Méthodes et outils de traitement des données en sciences sociales. Retours d'expériences

Eric Cahuzac, Marie Huyez-Levrat

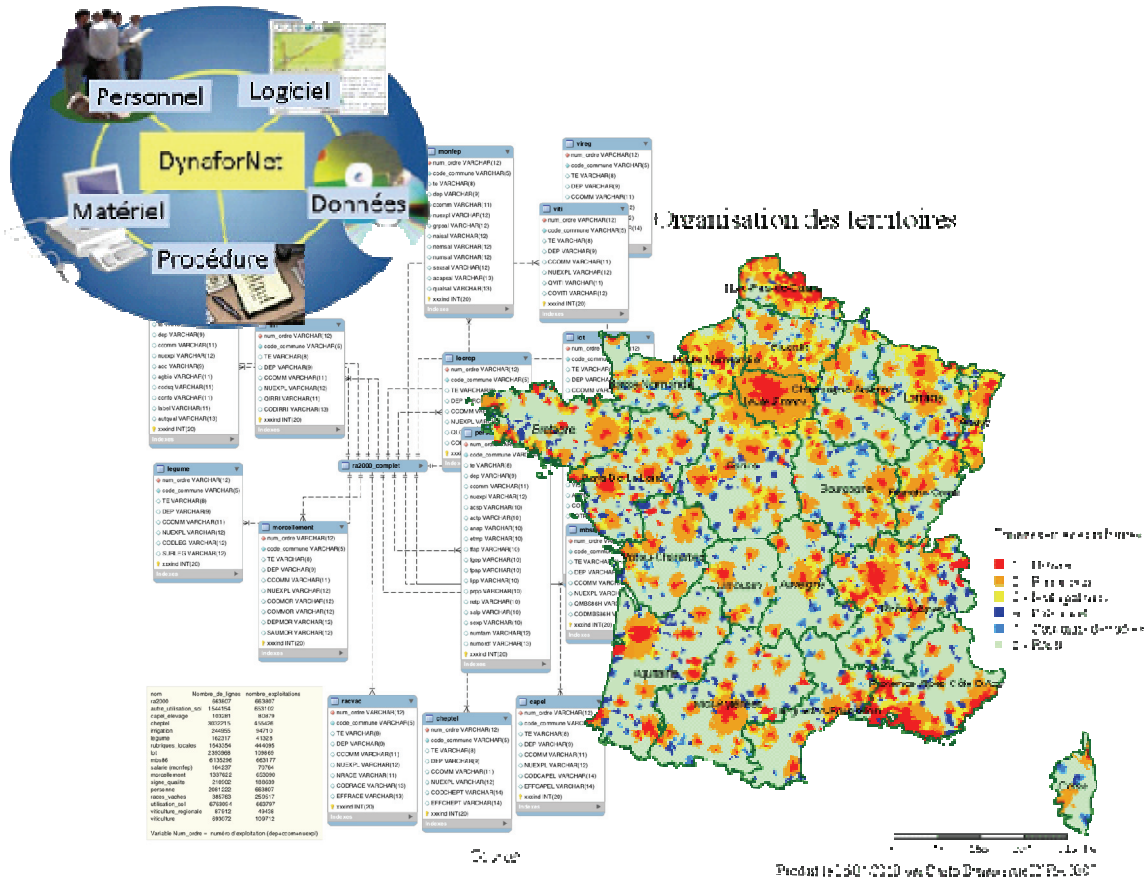
Le Cahier des Techniques de l'Inra

Méthodes et Outils

de traitement des données

en sciences sociales

Retours d'expériences



Délégation au Partenariat avec les Entreprises
Institut national de la recherche agronomique - 147 rue de l'Université - F-75338 Paris Cedex 07

Méthodes et outils de traitement des données en sciences sociales

Retours d'expériences

Numéro spécial 2010

Collectif[1]

Animateur : *Éric Cahuzac*[2]
Coordonnatrice : *Marie Huyez-Levrat*[3]

Sommaire

Cet ouvrage est téléchargeable à partir du site de l'Inra

https://intranet.inra.fr/cahier_des_techniques

[1] *Le Cahier des Techniques de l'Inra* est une publication destinée à échanger des méthodes et des savoir-faire ; ce n'est pas une revue à comité de lecture. Les auteurs sont seuls responsables de leurs textes.

[2] US0685 US ODR - Observatoire des programmes communautaires de développement rural – F-31326 Auzeville ☏ 05 61 28 53 41 ✉ EricCahuzac@toulouse.inra.fr

[3] UAR1185- Délégation au partenariat avec les entreprises DV/DPE - INRA - F-75338 Paris - ☏ 01 42 75 91 65 ✉ marie.huyez@paris.inra.fr

Les auteurs

Nom	**Prénom**	**page**
Gendre	Cédric	7
Harel	Monique	73
Heintz	Wilfried	23
Hilal	Mohamed	41
Hofstetter	Annie	65
Ladet	Sylvie	23
Le Roy	Cécile	73
Rasplus	Valéry	23
Rousselle	Jean-Marc	85

Introduction

> *« Les sciences de l'homme et de la société sont avant tout des sciences de l'observation et l'expérimentation au sens strict n'est que rarement possible pour elles. »* (Silberman, 1999)[1].

L'accès aux bases de données en sciences humaines et sociales est crucial. Il permet aux chercheurs de tester les hypothèses des modèles théoriques et d'être en mesure de répondre de façon empirique aux questions que se pose la société.

Longtemps ces données ont été difficiles à collecter et demandaient la réalisation d'enquêtes parfois coûteuses, réalisées sur de petits échantillons. Les progrès de l'informatique et de la statistique ont peu à peu amélioré la collecte de données – effectuée principalement par les services de l'État ou les instituts de sondages et les chercheurs – et leur traitement pour offrir actuellement une information conséquente et de meilleure qualité dans des domaines très variés. Dans le même temps, la représentativité s'est accrue donnant accès à des niveaux d'analyses géographique et temporelle plus fins, faisant ainsi la part belle aux analyses micro économiques, aux modélisations longitudinales et spatialisées. Enfin, l'accès aux bases de données administratives s'est peu à peu assoupli sous l'impulsion des pouvoirs publics désireux d'exploiter cette grande quantité d'information à des fins d'expertise et de pilotage économique. Dernière en date, la directive INSPIRE[2] qui vise à favoriser la production et l'échange des données dans le domaine de l'environnement au sein de l'Union européenne permet de mettre à disposition de la recherche et de mutualiser dans un grand nombre de services de l'État un ensemble de données spatialisées très important.

Après avoir comblé en partie son retard – par rapport aux principaux pays producteurs de données – dans le domaine de la collecte, la France est en train d'améliorer notablement son processus de mise à disposition des données. Tout en finalisant cette dernière étape, c'est un nouveau défi qu'elle devra relever dans un futur proche, celui de la valorisation de ce gisement extraordinaire de micros données. Sans soigner la réalisation de cette dernière étape, le processus restera au stade de la construction d'entrepôts de données qui seront critiqués pour leur coût de mise en œuvre et de ce fait abandonnés. Dans le contexte actuel de rationalisation des moyens, cette étape de valorisation ne pourra s'appuyer comme les précédentes sur un seul acteur, les services de l'État gestionnaires des données. Elle doit en faire intervenir de nouveaux, et la recherche est ici le partenaire historique à privilégier pour répondre à la demande d'expertise sociale et valoriser scientifiquement ce gisement de connaissance que renferment les données.

[1] « Les sciences sociales et leurs données », Rapport à l'attention de Monsieur le ministre de l'Éducation nationale, de la recherche et de la technologie.
[2] Infrastructure for Spatial Information in the European Community. Directive 2007/2/CE du 14 mars 2007.

Nous nous situons donc actuellement au milieu d'un processus qui s'inscrit dans la durée. Il va nous amener à développer de nouvelles alliances entre acteurs publics, de nouvelles stratégies de recherches et de nouveaux outils d'acquisition et de diffusion de connaissances. Ces mutations posent bien évidemment de nouvelles questions, de nature juridique (propriété des données et des résultats, confidentialité, secret statistique, droit du citoyen,…) ou techniques (archivage et sauvegarde,…). Ces nouveaux questionnements qui sont souvent liés à la crainte d'une entrave à la vie privée du citoyen, doivent être abordés sereinement afin qu'ils ne représentent pas des freins au développement et à la diffusion de la connaissance.

En tant qu'acteur de la recherche agronomique, l'Inra a mis au centre de ces réflexions la question de la place des données dans ses recherches[3]. Dans les départements et les unités, les chercheurs ont mis en place des processus de gestion et d'archivage de ce patrimoine scientifique. Fait nouveau, des plateformes spécialisées dans le traitement et la valorisation des données voient peu à peu le jour et acquièrent souvent une renommée nationale, voire internationale.

Ce numéro spécial du *Cahier des Techniques de l'Inra* se propose de mettre en lumière différentes contributions collectives, témoignant de l'intérêt que portent les unités dans le traitement et la valorisation des données qu'elles gèrent. Il est dédié à des développements techniques s'appuyant sur des données en sciences humaines et sociales, plus précisément dans les disciplines de l'économie de la sociologie, de la géographie, jusqu'à l'écologie du paysage. Les domaines d'application seront variés mais ils répondent tous à la même logique : mettre à la disposition du chercheur, du citoyen ou d'autres institutions des outils permettant de structurer, de gérer et de valoriser de volumineuses bases de données. Enfin, la composante spatiale tient désormais une place importante dans la description des phénomènes et la plupart de ces réalisations y feront référence.

Nous présenterons tout d'abord deux exemples de réalisation de plateforme de traitements de données géo-localisées. La première, centrée sur la thématique du développement rural, est née d'un besoin non couvert de suivi et d'évaluation en France des politiques européennes liées à cette thématique. Partant de cet objectif, nous verrons comment l'Observatoire du développement rural (ODR) est devenu un outil coopératif mutualisant de nombreuses bases de données, au service de la recherche en sciences sociales et couvrant plus largement les questions de politiques agricoles, de développement rural et d'agro environnement. La seconde plateforme mutualise quant à elle des données acquises sur le terrain afin de traiter des questions relatives à la biodiversité. Nous verrons comment pour la plateforme LTSER[4] l'apport d'outils tels que les SIG[5] couplés avec des outils SGBD[6] a permis de mettre en réseau et de mutualiser des données de type écologiques, socio-techniques et environnementales.

La composante géographique des données est à nouveau mise à l'honneur dans la troisième et quatrième contribution de cette revue. Avec ODOMATRIX, on découvrira comment à partir de bases de données géographiques de l'IGN et de couches d'informations géographiques, la modélisation mathématique et la prise en compte des contraintes socio économiques permet de réaliser un distancier performant capable de fournir l'itinéraire le plus court en kilométrage ou en temps – pour les heures creuses et les heures de pointes – pour atteindre des pôles ou

[3] « Analyse de l'inventaire des bases de données scientifiques », rapport de mission, C. Christophe (2010).
[4] Long term sociological and ecological research
[5] Systèmes d'Information Géographique.
[6] Système de Gestion de Bases de Données.

des communes équipées en commerces et en services. La quatrième réalisation décrit d'un point de vue technique un phénomène économique complexe à modéliser qui est celui de la diversité des acteurs dans l'analyse de l'impact d'une politique. Nous nous placerons dans le cadre précis de l'impact des politiques publiques sur la dynamique des paysages. En alliant à un système multi-agent les capacités d'un SIG, nous appréhenderons la complexité du couplage de bases de données à la fois géographiques et hiérarchisées, sur une entité spatiale très fine : la parcelle.

Enfin, les deux dernières contributions illustrent dans deux domaines différents – celui de la comptabilité des exploitations et celui des échanges internationaux des industries agroalimentaires – la mise à disposition pour la recherche d'interfaces Web dédiées à l'interrogation et à la description de bases de données volumineuses. En effet, l'environnement Web est devenu aujourd'hui un environnement accessible à tous, alors que les logiciels d'archivage de sauvegarde et de traitements statistiques de données ont tous un environnement spécifique qu'il est difficile de maîtriser. La mise en forme des données dans un SGBD et la réalisation d'une interface développée avec des outils orientés pour le Web (PHP, MySQL) offre à tous des moyens d'interrogation et de traitement simples sur les bases de données sans pour autant posséder les compétences informatiques ou statistiques requises. Cette méthode favorise aussi une réelle politique de gestion des données dans une équipe, à savoir, la sauvegarde, la mise à jour et la documentation des bases de données, en implantant ces bases sur un serveur dédié.

Les articles que vous pourrez découvrir dans ce numéro spécial ont pour la plupart fait l'objet d'une présentation dans un atelier technique lors des deuxièmes journées de recherches en sciences sociales INRA - SFER[7]- CIRAD, en décembre 2008. Ils ont été écrits par des ingénieurs des départements SAE2[8] et SAD[9] qui contribuent dans leur équipe respective à l'amélioration de l'accessibilité aux bases de données pour la recherche.

Eric Cahuzac[10]
Chargé de Mission « Bases de Données »
Pour le département SAE2

[7] Société française d'économie rurale
[8] Sciences sociales agriculture et alimentation, espace et environnement
[9] Sciences pour l'action et le développement

[10] US0685 US ODR - Observatoire des programmes communautaires de développement rural – F-31326 Auzeville
☏ 05 61 28 53 41 ✉ EricCahuzac@toulouse.inra.fr

Création d'un serveur de données

l'Observatoire du développement rural

Cédric Gendre[1]

Résumé : *Cet article présente l'organisation et les fonctionnalités de l'Observatoire du développement rural (ODR) au travers de son application web Carto-dynamique développée avec les outils et langages open source.*

Mots clés : sciences sociales, développement rural, environnement, serveur de données, informatique, cartographie et traitements statistiques.

Introduction

L'unité de service Observatoire du développement rural du département SAE2[2] a développé une application web intitulée **Carto-dynamique** avec les outils issus de l'open source PHP-MySQL. Cette application a pour vocation de rapprocher des données de sources variées centralisées sur un serveur, autour des thématiques de développement rural et de faciliter la production de résultats statistiques cartographiés. Carto-dynamique est intégré aujourd'hui dans une plateforme logicielle plus large dénommée *Observatoire des programmes communautaires de développement rural.*

Ce document est divisé en trois parties ; nous présenterons l'observatoire et ses fonctionnalités, puis l'outil Carto-dynamique développé pour répondre aux objectifs de l'observatoire et, en dernière partie, les trois grands types de données traitées dans l'ODR.

Page d'accueil de l'Observatoire du développement rural

[1], US0685 ODR -observatoire des programmes communautaires de développement rural – INRA -F-31326 Castanet-Tolosan ☎.05 61 28 53 48 ✉ cedric.gendre@univ-tlse2.fr

[2] Département des sciences sociales, agriculture et alimentation, économie, environnement

1. Observatoire des programmes communautaires de développement rural

L'observatoire des programmes communautaires de développement rural (ODR) est un système d'information collaboratif géré en partenariat. Il a eu pour objectif initial de contribuer à la préparation et au suivi en France des politiques européennes de développement rural, plus particulièrement des mesures financées par les règlements de développement rural (RDR) de 1999 (programmation 2000-2006) et de 2005 (programmation 2007-2013). C'est aujourd'hui un outil coopératif pour la recherche en sciences sociales sur les politiques agricoles, le développement rural et l'agro-environnement.

L'observatoire est régi par une convention cadre et un comité de pilotage. Les partenaires fondateurs sont l'Agence de services et de paiement (ASP, fusion du CNASEA et de l'AUP), l'Institut national de la recherche agronomique (Inra) et le ministère de l'agriculture, alimentation et pêche (MAAP). Il peut accueillir de nouveaux partenaires comme récemment la Caisse centrale de la mutualité sociale agricole (MSA) et des « tiers agréés », le Centre national de la recherche scientifique (CNRS) et le Cemagref susceptibles de fournir ou d'utiliser des données. Il est administré et développé par un chef de projet, Gilles Allaire, et une équipe opérationnelle, l'unité de service ODR, situé au centre Inra de Toulouse-Auzeville[3].

Les partenaires officiels de l'ODR

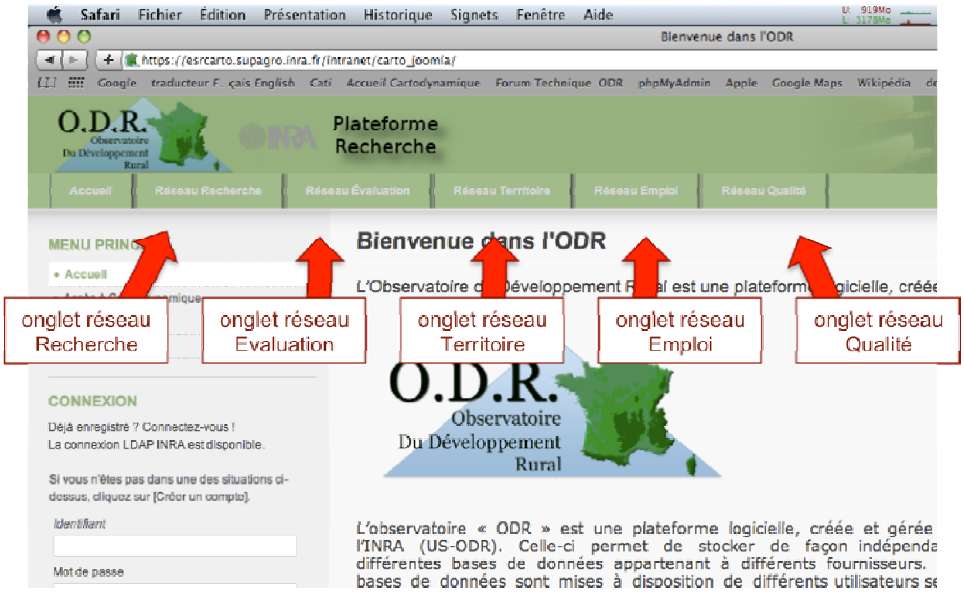
1.1 Un ensemble de réseaux gérés en partenariat

Outre les réseaux « évaluation et recherche » répondant aux objectifs initiaux de l'ODR, il héberge aussi un observatoire de l'emploi agricole et rural, géré en partenariat avec la MSA et un projet d'Observatoire des signes officiels de qualité en collaboration avec l'Institut national de l'origine et de la qualité (Inao). L'observatoire rassemble aujourd'hui un ensemble de réseaux sur son site web consultables en ligne :

- **Le réseau recherche** présente des projets menés par des équipes de recherches (Inra, Cemagref, CNRS) utilisant l'observatoire du développement rural et sa plateforme logicielle.

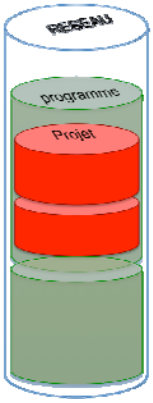
[3] lien web vers l'observatoire : https://esrcarto.supagro.inra.fr/

- **Le réseau évaluation** présente les résultats de l'évaluation du second pilier de la politique agricole commune (PAC) conduite avec l'assistance technique de l'équipe ODR.
- **Le réseau territoire** ouvre un portail territoire co-animé par le réseau des observatoires des activités agricoles dans les territoires (OAAT)
- **Le réseau emploi** en collaboration entre l'Inra et la MSA, via une convention cadre, ouvre sur la création de l'observatoire de l'emploi agricole et rural
- **Le réseau qualité** donne accès aux pages de l'Observatoire des impacts territoriaux et agro-environnementaux des signes officiels de qualité



Les réseaux accessibles dans l'ODR

Ces réseaux sont constitués de **programmes**, c'est-à-dire d'espaces de travail et de publications contractualisés avec un ou plusieurs partenaires. À l'intérieur de ces programmes, des utilisateurs particuliers, rédacteurs ou administrateurs de programmes, donnent accès à de la documentation consultable en ligne par les autres membres du réseau, ainsi qu'à des articles informatifs sur les thématiques de recherche développées[4].



Les membres du réseau peuvent aussi, suivant leurs droits, utiliser la plateforme logicielle pour créer, stocker et rendre disponible à la consultation, des traitements statistiques et cartographiques.

L'utilisation de la plateforme logicielle pour la production de résultats statistiques cartographiés passe par la création de projets.

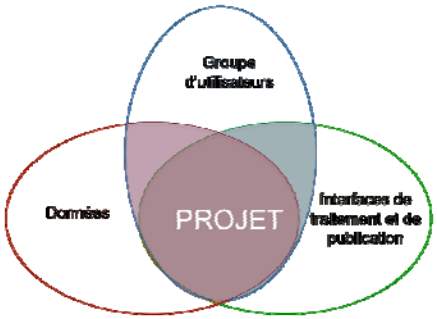
Un **projet** au sens de l'observatoire rassemble :

– un **groupe d'utilisateurs** dont nous verrons qu'il en existe de plusieurs types, autour d'une thématique de recherche ou d'étude, avec à la tête un responsable de projet ;

– des **données administratives ou statistiques, géographiques** déposées dans des zones de stockage de l'ODR et accessibles aux membres du projet ;

[4] Le gestionnaire de contenu utilisé pour gérer dynamiquement les pages des réseaux de l'observatoire est le CMS Joomla distribué sous licence GNU/GPL (gratuit).

– des **interfaces de traitement et de publication** sur internet utilisant des logiciels Open Source (Apache, PHP, MySQL, JOOMLA, MAPSERVEUR, POSTGIS/GRES, R) utilisables via l'ODR qui en garantissent la pérennité.

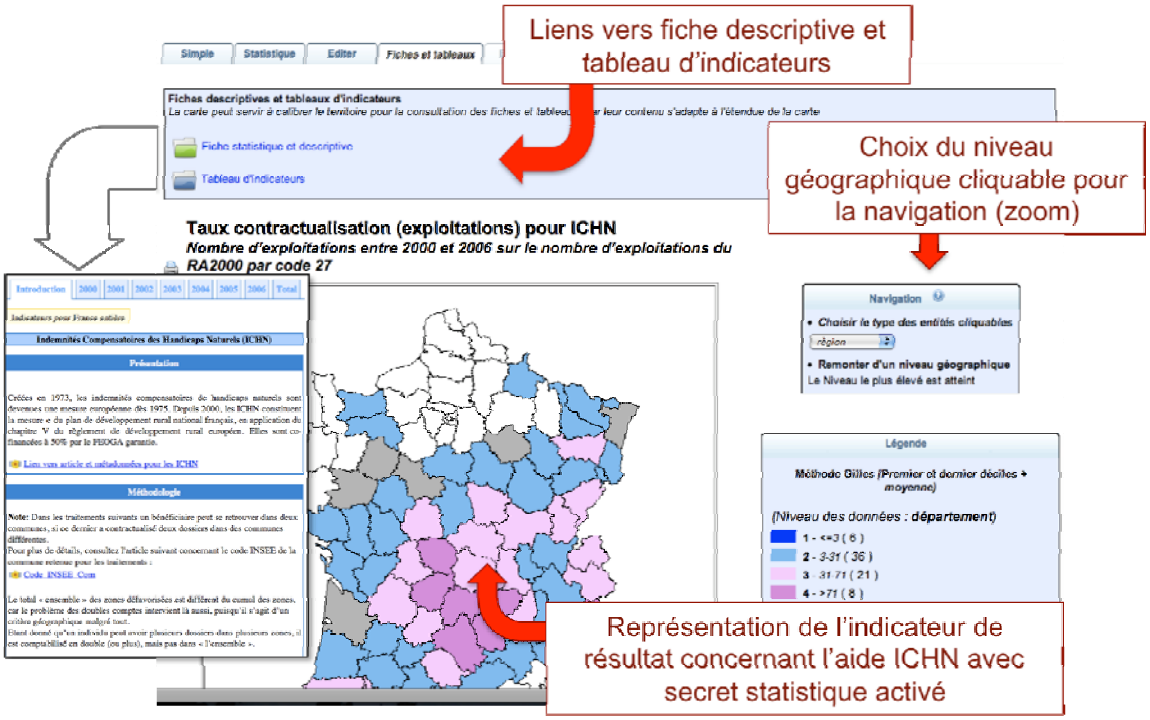


Éléments constitutifs d'un projet

Ces projets sont créés dans l'application **Carto-dynamique** outil central du fonctionnement de l'observatoire.
Les utilisateurs de l'observatoire ont la possibilité de publier certains résultats obtenus dans des **thèmes** du projet regroupés en **dossiers thématiques.**
Un **thème** est un ensemble de traitements cartographiés préprogrammés présentant les résultats d'un projet. Il peut contenir, des cartes dynamiques (agrégeables à différentes échelles spatiales et pour différents territoires), un ou plusieurs tableaux statistiques dynamiques, un document explicatif du thème.

La page répertoire des dossiers thématiques

Ces résultats sont visibles soit dans le projet concerné soit rassemblés dans un programme ce qui les rend accessibles en consultation en dehors du projet initial (cf. écran de consultation des dossiers thématiques ci-dessus).

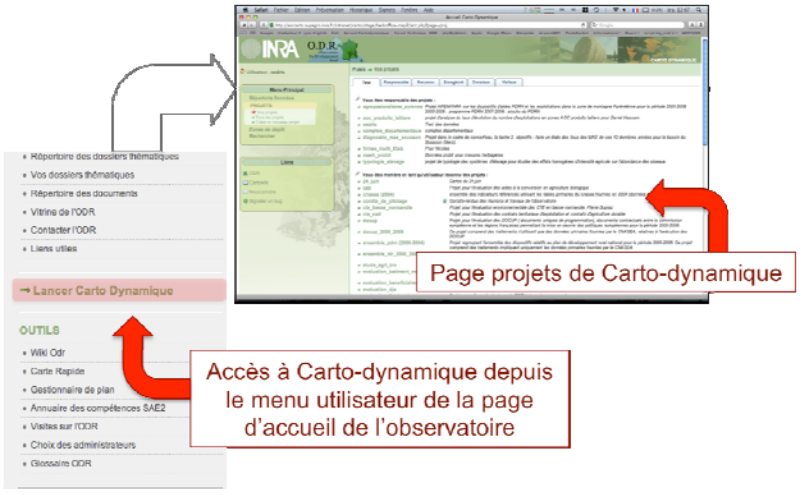


Exemple de thème publié par les membres de l'ODR

2. Principes de fonctionnement de Carto-dynamique

La plateforme logicielle de traitement et publication Carto-dynamique est principalement constituée d'un serveur de données utilisant des applications en PHP-MySQL, complété de logiciels open source permettant diverses actions sur les données : POSTGRES-GIS et MAPSERVER pour le traitement de données géolocalisées, R pour des traitements de statistiques descriptives ou de modélisation. Le serveur de données permet :

- le stockage de données repérées par un codage géographique ;
- la production de résultats (cartes et tableaux de synthèse) à partir des données du serveur ;
- la conservation de certains résultats en vue de leur consultation dynamique ultérieure.



Lancer Carto-dynamique

2.1 Carto-dynamique : serveur de données pour un système d'information collaboratif

L'application permet une centralisation du stockage des données .grâce à une gestion des utilisateurs en différents groupes avec des droits distincts d'accès au serveur et aux variables stockées.

2.1.a Droits d'accès des utilisateurs : visiteur ou titulaire

Tout membre de l'observatoire est utilisateur au sens qu'il a besoin d'être authentifié par un login/mot de passe pour entrer dans l'application. Les chercheurs Inra peuvent accéder à l'observatoire en utilisant leur login/mot de passe LDAP, tandis que les autres utilisateurs doivent être enregistrés préalablement dans le système.

Page d'accueil de l'observatoire

Les utilisateurs sont divisés en deux grands groupes, les utilisateurs visiteurs et les utilisateurs titulaires.

– Les utilisateurs visiteurs ont la possibilité de consulter et de télécharger, sous condition, des résultats publiés dans des programmes et des projets selon leur affiliation à des réseaux. Différentes interfaces sont dédiées à la consultation des résultats sous forme de dossiers thématiques.
Lors d'une première connexion à l'observatoire, l'utilisateur est automatiquement défini comme visiteur. Le statut de visiteur permet seulement la consultation de deux projets génériques (vitrine, zonages). Par la suite, il doit demander une habilitation au responsable d'un autre projet pour en consulter les publications.

– Les utilisateurs titulaires peuvent créer des projets ou y participer, déposer et gérer des données, produire, stocker et publier des traitements statistiques et cartographiques. Un

membre titulaire peut donc à la fois être dépositaire de données et/ou développer un projet soit comme responsable soit comme participant. Il existe un statut particulier et important de titulaires qui sont simplement fournisseurs de données dans un projet et ne produisent pas de traitement. Ils sont alors définis comme « donateurs » pour le projet concerné.

Visiteur	simple consultation des résultats publiés
Titulaire	utilisateur des données existantes, rapprochement avec ses propres données (importation sur serveur)

Gestion des membres et statuts dans un projet

2.1.b Droits d'accès aux données : réservés ou publiques

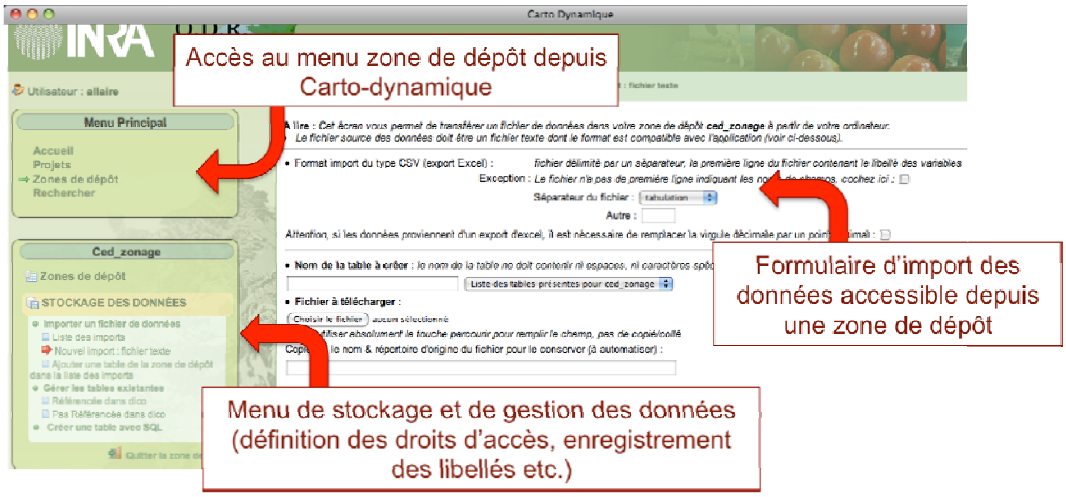
Tout fournisseur de données a une gestion fine des droits d'utilisation de ses données dans l'observatoire. Les dépositaires peuvent ainsi définir principalement deux types de droit d'accès aux données : réservé ou publique.

– Les données dites **publiques** sont déposées dans l'ODR et sont accessibles par tout membre enregistré et titulaire sans demande d'autorisation pour leur utilisation.

– Les données **réservées** sont accessibles aux utilisateurs titulaires selon des permissions attribuées par les dépositaires. Ces données peuvent être accessibles dans leur intégralité pour un projet précis ou soumises au secret statistique et à des restrictions selon les cas.

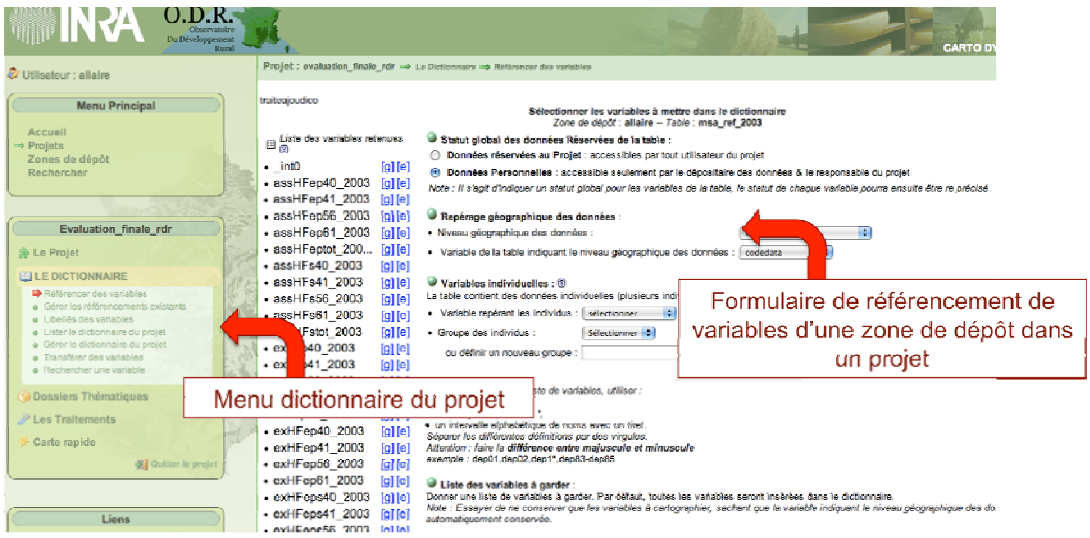
Données publiques	accessibles à tout utilisateur sans restriction.
Données Privées	possibilité d'ouverture de droits maîtrisés à d'autres utilisateurs par le dépositaire de la donnée

La gestion des droits sur les données autorise la publication de résultats suffisamment agrégés à un niveau géographique, même si les données brutes traitées sont à des niveaux très fins (par exemple : données individuelles comme unité statistique).
Dans le même ordre d'idée, l'utilisateur titulaire dispose d'un ensemble d'outils pour vérifier les étapes intermédiaires du calcul nécessaire à l'obtention des résultats sur le serveur sans avoir un accès direct aux tables de données brutes.
L'importation des données brutes est possible depuis le poste client dans une zone de stockage du serveur dite « **zone de dépôt** » appartenant au dépositaire.



Zone de dépôt

Pour être disponible dans un projet, les variables réservées doivent être référencées par le propriétaire des données dans le **dictionnaire** du projet. Elles deviennent, sous condition, accessibles pour des traitements aux membres titulaires du projet concerné.
☞Rappel : Les données publiques n'ont pas besoin de se référencement.

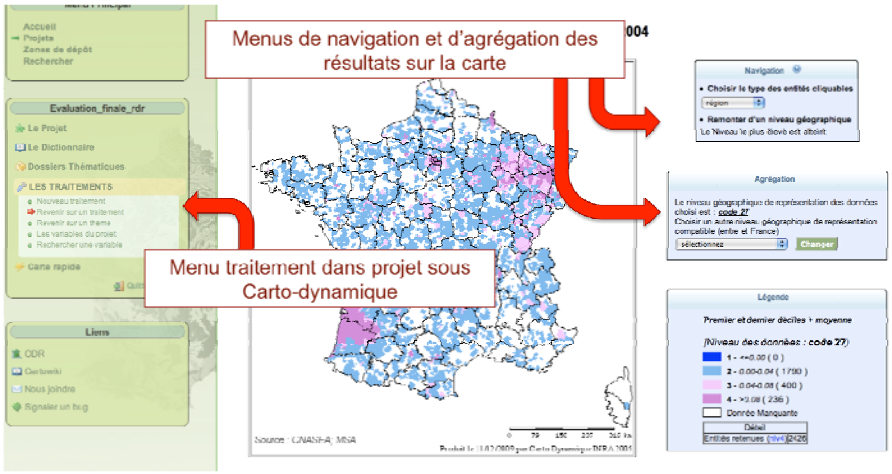


Formulaire de référencement du dictionnaire du projet

Les droits d'utilisation des données sont gérés par le dépositaire des données dans ces deux interfaces web de Carto-dynamique : le formulaire des zones de dépôt et le formulaire du dictionnaire de projet.

2.2 Connexion des données à différentes échelles spatiales

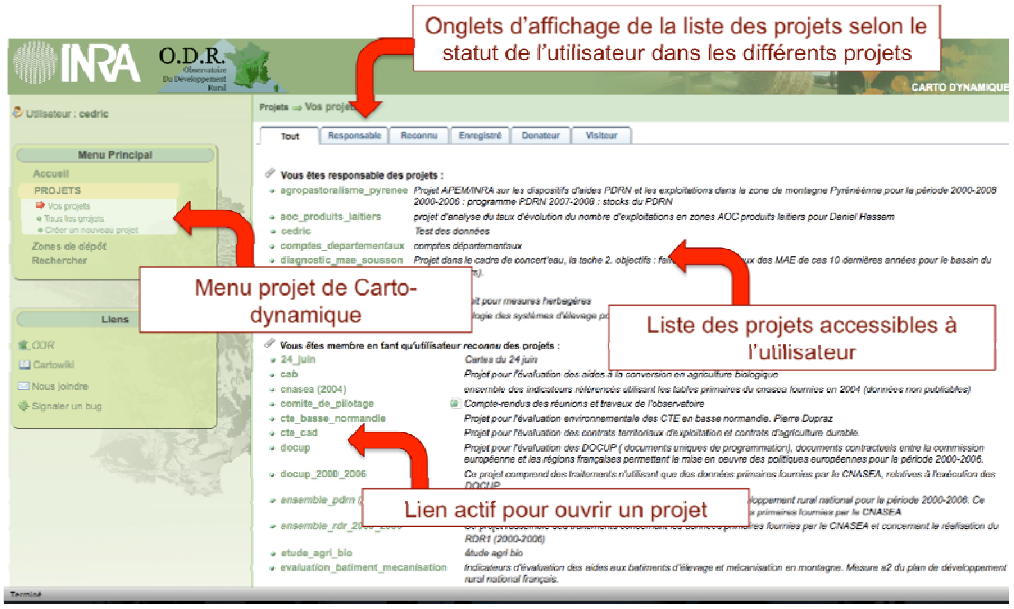
L'objectif de Carto-dynamique est de faciliter le traitement d'informations provenant de sources variées. Ceci est rendu possible par la mise à jour d'une table de correspondance entre tous les codages géographiques disponibles sur le serveur permettant leur rapprochement.
On désignera par « **géocodes** » ce codage géographique. Ce sont les attributs des polygones des fonds cartographiques enregistrés sur le serveur à un moment donné. Les géocodes de base sont ceux des communes présentes au recensement général de la population (RGP 1999).
Le serveur de données repose sur un système de gestion de base de données (SGBD). Le cœur de l'application est une métabase référençant à la fois les données brutes transférées sur le serveur, les utilisateurs et leurs droits, et les fonds de cartes disponibles. Via la métabase, le système peut alors rapprocher toute donnée stockée en tenant compte de la hiérarchie entre géocodes compatibles (exemple agrégation des données communales en données cantonales).
Il ne s'agit pas ici d'un système d'information géographique (SIG) bien que pouvant produire des résultats cartographiés dynamiques. On peut cependant enrichir la cartographie statistique de couches d'habillages Raster (image) pour en améliorer la lecture.
En pratique, cette spécificité permet, lors d'un traitement ou d'une consultation de résultat dans Carto-Dynamique, de changer facilement d'échelle géographique et aussi d'effectuer des traitements de manière automatique qui mélangent des données de différentes échelles géographiques par agrégation - par exemple produire une carte au niveau départemental en utilisant des données communales et cantonales.



