



HAL
open science

4 enseignements d'un projet de médiation aux données et algorithmes : le cas d'un outil de suivi de l'artificialisation des sols.

Loup Cellard, Clément Mandron, Maëlle Fouquenot

► To cite this version:

Loup Cellard, Clément Mandron, Maëlle Fouquenot. 4 enseignements d'un projet de médiation aux données et algorithmes : le cas d'un outil de suivi de l'artificialisation des sols.. Dataactivist. 2024. hal-04847381

HAL Id: hal-04847381

<https://hal.science/hal-04847381v1>

Submitted on 5 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

4 enseignements d'un projet de médiation aux données et algorithmes : le cas d'un outil de suivi de l'artificialisation des sols. (1/2)

Questions relatives à la consommation d'ENAF (Espaces Natu-

Questions relatives à l'artificialisation des sols

📅 26 novembre 2024

✍️ Blog

algorithmes médiation documentation apprentissage
transparence



Loup Cellard



Clément Mandron



Maëlle Fouquet



LA* MEDNUM

Une note de recherche avec quatre enseignements pour une meilleure information, documentation et médiation à propos des algorithmes en général

Référence(s)

[Publication et médiation des algorithmes de Mon Diagnostic Artificialisation](#)

La médiation aux algorithmes : une approche pour pallier le manque de transparence algorithmique.

Dans le cadre de notre projet de recherche-action co-porté avec La Mednum et dédié à la médiation aux algorithmes, nous souhaitons documenter des cas d'études concrets et tirer des leçons de nos rencontres, expérimentations et missions en cours. Nous saisissons l'occasion d'une mission de conseil réalisée par Maëlle Fouquenet et Clément Mandron pour la start-up d'Etat Mon Diagnostic Artificialisation — un outil de d'analyse et de visualisation de la consommation et artificialisation des sols — pour formuler des enseignements de la manière dont on peut améliorer les dispositifs de médiation aux données et algorithmes.

Pendant quatre mois, Maëlle et Clément ont travaillé à décrire les données et calculs au cœur de cet outil. Loup, coordinateur du projet sur la médiation aux algorithmes, a suivi le projet, assisté à un atelier avec des usagers et participé à l'édition du livrable. Dans cet article fleuve que nous publions en deux parties, nous tirons quatre enseignements de cette mission que nous mettons en relation avec les problématiques plus vastes autour de l'utilisation des algorithmes par les pouvoirs publics, et la manière dont on peut mieux les comprendre grâce à un travail de médiation. Ces grands enseignements sont les suivants :

- Expliquer un algorithme c'est d'abord expliquer ses données (Partie 1/2)
- Expliquer un algorithme public c'est expliquer la loi (Partie 1/2)
- Expliquer un algorithme grâce à un format déjà connu des usagers facilite sa compréhension (Partie 2/2)
- Expliquer un algorithme est important même s'il ne prend pas des décisions qui impactent directement les personnes (Partie 2/2)

Avant de commencer à explorer les détails saillants de ces enseignements, donnons quelques éléments de contexte sur Mon Diagnostic Artificialisation et la mission réalisée par Dataactivist. La plateforme Mon Diagnostic Artificialisation permet à une commune de connaître et maîtriser sa consommation d'espace et d'artificialisation de ses sols. L'outil aide les collectivités à s'inscrire dans la trajectoire du Zéro Artificialisation Nette (ZAN) à horizon 2050 et la réduction de consommation d'espaces Naturels Agricoles Forestiers (NAF) à horizon 2031.

effective d'espaces urbanisés sur le territoire concerné", c'est-à-dire à la conversion d'espaces naturels, agricoles ou forestiers en espaces urbanisés (article 194 de la loi Climat et Résilience). L'artificialisation est "l'altération durable de tout ou partie des fonctions écologiques d'un sol, en particulier de ses fonctions biologiques, hydriques et climatiques, ainsi que de son potentiel agronomique par son occupation ou son usage" (article 192 de la loi Climat et résilience). En résumé, il s'agit des bâtiments, des routes, des parkings et des espaces en pelouse comme un stade de foot par exemple. Elle entre dans le calcul de la consommation d'espaces.

