

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

Jean-Marc Rousselle¹

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 Analysis 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*

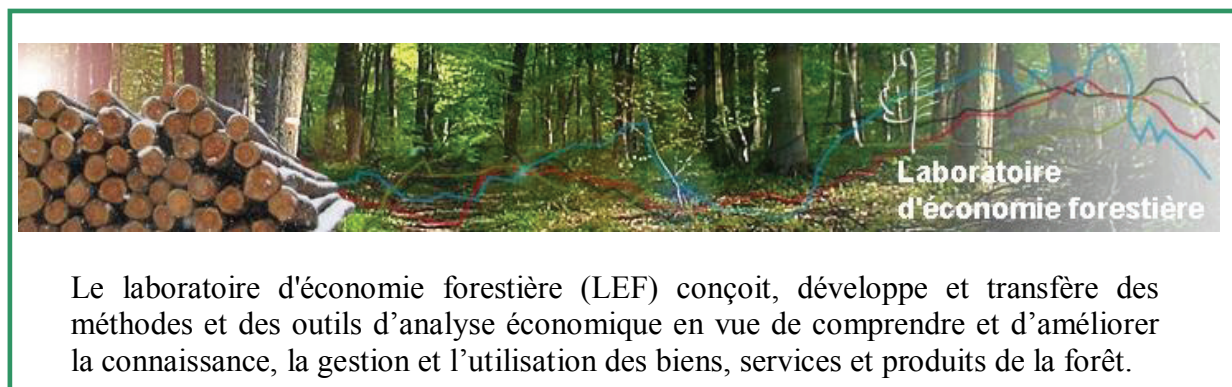
Mots clés : RICA, SAS, bases de données, PHP, Mysql, WEB

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 réseaux 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 est 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.



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 de 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) IBMTM de l'INSEE². 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 Analysis 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 IBMTM 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.

