ISSN 1620-4989

Géomatique

Expert

QUALITÉ DES DONNÉES
DATA QUALITY

FOSS4G Europe

SIG et IA / GIS and AI

Données OSM / OSM Data

N° 118 - Septembre-Octobre 2017 - 14 €

SIG 2017

Conférence Francophone Esri
11 & 12 octobre - Les Docks de Paris

Venez participer au 1er événement francophone dédié à l'Information Géographique

Un événement organisé par

esri France
THE SCIENCE OF WHERE™

Programme et inscriptions
https://sig2017.esrifrance.fr

N° 118 - Septembre-Octobre 2017

Image de couverture : Aqualagon (architecte Jacques Ferrier). Cliché L'Europe Vue Du Ciel pour Village Nature Paris.

Éditeur :
CiMax • 12, place G. Pompidou • 93167 Noisy-le-grand Cedex
Tél. : 01 45 92 96 96
Fax : 01 49 32 10 74
Email : geomatique@cimaxonline.fr

Directrice de la publication :
Anne-Caroline Prévot-Leygonie

Rédaction : Vincent Habchi

PAO et maquette : PAO Corlet

Merci à : OSGEO Europe, ENSG, 1Spatial, Isogéo, CG 14

Ont collaboré à ce numéro :
C. Carolin, D. Teixeira Alves da Sylva, A. Serradj, A. Do Vale Figueiredo, V. Backer, D. Garaud, R. Delhome, H. Mercier

Publicité : Sébastien Guénée
Tél : 01 45 92 99 96
Email : s.guenee@groupe-cimax.fr

Abonnements :
Ana Dos Santos
Tél : 01 45 92 98 98
France métropolitaine : 70 €
Communauté européenne : 85 €
Reste du monde : 95 €

Flashage et Impression :
CORLET Imprimeur
14110 Condé-sur-Noireau
ISSN 1620-4909

Commission paritaire : 0618 T 79345
Dépôt légal à parution

Géo | Plastique

Le plastique, c'est le matériau de l'ère moderne. Bouteilles, jouets de toutes sortes, appareils électroménagers, poêles, téléphones, ordinateurs, voitures et peut-être bientôt même trains et avions ? Bref, le plastique est partout : mou, dur, épais, fin, transparent, blanc, bleu, vert, rose, noir, imitation ronce de noyer, etc. Notre univers est noyé sous les polymères d'hydrocarbures.

Il faut dire que la matière présente des avantages : facile à produire, facile à usiner ou à mouler, isolante, avec des propriétés physicochimiques très différentes suivant le monomère de base. Bonne à (presque) tout faire, ses applications sont légions. On estime que, depuis les années 50, ce sont environ neuf milliards de tonnes de plastiques qui ont été produites.

Le problème, c'est de savoir que faire de ces neuf milliards de tonnes quand elles deviennent indésirables.

Tout le monde a present à l'esprit ces images (choquantes ?) de bouteilles ou sac déparant des sites naturels préservés, en montagne ou sur les plages. Et ce n'est rien en comparaison de ce que les scientifiques ont relevé en pleine mer. Les plages de l'île Henderson, un îlot inhabité du Pacifique sud, sont ainsi recouvertes de 38 millions de détritus plastiques représentant environ 18 tonnes. Treize mille nouvelles pièces s'y échouent tous les jours !

Actuellement, seuls 20 % de la production sont recyclés. Or, les plastiques se désagrègent en des morceaux minuscules, qui sont ensuite ingérées par la faune… ou pour certains s'envolent et se retrouvent dans l'air (une étude menée en 2015 à Paris a montré qu'en moyenne, tous les ans, ce sont dix tonnes de poussières plastiques qui se déposent dans la capitale) et même… dans nos verres ?

C'est du moins ce que semble prouver une récente étude commanditée par le quotidien anglais The Guardian. Celui-ci a fait procéder à des analyses d'eau du robinet partout sur la planète. Résultat : la plupart (83 %) des eaux données comme potables, et même certaines eaux « minérales », contiennent des résidus de plastique. Aux États-Unis, le taux de contamination atteint 94 %. La France, à égalité avec le Royaume-Uni et l'Allemagne, présente fort heureusement le taux le plus faible, à savoir… 72 % !

Certes, ce n'est pas encore une contamination intense. En moyenne, en Europe, on ne trouve environ que quatre fibres par litre d'eau. Mais c'est peut-être l'arbre qui cache la forêt. Car, disent les scientifiques, si les fibres ne sont peut-être pas nocives en elles-mêmes, elles peuvent transporter des produits toxiques ou des bactéries pathogènes qui pourraient ensuite être relâchés dans les intestins. En outre, s'il y a des microfibres plastiques dans l'eau, que dire des nanoparticules ?

Comment ces fibres sont-elles arrivées là ? C'est encore un mystère, mais certains pointent du doigt la confection, par exemple les habits en « polaire ». Cette matière, faite à partir de plastiques déchiquetés, peluche aisément. Une étude a ainsi montré que certaines polaires peuvent relâcher environ 250.000 microfibres par lavage. Sans parler des séchoirs qui larguent aux quatre vents des dizaines de milliers de ces poussières.

Changement climatique, déforestation, lutte contre le plastique. Ce ne sont pas les défis qui manquent pour la prochaine génération.

Vincent Habchi

Ce numéro étant distribué sur la conférence InterGeo de Berlin, certains articles sont écrits en français et en anglais. Retour à la normale au prochain numéro

1spatial
YOUR WORLD SMARTER

Nous rendons vos données plus intelligentes

1integrate

Logiciel breveté pour la création, la transformation, l'amélioration et le contrôle qualité de données géographiques
Existe aussi pour ArcGIS

elyxsuite

Logiciel pour l'exploitation de données géographiques visant la gestion territoriale et la gestion de réseaux

Pour en savoir plus : 1spatial.com
T : +33 (0) 1 71 33 01 00 | E : info@1spatial.com

sommaire / table of contents

n° 118 |
Septembre-Octobre 2017

Actus/News

SIG/GIS

Carto/Mapping

Mobilité/ Mobile data

QUALITE DES DONNEES / DATA QUALITY

Ce numéro comporte un encart abonnement broché 2 volets non foliotés.

La production numérique surpasse la lithographie

L'installation de la cinquième imprimante HP PageWide XL 8000 grand format achève la transformation vers le numérique

Dix ans après que le Bureau Hydrographique du Royaume-Uni n'ait entrepris la transition de la production d'impression analogique vers l'impression numérique, le voyage prend fin avec l'installation de la cinquième imprimante HP PageWide XL 8000 grand format – et en fait le site PageWide XL le plus important en Europe.

Le Bureau Hydrographique du Royaume-Uni produit et vend à l'échelle mondiale une série de 3500 cartes marines papier avec l'objectif principal de protéger les vies en mer.

« Nous ne vendons jamais de cartes obsolètes », affirme Paul Kelly, chef de production chez UKHO.
« Des mises à jour peuvent être effectuées sur n'importe quelle carte et ce, à tout moment, ce qui nous oblige à renouveler nos stocks. L'un des avantages évidents de passer à l'impression numérique et à « l'impression à la demande » consiste à se débarrasser de ces stocks.»

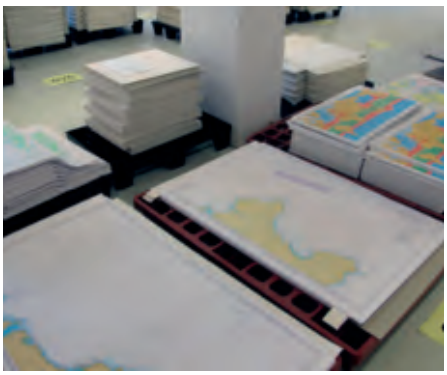
Opérateur du Bureau Hydrographique du Royaume-Uni avec une carte imprimée sur l'imprimante HP PageWide XL 8000.

Des avantages visibles immédiatement

Un autre effet indirect positif généré par le passage au numérique est la réduction de l'espace au sol. L'équipement d'impression qui remplissait précédemment l'équivalent de trois terrains de football (machines CTP machines, machines lithographiques et tous les autres produits de finition) est réduit à seulement onze m² par imprimante PageWide XL. En outre, le UKHO a pu se débarrasser de la plupart des « éléments néfastes » – les substances chimiques dangereuses appliquées dans le cadre du processus d'impression lithographique susceptibles d'avoir des conséquences sur la santé et qui nécessitent des prérequis spécifiques. « Il n'y a plus aucun danger et tout a été condensé en une quantité moindre de consommables à tous les niveaux », déclare M. Kelly.

Plus de dossiers d'impressions traités

Le UKHO a toujours contrôlé et testé les nouvelles technologies d'impression. Le passage à l'impression numérique a débuté il y a environ 10 ans avec l'installation de leurs premières imprimantes grand format portables à jet d'encre. Au fil du temps, ces imprimantes ont été remplacées par différentes autres technologies mais les coûts demeuraient 6 fois plus importants qu'une impression réalisée sur une presse lithographique. D'où la nécessité de maintenir deux flux de travail différents, y compris les presses lithographiques.



L'Introduction de la technologie HP PageWide en 2015 a marqué le début d'une nouvelle ère. « Bien que le volume d'impression reste le même – un processus de trois jours est désormais réalisé en un seul jour », résume M. Kelly. La Technologie HP PageWide comprend plus de 200 000 jets sur une barre d'impression immobile et couvre toute la largeur de la page, ce qui permet des vitesses d'impression révolutionnaires. La durée prolongée entre les cycles de nettoyage permet également une productivité soutenue. Suite à une procédure de tests rigoureuse, la première unité a été installée en août 2015.

Une qualité d'impression constante

En plus de la qualité d'impression constante et des avantages inhérents aux encres pigmentées (résistance à l'eau, à la lumière et à l'abrasion), un autre critère essentiel lié à la sélection de l'imprimante HP PageWide XL a été le traitement des médias. La capacité de charger jusqu'à 6 rouleaux de papier de différentes dimensions et le changement automatique de rouleau lorsqu'un rouleau est épuisé ou lors de la sélection d'une largeur de papier différente, ont contribué à une augmentation massive de la productivité. « Nous disposons de différentes largeurs de produits emballés. L'imprimante sélectionne le bac papier approprié sans aucun effet préjudiciable sur la vitesse de sortie et ceci a constitué l'élément clé du développement de la productivité », résume M. Kelly.

Pour plus d'informations :
www.ukho.gov.uk et
www.hp.com/go/LargeFormatPageWide

La conférence francophone ESRI 2017

Elle a déménagé ! Après plusieurs années passées à « pousser les murs » du Palais des congrès de Versailles, lui même choisi comme étant plus vaste que le Palais des congrès d'Issy-les-Moulineaux, la conférence francophone ESRI élit désormais domicile aux Docks de Paris, dans la commune de Saint-Denis. L'environnement ne sera certes pas le même, on peut donc espérer que les locaux seront plus spacieux.

Cette année les grands témoins seront Nicolas Vanier, écrivain, photographe et explorateur polaire, Yvan Bourgnon et Patrick Fabre, deux pionniers déterminés à nettoyer la mer de ces plastiques qui polluent et contaminent la faune marine, et Pierre Dufour, Directeur et Membre du Directoire d'*Inter Mutuelles Assistance*.

Bien entendu, *SIG* c'est aussi plus de deux cent cinquante communications utilisateurs, et l'endroit idéal pour rencontrer les partenaires d'ESRI France (le programme est disponible sur le site web *http://sig2017.esrifrance.fr*). Tout ce beau monde vous donne donc rendez-vous les 11 et 12 octobre prochains à Saint-Denis.

Les « sessions PostgreSQL » de Dalibo

Ces sessions sont une journée d'exposés/échanges autour de la base de données libres. Cette année, cette journée aura lieu le 17 novembre à Paris. *Dalibo* lance un appel à conférenciers. Chaque intervention doit durer 30 minutes (en comptant les éventuelles questions) et peut être donnée en français ou en anglais.

Les thématiques suivantes devraient être abordées (liste non exhaustive) :

- Nouveautés de PostgreSQL
- Retour d'expérience
- Migration vers PostgreSQL
- Optimisations
- Sauvegarde et Restauration
- Haute-Disponibilité
- Entrepôts de données / Big Data

Les interventions peuvent prendre plusieurs formes : témoignage utilisateur, *Proof of Concept*, tutoriel, comparatif, présentation de nouveautés, etc.

Toutes les propositions doivent nous parvenir avant le 30 septembre 2017 à l'adresse email : call-for-paper@postgresql-sessions.org.

Actus

Le FOSS4G Europe

Le *FOSS4G Europe* s'est tenu mi-juillet à l'ENSG. Que faut-il retenir de ces trois jours de conférence – échange entre acteurs de l'Open Source européen ?

Le *FOSS4G Europe*, c'est la conférence des développeurs impliqués dans l'*Open Source* géomatique en Europe. Cette conférence a lieu lorsque le *FOSS4G* mondial n'est pas organisé sur le Vieux Continent (cette année, il est prévu à Boston), ce qui permet aux développeurs qui n'ont pas les moyens de voyager loin de pouvoir quand même échanger avec leurs pairs.

L'événement débute traditionnellement avec une conférence plénière, qui a pour but de faire le point du développement de la géomatique libre, sous le patronage de l'*OSGeo*, un organisme créé en 2006 pour aider les projets et promouvoir l'utilisation des outils libres. S'il y a nul doute que la stratégie de l'*OSGeo* a connu un succès toujours grandissant (l'écosystème libre est en croissance constante), certains défis restent encore à résoudre pour assurer sa croissance dans les années qui viennent. On ne s'étonnera pas que le nerf de la guerre, ici aussi, soient les espèces sonnantes et trébuchantes. Les dirigeants de l'*OSGeo* souhaiteraient que le savoir-faire des équipes et des développeurs soit davantage mis en avant. Cela permettrait sans doute de trouver de nouveaux mécènes ou sources de financement. Autre piste : conclure des partenariats avec des structures internationales opérant dans le même secteur afin de générer des synergies. Il faudrait sans doute rassembler une communauté de développeurs encore plus vaste (au travers d'événements, de publications, d'autres conférences ?), mettre en avant des *success stories* ?

L'assistance lors de la conférence plénière d'ouverture.

Didier Richard, qui dirige le Valilab à l'IGN, était l'un des co-organisateurs de l'événement. Photo © Jody Garnett.

Quelques pistes de réflexion méritent d'être examinées : faut-il introduire des méthodes de type management d'entreprise au sein de l'*OSGeo* ? Rendre les conférences et événements plus accessibles, moins coûteux, pour y attirer davantage de monde ? Comment donner plus de moyens aux nouveaux projets pour qu'ils se développent plus vite ? Comment suivre l'évolution des projets une fois qu'ils ont atteint leur « maturité » ? Autant que questions auxquelles l'*OSGeo* devra répondre si elle veut continuer à jouer son rôle de catalyseur du monde foisonnant de l'open source géomatique.

Des exemples bien connus

Le *FOSS4G*, c'est aussi l'occasion de faire le point sur les dernières versions de logiciels que tous, ou presque, utilisent :

GDAL

GDAL, c'est LE couteau suisse du géomaticien libre. La bibliothèque est surtout utilisée pour convertir des données entre formats (vecteur et *raster*), mais, depuis quelques années, elle est également fournie avec un certain nombre d'utilitaires qui permettent de réaliser des opérations de reprojection, de découpe, de tuilage, de vectorisation, de *rasterisation*, d'ombrage, etc. La bibliothèque est développée par environ soixante développeurs permanents, et autant de développeurs occasionnels.

La dernière version, 2.2.1, apporte quelques améliorations par rapport à la version précédente (2.2). De nouveaux pilotes *raster*, par exemple DERIVED, un pilote qui permet de représenter les signaux radars sous forme de nombres complexes (amplitude + phase) ou encore des améliorations au pilote JPEG2000, pour encoder plus efficacement les nombres flottants. Du côté des pilotes vecteur, on note l'arrivée d'un lecteur DWG fondé sur la bibliothèque *Open Source LibOpenCAD*, permettant de lire des fichiers jusqu'au DWG R.2000. La prise en charge du DGN 8 se fait, quant à elle, au travers d'une bibliothèque tierce et payante. Enfin, *GDAL* intègre maintenant un pilote pour les schémas d'applications en GML *(GMLAS)*, qui préserve les relations et permet, par exemple, d'injecter le fichier dans une forme compréhensible par *PostGIS*. Le pilote fonctionnant dans les deux sens, il est aussi possible de transformer des tables en fichier GML.

GDAL 2.2.1 peut désormais gérer des types géométriques conformes à la norme SQL/MM part 3, comme les polyèdres, et effectuer des conversions entre tables *PostGIS* et des formats de type *Shapefile, GML* ou *DXF.* D'autres améliorations concernent l'ajout du traitement des *raster* virtuels dans l'API Python, des corrections concernant le traitement des jeux de données « creux » *(sparse data)*, l'ajout d'une constante NULL distincte de UNDEF, destinée à mieux gérer les tables SQL, un meilleur support de la norme GéoJSON, la lecture des géométries courbes à la norme FileGDB, et le développement d'un utilitaire supplémentaire, *ogrmerge*, permettant d'agréger plusieurs jeux de données vectorielles en une seule couche.

La prochaine version majeure, 2.3, attendue pour cet automne, apportera un lot supplémentaire d'améliorations et surtout plus de robustesse, avec la correction de quatre cents bugs ouverts. Au niveau des performances, l'utilitaire de découpe *gdal2tiles* devrait supporter le multitâche, un nouveau pilote *JPEG2K* utilisera du code GPGPU *(OpenCL)* pour accéder le décodage, *GDAL* prendra en charge des systèmes de fichier *Cloud*, la bibliothèque *Zstd*, trois fois plus rapide que *zlib*, remplacera cette dernière lors de l'écriture de fichiers TIFF compressés. Enfin, l'API sera étendue et deviendra 100 % compatible *Python*, ce qui signifie qu'il sera désormais possible de développer des pilotes *GDAL* ou *OGR* en *Python*. Tout ceci cependant, aura un léger coût, à savoir, pour ceux qui compilent *GDAL* à partir du code source, l'utilisation obligatoire d'un compilateur à la norme C++-11.

Grass

Grass, le SIG libre le plus ancien – mais également le plus complexe d'utilisation – continue d'évoluer. La version 7.2.1 a permis de fermer plus de cent cinquante bugs, et la version 7.4 est en vue, après une 7.2.2 intermédiaire qui viendra corriger d'autres problèmes.

Déjà, la version 7.2 a vu l'arrivée de nombreuses fonctionnalités : un nouveau catalogue de données, un éditeur *Python* simplifié, des nouvelles légendes pour les couches vectorielles, un traitement 3D (voxel) des écoulements d'eau *(r3flow, r3gradient)* et de nouveaux modèles hydrologiques (*ITZI,* capable par exemple d'évaluer le ruissellement à partir de données

radar de précipitations), une nouvelle algèbre temporelle pour le calcul sur des données 4D, des nouvelles fonctions de télédétection (*r.learn.ml* : classification et regression supervisées, *r.superpixels.slic* : segmentation SLIC) ou encore *v.decimate* pour réduire les nuages de points LiDaR.

Un projet *GSoC (Google Summer of Code)* actuellement en développement devrait aboutir à l'intégration de bibliothèques supplémentaires, comme *PDAL* pour la gestion des nuages de points.

La traditionnelle photo de groupe. © Gerald Fenoy.

Orfeo Toolbox

L'*Orfeo Toolbox* est une bibliothèque développée sous les auspices du CNES pour favoriser l'utilisation des images de la constellation *Pleiades*. Aujourd'hui à sa version 6.0, la bibliothèque s'est enrichie d'applications comme *Monteverdi* qui permettent la visualisation des images satellites en utilisant une accélération GPU, ainsi que de nombreux *plug-in* permettant l'utilisation de l'*OFT* dans des environnements comme *QGis*.

La version 5.6 de la bibliothèque avait déjà apporté des outils d'aide à l'extraction et à la sélection d'échantillons type pour les classifications supervisées, ainsi que le support géométrique pour les images *Sentinel-1*. Au niveau fonctionnel, l'exécution parallèle MPI *(message passing interface)* avait permis un gain appréciable en rapidité et initié le développement orienté cluster : vu la taille des images issues des satellites, il devient de plus en plus indispensable d'utiliser plusieurs ordinateurs pour traiter les données en des temps raisonnables. Avec la solution *MPI*, l'accélération constatée est quasi linéaire (utiliser 32 processeurs signifie diviser le temps de traitement par 30 environ), excepté sur les opérations d'entrées/sorties où les données doivent être sérialises. Le plus gros goulot à l'heure actuelle reste l'écriture de données, beaucoup plus lente que la lecture.

La version 5.8 a poursuivi dans la même direction (plus de parallélisme) et introduit le support pour les images *Spot 7*. La version 5.10 a ajouté un *framework* d'application composite, ainsi que des applications de type FFT/DWT.

La version 6.0 marque une évolution majeure. C'est tout d'abord le type de licence qui a été changé pour devenir de type *Apache*, ce qui correspond plus ou moins à une licence BSD 2 clauses (donc extrêmement permissive, non virale). Ceci a demandé l'accord des tiers qui avaient contribué au code. Au passage, un nettoyage systématique du code a été entrepris, et le nouveau code ajouté est à la norme C++-14, ce qui implique l'utilisation d'un compilateur adapté. Parmi les fonctionnalités nouvelles, on peut noter un module d'apprentissage automatique, la possibilité de traiter des images InSAR et les séries d'images temporelles (détection de changement). L'intégration avec *QGis*, qui reste l'un des principaux *front-ends* a été améliorée. La documentation a été améliorée, également, afin de rendre l'utilisation plus agréable.

Des produits moins bien connus ou nouveaux

D'un autre côté, le *FOSS4G* est toujours l'occasion de présenter des nouveaux logiciels ou des plateformes inédites. On pourra citer, par exemple, VTS, une plateforme intégrée pour le développement de cartes en 3D. VTS permet de *streamer* des représentations 3D grâce à des

composants côté serveur et côté client. L'intérêt de la solution est qu'elle est capable de diffuser de vastes quantités de données efficacement grâce à un système de tuiles 3D. Comme la solution est générique, elle peut même être appliqué en dehors de la Terre, par exemple pour représenter de futures colonies martiennes !

Certaines sociétés viennent aussi présenter des réalisations intéressantes. C'est le cas de *S2MAPS (http://s2maps.eu)* qui, grâce à l'utilisation d'images gratuites *Sentinel 2* à dix mètres de résolution a réalisé une image composite de l'Europe sans nuage. Le processus à consister à utiliser les masques produits par l'ESA et

Le *FOSS4G*, ce sont aussi des ateliers où les participants peuvent venir se former à de nouveaux logiciels ou découvrir des fonctions avancées de logiciels existants. © Milena Nowotarska.

Dans la même idée, mais sur des surfaces beaucoup plus réduites, *Mago3D* est une bibliothèque *Javascript* qui permet de représenter des modèles 3D BIM complexes, malgré le volume de données à transmettre. Pour cela, la bibliothèque emploie un certain nombre de pré-traitements pour accélérer les transferts, par exemple le pré-calcul des parties ombrées ou cachées. La bibliothèque peut être intégrée dans des environnements de *streaming* SIG, comme *Cesium* ou *Worldwind* par exemple. Dans les mois qui viennent, la bibliothèque sera capable de diffuser des nuages de points et de gérer les déplacements différemment selon que l'utilisateur se trouve à l'intérieur ou à l'extérieur des bâtiments modélisés.