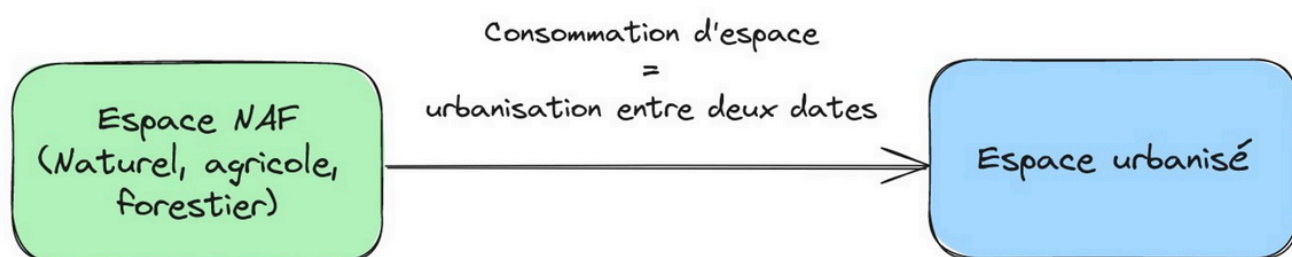


Illustration : Schéma qui explique ce qu'est un espace dit "urbanisé" (réalisé par Clément Mandron)

La plateforme permet de mesurer la consommation d'espace d'un territoire de 2011 à 2021, de simuler différentes trajectoires pour la décennie 2021-2031 sous forme de rapports, tableaux, graphiques ou cartes. Mon Diagnostic Artificialisation est un outil qui permet à deux ensemble d'acteurs de produire des rapports locaux : aux communes et intercommunalités dotées d'un document d'urbanisme, ainsi qu'aux services de l'Etat des communes couvertes par le règlement national d'urbanisme (RNU).

Différents enjeux écologiques se cachent derrière les problématiques de l'artificialisation des sols et notamment la capacité des élus à suivre son évolution et agir en conséquence. La biodiversité souffre de cette artificialisation des sols — des espaces qui abritent plus d'un quart de la biodiversité de la planète. L'érosion y est importante, le cycle de l'eau en est transformé, comme l'explique dans un récent article la Banque des Territoires :

Un sol artificialisé n'a plus les mêmes fonctions. Il ne permet plus à l'eau de s'infiltrer correctement (inondations), ni un rafraîchissement par la nature (îlots de

puisque le premier mètre de sol naturel renferme 2 à 3 fois plus de carbone que l'atmosphère et 3 à 7 fois plus que la végétation.

Pour que les élus prennent des décisions mesurées, il est important qu'ils comprennent les données et algorithmes qui nourrissent l'outil Mon Diagnostic Artificialisation.

Notre approche de la médiation a privilégié une documentation centrée sur l'utilisateur, des entretiens avec les producteurs de données (Cerema, IGN) et des analyses comparatives avec des startups d'État. En impliquant activement les utilisateurs par le biais d'ateliers et d'entretiens, notre objectif était d'intégrer la documentation dans la méthode de développement agile de Mon Diagnostic Artificialisation, favorisant des cycles courts et itératifs. Nous avons identifié les meilleures pratiques en matière de documentation, et proposé une refonte comprenant une [FAQ](#), un tutoriel et des évolutions sur la plateforme.

1. Expliquer un algorithme c'est d'abord expliquer ses données

À l'heure actuelle, et de manière quelque peu schématique, nous pouvons distinguer une première puis deuxième vague de mobilisation pour davantage de transparence et régulation des algorithmes [1]. La première a permis d'identifier les problèmes évidents de l'IA et des algorithmes, et a sensibilisé le public à ses biais et à ses limites. Il s'agissait de devenir vigilant et que cette vigilance se traduise dans des lois et la transformation des pratiques professionnelles. La seconde vague a contribué à ralentir suffisamment le déploiement de l'IA et de la robotique pour que les chercheurs, activistes et avocats de la première vague aient plus de temps et d'espace pour mettre en œuvre des réformes constructives et militer pour réparer les méfaits causés par les systèmes de calculs.

