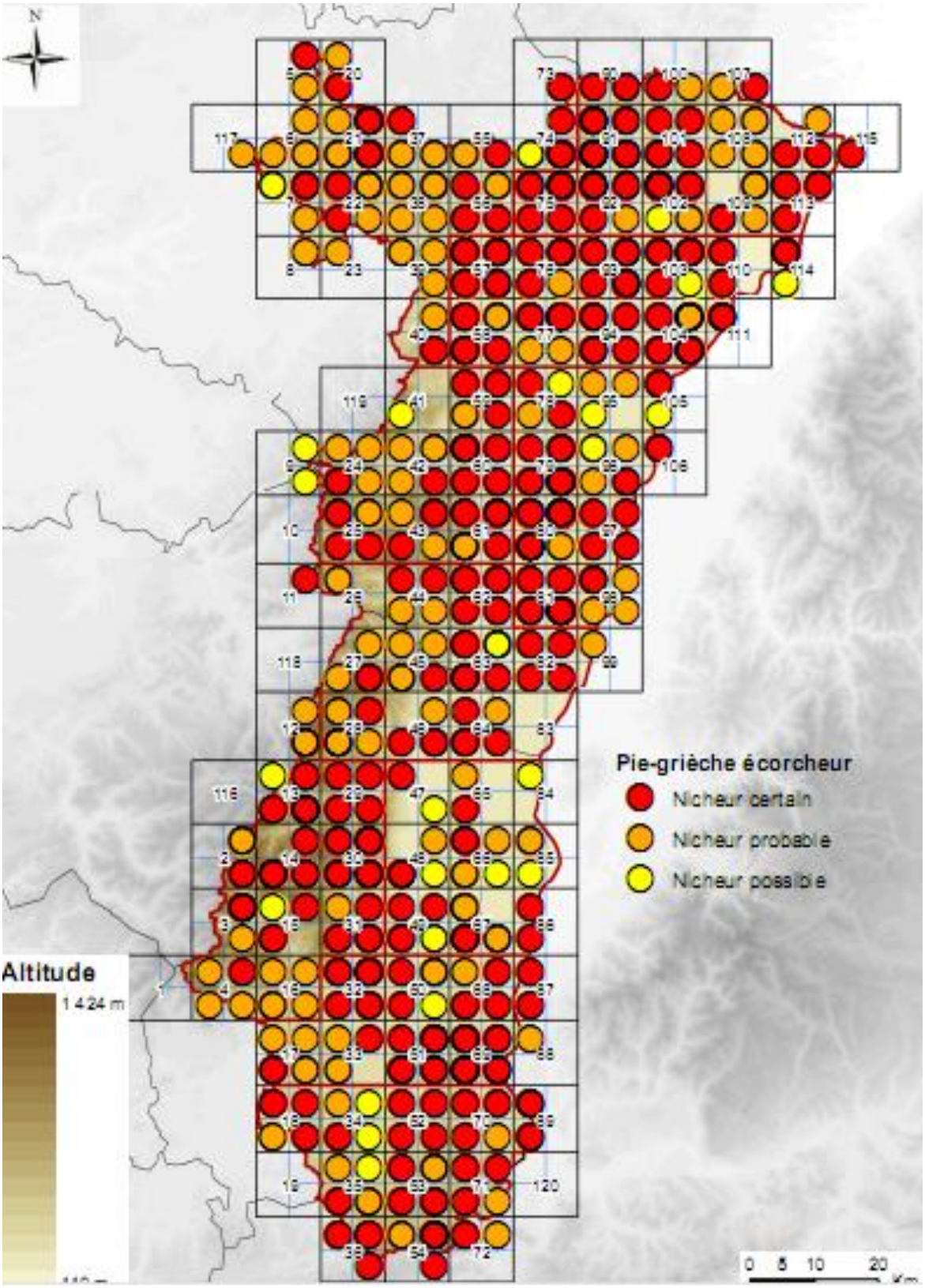
- Benoit Crabbé (2007) « *Régression Logistique* », Chapitre Modèles linéaires généralisés, Pages 1-32 sur le site : http://*www.linguist.univ-paris-diderot.fr/~b**crabbe**/LingExp/cours7.pdf*

- F.Husson, J. Josse, S. Le, J. Pages, « *Le package FactoMineR pour R* » sur le site : http://factominer.free.fr;

- A.B.Dufour & A. Viallefont, « *Un exemple de régression logistique* sous R », Pages 1-7.

