Insee
Mesurer pour comprendre

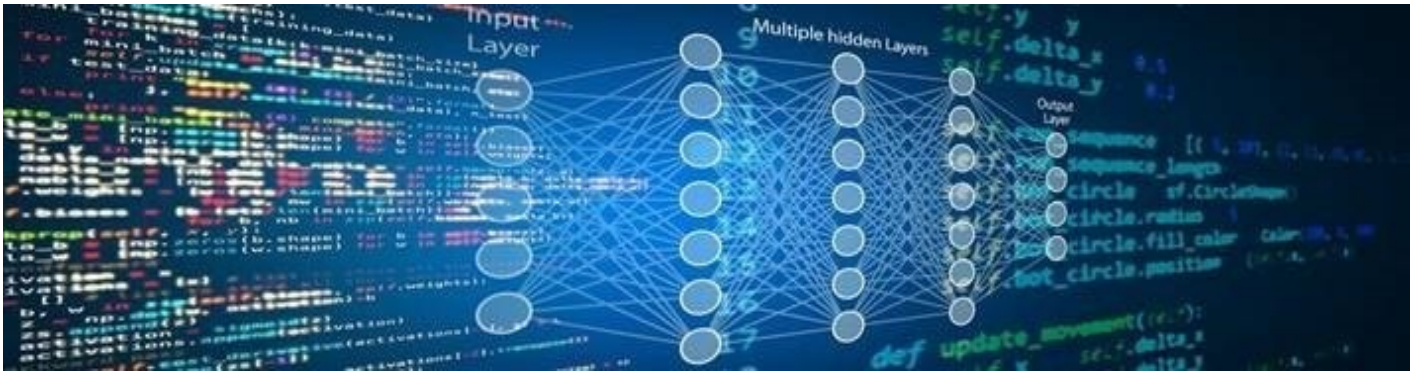
Big Data et Statistique Publique

Lettre d'information
N°12 – Mai 2022

Le SSP Hub, le réseau de *data science* du Service statistique public prend son envol

Le SSP hub est né d'une conviction : celle que le développement des compétences en data science constitue un enjeu stratégique pour l'Insee et les SSM et que l'animation d'une communauté ouverte, vivante dédiée à la data science peut utilement y contribuer. Encore fallait-il passer de l'idée à un projet concret, porteur de promesses attractives ....c'est l'offre que propose aujourd'hui le SSP Hub à tous ceux, experts, initiés, novices, curieux … intéressés par ce que la data science peut apporter à leur métier de statisticien public. Le réseau doit vivre pour tous et par tous. Gageons que l'essai sera transformé haut la main !

Danièle Bourlange, inspection générale, Insee



Le SSP Hub : origines et objectifs

Le SSP Hub – le réseau de *data science* de l'Insee et des services statistiques ministériels – déploie ses ailes progressivement. Son impulsion originelle remonte au travail de l'Inspection générale de l'Insee qui, dans la note N°2020_48/DG75-B001 "Les corps de l'Insee et les carrières de *data scientists*", préconise parmi les pistes à explorer la création d'un réseau de *data scientists* du Service Statistique Public (SSP). Une première pierre à l'édifice a été l'étude de la mise en place du réseau dont les esquisses sont synthétisées dans la note N°2020_86/DG75-B001. Enfin, sa préfiguration et le lancement de premiers canaux d'échange ont eu lieu grâce à une mission au SSP Lab amorcée fin 2021.

Le SSP Hub vise principalement deux objectifs :

• Faciliter l'échange entre pairs, tant en termes techniques que plus généralement sur les bonnes pratiques et nouvelles fonctionnalités de la *data science*. La discipline est en ébullition et voit naître de multiples *packages*, papiers académiques et outils, difficile de tout suivre tout seul !

• Promouvoir et démystifier la *data science*, acculturer un public non aguerri à cette thématique en montrant par l'exemple ses apports concrets mais aussi en redirigeant vers les ressources et formations disponibles pour son appréhension.

Le SPP Hub participe ainsi au développement et à l'amélioration des usages de *data science* au sein des services et directions participants. Ce réseau s'adresse aux agents du SSP, experts et débutants en *data science,* partageant leurs connaissances et pouvant s'entraider, aux non-initiés curieux de ces nouvelles techniques et désireux d'en apprendre plus, voire se former, et enfin aux encadrants souhaitant s'acculturer aux sujets en vue de favoriser l'utilisation de ces outils dans leurs services.

