L'acquisition des données : la télédétection[13] confrontée aux campagnes de terrain

La méthode de planification du travail de terrain en télédétection[13] consiste à identifier les écueils et problèmes et à sélectionner les solutions appropriées en avance (Joyce, 1978) L'échantillonnage a donc été réparti sur trois années chacune ayant un objectif spécifique et s'intégrant dans une méthodologie basée sur la meilleure connaissance possible de la zone d'étude.

En Camargue, le principal problème rencontré au cours des études sur le terrain est l'accessibilité limitée. Celle-ci dépend en premier lieu de la permission accordée par les propriétaires de domaines privés qui occupent 85 % de la superficie et en second lieu de la difficulté d'accès aux sites (chemin plus ou moins accessibles, optimisation du temps de marche par rapport au temps d'observation, profondeur de l'eau, vase...etc..). Afin de répondre au mieux à un objectif de planification de notre échantillonnage, nous avons entrepris en 2004 une saison de prospection dans le but de développer un premier contact avec les gestionnaires et propriétaires, élaborer, tester et améliorer notre protocole, et planifier la programmation de l'acquisition des images satellitales en fonction du développement de la végétation. Cela répond également aux objectifs méthodologiques de maximisation du nombre de sites d'échantillonnage et d'organisation du calendrier d'échantillonnage. Dans ce calendrier, nous avons intégré la contrainte d'accès relative au dérangement des populations d'oiseaux. Dans les marais de chasse, il est demandé de ne pas entrer dans le plan d'eau plus d'une ou deux fois et de stopper toute activité un ou deux mois avant le début de la période de chasse afin de ne pas nuire à l'installation des oiseaux d'eau. Pour les phragmitaies, il est préférable de ne pas entrer dans les roselières susceptibles d'accueillir une colonie de hérons pourprés avant la fin de la reproduction, soit fin juillet.

Les gestionnaires et propriétaires de domaines de chasse privée et d'espaces protégés ont été rencontrés lors de rendez-vous répartis sur l'ensemble de la période de développement des espèces végétales dominantes de Camargue. Cette consultation des gestionnaires par leurs avertissements, conseils et connaissances spécifiques du terrain a également contribué à un objectif d'organisation spatiale de l'échantillonnage. Il est en effet conseillé de placer un site de validation proche d'un chemin et en bordure de parcelle afin de limiter le temps d'accès, de

maximiser le nombre de sites et de permettre un repérage plus facile sur les images (Joyce, 1978). Seul une bonne connaissance du terrain permet d'appréhender ce type d'approche.

Nous devions également définir les communautés végétales à suivre et un protocole d'échantillonnage applicable à la fois sur le terrain et sur les images satellitales. Nous avons ensuite observé sur le terrain les regroupements des espèces dominantes de Camargue et leur phénologie[3] .Cette seconde étape a permis de définir en fonction du nombre d'habitats et de la superficie qu'ils représentent sur la zone d'étude, le type de communautés végétales que nous pouvions envisager de suivre. En effet, certaines espèces pouvant présenter un intérêt pour ce type d'étude, n'ont pas pu faire l'objet de suivi par manque de sites d'échantillonnage. La jussie (*Ludwigia spp.*), par exemple, espèce tropicale invasive sensible aux augmentations de salinité, se développe principalement dans des canaux de largeur inférieur à la taille des pixels de l'image et ne représentait qu'un seul site d'échantillonnage exploitable, détruit l'année suivante par traitement herbicide. Egalement le sénéçon en arbre (*Baccharis halimifolia*), autre espèce envahissante, n'a pu être suivie du fait des campagnes d'arrachages sur la zone au cours des années liées à ce travail de thèse. Ainsi les habitats retenus sont les phragmitaies, les herbiers aquatiques et les scirpaies. Le pic de développement des herbiers aquatiques selon les espèces s'étale de mai à fin juin (Grillas, 1992 ; Grillas et Roché, 1997 ; Mesléard, communication orale), les roseaux atteignent leur hauteur maximale fin juin début juillet (Poulin, Lefebvre, communication orale) et le scirpe est en général très développé la première quinzaine de juin avec un assèchement des feuilles à partir de la mi-juin (F. Mesléard et N. Yaverscovski, communication orale). Nous avions donc planifié d'échantillonner les premières espèces submergées en mai, de poursuivre avec les scirpaies en début juin, de continuer avec les espèces submergées plus tardives dans la deuxième quinzaine de juin et d'échantillonner les roselières de fin juin à début juillet. Cela nous a permis de combiner pic de croissance de chaque espèce, restriction temporelle d'accès due aux oiseaux nicheurs et à l'ouverture de la chasse, et maximisation du nombre de sites d'échantillonnage. L'annexe 3 retrace la phénologie[3] de ces communautés végétales influencée par la gestion des niveaux d'eau.

La seconde année d'échantillonnage avait pour objectif de servir de base aux tests de classification en s'appuyant sur les observations, essais et connaissances de l'année précédente. L'échantillonnage se devait d'être précis et organisé car il est le fondement de la méthodologie employée. Il devait prendre en compte les contraintes de temps, de

reproductibilité, d'applicabilité à la télédétection[13] en couvrant une grande zone d'étude. Il en va de même pour celui de la troisième année qui avait pour objectif la validation des indices et formules obtenues. Il est donc l'estimateur de la fiabilité du résultat final. A noter que l'utilisation d'un échantillon réalisé en 2006 pour estimer la qualité de la classification sur les deux années de suivi est rendu possible par le fait que les communautés végétales intéressant cette étude persistent d'une année sur l'autre à une même localisation. Windham (2001) a par exemple estimé que le phragmite commun peut persister, dans certaines régions, jusqu'à 4 années.

1 - Protocole d'échantillonnage de la végétation

a - Le plan d'échantillonnage

Cinq modèles de base doivent être considérés pour l'échantillonnage sur le terrain en télédétection[13] (McCoy, 2005): aléatoire, stratifié, systématique, systématique non aligné et groupé. Le choix du plan d'échantillonnage est effectué en fonction des objectifs de recherche, des contraintes spatiales du terrain et de temps imparti pour la prospection *in situ*. Girard et Girard (1999) conseillent comme solution la plus efficace de combiner des échantillons au hasard, systématiques et stratifiés.

Un plan d'échantillonnage de vérité terrain en télédétection[13] doit prendre en considération quatre éléments de bases dans la sélection des sites d'entraînement : la catégorisation des types d'occupation du sol, la taille et la configuration des sites d'échantillonnage, leur nombre et distribution et l'homogénéité et uniformité de la couverture végétale (Joyce, 1978). Le critère d'homogénéité est important dans la procédure de classification. En effet, l'ordinateur « s'entraîne » à partir des échantillons sélectionnés à reconnaître les mêmes types d'occupation du sol présents ailleurs.

Les principales contraintes d'échantillonnage rencontrées dans notre étude sont : l'accessibilité limitée, une grande surface de suivi qui introduit une contrainte de temps de travail et des communautés végétales sur d'importantes taches homogènes mais selon une configuration variable en terme de phénologie[3], le taux de recouvrement, le degré de monospécificité et les usages (scirpes pâturés, roseaux coupés). Notre but était de

cartographier au mieux toutes les configurations possibles qui nous étaient accessibles. Ces considérations ont ainsi défini la base de notre stratégie d'échantillonnage.

b - Les catégories d'occupation du sol et le nombre de sites

Dans le cadre d'une classification en présence/absence, on définit deux grands types de catégories d'occupation du sol que nous appelleront les « sites » et les « non-sites ». Dans le cadre de notre monitoring, les « sites » sont les différents types de communautés végétales dominantes que nous souhaitons identifier : phragmitaies, herbiers aquatiques et scirpaies. Les « non-sites » regroupent tous les autres types d'occupations du sol possibles. Les résultats de classification supervisée sont fortement liés à la qualité et la représentativité des données terrain utilisées (Friedl et al., 1999). Il est conseillé d'opter pour une approche d'échantillonnage par pixel unique pour les classes spectralement homogènes tandis que l'information spatiale et spectrale des classes hétérogènes peut être appréhendée plus facilement et plus efficacement par l'utilisation de pixels groupés (DongMei et Douglas, 2002). C'est pourquoi, dans le cadre de notre étude nous avons choisi d'apporter un maximum de précision à la caractérisation des communautés végétales nous intéressant (phragmitaies, herbiers aquatiques, scirpaies). Nous avons donc effectués deux types d'échantillonnages selon les catégories « site » et « non-site ».

Pour la catégorie « site », l'échantillonnage a été réalisé sur le terrain. Les difficultés rencontrées en Camargue nous ont contraints à intégrer une part d'échantillonnage intentionnel. Une première sélection des sites a été réalisée en fonction de l'accès et des autorisations obtenues lors de l'année de prospection. Du fait de l'irrigation en Camargue, chacune de ces surfaces se présente sous forme de parcelle ou bassin. Pour chaque bassin nous avons visualisé les taches les plus grandes, les plus homogènes et les plus accessibles à partir de photographies et survols aériens (avion et ULM). Selon la taille de ces bassins, nous avons sélectionné un ou deux emplacements. Les sites ont été placés à proximité d'un chemin à une distance minimale de 70 mètres des limites de la tache afin d'éviter l'influence de la réponse spectrale d'autres types d'occupations du sol à proximité (Figure 7).

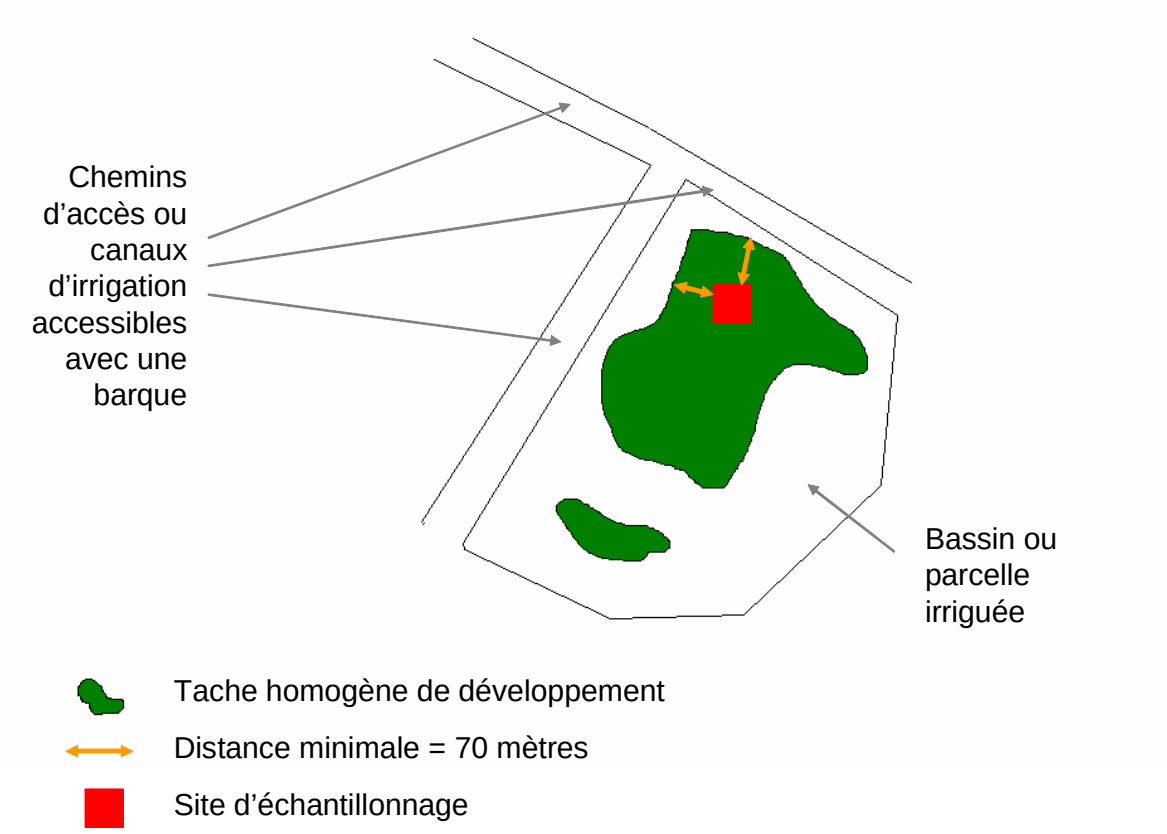

Figure 7: Localisation du site d'échantillonnage

Nous avons ainsi échantillonné 46 sites de phragmitaies, 25 sites d'herbiers aquatiques et 9 sites de scirpaies en 2005 et, 21 sites de phragmitaies, 91 sites d'herbiers, et 9 sites de scirpaies pour la validation en 2006.

Pour la catégorie « non-sites » nous avons fait le choix de combiner les numérisations réalisées en Camargue par le Parc Naturel Régional (PNRC) en 2001 et 2006 (PNRC) et par la Tour du Valat (non publié) avec une sélection de points sur photographies aériennes à partir de la BD ortho 1998 à notre disposition. Les données de la Tour du Valat concernent la délimitation des marais à marisques (cladiaies) et les zones envahies de *Baccharis*. Leur numérisation sous SIG a été réalisée à partir de photographies aériennes et de relevés sur le terrain (non publié).

