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 du 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

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'autre aspect évoqué à plusieurs reprises dans le règlement du concours et que nous avons souhaité exploiter au maximum est **l'interactivité** des interfaces graphiques. En parallèle des fonctions de la librairie R **« shiny »** qui permettent d'avoir un contenu web interactif (boutons, curseurs, liens), nous avons réalisé les graphiques à l'aide de **librairies particulières** qui permettent de générer du contenu interactif : leaflet (cartes interactives utilisant des fonds de carte externes), ggplot2 (graphiques complexes et conditionnels), plotly (permet d'animer les graphiques produits avec ggplot), streamgraph (graphiques interactifs de courbes pleines), treemap (cartes à cases dynamiques). Au final, une quinzaine de librairies ont été nécessaires à la réalisation de notre application web et de son contenu visuel varié : cartes, diagrammes en barres, courbes multiples, streamgraphes, bubbleplot, treemap.

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

Le chargement des librairies se fait au moment de l'ouverture de la page, conjointement au jeu de données, déjà compilé et sauvé en « *.RData », le format de sauvagarde 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 de codage 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) ;
- **Recodage** 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 (latitude, longitude) pour pouvoir les afficher avec la librairie « leaflet » ;
- **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, car compte tenu du grand nombre de mesures de concentration (2,8 millions de lignes), chaque agrégation nécessite un temps de calcul considérable. Au final, une table est produite par niveau géographique, et il ne reste qu'à faire une extraction conditionnelle à la volée (ANNEE==2007, FONCTION== « Herbicide »,…), une opération presque instantanée.

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 pastolérées 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)

Description de l'application

Arrivé sur la page d'accueil, l'utilisateur trouve les grands chiffres clés du réseau de suivi de la qualité des eaux (nombre de stations, nombre de pesticides, nombre d'années de suivi) et choisi d'avoir des informations sur la station la plus proche de chez lui (bouton permettant de se géolocaliser), de voir l'historique des données dans la France entière et d'en apprendre plus sur la toxicité des pesticides.

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

En regardant sa station, l'utilisateur voit d'un coup d'œil si l'eau près de chez lui est contaminée et l'évolution de la situation par rapport à la mesure précédente.

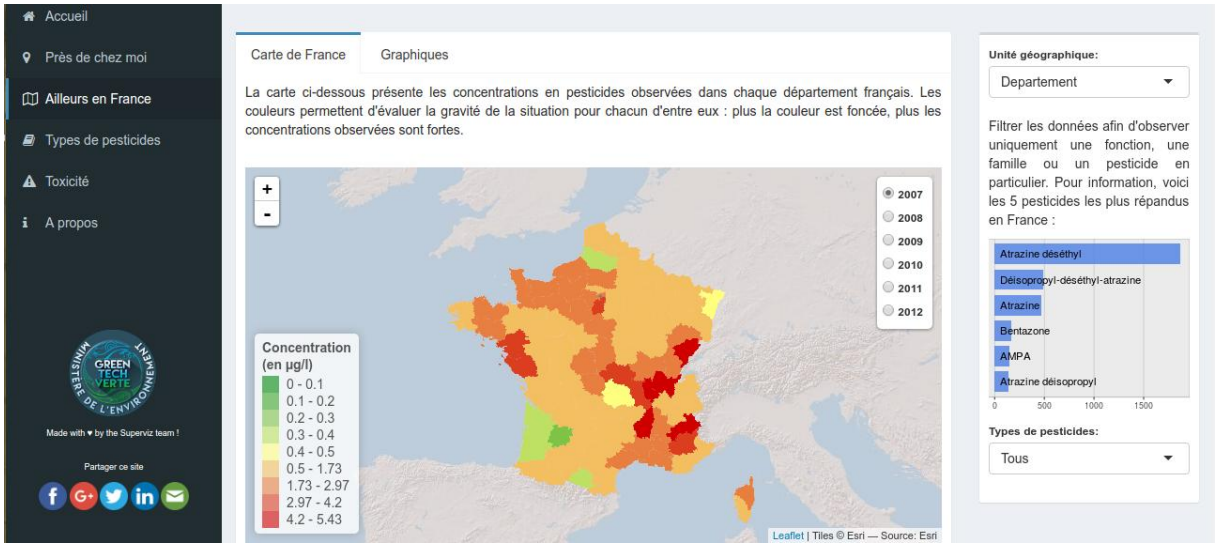
Onglet « Près de chez moi »

L'idée est aussi de connaitre l'évolution des quantités de pesticides en fonctions des années. Ici nous avons utilisé un streamgraph dynamique qui permet en un clin d'oeil de visualiser les quantités de chaque pesticide individuellement et les uns par rapport aux autres.