Dire qu'expliquer un algorithme consiste d'abord à expliquer ses données correspond au sursaut de la première vague qui souhaitait ainsi "ouvrir la boîte noire" : l'ouvrir pour y trouver des données — et donc des humains qui construisent et manipulent ces données. Cet angle d'attaque que nous avons adopté consiste à suivre une approche de l'analyse des algorithmes que l'on pourrait qualifier de généalogique ou génétique : c'est reconstruire ce qui

algorithmes sont une construction historique particulière informée par des pratiques, méthodes et métiers de l'analyse de données.

Comme analysé par le chercheur Valentin Goujon [2], la littérature en sciences sociales sur les données utilisées dans la recherche en IA peut être divisée en deux grands corpus qui se concentrent presque exclusivement sur les données d'apprentissage des modèles d'Intelligence Artificielle (IA). Le premier corpus se concentre principalement sur les étapes de collecte et d'annotation des données qui sont principalement réalisées, via des plateformes de crowdsourcing [3], par des travailleurs peu rémunérés et exposés à des contenus sensibles (sexisme, racisme, violence, désinformation, etc.). Dans ce corpus dédié au *digital labor*, l'accent est mis sur la partie amont du processus de développement des systèmes d'IA (c'est-à-dire les étapes de collecte et d'annotation) alors que la partie aval de ce même processus, leur évaluation comparative et leur déploiement « dans la nature », est plus rarement abordée [4].

Focus : Un modèle d'IA en cache un autre

LAION-5B est un très grand ensemble de données open-source d'images et de légendes extraites de l'internet, conçu pour les grands modèles d'intelligence artificielle. Il a été publié en 2022 par LAION, une organisation allemande à but non lucratif. LAION-5B est ce que nous appelons un « ensemble de données de base » pour l'intelligence artificielle générative. Midjourney et Stable Diffusion, deux grands modèles dont certaines sources de données sont connues, ont tous deux été formés en partie sur LAION-5B.

LAION-5B a lui-même été construit à partir d'un ensemble de données encore plus vaste, provenant d'une autre organisation à but non lucratif : Common Crawl. Un élément clé de la construction de LAION-5B a été d'essayer de sélectionner des images et des légendes de Common Crawl dont le texte de l'attribut ALT correspondait le mieux

Les chercheurs ont utilisé un modèle appelé CLIP (Contrastive Language-Image Pre-training), un réseau neuronal développé par des chercheurs d'Open AI.

Parce qu'ils doivent être si grands, la construction des modèles implique nécessairement l'utilisation d'autres modèles, qui ont eux-mêmes été formés sur des ensembles d'entraînement algorithmiques.

Pour aller plus loin: "[Models All The Way Down](#)". Christo Buschek & Jer Thorp. *A Knowing Machines Project*.

Au départ de notre mission, le besoin de la start-up d'Etat Mon Diagnostic Artificialisation était le suivant : rendre le plus compréhensible possible les données de consommations et d'artificialisation des sols, et répondre aux questions récurrentes. Ce gain de compréhension d'un système de calcul et d'un ensemble de données est bien sûr possible uniquement si nous avons accès à des données ouvertes.

Il a fallu comprendre la chaîne de traitement des données, ce qui a demandé la mobilisation des connaissances des différents chaînons. Notre rôle était d'améliorer la compréhension de ces informations par un travail de médiation afin d'expliquer d'où viennent les données qui sont utilisées par Mon Diagnostic Artificialisation, les traitements statistiques et algorithmiques qui leur sont appliqués ainsi que les calculs qui sont effectués par la startup.