Les menus de navigation dans l'interface de traitement

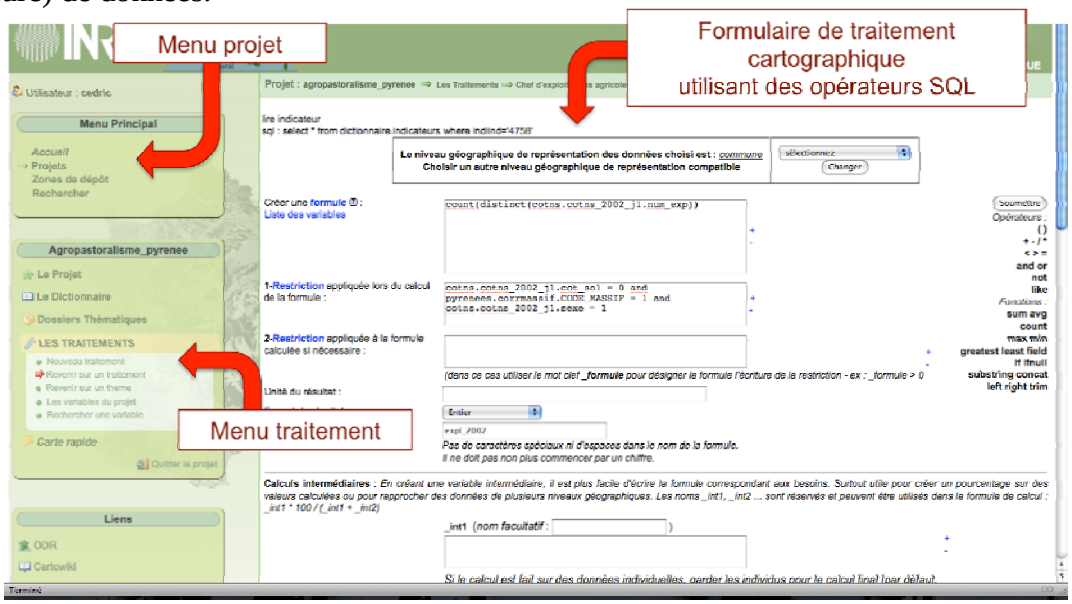
2.3 Mise en œuvre pratique du serveur de données.

En plus de rapprocher toutes données actuelles ou à venir stockées sur le serveur, le serveur de données offre la possibilité à tout utilisateur titulaire de mémoriser son travail en le sauvegardant dans ses projets. Il peut ainsi revenir facilement et rapidement pour consulter à nouveau ou modifier un traitement existant.



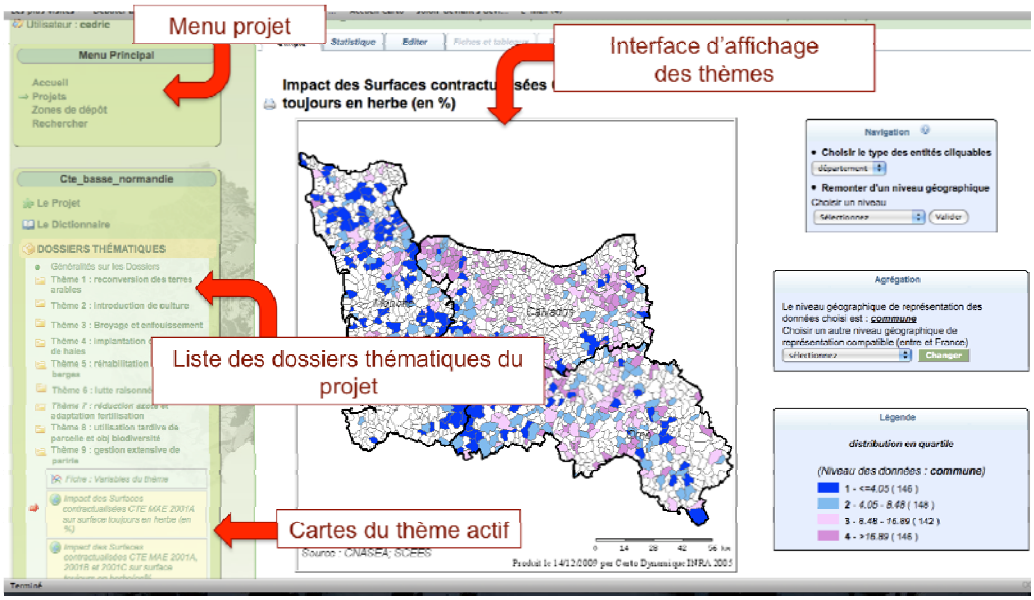
Menu des projets mémorisés dans Carto-dynamique

Les traitements sont réalisés dans une interface où il est possible de s'abstraire en partie des difficultés de mise en œuvre du langage SQL, langage de requête sur les tables du serveur et aussi, d'assurer un contrôle des données manquantes, incomplètes et des appariements (jointure) de données.



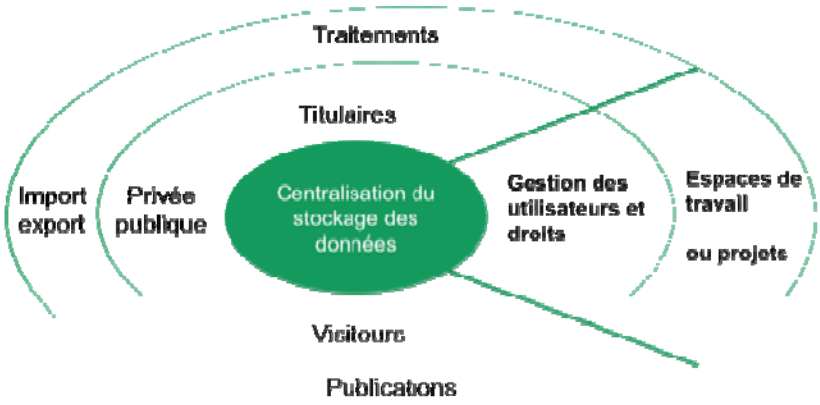
Menu « traitement » de Carto-dynamique accessible aux membres titulaires du projet

Une fois un traitement obtenu et validé, il peut être mémorisé dans le projet et même publié en ligne à travers l'interface du serveur sous forme de thème.



Menu « dossiers thématiques » accessible aux titulaires et visiteurs du projet

Selon les cas, et selon les droits de l'utilisateur, l'interface permet de faire l'export de résultat alliant les données agrégées (indicateurs) et leur géocode, en vue de traitements externes (Excel, logiciels statistiques).
En conclusion, le schéma suivant résume les fonctionnalités d'un serveur de données tel que Carto-dynamique développé pour l'observatoire.



Le cœur du système est la possibilité de centraliser et de rapprocher des données géocodées de différentes sources (par l'implémentation d'une métabase). Pour cela, des droits sont appliqués aux utilisateurs et sur les données et des services sont proposés aux utilisateurs pour traiter, publier et visualiser leurs résultats.
Pour répondre à ses objectifs, l'ODR doit aussi pouvoir proposer un ensemble d'informations lié aux thématiques de développement rural, c'est ce que nous proposons de décrire en dernière partie de cet article.

3. Les trois grands types de données disponibles dans l'ODR

L'ODR contient des données très diversifiées et des sources variées se rapportant aux thématiques du développement rural et à l'agro-environnement. Les informations gérées par l'ODR sont de trois types différents. **:** les fonds de carte, les données géocodées**,** les couches d'habillages.

3.1 Les fonds de carte

Les données cartographiques ou **fonds de carte** fournissent les contours de différents découpages géographiques, communes et dérivés, bientôt bassins versants et zonages environnementaux ou agricoles. Ces fonds de carte sont le support de la représentation cartographique des résultats statistiques. Leur interdépendance est maintenue dans une table de la métabase.

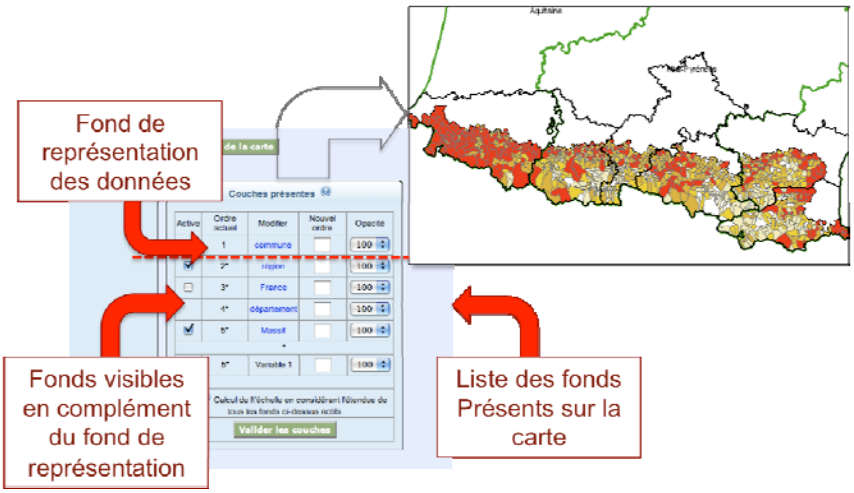


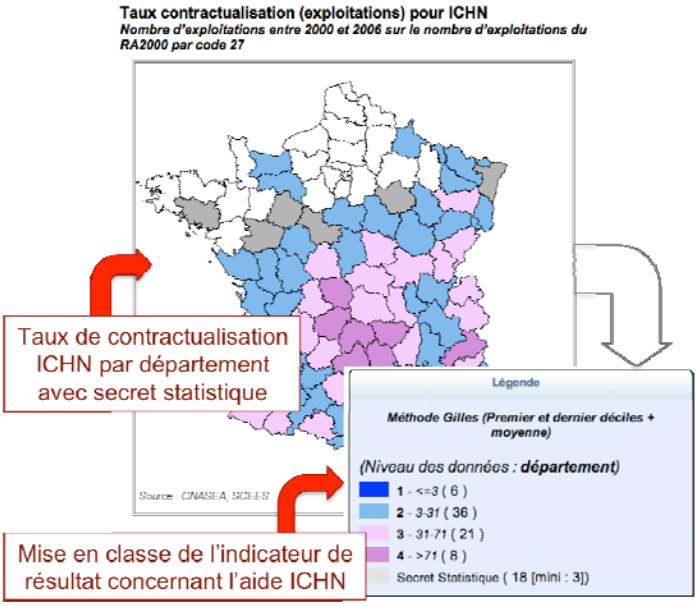
Tableau des fonds de carte actifs pour une carte

Les fonds de carte sont gérés par l'équipe d'administration. Toutefois il est possible d'ajouter sur demande d'un utilisateur titulaire un nouveau fond de carte créé facilement à partir d'une simple table de correspondance entre un fond existant et la cible (exemple création du fond des petites régions agricoles (PRA) à partir de la liste des communes qui les décrivent).

3.2 Les données géocodées

Les données **géocodées** sont le plus souvent administratives ou statistiques. Ces données doivent obligatoirement posséder un attribut géographique connu du serveur (en lien avec un fond de carte). Pour tout nouvel import de données, une vérification des géocodes est obligatoire, un outil de correction est disponible pour aligner les données géographiques importées sur les données géographiques de référence du serveur (le codage d'une commune peut varier dans le temps)
Ces données, possédant un attribut géographique, permettent la production de résultats, le plus souvent des statistiques descriptives ordinaires sous forme d'indicateurs, et leurs représentations cartographiques. Ces données peuvent être strictement géocodées (une seule valeur par géocode) ou individuelles (plusieurs valeurs pour un même géocode). Dans le cas des données individuelles, un identifiant est nécessaire pour différencier les groupes

d'individus (individus MSA, individus RDR etc.) et il doit être renseigné dans la métabase de Carto-dynamique afin de piloter le rapprochement correct de ces données lors du traitement.

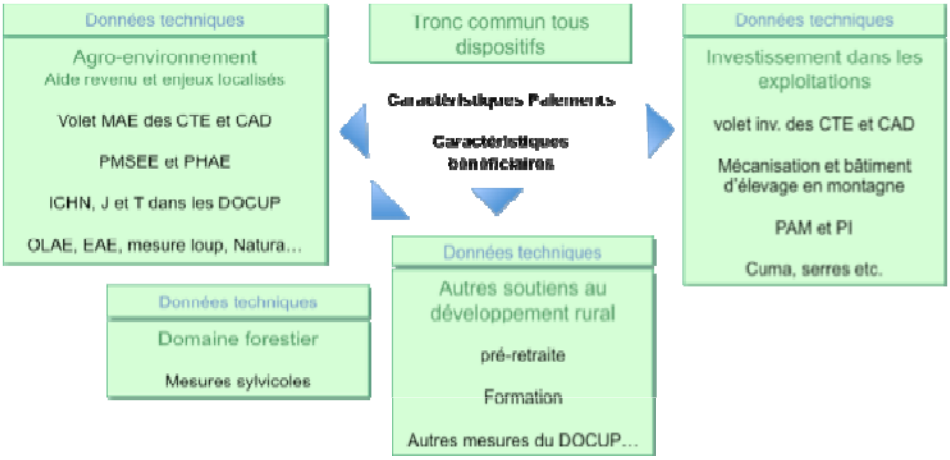


Exemple de traitement de données géocodées

3.2.1 Données géocodées gérées par les administrateurs

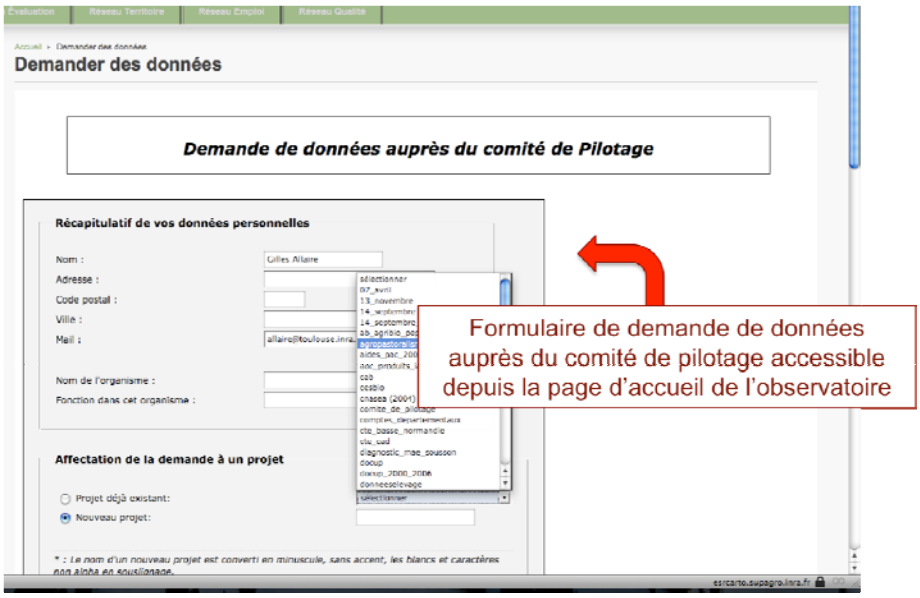
Majoritairement les données présentes sur le serveur concernent la réalisation des mesures européennes du règlement de développement rural en France, les indications géographiques (AOC, IGP), le nombre et le revenu des exploitants, les structures des exploitations agricoles, l'utilisation du sol etc., ainsi que des données de référence, de contexte ou d'impact dans différents domaines (démographie, revenus et emplois, environnement et systèmes de production)
La gestion des données se rapportant à l'évaluation du second pilier[5] de la PAC est déléguée à l'équipe d'administration par les partenaires de l'ODR.
Les données fournies par l'ASP concernent plus particulièrement l'ensemble des bénéficiaires des aides du Règlement de développement rural pour l'ancien programme (2000-2006) et le programme actuel (2007-2013). Localisées à la commune et agrégeables à d'autres niveaux géographiques, ces données regroupent par type d'aide des informations sur les caractéristiques des bénéficiaires, leur localisation, les caractéristiques des actions engagées et les paiements reçus.

[5] Le premier pilier de la PAC concerne le soutien des marchés agricoles, le second pilier de la PAC concerne le développement rural au sens large.



Exemple des dispositifs du RDR1 (2000-2006) présents dans l'ODR

Le statut de ces données n'est pas public, les utilisateurs titulaires doivent déposer une demande auprès du **comité de pilotage** de l'ODR pour obtenir le droit d'utiliser ces données dans un projet Carto-dynamique.



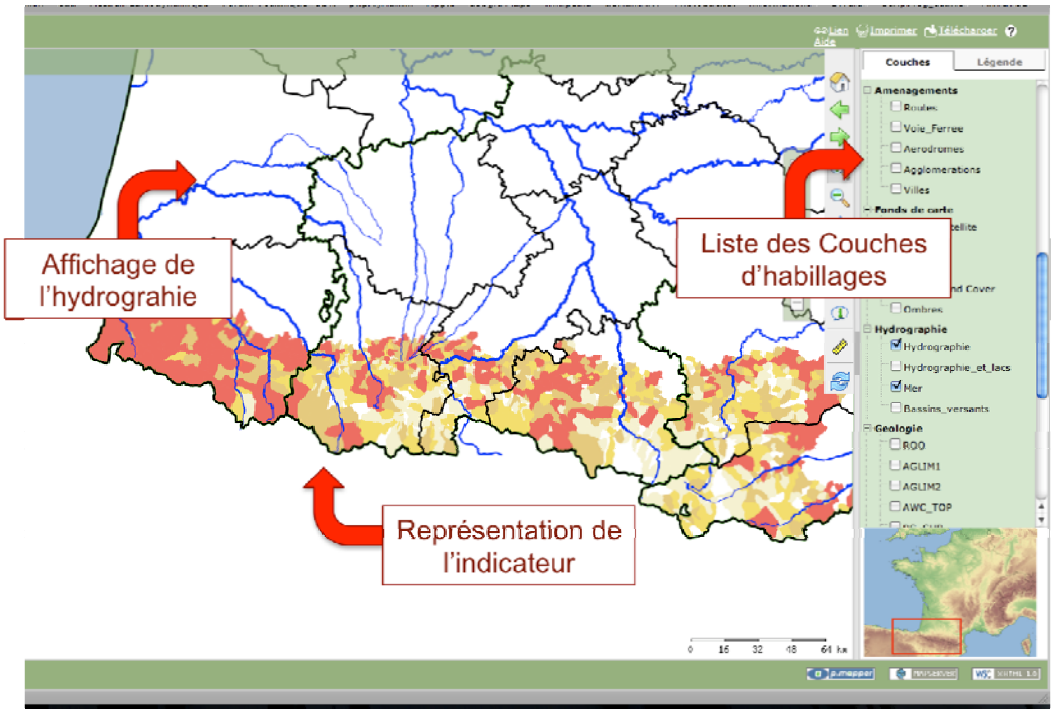
Formulaire de demande de données réservées

3.2.2 Données géocodées gérées par les utilisateurs titulaires

Un des intérêts du serveur est d'offrir la possibilité à tout titulaire de rapprocher ses données personnelles des données déjà présentes sur le serveur auxquelles il a droit. L'utilisateur titulaire peut ainsi transférer ses données (export Excel par exemple ou fichier texte) dans sa zone de dépôt et les référencer dans un projet pour de futurs traitements.
Même sans avoir à déposer et à référencer ses données, tout utilisateur peut avec la fonction « carte rapide » qui permet par copier/coller d'obtenir rapidement un résultat cartographié d'un lot de données ayant un géocode reconnu par le serveur.

3.3 Les couches d'habillages

Dernier grand type de données, les données géographiques (raster et vecteur) ou **couches d'habillages** permettent d'ajouter une information complémentaire visuelle à un traitement statistique en superposant son contenu à la carte obtenue (exemple : localisation des cours d'eau, dénivellation du terrain). Ces données n'interagissent pas sur le résultat statistique cartographié.



Les couches d'habillages

Conclusions

Les difficultés à rassembler des données homogènes et de qualité pour l'évaluation en France des politiques européennes de développement rural ont incité l'Inra et ses partenaires à investir dans l'observatoire et son outil Carto-dynamique. Aujourd'hui, cet outil, en constante évolution, favorise l'accessibilité de ces données dans des programmes de recherche. Il offre un ensemble de **services** aux utilisateurs : produire des traitements cartographiques, visualiser et naviguer dans des cartes et des tableaux d'indicateurs dynamiques à différentes échelles spatiales, ainsi que des **produits** : impression/export de carte sous forme d'image, export d'indicateurs en différents formats de fichiers, voire en fichiers géographiques sous format mif/mid.

Le développement de Carto-dynamique a été rendu possible par l'utilisation d'outils libres et les technologies web permettant facilement d'atteindre le public ciblé. Par la centralisation des données sur le serveur et l'incrémentation d'une métabase, Cartodynamique permet de piloter des appariements complexes de données qui demanderaient à l'utilisateur un investissement de temps important, et rend possible le traitement en ligne de grands volumes

d'informations comme les paiements destinés aux bénéficiaires des aides du RDR1 (table avec plus de 4 millions de lignes), ou les engagements des exploitations agricoles dans les mesures agro-environnementales (plus de 2 millions de lignes pour les MAE et MAET).
L'enjeu à venir pour l'observatoire est la normalisation des nomenclatures appliquées aux données et la standardisation des métadonnées dans le cadre d'une démarche de qualité des données.
Enfin, L'ODR a un rôle patrimonial de conservation dans le temps des données servant à produire des résultats de recherche.

Remerciements : Ce travail a reçu le soutien financier du ministère de l'Alimentation de l'agriculture et de la pêche ainsi que celui du département SAE2.

DynaforNet, un système d'information pour un site de recherche à long terme

Exemple de la gestion de données sur la biodiversité

Sylvie Ladet,[1], Wilfried Heintz[1], Valéry Rasplus[1]

Résumé : cet article décrit la chaîne de traitement informatique de données diverses acquises sur le terrain dans un espace géographique délimité au cours du temps et qui comprend la structuration de type relationnelle (SGBD) et spatialisée (SIG), le stockage (serveur et sauvegarde) et la valorisation (métadonnées et éventuellement diffusion vers l'extérieur).

Mots clés : SGBD, SIG, métadonnées, IDS, IDG, catalogage, stockage, sauvegarde, écologie du paysage, relevés de terrain, avifaune, modélisation.

Introduction

Les activités de l'unité mixte de recherche Dynafor du centre Inra de Toulouse sont centrées sur la gestion durable des ressources forestières et de l'espace rural dans le cadre de l'écologie du paysage. L'objectif principal de Dynafor est de comprendre et de modéliser les relations entre des processus écologiques, des processus techniques et des processus socio-économiques dans la gestion de ressources naturelles renouvelables (cf. **encadré n°1**). Par construction Dynafor est un projet de recherche interdisciplinaire qui mobilise des disciplines biologiques, techniques et socio-économiques. Cette unité travaille depuis plus de 30 ans sur deux sites d'études localisés dans le Sud-Ouest de la France. Nous allons ici aborder le site atelier des Vallées et Coteaux de Gascogne, situé au sud de Toulouse qui a été labellisé LTSER (Long Term Sociological and Ecological Research) en 2007. La thématique centrale des recherches effectuées concerne le suivi à long terme (i) des services écosystémiques rendus par la biodiversité dans les paysages agricoles, (ii) des facteurs qui les influencent (composition et structure du paysage, pratiques agricoles) et (iii) de leur évolution dans le contexte des changements globaux (changement d'utilisation des terres, changement climatique, changements dans l'environnement économique et réglementaire). Aussi chaque site est caractérisé par un périmètre géographique délimité au sein duquel seront mesurées in situ conjointement des données biologiques, des données sur les pratiques de gestion (parcelle et exploitation), des données sur les structures paysagères (mosaïque des occupations du sol, infrastructures écologiques, etc.). L'ensemble de ces données sont structurées numériquement au sein d'un système d'information pour faciliter l'intégration et la mise à disposition de ces données aux chercheurs de Dynafor. Pour illustrer concrètement nos propos, nous avons trouvé intéressant de faire écho à l'article de Laurent Raison (2007) qui porte sur l'inventaire des oiseaux nicheurs par la méthode des points d'écoute dans les paysages

[1] UMR1201 DYNAFOR – Dynamiques forestières dans l'espace rural - INRA - F-31326 Castanet Tolosan
☏ 05 61 28 52 55 ✉ sylvie.ladet@toulouse.inra.fr ; Wilfried.Heintz@toulouse.inra.fr ; Valéry.Rasplus@toulouse.inra.fr

agricoles. Il décrit la méthodologie de collecte de données sur le terrain des Coteaux de Gascogne avec la description précise de la phase de planification. Plusieurs fois il fait référence au système d'information en citant les mots-clés suivants SIG, bases de données… que nous gérons.
Nous allons décrire, au travers de ses différentes parties, la chaîne de traitement informatique de ces données acquises sur le terrain au cours du temps qui comprend la structuration de type relationnelle (SGBD) et spatialisée (SIG), le stockage (serveur et sauvegarde) et la valorisation (métadonnées et éventuellement diffusion vers l'extérieur).

Encadré n°1 : Présentation de l'unité Dynafor

L'unité Dynafor est une unité mixte de recherche INRA- INP/ENSAT regroupant chercheurs de l'INRA et enseignants-chercheurs de l'école d'agronomie de Toulouse y compris du personnel technique aussi bien en informatique qu'en observation de terrain. Elle s'est engagée dans le développement de recherches qui s'inscrivent résolument dans le champ d'une écologie du paysage « pour l'action ». L'écologie du paysage est vue comme science intégratrice des relations homme-nature et une science utile pour l'action. Dans nos travaux, nous considérons le paysage comme le fruit d'une coévolution entre les systèmes sociaux et les systèmes écologiques à des échelles spatiales et temporelles multiples qui correspondent aux différents niveaux d'organisation des processus étudiés (Allen et Star, 1982). Les processus étudiés concernent globalement la durabilité des entités de production agricole et forestière, le changement d'utilisation des terres dans les paysages, la dynamique de colonisation des arbres dans les paysages en déprise agricole, l'influence des pratiques de gestion agricole et forestière sur la biodiversité des paysages à différentes échelles, l'étude spécifique du rôle des lisières agriculture-forêt sur la biodiversité et les services rendus par la biodiversité à l'agriculture. Il s'agit de faire face aux enjeux actuels dans les espaces ruraux et forestiers, induits par les changements globaux qui concernent conjointement le climat, l'occupation des terres, la biodiversité et les activités humaines.
A Dynafor, nous sommes 3 responsables[1] techniques du systèmes d'information qui se construit « chemin faisant » sur ce terrain d'étude et qui est vu comme un ensemble de moyens humains, techniques, informationnels et méthodologiques, permettant d'acquérir, mémoriser, structurer, traiter, interpréter et diffuser entre les membres d'une organisation (ici, l'UMR), les informations recueillies par les disciplines en place. L'objectif est que chacun puisse consulter les informations produites par les uns et les autres, que chacun puisse utiliser et croiser les informations recueillies en y apportant des données produites dans sa discipline, afin de créer un « capital » de connaissances communes, matière première de la démarche interdisciplinaire de l'unité. Nous gérons donc un gros volume de données qu'il s'agit d'organiser et de mettre à disposition en respectant les standards informatiques.

[1] *Valery Rasplus (gestionnaire de bases de données) a la responsabilité technique de la conception et la maintenance des bases de données relationnelles.*
Sylvie Ladet (géomaticienne généraliste) a la responsabilité technique de la conception et la maintenance des systèmes d'information géographique et veille à la cohérence globale du système d'information.
Wilfried Heintz (géomaticien spécialisé sur la mise en réseau des SI) intervient sur la métadonnée pour mettre en œuvre la traçabilité.

1. Contexte informatique à Dynafor

1.1 Vallées et Coteaux de Gascogne

La plateforme LTSER de Dynafor est un paysage rural à vocation agricole situé à 80 kilomètres de Toulouse dans la région du Bas-Comminges (Deconchat et al, 2007). Il s'agit d'une zone vallonnée d'une altitude comprise entre 200 et 400m couvrant 250 km² avec une matrice agricole de polyculture-élevage marquée par la présence de nombreux bois fragmentés et d'éléments linéaires boisés de type haie, alignement d'arbres et un habitat dispersé. Ce terrain est l'objet, depuis les années 80 de nombreuses études menées *in situ* et répétée (avec une fréquence variable, par exemple tous les 10 ans par exemple pour un dispositif de suivi de communautés d'oiseaux ou tous les 20 ans pour le suivi des exploitations agricoles) sur l'évolution de la biodiversité des communautés végétales herbacées et animales (insectes, oiseaux), sur l'évolution des pratiques de gestion des exploitations agricoles et forestières. Aussi en 2007, Dynafor a obtenu la labellisation LTSER de ce territoire (numéro d'enregistrement LTER_EU_FR_003). C'est un label lancé en 1980 par la *National Science Fondation* aux Etats-Unis. Le but de ce label, maintenant international, est de mettre en réseau des sites de recherche à long terme, dans des écosystèmes importants et sensibles à travers le monde. Il en résulte une grande quantité de données de nature diverses qu'il convient de structurer informatiquement. Une des caractéristiques les plus importantes de ce label est le développement de bases de données, disponible pour tous les chercheurs, contenant des informations sur des données de types écologiques, socio-techniques et sur les conditions environnementales (occupation des sols, climat …) sur les sites étudiés. Nous identifions l'importance de la profondeur historique des données recueillies pour comprendre ces systèmes socio-écologiques complexes.

1.2 Les 4 composantes du système d'information à Dynafor

Nous avons mis en place progressivement un système d'information pour nous aider dans cet objectif de mutualisation et de partage. Ce système mis en place dans l'unité est un outil scientifique et technique fédérateur de type plateforme qui participe à établir un lien tangible entre les différents compartiments des systèmes étudiés. Cette plateforme est « chemin faisant » c'est-à-dire évolutive, modulaire. Il comporte 4 composantes qui interagissent : les outils (matériels et logiciels dédiés), les ressources humaines (disciplines des sciences et technologies de l'information et de la communication allant de la géomatique, à la modélisation en passant par la gestion de bases de données), les données (cartographiques et tabulaires) et les procédures (méthodologies, protocoles). Un effort particulier a été placé à Dynafor dans son développement et sa structuration au cours des dernières années. Il s'intègre peu à peu dans des réseaux d'échange de données (réseau LTER) qui nécessitent une organisation claire de son fonctionnement.

1.3 Exemple de la structure de la plateforme sur les données Oiseaux

Depuis 1981, chercheurs et techniciens de l'unité Dynafor effectuent des inventaires de communautés d'oiseaux en utilisant une méthode standardisée des points d'écoute (Blondel, J. *et al,* 1970) en réalisant concrètement des relevés de terrains comme l'explique Raison (2007).

Les recensements ornithologiques sont notés dans un premier temps sur une fiche de terrain construite préalablement (support papier) afin de recueillir des données sur l'abondance des oiseaux et les variables du milieu susceptibles d'expliquer cette abondance.

Analysons plus précisément le contenu de ces fiches qui indiquent le type de données à stocker numériquement. Un point d'écoute consiste à réaliser une observation stationnaire, géo-localisée, pendant une période totale de 20 minutes. Pour un point d'écoute donné, le temps global est divisé en quatre parties de cinq minutes (Raison L., 2007). On note sur la fiche les espèces d'oiseaux, vues, entendues, posées ou en vol dans un rayon préalablement défini (en moyenne 250 mètres). Des conditions environnementales sont également prises en compte, comme le vent, la pluie, le soleil, les nuages. Ces données couplées à des données d'habitat (forêt, jardin, verger, haie, lisière, pâture, type de culture), de préférence alimentaire (insecte, grain, etc.) permettent par la suite de modéliser les relations entre les oiseaux et les différents habitats. Pour le chercheur, l'évolution des communautés, l'apparition ou la disparition d'espèces et leur distribution sont des indicateurs de changements des structures paysagères (Balent G. et Courtiade B., 1992 ; Monteil et al, 2005). Les campagnes d'inventaire des populations d'oiseaux par la méthode des points d'écoute sur le terrain délimité des Coteaux de Gascogne, répétées dans le temps, ont exigé d'acquérir une compétence de gestion informatique fine afin de répondre au double objectifs de conservation et de structuration des données.

2. Structuration des données

2.1 SGBD : intégration des données de terrain

Ces informations manuscrites sont ensuite enregistrées dans un système informatique de gestion de données afin d'assurer à la fois une restitution rapide des données scientifiques pour les chercheurs intéressés et une réponse au besoin de sauvegarde sécurisée de données numériques. La collecte de données de terrain exige que celles-ci soient conservées et structurées dans un système de gestion approprié afin de répondre aux exigences de traitements et de diffusion de données volumineuses pour les recherches.

À l'unité Dynafor, l'ensemble des données ornithologiques recueillies sur le terrain sont intégrées dans une base de données relationnelle (Access©) ; elle a été déclarée dans l'inventaire des bases de données scientifiques structurées à l'Inra en 2009 sous le nom de *DynaBird*. Elle se démarque des bases de données traditionnelles en prenant en compte le facteur temps - dans l'intégration des données de campagne - et l'intégration de données spatiales via la localisation géographique des points d'écoute en X et Y.
Cette base de données géo-référencée comprend 13 tables et 130 champs. Elle comprend deux types de catégories de variables : les variables à expliquer comme l'abondance des espèces (137)[2], l'évolution de la répartition ; les variables explicatives comme les caractéristiques de terrain, les traits d'oiseaux (137), le nombre de campagnes (27), le nombre de communes (36), le nombre de stations (2981), le nombre de points d'écoute (5644), le nombre d'oiseaux contactés (167743). La volumétrie totale de cette base de données est de 113 546 lignes d'enregistrements.

[2] Entre parenthèses figure l'effectif rencontré pour donner une idée de la quantité de données manipulées

La **figure 1** donne une vision d'ensemble des tables physiques et des contraintes d'intégrités qui garantissent la cohérence des données lors des mises à jour de la base (Gardarin G., 2005).

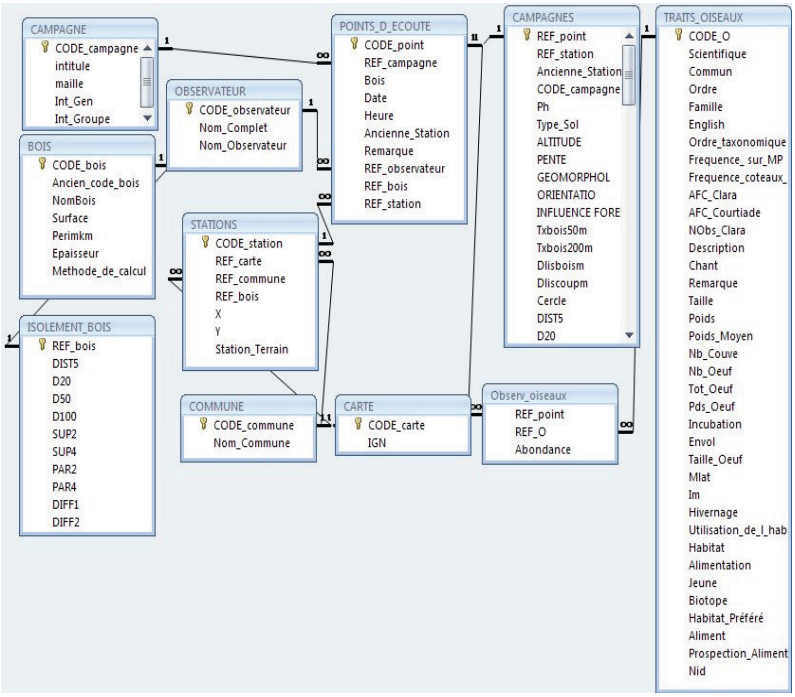


Figure 1 *: modèle physique des données de DynaBird*

2.2 Interfaçage des bases pour utilisation par les chercheurs

Un point important du cahier des charges était de construire un outil de gestion informatique facile à prendre en main afin de privilégier l'autonomie de ses utilisateurs. Son utilisation ne nécessite pas de connaissance particulière du logiciel Access©, ne demande pas de voir et de différencier les objets Access© (tables, formulaires, requêtes, états, macro, modules), ni de saisir directement dans les tables et aucune compréhension du langage de requête SQL (*Structured Query Language*) n'est exigée. Pour ce faire, un ensemble de formulaires, servant d'interface entre les données intégrées dans les tables et l'utilisateur, ont été créés en fonction des besoins et des pratiques de terrain, facilitant la rapidité de la saisie des données. L'utilisateur peut réaliser des opérations de consultation, de modification, de suppression, de mise à jour, de sélection et d'extraction de données croisées ou non.

La **figure 2** montre une partie du formulaire de saisie des campagnes d'écoute d'oiseaux.

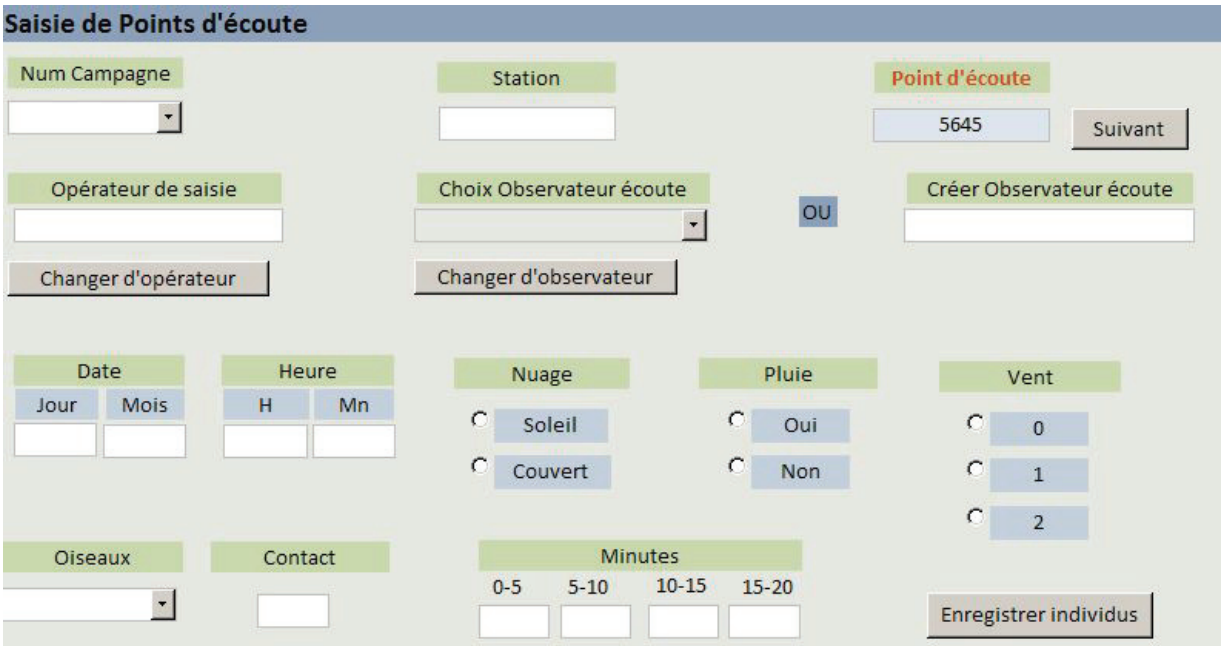


Figure 2 : *extrait du formulaire de saisie de points d'écoute*

En mode saisie, la tabulation permet de passer automatiquement d'un objet à un autre (zone de texte, case d'option, zone de liste déroulante). La quasi-totalité des données subissent un contrôle de validité. En cas de doute, une consultation visuelle de n'importe quelle campagne ou point d'écoute permet de modifier rapidement les données éventuellement erronées, comme le montre la **figure 3**.