Ensuite deux choix s'opèrent: soit on veut en savoir plus sur les pesticides en général, soit savoir quelle est l'état de l'eau ailleurs en France.

« Ailleurs en France » présente par défaut la concentration totale en pesticides observée dans chaque département français. Il est possible de faire varier l'unité géographique, de restreindre les données sources en fonction de catégories et voyager d'année en année.

Onglet « Ailleurs en France »



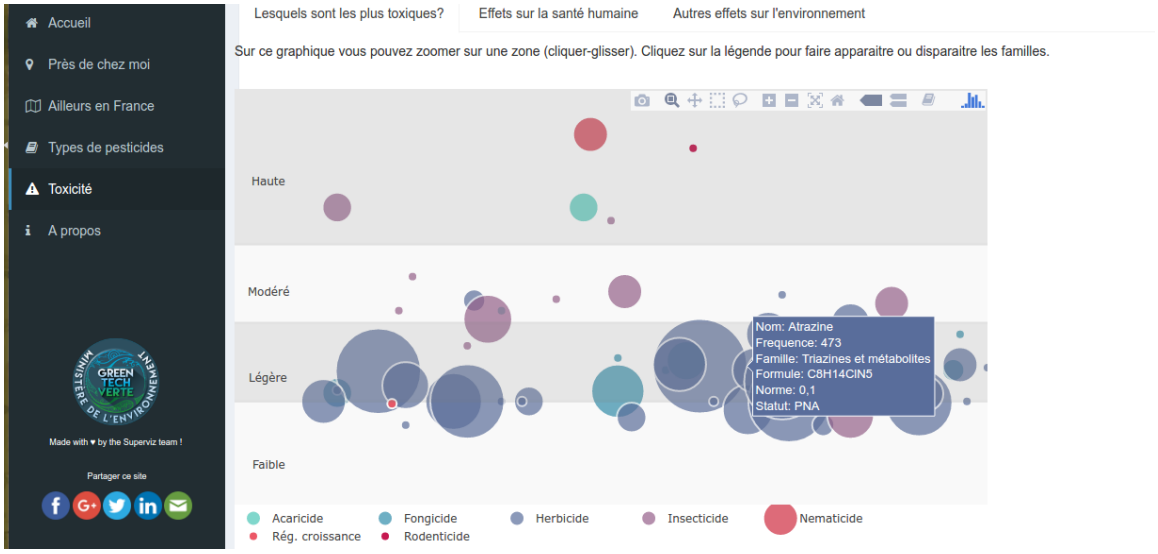
L'utilisateur du Lot-et-Garonne se dit ainsi que la qualité de son eau était ben supérieure à celle de l'habitant du Doubs en 2007 !

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

Onglet « Types de pesticides »

L'onglet des pesticides présente les informations totales mesurées sur la France. Le « treemap » est une représentation surfacique ou chaque carré représente la fréquence de détection de la molécule. On observe ainsi qu'une grande majorité des molécules retrouvées dans les nappes sont des herbicides.

Onglet « Toxicité »



Enfin, il nous a paru important de donner du sens au nom des molécules analysées et présentées dans la solution. C'est pourquoi le dernier onglet porte sur la toxicité des produits dans un « bubble chart » à ordonnée logarithmique accompagné de nombreuses informations sur les différents effets des pesticides.

Perspectives

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

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 : **prendre en compte les pratiques agricoles (**proposer un outil qui permette à l'utilisateur de parcourir les données de concentration des pesticides en parallèle de cartes de pratiques agricoles), **visualiser les masses d'eau en 3D** Un des axes proposés pour la visualisation des données suggérait de développer un outil permettant de visualiser la qualité des nappes d'eau, idéalement strate par strate (« en 3D »). En prenant le temps de nettoyer ces données, il serait possible de faire une carte avec un double curseur, le premier pour les années comme c'est déjà le cas, et un autre pour le niveau de profondeur de la masse d'eau ; **diminuer le temps d'affichage des cartes** (préparer à l'avance toutes les cartes possibles et les sauver en .RData) ; **charte graphique** (produire notre propre charte graphique) ; **bibliographie et information sur les pesticides** (continuer l'effort déjà mené de recherche).

Bibliographie et information sur les pesticides

, 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).
, 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