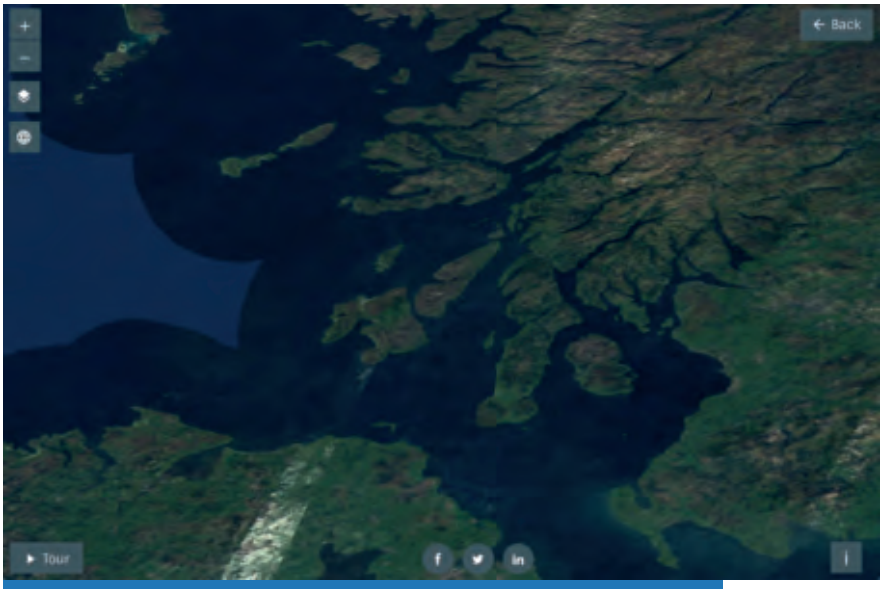
communiqués concomitamment avec l'image pour enlever les pixels correspondant aux nuages.

Plusieurs clichés ont été nécessaires sur chaque zone pour obtenir les valeurs moyennes de chaque pixel, qui ont été ensuite corrigés de l'atténuation atmosphérique puis corrigés géométriquement pour obtenir une image pan-européenne globale. Celle-ci est librement consultable sur le site de la société.

Le prochain *FOSS4G* Europe aura lieu l'année prochaine (le lieu n'est pas encore fixé), le *FOSS4G* mondial étant quant à lui prévu à Dar-es-Salaam, en Tanzanie. ◘

Retrouvez toutes les photos, le programme des conférences sur le site de la manifestation, https://europe.foss4g.org/2017/Home.

L'avancement des différents logiciels pilotés par l'OSGeo est disponible sur le site http://www.osgeo.org

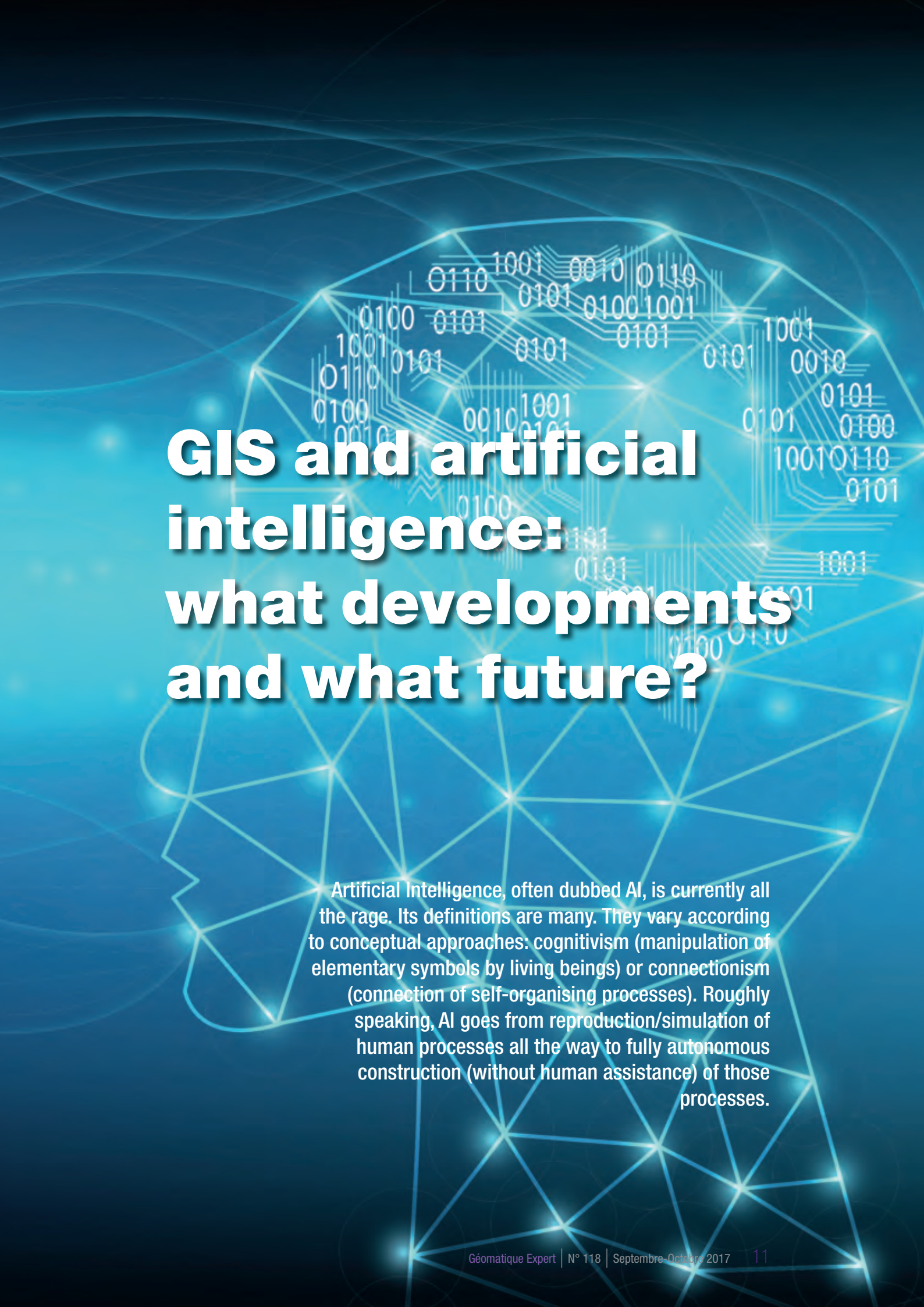


Malgré toute la bonne volonté de l'équipe de S2map, obtenir une image sans nuages de toute l'Irlande s'est révélé une tâche impossible à réaliser. Ici, l'Irlande est en bas, et l'Écosse en haut.

SIG et intelligence artificielle : quels développements et quel futur ?

Christian CAROLIN
Axes Conseil – carolin@axes.fr – http ://www.axes.fr

L'Intelligence Artificielle, souvent dénommée par ses initiales IA, constitue un sujet à la mode. Les définitions de l'IA ne manquent pas. Elles varient selon les approches conceptuelles : cognitivisme (manipulation des symboles élémentaires par le vivant) ou connexionnisme (connexion de processus qui s'auto-organisent). Globalement, l'IA va de la reproduction/simulation de processus humains jusqu'à la construction autonome (sans intervention humaine) de ces processus.



GIS and artificial intelligence: what developments and what future?

Artificial Intelligence, often dubbed AI, is currently all the rage. Its definitions are many. They vary according to conceptual approaches: cognitivism (manipulation of elementary symbols by living beings) or connectionism (connection of self-organising processes). Roughly speaking, AI goes from reproduction/simulation of human processes all the way to fully autonomous construction (without human assistance) of those processes.

IA, mythes et réalités

Contrairement à ce qui est souvent supposé, l'IA n'est pas une technologie nouvelle, ni même récente : la machine de Turing a été conceptualisée dans les années 40, les théories de logique floue furent élaborées en 1965 et Mycin, l'un des premiers systèmes experts, date de 1973. Après diverses tentatives de reproduction/imitation du fonctionnement du cerveau humain, l'IA s'est résolument tournée vers les mathématiques et la mise au point d'algorithmes sophistiqués. Les logiciels restent essentiellement basés sur les réseaux de neurones artificiels, la statistique probabiliste et sa déclinaison que constituent les réseaux bayésiens, ainsi que les systèmes de raisonnement à base de règles et de cas.

Pourquoi maintenant ?

En clair, ce qui constitue un réel progrès et permet de présenter les succès dont la presse nous abreuve régulièrement, réside davantage dans la puissance croissante des machines que dans la progression algorithmique fondamentale. Ce qui n'était concevable que d'un point de vue théorique il y a quelques décennies, est désormais rendu possible par la loi de Moore, même empirique et déclinante : des processeurs surpuissants, les super-calculateurs, les architectures distribuées, en attendant les processeurs neuromorphiques et les ordinateurs quantiques. Les progrès paraissent spectaculaires, parce qu'ils sont illustrés par des applications du quotidien : véhicules (plus ou moins) autonomes, chatbots sur les sites de vente en ligne, discussion avec son téléphone portable, via SIRI, Google Assistant, etc.

Quelques principes

L'IA sera de plus en plus présente dans la vie quotidienne, mais on constate déjà la disponibilité de nombreuses applications professionnelles. Les calculs de risque bancaire ou assuranciel, la prédiction dans le domaine marketing, par exemple, utilisent des fonctions d'intelligence artificielle depuis des années. S'il est difficile d'établir une approche ontologique de l'IA tant le champ de réflexion est vaste, il est possible de classifier quelques technologies de base :

- Les systèmes supervisés : il s'agit de fournir au système un modèle d'apprentissage constitué d'un ensemble de cas significatifs, sur lesquels le moteur s'appuiera pour apprendre puis pour déterminer des actions (phase de test). Ces systèmes sont utilisés pour obtenir un résultat sur la base de critères prédéterminés ;
- Les systèmes non supervisés : base du *deep learning*, ces algorithmes puisent dans de vastes sources de données pour constituer des classifications, sans base de référence obligatoire permettant l'apprentissage. Ils permettent par exemple la mise en évidence autonome de corrélations, sans déterminisme préalable (exemple : les cartes auto-adaptatives, ou *maps de Kohonen*).

Désormais, on utilise une sémantique évoquant les notions d'IA forte et IA faible, correspondant globalement, dans le premier cas, à la capacité de génération d'un raisonnement autonome (dit cognitif), là où l'IA faible (la plus développée à ce jour) raisonne sur des algorithmes déjà connus, maîtrisés par l'humain et ne progresse donc que dans les limites de la connaissance humaine (approche dite pragmatiste).

IA et SIG, « GeoIA »

Pour faire simple, on peut considérer que toute donnée géolocalisée, liée à une notion de déplacement et son corollaire, le temps, peut offrir une thématique géographique utilisable par l'IA. L'information statique est bien entendu également exploitable.

Fonctions IA intégrées ou non au SIG

Le SIG ne vit plus tout seul : il est interfacé. Il n'est plus de projet applicatif significatif sans une composante géographique et les données géolocalisées contribuent significativement à la masse d'informations stockées dans le contexte *big data*. Les sources de données peuvent être :

- Matérielles : caméra, radar, LiDAR… ;
- Les bases de données, les données non structurées ;
- Et de plus en plus, les données de source IoT (Internet des objets) et les réseaux sociaux.

À ce jour, les principaux SIG du marché n'intègrent pas de fonctions d'IA. D'une façon générale, on constate davantage la construction d'interfaces que du développement spécifique.

AI, myth and reality

Contrary to a widespread assumption, AI is neither a cutting-edge technology, nor even a recent one: the Turing Machine was formalised in the '40s, the first fuzzy logic theories appeared in 1965 and *Mycin*, one of the first expert systems, dates back to 1973. After trying to mimic or replicate the function of the human brain, AI development turned to mathematics and highly sophisticated algorithms. Software development still relies on neural networks, statistics and probability theory, together with their offspring known as "bayesian networks", and systems of reasoning based on rules and cases.

Why now?

To be honest, the real progress which leads to the success the press harps on is more likely due to the steady increase in computing power rather than any breakthrough in fundamental research. What was only dreamt of decades ago is now commonplace. Moore's Law looks like it will break down soon, but, in the meantime, it has given us ultra-powerful CPUs, supercomputers, distributed architectures, with neuromorphic CPUs and quantum processing in the offing. Progress appears outstanding, because it impacts everyday applications: (more or less) autonomous vehicles, chatbots used in various interactive websites, chat applications on mobile phones, *Apple Siri*, *Google Assistant*, and so on.

A quick AI primer

AI has yet to invade our daily life, but it has already given birth to numerous professional softwares. The fields of banking, insurance risk assessment and marketing prediction have been using AI engines for years. While an ontological approach to the vast field of AI can be tricky, some basic technologies can be listed:

- Supervised systems: the system is fed with a training model made up of a set of relevant cases, which the engine then uses to learn and later make decisions (test phase). These systems are used to get a result out of predetermined cases;
- Non-supervised systems: these algorithms are the essence of *deep learning*. They draw from huge data sources to create classifications, with no prior references given during a learning phase. They allow for autonomous discovery of correlations, without preexistent determinism (e.g. auto-adaptative maps, also known as *Kohonen's maps*).

Recently, a new taxonomy using the names "strong AI" or "weak AI" has emerged. The former concept corresponds to the ability to create an autonomous reasoning (dubbed "cognitive"). The latter operates only on known algorithms, designed by humans, and is thus bound to the limits of human knowledge (this approach is dubbed "pragmatistic"). It is currently the most widespread type of AI.

AI and GIS, "GeoAI"

To keep it simple, we can assume that every georeferenced or timestamped data can offer geographical information to an AI program. Statistical figures are obviously also processable.

AI function integrated in the GIS… or not!

No more is GIS a castaway: it communicates. There is currently no large data processing framework which does not include a tool to crunch geodata and where geodata do not represent a significant amount of stored information in a *big data* context. Data sources can be:

- Hardware: cameras, radars, lidars…;
- Databases, unstructured data;
- And more and more, IoT sources and social networks.

To date, the mainstream GIS softwares do not integrate any AI modules. Generally speaking, the trend is to offer interfacing rather than bespoken development. Several projects conflate GIS with AI engines. Their goals are:

- Improvement in analysis result quality, easing up decision making;
- Clearer interpretation of results, in a forecast context;
- Risk mitigation ahead of decision making.

A few outliers exist. The AI platform developed by the American company *Alteryx* offers functions that exploit geodata (geocoding, model building, mapping) and can also read or save geodata using most of the well-known formats (SHP, MIF/MID, MAP, KML…). The most common databases can be used as data sources.

Il existe de nombreux projets mixant les fonctionnalités natives SIG à des moteurs IA. Les objectifs du couple SIG/IA sont généralement :

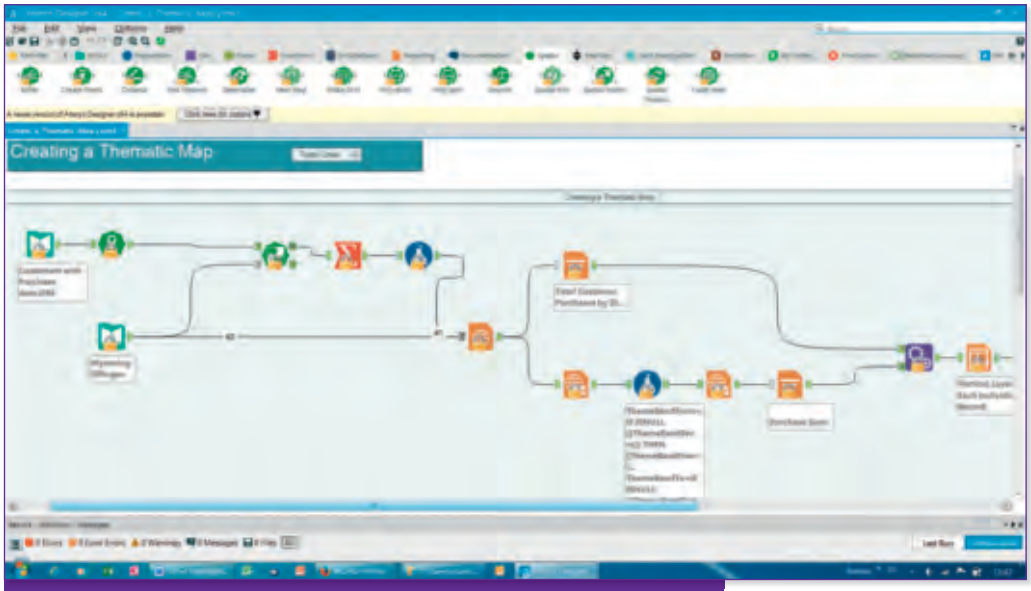
• L'amélioration de la qualité des résultats d'analyse, permettant une meilleure aide à la décision ;
• La progression des interprétations, dans un contexte prédictif ;
• La limitation des risques en amont des prises de décision.

Il existe quelques exceptions, telles que la plateforme IA de l'éditeur américain *Alteryx*, qui intègre des fonctions d'exploitation de l'information géographique (géocodage, modélisation, restitution) et peut également produire des données SIG dans les formats principaux du marché (SHP, MIF/MID, MAP, KML…). Les principaux SGBDR peuvent être utilisés en source de données.

Applications thématiques

Les applications sont constituées d'association des fonctions natives des SIG avec des algorithmes IA combinés : prédictifs, classification, etc. Quelques exemples de cartes interprétatives, souvent générées à l'aide de l'inférence bayesienne :

• Simulation de l'évolution de l'occupation des sols ;
• Estimation de la variation de la végétation ;
• Simulation des catastrophes naturelles résultant des changements climatiques, pour l'aide à la gestion de crise ;
• Simulation d'accidents industriels en 3D, dans un cadre de prévention des risques ;
• Prédiction des mouvements de terrain : à titre d'exemple, l'université canadienne de Sherbrooke a réalisé une étude portant sur l'utilisation de la télédétection, des SIG et de l'intelligence artificielle pour déterminer le niveau de susceptibilité aux mouvements de terrain, avec une application dans les Andes de la Bolivie (http://savoirs.usherbrooke.ca/handle/11143/2709) ;
• Il existe de nombreux projets de recherche dans le domaine de la prédiction de la déforestation, de la sécheresse ou *a contrario*, des inondations ;
• Utilisation de l'IA en géologie et pour la prospection minière et pétrolière, afin de déterminer des champs de prospection future ;
• La société canadienne *Globvision* propose un progiciel d'estimation de l'humidité et de la rugosité du sol, permettant des applications hydrologiques et agricoles. Sont cités : le suivi et prévision de la sécheresse, l'optimisation du rendement des cultures et la modélisation climatique. Le logiciel intègre une interprétation de données de télédétection pour l'estimation des paramètres de la surface du sol, à partir d'images satellite ;
• Quand la donnée n'existe pas : un éditeur SIG a réalisé une application permettant de prédire les conditions de trafic dans la ville de Koweit City, sachant qu'une surface significative de la ville ne dispose pas de données géographiques exploitables. L'IA permet, via des algorithmes appropriés, la prédiction et l'abstraction sur des zones cibles (même non intégralement couvertes par des données géographiques). Les solutions algorithmiques déployées permettent de déterminer les directions dans lesquelles la congestion de trafic s'étendra et quelles sont les mesures préventives à envisager.



Modélisation d'une carte thématique avec Alteryx.
Thematic map modelling with Alteryx.

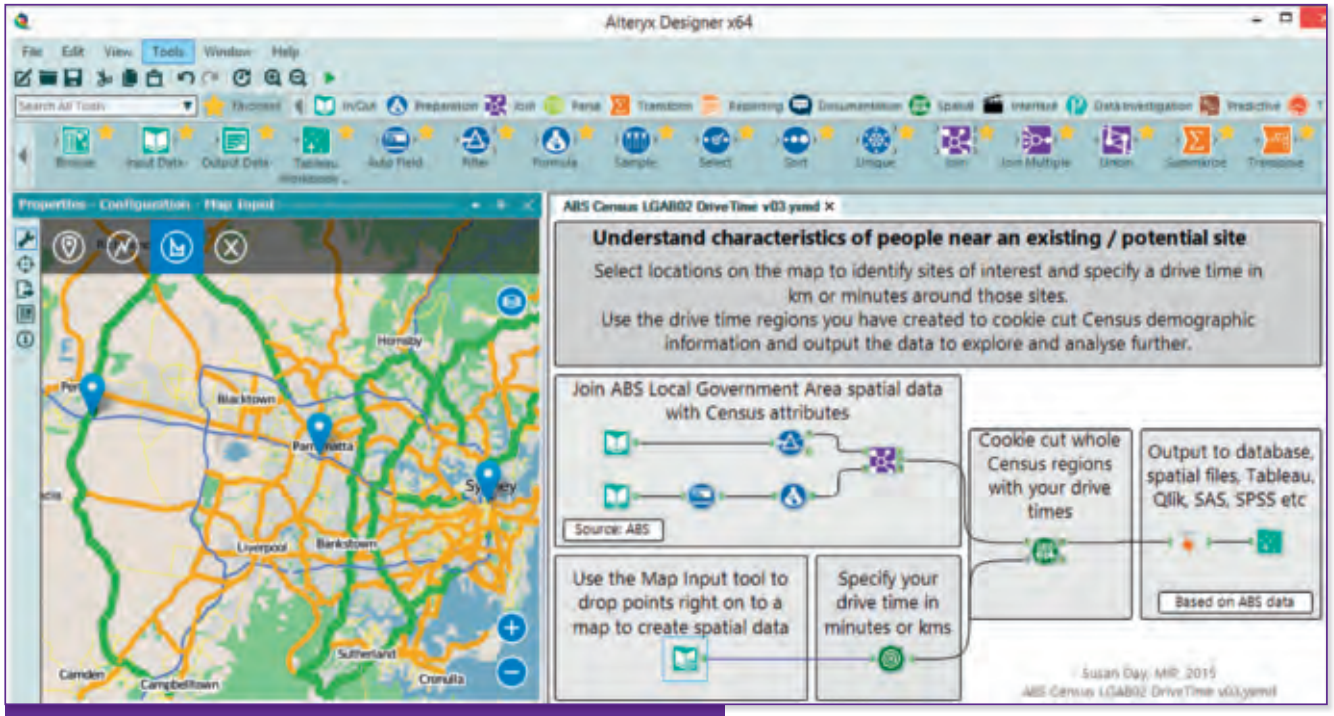
Thematic applications

These applications combine native GIS functions with AI algorithms: prediction, classification… Here are some examples of interpretative maps. Most of them are the result of bayesian inference:

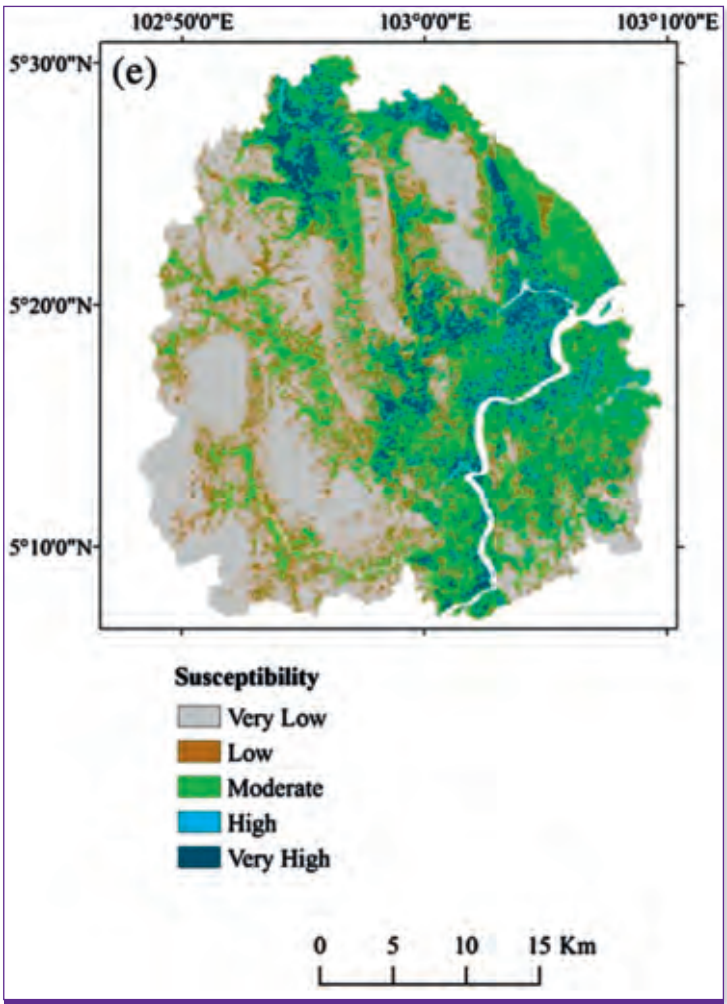
- Land use / land use evolution simulations;
- Vegetation variation assessment or estimation;
- Simulation of natural disasters caused by climate change, to aid in crisis management;
- 3D simulation of industrial disasters, in a risk mitigation context;
- Prediction of landslides: for instance, the Sherbrooke university has conducted a study over the use of remote sensing, GIS and AI to assess the landslide risk, with an application in the Bolivian Andes (http://savoirs.usherbrooke.ca/handle/11143/2709);
- Several research projects target deforestation, drought or flooding prediction;
- AI is used in geology and in the mining industry to help locate future prospection fields;
- The Canadian company *Globvision* sells a software which estimates soil wetness and ruggedness, for hydrological or agricultural applications. Are cited: management and forecast of droughts, crop yield optimisation, and climatic modelling. The software includes interpretation of remote sensing data and satellite imagery to estimate soil surface parameters;
- When data is unavailable or non-existent: a GIS software company has created an application able to predict traffic conditions in Koweit City, with the caveat that most of the city is uncharted. Using relevant algorithms, the AI engine allows prediction and abstraction over target areas (even in the absence of thorough geospatial data coverage). The software predicts how gridlocks expand and tells how to mitigate them.