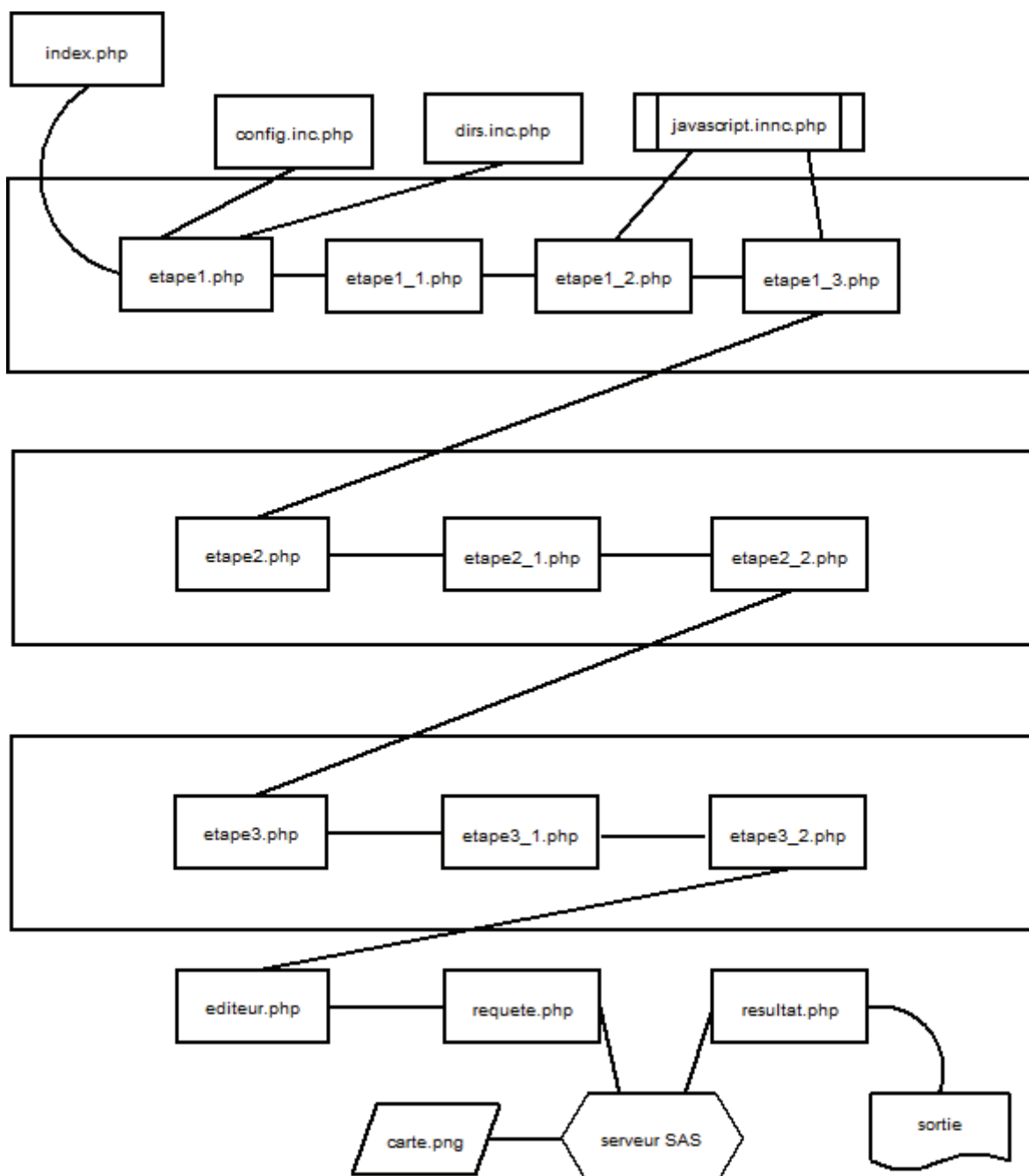
3.1 Développement

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.

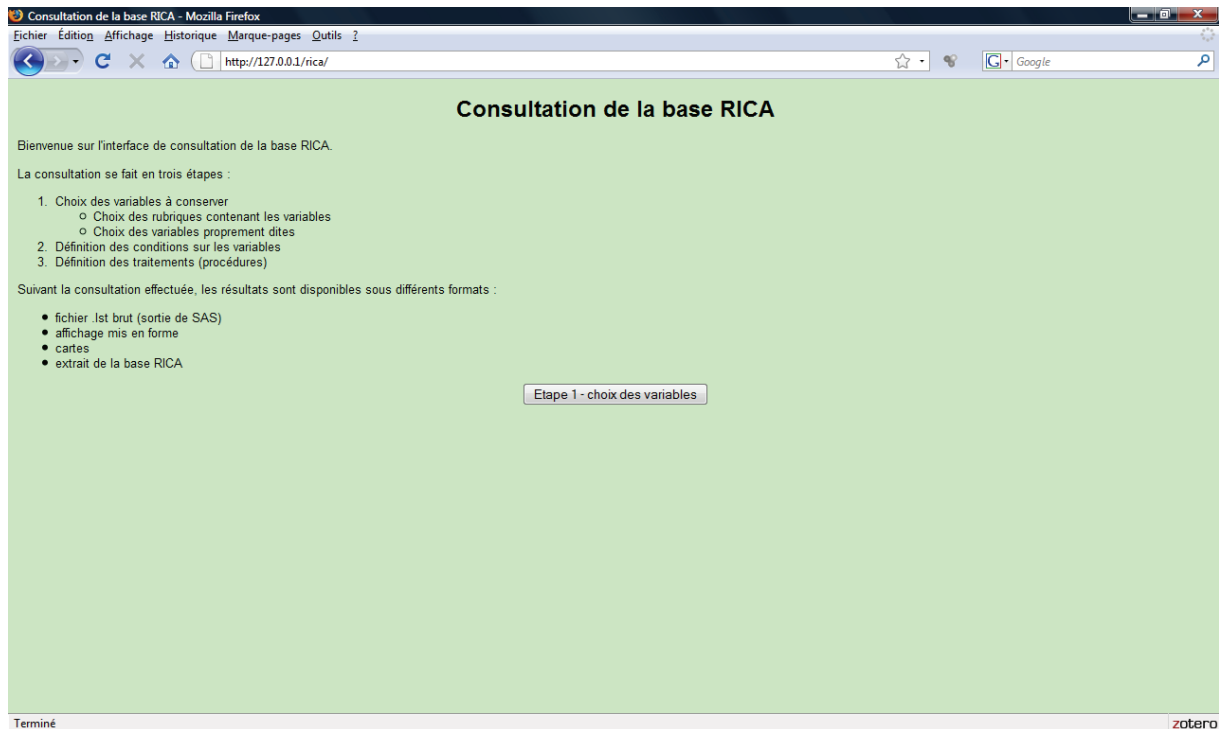
3.5 Utilisation

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.