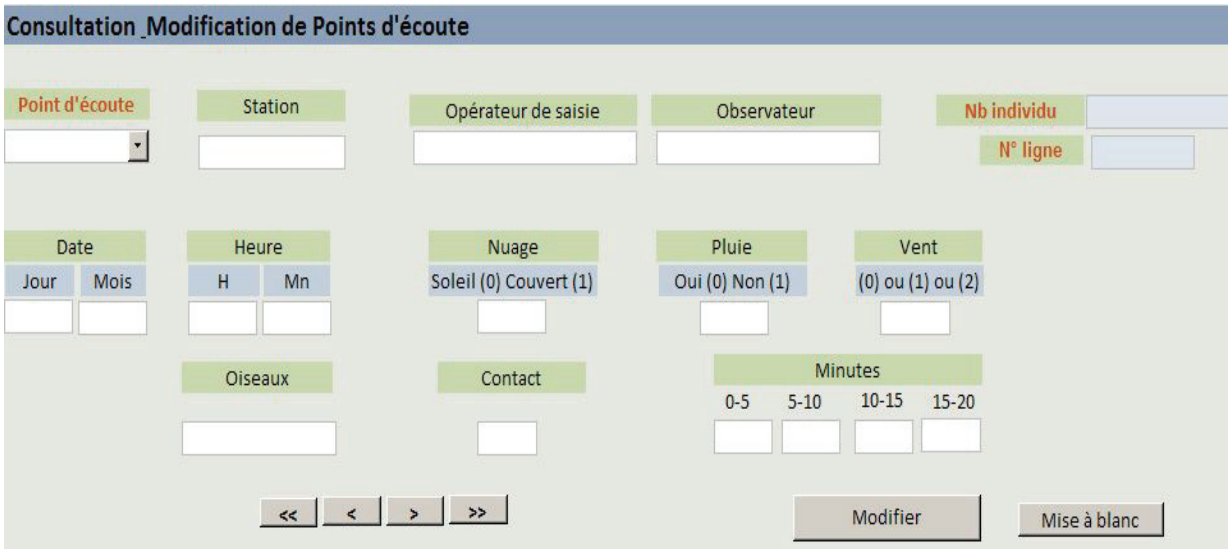


Figure 3 : *extrait du formulaire de consultation et modification de points d'écoute*

A tout moment, on peut consulter l'ensemble des données oiseaux point d'écoute par point d'écoute et ainsi suivre l'évolution des effectifs et des répartitions à un moment donné sur un point géo-localisé donné
Cette base de données construite sous Access© est interconnectée au logiciel de système d'information géographique (SIG) utilisé à Dynafor (ArcGis©) via une connexion ODBC (*Open DataBase Connectivity)*. C'est une technologie permettant d'interfacer de façon standard une application à n'importe quel serveur de bases de données, pour peu que celui-ci possède un driver ODBC (la quasi-totalité des SGBD possèdent un tel pilote).
Pour plus d'informations sur ce point particulier, se reporter à la note technique « *Comment connecter ArcGIS à des fichiers Microsoft Access 2007 (ACCDB)* » : http://support.esrifrance.fr/index.asp?page=/articles/arcgis/connexions%20aux%20bases%20de%20donnees/32976.htm (consulté le 12/03/2010). Dans le sens base vers SIG, ceci permet de spatialiser le résultat de requêtes effectuées dans la base de données. Dans le sens SIG vers base, ceci permet d'enrichir la base avec des données saisies dans le SIG (comme le pourcentage des occupations du sol autour des points d'écoute).

3. Spatialisation

Une partie des données acquises sur le site des Coteaux de Gascogne se présente sous la forme spatialisable (coordonnées X et Y des points d'écoute) ou cartographique (données brutes comme une photographie aérienne ou élaborées comme une carte d'occupation des sols). Nous touchons ici à la dimension spatiale qui est un des principaux facteurs des données manipulées par Dynafor. Le système d'information construit comprend donc un module géographique conséquent car la dimension spatiale est cruciale pour comprendre et modéliser les relations entre des processus écologiques, techniques et sociotechniques dans la gestion des ressources naturelles. Nous sommes alors dans le monde des SIG.
Mais au fait qu'est ce qu'un SIG ? Il existe des réponses multiples dans la littérature dédiée à la géomatique et la recherche d'une définition acceptée par tous est quasiment impossible tant elle varie en fonction des auteurs. Nous avons repris celle de l'université de géographie de Reno à savoir que le SIG est « *un outil unique intégrant des données diverses mais localisées dans le même espace géographique, relatives à la fois à la terre et à l'homme, à leurs interactions et à leurs évolutions respectives, quels que soient les domaines concernés: physiques, sociaux, économiques, écologiques, culturels, etc. Ce rassemblement permet d'élaborer les synthèses indispensables à la prise de décision dans tous les domaines aussi bien dans les situations de crise que dans le suivi des évolutions à long terme*». Ce qui nous intéresse dans cette définition est le fait que l'utilisation des SIG peut faire émerger les liens dans une perspective synchronique et diachronique dans l'organisation du territoire et synthétiser des informations provenant de sources et de domaines divers. À Dynafor, le SIG sert à plusieurs niveaux d'intervention en amont et en aval après la récolte des données sur le terrain.

3.1 En amont, la dimension spatiale des données collectées

Le SIG à Dynafor est un moyen efficace pour :

- aider à la planification des dispositifs d'observations de formes variées : au bureau nous construisons avec le chercheur responsable les plans d'échantillonnage régulier sur un maillage défini (campagne d'écoute sur plusieurs communes contigües) *versus* échantillonnage stratifié selon la nature d'occupation des sols (campagne d'écoute uniquement dans des bois). Pour se faire, il existe des scripts et d'autres extensions permettant de construire ces plans d'échantillonnage. Citons par exemple les outils gratuits proposés par l'extension *Hawth's Analysis Tools* (Beyer, 2004) qui sont développés pour le logiciel ArcGis© (**figure 4**).

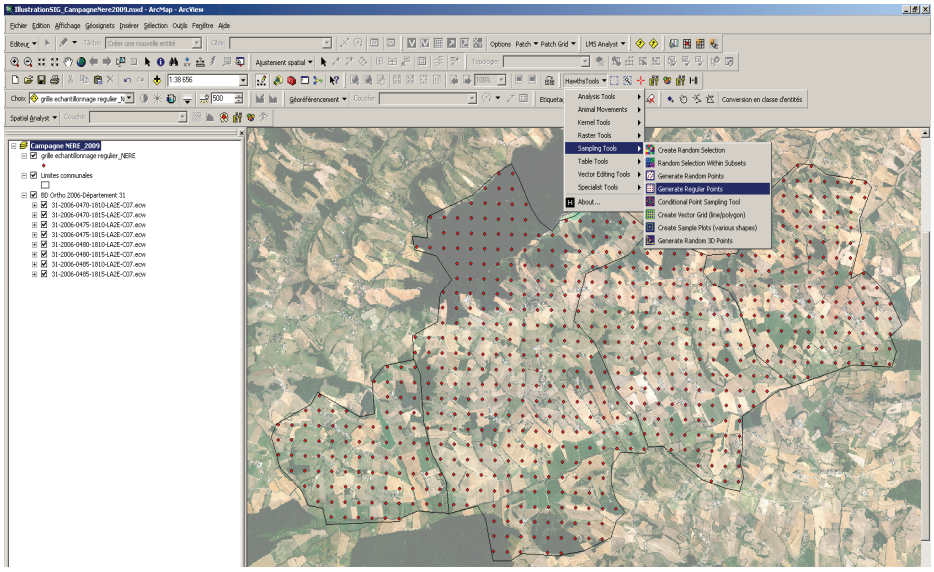


Figure 4 : *grille (676 points rouges) d'échantillonnage régulier d'un pas de 250 mètres sur 4 communes contigües construite sous ArcGis© avec la fonctionnalité « Generate regular points tool » de l'extension Hawth's Analysis Tools*

- aider à la coordination sur le terrain. Raison et al (2007) ont présenté dans leur article les documents cartographiques élaborées en amont de la campagne de terrain d'écoute des oiseaux (carte générale du dispositif sur fond de scan topographique de l'IGN, fiche de terrain recto avec fond photographie aérienne en couleur avec les points d'écoute géo-référencés). Effectivement, ces supports cartographiques peuvent aider à la coordination du dispositif sur le terrain en facilitant le repérage sur le terrain. Il est important pour l'ornithologue par exemple de se positionner précisément sur le terrain, de bien visualiser et de délimiter la zone dans laquelle il va dénombrer les oiseaux et décrire le milieu (zone tampon de type carré centré sur le point). C'est d'autant plus essentiel que les points d'écoute sont réalisés au même endroit à des dates différentes pour pouvoir comparer la composition de l'avifaune. La précision de la localisation géographique est donc primordiale. Pour se faire, nous recherchons des outils complémentaires pour personnaliser Arcgis© en rajoutant des extensions pour combler des manques de la version par défaut du logiciel.

Citons par exemple l'extension « *Créer un atlas cartographique* » disponible par téléchargement : http://support.esrifrance.fr/index.asp?page=/outilsscripts/arcgis/arcmap/miseenpage/atlas/atlas.html. (Consulté le 12/03/2010).

Cet outil permet de créer un document ArcMap©[3] contenant plusieurs mises en page cartographiques avec la même architecture. Dans une chaîne de production cartographique, l'objectif est d'automatiser la production de cartes (produire les 676 fiches recto de terrain automatiquement) en générant une grille d'index qui correspond à une classe d'entités utilisée pour diviser vos données en tuiles. En ce qui concerne les campagnes d'écoute d'oiseau, la grille d'index est la classe d'entités contenant les zones tampons autour de chaque point d'écoute. Dans l'atlas, nous pouvons ajouter des éléments (titre, échelle, carroyage, fichier tableur avec numéro des points, nom de communes...), qui seront pré-remplis de manière dynamique (**figure 5**).

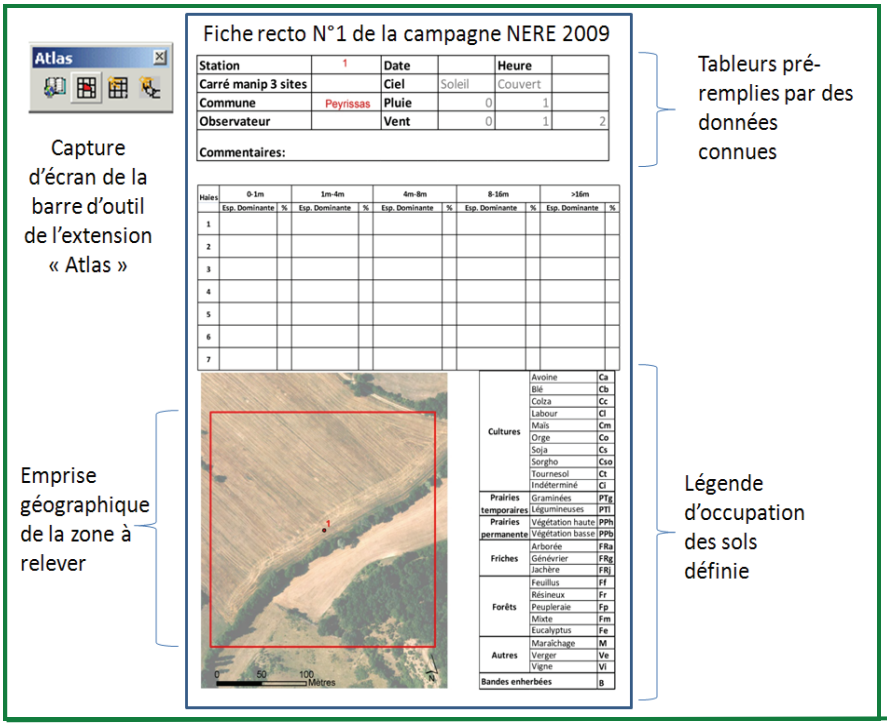


Figure 5 : *fiche recto du point d'écoute de la campagne Nère en 2009 (1ère page de l'atlas cartographique) réalisée sous ArcGis avec l'extension Atlas.*
L'ornithologue est assisté pour relever l'occupation des sols autour d'u point d'écoute. Il doit reporter sur l'image, le sigle de l'occupation du sol qu'il observe.

Ces méthodes actuelles servent à prendre en compte la dimension spatiale sur le terrain via l'utilisation de fonds cartographiques annotés. D'autres solutions plus abouties pourront être utilisés dans un futur proche comme les GPS embarqués ou le développement d'une application SIG nomade. Installé sur un récepteur PDA (*Personal Digital Assistant*), une tablette PC ou un

[3] ArcMap est l'application centrale d'ArcGis pour la visualisation de cartes, l'édition, l'analyse et la mise en page

micro-ordinateur portable, cela permet aux utilisateurs d'accéder aux bases de données (relationnelles et/ou géographiques) directement sur le terrain et d'utiliser des fonctions d'assistance à la localisation avec la possibilité d'intégrer un système GPS.

3.2 En aval : l'organisation des données géographiques

Dans le logiciel de SIG, l'intégration de données hétérogènes qui proviennent de sources de données variées, va nous aider : (1) à formaliser ces données à la fois dans leur géométrie, dans leur thématique et leur agencement temporel, (2) à combiner ces données entre elles grâce aux outils d'analyse spatiale, (3) à répondre aux interrogations des thématiciens par requêtes multi-critères et (4) à quantifier et modéliser les dynamiques spatio-temporelles de processus complexes (Collet C, 2005).

À Dynafor, nous disposons d'une base de données géographiques déjà structurée contenant des données dites brutes (comme les photographies aériennes de l'IGN, les images satellites et autres données images du Référentiel à grande échelle (cf. protocole de diffusion IGN/MAP/MEDAD de diffusion de données aux organismes de recherche …)) ou des données au format vecteur (cadastre, limites administratives, réseaux routiers…) et des données dites élaborées (comme des cartes d'occupations des sols, carte des réseaux de haies, carte des parcellaires d'exploitation agricole…). Mais notre objectif de suivi demande de mettre à jour cette base en intégrant des données plus récentes (carte d'occupation du sol de l'année en cours) ou des données récemment numérisées (cartes anciennes). Nous sommes alors confrontées à des problèmes spécifiques, à savoir le géoréférencement de cartes anciennes, la digitalisation de notations relevées sur des fiches de terrain (Joliveau T., 2004). Le traitement des données est ensuite relativement aisé ; les couches d'information étant superposables, à la manière de calques, les possibilités de croisements des données et de requêtes thématiques sont très grandes dans le logiciel SIG et permettent l'élaboration de cartes recoupant différents thèmes rendant l'analyse des données (**figure 6**) et de leur évolution spatio-temporelle (**figure 7**) à la fois souple et rapide.
L'atout majeur du logiciel SIG est sa capacité de navigation dans la base de données et son interrogation à la fois dans sa dimension spatiale (cartes) et dans sa dimension attributaire (tables). On peut donc établir des requêtes attributaires, fondées sur des relations logiques et des requêtes spatiales ou sur une combinaison des deux pour sélectionner des objets ayant une caractéristique particulière. Par exemple, nous pouvons analyser l'évolution des linéaires de haies au cours du temps. Pour rappel, les haies sont un objet important du paysage pour les oiseaux car ils assurent la connectivité des éléments boisés et sont donc essentiels à la circulation des oiseaux forestiers.

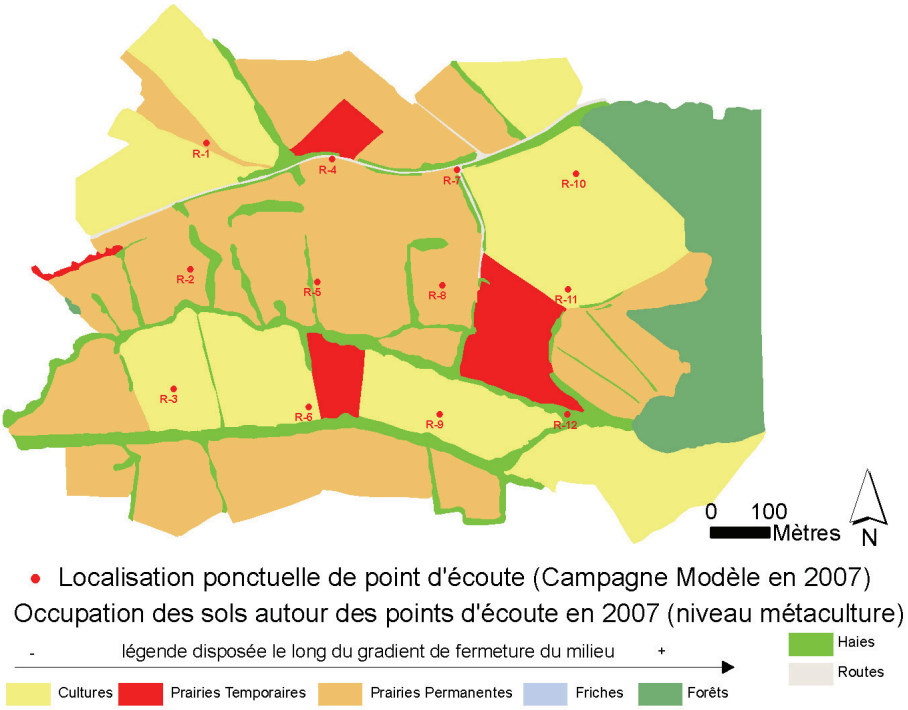


Figure 6 : *saisie sous ArcGis sous forme de polygone de l'occupation des sols relevée autour du point d'écoute de la campagne Modèle en 2007.*

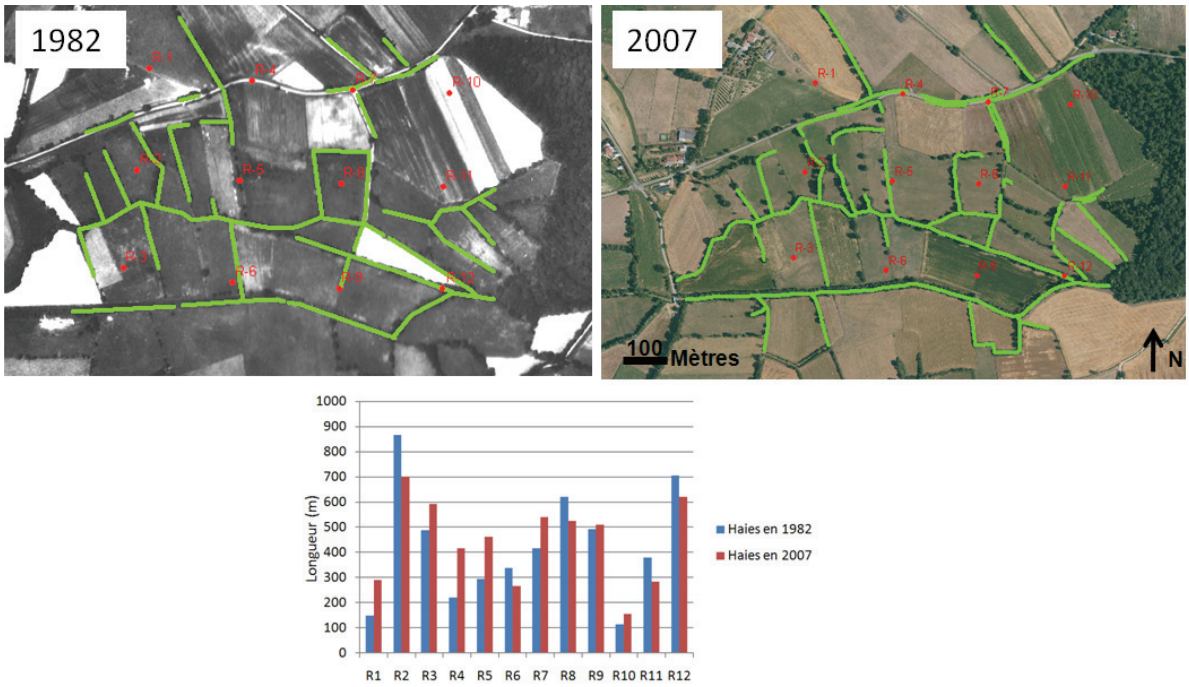


Figure 7 : *évolution des haies en 1979 et en 2007 sous ArcGis de la campagne Modèle représentée sous forme de cartes statiques en haut ou de graphique en bas.*

Le SIG apparaît comme un moyen d'organiser le traitement des données issues du dépouillement de données de terrain, de les analyser, de les confronter à d'autres données et de formaliser le tout. Il rend aussi possible le traitement des données à des échelles précises (résolution spatiale de l'ordre du mètre et résolution temporelle de l'ordre de l'année et résolution thématique de plus en plus précise). Ces données stockées atteignent alors des volumétries lourdes mais ces systèmes informatiques peuvent les gérer et les combiner sans limite de capacité liée à leur stockage. Une nouvelle question apparait alors, la description des données de plus en plus nombreuses.

4. Stockage et sauvegarde

4.1 Nécessité d'un serveur dédié

Le traitement de données volumineuses requiert avant toute chose un espace adéquat pour les stocker. Malgré les considérables évolutions technologiques dans le domaine du stockage de données numériques et l'augmentation permanente de la taille des disques durs, il n'en reste pas moins que les données géographiques telles que les images satellites ou aériennes voient leur taille grandir proportionnellement à leur résolution. Les données issues des nouvelles technologies comme le LIDAR (*Light Detection and Ranging*) sont extrêmement volumineuses. La notion de stockage suggère aussi implicitement celle de l'accès à la donnée. S'il est important de disposer d'un espace suffisant pour stocker la donnée, il est également déterminant de pouvoir la récupérer simplement et rapidement, depuis différents postes de traitement.
De nombreuses solutions techniques, parmi lesquelles les disques durs portables disposent désormais de grandes capacités. Néanmoins, un système optimisé tel qu'un serveur de données s'avère rapidement indispensable pour un fonctionnement mutualisé et durable. On peut se contenter d'un serveur de stockage tel un système NAS (*Network Attach Storage*) où les données brutes sont déposées et où chaque agent peut les consulter et manipuler. Une solution plus évoluée consiste à mettre en place un serveur où les données sont stockées dans une base de données, permettant une consultation et une manipulation plus poussée. Il est actuellement possible de stocker des données géographiques vectorielles sous forme d'enregistrements à l'intérieur d'une base de données. Ceci est permis par la cartouche spatiale PostGIS[4] du système de gestion de base de données (SGBD) PostgreSQL. Le requêtage s'en trouve considérablement optimisé. Le stockage des données de type matriciel n'est pas compatible avec un SGBD mais de nouveaux logiciels tels que PostGRID permettent d'indexer ces données pour un traitement optimisé. Ainsi, la solution retenue à Dynafor est la mise en place d'un serveur dédié au stockage et à la diffusion des données dites volumineuses.

4.2 Processus de sauvegarde des données

Un aspect primordial du stockage de données est d'en garantir la pérennité au sens « durabilité ». En clair, comment se prémunir d'une défaillance matérielle qui mettrait en péril la restitution des données numériques stockées sur le serveur ? Il s'agit ici du processus de sauvegarde des données, étape indispensable d'une chaîne de stockage correctement mise en œuvre. En effet,

[4] PostGIS est une extension du système de gestion de bases de données relationnel à objets PostgreSQL qui permet de stocker des objets SIG dans une base. L'ensemble des logiciels PostGIS et PostgreSQL est libre de droit.

une véritable sauvegarde se termine sur un support durable, à l'abri des aléas extérieurs. La solution retenue dans la plupart des cas consiste à écrire les données du serveur sur bande magnétique, elle-même stockée dans une armoire blindée et ignifugée.
L'équipe informatique du centre Inra de Toulouse utilise un robot de sauvegarde mutualisé qui réplique quotidiennement le contenu du serveur sur bande magnétique ; cette solution de sauvegarde reste onéreuse (de l'ordre de 400€ pour une licence annuelle).

5. Valorisation des données

5.1 Catalogage

Le préambule à la diffusion de données est la structuration et la description de ces données. En effet, il a été maintes fois constaté que beaucoup de données (géographiques, statistiques, etc.) étaient inutilisables car leurs méthodes d'acquisition ou certaines de leurs caractéristiques intrinsèques n'avaient pas été notées ou conservées. La traçabilité d'une ressource est donc avant tout une garantie de pérennité (Michener, 2005).

Pour mettre en œuvre cette traçabilité, il existe la métadonnée et c'est là qu'intervient le géomaticien spécialisé sur la mise en réseau de SI (Wilfried Heintz pour notre unité). La métadonnée est une « information sur l'information », composée d'un ensemble de champs descriptifs normalisés pour caractériser une ressource. S'il est vrai que ce concept est étroitement lié aux données de nature géographique, il reste néanmoins applicable à toute information quelle que soit sa nature. C'est toutefois dans le domaine de l'information géographique que les innovations sont les plus importantes en terme de gestion et valorisation des données. Ceci est particulièrement dû à la profusion de ce type d'information, elle-même permise par les avancées technologiques en matière de systèmes d'informations géographiques (Heintz et Guéro, 2005). L'usage des SIG s'est démocratisé ; les logiciels cartographiques ne sont plus l'apanage des spécialistes, justifiant notre démarche de diffusion des données auprès des chercheurs de l'unité. Néanmoins, face à cette déferlante de données aussi nombreuses que diversifiées, il est nécessaire de créer des outils de gestion, de diffusion et de valorisation. Ainsi, les outils de gestion de métadonnées - également appelés (géo)catalogues - sont devenus des outils de diffusion et de partage de l'information géographique (CNIG, 2007). Ils sont même devenus la pierre angulaire des infrastructures de données spatiales et d'autres géoportails. Mais le développement de ces derniers ne peut se faire qu'en respectant les lois sur la diffusion des données. Le contexte législatif actuel justifie donc pleinement la démarche de catalogage des données produites et utilisées dans le fonctionnement de l'unité. Ce catalogue de données doit se construire conformément aux recommandations techniques de la directive européenne Inspire ratifiée par la France (cf. **encadré n°2**).

Plusieurs outils de catalogage existent, dont deux implémentent le profil français de la norme ISO (*International Organization for Standardization)*. Ces logiciels sont Geosource, version « française » de l'outil Geonetwork, et MDWeb, développé par l'IRD[5]. Tous deux permettent de

[5] IRD = Institut de recherche pour le développement ayant pour vocation de mener des recherches pour l'hémisphère Sud.

générer des fiches de métadonnées 100% compatibles avec les directives et préconisations en vigueur. Le choix de l'un ou de l'autre dépendra des caractéristiques du serveur hébergeant l'outil et des compétences du futur administrateur du logiciel sur des langages de programmation comme le XML (*eXtensible Markup Language)* /XSL (*eXtensible StyleSheet Language)* ou sur la configuration du conteneur libre de servlets Tomcat.

Encadré n°2 : catalogage et législation européenne

La directive européenne *Inspire* (*Infrastructure for Spatial Information in Europe*) incite fortement les acteurs du service public à se doter d'un système de gestion de métadonnées. Elle vise à mettre en place une infrastructure de données éographiques (IDG) européenne pour favoriser l'harmonisation des pratiques, la mise en réseaux et l'élaboration d'accords de partage et d'accès aux données. Cette directive, approuvée au conseil des ministres de l'Union européenne et par le Parlement européen en avril 2007, a été transposée dans le droit français et mise en application en avril 2009 (CNIG, 2007).
La mise en œuvre d'un catalogue de données doit se faire conformément aux recommandations techniques de la directive qui portent essentiellement sur le choix de la norme de métadonnées. Dans cette directive, la norme ISO 19115 a été retenue, garantissant une interopérabilité totale avec les autres systèmes de catalogage. La norme ISO 19115:2003 a le statut de norme internationale depuis 2003. Cette norme abstraite de contenu définit en les organisant par classes toutes les informations que l'on peut mettre à dispositions pour décrire la donnée. Cette norme s'est affirmée comme une référence pour l'information géographique dans le domaine des métadonnées. Outre le fait qu'elle est enfin disponible après une longue gestation, ses atouts principaux pour la communauté résident dans son caractère modulaire et extensible qui la rend aisément adaptable. Son pendant « technique » est la norme ISO19139 qui définit un schéma XML et les balises correspondant aux différents champs descriptifs de la norme.

Bien que les intérêts de la traçabilité des données soient indéniables, il n'en demeure pas moins qu'une telle entreprise apporte une surcharge de travail. Outre les aspects de mise en place des outils qui nécessitent diverses compétences, c'est le remplissage des métadonnées qui s'avère être la tâche la plus lourde. En effet, ce travail incombe en toute logique à la personne « responsable » de la donnée décrite. Or le passif des métadonnées à remplir est tel que les personnes concernées se voient contraintes d'y consacrer beaucoup de temps. Une vision à long terme de cette tâche justifie d'y passer du temps, mais en pratique, cet aspect se révèle être le plus contraignant pour la mise en place des métadonnées.

5.2 Mise à disposition et diffusion externe

Les données décrites dans un catalogue de métadonnées sont facilement identifiables, il s'agit alors de les rendre consultables. En effet, la directive *Inspire* préconise également la mise à

disposition des données cartographiques via des protocoles spécifiques : la finalité étant la mise en œuvre d'une infrastructure de données spatiales.
Le principe d'une infrastructure de données spatiales est d'offrir la visualisation sur une plateforme commune de données hébergées sur des serveurs physiquement distants les uns des autres.
Ceci est possible grâce à des services web normalisés tels que le Web Map Service (WMS) ou le Web Feature Service (WFS). Ils en existent d'autres, tous définis par l'Open GIS Consortium (OGC). Ces services ont été créés pour faciliter le partage distant de données géographiques par serveurs interposés, en allant jusqu'à la manipulation et à l'édition des données. On mesure tout l'intérêt de pouvoir accéder à tout moment à des données stockées physiquement chez le producteur de ces données : pas de duplicata obsolètes ni de mises à jour régulières des données. Sur une telle plateforme, accessible à l'aide d'un simple navigateur web, tout membre de l'unité peut accéder à un ensemble de données géographiques jusqu'ici réservées à la seule utilisation des experts SIG.
Si l'on se replace dans notre chaîne de traitement, nous voici donc à l'ultime étape. Les données collectées sur le terrain et traitées par le responsable SIG ont été structurées et stockées sur le serveur dédié. Elles sont identifiables facilement via le catalogue de métadonnées, et chacun peut les manipuler à l'aide d'outils simples disponibles sur une interface intégrée (**figure 8).**

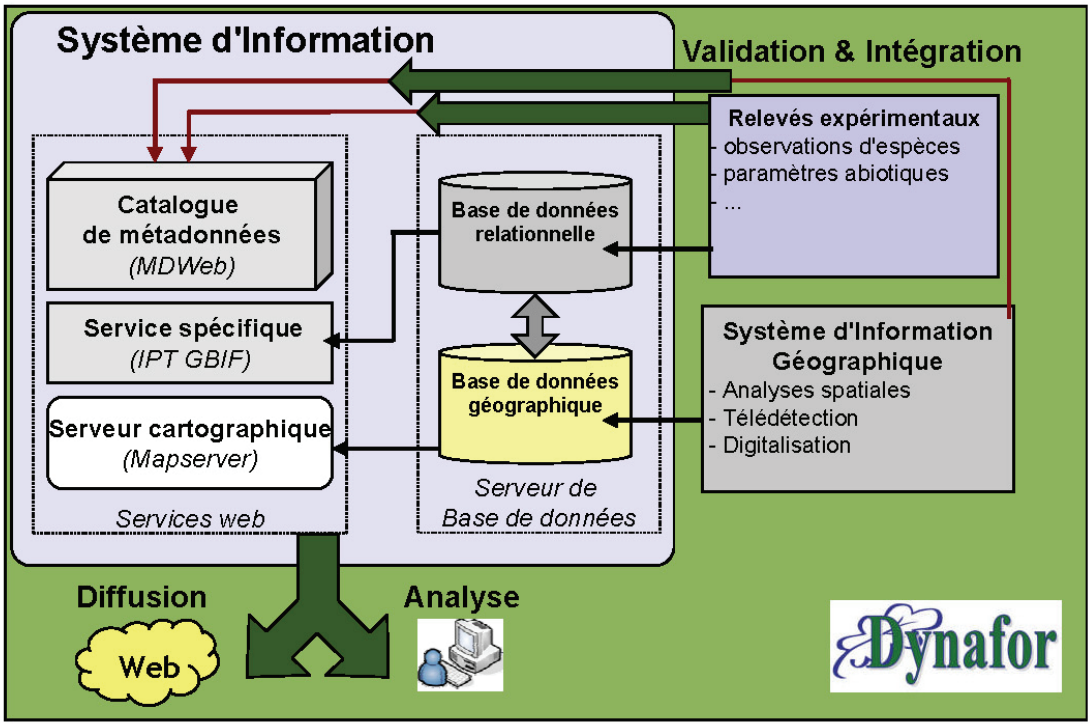


Figure 8 : *architecture du système d'information complet de Dynafor*

Des cas concrets de systèmes similaires existent déjà comme l'Observatoire du Développement Rural[6] (ODR) géré par une unité de service de l'Inra de Toulouse et Dynafor est membre de l'Observatoire Spatial Régional (OSR). L'OSR est aujourd'hui un service d'observation labellisé par l'Institut national des sciences de l'univers (INSU). C'est un dispositif d'organisation des systèmes d'observation dévolus aux suivis de long terme sur le fonctionnement et l'évolution des surfaces continentales aux échelles du paysage et de la région, et de la valorisation des données et informations à des fins scientifiques (fonctionnement des surfaces continentales) et appliquées (gestion durable et intégrée des territoires). Les principales fonctions et moyens de l'OSR sont :

- d'assurer la collecte et le traitement d'un ensemble de mesures et d'informations issues de mesures *in situ*, d'enquêtes et de données de télédétection ;
- d'organiser et de gérer ces données dans un système d'information ;
- de diffuser ces données, associées à des outils d'analyse, au moyen de l'Internet.

Conclusion

Le système d'information géré par Dynafor coordonne ainsi, grâce à l'information, les activités de recherche menée sur son site d'étude à long terme et lui permet ainsi d'atteindre ses objectifs de gestion durable des ressources naturelles et de l'espace rural dans le cadre de l'écologie du paysage. La constitution d'un SI mobilise des compétences variées et complémentaires dans le domaine de l'informatique. Le système d'information se construit autour de processus "métier" et leurs interactions, et non simplement autour de bases de données ou de logiciels informatiques.

Ce SI garantit une pérennité des recherches passées et présentes en facilitant la transparence et la qualité des données recueillies *in situ*. Il doit être chemin faisant, c'est-à-dire évolutif, modulaire pour appuyer les recherches scientifiques futures. C'est un vrai outil d'intégration scientifique. Sa mise en place requiert des compétences informatiques et techniques spécifiques qui mobilisent des métiers différents ainsi qu'un matériel informatique performant. Les métadonnées ont une fonction déterminante dans ce système d'information complet car elles servent de relais permettant l'utilisation adéquate des données par les autres utilisateurs. Il est important de noter qu'actuellement, la difficulté ne réside plus véritablement dans l'acquisition et l'analyse de données volumineuses, mais bien dans leur valorisation et leur traçabilité. Les performances techniques sont peu voire plus du tout restrictive, en revanche, la connaissance des données accumulées pose un véritable problème de gouvernance de l'information.

Remerciements : Ce travail a été financé par 2 projets ANR : POPULAR (ANR-06-PADD-014-004 ; 2006-2010) et BiodivAgriM (ANR-07-BDIV-002-03 ; 2008-2011) pour le recueil de données et par le département SAD pour l'achat du serveur de stockage.

[6] Il s'agit, pour l'essentiel, d'une plate-forme informatique associant une base de données et des fonctionnalités permettant d'une part de réaliser des traitements statistiques et cartographiques et d'autre part de gérer les modalités d'accès des différentes catégories d'utilisateurs aux informations contenues dans cette base. Ces dernières sont relatives à la mise en oeuvre des différents dispositifs du second pilier, aux caractéristiques des exploitations et des territoires.

Bibliographie

Allen TFH., Starr TB. (1982) Hierarchy Perspectives for ecological complexity. *The University of Chicago Press*, Chicago.

Balent G., Courtiade B. (1992) Modelling bird communities/landscape patterns relationships *in* a rural area of South-Western France, *Landsacpe Ecology*, 6.3, p. 195-211.

Beyer H. L. (2004) Hawth's Analysis Tools for ArcGIS. Disponible sur http://www.spatialecology.com/htools. Consulté le 12/03/10.

Blondel, J., Ferry C., Frochot B. (1970) Méthode des indices ponctuels d'abondance (I.P.A.) ou des relevés d'avifaune par « stations d'écoute », *Alauda*, Vol. 37, n°1, ,57-71.

Bucher B. (2002) *L'aide à l'accès à l'information géographique : un environnement de conception coopérative d'utilisations de données géographiques*. Thèse. Université Paris VI. p. 207.

Collet C. (2005) Analyse spatiale, géomatique et systèmes d'information géographique, *Revue Internationale de Géomatique*, 15 (4), 393-414.

Craglia M. (2003) *Towards a European Approach to Metadata for Geographic Information*. *In* European Commission GI&GIS Portal.