Strengths and weaknesses: a case study

In the *en vogue* field of urban security, the city of Zurich has been recently using a German software named *Precobs (PRE Crime OBservation System)*. The analysis criteria of this weak AI application (since the decisions to monitor or intervene are made by an officer) include: the type of offense, place, date, motive, and means used. The algorithms



Carte générée.
Generated map.



Carte « intelligente » de simulation de risques d'éboulement.
"Smart" map showing landslide risk simulation.

Points forts points faibles : étude de cas

Dans le domaine en vogue de la sécurité urbaine, la ville de Zurich utilise depuis peu un logiciel allemand nommé *Precobs (Pre Crime Observation System)*. Les critères d'analyse de cette application IA faible (les décisions d'intervention ou de surveillance sont prises par un policier), sont notamment : le type d'infraction, lieu, date, but du délit, moyens utilisés. Les algorithmes utilisés sont de type statistique (corrélation et régression) et intègrent également des caractéristiques topologiques et de psychologie criminelle. La probabilité des délits est évaluée sur un intervalle de temps allant jusqu'à sept jours.

La ville annonce une baisse de la criminalité de 30 %, cependant les résultats et l'efficacité du progiciel semblent être remis en question selon les interlocuteurs et usages :

- Insuffisance des sources de données (notamment, pas de données IoT ou issues des réseaux sociaux) ;
- Lien causal entre la diminution du nombre de crimes et utilisation du progiciel non prouvé ;
- En phase de test du progiciel, la ville allemande de Nuremberg a vu sa criminalité diminuer puis augmenter selon la période de l'année, sur des bases algorithmiques identiques ;
- Limitation du fonctionnement du progiciel aux actes criminels professionnels.

Les limites de la « Geo IA »

Au-delà de l'exemple précédent d'IA faible, une future *Géo IA* forte pose naturellement le problème du comportement humain. La simulation des déplacements dans un espace géographique implique en effet que l'algorithme reproduise le raisonnement humain. Or, celui-ci est souvent arbitraire, chaque personne ayant sa propre perception de son mode de déplacement, de sa vitesse, de son environnement géographique. Les notions de vitesse ou de proximité sont certes des principes assimilables par la logique floue, liés à l'environnement de chacun. Mais l'apport de sciences et approches additionnelles, telles que les neurosciences, semblent indispensables pour élaborer des raisonnements exploitables.

Il convient également d'intégrer le fait que la valeur de l'IA réside très largement dans la qualité de la donnée source, au-delà de la performance algorithmique, que l'IA soit forte ou faible. Par exemple, le poids synaptique des liens d'un réseau neuronal formel dépend totalement des

used are primarily statistical (correlation and regression), but they also take into account topological details as well as elements of criminal psychology. Forecast data is available up to seven days in advance.

The city boasts a 30% decrease in criminal activity, yet the software's results and efficiency vary according to its users or usages:

- Scarcity or insufficiency of data sources (especially no data from IoT or social networks);
- The statistical correlation between the decrease in crime activity and use of the software is moot;
- During the test phase, the city of Nurnberg saw its crime activity fluctuate along the year, despite the algorithms being unchanged;
- The software is limited to "professional crime" activity.

GeoAI limits

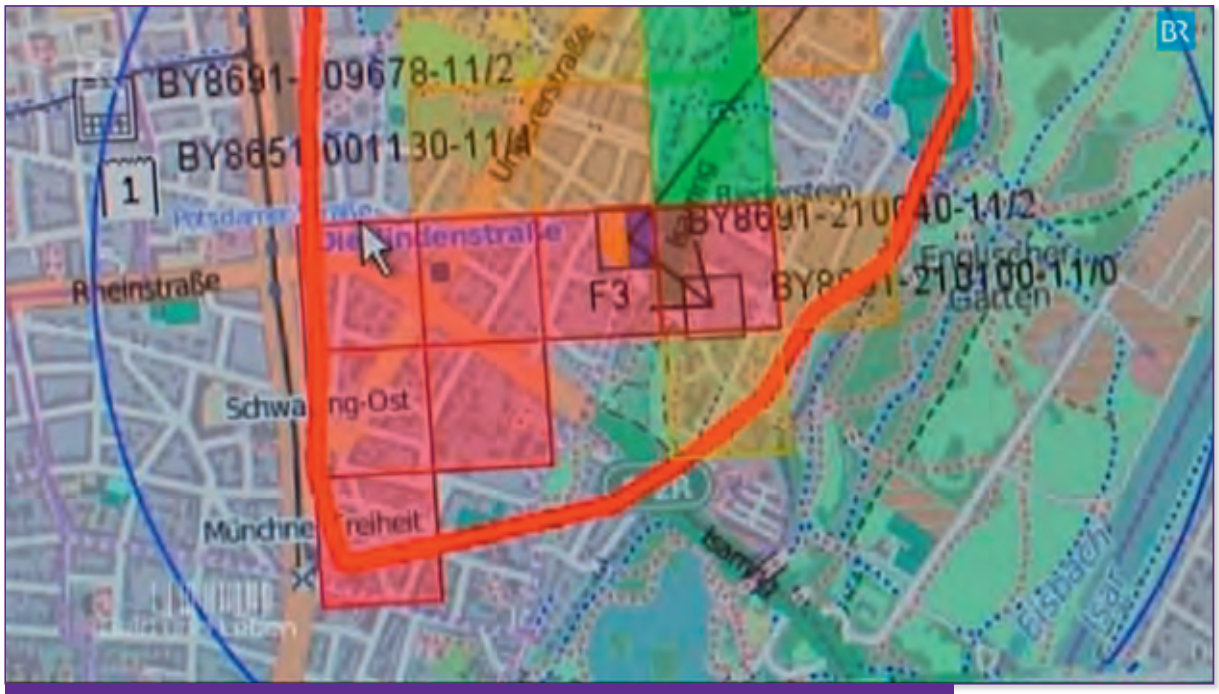
Beyond the former example of weak AI, a future strong AI poses the problem of emulating human behaviour. Simulation of movements in geographical space entails reliable reproduction of human thinking. However, that thinking is often arbitrary, each person having their own perception of their trajectory, speed or surroundings. While it is true that speed or proximity can theoretically be modelled in fuzzy logic, further scientific support or more complex approaches, such as neurosciences, seem required to mimic reasoning.

One should also consider that data quality is paramount to weak or strong AI's performance, irrespective of algorithm efficiency. For example, the synaptic weights of the connexions in a formal neural network depend on, and only on, the data used and its quality. All these considerations usher us into a vast field of investigation, mostly unexplored as of today.

Into the future: is strong AI inevitable despite the risks?

The pair GIS/AI is still in its early stage. As aforementioned, many applications coupling GIS with weak AI are available. Strong AI is still years down the line.

It is difficult to conclude without mentioning or pondering on the risks associated with autonomous AI. This goes beyond technology and flirts with ethics and law. A deep political reflection should be ordered, and measures taken so that in all circumstances a human could stop



Sur la carte, les zones représentant un risque fort de délit (ici, les cambriolages) sont en rouge (grille de 250 mètres).
On this map, areas with the highest risk of burglary are coloured in red (250 m grid).

données utilisées et de leur qualité intrinsèque. Tous ces aspects mettent en évidence un champ d'investigation vaste, à ce jour embryonnaire.

Le futur : l'IA forte, inéluctable mais risquée ?

Le couple SIG/IA est balbutiant. Comme présenté précédemment, de nombreuses applications couplant SIG et IA faible sont déjà disponibles mais amenées à progresser et se développer. La « Géo IA forte » reste à venir.

Il est difficile de conclure ce propos sans évoquer et s'interroger sur les risques potentiels inhérents à l'IA autonome. Ceci dépasse la technologie et est du ressort de l'éthique et de la loi. Une réflexion politique approfondie devrait être menée pour qu'en toutes circonstances, l'humain puisse interrompre unilatéralement le fonctionnement autonome d'une IA. Restreint à notre champ de réflexion géomatique, que penser d'un algorithme dédié au militaire qui déterminerait la trajectoire d'un missile, d'un avion, d'un drone, sans possibilité d'intervention d'un ingénieur ? Que penser d'une voiture autonome qui ne permettrait pas à l'homme de reprendre le contrôle du véhicule si le logiciel décidait d'une direction en contradiction avec la volonté du passager ? Le danger est d'ores et déjà largement annoncé par Elon Musk ou Stephen Hawking, personnalités parfois fantasques mais rarement fantaisistes. (https://futureoflife.org/ai-open-letter).

Il est étonnant de constater que la mesure de ce danger, même potentiel, ne semble faire l'objet d'aucune identification réelle, d'aucun groupe de travail dédié, dans le rapport national (https://www.economie.gouv.fr/files/files/PDF/2017/Rapport_synthese_France_IA_.pdf), qui a suivi la Journée de l'IA en mars 2017, conclue par le président de la République. La notion de risque y est évoquée mais majoritairement dédiée à la dimension économique. En France, pays du principe de précaution, l'autorité politique aurait pourtant une occasion, pas si fréquente, de réfléchir en amont et peut-être de légiférer de façon proactive et non en réaction à des événements passés, souvent dramatiques. Il faudrait penser un texte de loi simple, s'inspirant partiellement, pourquoi pas, des lois d'Asimov et imposer qu'il n'existe aucun cas dans lequel l'humain ne peut reprendre le contrôle d'une intelligence artificielle, qu'une application IA forte ne peut jamais être diffusée, commercialisée et exploitée sans respecter cette règle.

Cependant, des experts certifient que le danger est faible, voire inexistant. Alors, que penser des performances et de la progression de *DeepMind*, du programme *Watson* ou du programme *Google Brain*, IA qui a créé son propre langage, incompréhensible par les humains ? La théorie de la Singularité est-elle totalement stupide ? Ray Kurzweil est-il fou ? Souhaitons donc ces experts soient plus clairvoyants que ceux qui prédisaient qu'*Internet* ne remplacerait jamais le *Minitel*… ◘

Le club utilisateur Business Geographic

Tous les utilisateurs de (et personnes intéressées par) la suite ***GEO*** éditée par le français ***Business Geographic*** sont invités à se retrouver le 14 et 15 décembre (matin) prochain à Villeurbanne, dans les locaux du groupe, pour le club utilisateur 2017. Le programme est d'ores et déjà en ligne sur le site ***web*** de l'éditeur ***(https://www.business-geografic.com).*** Les participants qui souhaitent dormir sur place peuvent bénéficier de tarifs préférentiels dans un hôtel IBIS proche, donc dépêchez-vous de vous inscrire ! ◘

the function of any AI. What of a military algorithm able to compute the trajectory of a drone, a plane or a missile without the supervision of an engineer? What of an autonomous vehicle where the passenger cannot override the steering mechanism, even if the onboard computer is driving it astray? That danger is already trumpeted by Elon Musk or Stephen Hawking, two eccentric but not flighty people (https://futureoflife.org/ai-open-letter).

Amazingly enough, this danger, however remote, is not mentioned in the recent national report (https://www.economie.gouv.fr/files/files/PDF/2017/Rapport_synthese_France_IA_.pdf, in French) of the "AI's day", an event held in March 2017 and capped off with an address from the *Président de la République.* Risk is touched upon, but the focus is clearly on economical threats. In France, the country that invented the precautionary principle, authorities would have a unique opportunity to ponder ahead of time and not act in response to a past, and maybe dramatic, incident. A bill that could refer to Asimov's three laws of robotics should be passed, enforcing mandatory human control over any type of AI program. No strong AI application should ever be written, sold, deployed or operated that would violate this law.

However, some experts claim that this danger is insignificant, even possibly fanciful. If this is the case, what should we think of the progress and performance of *DeepMind, Watson*, or of *Google Brain*, an AI software which created its own language, obscure to any human? Is Ray Kurzweil a kook? Let's hope that these people are more insightful than those who predicted that the *Internet* would never overthrow the French *Minitel…* ◘

La carte géante de la biodiversité

À l'occasion du festival international de géographie de Saint-Dié (FIG), cet année dédié à la thématique « Territoires des hommes, mondes animaux », l'IGN va déployer une carte de quarante mètres carrés au sol de la gare de Saint-Dié permettant d'apprécier en détail les richesses de la biodiversité et des espaces protégés en France métropolitaine et en outre-mer.

Conçue conjointement par l'Agence française pour la biodiversité, le Muséum national d'histoire naturelle et l'IGN, cette troisième édition de la carte de la biodiversité a été réalisée sur les fonds cartographiques de l'IGN à partir des données de l'Inventaire national du patrimoine naturel (INPN). Elle représente les réserves naturelles et biologiques, les parcs nationaux, régionaux et marins, les ZNIEFF (Zones naturelles d'intérêt écologique faunistique et floristique), les sites Ramsar (relatifs aux zones humides d'importance internationale), dix-sept fiches d'espèces illustrées, animales ou végétales représentatives d'un milieu, d'une région ou d'un statut de protection particulier, ainsi qu'une cartographie au niveau européen du réseau Natura 2000. ◘

Vérifier la cohérence des données, un enjeu primordial

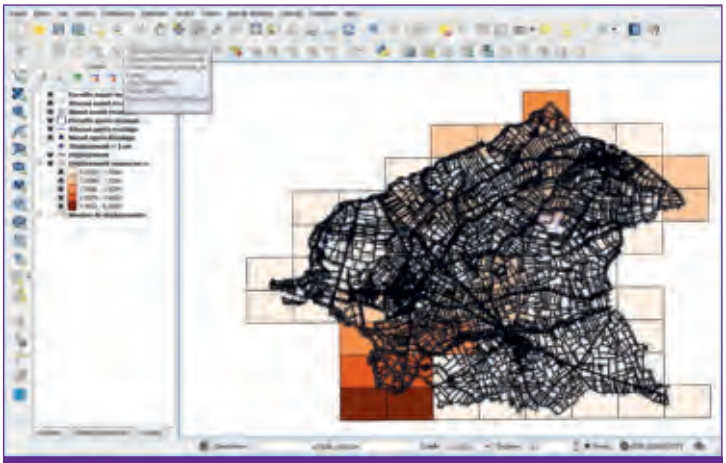
Lorsqu'il faut intégrer des données, ou bien les recaler sur des référentiels améliorés, de nombreux défauts peuvent apparaître. Certains outils automatiques, comme *1Integrate*, édité par *1Spatial*, peuvent aider les géomaticiens à automatiser les tâches fastidieuses de vérification/ajustement.

Avec la prochaine arrivée du PCRS, nombreuses sont les municipalités, EPCI ou autres structures publiques et/ou privées qui vont devoir reprendre l'intégralité de leurs documentations géographiques. Traditionnellement, les plans étaient référencés sur les limites cadastrales, avec les imprécisions et les défauts que chacun connaît. Le PCRS devrait mettre un terme à cette situation, et permettre à tous d'avoir un référentiel à la fois simple à manipuler, et précis. Le passage entre les deux ne va pas être instantané : si le PCRS corrige les erreurs du cadastre, alors tous les réseaux qui se trouvent référencés par rapport au cadastre vont devoir être corrigés également.

Or, les déformations du cadastre sont complexes et variées. Un recalage sur le PCRS ne peut se résumer à un simple déplacement en bloc, et ce d'autant que le décret sur les classes de précision oblige les réseaux à être localisés aussi précisément que possible. Souvent, le positionnement des réseaux résulte de considérations topologiques comme « au milieu de la rue », « à 15 cm du mur », etc., et donc impliquent de repositionner chaque tronçon pratiquement parcelle par parcelle.

Recaler automatiquement

L'une de ces solutions est éditée par la société anglaise *1Spatial*, sous le nom *1Integrate*. De quoi s'agit-il exactement ? « *Essentiellement d'un moteur de règles,* indique Seb Lessware, chef produit Royaume-Uni chez *1Spatial. Le moteur peut travailler sur n'importe quelle donnée, géographique ou non.* » *1Integrate* est couvert par un brevet aux États-Unis pour ce qui est de son algorithme de traitement des données, en particulier l'application automatique de règles de validation/transformation.



1Integrate est capable de calculer les vecteurs de déplacement de chaque tronçon lors d'opérations de recalage. Ces vecteurs peuvent ensuite être affichés dans un SIG, ici QGis, et faire l'objet de traitements statistiques.

Pour certaines petites communes, la pénitence peut paraître relativement bénigne. En revanche, les grandes communes et, *a fortiori*, les gestionnaires de réseaux département ou nationaux, se trouvent devant une tâche pharaonique. Si tout cela devrait être fait manuellement, la transition pourrait durer plusieurs années. Mais heureusement, il existe des solutions semi-automatiques qui permettent de soulager le travail des opérateurs et d'accélérer les temps de traitement.

1Integrate intègre en entrée les couches de données (référentiels ou données métiers sur lesquelles il doit travailler), qu'ils

L'éditeur de règles de *1Integrate*.

soient sous forme de fichiers de données, ou de tables dans des bases comme *Oracle* ou *PostGIS*. Le deuxième « élément » à fournir est un ensemble de règles, qui peuvent être spécifiées soit en alimentant une API au travers d'appels documentés, soit directement saisies dans un assistant. Ces règles peuvent être de nature géométrique (un segment doit être droit), topologique (un segment doit se terminer sur un autre segment, un polygone ne peut en chevaucher un autre, deux tronçons non raccordés ne peuvent se croiser, un segment doit être à telle distance d'un polygone, etc.) ou bien encore attributaire (tel objet doit avoir tel attribut renseigné, différent de zéro, correspondant à telle mesure, etc.).

Dans une opération de recalage d'une conduite d'eau potable, par exemple, chaque tronçon est repéré par sa distance au mur le plus proche. Si ce mur se décale suite à l'utilisation d'un référentiel plus précis que le cadastre, le tronçon va suivre. Parfois, les règles ne peuvent pas aboutir : par exemple, un tronçon indéformable dont la distance à un mur courbé doit être constante. Dans ce cas, *1Integrate* se contente de relever l'incohérence et propose à l'utilisateur d'agir manuellement. L'utilitaire fonctionne en plusieurs étapes : premièrement, calculer les « vecteurs de déplacements », c'est-à-dire les translations à appliquer à chaque objet ; deuxièmement, appliquer ces déplacements, tout en maintenant les éventuelles contraintes topologiques ; enfin, valider les résultats, en indiquant les cas problématiques où les règles n'ont pu aboutir.

Conflation de données

Les règles peuvent être utilisées dans d'autres opérations. Par exemple, des données spaghettis récupérées d'un plan type CAO peuvent être passées au crible pour repérer des polygones non fermés, ou des polygones auto-sécants, par exemple. Des emprises de polygones bâtis qui empiètent légèrement sur les parcelles voisines peuvent être détectées et alignées automatiquement sur la frontière parcellaire, etc. « *L'avantage principal de* 1Integrate *par rapport aux autres logiciels de traitement de données (qui reposent sur des requêtes, des processus manuels, des scripts, de la programmation ou des workbenches visuels) réside dans l'approche par règles. Les règles sont explicites, faciles à gérer via une interface utilisateur et évolutives. En outre, elles sont plus simples à organiser quand les traitements deviennent complexes et sont plus performantes sur les grands jeux de données.* »

Exemple de *reporting* graphique. Chaque point de non-conformité est matérialisé par une puce rouge.

Prenons le cas de deux référentiels contenant des informations différentes à des échelles différentes (on peut penser, par exemple, à la *BD Topo* d'un côté et à un plan de récolement de l'autre). Enrichir les données de la première couche avec les données de l'autre (ce que *1Spatial* appelle « *conflation* ») suppose de pouvoir apparier les objets des deux couches, mais la différence d'échelle signifie que les formes ne coïncident jamais parfaitement, donc une jointure spatiale ne rend pas de résultat. *1Integrate* intègre un moteur de logique floue qui permet au logiciel de reconnaître, et d'apparier, les objets homologues dans les deux couches, rendant ainsi la jointure possible. « *Recaler des données sur un territoire est une opération complexe, qui parfois nécessite l'intervention de géomètres pendant plusieurs mois.* 1Integrate *combine des algorithmes avancés et une architecture multiprocesseur pour traiter les problèmes complexes le plus rapidement possible, ainsi que des astuces pour pouvoir traiter des millions d'objets même avec peu de mémoire.* »

Mais *1Integrate* ne se contente pas de traiter de la donnée existante. Le logiciel est capable de détecter des absences, par exemple un tronçon manquant dans un réseau d'eau. Mieux, il peut inférer la donnée manquante à partir des règles. Par exemple, toujours dans le cas d'un réseau d'eau, où chaque abonné doit être relié au réseau par un raccord partant de sa résidence et aboutissant à la canalisation principale passant sous la rue, les raccords manquants peuvent être détectés et créés automatiquement. Bien entendu, cette donnée est « abstraite » et identifiée comme telle, si bien que les utilisateurs peuvent obtenir une liste des ces objets et programmer des visites terrain pour corriger leur position. « *Naturellement*, poursuit Seb Lessware, *plus les règles sont précises, plus les données créées seront "vraisemblables".* »

Actuellement, *1Integrate* fonctionne soit comme produit séparé (on envoie un jeu de données au travers d'une interface *web*, et l'utilitaire produit un jeu de données corrigées), soit comme *plug-in* intégrable dans l'environnement ESRI (directement sous *ArcGIS* ou bien dans *Collector4ArcGIS* afin de guider les saisies à la volée). Il est cependant prévu, à moyen terme, de développer un *plug-in* pour *QGis* et *Elyx*, fonctionnellement équivalant à celui d'ESRI. Diverses options sont également envisagées pour installer le produit sur des serveurs de type *cloud*, privé ou semi-privé, voire un mode SaaS où *1Spatial* pourrait facturer l'utilisation au volume.

Futurs développements

En attendant, plusieurs améliorations devraient être apportées au produit dans les mois qui viennent. La première consiste à intégrer le traitement des abscisses linéaires, une fonctionnalité indispensable aux gestionnaires de réseaux de transports linéaires, où tous les équipements sont repérés en PK. Les règles considéreront l'abscisse linéaire comme une dimension supplémentaire, et devraient permettre de vérifier la précision des PK par rapport à la géométrie des voies, par exemple.

Une deuxième amélioration consistera à gérer l'intégration de plans CAO de type BIM à l'intérieur de SIG. Ces opérations de fusion SIG/BIM devraient devenir de plus en plus fréquentes. Or, SIG et BIM correspondent à des échelles différentes (le BIM étant plus précis que le SIG), ce qui risque de conduire à des conflits. L'application de règles permettra de contraindre/rectifier les données SIG avant de procéder à l'intégration des données BIM.