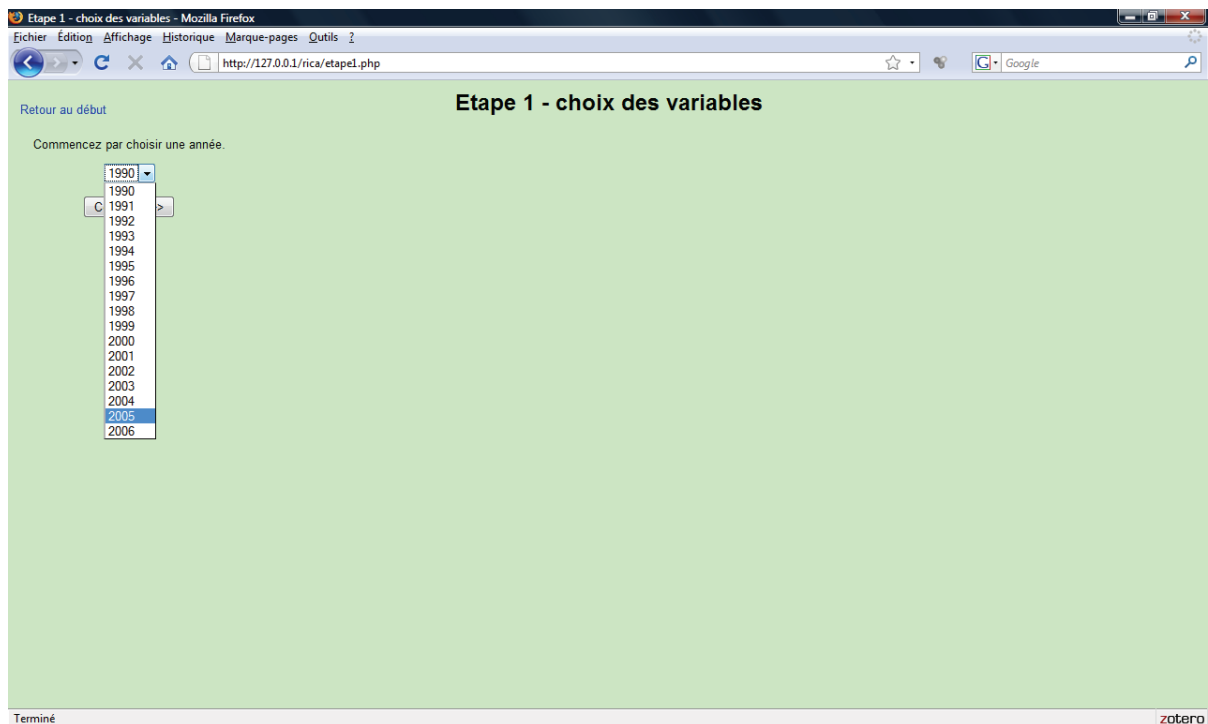
L'utilisateur se retrouve devant trois fenêtres qui s'enchaînent et qui demandent sur quelles données précises il souhaite travailler, sur quelle sélection de l'échantillon et quels résultats ou sorties il souhaite obtenir.

Tout d'abord le logiciel affiche une fenêtre expliquant la démarche à suivre par l'utilisateur pour lui indiquer ses choix.

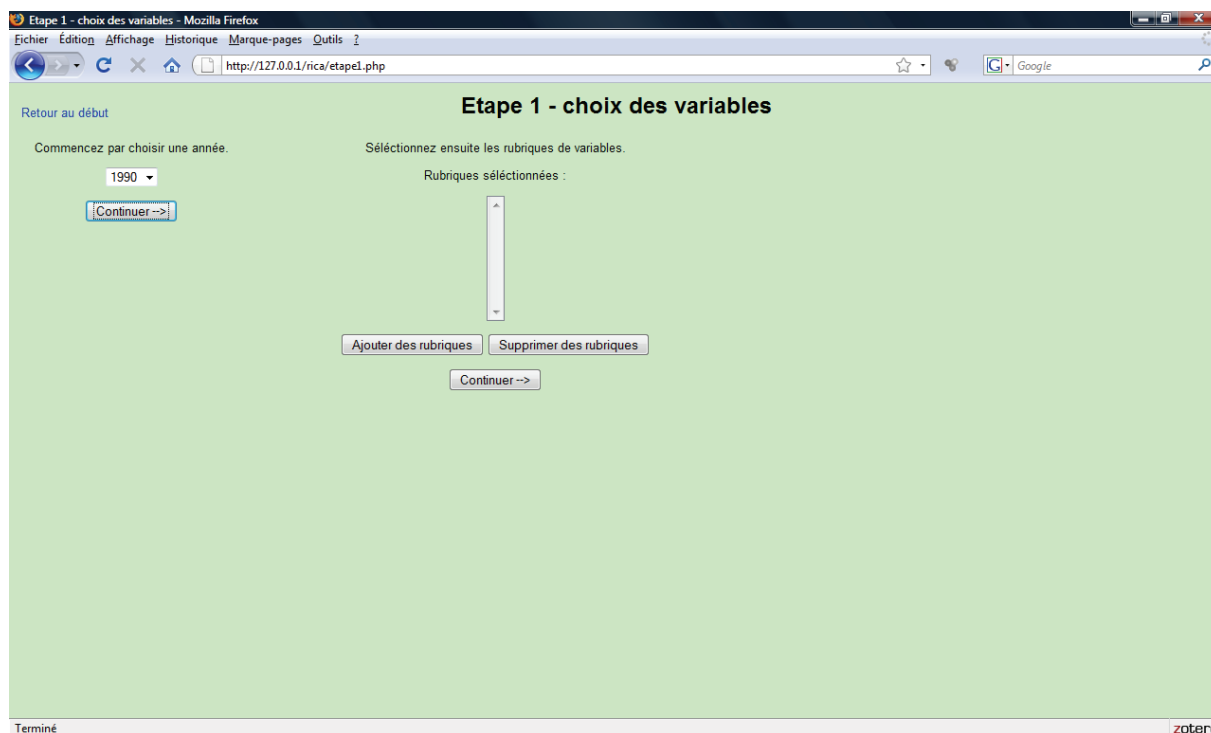
Etape 1 : Sélection de l'année, des rubriques et des variables