DATAACTIVIST

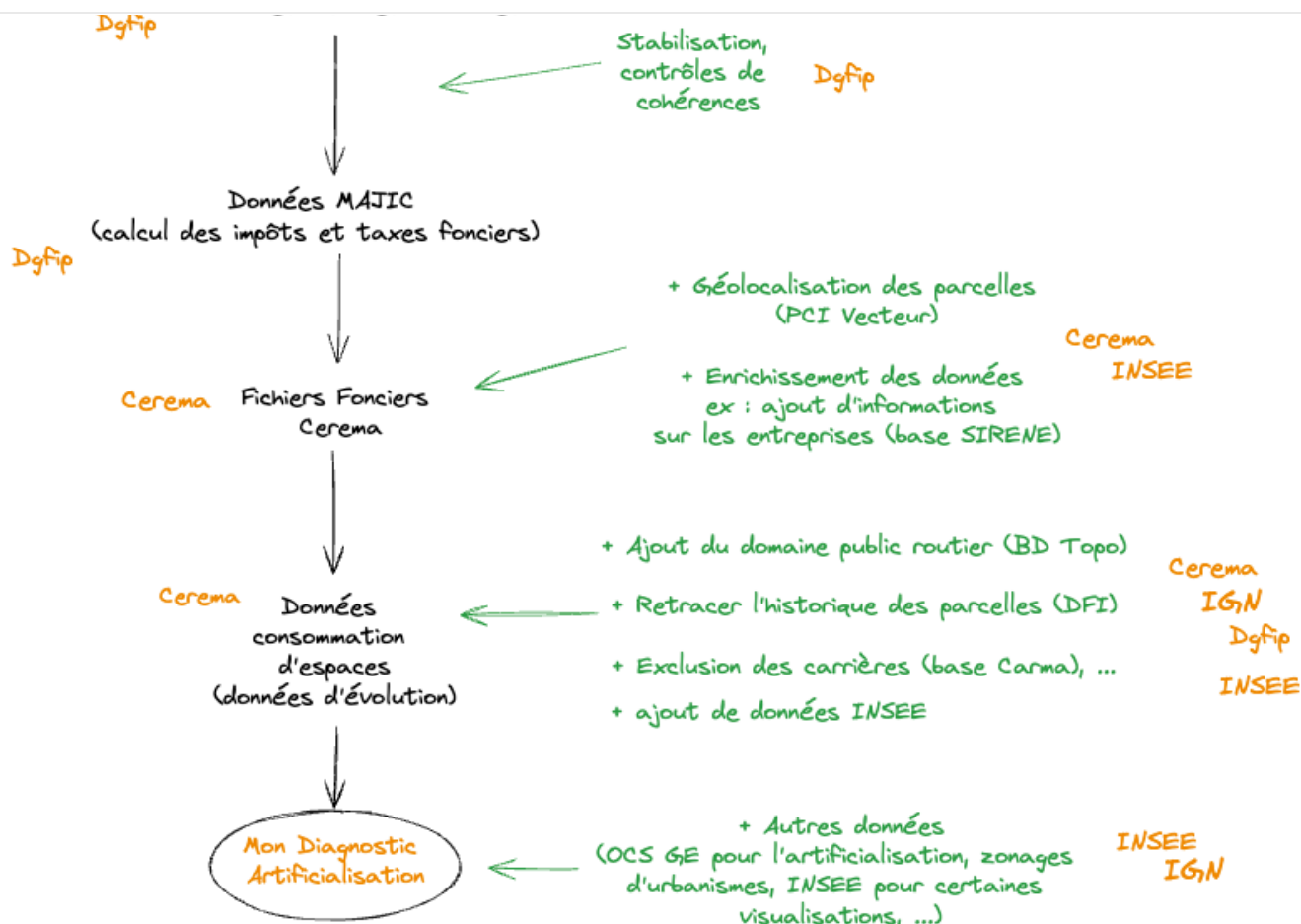


Illustration : Schéma de la chaîne de traitement des données de l'outil Mon Diagnostic Artificialisation (réalisé par Clément Mandron)

La consommation d'espaces NAF (naturel, agricole, forestier) par année et par déterminant, est fournie par le **CEREMA** — Le Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement, soit un expert technique public. Ces données proviennent de la DGFiP (Direction générale des Finances publiques) et sont le résultat d'un traitement des données issues du **système MAJIC** (Mise À Jour des Informations Cadastrales, qui permet de calculer et gérer l'impôt foncier).