Enfin, porté par l'OS Irlande, *1Spatial* planche actuellement sur une version 3D de son logiciel. « *Nous sommes en train de développer des bibliothèques pour étendre à la 3D ce que nous faisons en 2D. L'une des demandes les plus pressantes à l'heure actuelle est de pouvoir recaler des données 3D relevées par Lidar, par exemple des bâtiments, sur des emprises 2D comme le cadastre. Les futures versions de* 1Integrate *devraient permettre ce genre d'opération, ainsi que le calcul de quantités topologiques en 3D comme l'éloignement, la contiguïté, etc.* »

L'équipe de développement anglaise a donc de quoi s'occuper pendant les mois qui viennent ! ◻

Sur cet exemple, tous les branchements en rose/rouge (deux réseaux différents) ont été créés automatiquement par *1Integrate* à partir des règles métiers communiqués. Les créations nécessitent évidemment d'être revues, mais une bonne partie du travail est déjà effectuée.

ABONNEZ-VOUS !

Tarifs TTC	1 an	1 numéro
FRANCE	70 €	14 €
Education (TVA 2,1 % incluse)	55 €	
Union Européenne	85 €	16 €
Etranger	95 €	18 €

Le premier bimestriel français consacré entièrement aux technologies et aux applications des systèmes d'information géographique

OUI, je souhaite m'abonner 1 an à **Géomatique Expert** au prix de ..... €

Mon abonnement comprend : 6 numéros + 10 eNewletters

Je règle mon abonnement par : ☐ chèque ci-joint à l'ordre de CIMAX ☐ à réception de facture ☐ virement ou mandat administratif

Prénom ..... Nom ..... Fonction .....

Société .....

N° TVA intracommunautaire .....

Adresse .....

Code postal ..... Ville .....

Tél. ..... eMail (pour recevoir les newsletter) .....

L'abonnement ne sera pris en compte qu'accompagné de son règlement. ☐ Je souhaite recevoir une facture acquittée.

À renvoyer sous enveloppe affranchie au tarif en vigueur à : **Géomatique Expert** Service Abonnements

12, place G. Pompidou • 93167 Noisy-le-Grand cedex • Tél. : 01 45 92 98 98 - Fax : 01 49 32 10 74

Le Calvados distribue ses données

Avec l'aide de la PME Isogeo, le département normand va ouvrir un portail de diffusion de l'information géographique départementale à destination de ses communes/EPCI.

Le Calvados, département au cœur de la Normandie, peut se targuer d'avoir été précurseur dans le domaine géomatique. Il fut ainsi le premier à piloter la réalisation d'une orthophotographie en 2001, orthophotographie renouvelée en 2006. Par la suite, le département a supervisé la numérisation du cadastre, ce qui n'était pas une mince affaire dans un département essentiellement composé de petites communes rurales (plus de la moitié des communes comptent moins de cinq cents habitants) où les planches peinent à se raccorder entre elles. Néanmoins, cette campagne de numérisation est un succès, et le département en profite pour installer une plate-forme *web* proposant différents géo-services à destination des communes. « *Acquérir la donnée géographique, c'est louable*, commente Olivier Le Reste, responsable SIG au CD 14. *Mais la valoriser et la diffuser, c'est encore mieux !* »

Les acquisitions de données s'enchaînent, et le conseil départemental s'équipe avec les outils logiciels adéquats pour gérer cette quantité d'informations. À l'heure actuelle, celui-ci possède essentiellement des produits ESRI (*ArcGIS server, ArcGIS online*). Les métiers utilisent des outils conçus par *GéoMap/Imagis* pour la gestion des routes et du cadastre. D'autres applications, comme la gestion des espaces naturels sensibles ou la gestion des conventions d'occupation, sont en cours de déploiement.

La loi NOTRE, une remise à plat

Avec l'adoption de la loi NOTRE, c'est tout un pan de la géomatique départementale qu'il faut revoir. Les impacts opérationnels sont nombreux, notamment les EPCI, qui voient leur surface augmenter et leur nombre se réduire, et même les communes, dont plusieurs fusionnent. « *Globalement, les relations se sont clarifiées, ce qui est un bon point. Cela fait du travail imprévu, mais le SIG s'adapte aux nouvelles réalités territoriales.* » Certaines compétences, comme celle des transports scolaires, sont transférées à la région. « *Les utilisateurs de la région auront accès aux données en* Extranet », explique Olivier Le Reste.

Valoriser, c'est bien le maître mot. C'est pourquoi l'équipe géomatique envisage l'installation d'un portail, *GéoCalvados*, qui va permettre aux collectivités territoriales du département d'accéder aux données de leurs territoires facilement, grâce à un simple accès *extranet*. Ce projet, qui résulte donc d'environ dix-sept ans d'investissements continus dans l'information géographique, a démarré avec l'idée de dénombrer le patrimoine constitué depuis donc bientôt deux décennies, pour en faire profiter les nombreuses applications métier maintenant disponibles chez tous les agents du conseil départemental. Un accès modernisé à l'information géographique, en somme. Mais, bien entendu, *Inspire* est passé par là, et le département s'est retrouvé obligé de publier *a minima* ses métadonnées vers l'extérieur – et même ses données, le conseil départemental ayant adopté une politique *Open Data* volontariste. « *La directive Inspire*, poursuit Olivier Le Reste, *peut être envisagée à la fois en tant que contrainte et en tant qu'opportunité. Contrainte, parce que publier des métadonnées exige de connaître celles-ci, et donc de maîtriser parfaitement ses jeux de données ; opportunité, car il s'agit là d'un des meilleurs vecteurs de diffusion et de valorisation. Quant à l'*Open Data*, cela nous a conduits par exemple à élaborer* OpenEquip 14*, la carte des équipements publics, en partenariat avec l'agence d'urbanisme de Caen, dont le rôle a été essentiel dans l'établissement du référentiel, et avec laquelle nous contractualisons régulièrement.* »

Pour ce faire, un catalogage des données est indispensable. Or, les quelques expérimentations menées jusqu'alors s'étaient révélées peu concluantes. Elles

Calvados's data go online

The department of Normandy has worked with the French SME Isogeo to list and publish online the data it owns and their associated metadata.

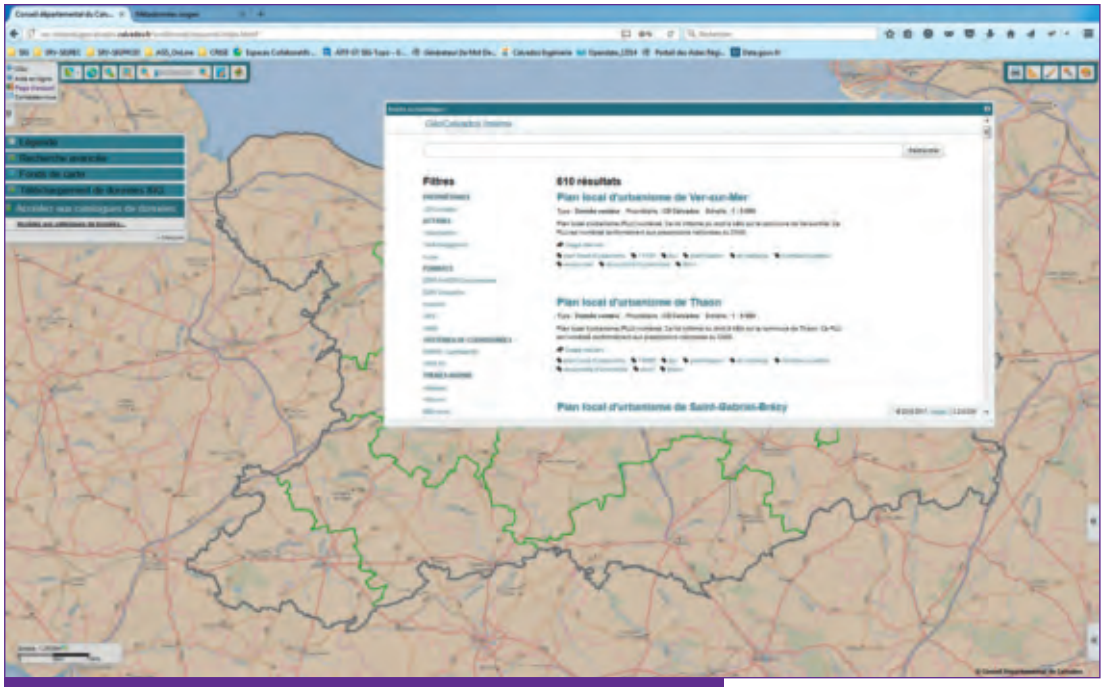
The Calvados, *département* in the very heart of Normandy, can boast his pioneering of the geomatics sector. Its local authority, the *Conseil Général du Calvados,* or *CG14,* was the first to order and supervise the taking of an orthophotography covering its entire territory in 2001, an operation it renewed in 2006. Soon after, it headed the digitalisation of the *département*'s cadaster. Not an easy task, since more than half of the communes tally less than five hundred people, and cadaster sheets' limits rarely line up. Nevertheless, this project succeeded. In its wake, the GIS service decided to set up an Internet server to offer geo-services to the communes. "*Acquiring geodata is fine*," Olivier Le Reste, head of the GIS service at the CG14 says. "*But promoting its use and handing it out are better yet!*"

Data acquisitions went on and on, and the *Conseil général* had soon to build an infrastructure to store and handle that mass of information. Nowadays, it is equipped with *ESRI* software (*ArcGIS server, ArcGIS online*) and *Geomap/Imagis's* – a French SME developing dedicated add-ons for the *ESRI* platform – products, that are used by the services in charge of road maintenance or cadaster. Other applications to help with the management of environmentally sensitive areas or leases of public property are under development.

The NOTRE law, a game changer

After the *NOTRE* (*Nouvelle Organisation du Territoire de la République, New Organisation of the Republic's Territory, NORET*) bill passage last year, much of the local GIS structure came into question. Operational impacts were plenty, as small communes and counties coalesced into wider entities. "*Relationships are clearer now, which is a good thing. Obviously, this means also unexpected work, but the GIS must adapt to the new territorial*

Les nouveaux EPCI du département.
Calvados's new county limits.



Résultat d'une interrogation du catalogue concernant le PLU.
Result of a urban planning document search.

avaient cependant permis de mettre en exergue la lourdeur des outils standards et surtout le côté fastidieux de la gestion des métadonnées. La documentation et le catalogage des données, pourtant prioritaire, avait toujours été différée, faute de temps et de motivation.

Olivier Le Reste entend alors parler d'une jeune *start-up, Isogeo,* qui propose des solutions de catalogage automatique. *Isogeo* avait déjà travaillé avec les départements voisins de l'Orne et de la Manche, mais, à l'époque, le Calvados était trop occupé avec la structuration de la base de données pour entreprendre des « bêta-tests » de la solution. Cette période de test passée, la solution *Isogeo* est désormais mature et apporte une vraie valeur ajoutée. Convaincu, Olivier Le Reste contacte *Isogeo* et la mise en œuvre du catalogage peut commencer.

Données et métadonnées toujours en phase

Techniquement, la solution *Isogeo* est distribuée en mode SaaS (Software as a Service). Elle se compose de deux parties : une installée en local sur le serveur SIG pour la génération automatique des métadonnées et une sur le cloud *Azure* de *Microsoft* pour la gestion et la publication des catalogues de données. L'outil installé en local, dénommé le *Scan*, utilise le célèbre ETL *FME* pour réaliser la lecture des données et ce quel que soit leur format (vecteur, raster, CAO). Le *Scan* parcourt périodiquement les données, crée, pré-remplit et met à jour les métadonnées. Il permet ainsi aux administrateurs de données de disposer facilement et rapidement d'un inventaire exhaustif à jour et documenté de leur base de données SIG. Un pré-requis indispensable au catalogage et à la valorisation de ce patrimoine.

Comment savoir si les données ont été modifiées ? La solution retenue est plus intelligente qu'une simple analyse des dates de dernière modification : le *Scan* calcule une signature pour chaque jeu de données et la compare à celui de la version précédente. En cas de non-concordance, cela signifie que les données ont été mises à jour, une nouvelle analyse est déclenchée et les métadonnées sont automatiquement mises à jour. Le service SIG décide lui-même de la fréquence de ces analyses.

Une fois enregistrées sur la plate-forme *Isogeo*, les métadonnées peuvent être diffusées au travers d'une API dans des applications, par exemple un *plug-in QGis*, un plug-in *ArcMap* ou bien un connecteur spécialement écrit pour la suite *Webville serveur*. Ce connecteur, actuellement en phase de test, sera ensuite disponible pour l'ensemble des utilisateurs

layout." A few former responsibilities, like school bus organisation, were transferred to the region. The GIS was unaffected, though. "*Region employees will access the data they need using our server's extranet service,*" Olivier Le Reste explains.

Promoting data usage is the clear priority. The extranet data access will be available not only to the region's employees, but also to all other local authorities operating in the *département*'s territory. This project, dubbed *GeoCalvados*, caps seventeen years of continuous dedication to geo-information. Initially, it consisted in collating the data stored over two decades and making them available to the dedicated software used internally by the *Conseil général* employees. But meanwhile the *Inspire* directive has changed the rules: the *département* now must also publish all the metadata it possesses. But not only, since the local assembly decided in favour of a strong open data policy. "*The Inspire directive,*" Olivier Le Reste comments, "*can be construed both as an opportunity and as a constraint. The latter, because publishing metadata entails perfect knowledge – and thus perfect mastery – of them. The former, because it is one of the most powerful means of dissemination and promotion. The* Open Data *policy led us to release* OpenEquip' 14*, the map of public facilities, in collaboration with Caen's urban planning agency. The agency's role was vital in the creation of the data. We routinely work with them.*"

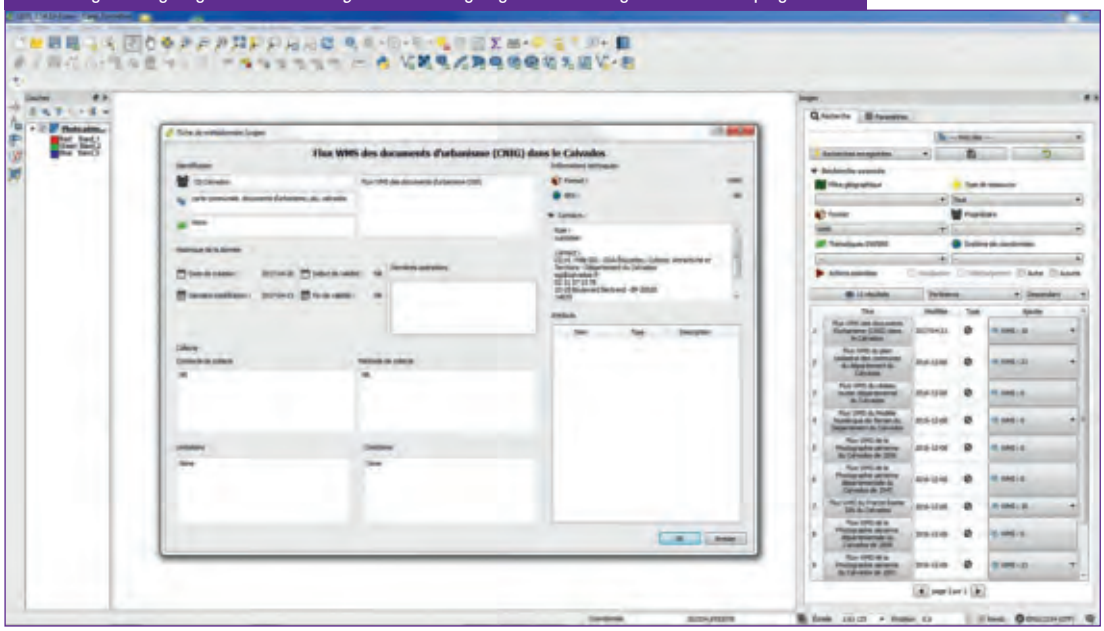
In order to publish data and metadata, cataloguing is paramount. However, the previous experiments were not fully successful: they had been hindered by the clumsiness of the existent tools and the tediousness of metadata processing. The task of structuring and documenting the data, urgent as it could be, had thus been delayed through lack of time and motivation.

At this point, Olivier Le Reste hears about a start-up called *Isogeo*, which develops automatic cataloguing software. As a matter of fact, *Isogeo* had already worked closely with the neighbouring *départements* of *Orne* and *Manche*. Le Reste's team, however, was too busy with the structuring of the database to join or even monitor that experiment. But when the head of the GIS department contacts the software company, the product is validated and fully functional. An agreement is signed and the installation begins.

Data and metadata always in sync

Technically speaking, *Isogeo*'s software uses the SaaS model. It is split into two different components: the first, running on the GIS server, automatically creates the metadata; the second, running on *Azure, Microsoft's* cloud infrastructure, manages and publishes them. The local routine, called "*the Scan*," relies on the ubiquitous ETL *FME* to read/convert data, whatever their

Dialogue *QGis* permettant de mettre à jour les métadonnées d'une couche WMS grâce au *plug-in Isogeo*.
QGis dialogue box giving access to the *Isogeo* data cataloguing software using the associated plug-in.



de la plate-forme *Géomap/Imagis*. Les métadonnées sont évidemment « moissonnables » en CSW par les organismes de plus haute importance, au sens de la directive *Inspire (GéoNormandie, GéoCatalogue National, datat.gouv.fr, etc.)*.

Trois mois de préparation

La mise en œuvre du catalogue de données a pris environ trois mois. Il a fallu analyser les données, en déduire des règles et des méthodes de catalogage et de publication. Ce sont plus de mille quatre cents jeux qui ont été passés au crible, ce qui représente environ seize millions d'objets et neuf mille attributs ! Certaines de ces couches sont des données historiques qui n'évoluent plus, et donc ont été cataloguées une fois pour toutes : « *Avant l'arrivée d'Isogeo,* indique Olivier Le Reste, *il n'existait pas vraiment de culture de l'historique, mis à part pour les photos aériennes et le réseau routier, qui est géré très spécifiquement.* »

GéoCalvados va donc diffuser ces données départementales à destination des communes et EPCI, qui deviendront sans doute les partenaires privilégiés de la plate-forme. « *Nous apportons des solutions pour les communes – beaucoup de communes – qui n'ont pas les moyens de s'équiper en SIG, même rudimentaire. Désormais, le cadastre, le Scot, etc., tout cela est à portée d'un seul clic. Nous ne voulons pas de "décrochage numérique", mais nous militons au contraire pour un renforcement de la coopération entre les différents échelons territoriaux. Par exemple, nous sommes le premier département à avoir initié le déploiement de la fibre en FTTH sur l'ensemble du territoire. Un projet pharaonique, mais nous ne voulons pas que les petites communes restent à l'écart en raison d'un débit insuffisant. L'accès à l'information géographique dépend de plus en plus du très haut débit.* »

Le serveur du conseil départemental pourrait-il également héberger les données géographiques des EPCI du Calvados ? « *C'est possible,* admet Olivier Le Reste, *mais cela demande réflexion.* »

Projets

Les prochaines étapes de *GéoCalvados* vont impliquer les partenaires externes du département, comme par exemple le Syndicat d'Énergie, ce qui pourrait conduire à une réflexion plus générale sur la « mobilité SIG ». Un autre projet consiste à intégrer le plan de déplacements cyclables départemental, ce qui suppose d'aller sur le terrain en utilisant des outils comme *Collector 4 ArcGIS*, et de monter un service *web* de type WFS-T pour autoriser les modifications à distance. Il faudra s'intéresser à cette occasion aux soucis de qualité des données. Plus généralement, Olivier Le Reste souhaite que les utilisateurs deviennent de plus en plus autonomes dans leur utilisation soit de l'*Intranet* géographique, soit des outils métiers. Enfin, la politique *OpenData* signifie que des API conformes aux protocoles standards WFS et WMS seront sûrement développés.

Du côté d'*Isogeo*, une nouvelle version des plug-in pour *Qgis* et pour *ArcMap* devraient voir le jour cet autonome. ◘

Le 9ème apéro géomatique et innovation

Organisé en partenariat avec *Business Geografic*, ce neuvième Apéro Géomatique & Innovation de l'AFIGEO sera accueilli à Lyon, au sein de l'espace événementiel de *Ciril GROUP*, le 21 septembre 2017. Partant de leur expérience, les intervenants témoigneront des clés de la réussite pour exporter le savoir-faire français dans le secteur du géo-numérique : actionner les leviers de l'innovation pour créer la valeur ajoutée, mobiliser les financements adaptés, développer des partenariats, s'inscrire dans des dynamiques collectives… Une large place dédiée aux échanges permettra d'élargir les retours d'expériences et d'ouvrir de nouvelles perspectives de collaboration !

Programme

- 18 h 30 : Accueil des invités et introduction
- 18 h 50 : Comment entreprendre l'innovation ?
- 19 h 10 : Comment trouver des financements ?
- 19 h 30 : Comment exporter l'innovation ?

format is: raster, vector, CAD. Using *Scan*, the database manager quickly and easily obtains a thorough, accurate and documented list of the database's contents. This is of course required for the cataloguing and dissemination of the data.

How does the *Scan* know that data have been altered? The underlying algorithm goes beyond a simplistic comparison between dates of file modification. Instead, the routine computes a checksum for each file or dataset, and compares the result with the previous (stored) one. If the two differ, then it means that a change has taken place. A new scan is triggered and the metadata are updated. It is up to the GIS team to set how often the scans are performed.

Once stored on the *Isogeo* server, the metadata are available through a documented API, and can be imported into various applications, for example *QGis* or *ArcMap*. The *WebVille server* also connects to *Isogeo* using a brand new software connector. This connector is currently under test. When validated, it will be made available free of charge to all other users of *WebVille*. The metadata can also be "harvested" by "higher level entities" (according to *Inspire*'s terminology), such as regional or national geodata infrastructures (*GéoNormandie, GéoCatalogue National, datat.gouv.fr, etc*).

A three month setup

The setup has been done over a three month period. Data had to be scrutinised and organised, rules and cataloguing methods created. More than a thousand four hundred data files were processed, representing sixteen million features and nine thousand attributes. Some of the data layers contain frozen historical data. The cataloguing of these data is done one and for all. "*Before we started to use Isogeo*, Olivier Le Reste carries on, *"there was no real concern to trace data history, barring aerial photography and road maintenance."*

GeoCalvados, the name of the future service, will push these data to the communes and counties. They will become natural project partners. "*We offer services for communes so small they cannot afford GIS software, even rudimentary. Henceforth, cadaster, urban planning documents, and so on, will just be one click away on a standard browser window. No need for additional software, no need for highly trained personnel. We are wary of the 'digital dropout' phenomenon and we do not want our smallest communes to be dumped on the wayside of the future 'digital highways'. Very high speed Internet access is paramount, and that's why we are currently installing FTTH everywhere in the department, a first in France. This titanic undertaking will spread over a couple of years, but we're proud of it.*" Could the server also host more detailed data on behalf of smaller local governments? "*It's an idea we should consider*."

And more to come!

The next step in *GeoCalvados* consists in bringing in partners, for example the power supply company, which could lead to an overall reflection on how to make data access compatible with mobile devices, commonly used by utilities. Similarly, the complete survey of bikeways should start shortly. That involves spending time in the field with mobile devices running *Collector-4ArcGIS* and opening a WFS-T type service to allow for remote modification of data. This will raise the problem of data quality. Generally speaking, Olivier Le Reste wants users to acquire more skill in using the internal geoservices or dedicated applications. Finally, the *Open Data* policy calls for the development of a WMS/WFS service.