Parmi les non-sites, les classes suivantes ont été prises en compte: les catégories « boisement », « pelouses », « milieu agricole », « sols nus ou sansouïres », « dunes ou plages », « urbain », « jonchaies » et « salins » de la cartographie du PNRC de 2001, ainsi que

la mer, les marais à marisques (cladiaies), la ripisylve, les zones envahies par le tamaris et le *Baccharis*, les pinèdes, l'eau claire, les typhaies et le riz. A partir de la cartographie du PNRC nous avons sélectionné aléatoirement une cinquantaine de polygones pour chacune des 8 catégories retenues (milieu agricole, salins, boisements, pelouses, sol nus ou sansouïres, jonchaies, étangs profonds, habitations). Pour les autres catégories (marais à marisques, *Bacccharis*, pinèdes, typhaies, ripisylve), nous avons sélectionné aléatoirement un nombre de points afin de couvrir les aires de développement connues. Nous avons ajouté des sites de mer et d'eau claire ainsi que des sites d'édifices industriels (bâtiments, cuves) que nous avons ajoutés à la catégorie urbain du PNRC pour créer une catégorie « habitations ». Quelques sites de rizières ont été visités en 2005 afin d'inclure ce type d'occupation du sol particulier en Camargue soupçonné d'être confondu avec les phragmitaies (G. Lefebvre, communication orale). Quelques sites de tamaris ont également été échantillonnés. La cartographie de 2001 utilisée pour une classification d'images de 2005 a limité le détail des catégories. En effet, il n'a pas été possible de différencier les types de milieux agricoles pour 2005 ni même de savoir si les zones de sols nus avaient ou non évoluées en sansouïres. C'est pourquoi nous avons conservé une désignation généraliste « milieu agricole » pour tous les types de cultures et une autre « sansouïres » pour les développements plus ou moins denses de salicorne.

En 2006, le PNRC a réalisé une mise à jour de la cartographie du « milieu agricole »*. Nous avons alors eu la possibilité de détailler les types de cultures. Ainsi cette catégorie se compose de culture maraîchère, riz, tournesol, blé, verger, colza, vigne, friche, maïs, terre, prés. Les aires de développement du *Baccharis* étant petites, peu nombreuses et en diminution constante du fait de l'arrachage massif contre son éradication, nous n'avons pas pu échantillonner de nouveaux sites en 2006. Il est à noter que les sites de tamaris, typhaies, cladiaies et ripisylves peuvent présenter un certain pourcentage de recouvrement en phragmites. En effet, ces espèces et/ou développements végétaux sont très souvent mélangés aux phragmites même s'ils occupent la plus grande part de la surface échantillonnée. Nous avons, par exemple, estimé le recouvrement des phragmites et des scirpes jusqu'à respectivement 15% et 10 % du recouvrement total des émergentes sur des sites de typhaies. Ces milieux étant peu développés en Camargue sur des tâches homogènes suffisantes pour notre protocole, nous n'avons pas eu la possibilité d'échantillonner de nouveaux sites en 2006. La catégorie « habitations » peut inclure des zones de végétation, les polygones n'excluant pas les jardins, par exemple. Le tableau 1 présente le nombre de pixels pour

chaque catégorie d'occupation du sol utilisés pour l'échantillon d'entraînement[4] (2005) et l'échantillon de validation[5] (2006).

Tableau 1 : Nombre de pixels de chaque catégorie d'occupation du sol dans l'échantillon d'entraînement[4] et de validation

Catégories	Nombre de pixels	
	échantillon d'entraînement[4]	échantillon de validation[5]
Mer	4272	6362
Tamaris	27	1264
Herbiers	30	99
Phragmitaies	57	25
Ripisylve	32291	8822
Cladiaies	25848	93
Jonchaies	9843	6236
Pelouses	7342	8631
Dunes et plages	12171	5370
Salins	278262	98047
Sansouïres	2598	42248
Habitations	6501	6669
Scirpaies	11	9
Boisements	2060	3017
Agricole	39300	27655
Baccharis	2378	
Pinèdes	25	
Typhaies	5	
Total	**423020**	***214547***

Les rubriques suivantes détailleront les caractéristiques des sites d'échantillonnage des communautés végétales que nous souhaitons identifier, c'est-à-dire la catégorie « sites » pour laquelle nous avons réalisé un relevé terrain détaillé.

c - La taille et la forme des sites d'échantillonnage

Les points importants à prendre en considération dans la taille du site d'échantillonnage sont la variabilité du terrain et la résolution de l'image (équivalent à la dimension du pixel). Afin de pouvoir associer un pixel sur le terrain à un pixel sur l'image, il est nécessaire de connaître les coordonnées géographiques de l'échantillon. Pour cela nous avons utilisé le système de positionnement global par satellite (GPS) nous permettant de situer nos sites avec une précision de 2 à 5 m. Nous avions planifié d'utiliser des images satellitales de type SPOT 5 ayant une résolution de 10 m. Compte tenu de la précision de notre GPS et du scan25 de la

Camargue (entre 2 et 9 m) (IGN, 2006) utilisé pour corriger géométriquement les images, nous avions estimé notre aire minimale aux alentours de 400 m², ce qui représente 4 pixels SPOT 5. La forme de notre site d'échantillonnage a ainsi était définie comme un carré de 20 m de côté localisé sur le terrain par une coordonnée GPS acquise en son centre à la croisée de ses diagonales. La taille et la forme de notre site d'échantillonnage permet, en tenant compte des précisions intervenant dans la localisation sur l'image, d'obtenir au minimum un pixel représentatif de la zone sur le terrain à l'intérieur du carré (Figure 8).

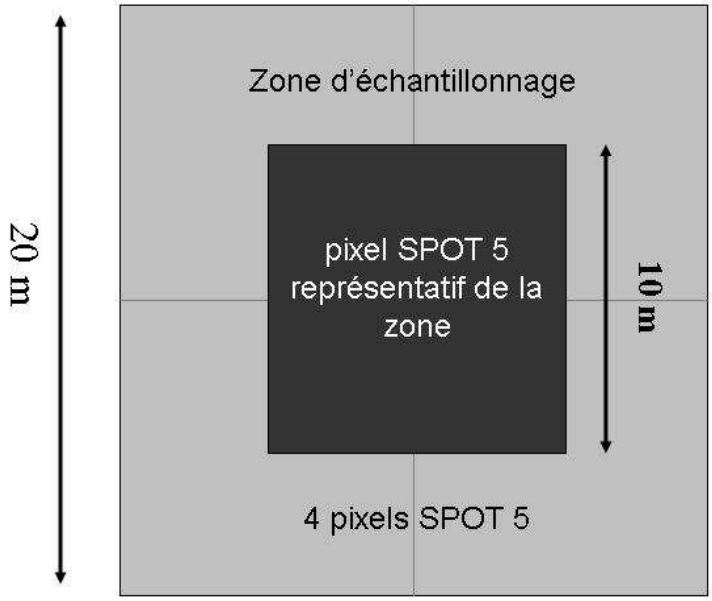

Figure 8 : Au moins un pixel SPOT 5, représentatif de la zone échantillonnée, est localisable sur l'image

d - Le nombre et la distribution des sites d'échantillonnage

Ces deux critères ont été définis selon l'accessibilité au terrain d'étude. Nous avons couvert l'ensemble des types de groupements homogènes pour chacune des communautés végétales à identifier sur l'ensemble des zones auxquelles nous avions accès chaque année. Nous avons porté un soin particulier à la distribution des sites afin de répartir l'échantillonnage sur l'ensemble de la zone d'étude (Figure 9). Les zones de scirpaies étant assez peu nombreuses, nous avons échantillonné la quasi-totalité de celles connues en Camargue. En effet, les neuf sites de scirpaies retenues correspondent aux seuls groupements homogènes à notre connaissance sur l'ensemble de la zone d'étude.

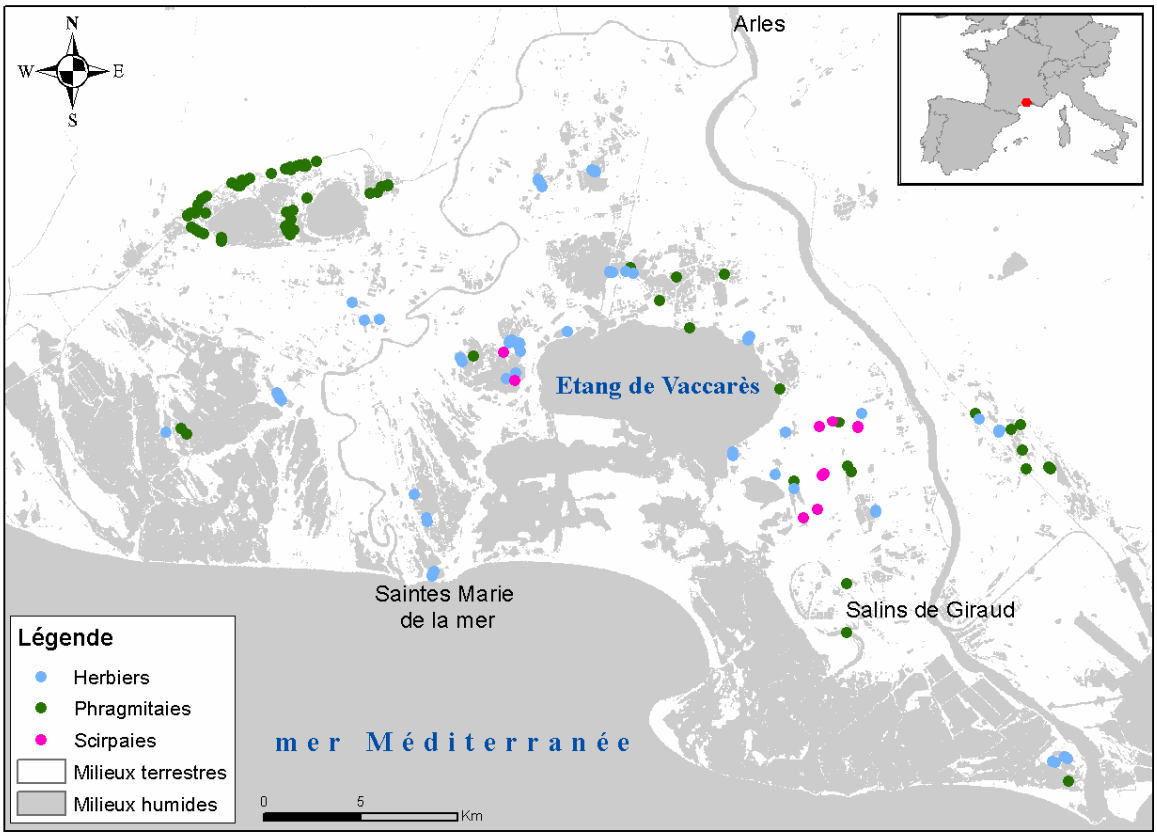

Figure 9: Localisation des sites d'échantillonnage

e - Les relevés botaniques

Mc Coy (2005) conseille, dans le but de préserver un caractère aléatoire au travail d'échantillonnage, de choisir une méthode permettant d'éliminer un possible biais grâce à la sélection d'un point de départ et une direction d'échantillonnage. Pour cela, il propose de sélectionner des points spécifiques le long de lignes ou transects.

Rappelons que notre site d'échantillonnage est un carré de 20 mètres de côtés. Le but est ici de couvrir une surface suffisamment grande de ce carré pour recenser les espèces dominantes formant la communauté végétale présente sur l'ensemble de la zone en optimisant le temps d'observation. Afin de couvrir au maximum la superficie de notre site, nous avons travaillé sur la base de transects représentés par les diagonales du carré. Selon la hauteur de la végétation, l'optimisation du relevé et afin de limiter les traces de notre passage sur le site, nous avons défini trois types de relevés selon les trois communautés végétales concernées.

- **Le relevé botanique des herbiers aquatiques**

Ce type d'habitat majoritairement en eau, peut contenir une épaisse couche de vase sous une végétation très dense. Le site d'échantillonnage devait être installé avec un minimum de piétinement pour limiter la perturbation du milieu. A partir d'un point de départ, nous avons délimité à l'aide de piquets et d'une boussole deux côtés et une diagonale nous permettant de repérer les quatre sommets du carré en limitant les passages sur une même zone (Figure 10).

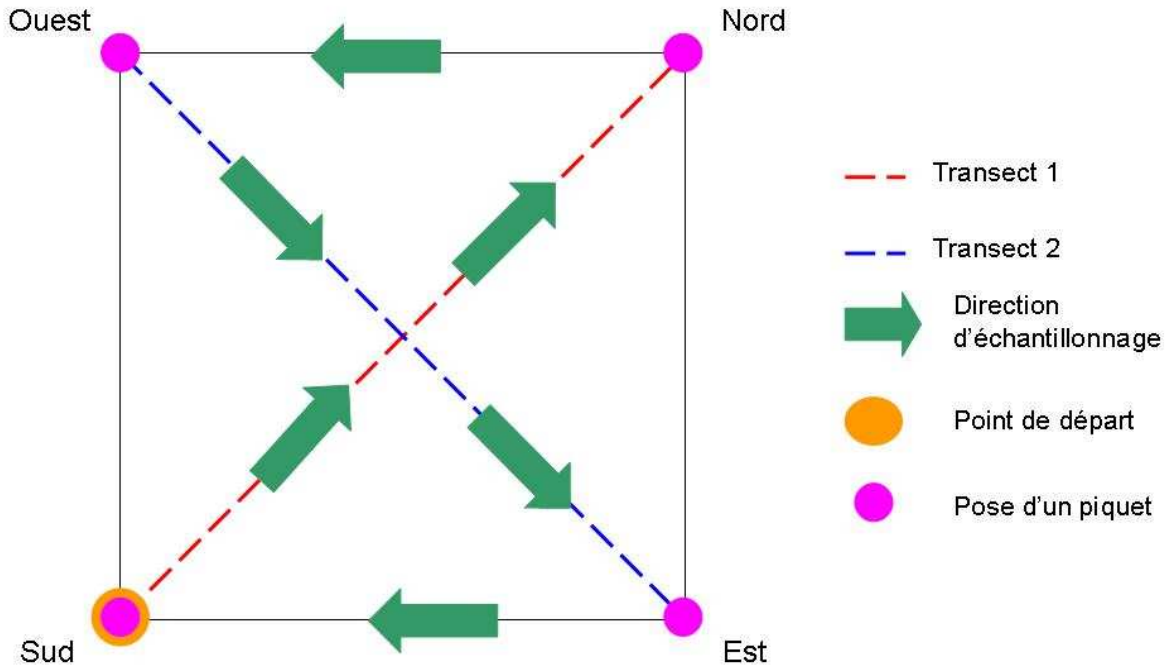

Figure 10 : Organisation des sites d'échantillonnage dans les herbiers aquatiques

Sur un mètre de largeur de part et d'autre de chaque diagonale, nous avons réalisé un inventaire et estimé visuellement le pourcentage de recouvrement des espèces végétales présentes en détaillant le pourcentage d'affleurement des herbiers. Nous avons également mesuré tous les quatre mètres sur chaque diagonale la tranche d'eau entre la surface et la hauteur maximale atteinte par la végétation (Figure 11), l'eau pouvant atténuer la réponse spectrale de la végétation aquatique selon sa profondeur.