Deconchat M., Gibon A. et coll. (2007) How to set up research framework to analyse socio-ecological interactive processes in a rural landscape. *Ecology and Society*, 12, 15. [online] URL: http://www.ecologyandsociety.org/vol12/iss1/art15

Fraisse S., Pornon H. (2008) Les métadonnées : corvée ou nécessité ? SIG expert - juin/juillet 2008, n°63, 29-35.

Gardarin G. (2005) *Bases de données*, Eyrolles, sixième édition.

Heintz W. et Guéro M.-C. (2005) *Projet SInPa : Systèmes d'Informations Partagées pour la gestion forestière régionale*. De l'observation des écosystèmes forestiers à l'information sur la forêt – 2 et 3fev. 2005, Paris, France.

Joliveau T. (2004) Géomatique et gestion environnementale du territoire. Recherches sur un usage géographique des SIG, *Mémoire d'Habilitation à Diriger des Recherches en Sciences Humaines*, Rouen, Université de Rouen, 2 vol.

Lembo A., Wagenet L., Schusler T., De Gloria S. (2007) *Creating affordable Internet map server applications for regional scale applications.* Journal of Environmental Management 85, 1120-1131

Michener W.K. (2005) Meta-information concepts for ecological data management. Ecological Informatics1

Monteil C, Deconchat M, Balent G. (2005) Simple neural network reveals unexpected patterns of bird species richness in forest fragments. *Landscape Ecology* 20, 513-527.

Najar C. *et al*. (2006) Spatial Data and Metadata Integration for SDI Interoperability. Article under Review for the International Journal of Spatial Data Infrastructures

Noucher M., De Sede-Marceau MH, Golay F. et Pornon H. (2007) *Contributions socio-cognitives aux dynamiques de coopération inter-organisationnelle autour de la donnée géographique.*, In GéoCongrès, Québec, octobre 2007.

Noucher M. (2007) *Coopérer autour des SIG : typologie et exemples en France et au Canada*, Atelier SIG Pyrénées (en visioconférence depuis le GéoCongrès de Québec), St Gaudens, octobre 2007.

Raison L. (2007) L'inventaire des oiseaux nicheurs par la méthode des points d'écoute dans les paysages agricoles – Regard d'un ornithologue - *in* Techniques et pratiques de recueil de données *in situ* - *Le Cahier des Techniques de l'Inra*, numéro spécial, 79-86.

ODOMATRIX

Calcul de distances routières intercommunales

Mohamed Hilal [1]

Résumé *: La question de l'accès aux services de proximité et aux emplois nourrit les débats régionaux et nationaux d'aménagement du territoire. Pour y répondre, l'Inra a développé un logiciel, dénommé ODOMATRIX, qui calcule des distances routières intercommunales et des zones d'accessibilité de pôles ou d'équipements. Les distances sont exprimées en kilomètres et en temps de trajet aux heures creuses et aux heures de pointe ; elles sont établies entre chefs-lieux de communes. Ce logiciel a été déposé à l'Agence pour la protection des programmes.*

Mots-clés : accessibilité, desserte, distancier, réseau routier, temps de transport

Introduction

L'unité CESAER (Centre d'économie et de sociologie appliquées à l'agriculture et aux espaces ruraux) du centre Inra de Dijon a développé le logiciel ODOMATRIX qui calcule les distances routières intercommunales et les zones d'accessibilité de pôles ou d'équipements.
Le réseau routier français est modélisé en utilisant plusieurs couches d'informations géographiques qui décrivent les routes, le relief, l'occupation du sol. La codification des tronçons routiers et l'attribution des vitesses de circulation sont réalisées de façon transparente. Le réseau routier se trouve ainsi décrit par une matrice creuse de 200 000 lignes et 200 000 colonnes. Le noyau de l'application, écrit en c++, calcule les plus courts chemins en utilisant l'algorithme de Dijkstra avec tas de Fibonacci. Les modules de gestion de données et d'interrogation sont des scripts Matlab, ce logiciel autorisant la gestion des très grandes matrices de données et l'écriture de code vectorisée pour accroître la vitesse d'exécution. ODOMATRIX calcule des distances intercommunales (matrices carrées ou rectangulaires), des distances pour des couples de communes (données bilocalisées), la distance ou le temps le plus court pour atteindre des pôles ou des communes équipées en commerces et services.
ODOMATRIX offre la possibilité de faire plusieurs milliers de requêtes simultanées (propriété du Dijkstra), avec une vitesse d'exécution très rapide (propriété du tas de Fibonacci et de l'écriture vectorisée du code) et une grande souplesse d'utilisation.

[1] UMR1041 CESAER, Centre d'économie et de sociologie appliquées à l'agriculture et aux espaces ruraux - INRA - F-21079 Dijon - ☏ 03 80 77 25 77 ✉ mohamed.hilal@dijon.inra.fr

Les applications commerciales spécialisées dans le calcul des itinéraires ne cumulent pas tous ces avantages[2].
La première section décrira les principales étapes de la modélisation du réseau routier, la deuxième section sera consacrée aux fonctionnalités du logiciel, la troisième présentera quelques exemples illustrant les utilisations potentielles de l'outil.

1. La modélisation du réseau routier

1.1 Principes généraux

Les outils de calcul de trajets routiers utilisent tous en entrée des bases de données géographiques routières : IGN (Institut géographique national), TéléAtlas, Michelin, etc. La modélisation du réseau repose sur les caractéristiques physiques (nombre et forme des tronçons, coordonnées des nœuds, ...) et les caractéristiques fonctionnelles (vocation de la liaison, nombre de voies, ...). La vitesse des véhicules qui empruntent le réseau est déterminée à partir de ces informations. La modélisation dans ODOMATRIX intègre, en plus de ces données, la nature géographique de l'environnement traversé (zones urbaines, périurbaines, rurales) et les formes du relief (sinuosité planimétrique et altimétrique). Ces informations exogènes sont ajoutées en tant qu'attributs aux tronçons ce qui constitue une originalité de l'outil.
Les vitesses, affectées aux types de tronçons aux heures dites creuses (HC) ou aux heures dites pleines ou de pointe (HP), sont fonction des caractéristiques physiques et fonctionnelles des tronçons et fonction du type de zone traversée et du relief.

1.2 Les bases de données géographiques utilisées

La modélisation utilise les bases de données géographiques de l'IGN décrites dans l'encadré suivant. Elle utilise également les nomenclatures géographiques de l'INSEE délimitant les unités urbaines et les aires urbaines, et des données de l'IFEN décrivant l'occupation des terres (CORINE Land Cover).
La description de l'environnement urbain combine les nomenclatures géographiques de INSEE, qui délimitent l'urbain et le rural tant du point de vue morphologique (unités urbaines) que du point de vue fonctionnel (zonage en aires urbaines et en aires d'emploi de l'espace rural), et le périmètre des taches urbaines identifiées dans la base CORINE Land Cover (CLC).

La notion d'unité urbaine repose sur la continuité de l'habitat : est considérée comme telle un ensemble d'une ou plusieurs communes présentant une continuité du tissu bâti (pas de coupure de plus de 200 mètres entre deux constructions) et comptant au moins 2 000 habitants. La condition est que chaque commune de l'unité urbaine possède plus

[2] L'auteur remercie l'INSEE et le CERTU pour leur participation à la mise au point finale d'ODOMATRIX dans le cadre d'un groupe de travail « Distancier ». Le CERTU a suggéré de tenir compte de la sinuosité altimétrique et d'utiliser une grille de vitesse empirique permettant de calculer les temps aux heures creuses et aux heures de pointes. L'INSEE, notamment par l'intermédiaire de ses deux Pôles de service de l'action régionale « Analyse territoriale » de Marseille et « Synthèse locale » de Lyon, a contribué à valider l'outil en comparant les distances calculées par ODOMATRIX avec les distances produites par des solutions commerciales (ChronoMap, Way Server, Mappy et ViaMichelin). Les divers types de distances présentent une très forte corrélation. L'analyse des écarts montre une sensible surestimation des distances routières calculées par Michelin, par rapport aux résultats des autres distanciers et des écarts très faibles entre les distances calculées par ODOMATRIX d'une part et les distances Mappy ou Wayserveur, d'autre part (INSEE PSAR Analyse territoriale (2006) - Distanciers AT30 : rapport au bureau du Copar, Juillet 2006).

de la moitié de sa population dans cette zone bâtie. Les unités urbaines sont redéfinies à l'occasion de chaque recensement de la population. Elles peuvent s'étendre sur plusieurs départements. Si la zone bâtie se situe sur une seule commune, on parlera de ville isolée et dans le cas contraire, on a une agglomération multicommunale.

Le zonage en aires urbaines et en aires d'emploi de l'espace rural (ZAUER) décline le territoire en six catégories. Les trois premières représentent l'espace à dominante urbaine qui comprend des pôles urbains, des couronnes périurbaines et des communes multipolarisées. Les trois autres constituent l'espace à dominante rurale : ce sont les pôles d'emploi de l'espace rural, les couronnes des pôles d'emploi de l'espace rural et les autres communes de l'espace à dominante rurale (Vallès, 2002).
Grâce au ZAUER, on peut délimiter un périmètre fonctionnel autour de chaque pôle urbain. Ce périmètre, dénommé aire urbaine, est un ensemble de communes d'un seul tenant et sans enclave, constitué par un pôle urbain, et par des communes rurales ou unités urbaines (couronne périurbaine) dont au moins 40 % de la population résidente ayant un emploi travaille dans le pôle ou dans des communes attirées par celui-ci.
Le nombre d'habitants des pôles urbains, qui permettra de construire trois catégories d'aires urbaines, est calculé à partir du **recensement de la population de 1999 (RP 1999)**.

La base de données géographique CORINE Land Cover est un inventaire biophysique de l'occupation des terres à l'échelle européenne mise en œuvre sous l'égide de l'Agence européenne pour l'environnement. ODOMATRIX utilise le millésime 2000, produit par l'IFEN, afin d'identifier les « taches urbaines ».

Ces bases permettent modéliser le réseau routier et l'environnement géographique traversé.

Encadré : *les bases de données géographiques de l'IGN*

ROUTE500® est une base de données routière décrivant 500 000 km de routes du réseau métropolitain classé (autoroutes, nationales, départementales), plus des tronçons donnant accès aux 36 000 communes, et des éléments d'habillage. Elle est dérivée de la BDCARTO® par généralisation. Sa date de validité est variable d'un département à l'autre et dépend de la date de mise à jour de la BDCARTO®. ODOMATRIX utilise le millésime 2004. La base est organisée en fonction de quatre centres d'intérêt ou « thèmes » : le thème administratif (limite administratives et communes), le thème habillage à des échelles cartographiques voisines du 1/250 000 (côtes et frontières, hydrographie, occupation du sol), le thème réseau ferré (nœuds ferroviaires et voies ferrées) et enfin le thème routes et infrastructures de transport (aérodrome, restrictions de circulation, nœuds des communes, nœuds routiers, tronçons de route). ODOMATRIX utilise les tables « TRONCON_ROUTE » « NŒUD_ROUTIER » et « NŒUD_COMMUNE », tirées du dernier thème, pour modéliser le réseau.

La base de données altimétriques de l'IGN BDALTI® 500 décrit le relief français sous forme d'un modèle numérique de terrain (MNT) qui correspond à une grille régulière de points d'altitudes connues ; dans la base utilisée, les points sont séparés de 500 m.

La base de données GEOFLA® décrit le découpage administratif simplifié de l'ensemble des communes françaises. Cette base servira à délimiter l'environnement urbain traversé par les tronçons de routes.
La base GEOFLA® est couplée au ***répertoire géographique des communes RGC®*** qui contient le nom, la position géographique et l'altitude des chefs-lieux de communes (mairies).

1.3 Enrichissement de la base de données routière

La vitesse de circulation sur un réseau routier dépend de la vocation de la route traversée (autoroute, liaison principale, liaison régionale, liaison locale) et de l'environnement géographique traversé (agglomération, relief). Afin de déterminer ces éléments, la base de données routière est enrichie par traitements géomatiques successifs. Ainsi, il devient possible : de classer la voierie en cinq types de tronçons ; d'intégrer l'environnement géographique traversé par les routes ; de créer un attribut de sinuosité globale après avoir déterminé la valeur des sinuosités planimétrique et altimétrique.

Tableau 1 : *caractéristiques physiques et fonctionnelles utilisées pour déterminer le type de voie*

Nom	Valeurs possibles	Description
Vocation	Type autoroutier	Autoroutes et routes express à chaussées séparées et carrefours dénivelés
	Liaison principale	Densification des autoroutes ayant pour fonction : - d'assurer les liaisons à fort trafic à caractère prioritaire entre agglomérations importantes - d'assurer les liaisons des agglomérations importantes au réseau autoroutier - d'offrir une alternative à une autoroute si celle ci est payante - de proposer des itinéraires de contournement des agglomérations - d'assurer la continuité en agglomération des liaisons interurbaines à fort trafic quand il n'y a pas de contournement possible
	Liaison régionale	Densification du réseau autoroutier et principal. Les liaisons régionales ont fonction, quand celle ci n'est pas assurée par des itinéraires à vocation plus élevée de : - relier les communes de moindre importance entre elles (les chefs-lieux de canton en particulier), - proposer des itinéraires de substitution aux autoroutes payantes, - proposer des itinéraires de contournement des agglomérations, - desservir les points de passage des obstacles naturels quand ils sont peu nombreux (cols routiers, ponts), - desservir les agglomérations d'où partent des liaisons maritimes et les embarcadères isolés, - desservir les localités et sites touristiques importants, - relier des voies de vocation plus élevée, - structurer la circulation en agglomération.
	Liaison locale	Valeur par exclusion des autres valeurs d'attribut
	Bretelle	Description détaillée des échangeurs et carrefours aménagés ou ronds points d'extension supérieure à 100 mètres.
Nombre de chaussées	Une chaussée	Les voies à chaussées séparées contiguës sont représentées par un seul tronçon à deux chaussées. Les voies à chaussées séparées éloignées de plus de 100 mètres sur moins de 1 kilomètre sont décrites par 2 tronçons à une chaussée à sens unique
	Deux chaussées et plus	
Accès au tronçon	Libre	
	A péage	
	Fermeture saisonnière	
Classement administratif	Autoroute	Classement administratif attribué à la route empruntant le tronçon routier.
	Route nationale	
	Route départementale	
	Sans objet	

1.3.1 Codage du type de voierie

Les informations physiques et fonctionnelles de la voierie (vocation, nombre de chaussées, accès payant ou gratuit) sont utilisées pour classer les tronçons routiers en cinq types de voie :

- les autoroutes à péage ;
- les 2 × 2 voies de type autoroutier ou voie rapide ;
- les liaisons principales et régionales à une voie ;
- les liaisons locales à une voie ;
- les bretelles d'accès

Les informations nécessaires sont tirées de la base ROUTE500® et conduisent à la classification décrite dans le **tableau 1**.

L'attribut type de voie (type_voie) est créé par une série de requêtes SQL :

- les autoroutes à péage
*SELECT * FROM route500 WHERE vocation= "Type autoroutier" AND acces = "à péage" AND classement_administratif_route= "autoroute" INTO autoroute*
UPDATE autoroute SET type_voie = "autoroute"

- les 2x2 voies de type autoroutier ou voie rapide
*SELECT * FROM rte500 WHERE type_voie <> "autoroute" AND nombre_chaussées = "2 chaussées" INTO voie_rapide*
UPDATE voie_rapide SET type_voie = "voie rapide"

- les liaisons principales et régionales à une voie ;
*SELECT * FROM rte500 WHERE (type_voie <> "autoroute" AND type_voie <> "voie rapide") AND (vocation <> "liaison locale" AND vocation <> "bretelle") AND nombre_nbaussées = "1 chaussée" INTO voie_regionale*
UPDATE voie_regionale SET type_voie = "voie regionale"

- les liaisons locales à une voie ;
*SELECT * FROM rte500 WHERE (type_voie <> "autoroute" AND type_voie <> "Voie rapide" AND type_voie <> "voie regionale") AND vocation = "liaison locale" AND nombre_chaussées = "1 chaussée" INTO voie_locale*
UPDATE voie_locale SET type_voie = "voie locale"

- les bretelles d'accès
*SELECT * FROM rte500 WHERE vocation = "Bretelle" INTO bretelle*
UPDATE bretelle SET type_voie = "bretelle"

1.3.2 Détermination de l'environnement géographique traversé

Les attributs utilisés pour définir l'environnement urbain des tronçons routiers sont tirés des sources du **tableau 2**.

Tableau 2 : *informations utilisées pour caractériser l'environnement géographique*

Sources	Valeurs possibles	Description
(A) Nomenclature des unités urbaines (INSEE)	0	Commune rurale
	11	Ville isolée
	21	Ville centre d'une agglomération multicommunale
	22	Banlieue d'une agglomération multicommunale
(B) Zonage en aires urbaines et aires d'emploi de l'espace rural (INSEE / INRA)	001..354	Identifiants géographiques des aires urbaines
	1	Pôle urbain
	2	Couronne périurbaine
	4	Pôle d'emploi de l'espace rural
	5	Couronne des pôles d'emploi de l'espace rural
(C) RP 1999 (INSEE)	 1 2 3	Nombre d'habitants du pôle urbain > 200 000 [100 000 ; 200 000] < 100 000
(D) CORINE Land Cover (IFEN)	1	Territoires artificialisés

Les attributs tirés de *(A)*, *(B)* et *(C)* sont couplés à la base géographique GEOFLA® et permettent d'obtenir une première couche géographique des parties urbanisées du territoire. Cette couche est constituée :

- des villes centres des aires urbaines, soit *(B) = 001..354 et (A) = 11 ou 21* ;
- du contour des aires urbaines, soit *(B) = 001..354 et (A) = 22* ;
- du contour des pôles ruraux et de leur couronne, soit *(B) = 4 et 5*.

Le découpage communal est utilisé pour délimiter la « tache urbaine » des villes centres des aires urbaines. En revanche, on utilise le périmètre issu de la couche CORINE Land Cover pour délimiter les « taches urbaines » dans les aires urbaines hors villes centres et dans les aires constituées de pôles ruraux et de leur couronne.

Grâce au nombre d'habitants du pôle urbain, on scinde le territoire urbanisé en trois catégories :

- tache urbaine d'un pôle ou à proximité d'un pôle de plus de 200 000 habitants ;
- tache urbaine d'un pôle ou proche d'un pôle compris entre 100 000 et 200 000 habitants ;
- tache urbaine d'un pôle ou à proximité d'un pôle de moins de 100 000 habitants. Les pôles ruraux et leur couronne sont considérés comme faisant partie de cette catégorie.

Le **tableau 3** récapitule ce traitement.

Tableau 3 : *grille d'environnement urbain en sept types de zone*

111	Les villes centres des aires urbaines dont le pôle urbain fait plus de 200 000 habitants
112	Les villes centres des aires urbaines dont le pôle urbain est compris entre 100 000 et 200 000 habitants
113	Les villes centres des aires urbaines dont le pôle urbain fait moins de 100 000 habitants
121	Les taches urbaines des aires urbaines dont le pôle urbain fait plus de 200 000 habitants, hors ville centre
122	Les taches urbaines des aires urbaines dont le pôle urbain est compris entre 100 000 et 200 000 habitants, hors ville centre
123	Les taches urbaines des aires urbaines dont le pôle urbain fait moins de 100 000 habitants hors ville centre ainsi que les taches urbaines des pôles des aires d'emploi de l'espace rural
130	Tout le reste de l'espace : reste des aires urbaines, reste des pôles des aires d'emploi de l'espace rural, reste de l'espace rural

La grille d'environnement urbain découpe l'espace métropolitain en sept types de zone (111 … 130) dans lesquels les conditions de circulation, et donc les vitesses, sur les tronçons routiers sont différentes.
Les sept types de zone sont ajoutés en tant qu'attributs à la table des tronçons routiers. Ensuite, une requête pour chaque tronçon donne, à chacune des sept variables créées, la proportion de la longueur du tronçon intersectant la zone. La vitesse affectée à ce tronçon sera donc une combinaison linéaire des vitesses des zones traversées.

1.3.4 Calcul d'un attribut de sinuosité globale

L'attribut de sinuosité globale décrit les modifications potentielles des conditions de circulation du fait des virages (sinuosité planimétrique) et des pentes (sinuosité altimétrique). Cette opération est réalisée en trois étapes décrites ci-après.

Etape 1 : calcul de la sinuosité planimétrique

La sinuosité planimétrique est estimée pour un tronçon par calcul de la différence entre sa longueur planimétrique et la distance à vol d'oiseau séparant les deux noeuds situés à ses deux extrémités. Ce calcul mobilise pour chaque tronçon :

- la longueur planimétrique *LP* qui est fournie par l'IGN avant simplification géométrique ;
- la distance à vol d'oiseau *DVO* calculée après avoir extrait les coordonnées des deux nœuds extrêmes.

Le ratio $S\_PLANI = 100 \times (LP - DVO) / DVO$ est un bon indicateur de la sinuosité planimétrique du tronçon. On considère que le tronçon est sinueux si ce ratio est supérieur à 30 % (valeur déterminée empiriquement). Cela concerne 4,1 % des tronçons situés majoritairement en montagne.

Etape 2 : calcul de la sinuosité altimétrique

Par analogie avec la sinuosité planimétrique, la sinuosité altimétrique peut être estimée pour un tronçon par calcul de la différence entre la longueur réelle et la longueur planimétrique fournie par l'IGN. En effet, la longueur planimétrique est une longueur projetée sur l'ellipsoïde de référence et ne tient pas compte du relief traversé par le tronçon. De fait, plus le relief est accidenté (succession de montés et descentes) et plus la longueur planimétrique sera différente de la longueur réelle du tronçon.
Le ratio $S\_ALTI = 100 \times (LR - LP) / LP$ est un bon indicateur de la sinuosité altimétrique du tronçon. Lorsqu'il tend vers 0 le tronçon est plat ; plus il s'en éloigne et plus le tronçon est sinueux. Le calcul mobilise pour chaque tronçon :

- la longueur planimétrique *LP* qui est fournie par l'IGN ;
- une estimation de la longueur réelle du tronçon qui tient compte du relief (MNT tiré de la BDALTI® 500). Cette estimation a nécessité les opérations suivantes :

1) fractionner chaque tronçon (ajout de nœuds) en plusieurs segments de façon à ce que la distance entre deux noeuds successifs ne soit pas supérieure à 500 m (pas du MNT) ;
2) draper la couche des tronçons de route sur le modèle numérique de terrain (superposition) de façon à pouvoir associer à chaque nœud une altitude extraite du MNT.
3) calculer la longueur réelle de chaque segment *lr* en utilisant la longueur planimétrique du segment *lp* et la différence d'altitude entre les nœuds extrêmes du segment *dz* : $lr = \sqrt{lp^2 + dz^2}$
4) sommer les longueurs réelles des segments qui composent chaque tronçon pour obtenir sa longueur réelle *LR*.

On considère que le tronçon est sinueux si le ratio *S_ALTI* est supérieur ou égal à 30 %. Cela concerne 4,2 % des tronçons.

Etape 3 : calcul de la sinuosité globale

Les calculs précédents permettent d'identifier les tronçons sur lesquels la circulation est affectée par le relief. On considère que la sinuosité planimétrique, respectivement altimétrique, est significative lorsque le ratio *S_PLANI*, respectivement *S_ALTI*, est supérieur à 30 %. Les deux ratios sont supérieurs à 30 % pour seulement 1,4 % des tronçons, alors que 2,7 % des tronçons dépassent le seuil uniquement pour *S_PLANI*, respectivement 2,8 % pour *S_ALTI*.

Ce résultat incite à calculer un indice de sinuosité globale qui tient compte à la fois de la sinuosité planimétrique et altimétrique. Un attribut est construit à cet effet : il prend la valeur 1 si le tronçon est sinueux et/ou pentu, 0 pour les tronçons plats et « droits ».

1.4 La base routière codifiée

Le temps de traversée de chaque arc est calculé en utilisant la longueur du tronçon, après correction planimétrique et altimétrique de la géométrie, et en appliquant une vitesse de circulation sur le réseau. La vitesse dépend de la vocation de la route (type autoroutier, liaison principale, liaison régionale, liaison locale), de l'environnement géographique traversé et du relief. Les conditions de circulation liées à la congestion du réseau sont partiellement prises en compte en intégrant une vitesse dite « en heure de pointe », sensée tenir compte du taux de charge de la voierie et qui varie selon le type de voie, la taille des pôles et l'environnement géographique.

Tableau 4 : *vitesses de circulation aux heures creuses et aux heures de pointe selon l'environnement géographique et le type de voie*

Environnement	Population pôle urbain	Type de voie	Vitesse en heure	
			creuse (HC)	de pointe (HP)
Ville centre des aires urbaines	plus de 200 000 habitants	autoroute	65	35
		2x2 voies	30	16
		principale et régionale	25	14
		locale à 1 voie	20	11
		bretelle	60	42
	entre 100 000 et 200 000 habitants	autoroute	65	41
		2x2 voies	30	19
		principale et régionale	25	16
		locale à 1 voie	20	13
		bretelle	60	47
	moins de 100 000 habitants	autoroute	65	53
		2x2 voies	30	25
		principale et régionale	25	20
		locale à 1 voie	20	16
		bretelle	60	54
Tache urbaine des aires urbaines hors ville centre	plus de 200 000 habitants	autoroute	70	53
		2x2 voies	40	28
		principale et régionale	30	19
		locale à 1 voie	20	12
		bretelle	20	42
	entre 100 000 et 200 000 habitants	autoroute	70	57
		2x2 voies	40	31
		principale et régionale	30	21
		locale à 1 voie	20	13
		bretelle	20	47
	moins de 100 000 habitants et taches urbaines des pôles d'emploi de l'espace rural	autoroute	70	64
		2x2 voies	40	36
		principale et régionale	30	26
		locale à 1 voie	20	17
		bretelle	20	54
Reste de l'aire urbaine, reste du pôle rural, espaces ruraux	sans objet	autoroute	130	130
		2x2 voies	85	85
		principale et régionale	70	70
		locale à 1 voie	60	60
		bretelle	60	60

Pour tenir compte de la sinuosité planimétrique et altimétrique nous appliquons les coefficients de pondération de la vitesse suivants : autoroute (0,75) ; 2x2 voies (0,70) ; principale et régionale (0,65) ; locale à 1 voie (0,60) ; bretelle (0,70). Les liaisons maritimes, assurant la jonction entre le continent et les îles (Corse, Iles du Ponant), sont intégrées dans la base routière. Elles comprennent les lignes de bac et les liaisons maritimes ouvertes aux automobiles et dont les embarcadères de départ et d'arrivée figurent parmi les nœuds routiers de ROUTE500®. La durée de traversée en minutes est fournie par l'IGN, les temps d'attente avant embarcation ne sont pas pris en compte.
La base codifiée est constituée de 199 586 nœuds et 571 365 tronçons. Pour chaque tronçon sont renseignées : les identifiants des noeuds des extrémités, la distance kilométrique et les temps de traversée en heure creuse et en heure de pointe déterminée d'après les vitesses de circulation estimées et d'après les coefficients de sinuosité.

Pour calculer les distances intercommunales, ODOMATRIX extrait de la base codifiée trois matrices creuses carrées de 199 586 lignes et 199 586 colonnes et ayant chacune 571 365 éléments non nuls.

Les Iles du Ponant

Les îles du Ponant sont, parmi la multitude d'îles et d'îlots qui jalonnent les côtes de la Manche et de l'Atlantique, les quinze îles habitées de façon permanente, constituant des collectivités locales et ne possédant pas de lien fixe avec le continent : Bréhat, Batz, Ouessant, Molène, Sein, Groix, Belle-Île, Houat, Hoëdic, Arz, l'île aux Moines, Yeu, Aix. Les deux archipels de Chausey et Glénan qui font partie des îles du Ponant ne sont pas pris en compte dans ODOMATRIX car ce sont des quartiers maritimes des communes continentales de Granville et Fouesnant.

2. Fonctionnalités d'ODOMATRIX

ODOMATRIX peut calculer les plus courts chemins entre les 199 586 nœuds du réseau routier en minimisant l'une des trois distances décrites dans la base de données routière codifiée, à savoir : la distance routière kilométrique, le temps de trajet en heure creuse ou en heure de pointe.

2.1 Architecture générale

2.1.1 Un noyau et quatre modules

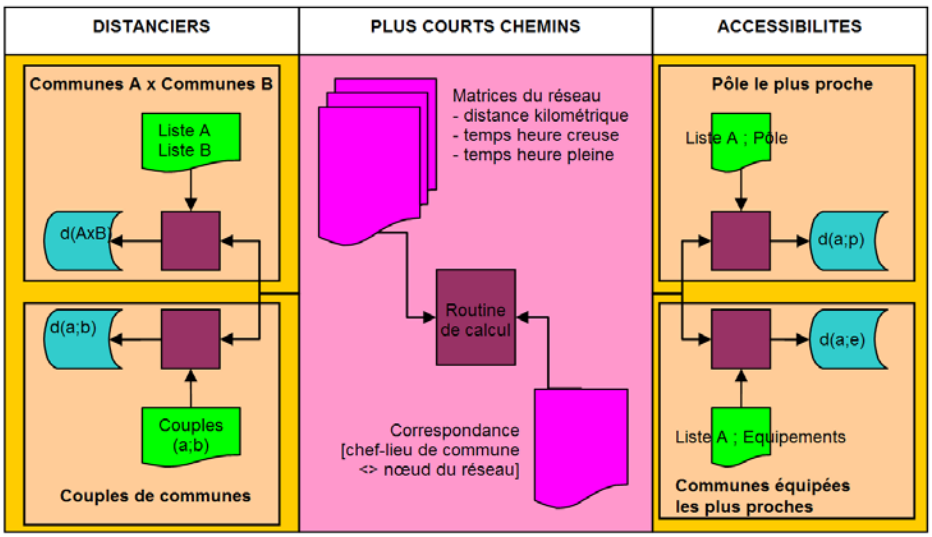
Le ***noyau*** de ODOMATRIX est une fonction de calcul des plus courts chemins qui utilise l'algorithme de Dijkstra (Dijkstra, 1959) avec Fibonacci (Fredman et Tarjan, 1987). Cette fonction calcule les chemins les plus courts d'un nœud donné à tous les autres nœuds du réseau.
La fonction est écrite en C++. Elle est compilée et peut-être appelée par Matlab retenu pour sa capacité à gérer les très grandes matrices et à augmenter la vitesse d'exécution des calculs en utilisant des techniques de vectorisation.

La fonction utilise en entrée :
- une des trois matrices de 199 586² éléments qui donne, pour chaque couple de nœuds correspondant à un tronçon routier, la distance routière kilométrique ou le temps de trajet en heure creuse ou en heure de pointe
- un dictionnaire géographique permettant d'associer à chaque commune le noyau routier le plus proche de son chef-lieu.

Bien que les calculs soient effectués en tenant compte des 199 586 nœuds du réseau, ODOMATRIX n'accepte en entrée que les codes des noeuds chefs-lieux de communes et, en sortie, il ne restitue que les distances entre nœuds chefs-lieux de communes. Cette limitation a été fixée au départ pour respecter les règles de diffusions des données IGN : la base géographique ROUTE500® n'étant pas publique, il n'est pas permis de diffuser à un utilisateur potentiel de ODOMATRIX la liste des noeuds du réseau et encore moins leur position géographique. De ce fait, les matrices creuses qui décrivent le réseau routier, ainsi que les autres données techniques (liste et position géographique des noeuds) sont encapsulées dans le logiciel, sans qu'il soit possible d'en faire une extraction totale ou partielle. Le dictionnaire des communes (INSEE), qui lui est public, permet d'identifier les noeuds du réseau qui sont chef lieu de commune et qui peuvent servir de points en entrée et en sortie du logiciel.
Les chefs lieux de commune correspondent à une géographie communale allant du 1er janvier 1999 au 1er janvier 2007, soit 36 624 communes existantes ou ayant existé entre ces deux dates en France métropolitaine.

Autour de ce noyau sont greffés ***quatre modules (*****tableau 5*****)***, développés en langage Matlab**,** qui appellent la fonction « noyau » de plus court chemin et qui calculent :
- des distanciers, en croisant deux ensembles de communes ;
- des distanciers pour des couples de communes pré-établis ;
- des accessibilités au pôle le plus « proche », à partir d'un tableau contenant en lignes des communes et sur une colonne une variable permettant de savoir si une commune de la liste est pôle ou pas ;
- des accessibilités aux équipements les « plus proches », à partir d'un tableau contenant en lignes des communes et en colonnes une variable donnant pour chaque équipement les effectifs de la commune.

Tableau 5 : *les modules d'ODOMATRIX*



2.1.2 L'interface graphique

L'interface graphique de ODOMATRIX est écrit en Matlab. Très simple, il se compose d'une barre de menus donnant accès aux quatre modules (menu Odomatrix), à la documentation, aux informations relatives à la licence et à l'auteur (menu Aide).

Interface d'ODOMATRIX

2.2 Description détaillée des modules du menu Odomatrix

Les modules présentés en 2.1.1 et dans le tableau 5 sont ici détaillés dans leur fonctionnement.

2.2.1 Les distanciers

Communes A × Communes B

Ce module calcule des matrices de distances intercommunales rectangulaires A×B ou carré A×A.

Il utilise, en entrée, deux fichiers contenant chacun une liste de codes communaux au format caractère sur 5 positions. Pour produire un distancier carré, on utilise deux fois le même fichier.

Exemple de fichier en entrée :

```
01001
01002
01004
01005
01006
01007
01008
01009
01010
```

Afin de réduire les temps de calcul des distanciers rectangulaires (A×B), le premier fichier doit avoir un nombre de lignes supérieur ou égal à celui du deuxième fichier.

Les distanciers peuvent être de très grande taille et difficilement manipulables. Le module refusera de lancer le calcul pour tout rectangle ayant plus de 37 millions de couples. Malgré cette limitation, on peut calculer des matrices contenant les distances entre les 36 624 communes, correspondant au référentiel géographique communal métropolitain disponible, et toutes les communes d'un ou de plusieurs départements (<1000).

En sortie, Le module fournit un fichier de données *.dat qui contient le distancier correspondant au produit cartésien A×B ou A×A. Les données sont stockées sous forme bilinéaire, avec les codes communaux entre guillemets et les champs séparés par des points-virgules :

dc_a : le code de la commune a
dc_b : le code de la commune b (lu dans le fichier B ou dans le fichier A si matrice AxA)
dvo : la distance à vol d'oiseau entre dc_a et dc_b exprimée en mètres
dr : la distance routière moyenne ([aller+retour]/2) entre dc_a et dc_b exprimée en mètres
hc : le temps de trajet moyen en heures creuses ([aller+retour]/2) entre dc_a et dc_b exprimé en minutes
hp : le temps de trajet moyen en heures de pointe ([aller+retour]/2) entre dc_a et dc_b exprimé en minutes

Couples de communes

Ce module calcule des distances intercommunales pour une liste de couples.
Il utilise, en entrée, un fichier contenant une liste de couples de communes définies par leurs codes communaux au format caractère sur 5 positions, avec comme séparateur « espace » ou « tabulation ».

Exemple de fichier en entrée :
13001 13008
13002 13008
13003 75054
13004 13008
13005 13005
13005 13006
13007 13001
13008 13001
13009 13010
13010 13009

Il n'y a pas de limitation de taille de fichier.
En sortie, le module produit un fichier de données *.dat qui contient pour chaque couple de communes (codes communaux entre guillemets et champs séparés par des points-virgules)
dc_a : le code INSEE de la commune a
dc_b : le code INSEE de la commune b
dvo : la distance à vol d'oiseau entre dc_a et dc_b exprimée en mètres
dr : la distance routière moyenne ([aller+retour]/2) entre dc_a et dc_b exprimée en mètres
hc : le temps de trajet moyen en heures creuses ([aller+retour]/2) entre dc_a et dc_b exprimé en minutes
hp : le temps de trajet moyen en heures de pointe ([aller+retour]/2) entre dc_a et dc_b exprimé en minutes

2.2.2 Les accessibilités

Pôle le plus proche

Ce module calcule la distance au pôle le plus proche, celui-ci pouvant être un pôle urbain, un pôle de bassin de vie ou toute autre commune polarisante.
Le module utilise, en entrée, un fichier contenant une liste de communes et une colonne précisant pour chaque commune si celle-ci est pôle (1) ou pas (0).

Exemple de fichier en entrée :
04001 0
04004 1
04005 0
04006 0
04007 0
04008 1
04009 0
04012 0
04013 0
04016 0
04017 0

En sortie, le module génère un fichier de données *.dat qui contient pour chaque commune le pôle le plus proche calculée en minimisant la distance kilométrique routière, les temps de trajet en heure creuse et en heure de pointe (codes communaux entre guillemets et champs séparés par des points-virgules) :

dc : code commune
dckm : pôle le plus proche d'après la distance routière kilométrique moyenne
km : distance routière moyenne ([aller+retour]/2) entre dc et dckm500 exprimée en mètres
dchc : pôle le plus proche d'après le temps de trajet moyen en heure creuse
hc : temps de trajet moyen en heure creuse ([aller+retour]/2) entre dc et dchc500 exprimée en minutes
dchp : pôle le plus proche d'après le temps de trajet moyen en heure de pointe
hp : temps de trajet moyen en heure de pointe ([aller+retour]/2) entre dc et dchp500 exprimée en minutes

Il n'y a pas de limitation de taille de fichier.

Communes équipées les plus proches

Ce module calcule la distance aux équipements les plus proches.
Il utilise, en entrée, un fichier qui contient une liste de communes et *n* colonnes. Chaque colonne correspond à un équipement. L'intersection ligne × colonne peut prendre comme valeur le nombre d'équipement(s) de la commune ou simplement 1 ou 0 indiquant la présence ou l'absence d'équipement.
L'utilisateur doit également préciser au programme le nombre de colonnes « équipement » du fichier en entrée et choisir une distance pour le calcul (ie : km pour distance routière kilométrique ; hc pour le temps de trajet en heure creuse et hp en heure de pointe).

Exemple de fichier en entrée :

13001 1 1
13002 1 1
13003 0 0
13004 1 1
13005 1 2
13006 0 0
13007 0 4
13008 0 1
13009 0 0
13010 0 0

En sortie le module produit un fichier *.dat qui contient pour chaque commune et pour chaque équipement le code de la commune équipée la plus proche et la distance à cette commune (codes communaux entre guillemets et champs séparés par des points-virgules) :

dc : code commune
dcequip1 : commune, possédant équip1, la plus proche selon la distance choisie par l'utilisateur (km, hc ou hp)
dist1 : distance moyenne ([aller+retour]/2) correspondant au choix de l'utilisateur (km, hc ou hp) entre dc et dcequip1.
dcequip2 : commune, possédant équip2, la plus proche selon la distance choisie par l'utilisateur (km, hc ou hp)
dist2 : distance moyenne ([aller+retour]/2) correspondant au choix de l'utilisateur (km, hc ou hp) entre dc et dcequip2.