Isogeo, on the other hand, plans to upgrade the services offered by its platform. A new version of the *QGis* and *ArcGIS* plugins should be rolled off this autumn. ◻

Évaluation de variables limnologiques grâce à des images Landsat

Danielle Teixeira Alves da Silva[1] (danielle_alves01@yahoo.com.br)
Aziz Serradj[1] (aziz.serradj@live-cnrs.unistra.fr)
Aline do Vale Figueiredo[2] (alinefigueiredo_@hotmail.com)
Vanessa Becker[3] (becker.vs@gmail.com)

[1] Université de Strasbourg, Faculté de Géographie et d'Aménagement
3 Rue de l'Argonne, 67000 Strasbourg, France.
[2] Université Fédéral du Rio Grande do Norte – UFRN,
Laboratoire de Ressources en Eau et Assainissement,
3000 Avenue Senador Salgado Filho, Campus Universitaire,
59078-970, Natal/RN, Brésil.
[3] Université Fédéral du Rio Grande do Norte – UFRN,
Département de Génie Civil,
3000 Avenue Senador Salgado Filho, Campus Universitaire,
59078-970, Natal/RN, Brésil.

Utilisation des images LANDSAT pour estimer la concentration de la chlorophylle-a et de la transparence de l'eau sur un territoire semi-aride du Nord-Est brésilien.

Introduction

Les ressources hydriques de surface ont la périodicité de recharge caractérisée par la variabilité spatiotemporelle du cycle de l'eau. Du fait de l'irrégularité temporelle du débit des masses d'eau, il n'y a pas de prévisibilité de la quantité totale d'eau qui peut être utilisée pour la demande humaine (Oki & Kanae, 2006). La pratique de construction de systèmes de stockage de l'eau, aussi appelés réservoirs, afin de favoriser la pérennisation des rivières, est devenue une pratique décisive pour réduire la vulnérabilité des zones soumises au manque et à l'irrégularité des précipitations (Medeiros, 2003 ; SERHID, 2006 ; Ventura, 2013). Malheureusement, l'action anthropique sur ces systèmes provoque la dégradation de la qualité environnementale et la perturbation des cycles d'eau et des nutriments.

L'eutrophisation artificielle est un processus dynamique d'origine humaine et peut être définie comme un type de pollution des eaux causé par un excès de nutriments, principalement en azote et en phosphore (Costa *et al.*, 2006 ; Esteves, 2011). Ce processus est intensifié dans les zones où il y a interférence humaine, du fait des émissions anthropogéniques de nutriments qui causent de profondes modifications qualitatives et quantitatives sur l'équilibre de l'écosystème aquatique :

a) Changements structuraux et fonctionnels ;
b) Baisse de la biodiversité, mortalité des poissons ;
c) Apparition de cyanobactéries, déplétion de l'oxygène et émission de toxines et de gaz à effet de serre (*Esteves*, 2011 ; Naselliflores, 2011).

Devant cette problématique, il est important d'élaborer des études sur la dégradation de la qualité de l'eau et les perturbations écologiques dues à la pollution hydrique. Les techniques conventionnelles de surveillance de la qualité de l'eau présentent quelques limitations d'ordre temporel et de couverture spatiale, ce que peut rendre difficile une analyse générale de tout le système en termes de qualité de l'eau (Nas *et al.*, 2010 ; Torbick *et al.*, 2013).

La télédétection à partir des images satellitaires représente une technique fiable et rentable pour les mesures synoptiques des principaux indicateurs de la qualité de l'eau, car elle est capable de mesurer la quantité du rayonnement solaire à différentes longueurs d'onde, réfléchie par la surface de l'eau, tout en offrant la possibilité d'estimer sa transparence, la concentration de solides en suspension et de la chlorophylle. Autrement dit, les informations spectrales des changements dynamiques telles que l'eutrophisation sont fournies avec une couverture spatiale et temporelle continue (Hellweger *et al.*, 2004 ; Olmanson *et al.*, 2008 ; Torbick *et al.*, 2013).

Plusieurs études antérieures montrent de fortes corrélations entre les données radiométriques obtenues par des systèmes de télédétection satellite et les variables de la qualité de l'eau. À partir de cela, il est possible d'établir des modèles d'estimation des principaux agents modificateurs du milieux aquatique (Harrington *et al.*, 1992 ; Zilioli *et al.*, 1994 ; Fuller & Minnerick, 2007 ; Nas *et al.*, 2010 ; Ventura, 2013 ; Santos & Pereira Filho, 2013 ; Gao *et al.*, 2015 ; Zhang & Han, 2015).

Dans ce contexte, l'objectif principal de cette étude est d'obtenir des modèles mathématiques afin d'estimer les concentrations de chlorophylle-a et la transparence de l'eau dans le réservoir Maréchal Dutra. À partir de la modélisation, nous cherchons également à créer des cartes d'estimation de la répartition spatiale de la concentration de chlorophylle-a et de la transparence de l'eau dans le réservoir Maréchal Dutra pendant la période 2010 à 2014.

La zone d'étude

Nous allons passer en revue ici, les caractéristiques climatiques et limnologiques de la région où le réservoir Maréchal Dutra est inséré.

Aspects climatiques

La région semi-aride du Brésil présente des aspects physico-climatiques et écologiques très spécifiques, en particulier si on considère que 92 % de la superficie totale du pays se caractérise par un étagement de climats : humide, intertropical subhumide et subtropical, de l'Amazonie au Rio Grande do Sul (Ab'Saber, 2003). Selon la classification climatique de Köeppen, le climat de la région est de type BSw'h (chaud et semi-aride, tendant à aride), avec un hiver sec, des températures élevées et des précipitations moyennes inférieures à 700 mm/an (Kottek *et al.*, 2006) avec répartition irrégulière des précipitations.

En termes de ressources en eau, la région semi-aride est fortement caractérisée par des rivières intermittentes dans des conditions naturelles, et la longue période de sécheresse est responsable des conditions climatiques de la région (Medeiros, 2003).

Caractéristiques du réservoir Maréchal Dutra

Le réservoir Maréchal Dutra est situé dans la région semi-aride du Nord-Est brésilien (cf. figure 1), aux coordonnées géographiques suivantes : longitude 36°35'08" W et latitude 06°23'55" E. Il est distant d'environ 13 km de la ville d'Acari. Sa construction a été terminée en 1959. Le barrage a été construit sur la rivière Acauã, sous la responsabilité du Département National Contre la Sécheresse (DNOCS) (SEMARH, 2015). Le réservoir Marechal Dutra a une capacité maximale d'accumulation d'eau de 44 421 480 m³, une profondeur maximale de 26,5 m et une surface de 805,67 ha (SEMARH, 2015).

Selon Panosso *et al.* (2007) et Petrovich (2009), les réservoirs des régions semi-aride brésiliennes sont associés au développement socio-économique régional, qui entraînent une forte demande d'usage en eau, tels que l'approvisionnement, la pêche, l'aquaculture et le tourisme. Toutefois, des phénomènes tels que l'eutrophisation affectent directement la qualité de l'eau dans les réservoirs de la région, ce qui représente un risque pour la santé publique en raison de la contamination liée à la présence d'algues toxiques, comme les cyanobactéries.

Depuis le début de son fonctionnement, le réservoir Maréchal Dutra présente des caractéristiques eutrophiques, avec de fortes concentrations de nutriments et une intense productivité de la biomasse algale, selon le rapport présenté par Okuda *et al.* (1963). Mesquita (2009) a évalué l'état trophique de six réservoirs du sous-bassin versant de la rivière Sérido, entre les années 2002 et 2008, et il a constaté que le réservoir Maréchal Dutra est considéré comme le plus eutrophisé, avec concentrations moyennes en PT de (148 ± 50,6), en Chl-a de (84,5 ± 63,8) et la plus petite transparence du disque de Secchi de (0,5 ± 0,3 m). Dans une étude plus récente (Figueiredo, 2015), le réservoir Maréchal Dutra a présenté variations de concentrations de PT entre 21,20 et 517,00 et de Chl-a entre 2,40 et 167,87. Les valeurs de référence pour caractériser des conditions eutrophiques en régions semi-arides sont 50 pour le PT et 15 pour la Chl-a (Thorton & Rast, 1993).

Ces conditions limnologiques, associées à l'intensité de la lumière et de la chaleur, aux eaux alcalines et à un fort temps de rétention hydrique, représentent une situation typique en réservoirs du semi-aride brésilien, et elles sont responsables de la prolifération de cyanobactéries. Celles-ci prédominent sur la biomasse algale et produisent d'importantes quantités de toxines (Costa *et al.*, 2006; Romo *et al.*, 2013).

Méthodologie

La méthodologie est basée sur la corrélation entre les données des deux importantes variables limnologiques pour l'identification trophique de plans d'eau avec les données de réflectance issues d'images satellitaires multi-spectrales.

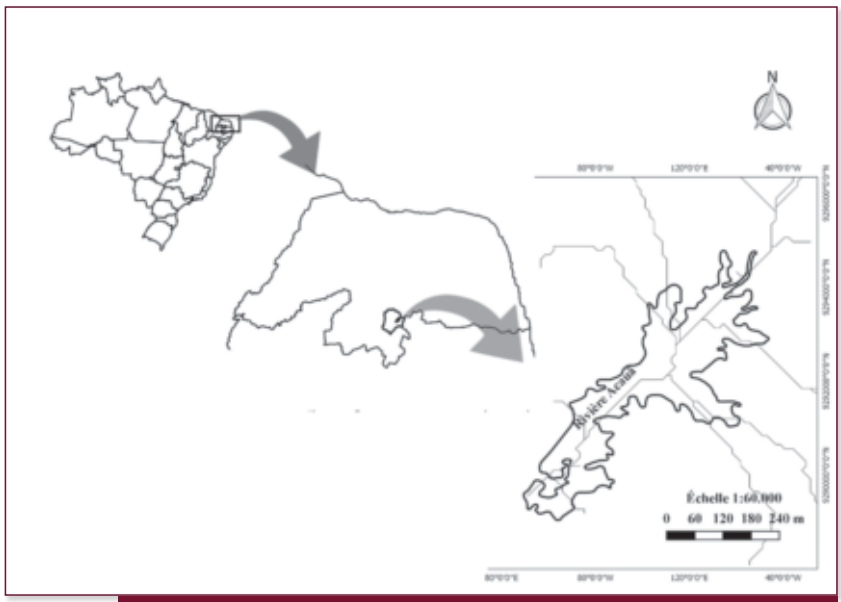


Figure 1. Localisation du réservoir Maréchal Dutra, Rio Grande do Norte, Brésil.

Prélèvement des échantillons d'eau

L'évaluation de la qualité de l'eau a été faite sur la base de données du projet de recherche MEVEMUC (*Monitoramento da Evaporação e Mudanças Climáticas no Rio Grande do Norte* / Surveillance de l'Évaporation et des Changements Climatiques du Rio Grande do Norte). Les échantillons de l'eau ont été collectés mensuellement pendant la période de juin 2010 à décembre 2014 à proximité du déversoir, considéré comme le lieu le plus profond du réservoir, ce qui a permis d'obtenir des échantillons tout le long de la colonne d'eau.

La transparence de l'eau a été mesurée *in situ* à l'aide du disque de Secchi (turbidimètre). Pour toute la profondeur de l'épilimnion, des échantillons intégrés ont été recueillis à l'aide de la bouteille de Van Dorn. Les échantillons ont été transférés aux bouteilles de polyéthylène, conditionnées dans des boîtes réfrigérées pour leur acheminement au laboratoire, où a été réalisée l'analyse de Chl-a en utilisant la méthode calorimétrique (Jespersen & Christoffersen, 1988).

Données radiométriques des images LANDSAT

Nous avons utilisé des images acquises par les plateformes orbitales Landat 5, 7 et 8, téléchargées à partir du site *http://earthexplorer.usgs.gov*, sous la responsabilité de l'*United States Geological Survey* (USGS/NASA). La sélection des images a été faite en tenant compte de la méthodologie appliquée par Olmanson *et al.* (2008). Les dates des collectes *in situ* et des images choisies par capteur sont répertoriées sur le tableau 1 ci-dessous.

Année	Date de collecte *in situ*	Date de capture de l'image	LANDSAT/ capteur
2010	23 juin	25 juin	5/TM
2011	25 juillet	22 juillet	7/ETM+
2012	19 avril	19 avril	
	19 juillet	24 juillet	
	19 septembre	10 septembre	
2013	12 mars	21 mars	
	22 mai	1 juin	8/OLI
	25 juin	17 juin	
	23 juillet	27 juillet	7/ETM+
	14 août	04 août	8/OLI
	16 octobre	23 octobre	
2014	15 janvier	11 janvier	
	11 février	20 février	7/ETM+
	21 mars	24 mars	
	16 avril	17 avril	8/OLI
	03 juin	25 mai	7/ETM+
	1 août	7 août	8/OLI
	1 septembre	23 août	
	1 octobre	10 octobre	

Tableau 1. Dates de collecte *in situ*, dates des images choisies et capteurs correspondants.

Pré-traitement numérique des images orbitales

Les images obtenues sur le portail de l'USGS ont été d'abord reprojetées de la projection initiale au système de coordonnées planes UTM de la zone 24 Sud Datum SIRGAS 2000, projection officielle au Brésil.

Avec le logiciel ENVI version 5.2, nous avons obtenu la Valeur Numérique (DN) de chaque pixel, et toujours avec la même application, nous avons réalisé la conversion des valeurs DN en valeurs de luminance au sommet de l'atmosphère ou réflectance Top of Atmosphere (TOA).

Les équations ci-dessous montrent les étapes de conversion des valeurs DN en réflectance TOA.

$$L_\lambda = \text{Gain} \times \text{DN} + \text{offset}$$

Où :

$\mathbf{L_\lambda}$ est la luminance spectrale vraie ;
DN est la valeur nombre digital ;
Gain est la valeur « gain » pour une bande spectrale déterminée ;
Offset est la valeur « offset » pour une bande spectrale déterminée.

$$\rho_\lambda = \frac{\pi \bullet L_\lambda \bullet d^2}{\text{ESUN}_\lambda \bullet \cos(\theta_{sz})}$$

Avec :

ρ_λ la réflectance TOA (sans dimensions) ;
$\mathbf{L_\lambda}$ la luminance vraie ;
d la distance Terre-Soleil exprimée en unités astronomiques ;
$\mathbf{ESUN_\lambda}$ l'éclairement solaire exo-atmosphérique (en haut de l'atmosphère) ;
θ_{sz} angle zénithal solaire (degrés) ($\theta_{sz} = 90^\circ - \theta_{se}$, θ_{se} étant l'angle d'élévation solaire).

Les équations 1 et 2 sont issues de *http://landsat.usgs.gov/how_is_radiance_calculated.php*.

Modèle de régression

La réflectance d'intérêt TOA a été extraite des images en utilisant une fenêtre d'extraction de 3 × 3 pixels, ce qui correspond à 8 100 m² sur le terrain, et le point central de la fenêtre coïncide avec la localisation du point de collecte *in situ* (Mushtag & Lala, 2016 ; Ekercin, 2007 ; Zhang & Han, 2015). Les valeurs de réflectance TOA obtenues correspondent aux bandes spectrales du visible,

du proche infrarouge (PIR) et du moyen infrarouge (SWIR).

Pour construire le modèle de régression entre les paramètres radiométriques et les variables limnologiques, nous avons utilisé les bandes spectrales :

a) Individuellement ;
b) Avec les résultats des combinaisons arithmétiques entre les bandes spectrales (Bee, 2009) ;
c) Avec des indices très connus en télédétection :

• L'indice de végétation par différence normalisée (NDVI), proposé pour la première fois par Rouse *et al.*, 1973 :

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

• L'indice de différence normalisée de l'eau (NDWI), proposé par Gao en 1996 ;

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

• L'indice de différence normalisée de l'eau transformé (NDWIt), proposé par Gao en 1996 ;

$$NDWI_t = \frac{NIR - SWIR}{NIR + SWIR}$$

Cette méthode nous a permis d'obtenir les corrélations entre les réflectances TOA et les paramètres Chl-a et transparence du disque de Secchi. Les combinaisons qui ont donné des corrélations statistiquement significatives (seuil de confiance ρ = 0,05) ont été soumises au calcul des modèles de régression. Avec les paramètres de l'équation de la droite obtenue, nous avons pu estimer les valeurs des variables limnologiques sur tout le réservoir Maréchal Dutra.

Répartition spatiale des variables limnologiques estimées

Pour l'estimation des variables limnologiques, nous avons pris plusieurs valeurs de réflectance à partir de différents points le long du réservoir, en plus du point de collecte (point de référence). À l'aide du logiciel *ArcGis* version 10.3, nous avons pu utiliser les outils d'analyse spatiale pour cartographier les variables limnologiques estimées par le modèle de régression. L'organigramme ci-dessous résume les principales étapes pour obtenir les cartes thématiques (cf. figure 2).

Résultats et discussions

Les résultats remettent à la répartition spatiale des deux importants paramètres limnologiques, et nous allons discuter leurs variations dans la période de cette étude.

Données limnologiques obtenues

Les graphiques de la figure 3 montrent les variations des variables limnologiques, leurs moyennes correspondantes et le volume cumulé du réservoir Maréchal Dutra.

Les données de volume cumulé montrent que l'année 2011 est considérée comme une année pluvieuse, avec des pluies intenses les premiers mois. En juillet de la même année, le réservoir a atteint sa capacité volumétrique maximale (cf. figure 3). À partir de cette date, nous observons une décroissance de l'accumulation d'eau dans le réservoir. Même après quelques jours de légères chutes de pluie pendant les périodes de la saison pluvieuse pour la région, qui correspond aux mois de mars, avril et mai, la situation ne s'est pas suffisamment inversée.

Le manque de pluie au cours de la période analysée dans cette étude a été accompagné par la recrudescence de la concentration de Chl-a et la diminution de la transparence de l'eau, ce qui indique une détérioration de la qualité de l'eau du réservoir. Au deuxième semestre de 2013, le réservoir a présenté

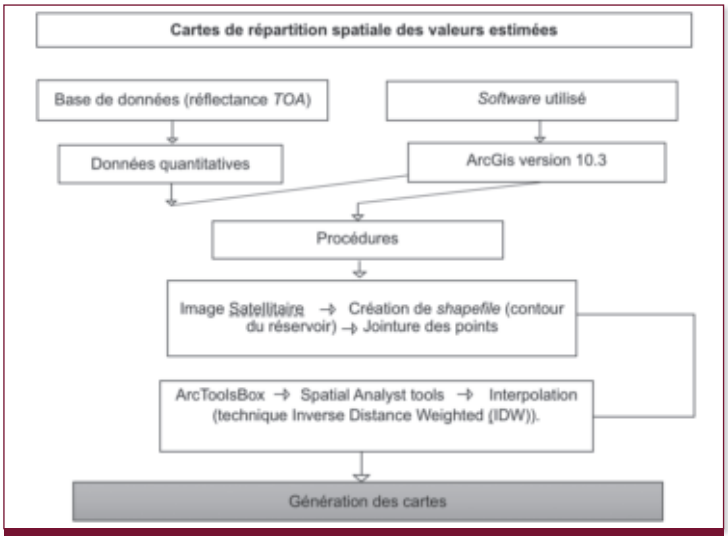


Figure 2. Organigramme de la méthodologie utilisée pour la création des cartes thématiques de répartition des variables limnologiques.

les plus hautes concentrations de Chl-a (cf. figure 3a) et les plus petites valeurs de transparence de l'eau (cf. figure 3b). Les données montrent la tendance des conditions plus eutrophiques en périodes de déficit hydrique pour ce réservoir (cf. figure 3).

Il faut remarquer que la période d'étude comprend la transition d'une période pluvieuse, caractérisée par l'accumulation maximale d'eau dans le réservoir en 2011, à une période qui a connu la crise hydrique la plus grave des soixante dernières années (MI, 2014). Des études antérieures sur les réservoirs du semi-aride au Rio Grande do Norte indiquent que les fluctuations du volume stocké sont l'un des plus grands modificateurs des conditions physico-chimiques de l'eau, qui sont, par exemple, le régime de mélange et d'accumulation de nutriments (Braga *et al.*, 2015 ; Figueiredo, 2015). En plus, la variation de volume stocké d'eau interfère sur la distribution temporelle de la biomasse du phytoplancton (Costa *et al.*, 2006).

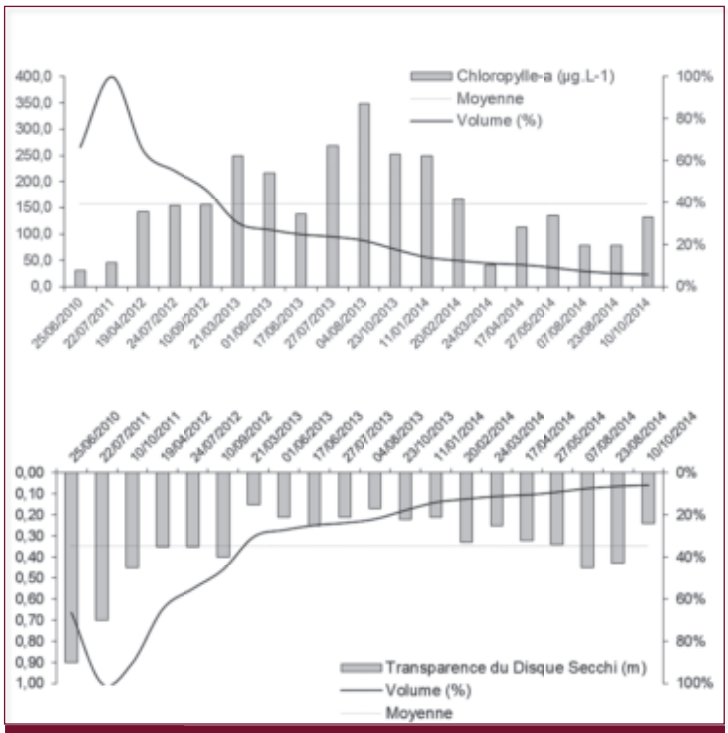


Figure 3. (A) Variations de Chlorophylle-a et du volume cumulé dans le réservoir Maréchal Dutra. (B) Variations de la transparence du disque de Secchi et du volume cumulé dans le réservoir Maréchal Dutra. Le volume cumulé a été calculé à partir de la capacité volumétrique maximale du réservoir, égale à 44,4 millions de m³ (source : SEMARH, 2015).

Validation du modèle

À l'aide de l'application statistique « *SigmaPlot* », nous avons testé toutes les combinaisons spectrales qui ont présenté des corrélations significatives pour la construction du modèle de régression. Les meilleurs coefficients de corrélation de *Spearman*, en tenant compte de $\rho < 0,05$, comprennent la Chl-a et le NDVI (0,60) et la transparence de Secchi et le NDVI (-0,49). Le coefficient de corrélation entre Chl-a et transparence de Secchi est égal à -0,76 (cf. tableau 2). Ainsi, le modèle de régression linéaire simple a été composé par le NDVI en plus des variables limnologiques étudiées.

Pour la relation entre la Chl-a et la transparence du disque de Secchi, le modèle de régression montre une relation statistiquement significative ($\rho < 0,001$; $r^2 = 0,503$) (figure 4). Le NDVI s'est révélé le meilleur prédicateur pour expliquer les variations de Chl-a ($\rho = 0,039$; $r^2 = 0,228$) et de transparence du disque de Secchi ($\rho = 0,022$; $r^2 = 0,272$) (cf. figure 5).

À partir de l'équation de régression, nous avons obtenu les modèles d'estimation des variables limnologiques :

$$\text{Chl-a } (\mu g \bullet L^{-1}) = 228,194 + (433,672 \times \text{NDVI})$$

$$\text{Transparence du disque de Secchi (m)} = 0,177 - (1,013 \times \text{NDVI})$$

Nous avons trouvé une relation négative entre Chl-a et la transparence du disque de Secchi (= -0,76). Nous pouvons donc suggérer que la recrudescence de la concentration de biomasse primaire a provoqué une diminution significative de la transparence de l'eau. L'analyse de corrélation suppose que 76 %

Variables	Chl-a (µg.L^{-1})	Transparence du disque de Secchi (m)
Chl-a (µg-L^{-1})	1,0000	-0,7550
Transparence du disque de Secchi (m)	-0,7550	1,0000
NDVI	0,6000	-0,4880

Tableau 2. Coefficients de corrélation de Spearman entre les variables limnologiques et les réflectances TOA (p value < 0,05).