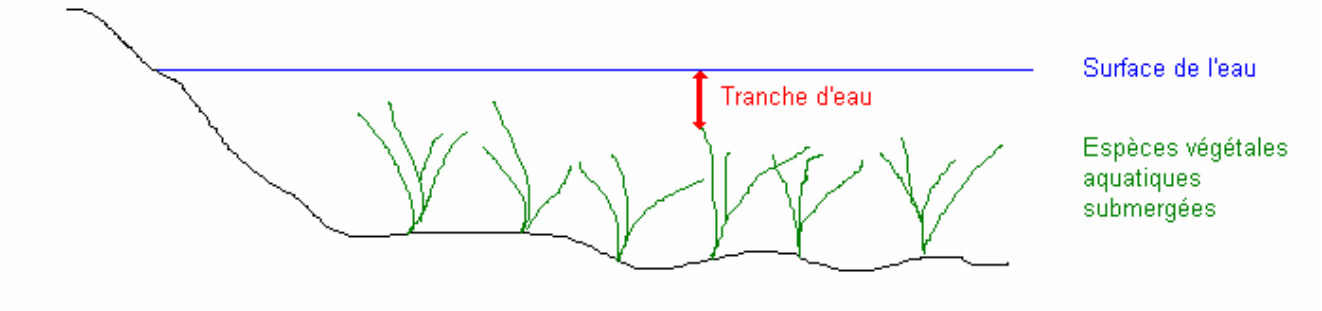

Figure 11 : Mesure de la tranche d'eau au dessus des herbiers

Nous avons estimé le pourcentage de recouvrement de la végétation émergente développée sur le site et pris des mesures de diamètre et de hauteur moyens des tiges lorsqu'elles représentaient plus de 10 % de la couverture végétale totale du site. Cela dans un but d'obtenir une indication sur la vigueur de cette végétation émergente.

Nous avons estimé le pourcentage de recouvrement de l'eau sans végétation et du sol pouvant apparaître selon la bathymétrie et/ou l'assèchement de la zone. Afin de caractériser au mieux le milieu, nous avons enfin estimé visuellement l'homogénéité du milieu codée de 1 à 4 du moins homogène au plus homogène.

- **Le relevé botanique des scirpaies**

Le point de départ de l'échantillonnage est identique à celui des herbiers. Les relevés se font également sur les diagonales. Les mesures prises tous les quatre mètres correspondent cette fois à une mesure de hauteur et de diamètre des plants de scirpes. Nous avons estimé le pourcentage de recouvrement du scirpe ainsi que des espèces végétales submergées et leur affleurement, ces dernières pouvant être très développées dans ce type de milieu. Nous avons également estimé le pourcentage d'eau sans végétation et de sol. Dans le cas d'un développement de plus de 10 % d'une autre espèce émergente sur le site, nous avons relevé leur diamètre et leur hauteur moyens. Enfin nous avons estimé l'homogénéité sur le site selon la même échelle semi-quantitative que pour les herbiers. Ainsi les sites échantillonnés sélectionnés pour leur homogénéité de développement du scirpe peuvent contenir jusqu'à 80 % d'herbiers.

- **Le relevé botanique des phragmitaies**

Pour les phragmitaies, nous avons travaillé par demi-diagonale en gardant comme point de départ le centre du carré. Cela a permis de limiter la perturbation du milieu et de nous orienter plus facilement dans une végétation pouvant parfois dépasser les trois mètres de hauteur. A l'aide d'une boussole, une personne positionnée au point central, indique la direction à une autre s'éloignant avec un piquet de trois mètres à la main. Nous avons ainsi virtuellement dessiné les quatre demi-diagonales servant de transects (Figure 12)

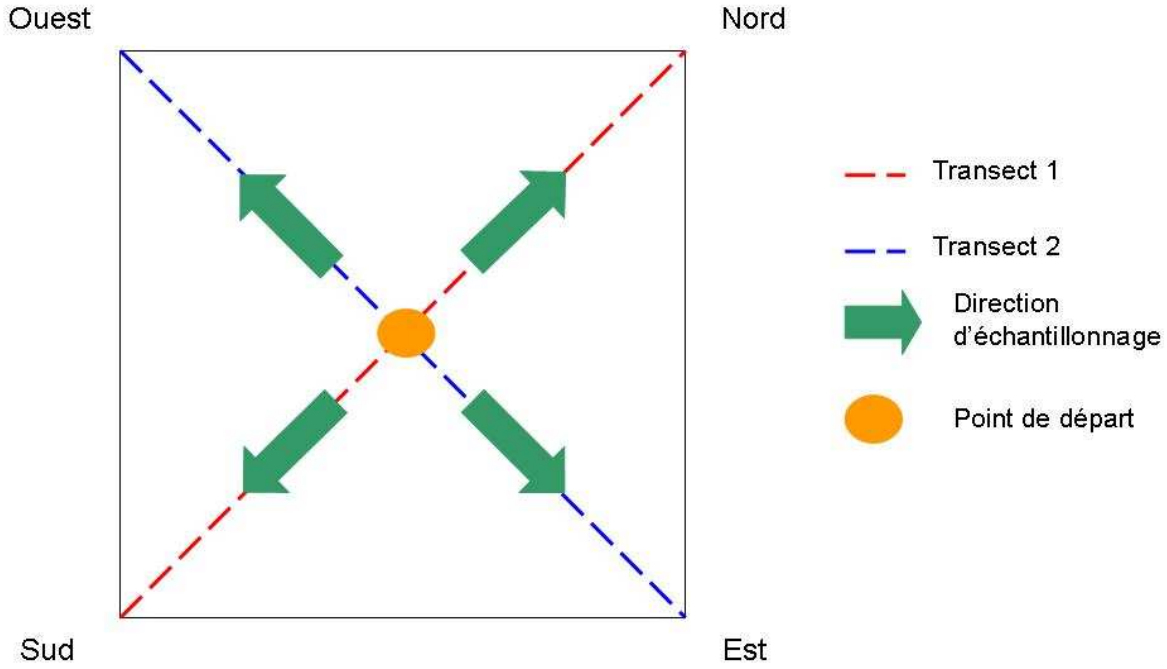

Figure 12 : Organisation des transects dans les phragmitaies

L'objectif du relevé botanique dans les phragmitaies était d'obtenir des mesures de leur structure et biomasse reflétant leur bonne santé ou vigueur et leur intérêt avifaunistique. Ces mesures s'inspirent du protocole d'échantillonnage des roselières mis en œuvre par la Tour du Valat et utilisé par les gestionnaires de réserves naturelles dans le cadre du Rézo du Rozo (Mauchamp et al., 2004).

Sur chacune des quatre demi-diagonales du carré, un inventaire floristique et une estimation visuelle du pourcentage de recouvrement du roseau et des autres espèces présentes ont été effectués. Afin d'obtenir également une estimation numérique de ce pourcentage de recouvrement, des survols en avion et en ULM ont été mis en œuvre, aux cours desquels des

photographies ainsi que des vidéos ont été acquises à la verticale. Aucune donnée numérique n'a pu être extraite de ces acquisitions du fait de l'altitude et de la vitesse minimales nécessaires aux différents engins aériens testés, qui limitent la qualité photographique. Une photographie du ciel à travers la végétation prise à l'aide d'un appareil photographique posé au sol à l'horizontal au milieu de chaque quadrat présenté ci-après, est finalement apparue comme étant la méthode efficace et reproductible la plus simple et la moins coûteuse pour l'acquisition de cette donnée numérique. Les spécificités de la prise de vue étaient une résolution de 4 megapixels, une distance focale de 7,8 mm - 23,4 mm et aucun zoom. Ces photographies ont permis d'obtenir, à l'aide du logiciel CANEYE (Baret et Weiss, 2004), une estimation du recouvrement moyen des espèces émergentes présentes sur l'ensemble du site d'échantillonnage (soit le carré de 20 m de côtés). CANEYE est un logiciel conçu pour caractériser la canopée à partir du traitement de plusieurs photographies. L'estimation de ces caractéristiques est basée sur la transmittance de la lumière à travers la végétation en considérant que celle-ci est opaque (Baret et Weiss, 2004).

A la moitié de chacune de ces demi-diagonales, dans une zone homogène représentative du développement du roseau sur le site, un quadrat de 50 cm de côté a été disposé au sol afin d'effectuer les mesures nécessaires à la détermination de la structure du roseau.
Dans ce quadrat ont été mesurés :

- le nombre de tiges vertes (= tiges de l'année)
- le nombre de tiges sèches (= tiges des années antérieures) cassées et coupées
- le nombre de tiges sèches entières (dont la panicule est toujours présente)
- la hauteur de deux tiges vertes choisies comme représentatives du site
- la hauteur de deux tiges sèches entières représentatives du site
- la hauteur d'une tige coupée
- le diamètre basal de deux tiges vertes
- le diamètre basal de quatre tiges sèches.

L'homogénéité a été estimée puis codée comme pour les herbiers et les scirpaies. Nous avons également noté le type de surface du sol avec plus ou moins de litière sèche ou humide pouvant contribuer à la part du roseau sec dans la réponse spectrale du site. Enfin, un GPS positionné au croisement des diagonales du carré a permis d'obtenir une localisation précise de chacun de nos sites d'échantillonnage.

2 - Protocole des relevés de niveaux d'eau

Les relevés de niveaux d'eau se composent de deux types de mesures. Une première consiste en un relevé périodique du niveau de chaque parcelle hydraulique à laquelle nous avions accès, la seconde en une moyenne de niveaux pris une seule fois sur le site d'échantillonnage, généralement lors du relevé botanique.

Le suivi du niveau des parcelles hydrauliques est tiré d'une base de données (non publiée) alimentée par les gestionnaires d'espaces protégés (Marais du Vigueirat, Domaine de la Palissade, Tour du Valat) produite à partir d'un réseau d'échelles et de piézomètres relevés mensuellement ou bi-mensuellement. Les suivis se font généralement à partir d'échelles disposées au point le plus profond des marais ouverts (visibilité non gênée par une végétation émergente haute) qui peuvent être observées de loin à l'aide de jumelles. Dans les marais à hautes émergentes qui offrent peu de visibilité comme les phragmitaies, les suivis se font plutôt à partir de piézomètres, soit des tubes de PVC de 200 cm de long enfoncé verticalement jusqu'à 50 cm sous la surface du sol dans un endroit facilement accessible en bordure du marais. Cette méthode, qui permet de suivre les niveaux d'eau et de salinité dans la couche du sol comprenant les rhizomes de roseaux, a également été utilisée dans certains marais de chasse dépourvu d'échelles.

Les relevés de niveaux d'eau sur le site d'échantillonnage étaient pris à l'aide d'un réglet tous les quatre mètres sur chacune des diagonales du carré. Dans les phragmitaies, nous avons noté le substrat où était prise la mesure en distinguant les trouées[14] des zones de roseau. Cela permet à la fois d'évaluer l'homogénéité de la couverture végétale et de mieux interpréter la bathymétrie du site car les amas de rhizomes ont pour effet de « surélever » le niveau du sol. Lorsque quelques points sur un site n'étaient pas inondés en surface, un petit puit était creusé jusqu'à apparition de l'eau sous-jacente et une mesure négative était notée. Une mesure au piézomètre ou à l'échelle selon le marais, a été noté le même jour afin de faire correspondre les deux relevés et ainsi disposer d'une mesure d'étalonnage. Cet étalonnage permet d'extrapoler le niveau d'eau moyen du site pour chaque date de relevé du bassin et de choisir ensuite les valeurs les plus appropriées en fonction de la date d'acquisition des images satellitales.

3 - Données satellitales

a - Les images satellitales

Les images utilisées dans ce travail proviennent du satellite SPOT 5. Le 4 mai 2002 à 1h31 min TU le satellite SPOT 5 a rejoint grâce à Ariane 4 la constellation du système SPOT à une altitude de 832 km. Il apporte, outre un nouvel instrument pour l'acquisition d'images stéréoscopiques, une meilleure résolution géométrique. L'orbite de SPOT est héliosynchrone, circulaire et quasi-polaire (avec une inclinaison à 98°). L'inclinaison de l'orbite a une influence sur l'orientation de l'image. Par exemple à une latitude d'environ 45° comme est située la Camargue, l'image est orientée selon un angle d'environ 15 °. Grâce à la rotation de la Terre, le passage du satellite du pôle nord vers le pôle sud décrit sur la Terre des traces à intervalles réguliers (Figure 13) tout en conservant des conditions d'éclairement solaire similaires permettant la comparaison des observations d'une même zone.