En les fusionnant avec d'autres types de données (comme les données de la base SIRENE de l'INSEE qui contient la liste des établissements), le CEREMA fait notamment le lien entre une parcelle et un/des occupants. Il transforme ensuite ces informations en données de consommation d'espaces par millésime.

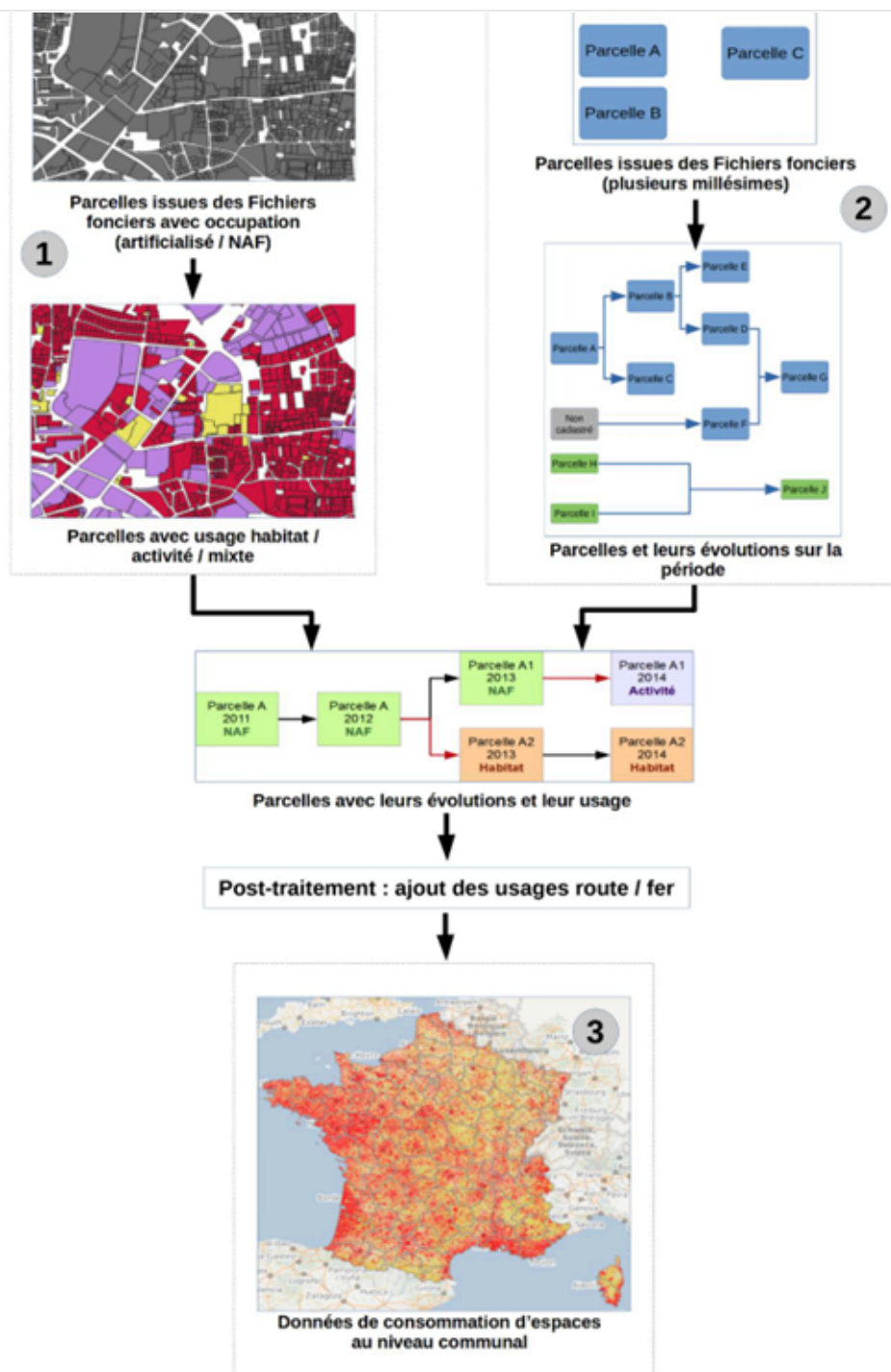


Illustration : Résumé de la méthode d'évaluation de la consommation d'espaces à partir des Fichiers fonciers. La partie en haut à gauche concerne l'usage (1), en haut à droite le multi-millésime (2) et le centre le traitement (3). Source : CEREMA

Mon Diagnostic Artificialisation utilise également les données démographiques de l'INSEE pour comparer la consommation d'espace et l'évolution de la population sur un territoire donné.

Ensuite, en ce qui concerne l'artificialisation : les données de l'occupation des sols à grande échelle (OCS GE) "nouvelles générations" sont fournies par l'IGN. Ces

des objets (bâti, arbre, surface de bitume...) et en délimiter les contours.

Le fait qu'il y ait recours à un modèle d'IA dans la chaîne de production des données doit attirer notre attention : les données qui (dans une approche génétique/généalogique de l'étude des algorithmes) pourraient être prises comme la source d'un outil de projection et visualisation comme Mon Diagnostic Artificialisation proviennent en réalité d'une infrastructure de calcul qui la précède. Données et algorithmes se révèlent donc indissociables.

En d'autres termes, **derrière les algorithmes il y a d'autres données et d'autres algorithmes** — un constat qui s'applique particulièrement au contexte de l'IA Générative, comme l'illustre l'encart ci-dessous.