Il n'y a pas de limitation de taille de fichier.

3. Exemples d'utilisation

Dans les exemples qui suivent, ODOMATRIX a été utilisé pour calculer des distanciers ou des accessibilités. Les résultats peuvent être utilisés soit tels quels (cartes d'isochrones par exemple), soit faire l'objet d'une exploitation statistique, notamment après un appariement avec d'autres données (recensement agricole, recensement de population, fichiers administratifs, etc.). La vocation d'ODOMATRIX n'étant pas de faire de la cartographie ou des traitements statistiques, les cartes et les tableaux statistiques présentés ci-après sont réalisés avec d'autres logiciels comme MAPINFO, ARCGIS, EXCEL, SAS, etc.

3.1 Cartes d'isochrones

Les cartes d'isochrones délimitent depuis un point et pour un temps de trajet donné l'ensemble des zones atteintes. De nombreux logiciels « métiers » (planification des transports) ou SIG (système d'information géographique) possèdent des modules pour calculer des isochrones. L'avantage de ODOMATRIX tient au fait que :

- il est capable de calculer quatre distances lors de la même requête (distance à vol d'oiseau, distance routière kilométrique KM, temps de trajet aux heures creuses HC, temps de trajet aux heures de pointes HP) ;
- pour un couple de points (a,b), les distances d(a,b) et d(b,a) n'étant pas nécessairement symétriques, il calcule la moyenne aller-retour ;
- il peut calculer des isochrones monocentriques (depuis un seul point) ou « polycentriques » depuis plusieurs points.

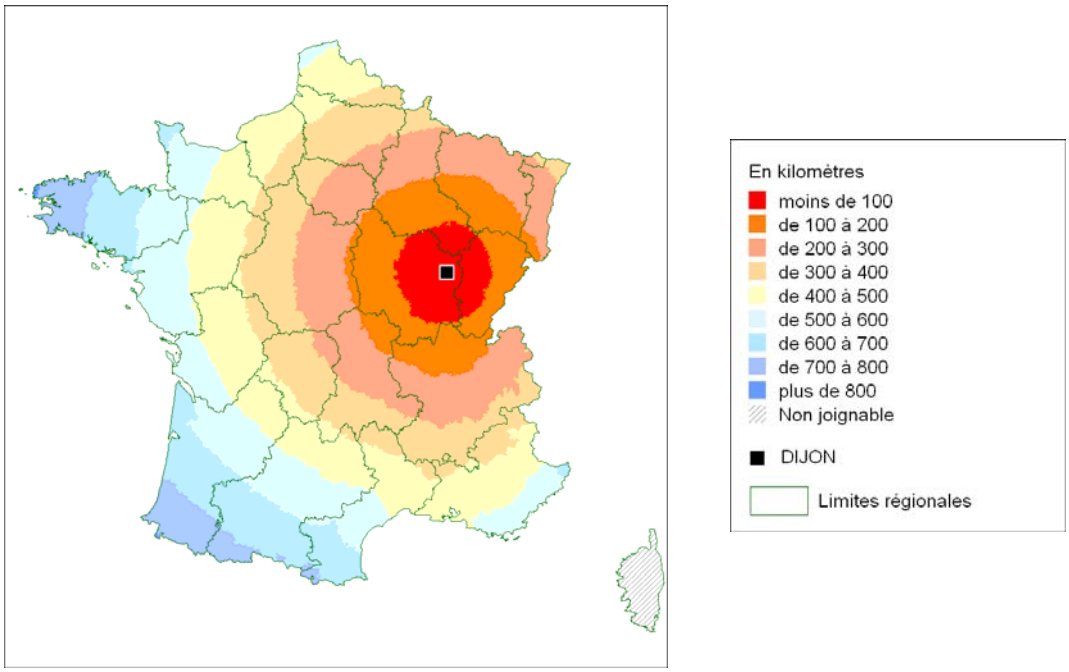
3.1.1. Carte d'isochrones « monocentriques »

Les cartes suivantes utilisent les données produites par le module Distancier (A×B) de ODOMATRIX. Le premier fichier (A) contient les codes INSEE de l'ensemble des communes métropolitaines, le second (B) uniquement le code de la commune origine à partir de laquelle sont calculées les distances, ici Dijon. Le calcul est quasi instantané pour ce type de requête.
Le fichier résultat contient les codes INSEE de l'ensemble des communes métropolitaines (DC_A), le code de Dijon (DC_B), la distance à vol d'oiseau (DVO exprimée en km), la distance routière kilométrique (KM) et les temps de trajet en heure creuse (HC exprimé en mn) et en heure de pointe (HP exprimé en mn). KM, HC et HP sont des distances moyennes aller-retour.

DC_A	DC_B	DVO	KM	HC	HP
01001	21231	130	144	95	105
01002	21231	149	163	104	114
01004	21231	153	164	107	119
…	…	…	…	…	…
95680	21231	272	303	217	236
95682	21231	277	309	221	237
95690	21231	310	346	256	284

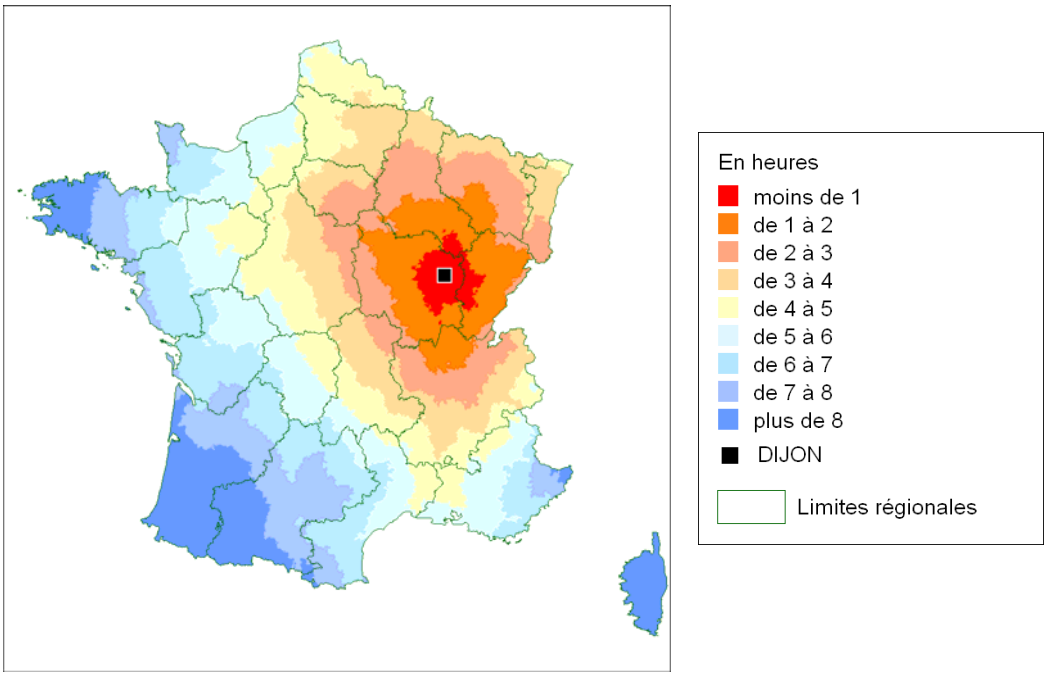
Les cartes, réalisées avec Mapinfo, permettent de comparer les trois indicateurs KM, HC et HP calculés par ODOMATRIX. La carte des distances routières (carte 1) montre des isodistances concentriques depuis Dijon, assez semblables à une carte des distances à vol d'oiseau. Ceci est du à l'homogénéité du réseau routier dans toutes les directions et notamment à la densité de la voierie de desserte locale. Sur les cartes 2 et 3, on voit une déformation des isochrones indiquant des temps de trajet plus faibles, à distance kilométrique identique, en direction de l'Île-de-France, du Nord-Est, et du Midi méditerranéen. La convergence des autoroutes Paris / Méditerranée et Europe du Nord / Méditerranée sur l'axe Dijon-Beaune explique que l'accessibilité Nord-Sud soit meilleure que l'accessibilité Est-Ouest.

Carte 1 : *distances routières depuis Dijon*



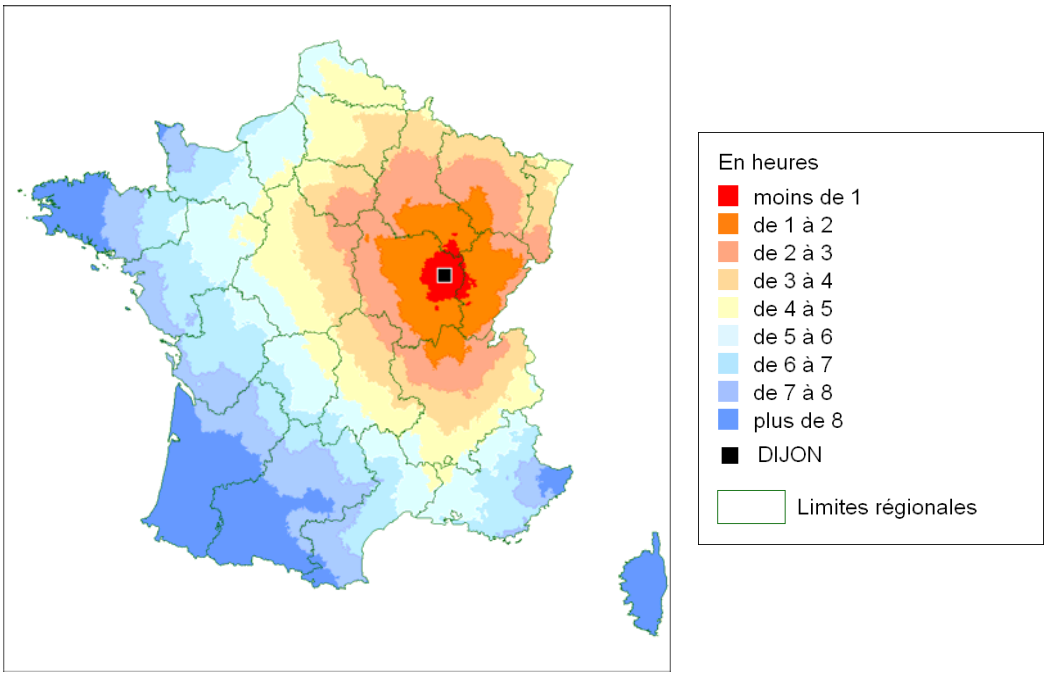
Source : Odomatrix, INRA, UMR1041 CESAER, F-21000 Dijon

Carte 2 : *temps d'accès en heure creuse depuis Dijon*



Source : Odomatrix, INRA, UMR1041 CESAER, F-21000 Dijon

Carte 3 *: temps d'accès en heure pleine depuis Dijon*

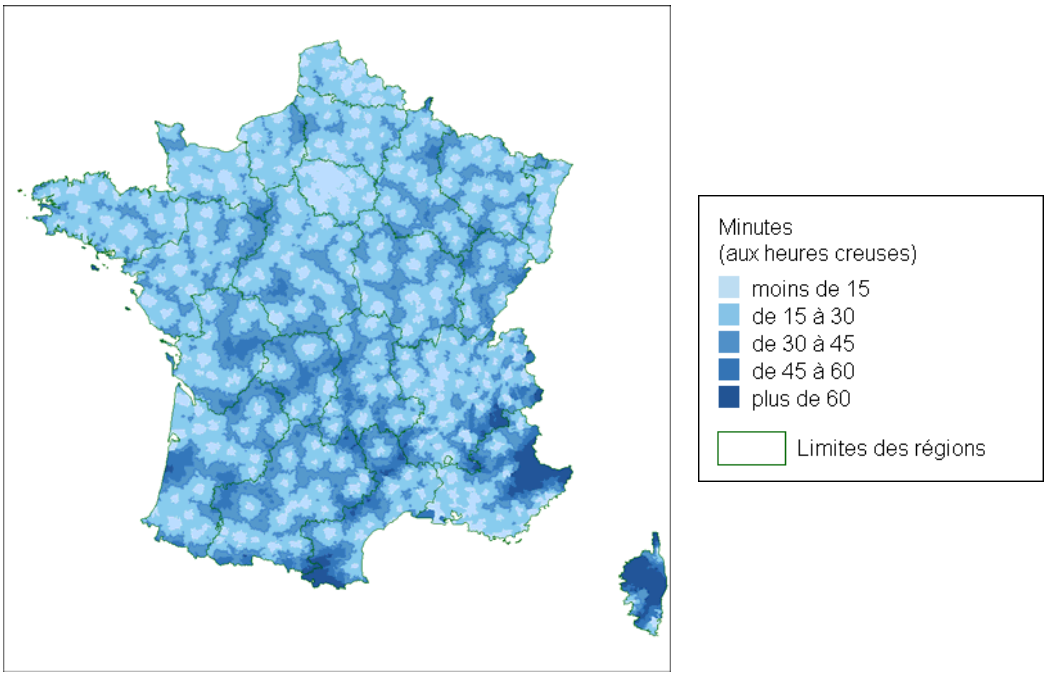


Source : Odomatrix, INRA, UMR1041 CESAER, F-21000 Dijon

3.1.2 Isochrones « polycentriques » : exemple de l'accès à la maternité la plus proche

Les données sont produites par le module Accessibilité (Pôle le plus proche) de ODOMATRIX. Le fichier utilisé en entrée contient : les codes INSEE de l'ensemble des communes métropolitaines (1ère colonne) ; la présence (=1) ou l'absence (=0) de maternité dans la commune (2ème colonne). ODOMATRIX commence par constituer la liste de communes équipées d'une maternité (les pôles), puis il calcule pour chaque commune métropolitaine, équipée ou pas, la distance à l'ensemble des pôles, enfin il retient le code du pôle équipé le plus proche (celui qui a la distance minimale) et la valeur de la distance. Les calculs sont répétés pour les trois distances KM, HC et HP.
Le fichier résultat contient les codes INSEE de l'ensemble des communes métropolitaines (DC) et pour chaque distance le code du pôle le plus proche et la distance (DCKM KM DCHC HC DCHP HP).
Le temps d'accès à la maternité la plus proche (heures creuses) figure sur la **carte 4**. Les zones les plus claires indiquent la présence d'une maternité à moins de 15 minutes. Elles se superposent avec le maillage urbain.

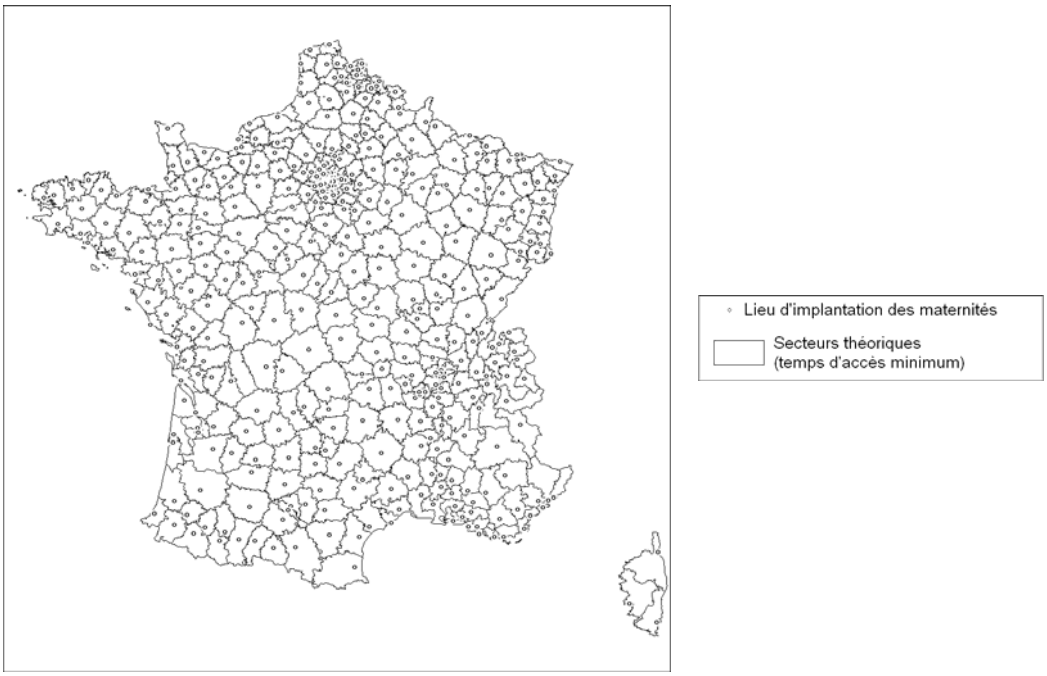
Carte 4 : *temps d'accès à la maternité la plus proche*



Sources : INSEE Base permanente des équipements 2007 ; Odomatrix, INRA, UMR1041 CESAER, F-21000 Dijon

Sur la **carte 5**, chaque commune métropolitaine est rattachée à la maternité dont elle est la plus proche. Le résultat aboutit à une sectorisation, théorique et sans existence administrative, basée sur l'accès au service de maternité.

Carte 5 : *sectorisation selon le temps d'accès à la maternité la plus proche*



Sources : INSEE Base permanente des équipements 2007 ; Odomatrix, INRA, UMR1041 CESAER, F-21000 Dijon

3.2. Exploitations statistiques

ODOMATRIX est principalement utilisé dans des travaux de recherche internes à l'Inra (Hilal, 2004, 2008 ; Aubert *et al.*, 2008 ; Gaigné et Détang-Dessendre, 2009 ; Fall *et al.*, 2009 ; etc.) et dans de nombreux travaux d'études de l'INSEE tant au niveau national que régional. Les exemples suivants sont tirés de deux publications nationales récentes de l'INSEE qui ont été déclinée dans de nombreuses publications régionales. Ils combinent ODOMATRIX avec des données issues des recensements de population, de la base permanente des équipements ou de fichiers administratifs.

3.2.1 Analyse des déplacements domicile-travail

L'exemple suivant est tiré de l'Insee Première n°1129 de mars 2007.
Les données utilisées proviennent des fichiers de Déclarations annuelles de données sociales (DADS) de 2004. La DADS est un document administratif que doit remplir tout employeur des secteurs privé et semi-public ayant rémunéré au moins un salarié au cours de l'année (les non-salariés et les agents de l'État ne font pas l'objet d'une déclaration). Ce document mentionne le lieu de résidence du salarié et l'adresse de son établissement de travail. A partir de ces deux informations, l'INSEE a établi une liste de couples (commune de résidence, commune de travail) qu'il a transmise à l'INRA. Les distances (routières en kilomètres, temps aux heures creuses et pleines) de tous les couples communaux ont ensuite été calculées par ODOMATRIX. Un extrait de la synthèse, sous forme de tableaux et graphiques, de ces résultats fait l'objet de l'encadré suivant.

Extrait de l'Insee Première n°1129 - Mars 2007

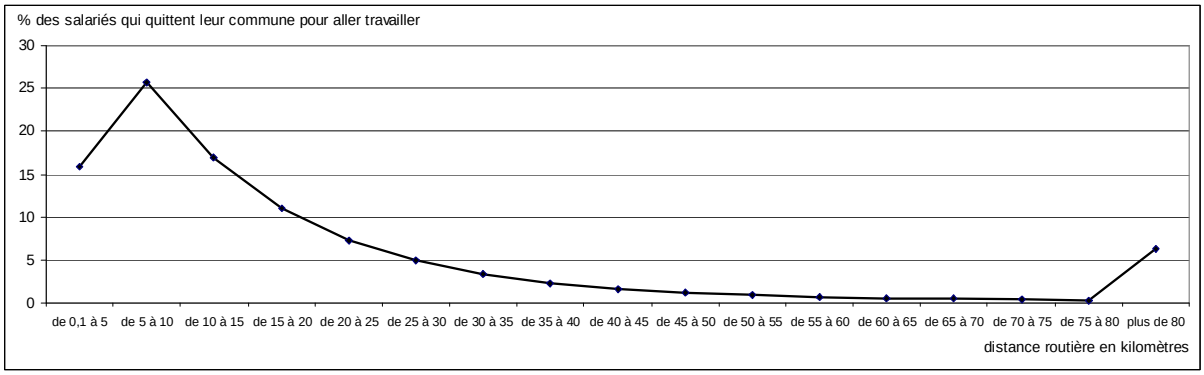
En 2004, près de trois salariés sur quatre travaillent hors de leur commune de résidence. Les actifs qui résident dans les communes périurbaines, moins bien pourvues en emplois que les pôles urbains, sont les plus mobiles : ils travaillent rarement dans leur commune de résidence et font des déplacements plus longs, tant en distance routière qu'en temps de trajet. Les cadres parcourent des distances nettement plus grandes que les autres catégories de salariés.
En incluant les personnes qui résident et travaillent dans la même commune (27 % des salariés), pour lesquels la distance domicile-travail et le temps de trajet sont conventionnellement considérés comme nuls, la distance domicile-travail moyenne est de 25,9 km (**tableau 6** et **graphiques 1 et 2**). Pour la moitié des salariés, la distance est inférieure à 7,9 km. La durée moyenne des navettes domicile-travail, si elles s'effectuaient toutes par la route, serait de 26 minutes en heure creuse et de 32 minutes en heure pleine. La moitié des salariés ont un trajet qui, en heure pleine, prendrait moins de 18 minutes par la route ; à l'autre extrême, pour 10 %, cette durée dépasserait 59 minutes.
En ne considérant que les salariés qui changent de commune ou d'arrondissement (Paris, Lyon et Marseille) pour aller travailler, la distance domicile-travail moyenne passe à 35,4 km et la moitié d'entre eux parcourt moins de 12,2 km (**tableau 6**). Le temps de trajet en heure pleine est de 43 minutes en moyenne par la route, mais il est de 25 minutes pour la moitié d'entre eux.

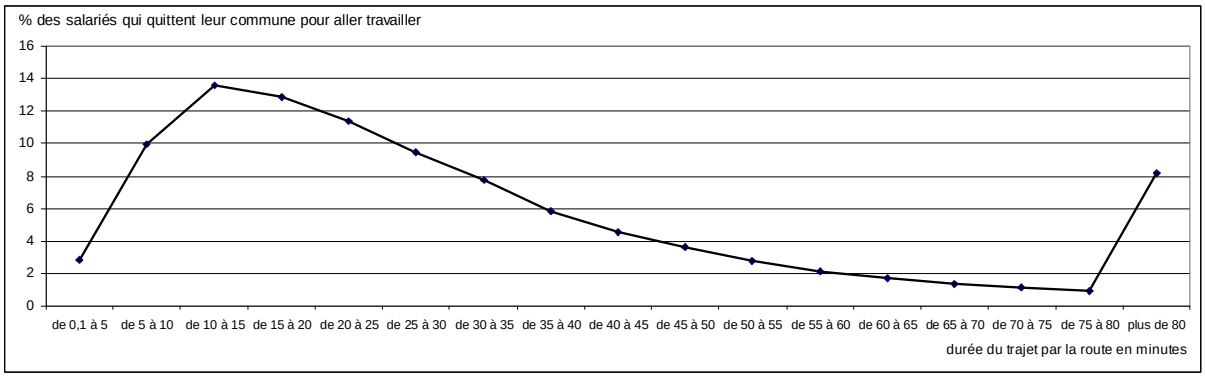
Tableau 6 : *ampleur des navettes selon l'espace de résidence*

	Distance routière (kilomètres)		Temps heure creuse (minutes)		Temps heure pleine (minutes)	
	Moyenne	Médiane	Moyenne	Médiane	Moyenne	Médiane
Navettes intracommunales incluses						
Pôles urbains	23,6	5,8	24,8	12,0	31,5	17,0
Communes périurbaines	30,5	13,6	30,5	18,0	35,0	21,0
Espace à dominante rurale	28,4	10,2	26,1	11,0	27,7	11,0
Ensemble	**25,9**	**7,9**	**26,0**	**13,0**	**31,7**	**18,0**
Navettes intracommunales exclues						
Pôles urbains	34,3	9,7	36,0	18,0	45,7	27,0
Communes périurbaines	35,1	15,6	35,1	21,0	40,2	20,0
Espace à dominante rurale	40,2	17,4	36,8	19,0	39,1	20,0
Ensemble	**35,4**	**12,2**	**35,9**	**19,0**	**43,3**	**25,0**

Sources : INSEE Déclarations annuelles de données sociales 2004 ; Odomatrix, INRA, UMR1041 CESAER, F-21000 Dijon

Graphiques 1 et 2 : *distance et durée par la route des déplacements domicile-travail des salariés migrants alternants en 2004*





Sources : INSEE Déclarations annuelles de données sociales 2004 ; Odomatrix, INRA, UMR1041 CESAER, F-21000 Dijon

Les salariés résidant dans les pôles urbains travaillent en moyenne à 23,6 km de chez eux, soit plus près que ceux des zones périurbaines (30,5 km) ou de l'espace à dominante rurale (28,4 km). Dans les pôles urbains, en effet, une proportion plus importante d'individus travaille dans leur commune de résidence. En ne retenant que ceux qui vivent et travaillent dans deux communes différentes, les distances parcourues par les habitants des pôles urbains et des couronnes périurbaines deviennent très proches (34,3 km et 35,1 km) et ce sont ceux qui résident dans l'espace rural qui parcourent les plus longues distances.
En raison d'une vitesse de circulation plus réduite dans les zones urbaines que dans le périurbain ou l'espace rural, les écarts de temps de trajet, d'un type d'espace à l'autre, sont plus faibles que ne le sont les écarts de distance. Ainsi, en heure pleine, la durée des trajets pour les salariés domiciliés dans l'espace rural est en moyenne inférieure à celle des résidants des pôles urbains (28 mn contre 32 mn), alors qu'ils parcourent des distances significativement plus longues.

3.2.2 Accessibilité aux commerces et services

Dans l'exemple suivant, tiré d'une publication nationale de l'INSEE (Le commerce en France - Édition 2009, Insee Référence – janvier 2010), l'auteur analyse l'offre de commerce de détail en France. Grâce à ODOMATRIX, elle montre que dans l'espace à dominante rurale, la population n'est pas trop éloignée des commerces, notamment des commerces alimentaires.
Les travaux sur l'accès aux commerces et services utilisent le module Accessibilité (Communes équipées les plus proches) de ODOMATRIX. Le fichier utilisé en entrée contient en première colonne les codes INSEE des communes suivis de plusieurs colonnes décrivant chacune le nombre d'équipement présent (0 pour l'absence). ODOMATRIX constitue, pour chaque colonne-équipement, la liste de communes équipées ; il calcule ensuite pour chaque commune (de la première colonne), équipée ou pas, la distance à l'ensemble des communes équipées, enfin il retient le code de la commune équipée la plus proche et la distance à cette commune. L'utilisateur ayant au préalable choisi une distance (parmi KM, HC et HP), les calculs ne sont effectués qu'une seule fois.
Le fichier résultat contient les codes INSEE de l'ensemble des communes métropolitaines (DC) et pour chaque équipement le code de la commune équipée la plus proche et la distance (DCEQ1 DISTEQ1 DC… DIST… DCEQn DISTEQn).
L'encadré suivant présente un extrait de cette étude.

Extrait de *Le commerce en France - Édition 2009, Insee Référence – décembre 2009*

En zone rurale, l'accès aux commerces est plus ou moins aisé selon l'équipement considéré (**tableau 7**). Seul moins de 1 % de la population de l'espace à dominante rurale n'a pas accès en moins de quinze minutes en voiture, à une épicerie-supérette, à une boulangerie-pâtisserie, à une boucherie-charcuterie ou à une pharmacie. En outre, 4,7 % de la population de l'espace à dominante rurale n'a pas accès à un supermarché en moins de quinze minutes. En revanche, 31,5 % de la population de l'espace rural n'accède pas à un commerce de produits surgelés en moins de trente minutes.
Le constat est similaire pour les commerces non alimentaires. Les magasins de vêtements sont accessibles rapidement par une grande majorité de la population de l'espace à dominante rurale : 4,4 % de la population de cette zone y accède en moins de quinze minutes. Les boutiques d'horlogeries-bijouteries, les magasins de meubles ou d'équipement du foyer ne sont pas accessibles rapidement pour plus de 16 % de la population de l'espace à dominante rurale ; ce taux est proche de 13 % pour les magasins de chaussures et les magasins d'articles de sports et de loisirs.

Tableau 7 : *part de la population selon l'éloignement des commerces dans l'espace à dominante rurale en 2007*

Éloignement	Commerce	Part de la population (%)
+ de 15 minutes en voiture	**Alimentaire**	
	Boulangerie, pâtisserie	0,2
	Alimentation générale, supérette	0,4
	Boucherie, charcuterie	0,5
	Supermarché	4,7
	Pharmacie	
	Pharmacie	0,6
	Non alimentaire (hors pharmacie)	
	Librairie, papeterie	2,5
	Fleuriste	3,0
	Quincaillerie, bricolage	4,1
	Magasin de vêtements	4,4
	Magasin d'électroménager	7,6
	Sport et loisirs	12,9
	Magasin de chaussures	13,9
	Magasin de meubles	16,9
	Horlogerie, bijouterie	19,7
	Magasin d'équipement du foyer	22,0
+ de 30 minutes en voiture	Parfumerie	3,9
	Poissonnerie	13,7
	Hypermarché	17,2
	Produits surgelés	31,5

Champ : France métropolitaine.
Lecture : 3 % de la population de l'espace à dominante rurale accède en moins de 15 minutes en voiture à un magasin de fleurs.
Sources : Insee, base permanente des équipements 2007 et recensement de la population 2006 ; Inra, UMR1041 CESAER ; Distancier Odomatrix.

Conclusion

La question de l'accès aux services de proximité et aux emplois nourrit depuis longtemps les débats régionaux et nationaux d'aménagement du territoire. L'enjeu est double. D'une part, le dynamisme économique et social d'un territoire est conditionné par l'attrait que celui-ci peut exercer sur les individus, attrait lié à la proximité de commerces et de services et aux opportunités d'emplois. D'autre part, les conditions d'accès des populations résidentes aux emplois et à tout un ensemble d'équipements et de services constituent une dimension essentielle de l'équité territoriale. Pour répondre à cette question, l'outil ODOMATRIX est mobilisé dans différentes opérations de recherche au sein de l'unité et dans les travaux d'études de l'INSEE tant au niveau national que régional. La valorisation de cet outil est en cours, en lien avec INRA Transfert, afin de répondre aux très nombreuses demandes d'utilisation émanant de la sphère publique (ministère de l'Écologie, de l'Énergie, du Développement durable et de la Mer ; ministère de l'Alimentation, de l'Agriculture et de la Pêche ; ministère de la Santé et des Sports ; Délégation interministérielle à l'aménagement du territoire et à l'attractivité régionale ; Collectivités territoriales ; Chambres consulaires ; Caisses régionales d'assurance maladie ; Agences régionales d'hospitalisation ; La Poste ; Réseau ferré de France, etc.) et privée (Sociétés d'autoroute, bureaux d'études, etc.).

Bibliographie

Aubert F., Dissart J.-C., Lépicier D. (2008) Localisation des services résidentiels. Analyse de la territorialisation de l'économie résidentielle à l'échelle intra-métropolitaine, Rapport d'étude pour la DIACT, septembre, 106 pages.

Baccaïni B., Sémécurbe F., Thomas G. (2007) Les déplacements domicile-travail amplifiés par la périurbanisation. Insee Première n° 1129, Mars 2007. http://www.insee.fr/fr/ffc/ipweb/ip1129/ip1129.pdf (consulté le 11 mai 2010)

Détang-Dessendre C., Gaigné C. (2008) Unemployment duration, city size and the tightness of the labor market., Regional Science and Urban Economics, 39(3): 266-276.

Dijkstra, E. W. (1959) A note on two problems in connexion with graphs. Numerische Mathematik 1: 269–271. http://www-m3.ma.tum.de/twiki/pub/MN0506/WebHome/dijkstra.pdf (consulté le 11 mai 2010)

Fall M; Hilal M., Selod H. (2007) Les déterminants locaux du chômage en France : le rôle de la faible densité d'emplois en zone rurale. Lea-WP 0702.

Fredman M. L. et Tarjan R. E. (1987) Fibonacci heaps and their uses in improved network optimization algorithms. Journal of the ACM 34(3) : 596-615.

Hilal M., Schmitt B., 1997 Les espaces ruraux : une nouvelle définition d'après les relations villes campagnes. INRA-Sciences Sociales, 1997(5) : 1-6.

Hilal, M. (2004) Accessibilité aux emplois en France : le rôle de la distance à la ville, Cybergeo, Revue Européenne de Géographie, 293, pp 1-15.

Hilal M. (2007) Temps d'accès aux équipements au sein des bassins de vie des bourgs et petites villes, Economie et Statistiques, 402, pp 41-57.

Solard G. (2009) A la campagne comme à la ville, des commerces traditionnels proches de la population in *Le commerce en France - Édition 2009, Insee Référence* – janvier 2010 & Insee Première n° 1245, Juin 2009. http://www.insee.fr/fr/publications-et-services/sommaire.asp?codesage=COMFRA09&nivgeo=0 (consulté le 11 mai 2010) http://www.insee.fr/fr/ffc/ipweb/ip1245/ip1245.pdf (consulté le 11 mai 2010)

Vallès V. (2002) Organisation territoriale de l'emploi et des services. Insee Première n°870, Novembre 2002. http://www.insee.fr/fr/ffc/docs_ffc /ip870.pdf (consulté le 11 mai 2010)

Couplage simple entre système d'information géographique et modèle multi-agents

Annie Hofstetter[1]

Résumé : *Cet article traite l'aspect technique d'un couplage simple entre un système d'information géographique (SIG) et un système multi-agent (SMA). Pour situer le contexte, nous faisons le point sur les données disponibles et le projet afin de comprendre pourquoi préférer utiliser un fond cadastral dans un système de simulation. Ensuite nous abordons la réalité du couplage à travers la phase d'initialisation du modèle multi-agent, en prenant quelques exemples de traitements spécifiques tels que le voisinage des entités spatiales élémentaires.*

Mots clés : modélisation, paysage, politiques publiques, couplage, SMA, SIG

Introduction

Le projet d'ensemble vise à construire un modèle pour analyser l'impact des politiques publiques sur la dynamique de végétation.
Le problème de modélisation est abordé comme un système complexe qui prend en compte la diversité des acteurs, les différentes interactions sur l'environnement physique et/ou social ou la hiérarchie spatiale. La mise en œuvre est conduite grâce à des outils adaptés à la représentation des relations agents-environnement hétérogènes que sont les simulateurs multi-agent (SMA), qui autorisent la représentation des interactions et de leurs effets sur la dynamique d'un système.
Pour initialiser la configuration spatiale, le couplage simple consiste à alimenter ce modèle à l'aide de données provenant d'un système d'information géographique (SIG). L'originalité de ce couplage réside surtout dans le fait d'utiliser le fond cadastral ainsi que les données associées.

1. Contexte

1.1 Cadre de l'analyse

Ce travail s'inscrit dans un programme[2] de recherches conduites au LAMETA (Laboratoire montpelliérain d'économie théorique et appliquée) du centre Inra de Montpellier, qui portait sur l'analyse de la dynamique naturelle de la végétation.
À la base, nous nous sommes inspirés du modèle ALAMO[3] Aui décrit les interactions entre les activités humaines telles que l'agriculture ou la forêt, et la dynamique naturelle de la végétation. Notre modèle vise à analyser et à évaluer l'impact des politiques publiques sur la dynamique des paysages.

---

1 INRA UMR 1135 LAMETA -Laboratoire montpelliérain d'économie théorique et appliquée- F-34060 Montpellier ☏ 04 99 61 24 99 ✉ annie.hofstetter@supagro.inra.fr

2 Cette recherche sur l'impact des politiques sur la dynamique des paysages au sud du Massif central a été financée par le ministère de l'écologie et du développement durable et s'inscrit dans le programme *Politiques publiques et paysages : analyse, évaluation, comparaisons*.

3 Agricultural Landscape Model, R. Lifran

L'objectif est de représenter la diffusion de l'innovation et de la végétation dans une communauté d'agriculteurs afin de comprendre l'impact des politiques publiques sur la dynamique des paysages une fois l'espace structuré. Nous tenterons d'observer le rythme d'embroussaillement en prenant en compte les différents rythmes d'innovations et les adaptations à l'application d'une politique publique. Un modèle de type multi-agents nous apporte une réponse car il existe des interactions dans tous les sens.
Historiquement, des enquêtes sur le terrain ont permis de construire une réflexion sur la dynamique du paysage. De précédentes approches ont conduit à une représentation géographique du paysage au niveau des lieux-dits définis comme unité de base structurant également la population. Le lieu-dit est un espace utilisé par une population regroupée dans un hameau.
Le polygone de Thiessen décrit le périmètre d'action autour d'un lieu-dit auquel il est rattaché. Il établit le lien entre le hameau et son espace. Chaque lieu-dit peut être caractérisé par un nombre de troupeaux, une population et un espace exploité. Il existe des terrains sectionnaux collectifs dont la description est exprimée en ces termes : *la propriété collective des habitants du hameau*.
Autour de cette définition, la structure sociale renforce le lieu-dit comme territoire d'exploitation.

Le cadre de l'étude est le Causse du Sauveterre où la production du lait de brebis est principalement destinée à la production fromagère. Sur cette zone nous pouvons observer :

- une concentration des élevages qui conduit à un abandon plus ou moins durable de fractions de territoires et qui entraîne une diminution de la population ;
- la modification des modes d'alimentation des troupeaux qui semble à l'origine du développement du boisement naturel ou artificiel. L'amélioration génétique a entraîné l'obtention de races plus sensibles à l'alimentation. Afin de répondre à ce nouveau besoin, l'intensification a dû se faire en se concentrant sur les meilleures terres au détriment des landes moins productives.

La dynamique du paysage est ainsi respectivement rattachée à la dynamique de la population d'une part, et à la dynamique des techniques et des pratiques, d'autre part.
Nous évoquerons la notion de transect qui représente une sélection des lieux-dits sur le Causse. Ce transect s'étend de l'ouest à l'est et il couvre une partie des communes de Saint-Georges-de-Lévejac, La Malène, La Canourgue, Laval-du-Tarn, Sainte-Enimie et Ispagnac. Cette zone d'étude est constituée de 10 000 parcelles sur 20 000 hectares.