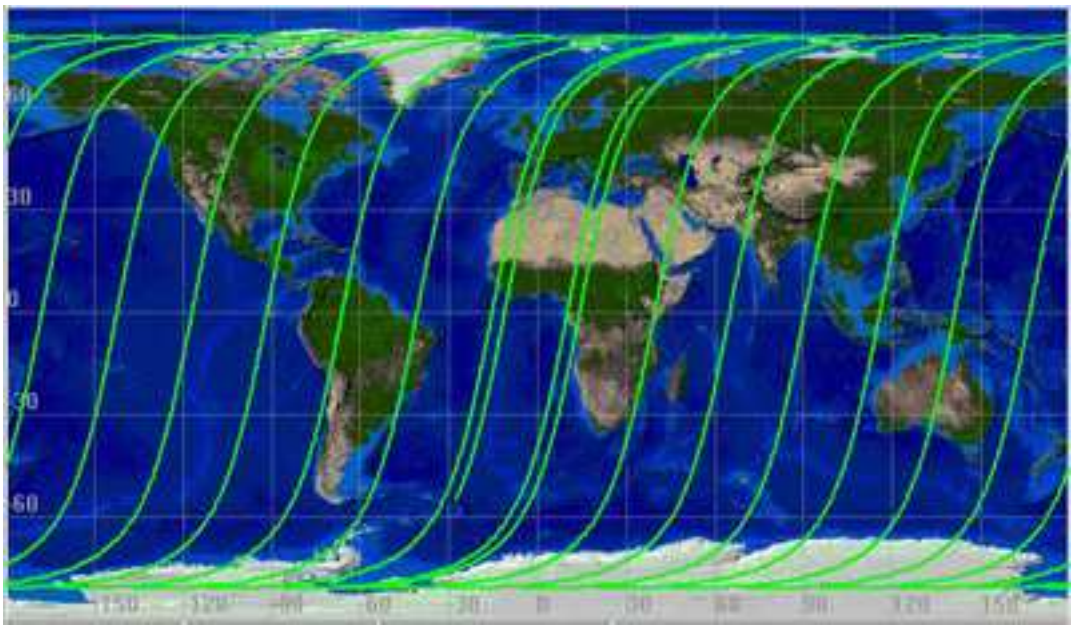

Figure 13 : Trace au sol de l'orbite SPOT (illustration CNES, http://spot5.cnes.fr/index3.htm)

La distance maximale (à l'équateur) entre deux traces est de 108 km. Il faut donc 26 jours à SPOT 5 pour couvrir l'ensemble de la Terre pendant lesquels il effectuera 369 orbites. On dit que son orbite est phasée. L'heure de passage au-dessus d'une région donnée est constante à plus ou moins 15 min. Les possibilités de visée oblique de Spot permettent l'acquisition de scènes à l'intérieur d'une bande de 900 kilomètres. Cette technique permet d'augmenter la

fréquence d'observation d'un même point au cours d'un même cycle. Cette fréquence varie en fonction de la latitude : à l'équateur, la même région peut être observée 7 fois pendant les 26 jours du cycle orbital. Située à une latitude d'environ 45°, la Camargue peut être observée 11 fois pendant un cycle orbital., soit 157 fois par an, ce qui correspond à une moyenne de 2,4 jours avec un intervalle se situant au maximum à 4 jours et au minimum à 1 jour. Les instruments HRG (haute résolution géométrique) de SPOT 5 ont un champ de vue de 4° soit une fauchée au sol de 60 kilomètres (Figure 14). Ces instruments disposent également d'une capacité de visée verticale latérale de 27° de part et d'autre de la verticale. Ceci permet d'observer des régions qui ne sont pas forcément à la verticale du satellite. Ils peuvent également travailler ensemble et couvrir des zones de 120 kilomètres de largeur.

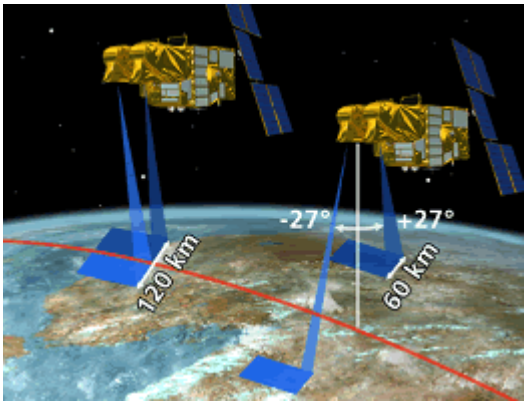

Figure 14 : Champ de vue des instruments HRG (illustration : CNES, http://spot5.cnes.fr/index3.htm)

SPOT 5 est capable de traiter 5 images au maximum dont 3 sont enregistrées et 2 sont transmises vers le sol. La transmission des données au sol peut être immédiate si le satellite est en visibilité d'une station de réception ou être différée après stockage à bord.

SPOT 5 comporte 5 bandes spectrales : B0 (panchromatique : 0.51 à 0.73 µm), B1 (vert : 0.50 à 0.59 µm), B2 (rouge : 0.61 à 0.68 µm), B3 (proche infrarouge ou PIR : 0.79 à 0.89 µm) et B4 ou MIR (moyen infrarouge ou MIR : 1.58 à 1.75 µm) (Spot image, 2005). Les scènes SPOT 5 10 m couleurs sont le produit d'une observation en mode multispectral effectuée simultanément dans quatre bandes spectrales. Les bandes B1, B2, B3 sont acquises avec une résolution de 10 m au sol et la bande B4 ou MIR est acquise à 20 m avec un ré-échantillonnage à 10 m qui permet d'obtenir une seule image (Spot Image, 2005). Le mode multispectral de SPOT 5 avec ses quatre bandes spectrales couvrant le visible et l'infrarouge

moyen est un point essentiel pour le monitoring des zones humides. En effet, des études ont montré que des canaux dans le moyen infrarouge sont utiles pour discriminer la végétation et l'humidité des sols et apporter une meilleure discrimination des différents types de zones humides (FGDC, 1992 ; Jensen et al., 1993a).

Le passage régulier sur une région et le jeu de miroir de SPOT permet d'optimiser le nombre d'acquisition d'images en fonction des conditions nuageuses parfois défavorables et de l'évolution de la végétation. Nous avions donc la possibilité de faire correspondre plus ou moins nos dates de relevés sur le terrain avec les dates d'acquisition des images. Ainsi nous avons planifié l'acquisition des images entre le 15 et le 30 des mois de développement de la végétation (Annexe 3) soit mai, juin, juillet et septembre, ainsi qu'une en décembre afin de comparer la réponse spectrale des sites entre la saison hivernale et la saison estivale et enfin une en mars, mois où les roselières ont été coupées pour l'exploitation de la sagne et où les tiges vertes ne sont pas encore sorties (Tableau 2). En 2006, l'image de septembre n'a pas été acquise mais une image en août et en octobre ont été achetées (Tableau 3). Ces images sont en mode multispectral (4 canaux : B1, B2, B3, MIR) avec un niveau de correction 1A. Le niveau 1A est un niveau de prétraitement où seules les corrections radiométriques ont été effectuées. La radiométrie est corrigée par un modèle linéaire qui égalise la sensibilité des détecteurs. Ce niveau de prétraitement est majoritairement utilisé pour les applications cartographiques, la stéréo-restitution ou pour les études radiométriques particulières. Des données complémentaires sont également fournies. Sans vouloir en dresser la liste complète, nous citerons ici celles qui peuvent être particulièrement utiles pour retravailler la radiométrie des images : le satellite (dans notre cas SPOT 5), l'instrument (HRG1 ou HRG2), les bandes spectrales, la date et l'heure de prise de vue (en temps universel), l'angle d'incidence et d'orientation (en degré), l'azimut et l'élévation du soleil, les angles de visée de début et de fin de ligne, les gains appliqués à chaque bande spectrale lors de la prise de vue, les histogrammes des valeurs radiométriques et les profils spectraux des bandes spectrales de l'image. Les produits de niveau 1A sont distribués au format des produits numériques scènes SPOT.

Tableau 2 : Caractéristiques des images acquises pour l'année 2005

Jour		30-déc-04	17-mars-05	19-mai-05	18-juin-05	31-juil-05	21-sept-05
Heure T.U		10,53	10,83	10,63	11,02	10,55	10,53
Latitude		43,52	43,52	43,52	43,05	43,52	43,52
Longitude		4,54	4,53	4,54	4,52	4,53	4,53
K-J identification		48-262	48-262	48-262	48-262	48-262	48-262
Instrument		HRG2	HRG1	HRG2	HRG1	HRG1	HRG1
Angle d'orientation		12,697861	15,654342	13,905931	17,527673	13,263458	13,284182
Angle d'incidence (degrés)		R20,277996	L8,641644	R8,798583	L25,117560	R14,828587	R14,661247
Angles solaires (degrés)	Azimut	162,340773	159,373668	147,721131	153,742727	141,843026	157,570559
	Elévation	21,098546	42,714105	63,022785	68,172755	60,633018	45,243817
Nébulosité (en %)		0	0	0	0	3	1
Vent		Sud	Sud ouest	Nord	Sud	Sud	Nord

Tableau 3 : Caractéristiques des images acquises pour l'année 2006

Jour		18-déc-05	16-mars-06	29-mai-06	23-juin-06	24-juil-06	30-août-06	15-oct-06
Heure T.U		10,65	10,79	10,41	10,73	10,80	10,60	10,86
Latitude		43,569722	43,568611	43,515833	43,568333	43,621389	43,516944	43,,516667
Longitude		4,552222	4,560556	4,6275	4,545833	4,335	4,523333	4,525278
K-J identification		48-262	48-262	49-262	48-262	48-262	48-262	48-262
Instrument		HRG2	HRG1	HRG2	HRG1	HRG2	HRG2	HRG2
Angle d'orientation (degrés)		14,475491	15,697234	12,151096	15,092436	15,567327	13,879777	16,32702
Angle d'incidence (degrés)		R3,518600	L8,628000	R25,283230	L2,589556	L7,037974	R9,289699	L14,414627
Angles solaires (degrés)	Azimut	165,477188	158,44473	139,166636	143,826014	146,560092	152,036122	168,788796
	Elévation	21,77996	42,06779	63,149179	66,476597	63,367976	52,964985	37,483181
Nébulosité (en %)		0	0	7	0	2	2	1
Vent		Nord	Nord	Sud ouest	Sud	Sud/Sud est	Nord	Nord

b - Le Global Positioning System (GPS)

Le GPS ou NAVSTAR (Navigation System by Timing And Ranging) est un système de radionavigation américain basé sur une constellation de satellites élaborés en 1970 et contrôlés par le département de la défense (DoD) des Etats-Unis. Utilisable librement par toute personne munie d'un récepteur GPS, il informe sur la position, le temps et la vitesse sur une large étendue mondiale quelques soient les conditions météorologiques ou le moment de la journée (jour et nuit). Le GPS est ainsi composé de trois parties : les satellites en orbite autour de la Terre, des stations de contrôle et de suivi sur la Terre et l'appareil de réception d'un nombre illimité d'utilisateurs. Le récepteur de chaque utilisateur capte les signaux

diffusés par les satellites et produit ainsi une localisation selon trois dimensions (latitude, longitude, altitude). Il calcule sa position par triangulation, en mesurant la distance entre lui-même et au minimum trois satellites. La composante spatiale du système est basée sur une constellation de 24 satellites placés en orbite quasi-polaire à 20 200 km d'altitude. Le GPS fournit une précision de l'ordre de 5 à 15 mètres. Le signal peut être perturbé par différents paramètres : la traversée de l'atmosphère, une densité forte des feuilles des arbres ou par un bâtiment en milieu urbain qui le cache ou répercute un écho, fournissant plusieurs signaux se traduisant par une localisation faussée. Afin de limiter l'influence de tels facteurs, nous avons immobilisé le GPS au centre du carré attaché à un piquet de 3 mètres de hauteur permettant de limiter la perturbation due au feuillage des phragmitaies par exemple. Le GPS restait ainsi immobilisé le temps du relevé botanique sur le site permettant d'obtenir la meilleure position possible de notre site d'échantillonnage. Nous avions alors la possibilité de positionner chaque carré par son point central sur les images satellitales géoréférencées.

B - Les prétraitements d'une imagerie multispectrale et multitemporelle

La précision et l'interprétation quantitative de données de télédétection[13] nécessitent que les images numériques soient corrigées radiométriquement et géométriquement au préalable à toute analyse. Ces prétraitements sont des éléments de base, dont dépendent la précision et la qualité du résultat de l'analyse d'images satellitales et sont particulièrement importants pour les étude en multi-date (Estes et al., 1983 ; Rosenfeld, 1984 ; Guindon et al 1981 cités par Teillet, 1986).