Si l'on prend l'exemple du diagnostic de la métropole Aix-Marseille, la trajectoire de consommation pour 2021-2030 (+2 076,6 ha) est calculée en divisant par deux le bilan de consommation des dix dernières années (+4 153,2 ha/2). Une division par deux ou une moyenne par années sont aisément compréhensibles par nombre d'utilisateurs, ce qui n'est pas le cas du processus de production et traitement des données qui demandent à la fois une compréhension technique et institutionnelle du sujet.

Complétons donc notre premier enseignement de la sorte : **expliquer un algorithme c'est d'abord expliquer ses données. Mais de manière contre-intuitive, ce n'est pas parce que les calculs de l'algorithme présentés à l'utilisateur sont simples qu'une chaîne de production des données sinueuse et sous-jacente ne doit pas faire l'objet d'une médiation.**

2. Expliquer un algorithme public c'est expliquer la loi

Bien souvent, si ce n'est pas quasi exclusivement, les algorithmes utilisés par les pouvoirs publics mettent en œuvre des règles de droit, qui se trouvent précisées dans des lois puis des codes, décrets, circulaires, et autres formes de doctrines juridiques. La loi n'est pas toujours limpide et cette traduction de celle-ci en code nécessite une interprétation et des ajustements techniques qui peuvent contredire l'esprit de la loi [5]. En ce qui concerne les algorithmes publics, l'équivalence entre code et loi est loin d'être évidente, ce qui se pose différemment dans le cadre des algorithmes des GAFAM [6].

enracinées dans les arts mécaniques de l'écriture et matérialisées dans des formats tels que le droit des brevets, les livres pratiques, les recettes et les livres de cuisine [7]. Cette idée des algorithmes comme des ensembles d'instructions ou de recettes correspond à leur conception prémoderne.