Par ailleurs, les politiques publiques peuvent être définies selon trois niveaux :

- les politiques publiques agissant directement sur la dynamique des paysages ;
- les politiques publiques agissant sur les pratiques d'utilisation de l'espace (utilisation ou non des parcours, pratiques d'intensification de l'élevage, etc.) ;
- les politiques publiques d'aménagement agissant sur la redéfinition des usages de la propriété collective ; selon les communes, les sectionnaux peuvent représenter jusqu'à un quart du territoire.

Une caractéristique importante du modèle est le pas de temps sur lequel repose le processus de diffusion de la végétation. Grâce à la photo-interprétation, nous pouvons traduire la dynamique du paysage et son évolution sur une période suffisamment pertinente.

De décennies en décennies apparaissent de nouvelles innovations, telles que l'alimentation concentrée entraînant l'utilisation des céréales sur les meilleures terres : les parcours alors moins utilisés, la broussaille s'installe. On comprend que l'intensification porte à la fois sur la gestion de cet espace et sur la gestion des troupeaux.

Encadré 1 : le Causse de Sauveterre

Le Causse de Sauveterre est l'un des deux grands causses de Lozère avec le Causse Méjan. D'une superficie totale de l'ordre de 30 000 hectares, il s'étend entre les deux grandes rivières de ce département en s'abaissant au nord-ouest vers le Lot et au sud vers le Tarn. Au milieu de cette masse calcaire, le plateau présente de petites dépressions, les dolines, où sont cultivées céréales et cultures fourragères. Aménagées avec des pavés elles constituent également les lavognes où s'abreuvent les animaux. Le plateau montre aujourd'hui un paysage végétal profondément modifié par la vie pastorale ; la forêt originelle a laissé la place à de vastes pelouses et à des landes plus ou moins piquées de buis et de genévrier. Des futaies de pins noir et sylvestre sont visibles surtout dans la partie occidentale. Ce plateau est fortement marqué par les activités d'élevage ovin. Par ailleurs, on dénombre des habitations dispersées et des hameaux qui témoignent, malgré l'exode rural, d'une activité agricole encore importante.

Sur des milliers d'hectares, le Causse de Sauveterre ne possède que très peu d'éléments d'artificialisation notables hormis quelques zones urbanisées et des cultures. Le caractère très dispersé de ses activités procure au site un aspect naturel marqué. Il constitue une entité paysagère originale et pittoresque : pelouses rares s'étendant à l'infini, vastes espaces déserts vallonnés, paysage ruiniforme dolomitique, etc.

La principale menace de dégradation repose sur les boisements qui, multipliés, modifieraient complètement la perception du paysage et annuleraient l'aspect désertique et ouvert qui caractérise et qui fait le charme du Causse.

1.2 Les données du SIG : analyser et décrire la dynamique des paysages

Un précieux travail d'interprétation a été réalisé sur la zone d'étude à partir des photos aériennes de 1963, 1977, 1989 et 2000[4]. Nous avons d'abord identifié les éléments structurants (champs, pelouses, parcours, bois) et leurs changements au cours des quatre dernières décennies. Puis nous avons identifié le rôle fondamental d'une gestion de l'espace par les habitants des hameaux. Ce rôle est lié aux contraintes posées par la production laitière. Les premières analyses montrent que la dynamique de l'embroussaillement a débuté avec les territoires sectionnaux. Les transformations du paysage recouvrent un double mouvement de progression des boisements et de réouvertures, combinées selon des logiques propres à chaque hameau. L'existence d'une grande propriété foncière collective dans chaque hameau et la façon dont les habitants gèrent ses usages est fondamentale pour comprendre ces logiques.

1.3 L'originalité du SMA : structure hiérarchique des entités

Selon Ferber, un système multi-agents se définit par plusieurs éléments dont l'espace représenté par les entités spatiales. Dans notre modèle, outre le fait que la grille spatiale repose sur la grille cadastrale issue du SIG, l'originalité tient à la hiérarchie spatiale de ces entités. La parcelle cadastrale représente l'entité spatiale élémentaire. Le lieu-dit est une agrégation de parcelles élémentaires. Il connaît ses parcelles. L'espace communal est une

4 Les techniques classiques de juxtaposition des photographies aériennes ont été utilisées grâce aux fonctionnalités d'orthorectification lors de la superposition des photos sur le fonds cadastral.

agrégation de lieux-dits et il connaît ses lieux-dits. Et enfin la zone d'étude représente l'ensemble des parcelles.
Le modèle met en jeu des agents qui réagissent et qui interagissent ; toutefois il n'y a pas de véritable dynamique de la population (les agents ont une vie simple et il n'y a pas de modèle démographique). Seule la réaffectation des parcelles lorsqu'un agent meurt, traduit une certaine dynamique de l'espace. De fait l'*habitant* pérenne conserve ses attributs et se renouvelle. Parallèlement à la structure hiérarchique des entités spatiales, les agents sont structurés en groupes au niveau des lieux-dits ainsi qu'au niveau des communes.

2. Le couplage simple

3.1. Aspects techniques

SIG : Arcview sur PC sous Windows 98
SMA : Cormas[5] sur PC sous Linux Mandrake
La plateforme de traitement étant différente, les fichiers d'échange doivent être au format ASCII pour Unix
Depuis le SIG, l'échange se fait par la procédure d'export au format MIF/MID[6] via MapInfo, puis le chargement de la grille spatiale s'effectue sur Cormas.

3.2. Initialisation

Dans le cas d'une configuration spatiale fournie par un SIG, Cormas travaille sur des entités spatiales vectorielles que sont les polygones. Une grille régulière est souvent utilisée dans les modèles multi-agents soit pour représenter des automates cellulaires soit pour simplifier la représentation spatiale. Nous avons malgré tout voulu tester une grille régulière de 95 × 110, c'est-à-dire le nombre quasi identique de nos parcelles initiales. Se pose alors principalement le problème de la représentation virtuelle pour laquelle le front de progression de l'embroussaillement ne peut plus s'expliquer d'ouest en est. Par ailleurs, les voisins ne sont plus ceux que l'on visualise, soit parce qu'on prend un voisinage constant et identique de 4 ou 8 cellules, soit parce que la représentation de la véritable liste des voisins ne correspond plus à ce que l'on observe sur une grille régulière. La lecture des résultats était quant à elle proche de celle obtenue sur une grille irrégulière.
Dans notre cas un couplage simple suffit car il n'y a pas de retours dans le SIG, d'une part il n'y a pas de mélange entre les données observées et les données simulées, d'autre part le chargement de la grille spatiale proche du cadastre dans le SMA permet de rendre compte directement des résultats de simulations. Ce principe économise également les ressources en calcul lors des échanges dans le cas d'un couplage dynamique.
Le fichier MIF contient les polygones (**encadré 2**), tandis que le fichier MID contient les données attributaires (**encadré 3**). Le fichier COR assure la correspondance entre le fichier de données et le SMA (**encadré 4)**.

5 Plateforme de simulation multi-agent du CIRAD : http://cormas.cirad.fr
6 Le format MIF/MID (Mapinfo Interchange Format/ Mapinfo Interchange Data) est un format natif de Mapinfo qui permet d'échanger des données graphiques et des données attributaires (non graphiques).

Encadré 2 : extrait du fichier sauveterre.mif

```
Version 300
Charset "WindowsLatin1"
Delimiter ","
CoordSys Earth Projection 1, 104
Columns 23
  Parc_id Char(16)
  Surface Decimal(16, 0)
  Veg_63 Decimal(16, 0)
  Nb Char(16)
  Veg_77 Decimal(16, 0)
  Veg_89 Decimal(16, 0)
  Nb77 Char(16)
  Nb89 Char(16)
  Nb00 Char(16)
  Veg_00 Decimal(16, 0)
  transect Char(3)
  voisin Char(254)
  Parc_id48 Char(12)
  Area Decimal(16, 14)
  Prcnt_tot Decimal(16, 14)
  Hameau_id Char(13)
  hamo_id Char(10)
  Hm_appart Char(16)
  hm_affect Char(10)
  Surf Decimal(16, 0)
  Nom Char(20)
  Collectif Char(12)
  Cah Decimal(16, 0)
Data

Region  1
  31
8.380904 17.373444
8.38081 17.373422
8.380423 17.373565
8.379978 17.373705
8.380013 17.374071
8.380127 17.374168
8.380203 17.374314
8.380173 17.37443
8.380101 17.374419
8.379935 17.374104
8.379847 17.37385
8.379742 17.373735
8.37944 17.373754
8.379082 17.373869
8.378938 17.373857
8.378725 17.373725
8.378282 17.373802
8.377939 17.373985
8.377827 17.374131
8.377761 17.374547
8.377908 17.374605
8.377858 17.374621
8.377649 17.376072
8.378045 17.3761
8.378096 17.376021
8.378325 17.37604
8.378547 17.376025
8.379527 17.376187
8.380034 17.376386
8.38031 17.375907
8.380904 17.373444
   Pen (1,2,0)
   Brush (2,16777215,16777215)
   Center 8.379277 17.374904
```

Encadré 3 : extrait du fichier sauveterre.mid

```
"48154A0001",80815,7,"",7,3,"","défriché","",6,"oui","154A0001 154A0004 154A0002 154A0003 154A0021
154A0006 154A0005 154A0008 154A0007 154A0412 154A0413 154A0414 154A0405 154A0403 154A0415
154A0382_B","154A0001",0.00000000000000,0.84166345482138,"PIG15460","","PIG15460","PIG15460",38
16810,"LA PIGUIERE","SECTX",1
"48154A0004",1226,5,"",3,3,"","","",4,"oui","154A0001 154A0004 154A0006 154A0005 154A0008
154A0007","154A0004",0.00000000000000,0.84166345482138,"PIG15460","","PIG15460","PIG15460",38168
10,"LA PIGUIERE","SECTX",1
"48154A0002",2510,7,"",7,7,"","","",7,"oui","154A0001 154A0002 154A0003 154A0005 154A0008 154A0411
154A0412 154A0413 154A0414 154A0405 154A0415
154A0382_B","154A0002",0.00000000000000,0.84166345482138,"PIG15460","","PIG15460","PIG15460",38
16810,"LA PIGUIERE","",1
```

La **figure 1** illustre le chargement de la grille spatiale dans Cormas. Nous avons choisi le point de vue du couvert dominant en début de période qui initialise le fond cadastral avec l'état de la végétation observée en 1963 correspondant au début de nos simulations.

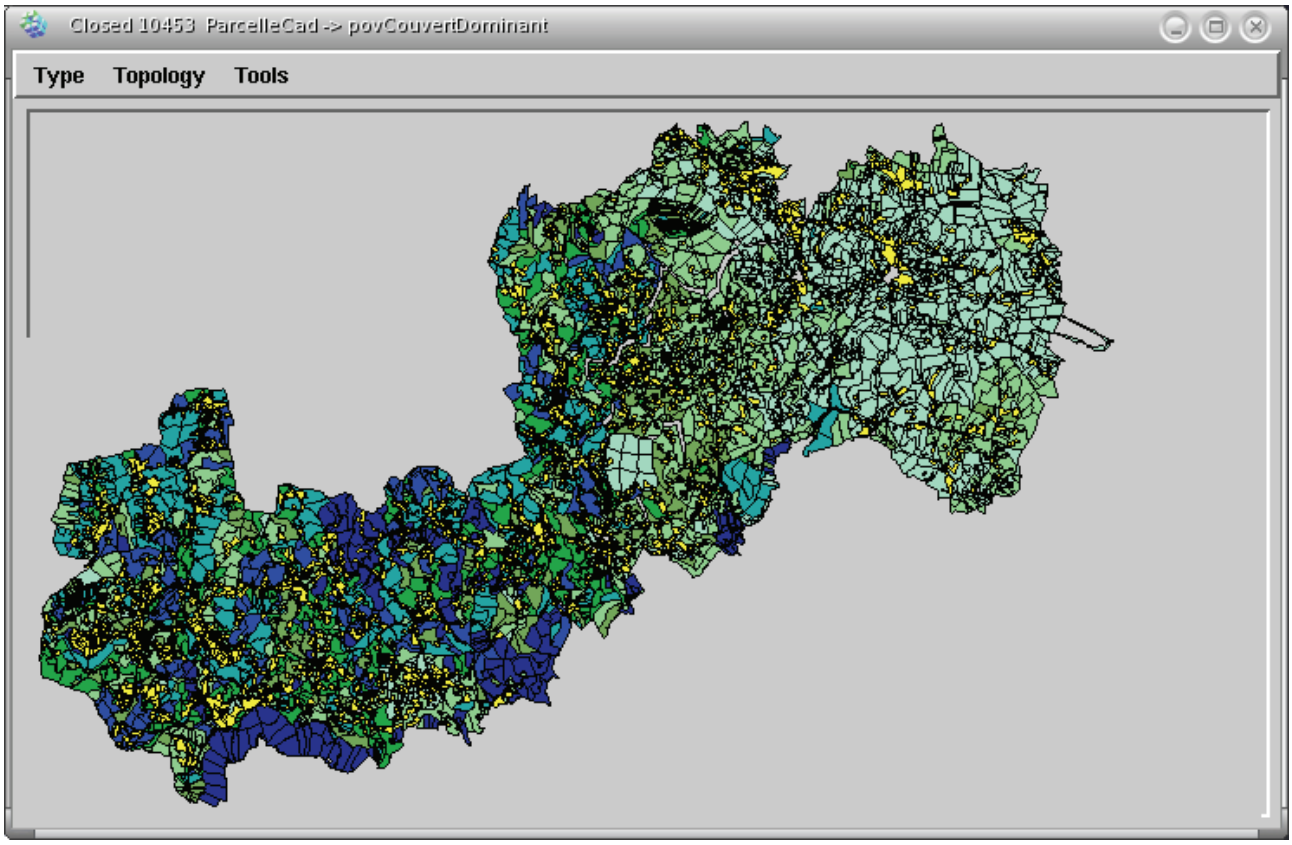


Figure 1 *: copie d'écran de Cormas.*
Point de vue sur l'attribut couvertDominant *à la suite de l'initialisation de la grille spatiale*

Encadré 4 : fichier sauveterre.cor

parcelleCadId	1	asString
surfaceCad	2	asNumber
couvertDominant	3	asNumber
lieuditAffect	19	asString
typeJuridique	22	asString

3.3. Traitement spécifique du voisinage

Une grille régulière sans couplage aurait implicitement initialisé les voisins de chaque cellule géographique élémentaire que constitue la parcelle. Le chargement d'une grille irrégulière ne s'effectue avec l'initialisation du voisinage que lorsqu'on le demande au moment de l'initialisation. Il est d'usage d'initialiser la grille, de calculer le voisinage puis de le sauvegarder. Le chargement des voisins au sens des voisins contigus calculés une première fois dans cormas sera utilisé à l'initialisation de la grille spatiale à chaque lancement de l'application par des méthodes appropriées (**encadrés 5 et 6**).

Encadré 5 : méthodes de traitement des voisins au niveau du modèle

```
saveNeighboors
	self spaceModel saveNeighboorsClass: ParcelleCad separator: $;
loadNeighboors
	self spaceModel loadNeighboorsClass: ParcelleCad separator: $;
```

Encadré 6 : exemple d'une méthode d'initialisation des instances de type parcelle

```
initParcelleCadsHeterogeneIEM
	|  stream line item tmp parc |
	self initCells: #init.
	self theParcelleCads do:
			[:c |
	"general"
			c modif1900.
			c initIndiceEmbroussaillementHomogene.
			c initValeurParcelle.
	"pature"
			c initTxPrelevementHasard.
			c initTxPrelevable.
	"bois de chauffage"
			c initTxPrelevementBois.
			c initTxConsumable.

			c initBiomasseTotale.
			c initBiomassePaturable.
			c initBiomasseConsumable].
	self loadNeighboors.
	"suite de l'initialisation des parcelles avec le fichier ascii parcelle.txt"
	stream :=  ((Cormas dataPath: self class name) construct: 'parcelle.txt' ) readStream.
	[stream atEnd] whileFalse:
			[line := (stream upTo: Character cr) readStream.
			tmp := OrderedCollection new.
			[line atEnd] whileFalse:
					[item := line upTo: $;.
					tmp add: item].
			parc := self theParcelleCads detect:[:cell| cell parcelleCadId = (tmp at:1) asString]
	ifNone:[nil].
			parc isNil ifFalse:[
			parc perimetre: (tmp at:2) asNumber.
			parc exploitation: (tmp at:3) asString.
			parc nbVoisinRayon: (tmp at:4) asNumber.
			parc distanceHameau: (tmp at:5) asNumber]].
	stream close.
	"Dialog warn: 'fin init parcelles'."
```

Pour compléter la définition du voisinage par rapport aux traitements que nous voulions effectuer, des scripts ont été écrits en *Avenue*, un langage semi compilé orienté objet spécifique à Arcview. En effet, deux parcelles séparées par un chemin ne sont pas considérées comme voisines si elles ne présentent aucune contigüité. Les scripts nous ont permis de prendre en compte le voisinage dans un rayon d'action autour du centre de gravité de chaque parcelle, et aussi, de contourner la difficulté liée au nombre de parcelles voisines lorsqu'on est dans un fond de vallée avec une multitude de petites parcelles ou de parcelle incluse dans une autre dans le cas d'une doline.

Conclusion

Dans un premier temps le choix d'un couplage nous permet de garder les données issues du système d'information géographique, choix d'autant plus pertinent qu'il s'agit de données réelles.
Bien qu'une grille irrégulière ne soit pas souvent utilisée dans les systèmes multi-agents, nous l'avons privilégiée à la grille régulière plus théorique quant à l'interprétation des résultats. Lorsque les données ne sont pas amenées à être modifiées dans le système d'information géographique, le couplage simple est facile à gérer lors de l'initialisation de la grille spatiale du système multi-agents. Dans le cas de notre modélisation, il représente un bon compromis en offrant la possibilité de visualiser directement les résultats des simulations lors des retours sur le terrain aux côtés des décideurs locaux, tout en optimisant les calculs intermédiaires sans échanges inutiles avec le système d'information géographique.

Bibliographie

Bommel, P., Lardon, S. (2000) Un simulateur pour explorer les interactions entre dynamiques de végétation et de pâturage. Impact des stratégies sur les configurations spatiales. *Géomatique* **1**(1) : 1-10.

Chassany J.-P. (1989). L'élevage ovin caussenard face aux marchés (1945-1985) : atouts et faiblesses actuels. *Annales du Parc national des Cévennes,* **4** : 55-89.

Chassany J.-P., Crosnier C., Cohen M., Lardon S., Lhuillier C., et Osty P.-L. (2002) Réhabilitation et restauration de pelouses sèches en voie de fermeture sur le Causse Méjan : Quels enjeux pour une recherche en partenariat ? Revue d'écologie (Terre et Vie), pp. 31-49

Lardon, S. Osty P.-L. (2003) Les éleveurs et leurs impacts sur le paysage. *Politiques publiques et dynamiques des paysages au sud du Massif central.* R. Lifran. Montpellier, INRA, UMR LAMETA **:** 46-54

Lepart, J., Marty P. *et al*. (2000) Les conceptions normatives du paysage. Le cas des grands causses. *Natures Sciences Sociétés,* **4** : 16-25

Lifran, R., Hofstetter A., Bommel P. (2003) Politiques publiques et dynamiques des paysages : analyse de leurs rapports par un modèle multi-agents spatialisés. Politiques publiques et dynamiques des paysages au sud du Massif central. Montpellier, INRA-UMR LAMETA : 110-164.

Lifran R., Hofstetter A. (2002) Atlas paysager du Causse de Sauveterre, INRA-UMR LAMETA, 36p.

Lifran, R., Editeur (2003) Politiques publiques et dynamiques des paysages au sud du Massif central. Montpellier, INRA-UMR LAMETA **:** 168p.

Lifran R., Hofstetter A. (2009) Quand les politiques publiques se heurtent au temps du paysage. *In:* Les Grands Causses, terre d'expériences. Chassany, J.-P., C. Crosnier, Florac, PNC, 2009 : 309-315

Marres P. (1935) Les Grands Causses : étude de géographie physique et humaine. Tours, Arrault.

MEDINA, une interface WEB de consultation de bases de données

Monique Harel [1], Cécile le Roy[1]

Résumé : MEDINA (Marchés extérieurs des industries agroalimentaires) *regroupe des bases de données implantées sous le SGBDR PostgreSQL localisé au centre Inra de Nantes et utilisées par les chercheurs du laboratoire LERECO. L'idée de créer un outil pour accéder facilement aux données et pour rendre l'utilisateur plus autonome est née de l'accroissement du volume des données et des demandes diverses et répétées d'extraction des bases.*

Mots-clefs : commerce international, bases de données, interface Web, PHP, PostgreSQL

Introduction

L'outil informatique *MEDINA (*Marchés extérieurs des industries agroalimentaires) est indispensable à la bonne conduite de nombreux travaux de recherche sur le commerce international. Cette gestion de données en amont de l'analyse statistique constitue une étape importante et incontournable dans le processus de recherche. Ce système regroupe 2000 produits de l'agriculture et de l'agroalimentaire, 245 pays et 15 années qu'il s'agit de mettre en relation avec la réglementation tarifaire et les informations sur les entreprises françaises. Il permet de fournir de la matière ordonnée directement utilisable par les chercheurs qui l'intègrent ensuite dans des modèles statistiques pour faire des simulations.

1. Contexte

1.1 Objectif MEDINA

Conçu pour répondre aux besoins des chercheurs en économie internationale, ce projet regroupe des informations officielles, bases de données, fichiers administratifs, enquête.
Ces données sont dispersées et hétérogènes.
L'objectif est de mettre en rapport des informations sur les échanges des produits, la tarification et les données individuelles des entreprises, permettre un suivi dans le temps et garantir la sécurité et la confidentialité des informations.

1.2 Sources et description des données

Les données volumineuses proviennent de sources officielles françaises, européennes et mondiales ; l'acquisition des données se fait à partir de fichiers fournis par les différents organismes. Certaines sont gratuites et facilement accessibles, d'autres sont financées par le département SAE2 (sciences sociales, agriculture et alimentation, environnement et espace) ou les unités et sont soumises à justification de projets, d'autres encore comportant des données individuelles confidentielles sont soumises à l'approbation du CNIS (conseil national de

[1] UR1134 LERECO - Laboratoire d'études et de recherches économiques - INRA - F- 44000 Nantes,
☏ 02 40 67 51 16- ✉ Monique.Harel@nantes.inra.fr ; Cecile.leroy@nantes.inra.fr

l'information statistique). Pour ces dernières, seuls les utilisateurs du projet sont autorisés à les exploiter. Principales sources :

- Douanes : INSEE – Institut national de la statistique et des études économiques
  Les échanges des entreprises françaises de 1995 à 2007 par pays, produits (NC8 et CPF6). L'utilisation de ces données est soumise à une convention entre la direction des Douanes et l'unité (LERECO, Nantes).
- EAE : Enquêtes annuelles d'entreprises (IAA) – Agreste-Ministère de l'agriculture et de la pêche. De 1995 à 2005 accès restreint car rattachées à un projet de recherche (Nantes).
- Comext : Eurostat – Office statistique des communautés européennes

  Les échanges mensuels européens par pays et par produit (NC8) de 1992 à 2007
- Taric : Commission européenne-Direction générale de la Fiscalité et Union douanière

  La réglementation douanière européenne (1995 à 2008)
- Comtrade (Commodity trade statistics database) : Nations Unies (UN)
  Les échanges mondiaux par pays et par produit (HS6) de 1992 à 2007 (financement département SAE2)

2. L'interface de consultation

L'objectif est de faciliter l'accès aux données et de rendre l'utilisateur plus autonome c'est-à-dire que la connaissance de la structure et de l'organisation des bases et la maîtrise du langage SQL (langage structuré de requêtes) d'interrogation des bases ne sont pas nécessaires.
Cette application regroupe une partie des informations de MEDINA, plus précisément les données du commerce international (Comext et Comtrade) et les nomenclatures correspondantes.

L'interface de consultation permet d'accéder aux données via un navigateur. Les principales fonctionnalités sont les suivantes :
- consultation des données et calcul d'agrégats
- extraction des données pour les traitements statistiques
- visualisation et extraction des nomenclatures pays (correspondance Comext, Comtrade)
- affichage de l'arborescence des produits de la nomenclature combinée (NC8)

2.1. Développement

L'application client/serveur a été développée en PHP et JavaScript permettant de créer des pages WEB dynamiques ; elle est couplée à une base de données PostgreSQL qui contient toutes les informations. La sélection de l'utilisateur (requête) est envoyée via le serveur Web au serveur des bases de données qui renvoie le résultat qui s'affiche à travers le navigateur
Le **schéma 1,** ci-dessous, représente l'architecture de l'application :
Les données sont hébergées sur le serveur de calcul sur lequel sont implantées les bases PostgreSQL ; les pages HTML et les programmes PHP sont installés sur un serveur d'applications WEB

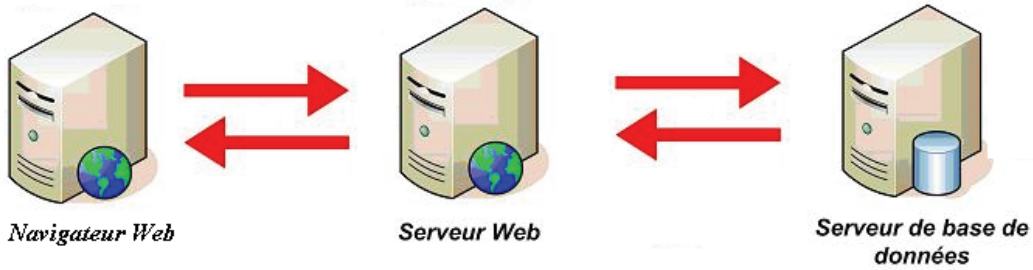


Schéma 1 : *architecture de l'application*

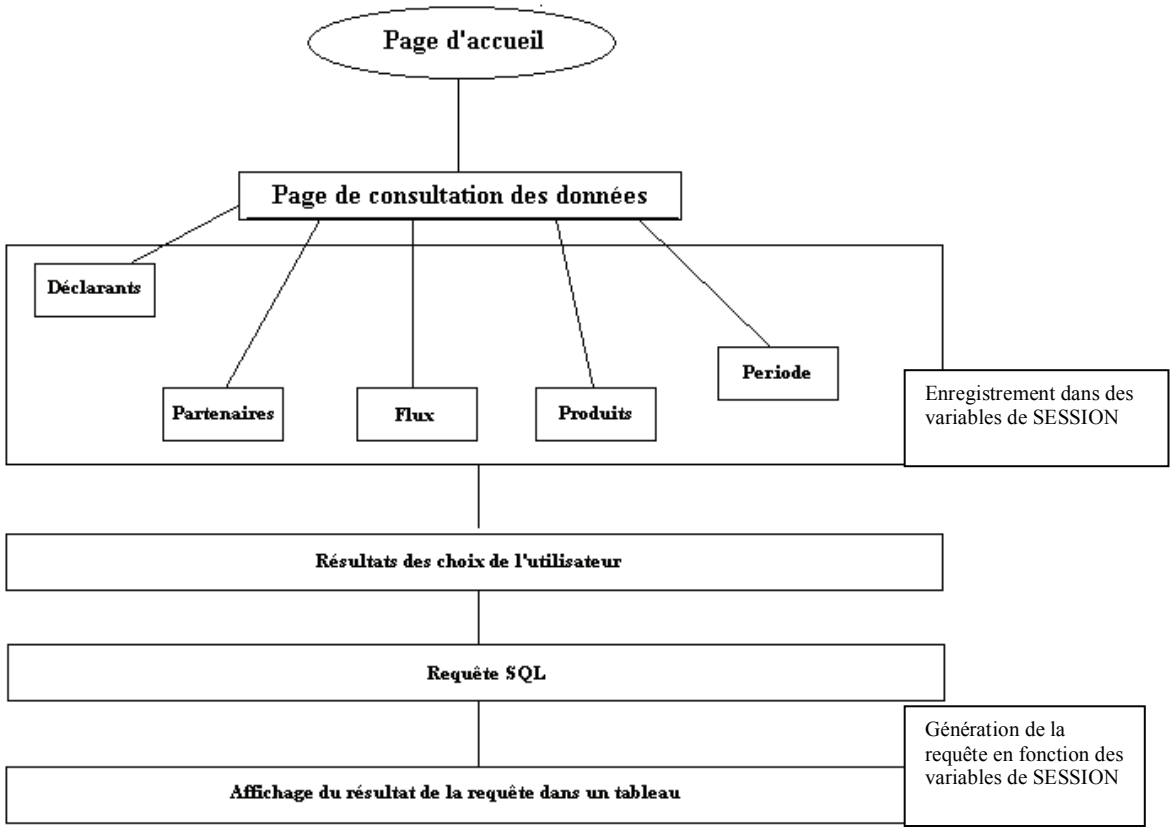
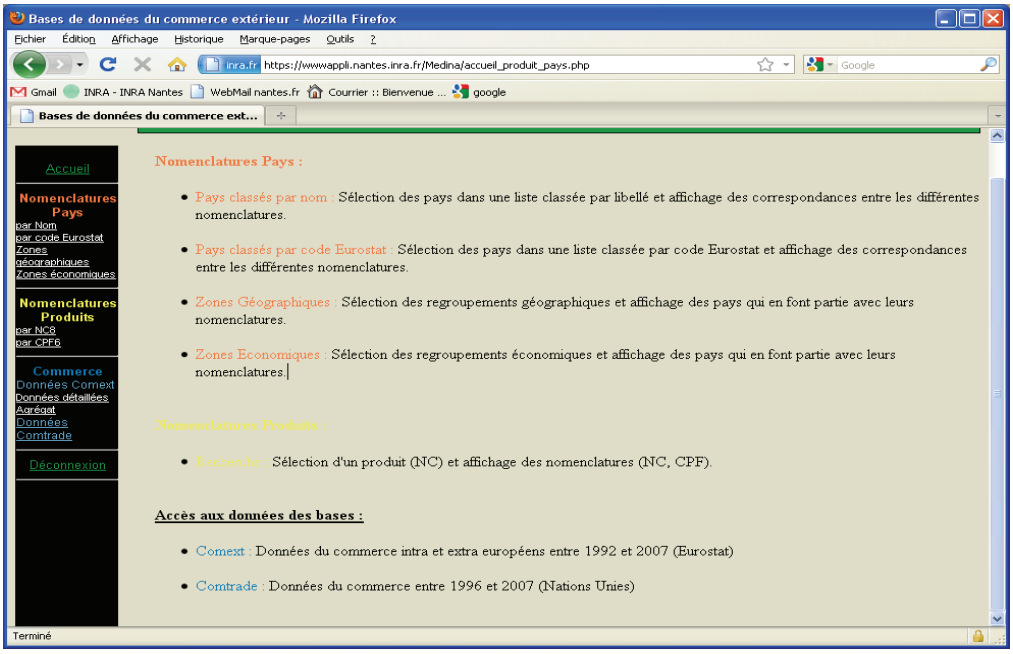


Schéma 2 : *principe de fonctionnement du projet*

2.2 Description de l'interface

La page d'accueil donne un accès direct aux nomenclatures des pays et des produits, ainsi qu'aux données du commerce Intra et Extra européens.



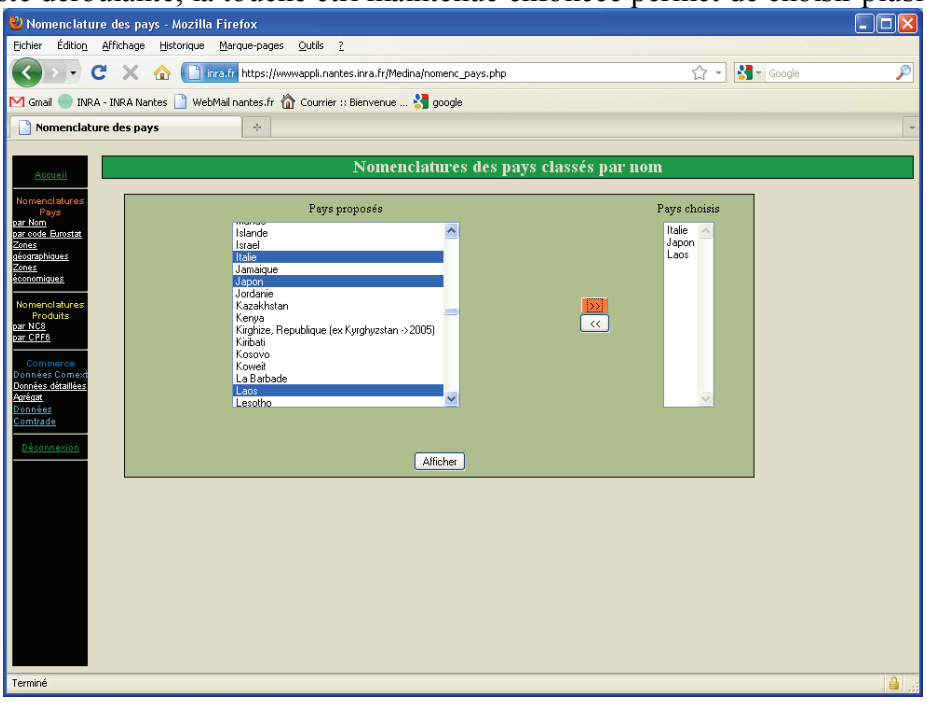
2.2.1 Les nomenclatures

La gestion des nomenclatures pays et produits est la clé de voûte du projet MEDINA ; en effet il est indispensable de s'assurer de la concordance entre les sources et de leur compatibilité dans le temps. Le croisement des différentes sources est réalisé grâce aux tables de correspondance

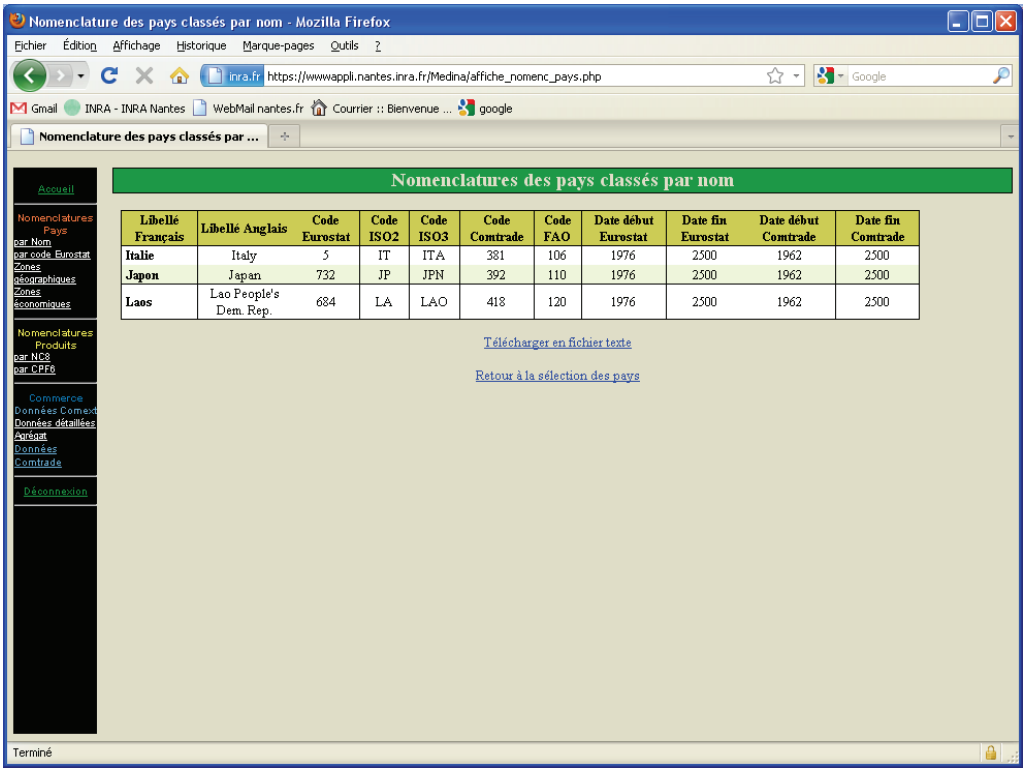
- Nomenclatures pays

Sélection des nomenclatures par nom

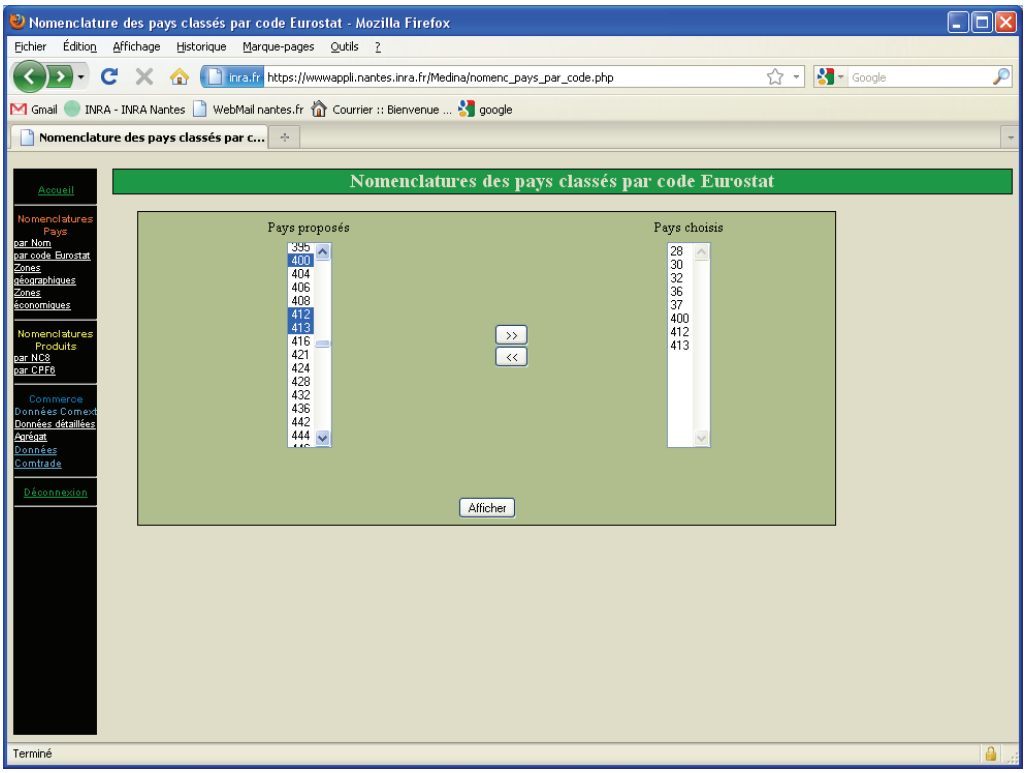
Pour obtenir le code correspondant à ces pays, l'utilisateur sélectionne les noms de ces pays dans une liste déroulante, la touche ctrl maintenue enfoncée permet de choisir plusieurs noms