1 - La correction des effets de l'atmosphère

Les images satellitales sont communément fournies à l'utilisateur sous forme de comptes numériques qui doivent être convertis en valeurs physiques relatives à la réponse de la surface de la Terre pour pouvoir mettre en relation les données de télédétection[13] et les caractéristiques des catégories d'occupation du sol. Cette étape est indispensable dans le cadre d'une étude en mode multi-temporel. En effet, les différences d'illumination, de propagation de l'atmosphère et de réponse du capteur d'une imagerie multi-date doivent être impérativement corrigées pour des objectifs d'utilisation des valeurs physiques de surfaces, de

comparaison ou combinaison de ces valeurs et de mise en place de méthodes reproductibles s'affranchissant d'un nouvel échantillonnage (Kergomard, 2000). Ces corrections peuvent être effectuées avec des modèles détaillés des conditions atmosphériques ou de simples calculs basés seulement sur les données de l'image. Il est parfois difficile de choisir la méthode la plus appropriée entre précision et simplicité sachant que, de toute façon, ces méthodes demeurent approximatives. Dans le cadre de cette thèse, nous avons testé deux approches dont les résultats sont explicités dans un article (Davranche et al, sous presse), dont voici un résumé:

L' imagerie satellitale multitemporelle est un outil potentiellement intéressant pour le suivi des zones humides méditerranéennes, milieux naturels parmi les plus menacés dans le monde. Six scènes SPOT-5 ont été acquises en 2005, grâce à une subvention du Centre National d'Études Spatiales, afin d'évaluer l'évolution de la superficie et de l'état de la végétation émergente et submergée des marais de Camargue. Les scènes SPOT-5 sont constituées de quatre canaux (couvrant des plages de longueurs d'onde du vert, rouge, proche et moyen infrarouge) et offrent une résolution de 10 m. Afin de rendre comparable des images acquises à différentes dates, une calibration en réflectance de surface ou à défaut une normalisation des comptes numériques est nécessaire pour limiter les effets des facteurs environnementaux, atmosphériques et liés au capteur. Dans cet article, nous comparons l'efficacité de deux méthodes de calibration radiométrique : le modèle 6S (Second Simulation of the Satellite Signal in the Solar Spectrum) élaboré dans le but d'une correction absolue des effets atmosphériques et environnementaux et l'utilisation de points pseudo-invariants (PPI) comme méthode relative. 6S est utilisé selon une méthode simplifiée permettant de s'affranchir de la mesure de l'épaisseur optique de l'atmosphère en évaluant la donnée de visibilité par itérations du modèle sous l'hypothèse d'une valeur de réflectance nulle de l'eau de mer profonde dans le canal moyen infrarouge. Quatre types de PPI ont été utilisés: eau profonde, pins, toitures et sable. L'eau profonde et les pins ont été les moins variants, tandis que les toits ont montré de grandes différences intra-image ; et le sable, les plus fortes variations inter-images tout en ayant des valeurs très similaires à une même date. Globalement, les deux approches ont présenté des résultats similaires en terme de variation des données de radiométrie telle qu'estimée par la distance euclidienne entre les valeurs moyennes obtenues pour chaque canal (6S =4.3%; PPI = 4.0%). L'exclusion des PPI les plus variants de l'analyse (eau profonde en mars et juin, pins en décembre, petits toits, sable en juillet et septembre) a permis de diminuer les variations des données radiométriques (PIF = 3.4%; 6S model = 2.9%), le modèle 6S offrant alors une valeur significativement plus basse que la

méthode des PPI. Ainsi, sous sa forme simplifiée d'utilisation, 6S reste une bonne méthode de calibration radiométrique. L'utilisation des points pseudo-invariants demeure une approche valable à la condition que les points soient sélectionnés de façon à couvrir un large éventail de luminance et que chaque type de points soit représenté par un minimum de cinq éléments dont la variation des valeurs radiométriques a préalablement été testée.

2 - Prétraitements géométriques

Les corrections géométriques permettent de corriger les erreurs générées au moment de l'acquisition de l'image et celles dues à la rotondité et au relief de la Terre (Girard et Girard, 1999). Ainsi corrigées radiométriquement, nos images ont ensuite été rectifiées géométriquement à l'aide d'une méthode polynomiale standard. Une trentaine de points d'appui répartis sur l'ensemble de la scène ont été utilisés. Le ré-échantillonnage a été effectué avec la méthode du plus proche voisin. Chaque image a été géoréférencée en Lambert II étendu conforme utilisant comme base le SCAN 25 de l'IGN. Afin de s'assurer de la qualité du géoréférencement, la réflectance et la position de points invariants[10] ont été comparées entre toutes les dates.

C - Analyses statistiques

1 - Classification en présence/absence

a - Revue des principales méthodes de classification utilisées en télédétection[13]

- **Méthodes traditionnelles non supervisées et supervisées**

Les méthodes de classement ont pour objectif commun la découverte d'un estimateur assurant l'affectation d'une classe parmi *c* classes disponibles à un individu inconnu sur la base de la connaissance d'un ensemble de *m* caractères le décrivant (appelés attributs descripteurs) (Brostaux, 2005). En télédétection[13], la classification est une méthode par laquelle des identifiants de classes sont attachés à des pixels produisant une image sur la base de leurs caractéristiques. Ces caractéristiques sont généralement des mesures de leur réponse spectrale

dans différentes plages de longueur d'onde. Les méthodes traditionnelles utilisées pour la classification de l'occupation du sol à partir de données de télédétection[13] sont généralement les procédures non supervisées telles que ISODATA et les méthodes supervisées dont la plus populaire est la classification par maximum de vraisemblance (Maximum Likelihood Classifiers, MLC). Cette dernière est basée sur une procédure de classification probabiliste qui suppose que chaque classe spectrale peut être décrite et modélisée selon une loi de distribution normale. La performance de ce type de classification dépend ainsi de la façon dont les données s'accordent au modèle prédéfini. Le modèle gaussien est caractérisé par le vecteur de moyenne et la matrice de covariance des classes. Si un échantillonnage a un nombre de sites fixe, la précision des estimateurs des éléments du vecteur de moyenne et de la matrice de covariance de l'échantillon diminue avec l'augmentation du nombre de zones d'occupation du sol. Ainsi on peut s'attendre à une dégradation des performances de la classification avec l'augmentation du nombre de types d'occupations du sol. L'hypothèse que les données de chaque classe suivent une distribution normale restreint les analyses à une certaine proportion de données. En d'autres termes, si les données sont complexes, leur modélisation devient difficile avec ce type de classification. Ce type de classification est également dépendant d'un échantillonnage systématique et n'offre pas la possibilité de construire un modèle permettant d'appliquer une formule reproductible sans avoir à ré-échantillonner chaque année.

Afin de palier à l'un et/ou l'autre de ces deux problèmes, d'autres méthodes de classification sont de plus en plus utilisées en télédétection[13]. On trouve les modèles linéaires généralisés (GLM) (Helfer et Métral., 2000) avec notamment la régression logistique (Borghys et 2004) basés sur une distribution binomiale des données (par exemple présence/absence), ou leur forme plus flexible des modèles additifs généralisés (GAM) (Miller et Franklin, 2002). Des classifications non paramétriques tels que les réseaux neuronaux (neural networks, ANN) (Bischof et al., 1992 ; Houet et al., 2003) ou encore la classification par arbre de décision (CT) (Pal et Mather, 2001 ; Brown de Colstoun et al., 2003 ; Pal et Mather, 2003 ; Baker et al., 2006) se rencontrent également de plus en plus souvent.

- **Modèles linéaires généralisés**

Les GLM sont utilisés pour la cartographie de la végétation parce qu'ils peuvent être manipulés pour produire une probabilité de surface en référence à la présence de la végétation

(Miller et Franklin, 2002). Ils permettent de répondre à un objectif de reproductibilité en s'affranchissant d'un nouvel échantillonnage terrain par la production d'un modèle. En effet, dans les GLM, la combinaison de facteurs de prédiction est apparentée à la moyenne de la réponse des variables à travers une fonction de lien. L'utilisation de cette fonction de lien permet une transformation linéaire et le maintien des prédictions parmi une tranche de valeurs cohérentes pour la réponse de la variable. Ainsi, les GLM peuvent traiter des distributions gaussiennes, de Poisson, binomiales ou Gamma. Les GLM en distribution binomiale (exemple : présence/absence codées en 1 et 0) ayant une fonction de lien LOGIT sont communément utilisés pour modéliser la distribution d'espèces sous le nom de régression logistique. Les modèles de GLM peuvent être implémentés dans un système d'information géographique à partir du moment où l'inverse de leur fonction de lien peut être calculé (Guisan et Zimmermann, 2000). Chaque modèle est généré en multipliant chaque coefficient de régression avec sa variable prédictive associée. Avec le modèle binomial, la transformation logistique inverse est :

$$p(y) = \exp(X)/(1+\exp(X)),$$

où X est le facteur de prédiction linéaire correspondant à la régression logistique.

$$\text{Soit, } X= a+x_1b_1+x_2b_2+x_3b_3\ldots$$

où x_i correspond à la valeur des variables prédictives sélectionnées et b_i à leur coefficient de régression.

Cette transformation est nécessaire pour obtenir une valeur de probabilité comprise entre 0 et 1. Afin d'augmenter la précision et le pouvoir prédictif du modèle, le nombre de variables explicatives utilisées doit être préalablement et raisonnablement réduit (Harrell et al., 1996). Alors intervient la partie la plus difficile de la modélisation par GLM qui consiste à sélectionner les variables ou combinaisons de variables descriptives les plus pertinentes. On peut faire une sélection arbitraire, automatique (procédure de pas à pas descente ou ascendante, régression PLS, ACP, etc.), ou en suivant des principes physiologiques. On peut par exemple sélectionner les variables en s'appuyant sur la réponse spectrale de la végétation à cartographier. La sélection pas à pas automatique peut être utilisée. Cette analyse consiste à ajouter et enlever les variables au modèle l'une après l'autre pour voir si elles sont significatives. Elle a tendance à retenir les premières variables introduites, ce qui influence le résultat de l'analyse et dans le cas d'une régression logistique le critère de sélection de la variable limite fortement le nombre de variables en entrée. On peut également chercher à tester toutes les combinaisons de variables possibles afin d'obtenir le meilleur modèle. Mais ce type de méthodologie demande un temps de traitement long. La forme GAM des GLM

permet de s'affranchir de l'hypothèse de linéarité mais au risque d'une plus grande complexité de la formule résultante.

- **Les réseaux neuronaux**

La classification par réseaux neuronaux est basée sur le neurone définit par McCulloch et Pitts en 1943 et ne nécessite aucune hypothèse statistique. La méthode neuronale la plus utilisée en télédétection[13] est le perceptron multicouche (Pal et Mather, 2003). Dans cette application, les neurones sont regroupés en trois classes. Elle consiste en une couche d'entrée, au moins une couche cachée et une couche de sortie. Chaque entité d'une couche est reliée exclusivement à une entité de la couche suivante (Figure 15).

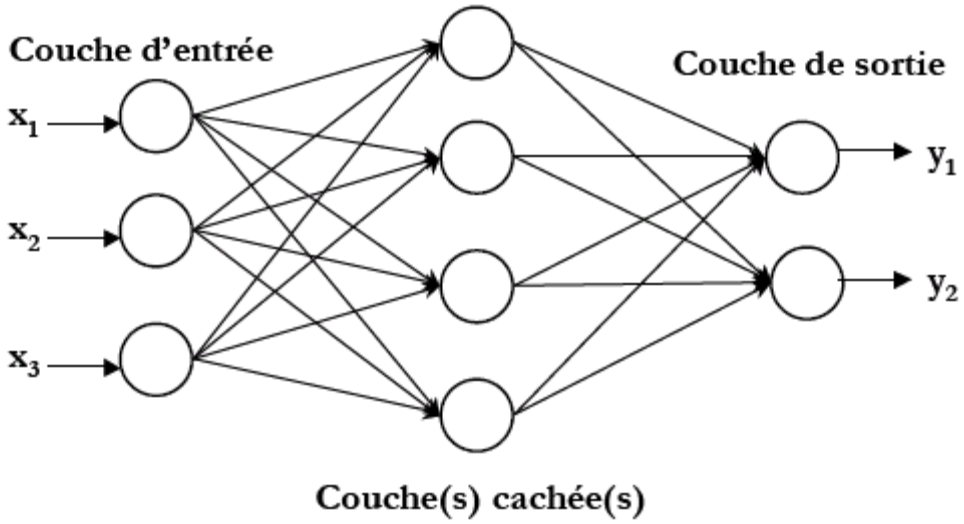

Figure 15 : Schéma général d'un perceptron selon Brostaux (2005)

Les méthodes de classification par réseaux neuronaux fonctionnent assez bien pour des échantillons de données plus petits que ceux nécessaires pour les procédures statistiques. Contrairement aux méthodes statistiques, la classification par réseaux neuronaux ne demande pas d'hypothèses statistiques spécifiques concernant la distribution des données ou l'échelle de mesure des entités utilisées dans l'analyse. Elle autorise ainsi l'utilisation d'échantillons non homogènes et bruités. Cette méthode présente certains inconvénients. La phase d'apprentissage consomme une puissance de calcul importante. Son application est gênée par la nécessité, pour l'utilisateur, de spécifier la configuration du réseau et de fournir des valeurs

pour un grand nombre de paramètres qui affectent la performance. Les réseaux neuronaux requièrent également une longue phase d'échantillonnage (Pal et Maher, 2003). Ils présentent également certains risques de sur-apprentissage, modélisant non seulement le concept mais également le bruit de fond qui l'accompagne. L'interprétation synthétique des estimateurs qu'ils fournissent est complexe du fait de la structure interne complexe et du syndrome de la boîte noire (Brostaux, 2005). Ils ne permettent donc pas d'obtenir une formule simple applicable à un nouveau jeu de données.