# XI. Annexes

## Annexe 1 : Atlas de répartition de la Pie-grièche écorcheur en Alsace



## Annexe 2. Programmes R de la régression logistique

**#Importation des deux tables de données.**

```
pge1<-read.csv(file="C:\\Users\\Cheikh Moustapha\\Desktop\\Cheikh Stat\\PGE1.csv",header=T,
sep=";",dec=",")

Pge2<-read.csv(file="C:\\Users\\Cheikh Moustapha\\Desktop\\Cheikh Stat\\PGE2.csv",header=T,
sep=";",dec=",")
```

**#Poser les deux modèles de régression logistique sur les deux tables de données**

```
mod1 -> glm(formula = presence ~ nois + npass + nidif + foret + milouv +  zonart + zonhum
+ annee + (nois* npass) + (nois * nidif) + (nois * foret) + (nois * milouv) + (nois * zonart) +
(nois *  zonhum) + (nois * annee) + (npass * nidif) + (npass * foret) + (npass * milouv) +
(npass * zonart) + (npass * zonhum) + (npass * annee) + (nidif * foret) + (nidif * milouv) +
(nidif * zonart) + (nidif * zonhum) + (nidif * annee) + (foret *milouv) + (foret * zonart) + (foret
* zonhum) + (foret * annee) + (milouv *  zonart) + (milouv * zonhum) + (milouv * annee) +
(zonart * zonhum) + (zonart * annee) + (zonhum * annee), family = "binomial")

mod2 ->glm(formula = presence ~ nois + npass + nidif + culann + culper +  eau + zonart +
foret + milouv + prailan + verg + annee + (nois * npass) + (nois * nidif) + (nois * culann) +
(nois *  culper) + (nois * eau) + (nois * zonart) + (nois * foret) + (nois * milouv) + (nois *
prailan) + (nois * verg) + (nois * annee) + (npass * nidif) + (npass * culann) + (npass * culper)
+  (npass * eau) + (npass * zonart) + (npass * foret) + (npass * milouv) + (npass * prailan) +
(npass * verg) + (npass * annee) + (nidif * culann) + (nidif * culper) + (nidif * eau) + (nidif *
zonart) + (nidif * foret) + (nidif * milouv) + (nidif * prailan) +  (nidif * verg) + (nidif * annee) +
(culann * culper) + (culann * eau) + (culann * zonart) + (culann * foret) + (culann * milouv) +
(culann * prailan) + (culann * verg) + (culann * annee) + (culper * eau) + (culper * zonart) +
(culper * foret) + (culper *milouv) + (culper * prailan) + (culper * verg) + (culper *annee) +
(eau * zonart) + (eau * foret) + (eau * milouv) + (eau * prailan) + (eau * verg) + (eau *
annee) + (zonart * foret) + (zonart * milouv) + (zonart * prailan) + (zonart *verg) + (zonart *
annee) + (foret * milouv) + (foret * prailan) + (foret * verg) + (foret * annee) + (milouv *
prailan) + (milouv * verg) + (milouv * annee) + (prailan * verg) + (prailan * annee) + (verg *
annee), family = "binomial")
```

**#Sélection BACKWARD sur les deux modèles de régression logistique**

```
mod1c -> step ( mod1, dir="backward")
mod2c -> step (mod2, dir="backward")
```

**#Résultats de la sélection BACKWARD sur les deux modèles de régression logistique**
**Modèle 1**

Coefficients:

| (Intercept) | nois | npass | nidif | foret | milouv |
|---|---|---|---|---|---|
| -6.950e+00 | -1.162e-02 | 3.269e-02 | 1.380e+01 | 8.041e-04 | 3.031e-03 |
| zonart | zonhum | annee2012 | nois:npass | nois:nidif | nois:foret |
| -1.095e-02 | 7.522e-03 | 3.449e+00 | -2.379e-05 | 1.233e-02 | 8.313e-05 |
| nois:annee2012 | npass:nidif | npass:milouv | npass:annee2012 | nidif:foret | nidif:milouv |
| 5.082e-03 | -1.062e-01 | 2.611e-04 | 7.427e-02 | -1.777e-02 | -1.938e-02 |
| foret:milouv | foret:zonart | foret:zonhum | zonart:zonhum | | |
| 2.354e-05 | 3.796e-05 | -1.897e-04 | -3.448e-04 | | |