Plus généralement, la compréhension pré-moderne d'un algorithme découle de l'arithmétisation de la logique, et des nombreuses tentatives de formalisation mathématique des jugements. Par exemple, dans un travail antérieur, Daston a fait valoir que la quantification poursuivie par les premiers probabilistes mathématiques était profondément enracinée dans la formalisation des concepts juridiques [8]. Elle a récemment repris l'exemple du calcul des « degrés d'incertitude » pour historiciser nos tentatives contemporaines de quantification du jugement. Pour elle, il s'agit d'«une doctrine très ridiculisée de l'arithmétique de la preuve, fortement développée dans la pensée juridique et occasionnelle du XVI^e siècle, dans laquelle le récit de différents témoins et différents types de preuves se voient attribuer des poids différents» [9]. L'histoire racontée par Daston nous apprend que les algorithmes sont des figures hybrides qui ont pu être formalisées grâce à l'héritage commun des formalismes juridiques et mathématiques. Leur essentialisation en tant qu'entités purement mécaniques ou technologiques est beaucoup plus récente et continue d'être contestée. Par exemple, la Commission d'Accès aux Documents Administratifs, le régulateur qui défend notre droit à l'information des algorithmes publics, ne considère pas que les algorithmes doivent être informatisés pour être qualifiés comme tels [10].

Ce qui est frappant après la découverte rapide de l'outil Mon Diagnostic Artificialisation est l'avalanche de rappels à la loi. Dès le bas de la page d'accueil il est rappelé l'article 192 de la loi Climat et Résilience qui fixe une trajectoire nationale de sobriété foncière en 2 étapes :

- D'ici 2031 : diminution d'environ 50% du rythme de consommation d'espaces naturels, agricoles, et forestiers par rapport au bilan des 10 dernières années (2011-2021)
- D'ici 2050 : Zéro Artificialisation Nette grâce à l'équilibre entre le total des surfaces artificialisées et celles renaturées

Grâce à de l'information et des visualisations, la raison d'être de l'outil est d'aider et d'orienter les acteurs publics à ce que ces éléments légaux soient traduits dans

- Un diagnostic de consommation d'espaces NAF sur la période 2011-2021, recommandé pour estimer l'objectif d'ici 2031
- Un rapport triennal de l'artificialisation des sols, obligatoire pour suivre la réduction progressive des surfaces artificialisées
- Un pilotage régulier, recommandé pour s'inscrire dans la trajectoire de la loi Climat et Résilience

Pour se conformer à la loi, ces diagnostics, rapports et éléments de suivi vont se trouver remplis de données et visualisations issus de Mon Diagnostic Artificialisation ou provenant des services des collectivités. Cependant, lors de nos ateliers collaboratifs avec les parties prenantes et usagers du service, nous avons compris que pour être juste et précise cette mise en données nationale et les critères de la loi qui l'accompagne doit s'accompagner d'un dialogue avec les acteurs de terrain. Ils nous expliquent par exemple que le seuil de densité pour qualifier un ensemble d'arbres comme d'une "forêt" a beau être précis et exhaustif dans la loi, les images satellites (construites pour partie avec de l'IA) ne permettent pas toujours cette qualification. En définitive, la loi est toujours interprétable et les images restent parfois ambiguës, il serait alors nécessaire de retourner sur le terrain pour constater les choses.

Comme l'ont montré les chercheuses Pauline Gourlet et Maud Barret Bertelloni à propos du projet Fonciers Innovants de la DGFIP, le recours aux algorithmes d'analyse d'images pour la segmentation et l'analyse du territoire a été contesté en interne à la DGFIP par les géomètres qui craignent de voir leur métiers peu à peu disparaître. Dans notre cas, rappelons que les données des cartes OCSGE (porté par l'État via l'IGN et ses projets d'IA) sont la référence pour de nombreuses bases de calculs (y compris celle de Mon Diagnostic Artificialisation). Afin de garder une unité nationale et comparer les territoires entre eux, la volonté de l'État serait que ce type de cartes prime sur celles produites par les bureaux d'études locaux : un désir de standardisation qui pose le statut des cartes produites avec de l'IA et des données comme source de l'autorité administrative.