Bref, le SSP Hub doit permettre de faire « plus et mieux » de *data science* en étant composé d'un public aux compétences et attentes diverses. Tout cela est résumé dans le manifeste du SSP Hub accessible depuis le site du réseau.

Data science, science des données, de quoi parle-t-on ?

La *data science* est un domaine interdisciplinaire combinant des techniques issues des mathématiques, de la statistique et de l'informatique pour produire de la connaissance utile à partir de données. Le SSP Hub, a vocation à aborder plus spécifiquement les thématiques faisant partie de la boîte à outils de la *data science* et pouvant amener des évolutions positives dans le métier de statisticien public. Certaines de ces techniques et pratiques sont par ailleurs déjà utilisées par des agents du SSP ne se définissant pas nécessairement comme *data scientists* :

- Visualisation esthétique, automatisée et interactive ;
- Acquisition, appréhension de nouvelles sources données et exploitation de données non structurées : images, textes, données issues d'*open data*, d'API ou de *webscraping ;*
- Utilisation d'algorithmes de *machine* et *deep learning* avec des jeux de données de nature variée dans une optique de production statistique récurrente ou d'études. Ces algorithmes, fréquemment appliqués dans le cadre du traitement de langage naturel (*NLP*) ou d'exploitations d'images (*computer vision*), peuvent aussi servir à des fins de modélisation plus classique (imputation de valeurs manquantes, prédiction d'une variable...) ;
- Bonnes pratiques issues du monde informatique et mise en production : standards de qualité de code, gestion de versions et travail collaboratif (Git), reproductibilité d'environnement (*Docker*), outils de déploiement automatisé (CI/CD, Kubernetes) ;
- Gestion et manipulation de données massives : format des bases, calcul distribué ou parallélisé (Spark, Dask, Cuda).

Qui peut rejoindre le réseau et comment ?

Tout agent du service statistique public est invité à se joindre au réseau ! On le répète qu'il ou elle soit expert en *data science* ou débutant, technique, encadrant ou simplement intéressé à un volet de ce domaine, chacun a sa place dans le réseau. Et, en pratique pour rejoindre le réseau il suffit d'écrire à ssphub-contact@insee.fr pour être ajouté à la liste de diffusion de l'infolettre, au canal Tchap et être tenu au courant des évènements à venir.

Quels moyens d'action se donne le réseau ?

Pour atteindre ses objectifs et traiter des sujets mentionnés, le réseau s'est doté d'une panoplie de moyens d'action :

- Un **site web** agissant comme un *hub* de redirection en indexant les travaux menés au sein du SSP, les formations et bonnes pratiques à mutualiser. Sa version est encore préliminaire et sera améliorée au fur et à mesure des compléments et retours des uns et des autres : https://ssphub.netlify.app
- Une **infolettre mensuelle** qui dresse l'actualité du réseau (nouvelles formations, calendrier des événements de *data science*, ...). La dernière infolettre présente par exemple les membres référents du réseau, propose des liens à des formations (utilitR, MOOC de l'Inria en M*achine Learning*, formations Sésam) et partage des initiatives de *data science* : rendez-vous communautaires du SSP Cloud, sélection de sessions aux Journées de la méthodologie statistique 2022.
- Un ensemble d'**évènements** aux publics, durées et finalités différentes afin de répondre aux besoins variés : séminaires techniques/thématiques et échanges plus informels d'entraide/partage de bonnes informations (Open Hour de la Donnée,..)
- Un **canal Tchap** qui permet un échange plus direct et quotidien entre les participants au réseau. Ce canal diffuse aussi des contenus et des évènements n'ayant pas pu être communiqués dans l'infolettre

Les modalités pratiques se dessinent petit à petit. L'infolettre et le site web ont été lancés sur un mode expérimental dès février, de façon à recevoir des retours constructifs et à s'adapter en conséquence. Retours qui peuvent se faire via le canal Tchap.
Une première occurrence de "l'**Open Hour de la Donnée** " a eu lieu courant avril. Ce rendez-vous régulier sous Zoom a pour but de discuter informellement d'une thématique de *data science* après une courte présentation d'agents ayant mené des travaux en lien avec la thématique définie. Pour cette première Open Hour le sujet a été choisi pour intéresser le public le plus large possible "Non-réponse, imputation et *machine learning*". Le prochain thème sera choisi collégialement selon les votes des membres du réseau !

Qui se cache derrière ces contenus ?