- **Les arbres de décision**

La combinaison de résultats de classification aisément interprétables avec la séparation précise des classes a certainement contribué à augmenter la popularité des méthodes par arbres de décision pour la classification de données multispectrales (Baker et al., 2006). Ce type de méthode produit une précision de 80 % dans la classification spécifique de certains types de zones humides ce qui correspond à 8 % de plus qu'une classification non–supervisée utilisée généralement (Sader et al., 1995). Les méthodes de classification par arbre utilisent l'information fournit par l'échantillon non plus globalement mais de manière hiérarchisée. Le principe de base d'un arbre de décision est le partitionnement successif d'un échantillon de données en des sous-échantillons de plus en plus homogènes en produisant des règles ou décisions optimales aussi appelées nœuds qui maximisent l'information gagnée et minimise le taux d'erreur dans les branches de l'arbre. Ces estimateurs issus des méthodes de segmentations récursives sont ainsi représentés sous forme d'arbres (Figure 16)

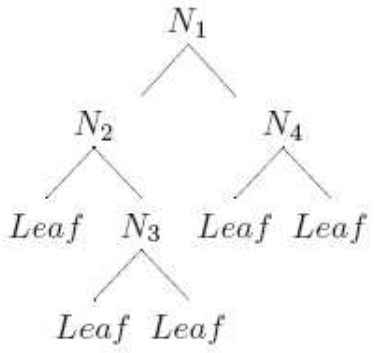

Figure 16 : Exemple d'un arbre de décision (Shi, 2006)

Ainsi chaque nœud interne de l'arbre (N_1, N_2, N_3, N_4) représente un choix entre un certain nombre d'alternatives et chaque nœud terminal aussi appelé feuille (*Leaf*) est marqué par une classification.

b - Choix de la méthode

Les GLM en régression logistique présentent l'inconvénient d'une hypothèse de linéarité sous jacente qui peut être parfois remise en cause car dans certains cas la relation physique entre données spectrales et variables à prédire ne peut pas être approchée de façon linéaire avec les données de télédétection[13] (Bertrand et al., 1999). Ils nécessitent une présélection des variables qui peut s'avérer fastidieuse et la formule obtenue reste compliquée. Les réseaux neuronaux et les arbres de décision ont tendance à produire des classifications de précision identique lorsqu'ils sont testés sur le même type de données de télédétection[13], et peuvent être parfois plus performants que d'autres types de classification (Strahler et al., 1999). Les réseaux neuronaux présentent l'inconvénient de l'effet « boîte noire » responsable du fait que le résultat de l'apprentissage est non interprétable par l'utilisateur (Brostaux, 2005). Les arbres de décision présentent un biais pour des échantillons à effectifs très élevés (Brostaux, 2005 ; Mahesh et Mather, 2003). Ils ne nécessitent donc pas de très larges échantillons pour être efficaces (Mahesh et Mather, 2003), Tout en conservant l'avantage de s'affranchir d'hypothèses statistiques préliminaires, ils sont moins demandeurs en temps de traitement informatique et fournissent aux analystes et utilisateurs, du fait de leur structure hiérarchique, une méthode simple et robuste pour interpréter, tester et analyser les résultats (Mahesh Pal et Mather, 2001 ; Brown de Colstoun et al., 2003 ; Gomez-Chova et al., 2003). Ils ne nécessitent pas de présélection des variables. Ils sont peu perturbés par les individus extrêmes et peu sensibles aux bruits des variables discriminantes. Ils permettent de sélectionner et de classer tout type, et un grand nombre, de variables explicatives (Gomez-Chova et al., 2003). Les bases de données présentant des données manquantes peuvent être utilisées. Ils représentent ainsi une méthode de mise en évidence de nouvelles connaissances notamment dans le domaine de la télédétection (Gomez-Chova et al., 2003). Ils présentent aussi l'avantage de pouvoir classifier efficacement de nouvelles sources de données. Cependant, il est conseillé d'appliquer ces méthodes, préférentiellement, à des échantillons ayant un nombre d'individus supérieur à 50 car la variabilité des petits échantillons peut favoriser une certaine instabilité

des arbres de décision (Brostaux, 2005). Le temps d'apprentissage et le choix des multiples méthodes existantes peuvent également présenter une contrainte à prendre en considération pour de nouveaux utilisateurs.

Dans le cadre de notre étude, nous souhaitions mettre en place des modèles reproductibles et facilement compréhensibles pour des gestionnaires d'habitats naturels. Il nous était également nécessaire de sélectionner les variables les plus pertinentes pour la discrimination des roselières, herbiers aquatiques et zones inondées. Les arbres de décision permettent de prendre en compte un grand nombre de variables descriptives, de les sélectionner en fonction du succès de la classification, et offrent des résultats facilement interprétables et ré-applicables à d'autres jeux de données. Cette méthode de classification a donc été choisie pour répondre à nos objectifs.

c - La classification par arbre de décision

L'idée de base des arbres de décision est la suivante : premièrement, on sélectionne un attribut à placer à la racine et on construit des branches pour cet attribut basées sur un critère (exemple l'indice de Gini). Deuxièmement, on sépare les sites d'échantillonnage en sous-échantillons, un pour chaque branche partant de la racine, le nombre de sous-échantillons étant le même que le nombre de branches. Troisièmement, on répète cette étape pour l'une des branches, en utilisant seulement l'échantillon disponible à cette branche. Un ordre est fixé pour étendre les nœuds (en général de la gauche vers la droite). Quatrièmement, si pour un échantillon donné, la procédure donne toujours la même règle, connu pour être un nœud pur, on stoppe l'expansion. Ce nœud sera dit terminal (ou feuille). Ce procédé de construction continue jusqu'à ce que tous les nœuds soient purs. Cette série de partition récursive (cf. Annexe 10) est basée sur un critère de pertinence qui évalue l'homogénéisation des échantillons résultants. L'arbre ainsi généré est ensuite éventuellement simplifié (ou élagué) pour éliminer les risques de sur-apprentissage (Brostaux, 2005) et ainsi pouvoir fournir une prédiction sur un nouveau jeu de données.

La méthode de classification par arbre comprend donc deux étapes : une phase d'élaboration de l'arbre maximal et une phase d'élagage.

- **Construction de l'arbre**

L'homogénéité des nœuds est définie par l'impureté, une mesure qui prend la valeur zéro pour un nœud complètement homogène et qui augmente en fonction de la diminution de l'homogénéité. Ainsi maximiser l'homogénéité revient à minimiser l'impureté. Il existe trois types de mesure de l'impureté pour les arbres de classification (De'ath et Fabricius, 2000). Cette mesure est basée sur la proposition de réponses dans chaque catégorie. L'indice d'information ou entropie (Quinlan, 1986) est calculé sous la forme : $-\Sigma\ p \ln p$ (où p est la proportion de réponses dans chaque catégorie). Il est équivalent à l'indice de diversité de Shannon-Weiner. Cet indice forme des groupes en minimisant la diversité intra-groupe. L'indice de Gini (Breiman et al., 1984), noté $1-\Sigma\ p^2$, tente de séparer la plus importante catégorie de l'échantillon en groupes séparés tandis que l'indice d'information tente de former des groupes comprenant plus d'une catégorie dans le nœud précédent. L'indice Twoing (Breiman et al., 1984) peut être utilisé pour plus de deux catégories. Il définit deux « super-catégories » à chaque séparation pour lesquelles l'impureté est définie par le critère de Gini. Il peut aussi être utilisé pour ordonner les catégories. Le choix du critère de partition est d'une importance secondaire pour les performances finales du classificateur, cela simplifie son choix car la gamme de critères disponibles est étroitement liée à chaque algorithme. Les méthodes d'arbres de classification les plus utilisées, et notamment avec des données de télédétection[13] sont:

- CART (Classification And Regression Tree) développé par Breiman et al. en 1984 associé au critère de Gini, utilisé pour associer un test à un nœud.
- ID3 développé par Quinlan en 1983 et amélioré en 1993 par une nouvelle version C4.5. Ici, le choix du test associé au nœud se fait à l'aide de la fonction Gain basée sur la fonction entropie.

La méthode de Quinlan (1983) autorise les divisions multivariées soit une branche par item de l'attribut sélectionné (Figure 17).

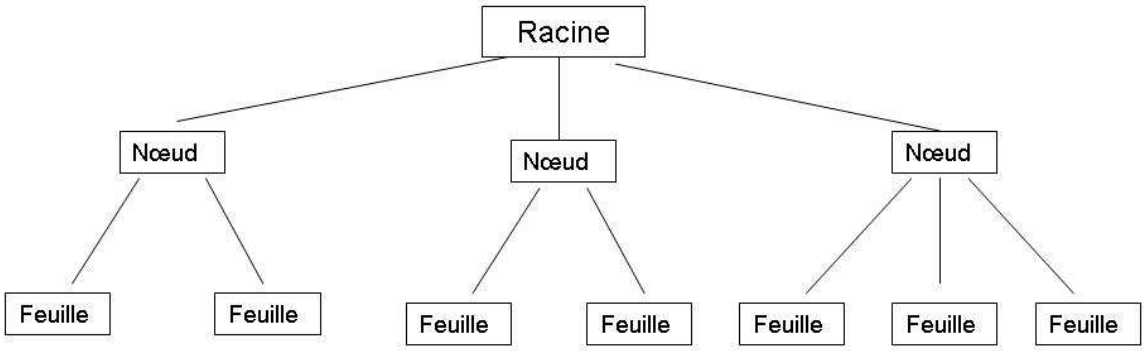

Figure 17 : Schéma d'un exemple d'arbre de classification autorisant les divisions multivariées

La méthode CART impose, quant à elle, une partition dichotomique (Figure 18), chaque nœud engendrant deux branches filles créées par regroupement des items de l'attribut sélectionné (Breiman et al., 1984). Elle présente ainsi l'avantage de la simplicité d'interprétation et est moins sensible au problème de fragmentation des données.

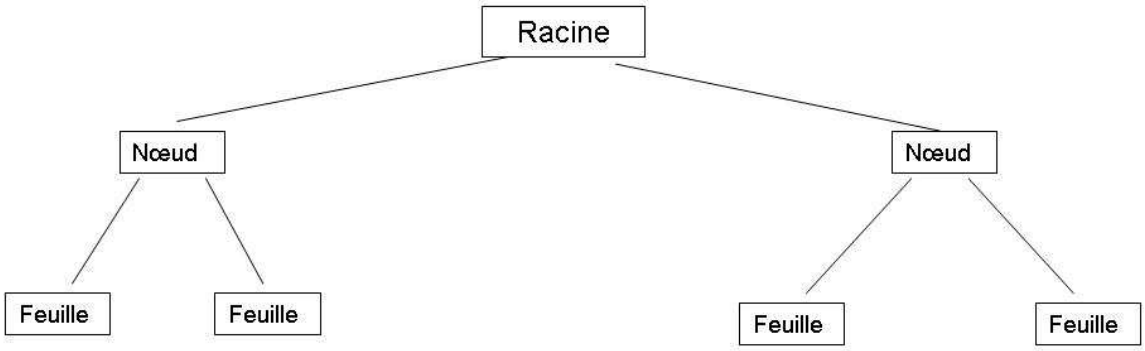

Figure 18 : Schéma d'un exemple d'arbre de classification à partition dichotomique

Les critères de partition comme décrits ci-dessus ne permettent pas d'arrêter la partition prématurément c'est-à-dire avant d'obtenir des nœuds purs. Dans la plupart des cas, construire un arbre jusqu'à ce que toutes les feuilles contiennent des données pour une seule classe peut entraîner un problème de sur-apprentissage. Ce peut être le cas, par exemple, d'un échantillon bruité par l'influence de facteurs qui ne peuvent pas être mesurés comme c'est fréquent en situation réelle (Mingers, 1989) ou du fait d'un échantillon non représentatif de la population. Si l'échantillon contient des erreurs, le sur-apprentissage des données peut conduire à une faible performance dans certaines situations. Afin de minimiser ce problème,

l'arbre original (ou arbre maximal) doit être élagué afin de réduire l'erreur pour que de nouvelles données soient classifiées correctement.

- **Elagage**

Considérer comme meilleur résultat l'arbre étendu jusqu'à minimisation maximale du taux d'erreur en resubstitution[12] présente deux inconvénients notables (Breiman et al., 1984). Si la règle d'arrêt est basée sur un gain en performance trop petit du prédicteur, alors un arbre trop étendu en résultera. Egalement, si le calcul du critère d'impureté donne une grande valeur, alors les divisions basées sur les interactions entre les variables ne seront pas trouvées sans qu'au moins un des principaux effets associés soit assez important pour générer une partition.
L'élagage d'un arbre consiste donc à supprimer certains sous-arbres (Figure 19) augmentant le taux d'erreur en resubstitution[12] mais, dans un domaine bruité ou mal échantillonné, permettant d'augmenter la précision pour l'application du modèle sur de nouveaux jeux de données (Minger, 1989 ; Quinlan, 1987 ; Breiman et al., 1984).

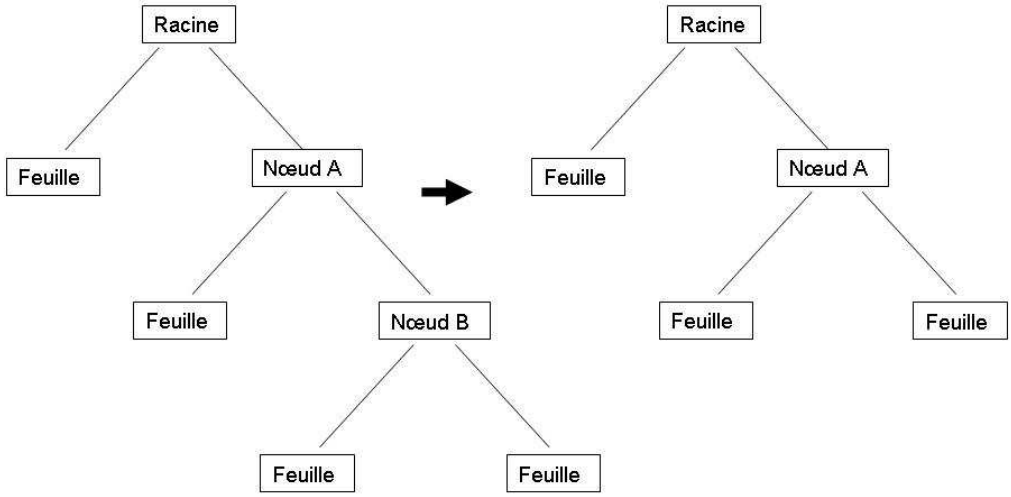

Figure 19 : Elagage du sous-arbre issu du nœud A, remplacé par une feuille

Les algorithmes d'élagage ont ainsi été créés dans le but de prévenir le sur-apprentissage et de simplifier la variance. Il y a deux types d'élagage. Le pré-élagage (pre-pruning) est basé sur l'arrêt en cours de croissance de l'arbre, alors que le post-élagage (post-pruning) s'attache à éliminer les branches qui n'améliorent pas les performances du modèle. Trois paramètres peuvent permettre le choix des branches à élaguer : l'utilisation d'un échantillon indépendant, une validation croisée[15] ou une correction de l'erreur de resubstitution (Brostaux, 2005). Le choix de la méthode reste lié à l'utilisation de l'algorithme choisi (Brostaux, 2005). Les

méthodes intégrées à CART (Breiman et al., 1984) et C4.5 (Quinlan, 1993) se sont donc imposées. Notons que des études comparatives les ont évaluées comme étant les plus efficaces (Mingers, 1989 ; Esposito et al., 1999). Nous nous attacherons donc, à décrire seulement ici, ces deux méthodes.