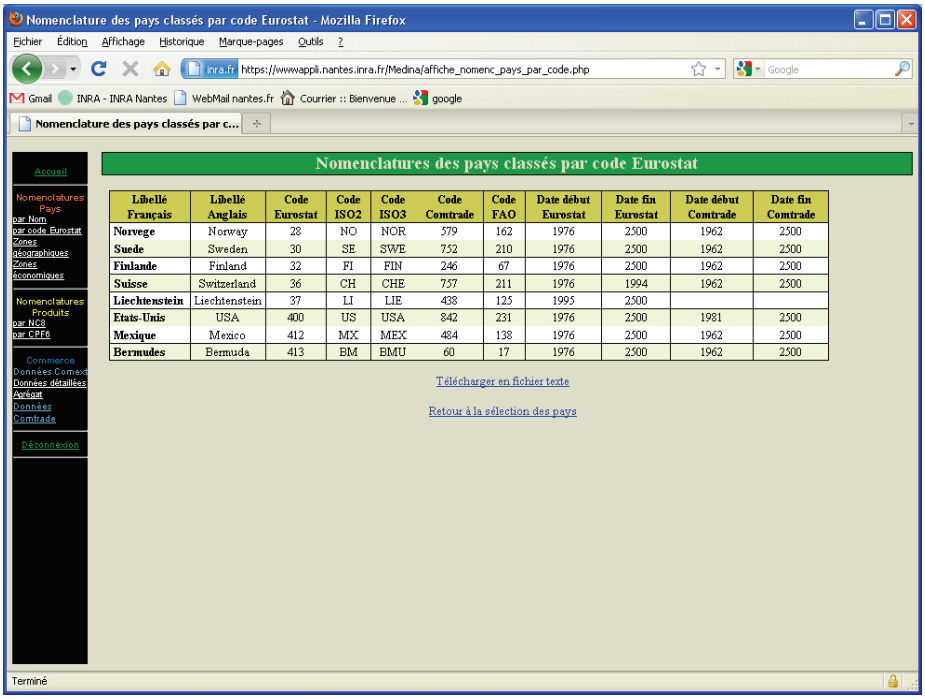


L'option « afficher » permet d'obtenir le code de ces pays dans les différentes nomenclatures



L'utilisateur peut aussi interroger à partir des codes numériques et ainsi obtenir le libellé du pays et la correspondance avec les autres nomenclatures ; il doit choisir dans le menu nomenclature pays, l'option « par code eurostat »



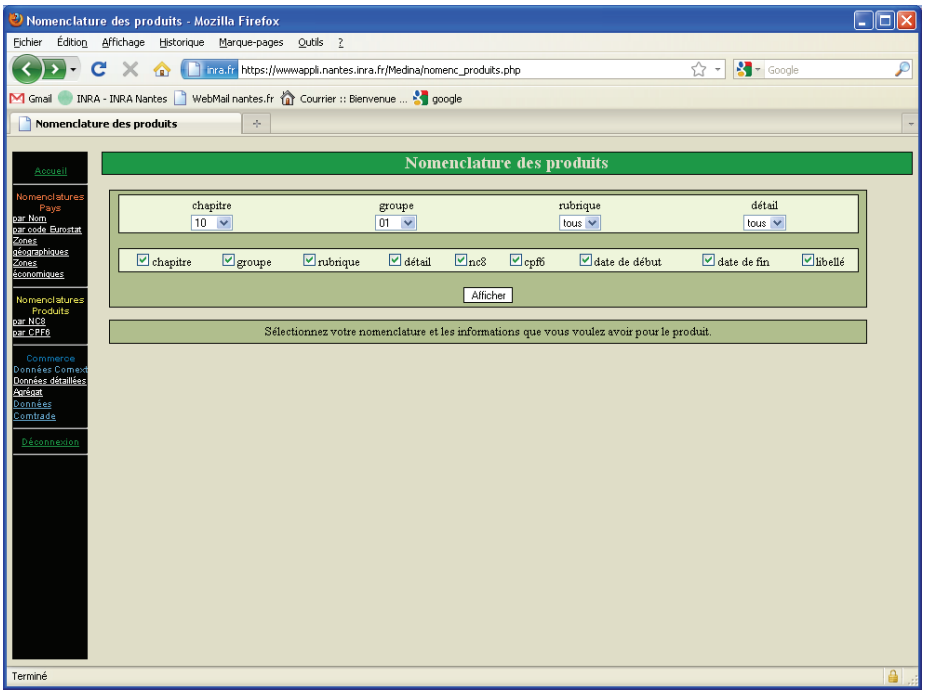


- Nomenclatures produits

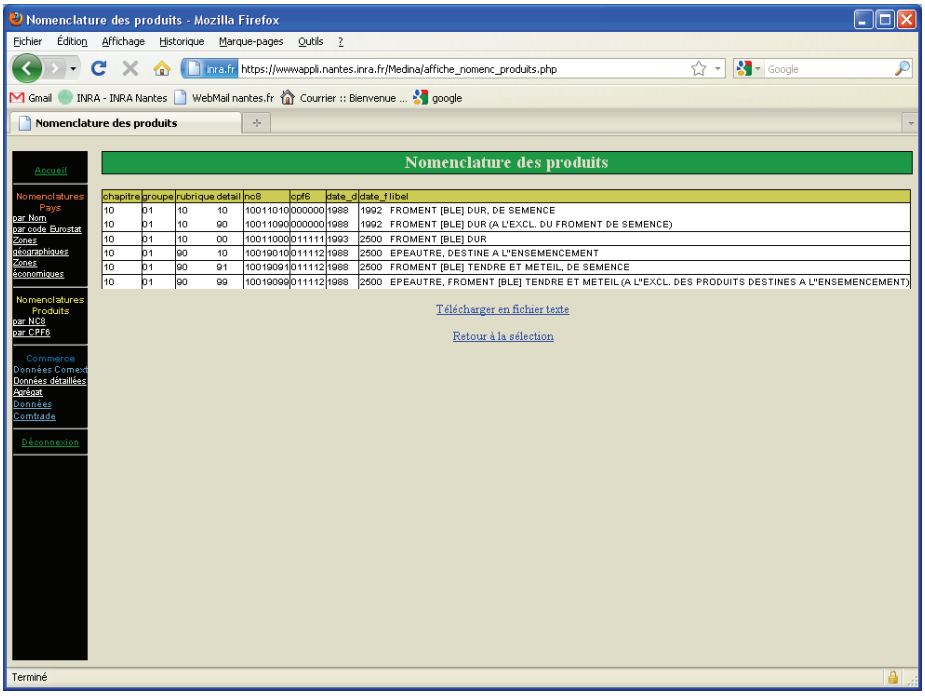
Les statistiques du commerce international utilisent plusieurs niveaux de codification des marchandises : le système harmonisé (SH) codifié sur 6 positions numériques, la nomenclature combinée (NC) à 8 chiffres correspondant à la SH plus 2 chiffres et la classification française des produits sur 6 positions numériques (CPF6).

La recherche des équivalences dans les nomenclatures citées peut se faire à partir d'un code NC8 ou CPF6. La nomenclature NC8 est formée par l'agrégat chapitre (1 à 24 pour les produits agricoles et agroalimentaires), groupe, rubrique, détail à renseigner ou non pour la demande.

On renseigne aussi les rubriques résultats souhaitées (libellé, date de validité…).



Dans cet exemple, on demandait les nomenclatures NC8 et CPF6, le libellé de tous les produits du chapitre 10 et du groupe 01



chapitre	groupe	rubrique detail	nc8	cpf6	date_d	date_f	libel
10	01	10	10	10011010	000000	1988	1992 FROMENT [BLE] DUR, DE SEMENCE
10	01	10	90	10011090	000000	1988	1992 FROMENT [BLE] DUR (A L'EXCL. DU FROMENT DE SEMENCE)
10	01	10	00	10011000	011111	1993	2500 FROMENT [BLE] DUR
10	01	90	10	10019010	011112	1988	2500 EPEAUTRE, DESTINE A L'ENSEMENCEMENT
10	01	90	91	10019091	011112	1988	2500 FROMENT [BLE] TENDRE ET METEIL, DE SEMENCE
10	01	90	99	10019099	011112	1988	2500 EPEAUTRE, FROMENT [BLE] TENDRE ET METEIL (A L'EXCL. DES PRODUITS DESTINES A L'ENSEMENCEMENT)

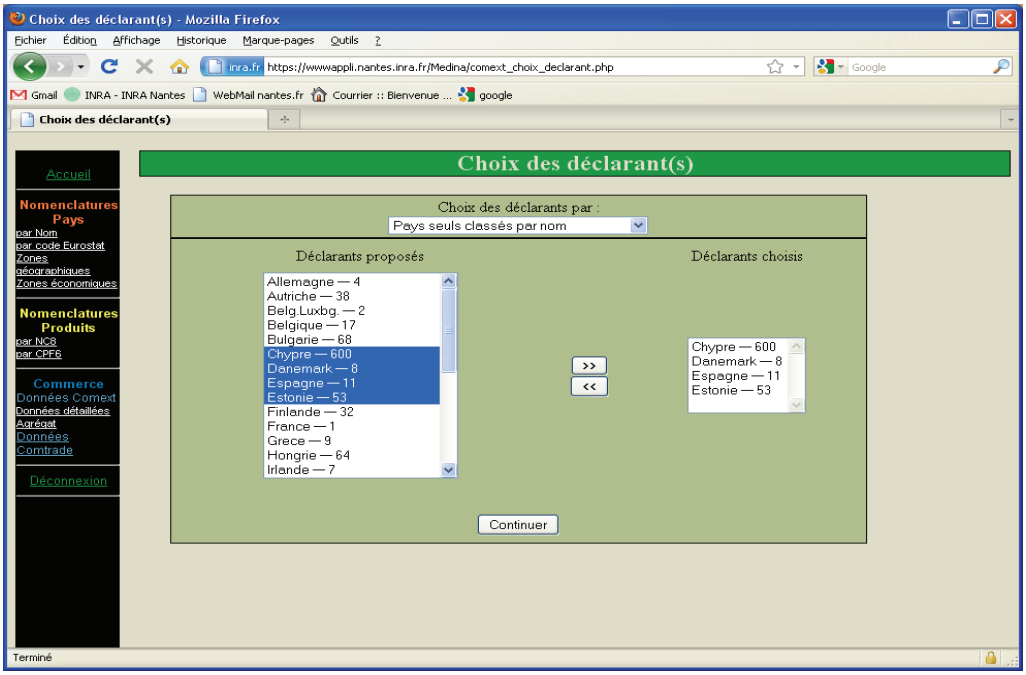
Pour une demande à partir de la nomenclature CPF6, plus agrégée, on choisit les codes parmi la liste des produits existants et on obtient le même type de résultat.

2.2.2 Les données

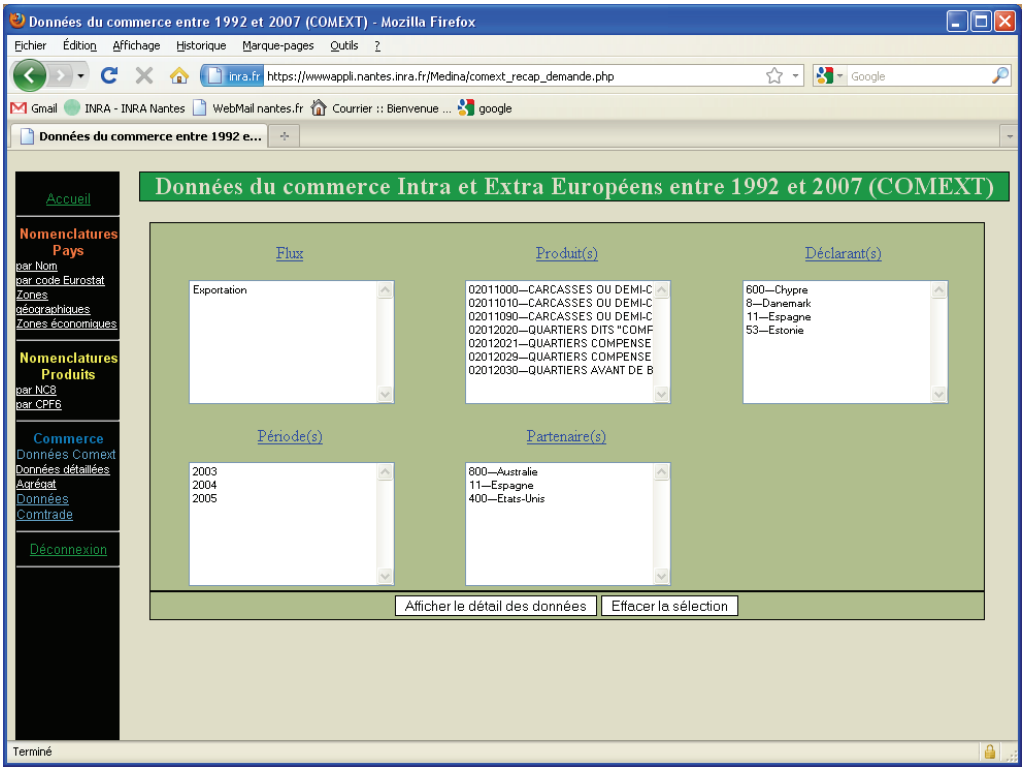
Avec le lien données Comext, données détaillées, on accède à une nouvelle page pour la sélection des informations à partir de différentes listes (Déclarants, Partenaires, Produits, Flux, Période) :
- déclarants : pays de l'Union Européenne importateur ou exportateur
- partenaires : pays destinataires
- produits : produits de l'agriculture et de l'agroalimentaire
- flux : importation ou exportation
- période : année(s) concernée(s)

Choix des pays déclarants

Pour sélectionner un ou plusieurs pays, en les mettant en surbrillance, ces derniers sont stockés dans un tableau de session utilisé par PHP

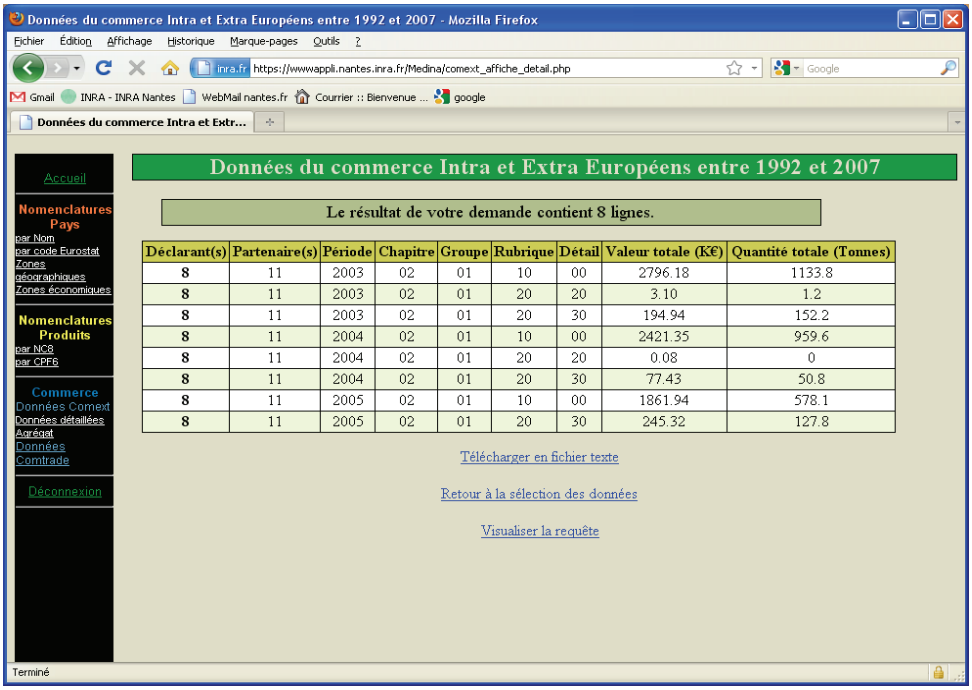


Lorsque les choix sont effectués, ils sont affichés sur la page principale dans les listes prévues à cet effet



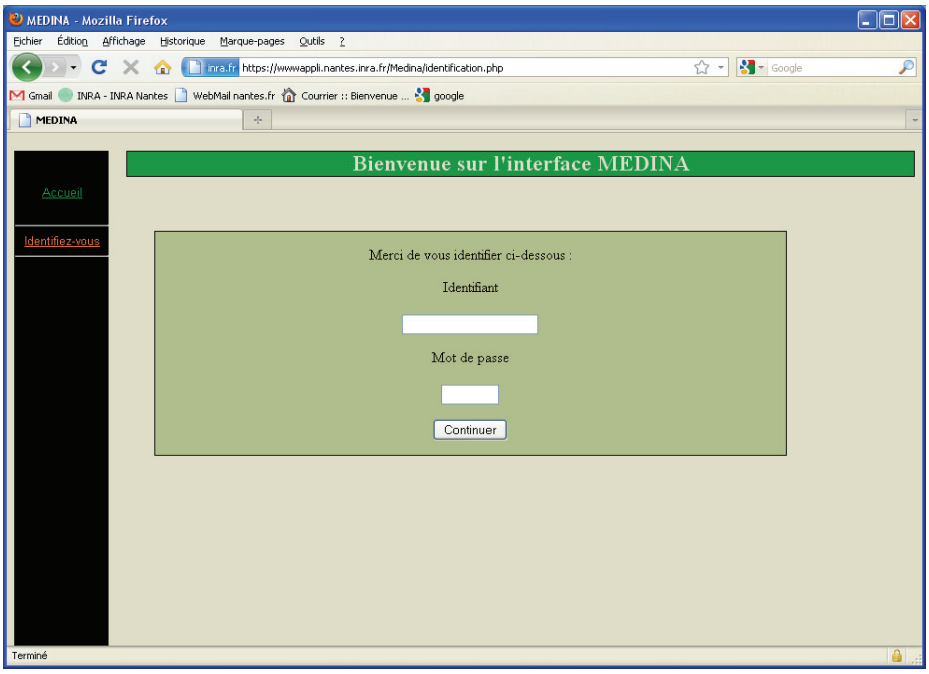
De cette page on peut générer une requête SQL en fonction des choix donnés par l'utilisateur.

Le résultat de la requête s'affiche dans un tableau. Il est possible de télécharger ces informations dans un fichier texte (type .txt)



3. Sécurité de l'application

La sécurité du projet MEDINA se situe à 2 niveaux (serveur et base de données). Au niveau serveur, la sécurité est gérée par un fichier de configuration PostgreSQL. Au niveau bases de données, certaines étant financées par le département SAE2, il est indispensable de restreindre l'accès aux personnes appartenant à ce département via un système d'authentification par login/mot de passe qui doit être reconnu au niveau des bases de données. Pour le moment, il existe un écran d'identification avec saisie d'un identifiant et d'un mot de passe fournis par l'administrateur des bases.



4. Limites de l'application

A ce stade, l'application ne donne pas accès à l'ensemble du projet MEDINA, néanmoins elle permet la consultation et l'extraction d'informations sur le commerce extérieur et sur les nomenclatures pays et produits.
Pour rendre accessible l'ensemble des données (comme les « douanes »), il reste à mettre en place un processus d'identification pour les membres du projet associé à ces données.

Conclusion

L'application Web MEDINA est un outil qui peut facilement évoluer en fonction de la demande.
De nouvelles fonctionnalités telles que le calcul d'agrégats (c'est à dire, calculer pour un produit ou groupe de produits donné, un flux, un ou plusieurs partenaires, une ou plusieurs années, la quantité et la valeur échangée par l'ensemble des pays de la zone UE) et les sélections multiples sont maintenant proposées. On pourrait ajouter, des informations concernant la réglementation douanière et l'évolution des nomenclatures produits qui évoluent en permanence.
Par ailleurs, le CATI[2] IATISS[3] outil d'animation informatique a pour objectif de mettre en commun les applications développées dans le département. Ainsi, les données gérées dans les unités seront mises à disposition à l'ensemble des chercheurs.
C'est ainsi qu'une version test a été portée sur le serveur d'application dédié au CATI localisé à Toulouse.

Bibliographie

Rigaux P. (2002) Pratique de MySQL et PHP

Geschwinde E., Schönig H. (2005) Pratique de MySQL et PHP

Messager J., Rousseau R. pour Learning Tree International (2007) Développement Web avec PHP (cours)

INSEE : guide d'utilisation des nomenclatures d'activités et de produits

Documentation PHP, http://www.php.net/manual/fr/

Documentation PostgreSQL, http://docs.postgresql.fr/8.3/

[2] Centre automatisé de traitement de l'information
[3] Informatique Appliquée au Traitement de l'Information en Sciences Sociales

Encadré 1 : extrait d'une page PHP

```
<h2>Nomenclatures des pays classés par nom</h2>
<?
/*initialisation de la connexion avec la base de données*/
require_once('dblogin_intercom.php');
/*définition de la requête*/
$req="SELECT * from pays_9207_appli where";
$i=1;

foreach($_SESSION['panier_pays'] as $code_geo)
{
	$req.=" code_geo=".$code_geo;
	if($i!=count($_SESSION['panier_pays']))/*count donne le nombre d'éléments d'un tableau*/
	{
		$req.=" or";/*si l'élément lu n'est pas le dernier du tableau on ajoute l'élément
suivant*/
	}
	$i++;
}
$req.=" order by libel_fr;";
/*exécution de la requête et récupération du résultat*/
$resultat=pg_query($req);
/*génération du nom de fichier temporaire : le fichier doit avoir un nom unique au cas ou 2 utilisateurs
sont présents en même temps sur la page(le fichier du 1er utilisateur serait écrasé par celui du 2e dans
le cas contraire).*/
srand();/*initialisation du générateur de nombres aléatoires*/
$n=rand();/*génération d'un nombre aléatoire*/
$d=date('His');/*récupération de l'heure (HeuresMinutesSeccondes)*/
$nom_fichier="tmp/nomenc_pays-".$d.$n.".txt";
/*ouverture du fichier texte en écriture*/
$fichier=fopen($nom_fichier,"w"); /*w est le mode "écriture"  d'ouverture du fichier*/
/*affichage des informations dans un tableau*/
echo "
<table class='result' border='1'>
<thead><tr>
<th>Libellé Français</th>
<th>Libellé Anglais</th>
<th>Code Eurostat</th>
<th>Code ISO2</th>
<th>Code ISO3</th>
<th>Code Comtrade</th>
<th>Code FAO</th>
<th>Date début Eurostat</th>
<th>Date fin Eurostat</th>
<th>Date début Comtrade</th>
<th>Date fin Comtrade</th>
</tr></thead>
";
```

RICA, outil d'interrogation et de traitements SAS via le Web

Jean-Marc Rousselle[1]

Résumé : *L'accès aux données du RICA (Réseau d'information comptable agricole) est lourd et complexe ; par ailleurs le langage de programmation de SAS (Statistical Analisys System) est ardu ; le laboratoire d'économie forestière (LEF) du centre Inra de Nancy a développé un environnement d'interrogation du RICA pour réduire au maximum les contraintes liées à l'accès à ses données*

Mots clés : RICA, SAS, bases de données, PHP, Mysql, WEB

Introduction

L'accès aux données du RICA (Réseau d'information comptable agricole) est complexe tant d'un point de vue matériel par l'achat du logiciel d'émulation de terminal IBM et par un abonnement onéreux aux réseau TRANSPAC que par un gros investissement de l'utilisateur qui doit connaître l'organisation des bases et maîtriser la syntaxe du langage de programmation de SAS.
Le but de l'environnement d'interrogation du RICA développé au LEF de Nancy et qu'une personne qui ne connaît ni le langage SAS ni l'organisation des bases RICA soit capable de pouvoir sortir des résultats assez rapidement et simplement. Seule l'utilisation d'un navigateur WEB est nécessaire.
Cet article présentera rapidement le RICA et le SAS ainsi que l'outil mis en place par le LEF, ses astuces et ses limites.

Le laboratoire d'économie forestière (LEF) conçoit, développe et transfère des méthodes et des outils d'analyse économique en vue de comprendre et d'améliorer la connaissance, la gestion et l'utilisation des biens, services et produits de la forêt.

[1] INRA - UMR0356 LEF - Laboratoire d'économie forestière – F-54042 Nancy
☏ 03 83 39 68 62 ✉ Jean-Marc.Rousselle@nancy-engref.inra.fr

1. Le RICA

1.1 Descriptif

Le RICA (Réseau d'information comptable agricole), en tant qu'opération statistique communautaire, est instauré en France, en 1968, par application du règlement 79/65/CEE. Il a pour objectif de fournir des informations sur le fonctionnement technico-économique des exploitations agricoles pour suivre leur revenu et pour éclairer les décisions de la PAC (Politique agricole commune). Le RICA couvre l'ensemble des exploitations agricoles d'une dimension économique supérieure à 9 600 UCE "1986" soit 12 équivalents hectares de blé et employant au moins 0,75 UTA (Unité de travailleur annuelle). La production des exploitations agricoles appartenant au champ de l'enquête représente près de 90 % de la production de la branche agriculture. Les règlements communautaires fixent le nombre minimum d'exploitations à sélectionner par pays membre (de l'ordre de 7 100 pour la France avec une ventilation régionale précise).
L'échantillon est extrapolé, en s'appuyant sur les données des enquêtes de structure, afin d'être représentatif de l'agriculture professionnelle.
Le RICA est une base de données très utilisée par les chercheurs notamment à l'Inra. Les recherches portent sur les disparités de revenus, l'économie de la production, la modélisation des exploitations, l'analyse des effets de la PAC. De nouvelles questions concernent l'environnement, ce qui nécessite une adaptation du RICA.

1.2 Organisation des données

Les bases de données du RICA sont stockées sous le format spécifique du logiciel statistique SAS. L'organisation des ces données est de type hiérarchique. Organisées dans plusieurs tables SAS suivant leur fonction, il est nécessaire de bien connaître cette organisation pour pouvoir les manipuler. Les tables SAS référençant les informations comme, les caractéristiques générales, le bilan comptable, et les résultats des exploitations agricoles contiennent une observation par exploitation. En revanche, les autres tables comme les données sur les animaux, les produits animaux, les végétaux, les produits végétaux et les produits végétaux transformés, peuvent avoir plusieurs observations par exploitation, en fait autant que de produits ou d'animaux. Toutes ces informations peuvent être réorganisées grâce à une variable commune à toutes ces tables, qui est en fait un code d'identification de l'exploitation.
Il est donc indispensable de bien connaître l'organisation des données du RICA avant de pouvoir l'exploiter.

1.3 Accès aux données

Historiquement les bases de données du RICA étaient hébergées sur les gros systèmes (mainframe) IBM™ de l'INSEE[2]. Pour accéder à ces données depuis un poste Inra, il était indispensable d'utiliser un logiciel spécifique qui transforme un Micro-ordinateur de type PC sous Windows en terminal reconnu par les systèmes IBM (type terminal 3278). Une fois ce terminal configuré, il était nécessaire ensuite de se connecter sur le réseau interne de l'INSEE en transitant via le réseau privé (et payant) TRANSPAC. Une fois connecté sur le système de l'INSEE, le RICA pouvait alors être interrogé sous l'environnement d'exploitation spécifique IBM.

[2] Institut national de la statistique et des études économiques

2. SAS

Les bases de données du RICA sont créées et enregistrées au format propriétaire de SAS (Statistical Analisys System). Donc pour éviter les transformations de celles-ci, SAS est le langage naturel pour exploiter les données. Ce qui fait que tout utilisateur potentiel des informations du RICA se doit de connaître le langage SAS. N'ayant pas d'environnement graphique sous le système IBM™ de l'INSEE, le nouveau produit de programmation intuitive « Enterprise guide »[3], implémenté depuis la version 9 de SAS, n'est pas disponible. Il est donc nécessaire de connaître les principes, les règles, le fonctionnement et surtout la syntaxe du langage SAS.

3. L'outil

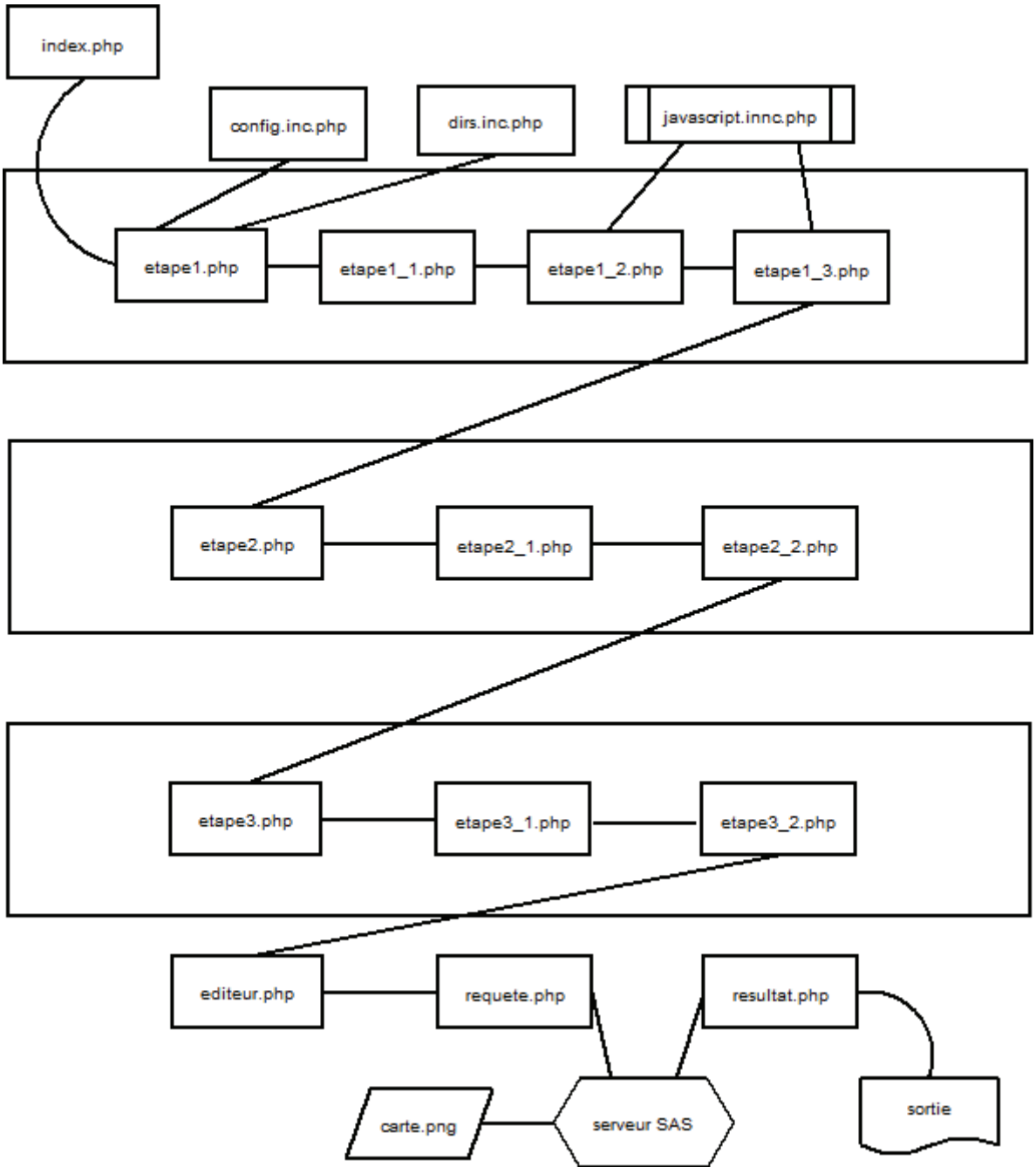
Comme il est possible de le constater, l'accès aux données du RICA est lourd et complexe. Non seulement d'un point de vue matériel par l'achat du logiciel d'émulation de terminal IBM et un abonnement onéreux aux réseaux TRANSPAC, mais également par un grand investissement de la part de l'utilisateur qui doit connaître l'organisation des bases mais aussi maîtriser la syntaxe du langage de programmation de SAS.
Le but de l'environnement d'interrogation du RICA développé au LEF de Nancy, est d'essayer de se soustraire au maximum de ces contraintes. Normalement une personne qui ne connaît ni le langage SAS ni l'organisation des bases RICA est capable de sortir des résultats assez rapidement et simplement. Seule l'utilisation d'un navigateur WEB comme par exemple Firefox ou Internet Explorer est nécessaire.

3.1 Développement

L'outil a été développé en PHP, langage permettant de générer du code HTML. Le produit est également couplé à une base de données sous MySQL qui recense toutes les caractéristiques de la base de données RICA, comme le nom des variables disponibles, dans quelle table SAS elles se trouvent, sous quelle catégorie elles sont référencées, leur période de validité etc. Initialement le logiciel tournait sur un serveur dédié WEB qui envoyait des requêtes SAS à une seconde machine hébergeant SAS, laquelle machine renvoyait les résultats à la première. Maintenant, avec l'utilisation de produit permettant de transformer n'importe quel ordinateur en pseudo serveur WEB (comme par exemple EasyPHP, LAMP ou WAMP)[4], l'ensemble du dispositif peut être exploité sur une seule et même machine. Si bien que pour exécuter le logiciel, il suffit d'avoir une machine pouvant faire tourner SAS, d'y copier les bases du RICA et de la transformer en serveur WEB.

[3] Enterprise Guide est un outil de la gamme SAS à destination des nouveaux utilisateurs, permettant de programmer le langage SAS à l'aide de menus, de choix et d'options et d'éviter ainsi les contraintes syntaxiques.
[4] LAMP = Outil associant Linux-Apache-MySql-PHP, WAMP = Outil associant Windows-Apache-Mysql-PHP

3.2 Organigramme des programmes



3.3 Les bases et fichiers de travail

Le programme s'appuie sur les données du RICA qui sont au format natif SAS, mais également sur des bases de données décrivant l'organisation et les caractéristiques des variables disponibles. Ces informations sont enregistrées au format base de données sous MySql. Il existe deux bases de données de travail décrivant les variables du RICA : Les bases rubrique et variable

Serveur: localhost ▸ Base de données: rica ▸ Table: rubrique

rubrique

Champ	Type	Null	Défaut
rubco	char(255)	Oui	*NULL*
rubrique	char(255)	Oui	*NULL*

Avec : Rubco : Le code de la rubrique
Rubrique : Le libellé de la rubrique

Serveur: localhost ▸ Base de données: rica ▸ Table: variable

variable

Champ	Type	Null	Défaut
rubco	char(255)	Oui	*NULL*
nom	char(10)	Non	
fichier	char(10)	Non	
type	char(255)	Oui	*NULL*
label	char(255)	Oui	*NULL*
libelle	char(255)	Oui	*NULL*
unite	char(255)	Oui	*NULL*
debut	int(11)	Oui	*NULL*
fin	int(11)	Oui	*NULL*

Avec :

Rubco : Le code de la rubrique
Nom : Le nom de la variable
Fichier : La table SAS dans laquelle se trouve la variable (CAR,RES,BIL etc.)
Type : Le type de la variable (numérique ou alphanumérique)
Label : Le label sous lequel est référencée la variable sous SAS
Libelle : Le libellé complet de la variable
Unité : L'unité de la variable (monétaire, quintal etc.)
Début : Année d'apparition de la variable
Fin : Année de fin de la variable

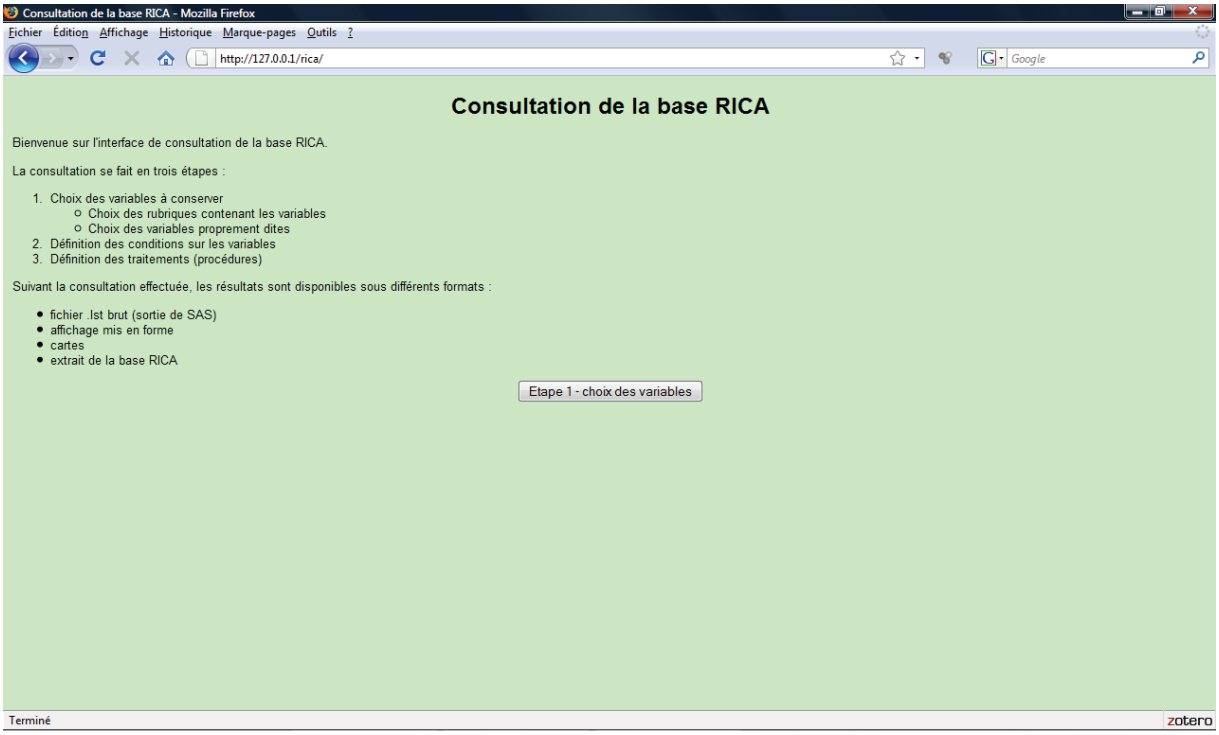
3.4 Installation

Pour pouvoir utiliser notre logiciel, la machine doit donc avoir le logiciel SAS installé, les bases de données du RICA doivent être accessibles depuis cette machine. Il existe un CD contenant le reste des outils nécessaires pour une utilisation sous un environnement Windows, à savoir le logiciel easyPHP, les programmes du logiciel lui même, ainsi qu'une notice détaillée d'installation au format PDF qu'il suffit de respecter. L'installation a déjà été faite sur plusieurs sites et n'a posé aucun problème.