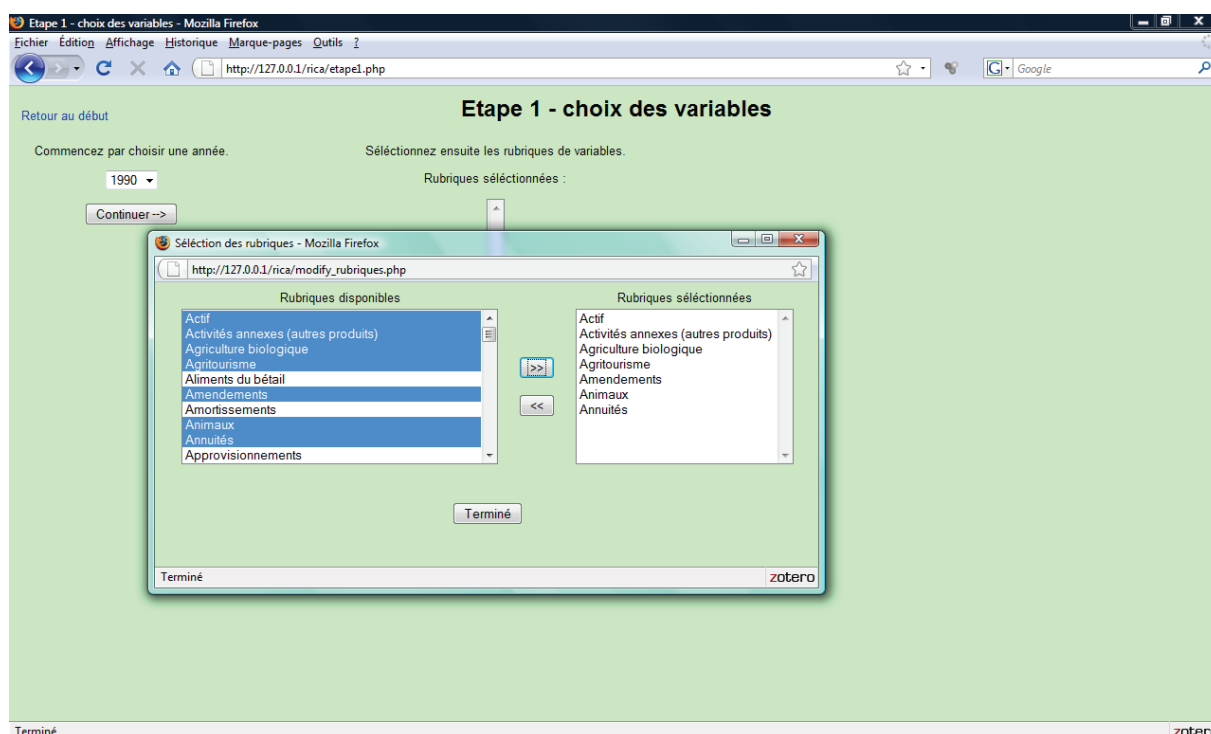
En premier lieu l'utilisateur mentionne sur quelle année il souhaite faire ses traitements. Il lui suffit de sélectionner cette année dans la liste déroulante proposée.



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 ».