Les outils et événements, conçus et mis en œuvre à la suite d'entretiens et échanges lors de la préfiguration, sont animés par Titouan Blaize au SSP Lab jusqu'à l'arrivée d'un futur coordonnateur à partir de septembre prochain. Un groupe de participants "référents" volontaires issus de SSM et de différentes directions de l'Insee aide à l'animation en orientant les thèmes abordés par le réseau et améliore le format des évènements. Il s'agit de Raphaele Adjerad (SSM DGFIP), Fabien Arnould (Insee - CSSL Metz), Romain Avouac (Insee - DIIT), Laura Gaimard (Insee - DRISS), Lino Galiana (Insee - DEE), Ronan Le Saout (SDES), Olivier Meslin (Insee - DEE).

Mais chacun est évidemment invité à contribuer en apportant des critiques constructives ou proposant des sujets, notamment via le canal Tchap ou en écrivant à ssphub-contact@insee.fr. Ces contributions doivent permettre de compléter le site web (sujets manquants, formations à ajouter, mise en page défaillante), choisir le thème d'un prochain évènement, etc..

Quel est le futur proche ?

Pratiquement, tous les moyens d'action du réseau cités précédemment ont été déployés (site web, infolettre,canal Tchap). C'est désormais aux évènements de se mettre petit à petit en place. L'Open Hour doit être pérennisée et éventuellement adaptée selon le déroulement de ses premières occurrences, la deuxième occurrence devrait d'ailleurs avoir lieu début juin. Des formats supplémentaires pourraient également voir prochainement le jour d'ici la fin de l'année : *masterclass* sur un thème précis, séminaires techniques pour présenter exhaustivement un travail de *data science*.

Une fois ces éléments mis en place et installés dans la durée, un événement d'officialisation, élargi à un public n'ayant pas forcément vocation à intégrer le SSP Hub, devrait avoir lieu vers la fin de l'année 2022. Le format et la date précise restent à définir par le futur coordinateur. À la fin de l'année 2022 aura également lieu une présentation en Comité du Programme Statistique restituant l'avancement du réseau et ses concrétisations.
De manière générale, le réseau évoluera au gré des réussites, des intérêts manifestés par les participants et des manques collectifs sur certains aspects de la *data science.* Coordinateur et référents sont à l'écoute de vos besoins, n'hésitez pas à écrire à ssphub-contact@insee.fr ou sur le canal Tchap !

Ressources/Actualités / Brèves

- **SAVE THE DATE : 2e édition du Funathon les 20 et 21 juin prochains**. Le "Funathon" est un événement de formation collaboratif organisé par le SSP Lab. L'idée est de proposer à tous une occasion de s'initier, se perfectionner, ou simplement de se confronter aux techniques de *data science*. **Une réunion d'information est prévue le 25 mai prochain à 14h30**, au lien Zoom suivant, réunion pendant laquelle le concept et les sujets seront présentés. En attendant, n'hésitez pas à regarder la session de présentation de la 1ère édition, la vidéo de restitution finale, ou simplement échanger sur le canal Slack dédié, voire à vous **inscrire** dès maintenant **auprès de votre référent formation**.
- Un nouveau **site internet pour le SSP Lab**. Pour faciliter l'accessibilité de ses travaux, le SSP Lab se dote d'un site internet. Vous y retrouverez aussi un blog alimenté par l'équipe avec leurs sources d'étonnement, leurs astuces, leurs résultats.
- "Introduction à la géomatique pour le statisticien : quelques concepts et outils innovants de gestion, traitement et diffusion de l'information spatiale", document de travail No M2022/01, série méthodologie, Insee. La **géomatique** c'est la combinaison de la **géographie** et de l'**informatique.** Ce DT présente les outils géomatiques récents permettant de stocker, traiter et diffuser l'information spatiale en R, Python et PostGIS, et réaliser des cartographies thématiques percutantes !
- Un tour des papiers présentés aux Journées de méthodologie statistique 2022 où beaucoup de présentations tiraient bénéfice de la datascience !
- Et toujours utilitR, documentation R destinée à tous sans pré-requis de niveau, les autoformations déployées sur le datalab-SSP Cloud, Spyrales : Soutien python et R entre agents de l'État.

Cette lettre est une occasion d'informer largement et d'échanger. N'hésitez pas à nous transmettre vos réactions et suggestions d'articles à ssplab@insee.fr.
Les archives de la lettre sont disponibles ici et là.
Demande d'inscription individuelle à la lettre : dg75-l001@insee.fr