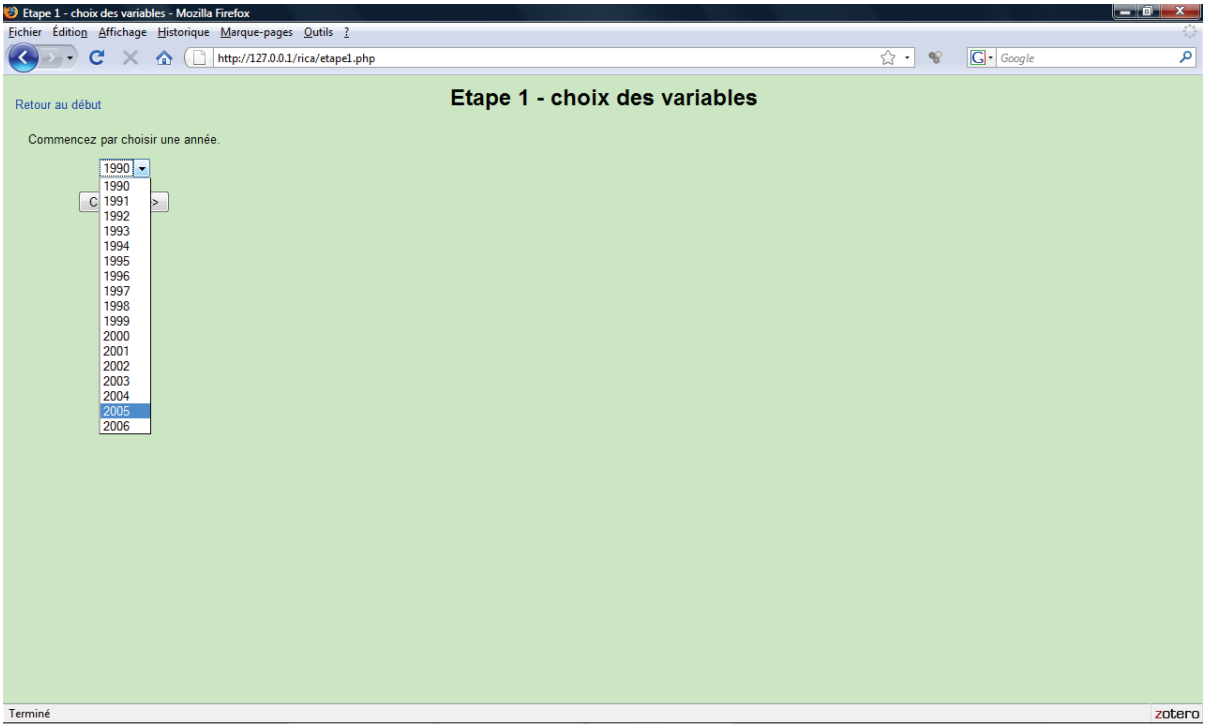
3.5 Utilisation

Nous avons essayé de développer l'outil grâce à notre expérience de l'utilisation et de l'exploitation des données du RICA, avec pour objectif que sa simplicité d'utilisation le mette à la portée de personnes non familiarisées avec le langage SAS ou avec l'organisation des données.
L'utilisateur se retrouve devant trois fenêtres qui s'enchaînent et qui demandent sur quelles données précises il souhaite travailler, sur quelle sélection de l'échantillon et quels résultats ou sorties il souhaite obtenir.
Tout d'abord le logiciel affiche une fenêtre expliquant la démarche à suivre par l'utilisateur pour lui indiquer ses choix.

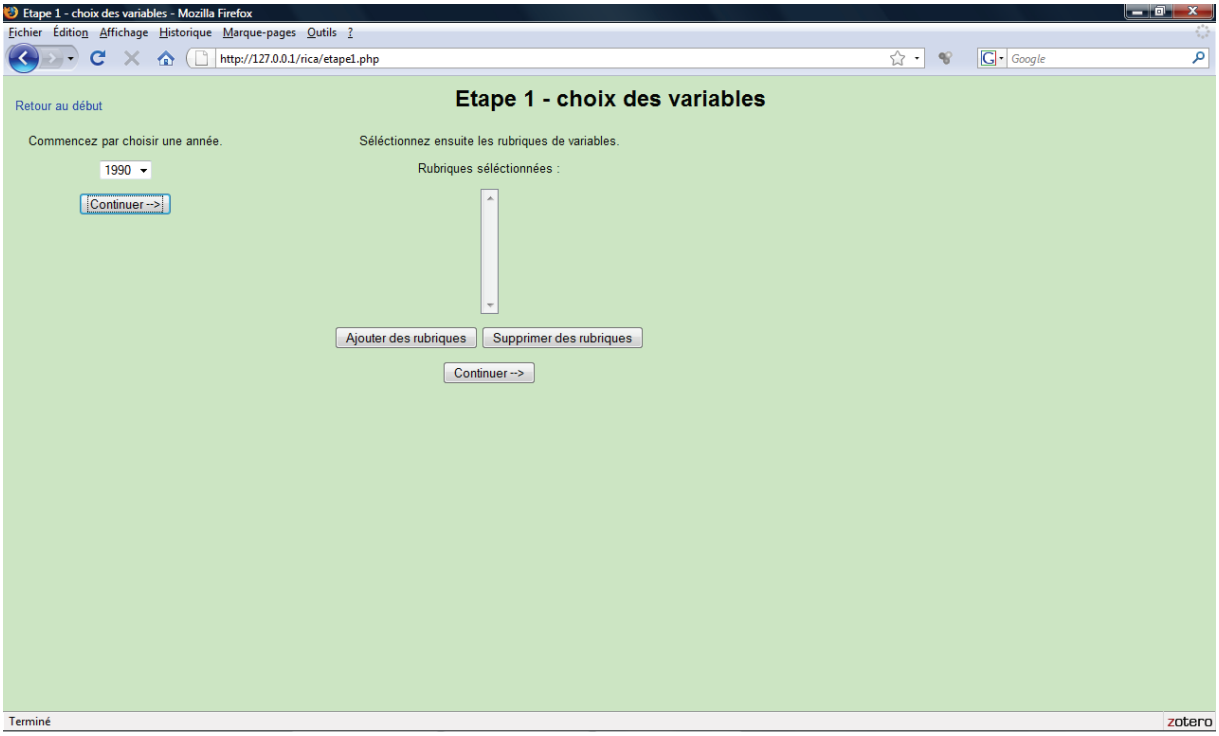
Etape 1 : Sélection de l'année, des rubriques et des variables



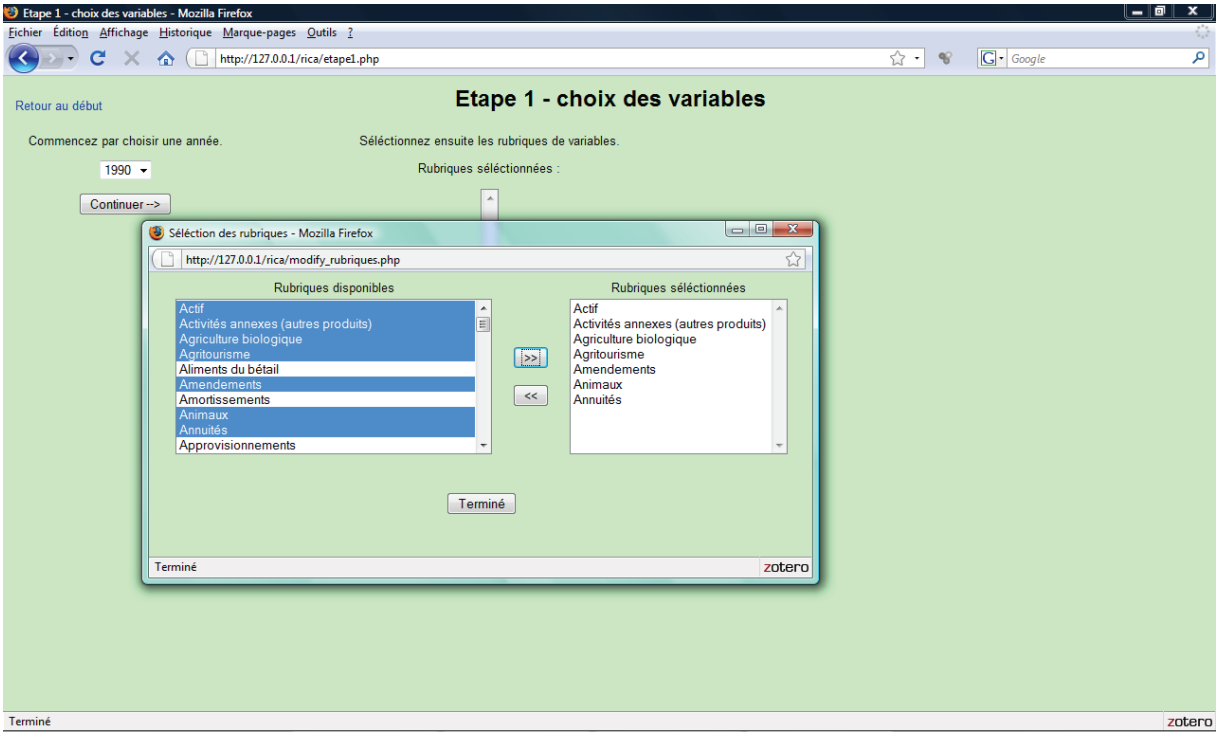
En premier lieu l'utilisateur mentionne sur quelle année il souhaite faire ses traitements. Il lui suffit de sélectionner cette année dans la liste déroulante proposée.



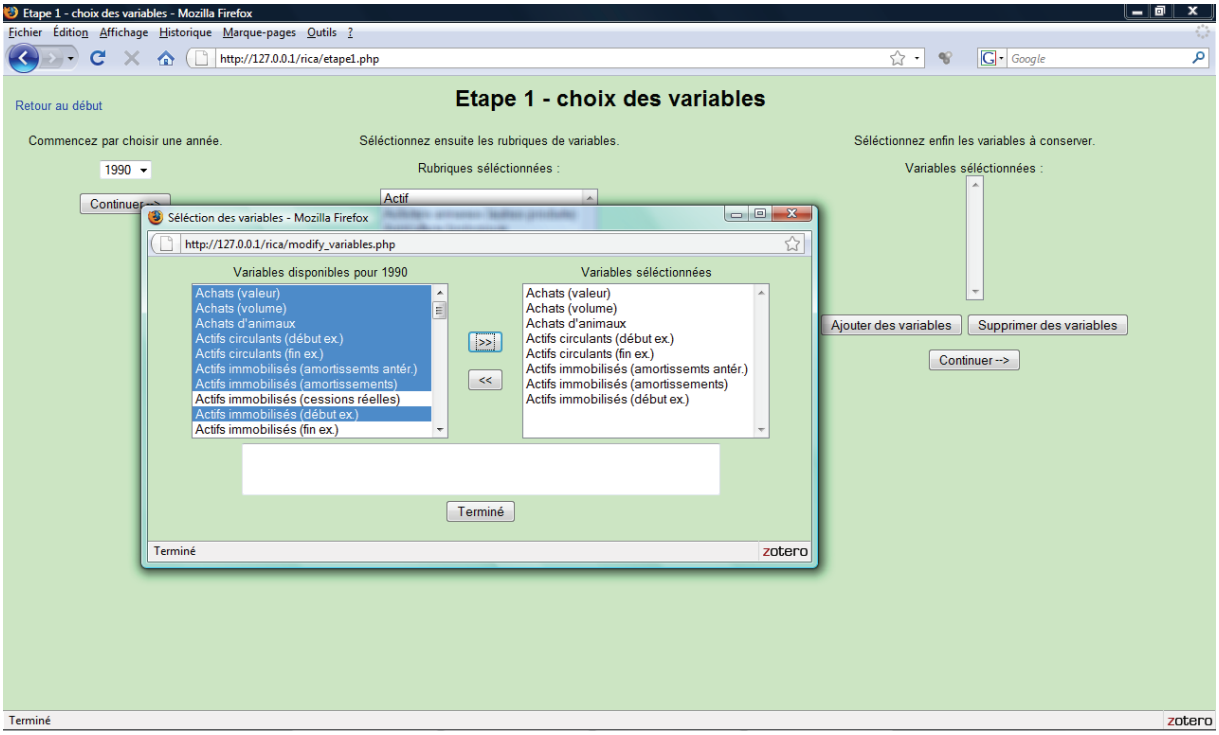
Après avoir cliqué sur le bouton « Continuer », l'utilisateur pourra affiner sa sélection en choisissant la ou les rubriques qui l'intéressent en cliquant sur le bouton « Ajouter des rubriques».



A ce moment là, une petite fenêtre s'ouvre et fournit un outil permettant de sélectionner une ou plusieurs rubriques dans la partie gauche (Il est possible de cliquer en maintenant la touche « Majuscule » du clavier enfoncée pour sélectionner des rubriques contiguës dans la liste ou sur la touche « CTRL » pour des rubriques non contiguës). Une fois la sélection faite, il suffit de cliquer sur la touche « → » pour passer celle-ci de la fenêtre « Rubriques disponibles » vers la fenêtre « Rubriques sélectionnées ». Il est possible de faire cette manipulation plusieurs fois de suite, voire revenir sur la sélection faite à tout moment en passant les rubriques d'une zone à l'autre avec les boutons centraux de la fenêtre.



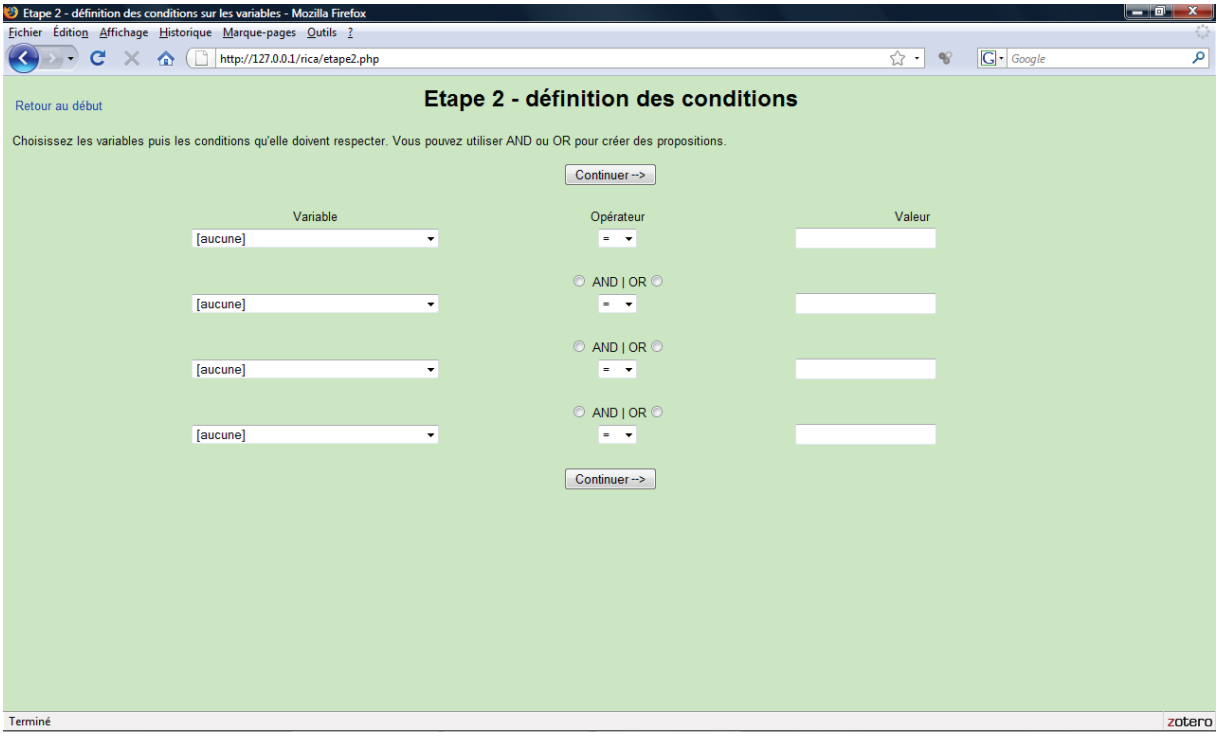
Enfin il suffit de cliquer « Terminé » pour fermer la fenêtre. Il est encore possible de revenir sur les sélections, tant que le bouton « Continuer » n'a pas été cliqué. Dans ce cas on passe à la dernière phase de l'étape 1, à savoir le choix des variables.
De la même manière qu'on a sélectionné les rubriques, on va maintenant pouvoir ne garder que les variables qui intéressent l'utilisateur dans son traitement. Il est à noter que seules les variables associées aux rubriques déjà sélectionnées apparaissent ; il n'est pas possible de revenir en arrière et d'ajouter une nouvelle rubrique.
En cliquant sur « Ajouter des variables », une nouvelle fenêtre s'ouvre.



Faire donc comme pour les rubriques, les mêmes manipulations pour affiner la sélection des variables. Cliquer sur « Terminé » pour finaliser cette sélection et « Continuer » pour passer à la seconde étape.

Etape 2 : Sélection des exploitations

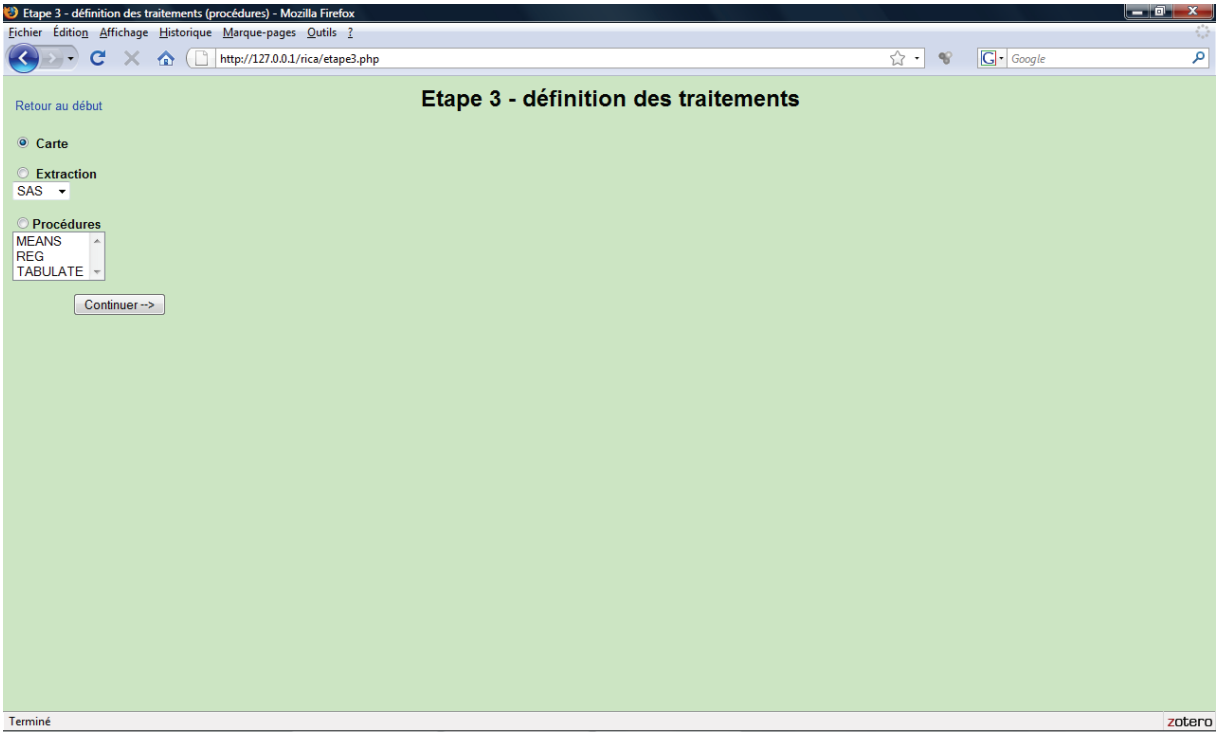
La deuxième étape va permettre à l'utilisateur de faire une sélection précise du sous échantillon sur lequel il va vouloir travailler en précisant différents critères de sélection (ou pas, pour prendre l'échantillon complet).



La sélection du sous échantillon ne pourra se faire que sur les variables sélectionnées dans l'étape 1 ou sur les variables qui sont automatiquement ajoutées par le programme comme la région, les OTEX (orientations technico-économiques), CEDEX (classes de dimension économique) etc.

Etape 3 : Le choix du traitement

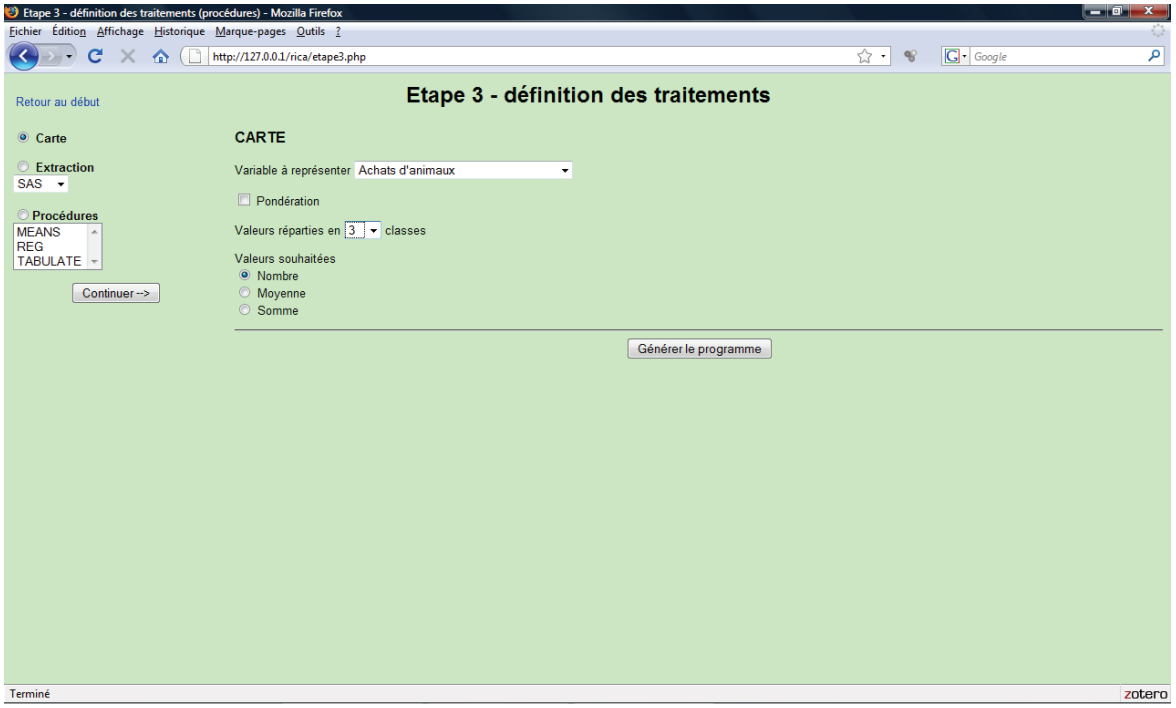
Enfin, la troisième étape, où l'utilisateur pourra choisir le traitement qu'il attend recevoir de l'exploitation des données ainsi sélectionnées. Il pourra, par exemple, demander des statistiques de base, faire une régression, demander d'obtenir une extraction de ses données en format SAS ou en format récupérable sous un tableur. Il pourra enfin demander la génération d'une carte des régions françaises représentant des informations sur une donnée particulière.



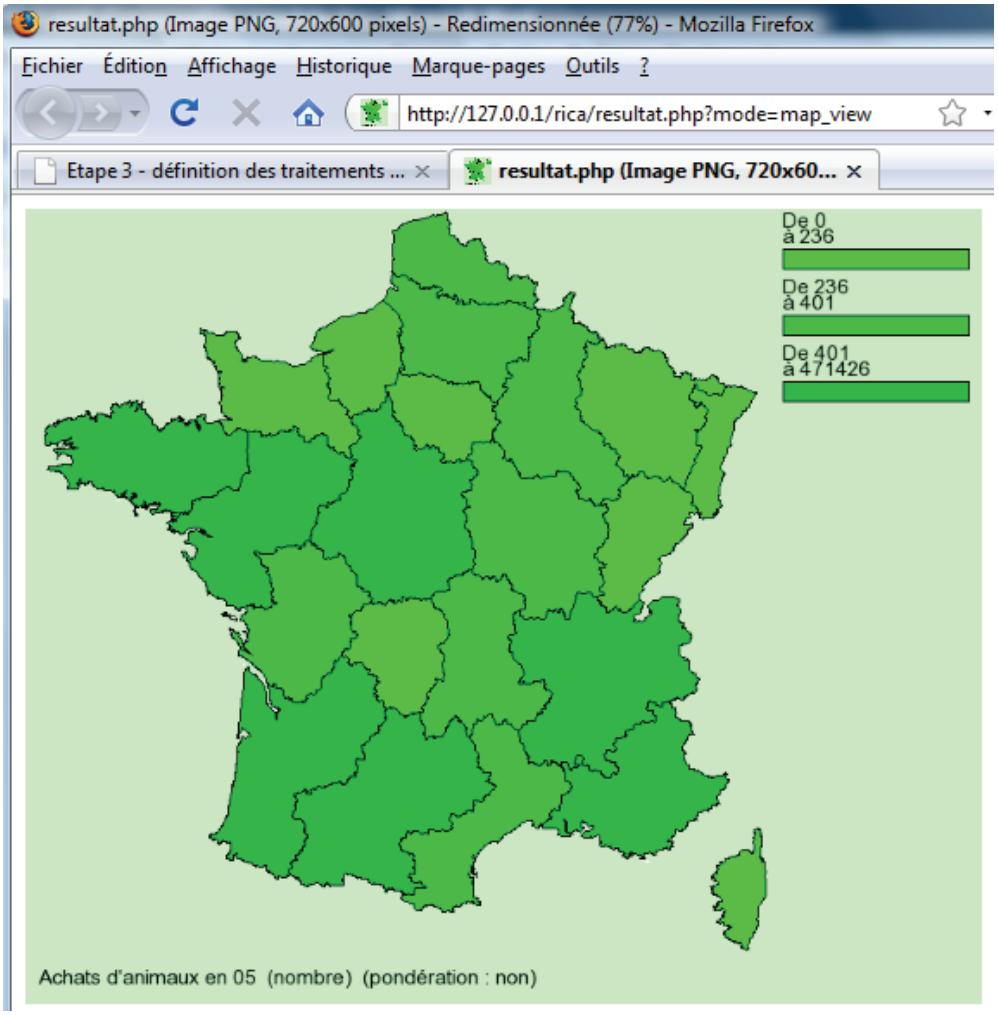
Trois grands choix sont possibles pour l'utilisateur :
- soit sortir une carte ;
- soit faire une extraction des données brutes (au format SAS ou au format texte qui permet une importation sous un tableur) ;
- soit lancer des procédures statistiques simples.

Ces traitements se feront bien entendu sur les variables sélectionnées dans l'étape 1 et sur l'échantillon choisi dans l'étape 2.

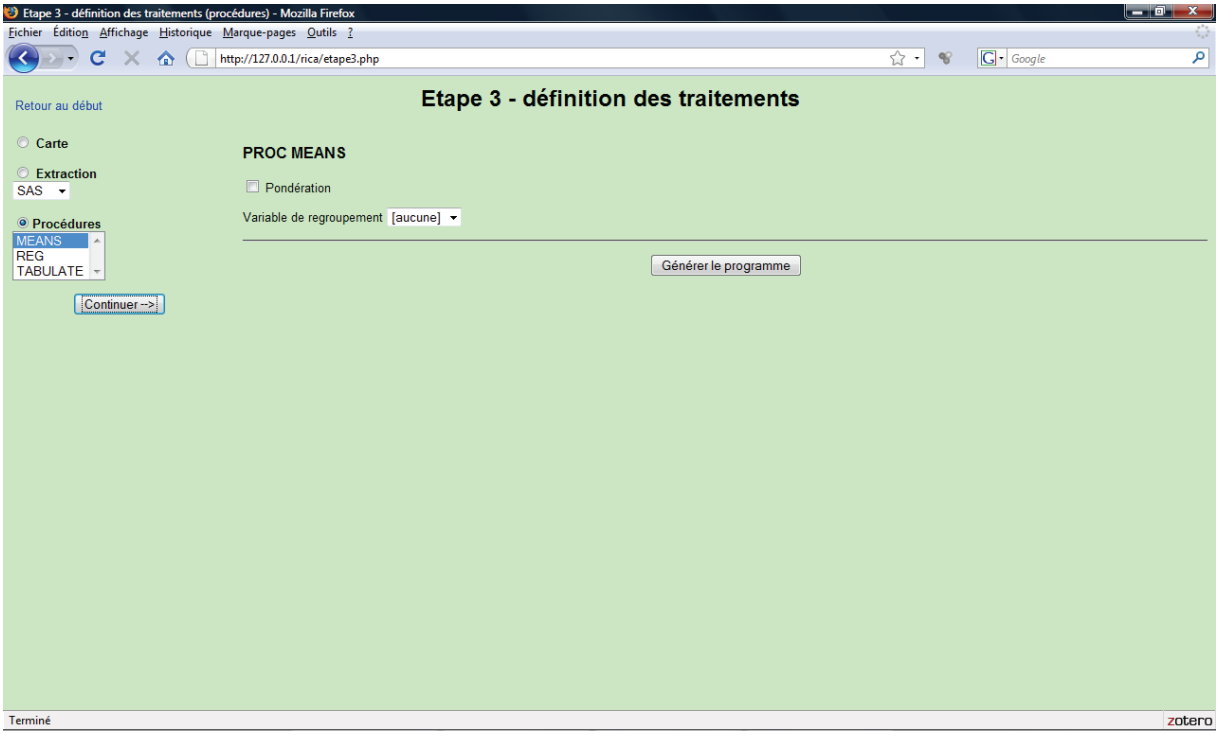
Dans l'exemple qui suit, l'utilisateur demande la sortie d'une carte.



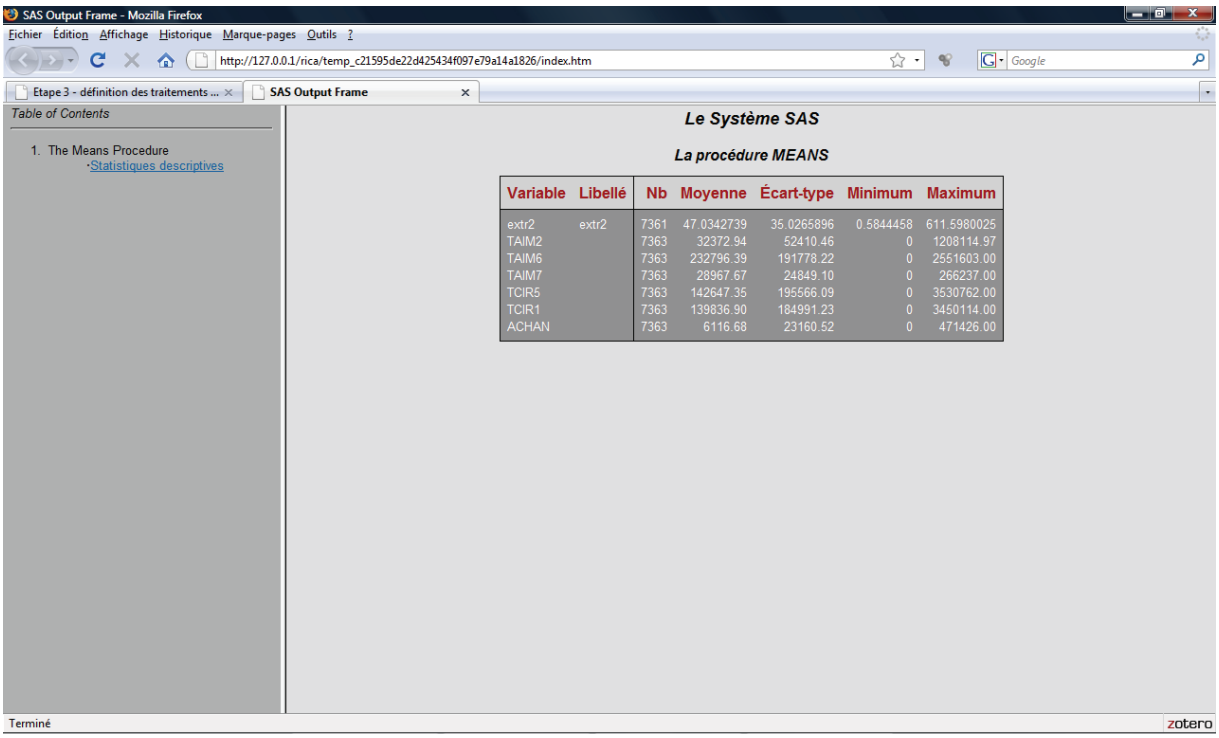
Il obtient une carte comme celle-ci :



S'il choisit un traitement statistique comme dans l'exemple suivant



Il obtient une sortie comme celle-ci :



Il est à noter que tous les traitements peuvent être effectués avec l'application de la variable de pondération et peuvent être lancés en spécifiant une variable de regroupement (sorties par région ou OTEX par exemple).

3.6 Astuces

Entre la sélection du traitement et les résultats, l'utilisateur verra un écran et une étape intermédiaire, lui donnant accès, via un pseudo éditeur de texte, au programme SAS généré par le logiciel. Cette étape, initialement utilisée pour le développement du programme et son débogage, a été conservée pour son aspect pratique. En effet, l'utilisateur peut intervenir directement dans le programme avant que celui-ci ne soit exécuté (ajout d'une variable oubliée, modification du choix de l'échantillon etc.) voire même le récupérer dans son intégralité par une opération « sélection-copier-coller », et l'insérer dans l'environnement du logiciel SAS et ainsi se soustraire de l'écriture d'un programme avec ses règles syntaxiques étant donné que le logiciel le génère lui-même.

3.7 Limites

Comme on peut le constater notre logiciel ne prétend pas vouloir remplacer l'utilisation du logiciel SAS pour exploiter les données du RICA mais il propose un outil simple à ceux qui souhaitent s'initier aux bases du RICA, de le faire sans un gros investissement.
L'utilisation de cet outil reste quand même limitée à l'exploitation des données générales du RICA, car il n'est pas capable pour le moment d'exploiter les données hiérarchisées de productions des exploitations comme les animaux, les produits animaux, les végétaux ou les produits végétaux. Il ne peut également travailler que sur une seule année à la fois

Conclusion et évolution

Le produit pourrait évoluer sans problème en y ajoutant des fonctionnalités comme par exemple travailler sur plusieurs années simultanément et en proposant la possibilité de constituer un échantillon constant. Pour l'instant le logiciel propose uniquement de sortir des cartes, faire des statistiques de bases, des régressions ou des tableaux croisés. Il pourrait donc se voir ajouter des traitements statistiques supplémentaires ou donner les possibilités à l'utilisateur d'exploiter les données concernant les animaux et les végétaux entre autres (avec la possibilité de créer ses propres agrégations).
Pour le moment, au niveau des sorties, apparaît uniquement le nom des variables SAS. Il serait assez facile de mentionner le libellé de ces variables, ce qui serait plus parlant pour les nouveaux utilisateurs du RICA.
Une autre évolution qu'il serait intéressant d'envisager, est la génération de code du logiciel de statistique libre et open-source R. Bien que les données RICA soient au format SAS, R est maintenant capable de lire ce type de données. L'intérêt serait de se soustraire de la contrainte d'avoir une licence du logiciel SAS et ainsi d'utiliser un logiciel gratuit pour exploiter ces données.
L'intérêt de l'utilisation d'un tel outil réside essentiellement dans la possibilité d'avoir un premier contact avec les données des bases du RICA, ceci assez rapidement et en tout cas beaucoup plus facilement que par l'exploitation habituellement complexe des données avec les contraintes d'écriture de programmes sous l'environnement du logiciel SAS.

Annexe

Programme de configuration du logiciel

Ce programme recense les informations nécessaires au bon fonctionnement du logiciel comme les caractéristiques de connexion à la base MySql, qui décrit les variables du RICA, les années disponibles et les couleurs servant au traitement de la sortie carte.

Config.inc.php

```
<?
//
// Fichier de paramétrage de l'interface
// Tous les réglages sont commentés
// Dans les fichiers concernés, on include("config.inc.php"), et les variables
// qui proviennent d'ici sont marquées //PARAMETRE quand elle sont utilisées
//

//années de début et fin des données disponibles, concernent la liste déroulante dans
etape1_1.php
$annee_debut = 1990;
$annee_fin = 2006;

//age maximal (en heures) d'un répertoire temporaire, avant suppression,
//concerne la partie de ménage dans requete.php
$age_limite_rep_temp = 1;

//nombre de conditions maxi définissables dans l'étape 2
//concerne etape2_1.php
$nb_max_conditions = 4;

//Serveur de SGBD, login, password, base à utiliser
$sgbd_server = "localhost";          //serveur MySQL
$sgbd_user_name = "root";            //login
$sgbd_user_pass = "";          //password
$sgbd_basename = "rica";             //nom de la base

//composantes RGB de la couleur de fond de la carte,
//de préférence la même chose que celle de l'interface (cf. ci-dessous)
//mais pas forcément, concernent la carte dans carte.php
$couleur_fond_carte_red = 204;              //équivalent CC
$couleur_fond_carte_green = 255;   //équivalent FF
$couleur_fond_carte_blue = 204;             //équivalent CC

//couleur de fond de l'interface, dans le format RGB hexa (de 000000 à FFFFFF)
//cette couleur se modifie dans normal.css, paragraphes BODY et TD !!!!
?>
```

Numéro spécial

Année 2010

Méthodes et outils de traitement des données en sciences sociales

Retours d'expériences

Sommaire

Participation du département Sciences sociales agriculture et alimentation, espace et environnement - SAE2 -
et du département Sciences pour l'action er le développement - SAD -
Animateur : Éric Cahuzac
Directeur de la Publication : Jean-François Quillien , Responsable du *Cahier des Techniques de l'Inra* : Marie Huyez-Levrat
Photos ©Inra - Impression : Jouve 18, rue saint Denis 75001 Paris - **N° ISSN 0762 7339**
https://intranet.inra.fr/cahier_des_techniques