GreenTech Verte

Concours data visualisation

Equipe Superviz

Découvrez le projet dès maintenant sur :

www.agap-sunshine.inra.fr/holtz-apps/GreenTech_Challenge/

L'équipe Superviz
Yan Holtz – Guilhem Marre – Charles Moszkowicz – Jean-Charles Simonin

Présentation de la démarche suivie

Avec une telle quantité de données brutes, les champs d'investigations sont très vastes. Pour ce challenge il s'agissait de faire parler et surtout comprendre l'information au grand public et nous avons dû faire des choix. L'équipe de Superviz a pris cela comme un défi, habitués que nous sommes au raisonnement scientifique et à la représentation des données vraies. Le grand objectif du challenge est de mettre en forme une grande quantité de données brutes afin que le rendu parle de lui-même. L'aspect dynamique du projet nous a tout de suite guidé vers un site web dédié où le visiteur est guidé par des informations d'intérêt tout en laissant aussi la liberté de naviguer choisir ses propres paramètres. Compte tenu des données à notre disposition et de nos connaissances sur les pesticides, l'aspect territorial de la qualité de l'eau nous est apparue comme une part importante du rendu final. D'autre part, nous avons appuyé l'aspect pédagogique sur la compréhension de ce que sont effectivement ces molécules, leur toxicité (LD50) et les nuisances qu'elles peuvent causer à l'homme et à l'environnement. Nous même en tant qu'agronomes savions assez peu de chose sur les pesticides et avons pris plaisir à analyser en profondeur les données proposées. En effet, l'équipe de Superviz est composée d'agronomes dont deux sont spécialisés en développement d'applications pour l'agriculture, le troisième est data scientist expert des graphiques sous R, le dernier est spécialiste en géomatique.

Principes de la solution de data-visualisation proposée

Le projet a été pensé dès le départ comme un portail d'information pour le **grand public**, c'est-à-dire un site qui présente des informations utiles par des représentations originales et facilement **compréhensibles**. Le second objectif du projet était de réaliser un portail **interactif**, dans lequel le visiteur est guidé mais reste libre de s'interroger, rechercher, comparer et apprendre. C'est dans l'action que l'on retient le mieux l'information.

Choix d'implémentations identifiés et retenus

Une des prérogatives du concours était de proposer une solution technique qui soit à la foi visuelle, dynamique et développée dans un langage libre et répandu. Nous nous sommes donc naturellement tournés vers le langage R (R Core Team, 2016) pour plusieurs raisons :

- C'est le meilleur langage pour faire de la manipulation, de l'analyse de données et qui offre un panel de visualisation très varié.
- C'est un langage **open source** et utilisé par une large communauté de scientifiques et de programmateurs, et pour lequel il existe de **très nombreuses librairies** (« packages ») développés par les utilisateurs, qui permettent de réaliser des tâches très variées ;
- De plus, compte tenu du temps imparti pour réaliser le projet, il était important de simplifier au maximum l'architecture informatique et le nombre de langages devant communiquer entre eux. Le socle de notre solution est donc basée sur la librairie « shiny » et le framework de tableau de bord « shinydashboard » qui permettent d'utiliser R, et uniquement du langage R, pour produire aussi bien les analyses, les graphiques et la traduction du code en langages web (HTML, JavaScript).

L'équipe Superviz
Yan Holtz – Guilhem Marre – Charles Moszkowicz – Jean-Charles Simonin

Les derniers aspects évoqués à plusieurs reprises dans le règlement du concours et que nous avons souhaité exploiter au maximum sont **l'interactivité** et **l'originalité** des interfaces graphiques. Au final, une quinzaine de librairies ont été nécessaires à la réalisation de notre portail web et de son contenu visuel varié. Chaque représentation a été pensée pour être la plus visuelle et intelligible possible. Par exemple la cartographie permet de situer les stations proches de son domicile, mais aussi de comparer des zones de répartitions ; les streamgraphs montrent une évolution temporelle ; dans le treemap, la surface est proportionnelle à la présence de chaque pesticide ou famille de pesticide.

Enfin parmi les choix techniques effectués, nous avons porté une attention particulière au temps de chargement des jeux de données. L'ensemble est déjà compilé et sauvé en « *.RData », le format de sauvegarde de R qui a le double mérite d'être très léger et très rapide à charger (environ 3 secondes pour charger l'ensemble des données nécessaires au projet).