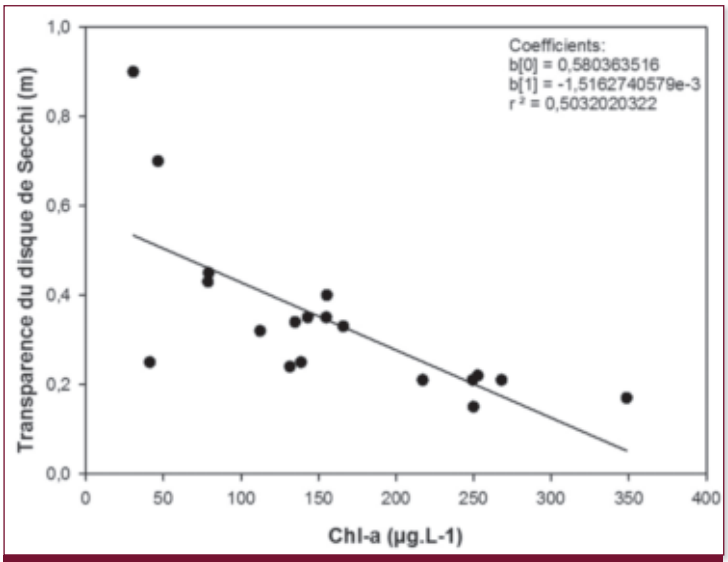


Figure 4. Relation entre Chl-a et transparence du disque de Secchi ($\rho < 0,001$; $r^2 = 0,503$). La droite de la régression linéaire s'ajuste aux semis de points.

de la variabilité de la transparence de Secchi peut être inversement estimée par la concentration de Chl-a, ce qu'indique une prédominance de turbidité organique dans le réservoir Maréchal Dutra pendant la période étudiée.

Les organismes photosynthétiques qui contiennent la chlorophylle-a absorbent la lumière principalement dans les longueurs d'onde des bandes bleue et rouge du spectre électromagnétique et réfléchissent celle-ci dans la bande correspondant à la bande verte et la bande du PIR du spectre (Ekercin, 2007 ; Bee, 2009). Méléder *et al.* (2003) soulignent cependant que la biomasse algale peut présenter une faible absorption à 673 nm (bande spectrale rouge) due à la chlorophylle-a et à ses produits de dégradation dans les cellules mortes, qui restent à la surface et caractérisent aussi visuellement le phénomène d'eutrophisation des plans d'eau. Bee (2009) souligne encore que la réflexion dans le proche infrarouge est liée à la structure de la paroi cellulaire des algues.

Le phénomène d'eutrophisation est bien caractérisé par la haute production primaire, c'est-à-dire la haute densité d'organismes chlorophylliens. La chlorophylle-a est présente dans toutes les espèces végétales, ainsi le NDVI utilisé comme prédicateur des variables limnologiques nous indique une remarquable prolifération algale dans le réservoir Maréchal Dutra. Au semi-aride du nord-est brésilien, la densité algale dans les plans d'eau est si évidente que nous avons l'habitude d'appeler « tapis verdâtre » l'apparence de la surface d'eau.

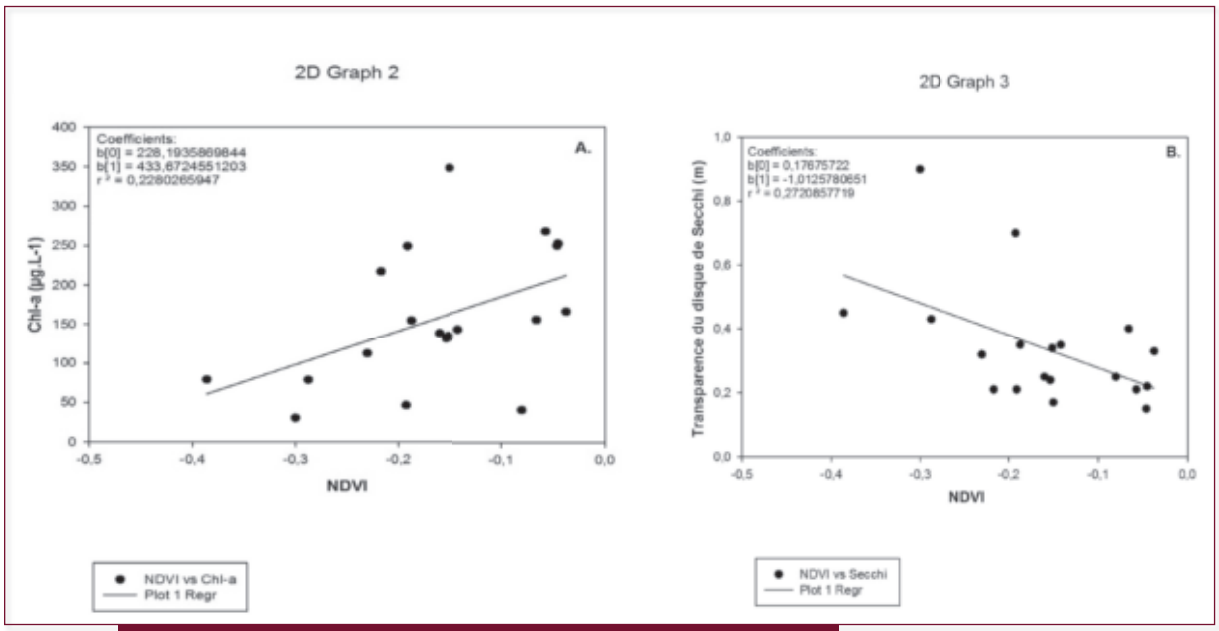


Figure 5. Modèle de régression entre (A) Chlorophylle-a et NDVI ($r^2 = 0,228$) et (B) transparence du disque de Secchi et NDVI ($r^2 = 0,272$). La droite de la régression linéaire s'ajuste aux semis de points.

Répartition des variables limnologiques

Les figures 6 et 7 ci-dessous montrent, en séquence chronologique, les cartes de répartition spatiale plus représentatives pour cette étude, et mettent en évidence l'évolution du phytoplancton et de la transparence de Secchi, respectivement, après l'application des modèles de régression sur les résultats du traitement des images satellitaires.

Sur les cartes de répartition spatiale, nous pouvons observer que la variabilité des paramètres limnologiques est plus sensible durant les périodes de sécheresse, contrairement aux périodes d'accumulation maximale d'eau, où les concentrations de Chl-a et les profondeurs de Secchi présentent les meilleures conditions observées.

Pour des raisons logistiques et financières, la surveillance de la qualité de l'eau dans le réservoir Maréchal Dutra est réalisée à partir d'un seul point de collecte, considéré comme le plus stratégique du fait de sa localisation dans la zone la plus profonde du réservoir et de sa proximité avec la prise d'eau pour les réseaux de distribution. En prenant en compte les résultats des cartes de répartition spatiale, nous pouvons suggérer que, au cours des périodes de sécheresse, il serait intéressant d'avoir plus de points d'échantillonnage sur l'ensemble du réservoir. Cela pourrait conduire à une meilleure compréhension de la dynamique aquatique et à identifier les sources de pollution.

La carte de répartition de chlorophylle-a montre l'évolution des concentrations de Chl-a avec la diminution de la quantité d'eau stockée dans le réservoir.

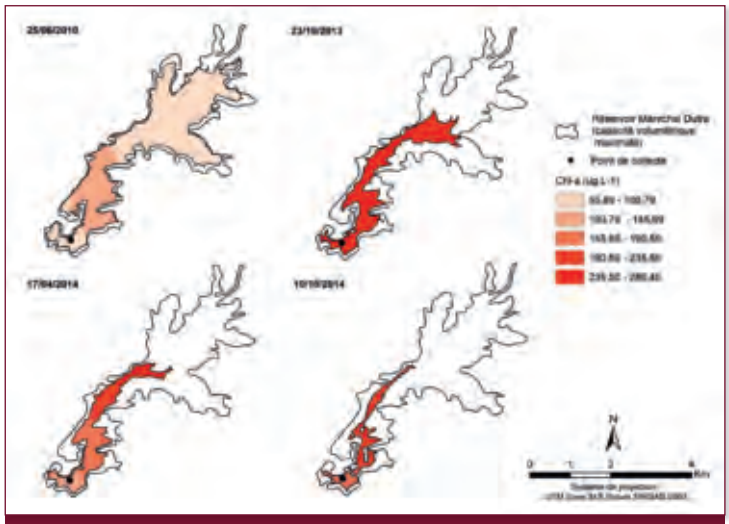


Figure 6. Cartes de la répartition spatiale de l'estimation de la chlorophylle-a dans le réservoir Maréchal Dutra pour la période d'étude.

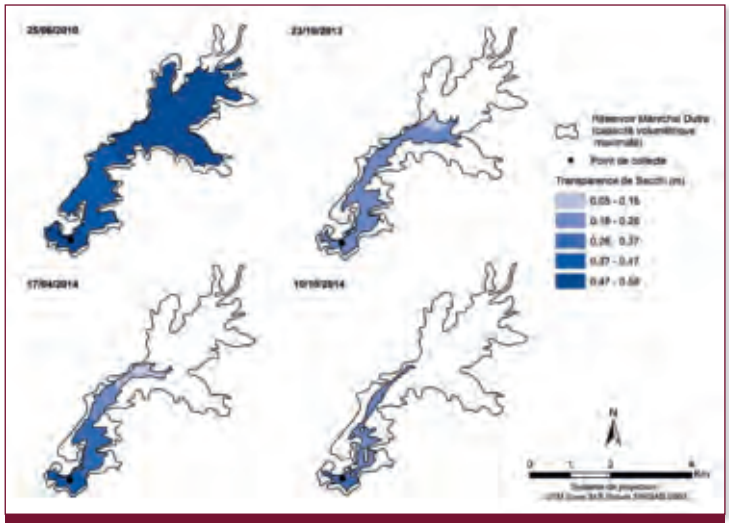


Figure 7. Cartes de la répartition spatiale de l'estimation de la transparence du disque de Secchi dans le réservoir Maréchal Dutra pour la période d'étude.

Nous attendions des concentrations plus fortes lors de la chute du volume d'eau stockée. Cependant, nous observons une décroissance relative des concentrations de cette variable entre le changement de l'année 2013 et 2014. En 2014, une augmentation des totaux pluviométriques a été enregistrée dans la région, mais néanmoins inférieure à la moyenne attendue pour la saison pluvieuse (ANA, 2015). Ceci est bien insuffisant pour caractériser l'année 2014 comme année pluvieuse, mais l'entrée d'eau de ruissellement a atténué les fortes concentrations de Chl-a.

Par rapport à la transparence de l'eau, les plus grandes valeurs de profondeur de Secchi sont observées dans les années d'accumulation d'eau plus élevée. Avec la diminution du volume d'eau du réservoir, la transparence d'eau a été affectée négativement. Ce comportement était attendu étant donné que les intenses

concentrations de Chl-a sur la couche d'épilimnion augmentent le degré de turbidité de l'eau. Les précipitations de l'année 2014 ont présenté un aspect positif pour la mesure de la transparence de Secchi, mais avec la sécheresse prolongée au cours du dernier trimestre de 2014, nous pouvons remarquer une perte de volume considérable d'eau, et la diminution de ± 0,3 m de profondeur de Secchi par rapport aux valeurs du début d'étude, en 2010.

Le modèle d'estimation obtenu dans cette étude a fourni la valeur de concentration maximale de Chl-a égale à 280,40 $\mu g \bullet L^{-1}$, ce qui correspond à une erreur relative de plus de 20 % en comparaison de la valeur maximale obtenue en laboratoire (348,68 $\mu g \bullet L^{-1}$). Les principales limitations de ces modèles mathématiques sont liées à la modification sur les données radiométriques en raison des phénomènes atmosphériques (Polidório *et al.*, 2004) et aux changements sur les conditions environnementales des réservoirs, identifiés par des échantillonnages représentatifs capables de caractériser les différents degrés trophiques dans la zone d'étude (Urbanskia *et al.*, 2016).

Conclusion

Les avantages qui justifient l'investissement dans les nouvelles techniques de télédétection des milieux aquatiques sont les suivants : réduction des coûts ; accès rapide et régulier aux données radiométriques et couverture spatiale plus large, ce qui permet d'analyser avec facilité et précision les relations d'un plan d'eau avec d'autres éléments présents / absents dans la zone d'étude.

Pour le réservoir Maréchal Dutra, les résultats trouvés montrent la viabilité et le potentiel exploratoire de l'utilisation des images satellitaires sur la détection de la répartition spatiales des variables qui caractérisent la qualité de l'eau. L'utilisation d'images *Landsat* multi-spectrales a permis la cartographie des concentrations d'organismes photosynthétiques et la cartographie de la profondeur de Secchi dans le réservoir Maréchal Dutra.

La surveillance à travers de la télédétection peut être un allié important pour la reconnaissance globale de la zone d'étude, ce qui peut aider la prise de décision pour la mise en œuvre des techniques de lutte contre les processus de détérioration de la qualité de l'eau. ◻

Bibliographie

Ab'sáber, A., 2003. *Os domínios de natureza no Brasil: potencialidades paisagísticas.* São Paulo: Ateliê Editorial.

AGÊNCIA NACIONAL DE ÁGUAS – ANA. (2015). *Encarte Especial sobre a Crise Hídrica.* Brasília: Superintendência de Planejamento de Recursos Hídricos.

Bee, S., 2009. *Seasonal and Annual Changes in Water Quality in the Ohio River Using Landsat based measures of Turbidity and Chlorophyll-A.* Thèse de doctorat, University of Cincinnati.

Braga, G. G. ; Becker, V. ; Oliveira, J. N. P. ; Mendonça JR, J.R. ; Bezerra, A. F. M. ; Torres, L. M. ; Galvão, A. M. F. ; Mattos, A., 2015. *Influence of extended drought on water quality in tropical reservoirs in a semiarid region. Acta Limnologica Brasiliensia*, v. 27, n. 1, p. 15-23.

BRASIL. MINISTÉRIO DA INTEGRAÇÃO NACIONAL. *Construção do Plano Nacional de Segurança Hídrica.* Brasília: MI (2014). Disponible à l'adresse *http://www.mi.gov.br/web/guest/pagina-integracao-nacional-inicia-construcao-do-plano-nacional-de-seguranca* (au 22 mai 2017).

Costa, I. A. S. ; Santos, A. P. ; Silva, A. A. ; Melo, S. G. ; Mendonça, J. M. S. ; Panosso, R. F. ; Araújo, M. F. F., 2006. *Florações de Algas Nocivas: Ameaça às Águas Potiguares. Revista da FAPERN*, v. 1, n. 4, p. 14-16.

Ekercin, S., 2007. *Water quality retrievals from high resolution IKONOS multispectral imagery: A case study in Istanbul, Turkey. Water, Air & Soil Pollution*, v. 183, p. 239-251.

Esteves, F. A., 2011. *Fundamentos da Limnologia.* In: Esteves, F. A. ; Meirelles – Pereira, A. *Eutrofização Artificial.* Cap. 27, p. 625-655.

Figueiredo, A. V., 2015. *Influência de eventos hidrológicos extremos na qualidade da água de reservatórios na região tropical semiárida.* Dissertation de maîtrise (UFRN).

Fuller, L. M. ; Minnerick, R. J., 2007. *Predicting Water Quality by Relating Secchi-Disk Transparency and Chlorophyll A Measurements to Landsat Satellite Imagery for Michigan Inland Lakes, 2001-w2006. United States Geological Survey*, 4 p.

Gao, Y. ; Gao, J. ; Yin, H. ; Liu, C. ; Xia, T. ; Wang, J. ; Huang, Q., 2015. *Remote sensing estimation of the total phosphorus concentration in a large lake using band combinations and regional multivariate statistical modeling techniques. Journal of Environmental Management*, v. 151, p. 33-43.

Harrington, J. A. ; Schiebe, F. R. ; Nix, J. F., 1992. *Remote Sensing of Lake Chicot, Arkansas: Monitoring Suspended Sediments, Turbidity, and Secchi Depth with Landsat MSS Data. Remote Sens. Environ.*, v. 39, p. 5-27.

Hellweger, F. L. ; Schlosser, P. ; Lall, U. ; Weissel, J. K., 2004. *Use of satellite imagery for water quality studies in New York Harbor*. Elsevier, v. 61, p. 437-448.

Jespersen, A. M. & Christoffersen, K., 1988. *Measurements of chlorophyll-A from phytoplankton using ethanol as extraction solvent. Hydrobiologia*, v. 109, p. 445-454.

Kottek, M. *et al.*, 2006. *World Map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift*, v. 15, n. 3, p. 259-263.

Medeiros, W. D. A., 2003. *Sítios geológicos e geomorfológicos dos municípios de Acari, Carnaúba dos Dantas e Currais Novos, região Seridó do Rio Grande do Norte*. Dissertation de maîtrise (UFRN).

Méléder, V. ; Launeau, P. ; Barillé, L. ; Rincé, Y., 2003. Cartographie des peuplements du microphytobenthos par télédétection spatiale visible-infrarouge dans un écosystème conchylicole. Comptes Rendus Biologies, V. 326, n. 4, pages 377-389.

Mesquita, T. P. N., 2009. *Eutrofização e Capacidade de Carga de Fósforo de Seis Reservatórios da Bacia do Rio Seridó, Região Semi-Árida do Estado do RN*. Dissertation de maîtrise (UFRN).

Nas, B. ; Ekercin, S. ; Karabörk, H. ; Berktay, A. ; Mulla, D. J., 2010. *An Application of Landsat-5TM Image Data for Water Quality Mapping in Lake Beysehir, Turkey. Water, Air and Soil Pollut*, v. 212, p. 183-197.

Naselli-Flores, L., 2011. *Mediterranean Climate and Eutrophication of Reservoirs: Limnological Skills to Improve Management. Eutrophication: Causes, Consequences and Control*. v. 2. c. 6.

Oki, T. & Kanae, S., 2006. *Global Hydrological Cycles and World Water Resources*. Freshwater Resources, v. 313, p. 1068-1072.

Okuda, T., 1963. *Nota sobre as condições hidrográficas no açude Acarí – R.G. do Norte. Departamento Nacional de Obras Contra as Secas (DNOCS). Ministério da Integração Nacional: Trabalhos Técnicos*, v. 3, n. 1, p. 33-38.

Olmanson, L. G. ; Bauer, M. E. ; Brezonik, P. L., 2008. *A 20-year Landsat water clarity census of Minnesota's 10, 000 Lakes. Remote Sensing of Environment*, v. 122, n. 11, p. 4086-4097.

Panosso, R. ; Costa, I. A. S. ; Souza, N. R. ; Attayde, J. L. ; Cunha, S. R. S ; Gomes, F. C. F., 2007. *Cianobactérias e cianotoxinas em reservatórios do Estado do Rio Grande do Norte e o potencial controle das florações pela Tilápia do Nilo (Oreochromis niloticus). Oecologia Brasilienses*, v. 11, n. 3, p. 433-449.

Pereira, R., 2007. *Aplicabilidade de métodos de Sensoriamento Remoto na avaliação e monitoramento do estado trófico de lagoas costeiras do Rio Grande do Sul – Brasil*. Dissertation de maîtrise (UFRGS).

Polidório, A. M. ; Imai, N. N. ; Tommaselli, A. M. G., 2004. *Índice indicador de corpos d'água para imagens multiespectrais. In: I Simpósio de Ciências Geodésicas e Tecnologias da Geoinformação*.

Torbick, N. ; Hession, S. ; Hagen, S. ; Wiangwang, N. ; Becker, B. ; Qi, J., 2013. *Mapping inland lake water quality across the LowerPeninsula of Michigan using Landsat TM imagery. International Journal of Remote Sensing*, v. 34, p. 7607-7624.

Romo, S. ; Soria, J. ; Fernandez, F. ; Ouahid, Y. ; Baron-sola, A., 2013. *Water residence time and the dynamics of toxic cyanobacteria. Freshwater Biology*, v. 58, p. 513-522.

Santos, F. C. ; Pereira Filho, W. 2013. *Reflectância espectral relacionada aos constituintes opticamente ativos da água do reservatório Passo Real, RS, Brasil. In: XVI Simpósio Brasileiro de Sensoriamento Remoto*.

SECRETARIA DO MEIO AMBIENTE E DOS RECURSOS HÍDRICOS DO RIO GRANDE DO NORTE – SEMARH, 2015. Programa de Monitoramento Volumétrico e Fiscalização. 141 p.

Urbanskia, A. ; Wochnaa, A. ; Bubakb, I. ; Grzybowskib, W. ; Lukawska-Matuszewskab, K. ; Łackad, M., 2016. *Application of Landsat 8 imagery to regional-scale assessment of lake water quality. International Journal of Applied Earth Observation and Geoinformation*, v, 51, p. 28-36.

Ventura, D. L. T., 2013. *Uso do Sensoriamento Remoto para Monitoramento da Concentração de Clorofila-A em Açudes do Semiárido*. Thèse de doctorat (UNB).

Zhang, C. ; Han, M., 2015. *Mapping Chlorophyll-a Concentration in Laizhou Bay Using Landsat 8 OLI data. In: XXVI IAHR World Congress*, Pays-Bas.

Zilioli, E. ; Brivio, P. A. ; Gomarasca, M. A., 1994. *A correlation between optical properties from satellite data and some indicators of eutrophication in Lake Garda (Italy).The Science of the Total Environment*, v. 158, p. 127-13.

Une chaîne de traitement pour évaluer la qualité des données OSM

Seconde partie : Évaluer la qualité en regroupant les contributeurs

Damien Garaud <damien.garaud@oslandia.com>, Raphaël Delhome <raphael.delhome@oslandia.com>,

A data-oriented framework to assess OSM data quality

Part 2 : Evaluate element quality by clustering contributors

Hugo Mercier <hugo.mercier@oslandia.com>, Oslandia

Après un premier article consacré à la présentation des outils et l'extraction des données, cette suite (et fin) détaillera la manière dont nous utilisons l'historique des données OSM pour évaluer leur qualité.

Nous avons vu dans le premier article (Garaud *et al.*, 2017) comment interpréter la donnée OSM, puis l'intégrer dans une chaîne de traitements *Python*. Cette seconde partie explique comment utiliser les données interprétées pour alimenter des modules de type *machine learning* qui formeront des groupes de contributeurs. À leur tour, ces groupes nous donneront une idée de la qualité des données expertisées.

Dans cet article, nous allons tout d'abord générer des informations utilisateurs à partir de l'historique, c'est-à-dire constituer des métadonnées. Puis, nous allons utiliser une classification automatique fondée sur des algorithmes classiques de *machine learning* pour regrouper ces utilisateurs en classes. Enfin, nous conclurons cet examen par la réalisation de nouvelles cartes.

Génération des métadonnées

Nous avons déjà couvert tout le processus d'extraction, d'analyse et de stockage des données dans la première partie de cet article. Nous allons donc supposer que nous avons à disposition une table **elements**, par exemple celle des environs de Bordeaux.

Des opérations d'agrégation sont nécessaires pour mieux décrire l'évolution des objets OSM et ses différents contributeurs.

Définition des métadonnées

Nous allons nous intéresser à des critères plus complexes que la simple géométrie, vu que d'autres informations sont aisément disponibles en examinant l'historique. Trois cas principaux se présentent :

• Métadonnées des objets eux-mêmes : même s'il paraît évident de considérer ces métadonnées, elles sont trop verbeuses pour apporter quoi que ce soit en termes de mesure de qualité, et leur verbosité les rend coûteuses à traiter (en termes de temps machine : trois millions de données pour une ville moyenne comme Bordeaux) ;

• Métadonnées de mise à jour *(changeset)* : chaque modification de la donnée est caractérisée par une opération particulière (création, destruction, mise à jour). On peut isoler des lots de changements « productifs », qui correspondent à des zones étendues et des modifications durables ;

• Métadonnées utilisateur : nous pouvons classer chaque utilisateur en fonction du nombre de ses contributions, de la durée de validité de celles-ci ou encore de la diversité des éléments modifiés.