Degrees of Freedom: 4563 Total (i.e. Null); 4542 Residual
Null Deviance: 3192
Residual Deviance: 1733 AIC: 1777

**Modèle2**
Coefficients:

| (Intercept) | nois | npass | nidif | culann | culper |
|---|---|---|---|---|---|
| -6.963e+00 | -1.429e-02 | 3.624e-02 | 1.417e+01 | 1.621e-03 | 5.620e-03 |
| eau | zonart | foret | milouv | prailan | verg |
| 7.913e-03 | -8.631e-03 | 7.919e-04 | 4.167e-04 | 6.520e-04 | -5.836e-03 |
| annee2012 | nois:npass | nois:nidif | nois:culper | nois:zonart | nois:foret |
| 3.465e+00 | -2.788e-05 | 1.438e-02 | -1.388e-04 | 1.850e-05 | 8.014e-05 |
| nois:verg | nois:annee2012 | npass:nidif | npass:culann | npass:culper | npass:prailan |
| 7.476e-04 | 5.898e-03 | -1.133e-01 | 1.464e-04 | 1.146e-03 | 7.442e-04 |
| npass:verg | npass:annee2012 | nidif:culann | nidif:culper | nidif:foret | nidif:milouv |
| -2.101e-03 | 7.044e-02 | -1.648e-02 | -1.930e-02 | -1.755e-02 | -2.199e-02 |
| nidif:prailan | culann:zonart | culann:prailan | culper:eau | culper:foret | culper:milouv |
| -2.663e-02 | -4.839e-05 | 4.243e-05 | 1.540e-03 | 3.796e-05 | -6.435e-04 |
| culper:prailan | culper:verg | eau:foret | foret:prailan | prailan:verg | verg:annee2012 |
| 1.315e-04 | -3.823e-04 | -1.526e-04 | 5.041e-05 | -2.894e-04 | 6.792e-02 |

Degrees of Freedom: 4563 Total (i.e. Null); 4522 Residual
Null Deviance: 3192
Residual Deviance: 1607 AIC: 1691

**#Calcul des matrices de confusion sur les prédictions des deux modèles**

```
prediction<-ifelse(predict(mod1c, type="response") > 0.5, "Présence", "Absence")
print(table(PGE1$presence, prediction))

prediction<-ifelse(predict(mod2c, type="response") > 0.5, "Présence", "Absence")
print(table(PGE2$presence, prediction))
```

**#Codes pour tester la significativité des variables du modèle retenu par sélection BACKWARD**

**Modèle1**

```
mod1c -> step ( mod1, dir="backward")
summary (mod1c)
```

**Modèle2**

```
mod2c -> step (mod2, dir="backward")
summary (mod2c)
```

## Annexe 3: Codes R sur l'Analyse en Composantes Principales

**#Utilisation de la procédure PCA de FactoMineR sur les deux tables de données**

#centrage et réduction des données –> scale.unit = T

**#numéro des colonnes des variables quantitatives supplémentaires (variable presence)** –> quanti.sup

**#numéro des colonnes des variables qualitatives supplémentaires (variable annee)** –
> quali.sup

**#Décompression du package FactoMineR**
Library (FactoMineR)
acp1.pca<-PCA (PGE1, quanti.sup=8,quali.sup=9)
acp2.pca<-PCA(PGE2, quanti.sup=12,quali.sup=13)