Traitement des données

La production de cette application web a d'abord nécessité une **phase importante de prétraitement des données**. Il a en effet fallu comprendre la structure des différentes sources de données, corriger les erreurs d'encodage de certains tableaux de données, ne conserver que les données effectivement utilisées dans l'application,… Parmi les différentes étapes du traitement des données :

- **Compilation** des fichiers des moyennes historiques en un seul tableau ;
- **Nettoyage** des quelques valeurs aberrantes de concentration moyenne *MA_MOY* (i.e largement supérieures au 95ème percentile) ;
- **Re-encodage** de la table des pesticides en UTF-8 (problèmes d'accents) ;
- **Nettoyage** de la table des pesticides (3 doublons avec ID identique et fautes d'orthographe) ;
- **Transformation** du codage de fonction de pesticides (« H » = Herbicide ; « IA » = Insecticide et Acaricide,…) en autant de colonnes que de fonction, remplies avec des 0 et des 1 ;
- **Transformation** des coordonnées des stations (Lambert 93) en coordonnées géographiques (WGS84) pour pouvoir les afficher avec la librairie « leaflet », qui ne gère pas les coordonnées projetées ;
- **Suppression** des stations pour lesquelles aucune mesure n'a été faite ;
- **Chargement** des shapefile des régions et des départements, modification des noms de région avec les nouveaux noms (*sources externes*) ;
- **Jointure** des données de toxicité à la table des pesticides (*source externe*);
- **Agrégation** des données de concentration par niveau géographique (station, département, région), par année (2007-2012), par fonction (herbicide,…) et par famille (organochlorés,…).

Cette dernière étape est extrêmement importante. En effet, compte tenu du grand nombre de mesures de concentration (près de 3,5 millions de lignes), chaque agrégation nécessite un temps de calcul considérable. Au final, une table est produite par niveau géographique (station, département, région), et il ne reste qu'à faire une extraction

conditionnelle à la volée (ANNEE==2007, FONCTION== » Herbicide »,…), une opération presque instantanée. Par ailleurs, d'autres indicateurs comme la concentration totale en pesticides ou le nombre de molécules non autorisées détectées sont calculés à l'avance et intégrés à ces bases de données.

NB : concernant le shapefile des masses d'eau souterraines, nous ne nous en sommes finalement pas servi. Les **erreurs dites « géométriques »** qu'il contenait ne sont pas tolérées par la librairie « leaflet » utilisée pour toutes les cartes interactives.

NB : nous avons utilisé plusieurs **bases de données externes**, mais toutes en accès libre :

- Shapefile des régions et des départements (OpenStreetMap) ;
- Fond de carte Google et OpenStreetMap via la librairie « leaflet » ;
- Toxicité des pesticides (LD50)

Prérequis techniques

L'ensemble de l'application est développé en langage « R » et ne nécessite donc que l'installation du logiciel R et des librairies utilisées. Par ailleurs, aucune connaissance particulière du langage R n'est requise pour faire fonctionner l'application. Dans le cas d'une mise à jour des données (ajout d'années de mesures ou ajout de nouveaux pesticides) il suffit d'exécuter un script existant. Là encore pas besoin de connaissances particulières si ce n'est une certaine rigueur dans la gestion des données (fichiers construits de manière identique, dans le même format, …).

Pour installer R et les librairies utilisées :

- Télécharger et installer R depuis le site du CRAN : https://cran.r-project.org/
- Ouvrir une fenêtre R et exécuter les lignes de code suivantes

```
install.packages(c( "shiny", "sp", "plotly", "ggplot2", "DT",
"RColorBrewer", "devtools", "leaflet", "ggmap", "tidyr",
"shinydashboard", "shinyjs"))

library(devtools)

install_github("hrbrmstr/streamgraph")
install_github("mtennekes/treemap", subdir="pkg")
install_github("timelyportfolio/d3treeR")
```

Lancer l'application en local :

- Télécharger et placer le dossier « GreenTech_Challenge » à l'endroit désiré
- Ouvrir une fenetre R et exécuter les lignes suivantes

```
library(shiny)

runApp("mon_chemin/GreenTech-Challenge")
```

Installer l'application sur serveur :

En plus de l'installation de R, il est nécessaire de télécharger et installer un **serveur shiny**. Vous trouverez toute la documentation sur le site de RStudio :
https://www.rstudio.com/products/shiny/download-server/

Intégration de nouvelles données

Le code a été écrit de telle sorte qu'il soit indépendant du nombre de données d'entrée, si bien que l'ajout d'un nouveau fichier de mesures ou la modification de la liste des pesticides sont pris en compte de manière transparente. Il suffit à un administrateur de placer les nouveaux fichiers dans les dossiers adéquats et d'exécuter le script de préparation des données (RESSOURCES/0_Prepare_data_Greentech.R). Celui-ci fait appel aux différentes sources de données : liste des pesticides, liste des stations, et les mesures des concentrations par station et par année. L'exécution du script met à jour le fichier d'environnement R « env_greentech.R ». Pour ce faire, ouvrir une fenêtre R et lancer la ligne de commande suivante :

```
source("CHEMIN_RACINE/RESSOURCES/0_Prepare_data_Greentech.R")
```

D'autre part, la visualisation des données dans l'application est elle aussi gérée de manière totalement indépendante de la quantité de données. Ainsi la légende s'adapte automatiquement aux valeurs reçues, les paramètres des graphiques (années de mesures, fonctions, familles et nom des pesticides) se mettent à jour en fonction des données de l'environnement R. Par exemple, les cartes du package « leaflet » fonctionnent par couche additives et nécessitent une couche par année de mesure. Au lieu de construire l'objet en écrivant « en dur » chaque couche dans le code, une boucle génère les portions de code en fonction des années disponibles.