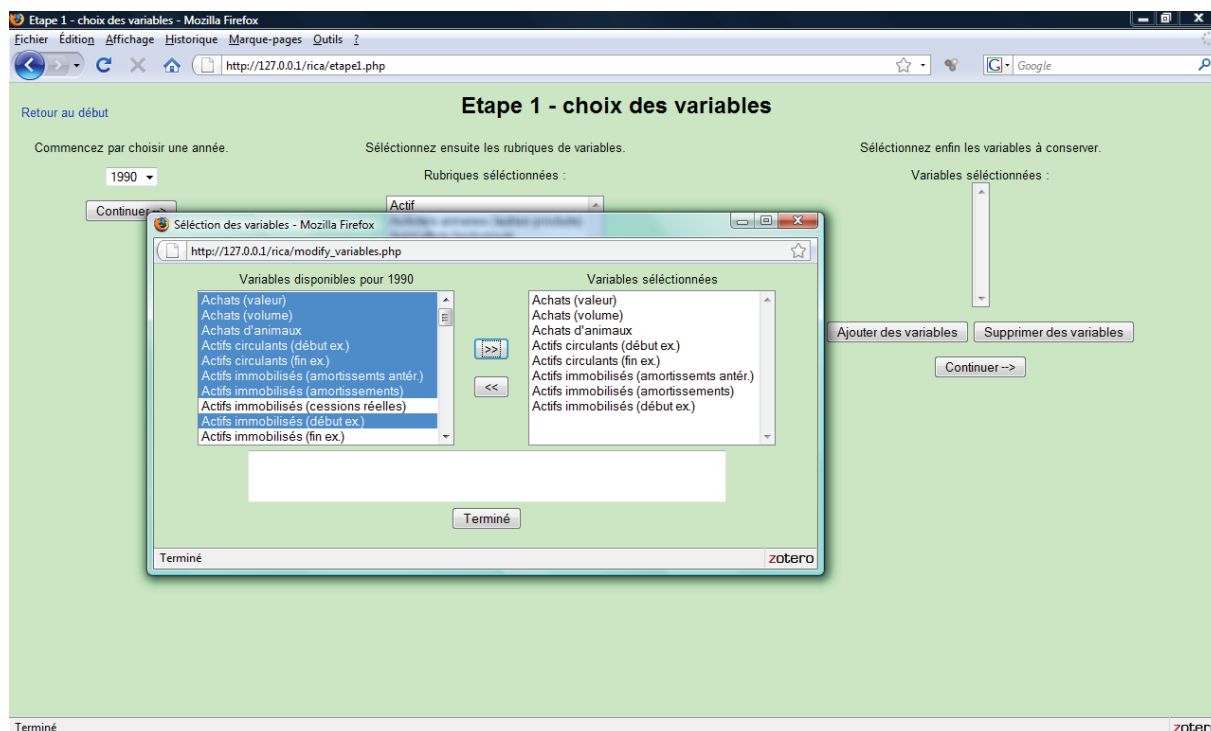
A ce moment là, une petite fenêtre s'ouvre et fournit un outil permettant de sélectionner une ou plusieurs rubriques dans la partie gauche (Il est possible de cliquer en maintenant la touche « Majuscule » du clavier enfoncée pour sélectionner des rubriques contiguës dans la liste ou sur la touche « CTRL » 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).

Etape 2 - définition des conditions sur les variables - Mozilla Firefox
http://127.0.0.1/rica/etape2.php

Retour au début

Etape 2 - définition des conditions

Choisissez les variables puis les conditions qu'elle doivent respecter. Vous pouvez utiliser AND ou OR pour créer des propositions.

Continuer-->

Variable	Opérateur	Valeur
[aucune]	=	
[aucune]	AND OR	
[aucune]	=	
[aucune]	AND OR	
[aucune]	=	
[aucune]	AND OR	
[aucune]	=	