Ces métadonnées peuvent être collectées encore plus efficacement en considérant les contributeurs eux-mêmes. Nous supposerons que la qualité des données OSM peut être déduite du type d'utilisateur qui contribue le plus à la mise à jour d'un nœud, chemin ou relation. En d'autres termes, nous pouvons qualifier la qualité de la donnée en ne retenant que la « réputation » du meilleur contributeur. Une autre solution, que nous retiendrons par la suite, consiste à associer la qualité d'un élément à la réputation du dernier contributeur à l'avoir changé.

Extraction des métadonnées

Plusieurs types d'informations peuvent être extraites de l'historique des données OSM :

• Temporelles : durée de vie, dates de mise à jour ;

• Spatiales : quels sont les changements faits, y a-t-il un « motif » répétitif sur lequel les modifications portent ;

• Fréquences : nombre moyen de changements par élément, etc. ;

• Numérales : combien de nœuds, voies ou relations l'utilisateur a changé ;

• Descriptives : informations additionnelles sur les éléments modifiés.

Exemple : un utilisateur OSM typique

Considérons par exemple le cas de l'utilisateur possédant l'identifiant *4074141*.

```
uid
4074141
first_at                    2016-06-06
00:00:00
last_at                     2016-06-09
00:00:00
lifespan
3
n_inscription_days
```

After a first article dedicated to the framework presentation and to data extraction, this sequel will carry on explaining the methodology that links OSM data history and quality assessment.

We have seen in the first article (Garaud *et al.*, 2017) how to parse *OpenStreetMap* data, and how to integrate it in a *Python* workflow. We now look at feeding the parsed data into a machine learning algorithm to cluster similar contributors; these clusters will allow us to assess data quality.

We will first create contributor-focused information from the data history, *i.e.* user metadata. Then we will set up an unsupervised learning procedure to group users by using classical machine learning algorithms. In the last section of the paper, we will conclude about OSM data quality by drawing derived maps.

User Metadata Production

As the extraction of OSM data history has already been explained in the previous article, we posit that we have at hand an element table that mirrors the elements' history. We continue to use the Bordeaux area to illustrate our workflow.

Some data clustering operations will be neededto help describe the evolution of both OSM objects and their creators.

Metadata Definition

We will take into account other characteristics beside geometric data: additional information is available within contributions' history. Three main cases may be distinguished:

• Element metadata: while it sounds trivial to consider it, element metadata is often too verbose to provide any improvement in terms of quality assessment, and its processing is fairly intensive in terms of computing (circa three million chunks for a medium city like Bordeaux);

• Changeset metadata: each changeset corresponds to a number of alterations (creation, updates, deletion). One may pick out "productive" changesets, that contain a large amount of enduring modifications;

• User metadata: each user may be productive or not, depending on the number of modifications authored, the amount of time the contributions lasts, and the diversity of modified elements.

The information may be gathered more efficiently by considering the contributors themselves. We posit that OSM data quality can be qualified by the class of users who contribute the most to each node, way or relation. In other words, we consider the most experienced user who has contributed on an element as a clue to the element's quality. The quality of an element may also be tied to the class (more or less experienced) of its last editor. We will based the next section on this latter assumption.

Metadata Extraction

Several kind of features may be extracted from OSM contribution history:

• Time-related features: how users contribute through time (e.g. *lifespan*);

• Changeset-related features: they describe the user's modification "strategy" (e.g. number of local changesets);

• Contribution frequency (average number of modifications per element);

• Element-related features: they trivially disclose how many nodes, ways and relations the user modified;

• Modification-related features: they provide more details about modifications' types (e.g. how many way were created).

A typical OSM user case

To illustrate the various metadata features, let's take the user whose ID is 4074141:

```
uid
4074141
first_at                    2016-06-06
00:00:00
last_at                     2016-06-09
00:00:00
lifespan
3
n_inscription_days
258
n_activity_days
2
n_chgset
3
dmean_chgset
```

```
258
n_activity_days
2
n_chgset
3
dmean_chgset
0
nmean_modif_byelem
2.94061
n_total_modif
1832
n_total_modif_node
1783
n_total_modif_way
46
n_total_modif_relation
3
n_node_modif
1783
n_node_modif_cr
0
n_node_modif_imp
1783
n_node_modif_del
0
n_node_modif_utd
0
n_node_modif_cor
598
n_node_modif_autocor
1185
n_way_modif
46
n_way_modif_cr
0
n_way_modif_imp
46
n_way_modif_del
0
n_way_modif_utd
0
n_way_modif_cor
23
n_way_modif_autocor
23
n_relation_modif
3
n_relation_modif_cr
0
n_relation_modif_imp
3
n_relation_modif_del
0
n_relation_modif_utd
0
n_relation_modif_cor
2
n_relation_modif_autocor
1
```

Cet utilisateur est enregistré comme contributeur depuis 258 jours. Son *lifespan* (durée de vie) sur le site *OSM* est d'environ trois jours. Il a effectué deux modifications à deux dates différentes, et trois changements durant sa durée de vie, d'une durée moyenne d'environ vingt-deux minutes. Il semble modifier les objets sur lesquels il travaille trois fois par session. C'est trop peu pour conclure qu'il s'agit d'un robot, quoiqu'il paraisse très peu confiant dans ses modifications…

Cet utilisateur est très actif dans la région de Bordeaux : il a proposé 1 832 modifications, qui se répartissent en 1 783 sur des nœuds, 46 sur des chemins et 3 sur des relations. L'intégralité des modifications sur les nœuds correspond à des améliorations (pas de création, pas de destruction). 598 de ces modifications ont été reprises par d'autres utilisateurs, et 1 185 ont fait l'objet d'auto-corrections ; aucun changement n'a été définitif. Le tableau pour les routes et relations n'est pas sensiblement différent. En somme, nous avons identifié un utilisateur qui contribue activement à l'amélioration de la base, mais pas suffisamment bien pour que son travail soit considéré comme « définitif ».

Nous pourrions également ajouter quelques informations concernant le type de logiciel utilisé pour les éditions de données, ce que nous ne ferons pas pour des raisons de concision.

Nous pouvons aisément imaginer que l'examen de l'ensemble des contributeurs selon ces critères permet de les regrouper en classes.

Classification automatique à partir des métadonnées

Nous venons de voir comment les métadonnées peuvent être constituées à partir de l'agrégation de données disponibles dans l'historique. Nous avons donc défini un ensemble de quarante indicateurs décrivant le comportement des utilisateurs, puis appliqué ces critères sur les 2 073 utilisateurs ayant contribué à la carte des environs de Bordeaux.

Nous allons maintenant détaillé comment classer ces contributeurs à l'aide d'algorithme de *machine learning* automatisés.

Transformation des métadonnées utilisateur

Comme illustré figure 1, les métadonnées ne suivent pas une loi normale. Il faut donc les « normaliser » avant de pouvoir les utiliser dans les algorithmes d'apprentissage.

```
0
nmean_modif_byelem
2.94061
n_total_modif
1832
n_total_modif_node
1783
n_total_modif_way
46
n_total_modif_relation
3
n_node_modif
1783
n_node_modif_cr
0
n_node_modif_imp
1783
n_node_modif_del
0
n_node_modif_utd
0
n_node_modif_cor
598
n_node_modif_autocor
1185
n_way_modif
46
n_way_modif_cr
0
n_way_modif_imp
46
n_way_modif_del
0
n_way_modif_utd
0
n_way_modif_cor
23
n_way_modif_autocor
23
n_relation_modif
3
n_relation_modif_cr
0
n_relation_modif_imp
3
n_relation_modif_del
0
n_relation_modif_utd
0
n_relation_modif_cor
2
n_relation_modif_autocor
1
```

This user has been registered as an OSM contributor for 258 days; their lifespan on the OSM website is almost three days; they have changed the data on two different days. During their lifespan, they produced three changesets, the mean duration of which is around 22 minutes. They seem to modify each OSM element almost three times. That's not enough to indicate a bot, however they look quite unsure about their contribution.

This user is very active around the Bordeaux area: they proposed 1,832 modifications, 1,783 to nodes, 46 to ways and three to relations. Amongst the 1,783 modifications on nodes, there are 1,783 updates (no creation, no deletion). 598 modifications have been touched up by other users, the rest being self-corrections. No node modification was found to be definitive.

We can draw a comparable picture for ways and relations. We thus think we have picked a user that contributes much to improve OSM elements; however their contributions are never good enough to remain in force for long. We could tack on information about the OSM editors used by this contributor (not done here for brevity's sake).

If we consider every single user who has contri buted on a given area, we can easily figure out that groups could arise.

Unsupervised Learning With User Metadata

In the last section, we have seen that user metadata can be easily created from OSM data history. Using this method, we created a set of forty "features" describing the user's behaviour and analysed 2,073 contributors to the Bordeaux area. We will now detail how to use this metadata to classify OSM users, with the help of some classical machine learning procedures.

User Metadata Transformation

As illustrated by figure 1, the metadata features are not normallydistributed. However, all standard machine learning algorithms need normalised variables. To comply to this rule, we need to scale our variables to simple bounds (e.g. [0, 100] or [-1, 1]). Some of those variables can be recast as percentages of other variables:

- The number of node/way/relation modifications w/r to all modifications;
- The number of created/improved/deleted elements w/r to all changes, for each element type;
- The number of changesets made with a given editor w/r to all changesets.

Une méthode plus intelligente pour représenter les attributs utilisateur consiste à utiliser un système de conversion pour ramener les variables entre des bornes précises (par exemple [0, 100] ou [-1, 1]). Premièrement, certains indicateurs peuvent être exprimés comme pourcentage d'autres indicateurs :

• Le nombre de modifications des nœuds/chemins/relation par rapport au total ;

• Le nombre de créations/modifications/destructions parmi toutes les modifications ;

• Le nombre de changements réalisés avec un éditeur particulier parmi tous les éditeurs possibles.

D'autres indicateurs peuvent être normalisés par définition. Par exemple, l'ancienneté d'un utilisateur ne peut dépasser celle de la base de données elle-même.

Enfin, les indicateurs peuvent être normalisés en comparant les utilisateurs entre eux. Par exemple, savoir qu'un utilisateur a effectué *n* modifications est intéressant, mais ne dit rien quant à son activité par rapport aux autres contributeurs. C'est pour cela que nous utiliserons ici des distributions cumulatives.

Une fois les fonctions de distribution calculées, nous savons que l'utilisateur n° 4074141 a modifié plus de nœuds, de chemins et de relations que 97,3, 2,5 et 0,2 % de tous les autres utilisateurs et que, parmi ses modifications, 100 % étaient des améliorations, etc.

L'étape finale consiste à ramener tous les indicateurs dans la même gamme de valeur [*min, max*]. Comme les indicateurs sont assez biaisés, nous le faisons avec un simple algorithme de type *Min-Max*, afin d'éviter trop de distorsion.

```
from sklearn.preprocessing import
RobustScaler
scaler = RobustScaler(quantile_
range=(0.0, 100.0))
X = scaler.fit_transform(user_md.values)
```

Analyse en composantes principales

Réduire la dimensionnalité d'un problème apparaît comme un pré-requis indispensable avant de procéder à une classification. Comme nous l'avons indiqué, nous disposons d'un panel de quarante variables. Cela paraît relativement faible pour réaliser une PCA (nous pourrions effectuer un regroupement direct en utilisant nos variables normalisées) ; toutefois, pour des questions de clarté, nous avons décidé d'implémenter cet algorithme.

Le principe de l'analyse en composantes principales est de projeter des variables multi-dimensionnelles (plusieurs coordonnées) sur un espace de dimension inférieure. Ceci permet d'isoler les coordonnées indépendantes et d'éliminer les redondances. Deux règles empiriques permettent de choisir le nombre de composantes : les valeurs propres des coordonnées (≥ 1) et la variance expliquée (≥ 70 %). Ici, la première des deux règles ne s'applique pas, car nous n'appliquons pas une normalisation standard (soustraction de la moyenne et division par l'écart type) ; toutefois, la seconde nous donne six coordonnées (qui expliquent 72 % de la variance).

L'algorithme PCA est importé du module **sklearn**, et prend le nombre de composantes comme paramètre. La nouvelle projection linéaire s'obtient en appliquant la fonction **fit_transform**. La contribution de chaque objet pour chaque nouvelle coordonnée est ensuite stockée dans le modèle **model**.

```
from sklearn.decomposition import PCA

model = PCA(n_components=6)
Xpca = model.fit_transform(X)
pca_cols = ['PC' + str(i + 1) for i in
range(6)]
pca_ind = pd.DataFrame(Xpca,
     columns=pca_cols,
        index=user_md.index)
pca_var = pd.DataFrame(model.compo-
nents_,
     index=pca_cols,
    columns=user_md.columns).T
pca_ind.query("uid == 4074141").T
uid   4074141
PC1 -0.117666
PC2  1.145482
PC3  0.273033
PC4 -0.095888
PC5 -0.151745
PC6  0.931930
```

Après l'analyse, l'information pertinente de chaque utilisateur est résumé en six valeurs, dont la signification reste à trouver. Ces valeurs varient entre -1 (contribution fortement négative) et 1 (contribution fortement positive). Ces contributions sont figurées dans le tableau 3.

Nos six variables peuvent être décrites comme suit :

• PC1 (28,5 % de la variance totale) est fortement dépendant des modifications des relations. Cette

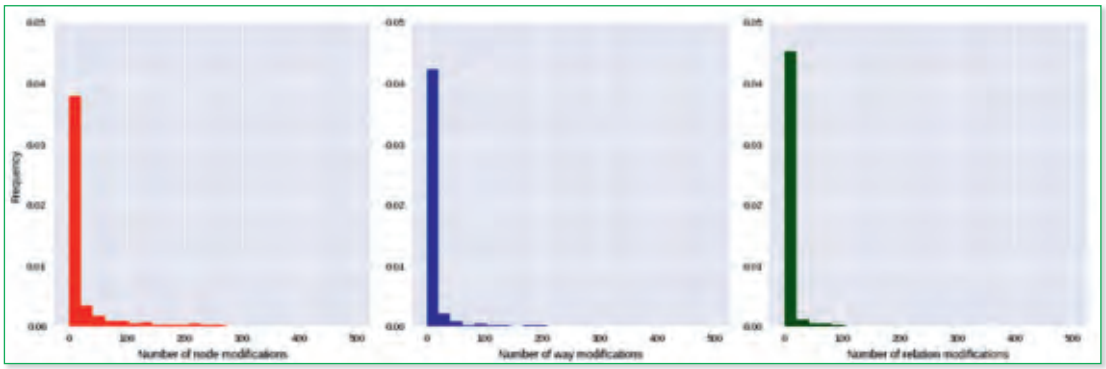


Figure 1 : Histogramme du nombre de nœuds, chemins et relations autour de Bordeaux.
Figure 1: Histogram of node, way and relation modification amounts around Bordeaux.

Other features can be normalised using intrinsic boundaries; for instance, user lifespan can not be larger than the OSM website lifespan itself.

Finally, features can be scaled by comparing users: knowing that a user did N changes is interesting, however it tells nothing about how many users are more or less active. This the definition of a cumulative distribution function, that we shall apply on all the other features.

After this phase, we find that the ID 4074141 user did more node, way and relation changes than 97.3, 2.5 and 0.2% of other users respectively, thathis node modificationswere all updates, and so on.

A final scaling ensures that all features have the same range. As they are skewed, we use a simple Min-Max rule, in order to avoid to distort our data too much:

```
from sklearn.preprocessing import
RobustScaler

scaler = RobustScaler(quantile_
range=(0.0, 100.0))
X = scaler.fit_transform(user_md.values)
```

The Principle Component Analysis (PCA)

Classifying data often requires a prior reduction in data dimensionality. In our case, we start from forty variables, a number which hardly justifies the recourse to PCA (we could directly run a clustering algorithm on our normalised dataset); however for clarity's sake during the result interpretation, we decided to insert this step into the workflow.

The Principle Component Analysis consists in a linear projection of multi-dimensional individuals on a smaller dimension space. It identifies uncorrelated coordinates and reduces redundancy by discarding coordinates that turn out to be linear combination of other ones. Two rules of thumb give the dimension of the space to project on: the explained variance proportion (at least 70%) and the eigenvalues of components (larger than 1).

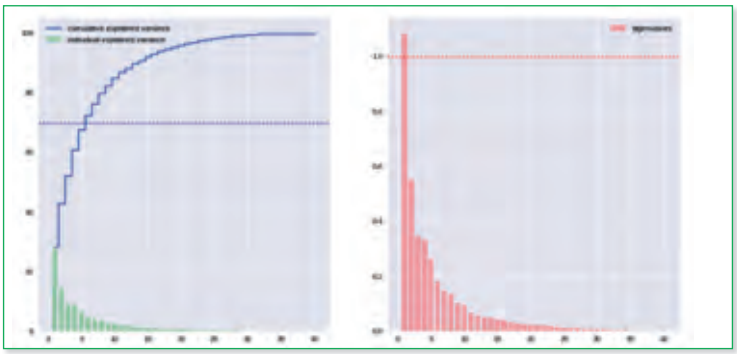


Figure 2 : Analyse de la variance des métadonnées utilisateur et nombre idéal de composants PCA.
Figure 2: User metadata variance analysis and ideal number of PCA components.

In our case, the second rule of thumb is inapplicable, as we do not use a standard scaling process (e.g. value less mean, divided by standard deviation). The first rule, however points to the use of six reduced components (which explain around 72% of the total variance).

The PCA algorithm is imported from the **sklearn** module. It takes the number of components as a parameter. The **fit_transform** function outputs the new linear projection. The contribution of each feature to the new components is stored into the model variable.

```
from sklearn.decomposition import PCA

model = PCA(n_components=6)
Xpca = model.fit_transform(X)
pca_cols = ['PC' + str(i + 1) for i in
range(6)]
pca_ind = pd.DataFrame(Xpca,
```

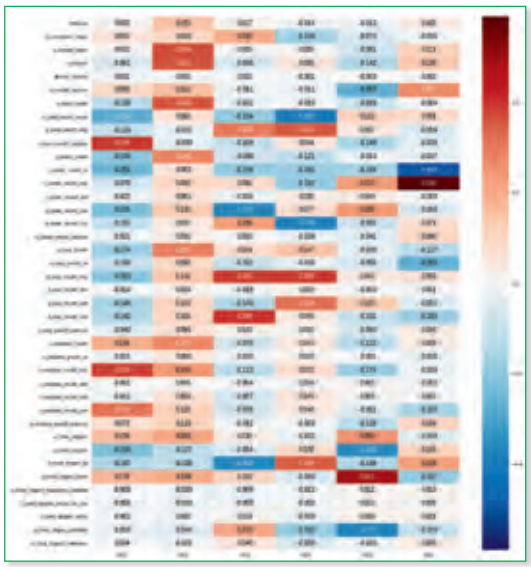


Figure 3 : Contribution de chaque objet aux composantes PCA.
Figure 3: Feature contribution to each PCA components.

variable est élevée lorsque l'utilisateur a réalisé un grand nombre d'améliorations sur les relations (ainsi que, accessoirement, quelques modifications de nœuds et de chemins), et que ces améliorations ont été reprises par d'autres utilisateurs. C'est donc un indicateur de spécialisation sur des structures complexes. Cette variable se rapporte également aux contributions effectuées par des utilisateurs « étrangers » (hors de la région d'intérêt, ici Bordeaux), habitués à *JOSM* ;

• PC2 (14,5 % de la variance totale) caractérise l'expérience et la polyvalence de chaque utilisateur : cette variable sera élevée pour les utilisateurs ayant une grande activité, beaucoup de contributions locales et globales, sur tous les types d'objets. Cette seconde composante caractérise également assez bien *JOSM* ;

• PC3 (9,1 % de la variance totale) décrit les contributions liées aux chemins effectués par des utilisateurs anciens mais pas très actifs. Une valeur élevée est liée à un taux élevé de correction, mais cela ne veut pas dire grand-chose : un contributeur ancien a de fortes chances d'avoir entrée une donnée maintenant périmée, donc mise à jour. Cette variable est corrélée à *Potlatch* et *JOSM*, les éditeurs OSM les plus communs ;

• PC4 (8,7 % de la variance totale) ressemble à PC3 en ce sens qu'il est fortement corrélé à la modification des chemins. Cependant, il s'applique aux nouveaux utilisateurs : inscription plus récente, contributions moins corrigées, et plus souvent valides. L'éditeur associé à cette variable est plutôt *iD* ;

• PC5 (6,9 % de la variance totale) dénote une spécialisation d'utilisateurs très actifs sur les nœuds. Les modifications associées sont toujours d'actualité. Cependant PC5 est lié à des utilisateurs qui ne sont pas très à l'aise sur notre zone d'étude, même s'ils sont l'auteur de nombres de changements ailleurs. *JOSM* est clairement l'éditeur correspondant ;

• PC6 (4,8 % de la variance totale) dépend fortement de la mise à jour des nœuds, par opposition à leur création (un comportement similaire émerge également pour les chemins). Tout ceci pointe vers les spécialistes de la géographie locale : une quantité appréciable de contributions locales, mais un faible nombre de contributions globales. Comme PC4, l'éditeur associé est plutôt *iD*.

Nous pouvons maintenant revenir au cas de l'utilisateur 4074141.

Cet utilisateur est très expérimenté (PC2 élevé) même si cette expérience est locale (PC5 négatif). La valeur de PC6 confirme cette hypothèse. Nous pouvons suspecter que cet utilisateur est assez polyvalent (PC2) avec une spécialité sur les corrections de nœuds (PC6).

Même si cet exercice semble plutôt abstrait, la comparaison entre cette interprétation et les faits relatés dans le premier paragraphe semble satisfaisante.

Regroupement des utilisateurs en fonction de leur activité passée

Dans cette sous-section, nous allons classifier l'ensemble des utilisateurs sans préjuger de leur aisance à manipuler les données géo-spatiales ou leur maîtrise de l'API OSM. Nous allons calculer les groupes avec l'algorithme des k-moyennes, le seul paramètre d'entrée étant les informations contenues dans les variables réduites.

Comme l'analyse en composantes principales, l'algorithme des *k-moyennes* est caractérisé par un paramètre, le nombre de groupes. Combien de ceux-ci peuvent être identifiés ? Nous ne pouvons nous référer qu'aux conseils donnés par des implémentations modernes. Comme illustré figure 4, nous avons choisi la méthode « coude et silhouette ». Le premier paramètre représente la variance au sein du groupe, c'est-à-dire la dispersion des observations dans les groupes. Il diminue

```
columns=pca_cols, index=user_md.index)
pca_var = pd.DataFrame(model.compo-
nents_, index=pca_cols,
          columns=user_md.columns).T
pca_ind.query(«uid == 4074141 »).T

uid   4074141
PC1 -0.117666
PC2  1.145482
PC3  0.273033
PC4 -0.095888
PC5 -0.151745
PC6  0.931930
```

After running the PCA, the information about each user is summarised in six values, whose meaning has to be clarified.

The feature factors in each components vary from -1 (a strong negative contribution) to 1 (a strong positive contribution). These contributions are plotted in the figure 3.

Our six components may be described as follows:

• PC1 (28.5% of total variance) depends on relation modifications: this component will be high if the user did a lot of relation updates (and very few node and way modifications) that have been further corrected by other users. It denotes a specialisation in complex structures. PC1 also reflects contributions from "foreign" users (*i.e.* not from the area of interest, here the Bordeaux area), familiar with *JOSM*;

• PC2 (14.5% of total variance) indicates how experienced and versatile users are: this component will be high for users tallying a high number of activity days, a lot of local as well as global changesets, and high numbers of node, way and relation editions. This second component is linked to *JOSM* too ;

• PC3 (9.1% of total variance) links to way-centric contributions of old, occasional users. A high value correlates to many corrected contributions. However, that's no surprise: old editions are more likely to have been updated. *Potlatch* and *JOSM* are the most used editors;

• PC4 (8.7% of total variance) is akin to PC3, in the sense that it is strongly correlated with way modifications. However, it points to newer users: more recent inscription date, less overwritten, more up-to-date contributions. The associated preferred editor is *iD*.