Perspectives

Comme dans tout projet, il existe malheureusement une différence entre ce que l'on souhaite réaliser dans un idéal pourtant vraisemblable, et ce qu'il est réellement possible de faire compte tenu des contraintes de temps et de disponibilité de la donnée. Parmi les perspectives d'évolution envisagées pour notre application, nous avons pensé à :

1. **Prendre en compte les pratiques agricoles**

Agronomes de formation, la corrélation entre les concentrations en pesticides dans les eaux souterraines et les pratiques agricoles nous est apparue un sujet important. Après quelques investigations sur la répartition des cultures, les types de sols et les maladies caractéristiques des cultures, nous avons fait le bilan que ce travail était d'une grande ampleur et que la durée du concours ne nous permettait pas d'aller assez loin. C'est cependant un champ d'investigation que nous aimerions creuser.

2. **Visualisation des masses d'eau souterraines**

Nous n'avons pas pu utiliser le shapefile des masses d'eau souterraines à cause des erreurs de structure géométrique des polygones. Aucun des algorithmes de traitement de ce type d'erreurs n'est venu à bout de l'ensemble des erreurs du fichier. Pour autant, avec un shapefile débarrassé d'erreurs géométriques, il serait tout à fait envisageable d'avoir un rendu cartographique par masse d'eau, de la même manière que nous l'avons fait par département ou par région.

L'équipe Superviz
Yan Holtz – Guilhem Marre – Charles Moszkowicz – Jean-Charles Simonin

3. Améliorer le temps d'affichage des cartes

Malgré nos efforts pour préparer au mieux les données, faire les calculs les plus lourds en amont et éviter les chargements trop longs certaines cartes mettent toujours environ 5 secondes à s'afficher. Une solution pour gagner 2 ou 3 secondes à l'affichage des cartes les plus lourdes serait de préparer à l'avance toutes les cartes possibles (combinaisons de chaque unité géographique, chaque niveau (tous, molécule, famille, fonction) pour chaque année et de les enregistrer au format *.RData.

4. Apparence

Concernant l'apparence générale l'application, nous avons utilisé le framework de la librairie « shinydashboard » ajusté à l'aide d'un fichier de mise en forme au format *.css. Avec plus de temps, nous aurions produit notre propre modèle de mise en page web pour un rendu plus original, ergonomique et souple d'utilisation.

Bibliographie et information sur les pesticides

T. Blacquière, G. Smagghe, C. A. M. van Gestel, V. Mommaerts, Ecotoxicology 21, 973–992 (2012).
M. A. Fleischli, J. C. Franson, N. J. Thomas, D. L. Finley, W. Riley Jr., Arch. Environ. Contam. Toxicol. 46, 542–550 (2004).
F. Brucker-Davis, Thyroid 8, 827–856 (1998).
R. McKinlay, J. A. Plant, J. N. B. Bell, N. Voulvoulis, Environ. Int. 34, 168–183 (2008).
T. S. Galloway, M. H. Depledge, Ecotoxicology 10, 5–23 (2001).
L. Gawade, S. S. Dadarkar, R. Husain, M. Gatne, Food Chem. Toxicol. 51, 61–70 (2013).
P. C. Lin, H. J. Lin, Y. Y. Liao, H. R. Guo, K. T. Chen, Basic Clin. Pharmacol. Toxicol. 112, 282–286 (2013).
H. R Kohler, R. Triebskorn, Wildlife Ecotoxicology of Pesticides: Can We Track Effects to the Population Level and Beyond? Science (2013).
S. Mañosa, R. Mateo, R. Guitart, Environ. Monit. Assess. 71, 187–205 (2001).
D. A. Crain, L. J. Guillette Jr., Anim. Reprod. Sci. 53, 77–86 (1998).
T. Farooqui, Neurochem. Int. 62, 122–136 (2013).
L. P. Belzunces, S. Tchamitchian, J.-L. Brunet, Apidologie (Celle) 43, 348–370 (2012).
R. J. Gill, O. Ramos-Rodriguez, N. E. Raine, Nature 491, 105–108 (2012).
P. R. Whitehorn, S. O'Connor, F. L. Wackers, D. Goulson, Science 336, 351–352 (2012)

L'équipe Superviz
Yan Holtz – Guilhem Marre – Charles Moszkowicz – Jean-Charles Simonin