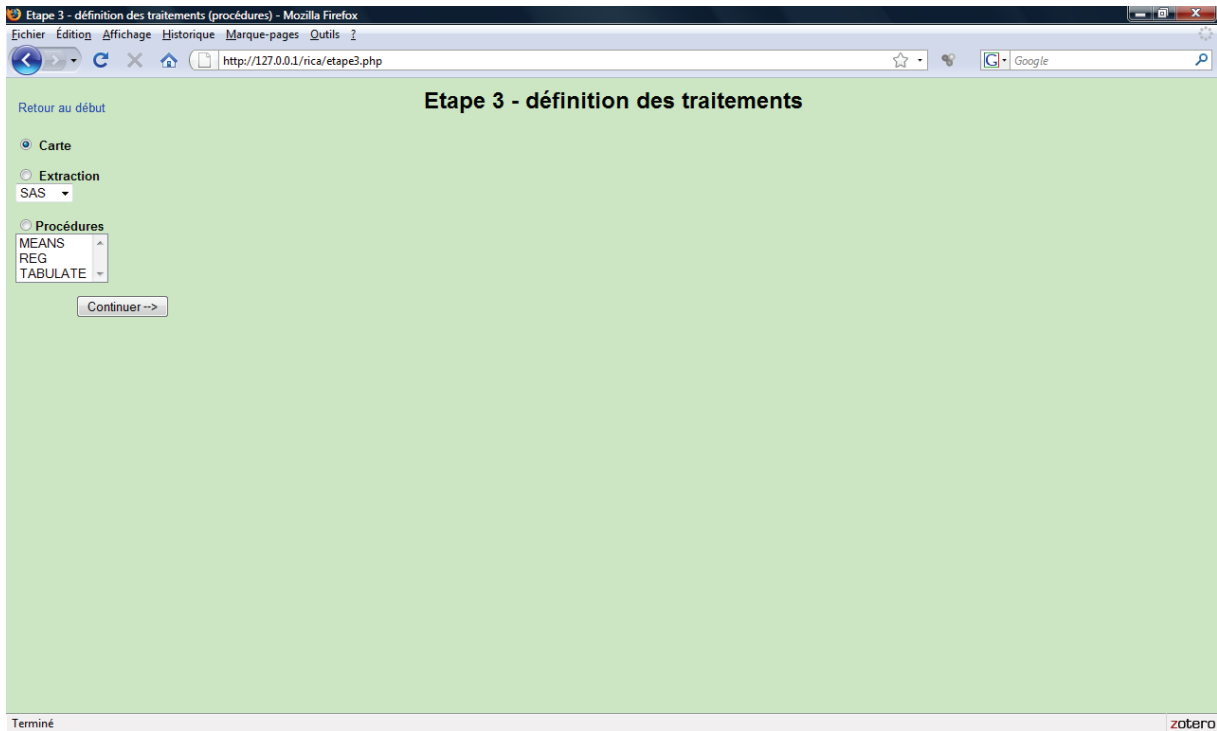
Continuer-->

Terminé zotero

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.

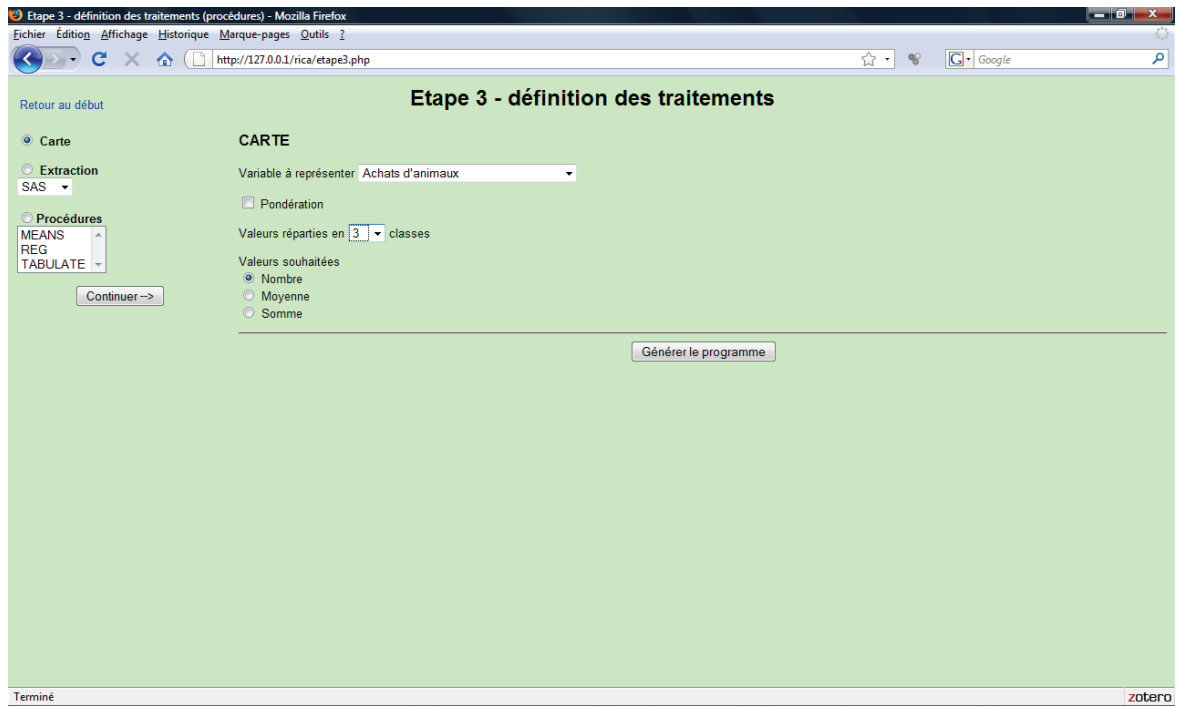


Trois grands choix sont possibles pour l'utilisateur :