**#obtenir les propriétés des objets acp1.pca et acp2.pca**
summary(acp1)
summary(acp2)

**#Eboulis des valeurs propres**
**#obtenir les variances associées aux axes c'est-à-dire les valeurs propres**
Val.propres1 <- acp1.pca$eig[,1]

| | eigenvalue | percentage of variance cumulative | percentage of variance |
|---|---|---|---|
| comp 1 | 2.1829007 | 19.844552 | 19.84455 |
| comp 2 | 1.7267338 | 15.697580 | 35.54213 |
| comp 3 | 1.3428345 | 12.207587 | 47.74972 |
| comp 4 | 1.2767290 | 11.606627 | 59.35635 |
| comp 5 | 1.1001816 | 10.001651 | 69.35800 |
| comp 6 | 0.8523557 | 7.748688 | 77.10668 |
| comp 7 | 0.7961733 | 7.237939 | 84.34462 |
| comp 8 | 0.6816504 | 6.196822 | 90.54145 |
| comp 9 | 0.5328267 | 4.843879 | 95.38532 |
| comp 10 | 0.2926027 | 2.660025 | 98.04535 |
| comp 11 | 0.2150115 | 1.954650 | 100.00000 |

Val.propres2 <- acp2.pca$eig[,1]

| | eigenvalue | percentage of variance cumulative | percentage of variance |
|---|---|---|---|
| comp 1 | 2.1829007 | 19.844552 | 19.84455 |
| comp 2 | 1.7267338 | 15.697580 | 35.54213 |
| comp 3 | 1.3428345 | 12.207587 | 47.74972 |
| comp 4 | 1.2767290 | 11.606627 | 59.35635 |
| comp 5 | 1.1001816 | 10.001651 | 69.35800 |
| comp 6 | 0.8523557 | 7.748688 | 77.10668 |
| comp 7 | 0.7961733 | 7.237939 | 84.34462 |
| comp 8 | 0.6816504 | 6.196822 | 90.54145 |
| comp 9 | 0.5328267 | 4.843879 | 95.38532 |
| comp 10 | 0.2926027 | 2.660025 | 98.04535 |
| comp 11 | 0.2150115 | 1.954650 | 100.00000 |

**#graphiques de l'éboulis des valeurs propres**
plot (1 :7, val.propres1, type= « b »,ylab= « Valeurs propres »,xlab= « Composante»)
plot (1 :11, val.propres2, type= « b », ylab= « Valeurs propres », xlab= « Composante»)

**#Corrélations des variables avec les composantes principales**
acp1.pca$var$cor[,1:2]
acp2.pca$var$cor[,1:2]

**#Contributions des variables aux composantes principales**
**#directement fournis par la procédure PCA**
**#ici pour les deux premières composantes**

**#Contributions relatives des variables aux composantes principales**
acp1.pca$var$cos2[,1:2]
acp2.pca$var$cos2[,1:2]

**#Contributions absolues des variables aux composantes principales**
acp1.pca$var$contrib[,1:2]
acp2.pca$var$contrib[,1:2]

**#Carte des individus sur les 2 axes factoriels.**
plot (acp1.pca, choix= « ind », title= «individus-1er plan»)
plot (acp2.pca, choix= « ind », title= «individus-1er plan»)

**#Contributions relatives des individus aux composantes principales et exportation du tableau sur Excel**
write.table(acp1.pca$ind$cos2[,1:2]
,"C:\\Users\\CheikhMoustapha\\Desktop\\CR1.csv",sep=";",dec=",")

write.table(acp1.pca$ind$cos2[,1:2]
,"C:\\Users\\CheikhMoustapha\\Desktop\\CR2.csv",sep=";",dec=",")

**#Contributions absolues des individus aux composantes principales et exportation du tableau sur Excel**
write.table(acp2.pca$ind$cos2[,1:2]
,"C:\\Users\\CheikhMoustapha\\Desktop\\CA1.csv",sep=";",dec=",")

write.table(acp2.pca$ind$cos2[,1:2]
,"C:\\Users\\CheikhMoustapha\\Desktop\\CA2.csv",sep=";",dec=",")

**#Représenter les individus selon les modalités de la variable qualitative illustrative**
**#coloriage différents des individus pour chaque modalité de la variable « annee » colonne 9**
plot(acp1.pca, choix= « ind», title= « Individus – 1e plan», habillage=9)

**#coloriage différents des individus pour chaque modalité de la variable « annee » colonne 13**
plot(acp1.pca, choix= « ind», title= « Individus – 1e plan», habillage=13)