L'élagage dit du coût-complexité (cp) minimal (Minimal cost complexity pruning) a été mis au point par Breiman et al. en 1984 pour l'algorithme CART. Cette méthode comprend deux étapes. La première consiste à générer une série d'arbres élagués à différents stades et la deuxième permet de sélectionner parmi l'ensemble de ces arbres selon l'erreur de classification pour chacun d'eux en utilisant un échantillon indépendant. Cette méthode prend donc en compte à la fois le taux d'erreur en généralisation[11] et la complexité de l'arbre, en d'autres termes sa taille. Avec un échantillon d'apprentissage[4] l'élagage conduira toujours à une augmentation du pourcentage d'erreur. Le paramètre de coût-complexité est ainsi calculé en divisant cette augmentation par le nombre de feuilles dans le sous-arbre, ce qui donne une mesure de la réduction de l'erreur par feuille pour ce sous arbre (voir Mingers, 1989 pour un exemple détaillé du calcul du cp). Ainsi, ce paramètre offre une mesure de la valeur du sous-arbre c'est-à-dire de la réduction de l'erreur par feuille (Mingers, 1989). Le cp est calculé pour chaque sous-arbre et présente une valeur croissante au fur et à mesure de la diminution des sous-arbres. La seconde étape consiste à sélectionner un de ces sous-arbres à l'aide d'un échantillon test[5] ou de la validation croisée[15]. Pour cela on utilise le critère de mauvaise classification. L'arbre optimal est sélectionné grâce au plus petit taux de mauvaise classification estimé sur l'échantillon test[5] ou par validation croisée[15]. Il y a alors deux règles de sélection. La première est la règle du CV0-SE (Esposito et al., 1999) avec laquelle l'arbre sélectionné sera celui dont le taux de mauvaise classification sera le plus petit. La seconde est la règle du règle 1-SE (Breiman et al., 1984) qui sélectionne le plus petit arbre ayant une erreur inférieure ou égale au taux d'erreur minimal additionné à son erreur standard. L'erreur standard de la mesure de mauvaise classification (Mingers, 1989) est calculée ainsi :

$$SE = \sqrt{\frac{R \times (100 - R)}{N}}$$

où R est la mesure de mauvaise classification de l'arbre élagué et N le nombre d'observations dans l'échantillon test[5] (ou dans l'échantillon E-Ei dans le cas de la validation croisée[15], méthode CV1-SE décrite par Esposito et al. 1999). Dans cette méthode la taille des échantillons peut influer sur la méthode d'élagage. Dans le cas de grands échantillons, ceux-ci

sont en général séparés en deux pour former un échantillon d'apprentissage et un échantillon test[5]. Dans le cas d'un petit échantillon, l'utilisation de la validation croisée[15] est obligatoire. Cependant, l'utilisation systématique de celle-ci, peu importe la taille de l'échantillon, fait partie des meilleures méthodes d'élagage testées par Esposito et al. (1999) et Mingers (1989).

La méthode d'élagage basée sur l'erreur (error-based pruning) a été développée par Quinlan (1993) pour C4.5 et est une amélioration de la méthode de l'erreur pessimiste qu'il avait développée en 1987 pour l'algorithme ID3. Cette méthode utilise l'information de l'échantillon d'apprentissage à la fois dans l'élaboration et la simplification de l'arbre. L'élagage est basé sur l'erreur de l'arbre et l'erreur du nœud. Le choix de la branche à élaguer se compose de deux aspects. Dans un premier cas on calcule l'erreur du nœud *i* et de son sous-arbre *a* (ayant *i* comme racine), et l'erreur du sous-arbre *a'* issu de *i* et présentant le plus grand effectif. Si l'erreur de *i* est inférieure à celle de *a* et de *a'*, on transforme ce nœud en feuille (le sous-arbre est élagué). Dans le cas ou l'erreur de *i* est supérieure, mais que l'erreur de *a'* est inférieure à celle de *a,* alors *a'* sera conservé et *a* sera supprimé. Si l'erreur de *a* est inférieure, le sous-arbre sera conservé. Cette procédure est réalisée pour chaque nœud. Le calcul des erreurs est réalisé ainsi : l'erreur du sous-arbre est la somme des erreurs de ses feuilles, l'erreur du nœud est calculé à partir du nombre d'observations mal classifiées et du nombre total d'observations à ce nœud. Le calcul de l'erreur est basé sur une hypothèse de distribution binomiale et la probabilité des observations mal classifiées. Cette dernière ne pouvant être estimée dans la réalité, le calcul se base sur une limite supérieure de confiance autour de l'erreur sur l'échantillon d'apprentissage utilisé pour construire l'arbre.

- **Inconvénients des arbres de classification**

Les processus d'élaboration des arbres de classification présentent certains inconvénients. Les arbres de classification peuvent être considérés comme une méthode non-optimale du fait de « l'effet papillon » qui caractérise leur instabilité. En effet, si des variables ont un pouvoir prédictif équivalent, la sélection de celle correspondant au maximum est dépendante de l'échantillon d'apprentissage. L'instabilité se marque surtout sur les niveaux inférieurs (feuilles) mais pas exclusivement. Ainsi un arbre très différent visuellement peut être obtenu en modifiant seulement quelques observations de l'échantillon. Cependant, si l'arbre de classification semble très différent, la variabilité de la prédiction sur un individu pris au hasard dans la population n'est pas aussi marquée et, généralement, on lui attribuera la même

étiquette. L'utilisation de règles heuristiques peut également poser le problème de la non remise en question des choix dans la construction de l'arbre (pas de retour en arrière). Enfin les arbres de classification sont construits sur la base de l'optimisation par maximisation de la précision sur l'échantillon qui leur est fourni en entrée. Ils considèrent donc que cet échantillon est représentatif de la variabilité et des caractéristiques de la population sous-jacente et que la fréquence relative de chaque classe est représentative de cette population (Breiman et al., 1984).

Mais en réalité, et notamment pour l'échantillonnage de l'occupation du sol en télédétection[13] (par exemple en cas de difficultés d'accès au terrain), il est parfois difficile de traduire, à travers l'échantillonnage, la distribution réelle des classes sur l'ensemble de la population. Un paramètre supplémentaire peut donc intervenir dans la construction de l'arbre : le paramètre de priorité. Sa valeur par défaut est calculée en fonction de l'occurrence des éléments dans la base de données mais, ajusté intelligemment, ce paramètre peut aider considérablement à la construction du meilleur arbre de classification (Breiman et al., 1984). Il peut être utilisé pour ajuster le pourcentage de mauvaise classification de chaque classe. Ainsi, la méthode est de construire plusieurs arbres en faisant varier le paramètre de priorité jusqu'à l'obtention de la meilleure classification possible pour chaque classe. Dans le cas de données de télédétection[13], il peut améliorer les résultats de la classification en aidant à résoudre la confusion entre classes qui sont difficilement séparables et en réduisant le biais quand un échantillon n'est pas représentatif de la population à classifier. Il est à noter cependant, que seuls les domaines se trouvant dans des zones de chevauchement de deux classes sont affectés par la modification de ce paramètre (McIver et Friedl, 2002).

d – Sélection du logiciel

Nous avons fait le choix d'utiliser la fonction rpart (Recursive PARTitioning) (Therneau et Atkinson, 1997) sous le logiciel R. Le projet R est une conception libre du langage S dont « *l'objectif est de fournir un environnement interactif d'analyse de données, doté d'outils graphiques performants et permettant une adaptation aisée aux besoins des utilisateurs, depuis l'exécution de tâches routinières jusqu'au développement d'applications entières* » (Brostaux, 2005). Rpart a été élaboré sur la base de CART (Breiman et al., 1984). Cet algorithme permet à la fois de créer et d'élaguer un arbre sous la forme binaire. Il intègre

entre autres les paramètres de validation croisée[15], d'élagage par coût-complexité et la possibilité de modifier simplement le facteur de priorité des classes. Ces paramètres répondent aux particularités de nos échantillons de petites tailles présentant des classes à effectifs non-équilibrés.

e – Mise en évidence des paramètres pouvant expliquer les sources d'erreur de la classification

Nous avons utilisé les GLM avec une sélection pas à pas ascendante des variables en entrée, d'une part sous forme de régression logistique pour expliquer si les facteur structuraux des phragmites et les paramètres caractérisant les marais à espèces submergées pouvaient entraîner des erreurs d'identification des roselières et herbiers aquatiques et d'autre part sous forme de régression multiple pour comprendre comment les mesures relevées sur le terrain pouvaient expliquer l'imprécision de la classification de la présence d'eau.

2 - Modélisation de la qualité des roselières

a – Approche multitemporelle

Contrairement à la classification qui vise à expliquer une variable qualitative, la modélisation de la qualité des roselières porte sur des variables dépendantes quantitatives permettant l'utilisation de GLM sous la forme de régressions multiples ou de méthodes adaptées comme la forme « régression » des arbres de décision. La régression peut être abordée selon différents objectifs. Elle peut servir à décrire en cherchant les liaisons entre une variable dépendante (également nommée réponse ou variable exogène) afin d'évaluer, justement, la dépendance entre cette variable et d'autres variables indépendantes et potentiellement explicatives. Elle peut être envisagée dans le cadre explicatif comme dans le cas de la confirmation ou de la précision de résultats théoriques. Enfin, la régression peut être utilisée à des fins prédictives avec la constitution d'un modèle ou équation linéaire de prédiction permettant d'évaluer la précision de ces prédictions.

Les algorithmes d'arbres de régression sont de bonnes méthodes dans un objectif de description ou d'explication de la variable dépendante, mais ne sont pas optimaux pour la prédiction quantitative d'une variable à partir de variables explicatives. En effet, ils donnent une estimation à partir d'une moyenne ou d'une médiane (selon l'algorithme utilisé) des prévisions pondérées par la qualité respective de chacune de ces prévisions et n'offrent donc pas une bonne précision de la valeur prédite. Aussi, dans le cadre de notre suivi de l'évolution de la qualité des roselières impliquant des mesures sur les tiges de roseaux, nous avons retenue la méthode traditionnelle de régression multiple. « Cette méthode est en effet, l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles » (Besse, 2003).

Le principe de la régression multiple repose sur l'équation de la régression linéaire simple de type :

$$Y = ax + b$$

où Y est la variable dépendante et x la variable explicative. Dans le cas de la régression multiple à *p* variables explicatives, l'équation s'écrit :

$$Y=a_1x_1+a_2x_2+\ldots+a_ix_i+\ldots a_px_p+b.$$

Sa représentation graphique est alors un espace à *p* dimensions.

Comme pour une régression linéaire simple, on mesure la variance expliquée par la régression à l'aide d'un coefficient de détermination multiple R^2, calculé selon le rapport :

$$R^2 = SSR/SST$$

où SSR est la somme des carrés de la régression, SSE la somme des carrés des résidus et SST la somme totale des carrés (SST = SSE + SSR). Plus R^2 est proche de 1, plus la proportion de variance expliquée par l'équation s'approche de 100%. Idéalement, le meilleur modèle est celui qui combine parcimonie et fiabilité. Ainsi un modèle plus performant pour prédire une relation pourra être préféré à un modèle présentant une valeur de R^2 plus élevée et présentant potentiellement des risques de colinéarité[2]. Pour un nombre *n* de variables explicatives, il n'est pas envisageable d'examiner les 2^n modèles possibles. Différentes méthodes de sélection existent selon le critère de sélection du meilleur modèle, du temps de calcul, de l'optimalité du processus et du logiciel utilisé. Etant donné le nombre de variables explicatives en entrée dans le cadre de notre étude, nous avons utilisé une sélection des variables à l'aide d'une méthode pas à pas ascendante. Dans cette procédure, à chaque « pas » (étape), la variable la plus significative entre dans le modèle. Sa significativité est définie par le minimum de la valeur de F reliée à la statistique de Fischer qui compare deux modèles. La

fin du processus est dictée par une valeur seuil de F (dans notre cas F=1) en entrée comme en sortie. On teste ensuite la pertinence de toutes les variables du modèle pour éliminer celles devenues non significatives suite à l'entrée de cette nouvelle variable. L'obtention du meilleur modèle est donc atteinte lorsqu'aucune des variables restantes ne peut expliquer une partie significative de la variance restante ($p < 0.05$). La performance du modèle repose donc sur celle de quelques variables sélectionnées. Cette méthode a tendance à retenir les premières variables introduites, ce qui influence le résultat de l'analyse. Considérant le très grand nombre de variables explicatives à notre disposition et le faible nombre de nos effectifs, nous avons effectué une première sélection de variables afin de maximiser la robustesse des modèles produits.