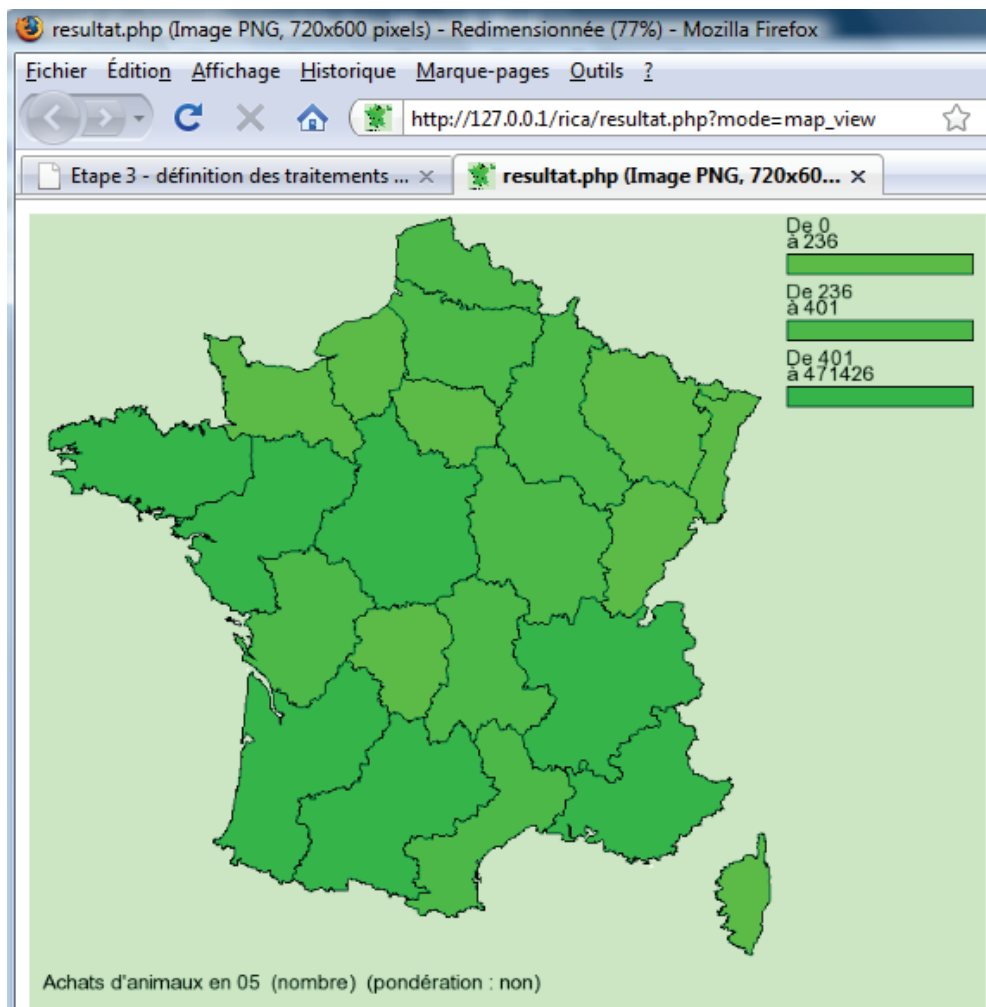
- soit sortir une carte ;
- soit faire une extraction des données brutes (au format SAS ou au format texte qui permet une importation sous un tableur) ;
- soit lancer des procédures statistiques simples.

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

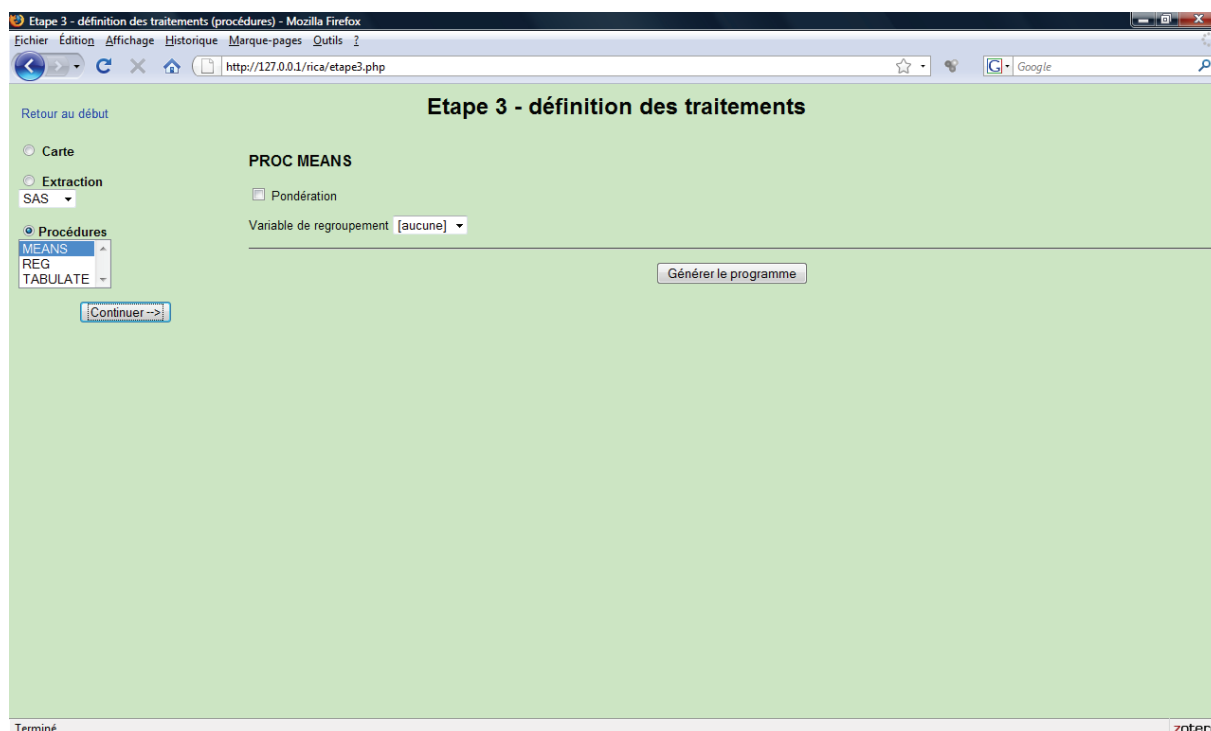
Dans l'exemple qui suit, l'utilisateur demande la sortie d'une carte.