• PC5 (6.9% of total variance) refers to node specialisation in very productive users. The associated modifications are overall improvements that are still valid. However, PC5 is linked with users that are not familiar with our area of interest, even if they authored a lot of changesets elsewhere. *JOSM* stands out as the corresponding editor.

• PC6 (4.8% of total variance) strongly depends on node improvements, in contrast to node creations (a similar behaviour emerges for ways, too). This less important component pins down local specialists: a fair amount of local changesets, but a small total of changeset quantity. As in PC4 case, the editor used for those contributions is *iD*.

Now that we have explained the signification of the new components, we can apply them to user 4074141.This user is really experienced (PC2 high), even if this experience tends to be local (PC5 negative). The fairly high value for PC6 bolsters that analysis. We can figure that the user is quite versatile (PC2), and focuses on node improvements (PC6).

Even if this exercise in profiling may look quite abstract, the results match the description given a few sections before reasonably well.

Cluster the User out of Their Past Activity

In this section the set of users will be classified without any prior knowledge of their identity, their experience with geospatial data or their mastership of OSM API. We will create clusters using the *k-means algorithm* with only one input, the information about user past contributions contained into PCA components. Similar to PCA, the k-means algorithm takes a parameter, the number of clusters.

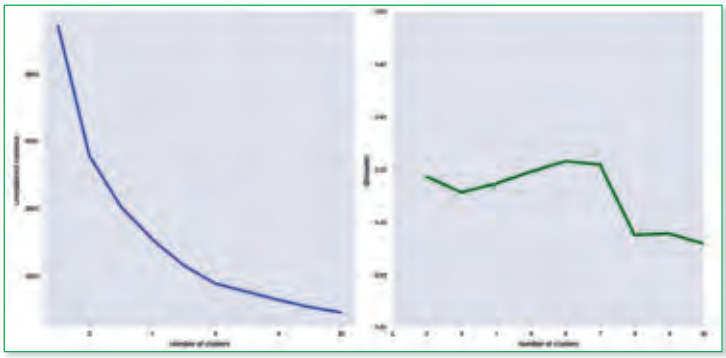


Figure 4 : Nombre optimal de groupes donné par la méthode « coude et silhouette ».
Figure 4: Optimal cluster number according to the elbow and silhouette method.

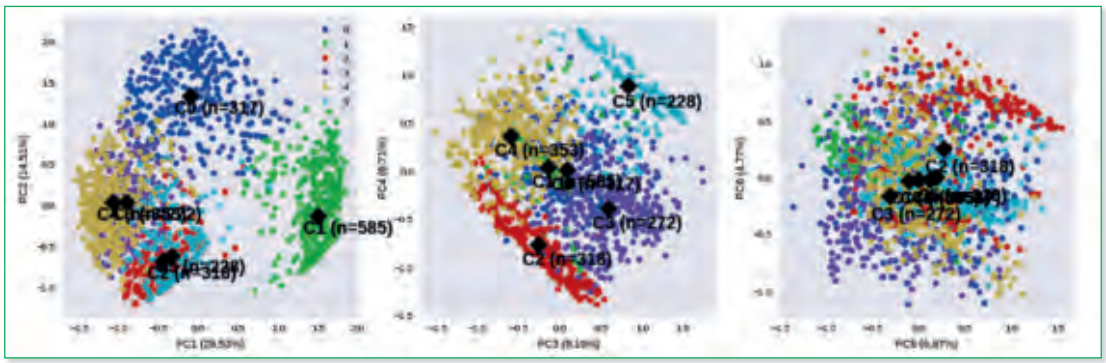


Figure 5 : Regroupement des individus suivant leurs composantes principales.
Figure 5: Clustered individuals positioning w/r to PCA components.

bien évidemment quand le nombre de groupes augmente. Pour garder un modèle simple et ne pas sur-modéliser, cette quantité doit rester aussi faible que possible. Le « coude » réfère à un point d'inflexion de la courbe d'explication de la variance marginale, laquelle est un indicateur synthétique dénotant l'adéquation entre un individu et sa classe de rattachement. Elle varie de 0 (classification totalement erronée) à 1 (classification parfaite).

Le premier critère suggère de former deux ou six groupes, alors que le second pointe plutôt vers six ou sept. Nous choisissons donc six.

Comment interpréter les six groupes choisis à partir du jeu de données bordelais ?

```
from sklearn.cluster import KMeans

model = KMeans(n_clusters=6, n_init=100,
max_iter=1000)
kmeans_ind = pca_ind.copy()
kmeans_ind['Xclust'] = model.fit_
predict(pca_ind.values)
kmeans_centroids = pd.DataFrame(model.
cluster_centers_,
      columns=pca_ind.columns)
kmeans_centroids['size'] = (kmeans_ind
     .groupby('Xclust')
     .count())['PC1']
round(kmeans_centroids, 2)
   PC1   PC2   PC3   PC4   PC5   PC6
size
0  1.51 -0.14 -0.14  0.03 -0.12 -0.03
585
1 -0.11  1.32  0.08  0.01  0.12 -0.02
317
2 -0.45 -0.68 -0.27 -0.76  0.26  0.25
318
3 -1.08  0.03 -0.60  0.37 -0.01 -0.02
353
4 -0.35 -0.62  0.84  0.87  0.18 -0.00
228
5 -0.90  0.03  0.59 -0.40 -0.32 -0.17
272
```

L'algorithme produit six groupes relativement équilibrés (le premier est un peu plus grand que les autres, mais pas énormément plus).

• Le groupe 0 représente les utilisateurs les plus polyvalents et les plus expérimentés. Il s'agit clairement des contributeurs OSM clef ;

• Le groupe 1 se réfère aux spécialistes des relations, qui sont assez productifs sur OSM ;

• Le groupe 2 regroupe des utilisateurs occasionnels, qui ne viennent que pour modifier des nœuds de temps en temps ;

• Le groupe 3 contient les anciens contributeurs ponctuels, plutôt intéressés dans la géométrie des chemins ;

• Le groupe 4 se rapproche du groupe 3, mais ses individus sont moins « anciens » ;

• Le groupe 5 contient les contributeurs locaux novices, qui font surtout des modifications de chemins.

Pour compléter ce tableau, nous pouvons tracer les individus selon leur groupe, en fonction des composantes les plus importantes.

Les deux premières composantes permettent de discriminer clairement entre 0 et 1. Nous avons besoin des deux composantes suivantes pour différentier entre 2/5 d'une part et 3/4 d'autre part. Les deux dernières composantes n'apportent pas d'information supplémentaire.

Cette classification a été effectuée sans connaissance préliminaire des utilisateurs ou de leur expertise. C'est l'avantage de la classification

How many clusters can be identified? We only have access to guidelines given by state-of-the-art procedures. As illustrated in figure 4, we use the elbow and silhouette method.

The former represents the intra-cluster variance, *i.e.* the sparsity of observations within clusters. It obviously decreases as the number of clusterincreases. To keep the model straightforward and not overfit it, this quantity must remain as small as possible. Hence the "elbow", that refers to an inflexion point in the explained variance marginal gain curve. "Explained variance" is a synthetic figure that indicates how well each individual fits within its cluster. It varies between 0 (bad clustering representation) and 1 (perfect clustering).

The first criterion suggests to choose either two or six clusters; the second criterion maxes out at six or seven clusters. We thus decide to use six clusters.

How to interpret the six chosen clusters in light of the Bordeaux area dataset?

```
from sklearn.cluster import KMeans

model = KMeans(n_clusters=6, n_init=100,
max_iter=1000)
kmeans_ind = pca_ind.copy()
kmeans_ind['Xclust'] = model.fit_
predict(pca_ind.values)
kmeans_centroids = pd.DataFrame(model.
cluster_centers_,
       columns=pca_ind.columns)
kmeans_centroids['size'] = (kmeans_ind
         .groupby('Xclust')
         .count())['PC1']
round(kmeans_centroids, 2)
   PC1   PC2   PC3   PC4   PC5   PC6
size
0  1.51 -0.14 -0.14  0.03 -0.12 -0.03
585
1 -0.11  1.32  0.08  0.01  0.12 -0.02
317
2 -0.45 -0.68 -0.27 -0.76  0.26  0.25
318
3 -1.08  0.03 -0.60  0.37 -0.01 -0.02
353
4 -0.35 -0.62  0.84  0.87  0.18 -0.00
228
5 -0.90  0.03  0.59 -0.40 -0.32 -0.17
272
```

The k-means algorithm makes six well-balanced groups (group 0 is larger than the others, however the difference is not so high):

• Group 0 (15.3% of users) gathers the most experienced and versatile users. The users are seen as *the* OSM key contributors;

• Group 1 (28.2% of users) contains relation specialists, users who are fairly productive;

• Group 2 (15.3% of users) corresponds to very unexperienced users, who come just a few times to modify nodes (mostly);

• Group 3 (13.2% of users) contains old one-off contributors, mainly interested in way modifications;

• Group 4 (17.0% of users) is very similar to the previous one, but for the contributors being less old;

• Group 5 (11.0% of users) identifies contributors who are locally unexperienced, and have mainly committed way modifications.

To complete this overview, we can plot individuals according to their group, with respect to the most important components:

The first two components allow to discriminate between cluster 0 and 1. We need the third and the fourth components to differentiate the pair of clusters 2, 5 from the pair 3, 4. The last two components do not provide any additional information.

This user classification has been carried out without any prior knowledge of identity or expertise. It demonstrates the power of unsupervised learning. We will now try to apply this clustering to OSM data quality assessment.

Data Quality Visualisation

Description of OSM element

What is an OSM feature? How to extract its associated metadata? As it turns out, in a similar way to what we have done before.

"Feature" is a generic term meaning either a node, a way or a relation. A feature is created during a changeset by a given contributor. It then can be modified several times, and possibly deleted. The OSM data history registers all operations carried out on all features.

```
elem          node
id       1669353159
version           1
n_user            1
last_uid     219843
```

automatique : nous allons l'appliquer à la qualité des données OSM dans la section suivante.

Visualisation de la qualité

Description d'un élément OSM

Qu'est-ce qu'un élément OSM ? Comment extraire les métadonnées qui lui sont associées ? La réponse est : assez facilement, par un processus similaire à celui utilisé pour les contributeurs.

Un élément est créé lors d'une session d'édition *(changeset)* par un contributeur donné, puis peut être modifié arbitrairement, voire détruit, par la suite. Cet élément est soit un nœud, soit un chemin, soit une relation. L'historique d'OSM contient toutes les opérations réalisées sur l'élément, par exemple ici un ancien nœud à versionnement unique.

```
elem           node
id       1669353159
version           1
n_user            1
last_uid     219843
```

Si nous supposons maintenant que le regroupement des utilisateurs permet de distinguer plusieurs classes d'expertise, à savoir : novices, avancés ou experts, chaque élément OSM peut avoir été édité par des contributeurs de n'importe quel groupe. Nous considérerons ici que la qualité de l'élément est bonne si le dernier contributeur à l'avoir modifié est expérimenté. Cela nous permet de classer les éléments à leur tour :

```
uid      4074141
PC1    -0.117667
PC2     1.145494
PC3     0.272941
PC4    -0.095759
PC5    -0.151579
PC6     0.932229
Xclust  2.000000
```

Le groupe le plus approprié pour l'élément est stocké dans la variable **Xclust** qui vient ensuite compléter le reste des métadonnées.

Les géométries ont été directement injectées dans une base de type *PostGIS* à l'aide de l'utilitaire **osm2pgsql**. Par exemple, si la base OSM s'appelle **osm**, appartient à l'utilisateur **user** et le fichier **pbf** correspondant s'appelle **bordeaux-metropole.osm.pbf** :

```
osm2pgsql -E 27572 -d osm -U user -p
bordeaux_metropole --hstore --extra-
attributes bordeaux-metropole.osm.pbf
```

Le SRID 27572 correspond à l'ancienne projection *Lambert Zone II* et l'option derrière **-p** permet de spécifier un préfixe au nom de tous les objets *point, polyline, polygon* et *roads*.

Nous allons nous intéresser, par exemple, au type *polyline*, qui contient, entre autres, à peu près toutes les routes (correspondant au type *ways* d'OSM), pour lequel nous souhaitons générer des métadonnées enrichies, en plus de la géométrie.

Nous allons commencer par créer une table **bordeaux_metropole_geomelements**qui contiendra nos métadonnées :

```
DROP TABLE IF EXISTS bordeaux_metropole_
elements;
CREATE TABLE bordeaux_metropole_
elements(
       id int,
       elem varchar,
       osm_id bigint,
       version int,
       n_users int,
       last_user int,
       last_user_group int
);
```

Puis nous la remplissons avec le fichier **csv** adéquat :

```
COPY bordeaux_metropole_elements
FROM '/home/rde/data/osm-history/output-
extracts/bordeaux-metropole/bordeaux-
metropole-element-metadata.csv'
WITH(FORMAT CSV, HEADER, QUOTE '"');
```

Et nous fusionnons ces métadonnées avec la géométrie issue de l'utilitaire **osm2pgsql** :

```
DROP TABLE IF EXISTS bordeaux_metropole_
geomelements;
SELECT l.osm_id, h.version, h.n_users,
h.last_user_group, l.way AS geom
INTO bordeaux_metropole_geomelements
FROM bordeaux_metropole_elements as h
INNER JOIN bordeaux_metropole_line as l
ON h.osm_id = l.osm_id AND h.version =
l.osm_version
WHERE l.highway IS NOT NULL AND h.elem =
'way'
ORDER BY l.osm_id;
```

Nous avons maintenant à disposition le groupe du dernier contributeur pour chaque élément, ce qui va nous permettre de générer des cartes de qualité.

Cartographie de la qualité des données

Vu que nous pouvons maintenant caractériser chaque élément, il est facile de dresser une carte qui nous indique quelles sont les données fiables

As an illustration we have above an old single-versioned node.

Let's recall we assumed that clustering the users allowsevaluating their trustworthiness as OSM contributors. They are either beginners, or intermediate users, or even OSM experts.

Each OSM entity may have been edited by one or many users of each group. Let's posit the entity quality is good if the last contributor was experienced. We can now tag the OSM entities themselves.

```
uid      4074141
PC1     -0.117667
PC2      1.145494
PC3      0.272941
PC4     -0.095759
PC5     -0.151579
PC6      0.932229
Xclust   2.000000
```

Each user's fitting cluster is saved into the column **Xclust**, which has to be tacked on to the other element metadata.

OSM element geometries have been decoded using **osm2pgsql**, another OSM data parser. Assuming the existence of an **osm** database owned by user**rde** and a file named **bordeaux-metropole.osm.pbf**:

```
osm2pgsql -E 27572 -d osm -U rde -p
bordeaux_metropole --hstore --extra-
attributes /home/rde/data/osm-history/
raw/bordeaux-metropole.osm.pbf
```

A French specific SRID (27572) and a prefix for naming output points, polylines, polygons and roads are specified as well.

Let's take the specific case of polylines, which represent the physical roads, among others (they roughly corresponds to OSM ways). We want to build an extended version of element metadata, with geometries.

First we can create the table **bordeaux_metropole_geomelements**, that will contain our metadata:

```
DROP TABLE IF EXISTS bordeaux_metropole_
elements;
CREATE TABLE bordeaux_metropole_
elements(
       id int,
       elem varchar,
       osm_id bigint,
       version int,
       n_users int,
       last_user int,
       last_user_group int
);
```

Then populate it with the accurate **csv** file:

```
COPY bordeaux_metropole_elements
FROM '/home/rde/data/osm-history/output-
extracts/bordeaux-metropole/bordeaux-
metropole-element-metadata.csv'
WITH(FORMAT CSV, HEADER, QUOTE '"');
```

And finally, merge the metadata with the geometry processed by **osm2pgsql**.

```
DROP TABLE IF EXISTS bordeaux_metropole_
geomelements;
SELECT l.osm_id, h.version, h.n_users,
h.last_user_group, l.way AS geom
INTO bordeaux_metropole_geomelements
FROM bordeaux_metropole_elements as h
INNER JOIN bordeaux_metropole_line as l
ON h.osm_id = l.osm_id AND h.version =
l.osm_version
WHERE l.highway IS NOT NULL AND h.elem =
'way'
ORDER BY l.osm_id;
```

From now on, we can use the last contributor cluster as an additional information to generate maps, so as to study data quality.

Quality Assessing Through Map Production

Since each OSM entity can be qualified, we can draw thematic maps of element quality using different colours according to their level of reliability. In this section we will draw maps of the Bordeaux area using *QGis*.

To begin with, the figure **bm_users** shows each OSM road w/r to a simple feature, here the number of users who have edited it. We see that the "ring" around Bordeaux is the most worked-on part of the road network: more contributors participated in the way completion. Some major roads within the city center are in the same case.

A similar map may be drawn to represent users' classification, as in figure 7:

According to the clustering done in the previous section (watch out, as the legend entries have been shuffled during map design…), we can make additional hypotheses:

• Light-blue roads are fine, they correspond to the cluster of the most trustworthy contributors (91.4%);

et celles qui semblent plus douteuses. Nous allons continuer à travailler sur la zone de Bordeaux, et utiliser *QGis* pour réaliser nos cartes.

La première carte représente le nombre de contributeur par élément. Nous voyons que le périphérique bordelais est le plus concerné, de nombreux contributeurs ont œuvré à son dessin. Quelques artères majeures du centre-ville ont reçu le même traitement.

Représentons maintenant la qualité, c'est-à-dire le groupe auquel appartient le dernier contributeur, par élément. Nous obtenons la carte de la figure 7.

Que pouvons-nous déduire de cette carte ?

• Les routes bleu clair sont correctes, elles correspondent aux utilisateurs les plus fiables (91,4 % de celles-ci) ;

• Il n'y a aucune contribution des utilisateurs novices ! Ce fait, rassurant en lui-même, tient peut-être au fait qu'aucun novice ne s'intéresse aux routes, mais plus probablement au fait que chacune des contributions est promptement corrigée par un « gradé » ;

• Le reste des contributions est issu d'utilisateurs intermédiaires, et leur qualité doit être jugée au cas par cas. Nous pensons qu'il n'y a pas de problème spécifique, même si des tendances locales fortes apparaissent, ce qui ne signifie pas un problème de qualité en lui-même.

Conclusion

Dans cette seconde partie nous avons détaillé une méthodologie pour générer des métadonnées sur chaque contributeur. Nous avons ensuite alimenté un algorithme de *machine learning* pour classifier ces contributeurs, après avoir réduit la dimensionnalité des variables explicatives grâce à une analyse en composantes principales. Ce processus permet de grouper les utilisateurs en classes homogènes, sans connaissance préalable des contributeurs ou de leurs habitudes.

Enfin, nous avons caractérisé la qualité des données OSM en utilisant ces groupes d'utilisateurs : pour chaque élément, nous avons considéré que sa qualité était directement liée à l'expertise du dernier contributeur l'ayant modifié.

Naturellement, il reste encore à améliorer ce processus, dont les grandes lignes ont néanmoins été jetées. Si vous êtes intéressés par ce sujet, n'hésitez pas à nous contacter. □

Figure 7 : « Chemins » OSM de Bordeaux, coloriés selon le groupe du dernier contributeur (1. expert en relations ; 2. experts polyvalents ; 3. nouveaux contributeurs ponctuels sur les chemins ; 4. anciens contributeurs ponctuels sur les chemins ; 5. contributeurs chemins non bordelais.)

Figure 7: OSM roads around Bordeaux, according to the last user cluster (1: C1, relation experts; 2: C0, versatile expert contributors; 3: C4, recent one-shot way contributors; 4: C3, old one-shot way contributors; 5: C5, locally-unexperienced way specialists).

Figure 6 : Nombre de contributeurs actifs par « chemin » OSM à Bordeaux.
Figure 6: Number of active contributors per OSM way in Bordeaux.

• There are no group0 roads… and that's comforting! It seems that "unreliable" users do not contribute to road or - more likely - that their contributions are quickly amended;

• Other contributions are made by intermediate users: a finer analysis should indicate if the corresponding features are valid. For now, we can consider everything is fine, even if local patterns seem strong. Areas of interest should be checked (they are not necessarily of low quality!)

Conclusion

In this second paper, we explained how to generate contributor metadata, i.e. information related to OSM users. We then fed this metadata into a machine learning framework. After reducing the dimensionality of the data using a Principle Component Analysis, we were able to squeeze the information into a small set of synthetic components, that we sorted into clusters of similar users. This workflow ran without any prior knowledge about users and their contribution patterns. OSM data quality was finally assessed using the previous user clusters, positing that the quality of a given feature was directly related to the "trustworthiness" of its latest editor.

Of course there is still work to be done, but the main path has been blazed. We hope you will be able to reproduce the proposed workflow, and to design your own maps.

Feel free to contact us if you are interested in this topic. ◘

References

Articles

• Garaud, D., Delhome, R., Mercier, H. 2017. A data-oriented framework to assess OSM data quality (part 1): data extraction and description. Geomatique Expert. 117, July 2017.

Websites

• Python Software Foundation. Python Language Reference, version 3.5. Available at http://www.python.org

• OpenStreetMap API: Available at http://www.openstreetmap.org

PostGIS 2.4

La nouvelle version majeure de votre cartouche spatiale préférée arrive incessamment sous peu ! La version 2.4 de *PostGIS* sera également la toute première à prendre en charge *PostGIS 10* qui, lui aussi, ne devrait pas tarder à pointer le bout de son nez. En contrepartie, *PostGIS 2.4* ne sera plus compatible avec les versions 9.2 et antérieures de *PostGreSQL*. Il est donc temps de migrer, si vous ne l'avez déjà fait.

Cette nouvelle version est également pleine de nouveautés. Citons pêle-mêle :

• Prise en charge des géométries curvilignes dans *ST_Reverse* ;
• *ST_Centroid* fonctionne sur le type *GEOGRAPHY* ;
• *ST_CurveToLine* convertit les courbes en lignes avec une tolérance maximale ;
• Prise en charge du format *Mapbox vector tile* en écriture *(ST_AsMVT)* ;
• Plus de fonctions éligibles au parallélisme ;
• Distance de Fréchet *(ST_FrechetDistance)* ;
• Prise en charge du catalogue de projections EPSG v.9

Cette nouvelle version devrait être disponible courant septembre.

Géocliché

Image GOES 6 de l'ombre de la Lune sur la surface terrestre lors de l'éclipse totale de soleil du 17 août dernier. Cliché NOAA

INTERGEO®

GLOBAL HUB OF THE
GEOSPATIAL COMMUNITY

BERLIN 2017

26 – 28 SEPTEMBER

GEOSPATIAL 4.0

OPEN GOVERNMENT

DIGITAL CONSTRUCTION

SMART CITIES

BE PART OF IT!
>>> WWW.INTERGEO.DE <<<

Along with:
GERMAN CARTOGRAPHIC CONFERENCE

DVW

Host: DVW e.V.
Conference organiser: DVW GmbH
Trade fair organiser: HINTE GmbH

SPONSORS: HEXAGON Trimble

COVADIS

Nouvelle version 16

Obtenez les outils BIM pour vos projets **d'aujourd'hui et de demain...**

GEOMEDIA SAS
Immeuble "La Vigie" - 20, quai Malbert - CS 42905 29229 BREST Cedex 2 - France
Tél. 02 98 46 38 39 - Fax 02 98 46 46 64
E-mail : contact@geo-media.com - Site Web : www.geo-media.com

AUTODESK
Gold Partner
Architecture, Engineering & Construction

NOUVELLE VAGUE