Les variables descriptives ont donc été sélectionnées à partir de données issues de la réponse spectrale des mêmes sites sur les mêmes mois de deux années consécutives. Deux constats ont autorisé cette présélection des variables descriptives. D'une part, il est difficile d'imaginer que la réponse spectrale d'un pixel de phragmitaie ne puisse pas, entre deux années consécutives, présenter une relation significative ($r = 0.325$; $dl=38$; $p>.05$). Dans notre cas cela reviendrait à dire que l'auto-corrélation spatiale d'une comparaison interannuelle de mêmes sites devrait représenter au moins 10% de la variance. D'autre part, la réponse moyenne de l'ensemble des mêmes sites de phragmitaies ne peut montrer des différences hautement significatives ($t = 2.75$, $dl=38$, $p<0.01$) après une année. Il y a alors au minimum 1 chance sur 100 que la réponse interannuelle moyenne des mêmes sites soit tirée de la même population statistique. Une variable descriptive dite stable doit permettre de combiner ces deux constats.

b - Approche mono-date

Cette modélisation par GLM constitue une approche multitemporelle permettant de produire des modèles prédictifs spécifiques à chacun des paramètres de la qualité et de l'intérêt des phragmitaies en utilisant les données issues de l'échantillonnage réalisé en 2005. Nous avons également abordé une approche en monodate apportant une identification des indices multispectraux en relation aux facteurs de qualité et d'intérêt avifaunistique des phragmitaies. Ceci a été évalué par test de la corrélation (*r* de Pearson) entre les paramètres mesurés dans

les phragmitaies et les valeurs des canaux et des indices multispectraux pour chaque mois d'acquisition des images.

D - Constitution de la base de données

Nous avons trois types de données : les données de réflectance extraites des images corrigées radiométriquement et géométriquement, les données des relevés botaniques et les données de GPS. Il s'agit ici de définir les variables descriptives et dépendantes utilisées pour chaque test statistique relatif à nos objectifs. La constitution de la base de données se compose donc de trois étapes : repérage des sites sur les images satellitales, extraction de la réflectance correspondante pour chaque site à chaque date et canal de l'image, et calculs des paramètres et indices permettant de répondre à nos objectifs.

1 - Les variables descriptives

Le repérage des sites sur les images se base sur les relevés de points GPS effectués au centre des carrés. Nous avons créé une couche dans un système d'information géographique (SIG) situant les points centraux de nos sites d'échantillonnage que nous avons superposé aux images afin d'extraire la donnée des pixels. L'extraction a été réalisée sous ArcGIS avec la commande statistiques spatiales du module d'analyse spatiale. Nous avons ainsi extrait pour chaque site échantillonné et chaque catégorie « non-site » une valeur de réflectance pour chaque canal de chaque image.

A partir de ces données de réflectance, nous avons également calculé des indices multispectraux adaptés aux bandes de SPOT 5. Les indices ci-après ont été ajoutés à la liste des variables descriptives (canaux SPOT 5 de chaque date) :

- le simple ratio (SR - simple ratio) qui est le rapport de la réflectance de la bande rouge (R) et proche infrarouge (PIR) (Pearson et Miller, 1972, cités par Bannari et al., 1995). Il peut être utilisé pour mettre en évidence le contraste entre le sol et la végétation :

$$SR = \frac{R}{PIR}$$

- Le simple indice de végétation (VI - vegetation indice) (Lillesand and Kiefer, 1987) est obtenu par un rapport du canal proche-infra rouge sur le canal rouge. Il est également un outil de mise en valeur du contraste entre le sol et la végétation:

$$VI = \frac{PIR}{R}$$

- L'indice de végétation par différence (DVI - differential vegetation indice) (Richardson and Everitt, 1992) est la différence entre le canal proche infrarouge (PIR) et le canal rouge (R). Le DVI est plus faible pour une végétation sèche. Pour une végétation saine et verte, la chlorophylle absorbe la part rouge du spectre électromagnétique. Le proche-infrarouge, fortement diffusé par la structure des feuilles, est caractérisé par une réflectance importante de la végétation. En situation de stress hydrique, cela est inversé : le rouge présente une réflectance plus importante que le proche infrarouge :

$$DVI = PIR - R$$

- L'indice de stress hydrique (MSI - moisture stress index) est un rapport du canal infra-rouge moyen (MIR) sur le canal proche-infra-rouge (PIR). Il a été mis au point pour Landsat par Hunt and Rock (1989). Il est corrélé à la réflectance de l'eau liquide dans la feuille et possiblement dans la canopée. Il est ainsi lié à la teneur en eau de la végétation :

$$MSI = \frac{IRM}{PIR}$$

- L'Indice de végétation normalisé (NDVI - Normalized Difference Vegetation Index), proposé pour la première fois par Rouse et al en 1973 (Bannari et al , 1995) est l'indice de végétation le plus connu et le plus utilisé en télédétection[13] en relation avec la signature spectrale de la végétation. Il est le rapport entre la différence du proche-infrarouge (PIR) et du rouge (R) :

$$NDVI = \frac{PIR - R}{PIR + R}$$

Malgré sa normalisation, le NDVI est sensible à la géométrie de vue et d'illumination, notamment dans les régions où la densité de végétation est faible et

où la présence de sol est importante (Bannari et al ,1995). Le NDVI est également rapidement saturé en présence de végétation dense et la contribution du sol en région de faible densité végétale peut rendre son interprétation douteuse.

- L'indice de végétation ajusté du sol permet de minimiser son influence (SAVI - Soil Adjusted Vegetation Index) (Huete, 1988). Cet indice, basé sur le NDVI, introduit une correction L de la brillance pour le signal au satellite. Une correction pour le sol donne une information plus exacte de la condition de la végétation propre. Huete(1988) a démontré qu'une valeur de 0.5 pour l'ajustement offre une correction optimale de la rétrodiffusion du sol à travers le couvert végétal. Le SAVI est ainsi calculé :

$$SAVI = \frac{(1 + L)\,(PIR - R)}{PIR + R + L}$$

Soit avec L = 0.5 :

$$SAVI = \frac{1{,}5*(PIR - R)}{PIR + R + 0{,}5}$$

- L'indice de végétation ajusté du sol optimisé (OSAVI – Optimized SAVI), vient du fait que l'ajustement du SAVI n'est en réalité pas constant et demande une correction adaptée à la végétation étudiée (Rondeaux et al., 1996). Le OSAVI est plus adapté aux régions agricoles de moyenne latitude présentant un développement homogène de la végétation. Il est exprimé ainsi :

$$OSAVI = \frac{PIR - R}{PIR + R + 0{,}16}$$

- L'indice normalisé de différence d'eau (NDWI – normalized difference water index) existe sous deux formes. Le NDWI de Gao (1996) est calculé à partir des canaux infrarouge moyen et proche infrarouge et est corrélé au contenu en eau de la végétation. Sa valeur augmente d'un sol sec vers l'eau libre et avec le pourcentage de recouvrement de la végétation car il est sensible à la quantité totale d'eau liquide dans la superposition des feuilles. Il est lié ainsi à la quantité de molécules d'eau liquide dans la canopée (Gao, 1996). Le NDWI de McFeeters

(1996) est formé à partir des canaux vert et proche infrarouge et est utilisé pour identifier les surfaces aquatiques. Les zones d'eau libre ont des valeurs positives tandis que les zones de sol et de végétation terrestre ont des valeurs inférieures ou égales à 0. McFeeters (1996) suppose qu'il pourrait également apporter une estimation de la proportion en matière en suspension et en chlorophylle *a* dans l'eau. SPOT 5 ayant deux canaux dans l'infrarouge, en nous basant sur l'indice NDWI de Mc Feeters, nous avons défini deux types d'indice : l'un calculé à l'aide de la bande 3 (PIR) que nous appellerons NDWIF1 (correspondant au NDWI de Mc Feeters, 1996) et l'autre à partir de la bande MIR nommé NDWIF2. Le NDWIF2 s'apparente au MNDWI (modified normalized difference water index, Hanqiu, 2006) qui remplace un canal du proche infrarouge utilisé par McFeeters en 1996 par un canal du moyen infrarouge de LANDSAT. Ainsi nous avons utilisé trois types d'indice normalisé de différence d'eau selon les formules :

$$NDWI = \frac{PIR - IRM}{PIR + IRM}$$

$$NDWIF1 = \frac{V - PIR}{V + PIR} \quad NDWIF2 = \frac{V - IRM}{V + IRM}$$

- L'indice de différence entre la végétation et l'eau noté DVW par Gond et al (2004) permet d'accroître à partir du NDWI de Gao (1996) et du NDVI (Rouse et al.,1973 in Bannari et al , 1995) la mise en évidence de l'eau libre et de zones humides dans une région aride. Nous avons utilisé la différence entre l'eau (water) et la végétation appelée l'indice DWV. Ainsi le DWV sera positif pour les territoires inondés et négatif pour les sols et la végétation. Il est exprimé ainsi :

$$DWV = NDWI - NDVI$$

Afin de renforcer les différences de réponse spectrale entre les saisons, nous avons également ajouté les différences de canaux entre chaque date. Par exemple la différence entre la bande 3 (ou canal 3) de juin et la bande 3 de mars que nous avons noté c30603.

L'ensemble de ces variables a été pris en compte pour la reconnaissance des communautés végétales. Le suivi des zones inondées étant mensuel, nous avons considéré l'ensemble des

bandes spectrales (B1, B2, B3, B4) et indices multispectraux présentés ci-dessus. Pour la qualité des roselières, les résultats obtenus avec l'ensemble des canaux et indices n'étaient pas satisfaisants. Nous avons donc cherché d'autres variables afin d'améliorer la qualité du suivi et avons ainsi ajouté les différences entre indices de chaque date sur la base des travaux réalisés par Yazdani et al (1981, in Bannari et al., 1995) à partir de l'indice de végétation multitemporel. Par exemple, la différence entre le NDVI de mai et le NDVI de juin que nous avons noté NDVI05-06.

2 - La variable dépendante pour le suivi de la végétation

Pour le suivi en présence/absence de la végétation, la variable dépendante a consisté en un codage des valeurs 1 et 2 correspondant respectivement à absence et présence ; les « sites » étant codés 1 et les non-sites codés 2.
Pour la reconnaissance de la végétation, la formation à identifier (ex : phragmitaies) a été codée 1 tandis que les autres formations (ex : herbiers, scirpaies) et les non-sites ont été codés 2.

3 - La variable dépendante pour le suivi de l'eau

Trois étapes ont constitué l'élaboration de la variable dépendante pour le suivi de la présence d'eau. Une première étape a consisté en un calcul du niveau d'eau au site à chaque date d'image à partir des relevés périodiques (piézomètre ou échelle) et des mesures d'étalonnage. Dans une seconde étape et dans le cas où il y avait plus d'un jour entre la date du relevé et la date d'acquisition de l'image, une extrapolation du niveau d'eau était faite en présumant une évolution linéaire en fonction du temps. Dans une troisième étape, nous avons isolé le niveau minimal et le niveau maximal parmi les huit mesures effectuées sur les diagonales de notre carré d'échantillonnage et les avons regroupées selon les catégories suivantes :

- niveau minimal et maximal ≤ 0 : classe 1
- niveau minimal et maximal > 0 : classe 2
- niveau minimal < 0 et maximal >0 : classe 3

Nous avons retenu la classe 1 afin d'évaluer les potentialités des images satellitates pour reconnaître les zones sèches et la classe 2 pour la reconnaissance des zones en eau (milieu en

eau). La classe 3, bien qu'intéressante d'un point de vue écologique, n'a pas fait l'objet d'analyses car cette classe est susceptible d'apporter des pixels mixtes et donc une source d'erreur.

En 2005, 215 sites au total ont ainsi pu être retenus dont 34 sites secs et 181 sites en eau. En 2006, 248 sites au total ont été utilisés dont 77 sites secs et 171 sites en eau. Le nombre de pixels de chaque catégorie de l'échantillon d'entraînement[4] (comprenant les données relatives aux images de décembre 2004, mars, mai, juin, juillet et septembre 2005) et de l'échantillon de validation[5] (construit à partir des données relatives aux images de décembre 2005, mars, mai, juin, juillet, août et octobre 2006) est présenté au tableau 4.

Tableau 4 : Nombre de pixels dans chaque catégorie de l'échantillon d'entraînement[4] et de validation[5]

Catégories	Nombre de pixels	
	échantillon d'entraînement[4]	**échantillon de validation[5]**
Sec	41	97
En eau	219	271
Total	**260**	**368**

4 - Les variables dépendantes pour le suivi de la qualité et de l'intérêt avifaunistique des phragmitaies

Le suivi de la qualité et de l'intérêt avifaunistique des roselières s'appuie sur une variable dépendante continue et non binaire comme dans les cas précédents. Six variables continues sont particulièrement intéressantes pour caractériser les phragmitaies (Tableau 5).

Tableau 5 : Récapitulatif des variables permettant un suivi de la qualité et de l'intérêt avifaunistique des roselières

Désignation des variables	Paramètres mesurés dans les roselières
Hauteur des tiges vertes	Hauteur des tiges vertes en cm
Pourcentage de trouées	Pourcentage de trouées[14] (T/(T+R)*100)
Nombre de tiges vertes	Nombre total de tiges vertes dans les 4 quadrats de 50 X 50 cm
Rapport sec/vert	Rapport du nombre de tiges sèches sur le nombre de tiges vertes
Nombre de tiges sèches	Nombre total de tiges sèches (tiges entières et cassées) dans les 4 quadrats de 50 X 50 cm
Nombre de panicules	Nombre total de tiges sèches entières, c'est-à-dire avec une panicule de l'année précédente, dans les 4 quadrats de 50 X 50 cm