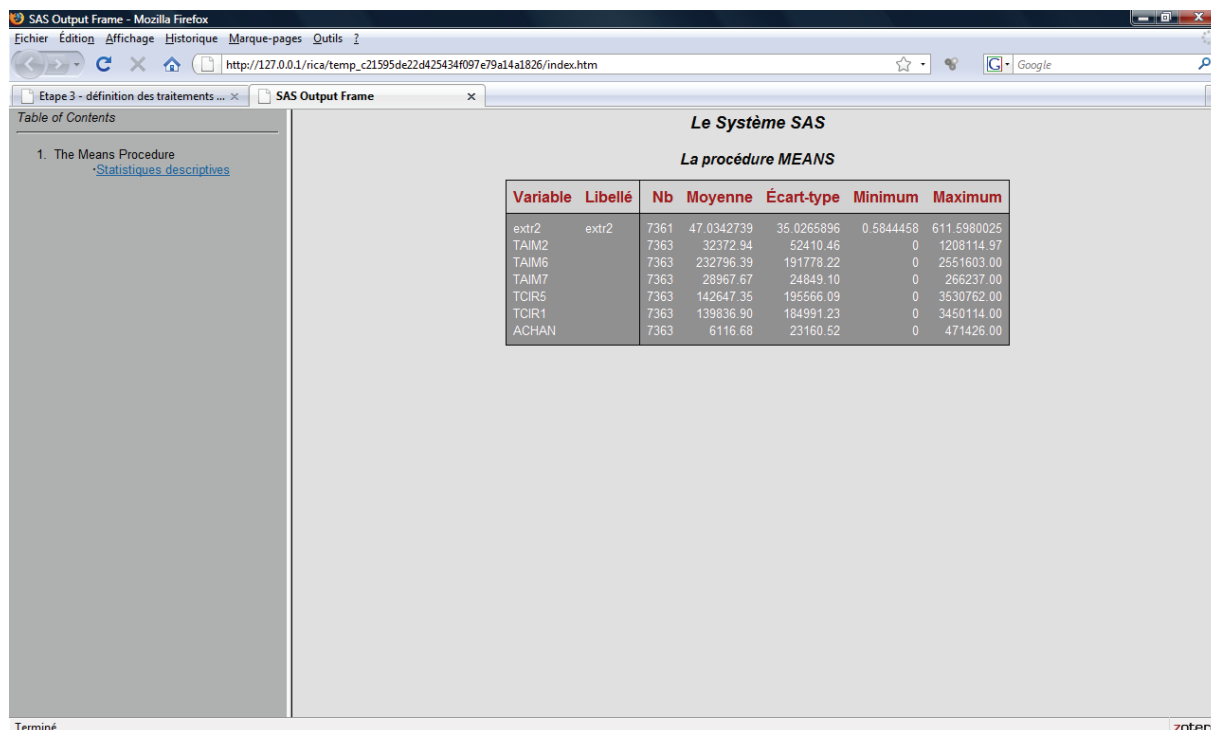
Il obtient une carte comme celle-ci :



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
$sghbd_server = "localhost"; //serveur MySQL
$sghbd_user_name = "root"; //login
$sghbd_user_pass = ""; //password
$sghbd_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 FFFFFFFF)
//cette couleur se modifie dans normal.css, paragraphes BODY et TD !!!!
?>
```