Focus : Expliquer un algorithme à partir de ses règles de droit : est-ce

Un groupe interdisciplinaire de chercheurs et chercheuses ont tenté l'exercice d'expliquer un algorithme à partir de ses règles de droit sur le cas de l'allocation au logement. Bien que cette explication par l'analyse du droit s'avère individualisée à la situation de l'intéressé et détaillée, la question de son intelligibilité reste discutable. Le cas pratique est écrit dans le langage du droit pendant 8 pages, ce qui pose deux défis majeurs. D'abord, les textes juridiques eux-mêmes définissent des calculs et provisions complexes avec une rédaction souvent alambiquée. Ensuite, la lecture d'un cas pratique nécessite des capacités analytiques normalement apprises durant la formation juridique.

Dans ce contexte, il y a un réel besoin de médiation entre les règles de droit et leurs traitements algorithmiques. Les chercheurs·euses concluent que cette explication serait plus adaptée si effectuée par des travailleurs sociaux, associations et ONG qui savent naviguer dans les décisions de justice et accompagnent des allocataires affectés par les résultats des algorithmes de la CAF. Un tel travail est actuellement mené par Soizic Pénicaud et Valérie Pras au sein de l'association Changer de Cap, en coordination avec d'autres associations (La Quadrature du Net, ATD Quart Monde, la Ligue des droits de l'Homme) impliquées dans la lutte contre le contrôle social algorithmique.

Pour aller plus loin : Denis Merigoux, Marie Alauzen, Justine Banuls, Louis Gesbert, Émile Rolley. "De la transparence à l'explicabilité automatisée des algorithmes : comprendre les obstacles informatiques, juridiques et organisationnels". RR-9535, INRIA Paris. 2024.

compliquent si la loi a besoin d'être à la fois interprétée, précisée par des personnes et mise à l'épreuve de l'observation de terrain.

Dans la deuxième partie de l'article nous aborderons en détail les deux enseignements suivants :

- Expliquer un algorithme grâce à un format déjà connu des usagers facilite sa compréhension (Partie 2/2)
- Expliquer un algorithme est important même s'il ne prend pas des décisions qui impactent directement les personnes (Partie 2/2)

Stay tuned !

Notes et références

N°	Référence
1	Voir le texte de Frank Pasquale : https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability
2	Valentin Goujon, "Halving the AI black box: towards benchmark datasets as new obligatory passage points for AI studies", <i>Politics of Machine Learning Evaluation Workshop</i> , 16-17/11/2023. University of Amsterdam.
3	Miceli, M., Schuessler, M., & Yang, T. (2020). Between subjectivity and imposition: Power dynamics in data annotation for computer vision. <i>Proceedings of the ACM on Human-Computer Interaction</i> , 4(CSCW2), 1-25. https://dl.acm.org/doi/abs/10.1145/3415186 Posada, J. (2022). Embedded reproduction in platform data work. <i>Information, Communication & Society</i> , 25(6), 816-834.
4	Voir le dossier à venir sur la politique de l'évaluation du ML dans la revue Digital Societies : https://link.springer.com/collections/bbbehaijbcj
5	Denis Merigoux, Marie Alauzen, Justine Banuls, Louis Gesbert, Émile Rolley. De la transparence à l'explicabilité automatisée des algorithmes :

	comprendre les obstacles informatiques, juridiques et organisationnels. RR-9535, INRIA Paris. 2024, pp.68. <hal-04391612>
6	Jiménez González, A. (2022). Law, Code and Exploitation: How Corporations Regulate the Working Conditions of the Digital Proletariat. <i>Critical Sociology</i> , 48(2), 361-373. https://doi.org/10.1177/08969205211028964
7	Lauren Daston, Rules: A Short History of What We Live By. Lorraine Daston. <i>Princeton University Press</i> , 2022.
8	Daston L (1988) Classical Probability in the Enlightenment. Princeton, N.J: Princeton University Press
9	Gross J (2020) Historicizing the self-evident: an interview with Lorraine Daston. Available at: https://lareviewofbooks.org/article/historicizing-the-self-evident-an-interview-with-lorrainedaston/ (accessed 23 October 2020).
10	Loup Cellard, « Algorithms as Figures. Towards a post-digital ethnography of algorithmic contexts »._ <i>New Media & Society</i> _, 2022.



Nos partenaires

Ressources externes