

L'inspection générale des finances et la science des données

Quelles méthodes pour quels usages ?



Quentin Bolliet*, Aymeric Floyrac**, Sophie Maillard***
et Agathe Rosenzweig****

Le décideur montre un intérêt croissant pour la prise de décision basée sur les données. L'administration, de son côté, dispose d'une grande quantité de données, d'infrastructures et de ressources humaines capables de les exploiter. Ces développements récents permettent aux inspections, comme l'inspection générale des finances (IGF), de renforcer leurs missions d'évaluation et de conseil par des études spécifiques mobilisant des données déjà disponibles. L'IGF utilise celles-ci pour quantifier les constats et consolider les analyses. À la différence du système statistique public, cependant, elle ne produit pas de données et n'en assure pas le suivi dans le temps.

À titre d'exemple, une mission réalisée par le pôle science des données de l'IGF sur l'assurabilité des biens des collectivités territoriales est présentée. Elle illustre comment une utilisation large des données de l'Institut national de l'information géographique et forestière (IGN) et de la Direction générale des Finances publiques (DGFIP) permet, grâce aux techniques de *data science*, d'éclairer de manière très fine un questionnement adressé à l'IGF. Si la demande pour des études sur mesure augmente et que le modèle d'un pôle dédié à la science des données essaime dans les inspections, il comporte des conditions de réussite, tant sur la culture de l'analyse quantitative que sur la qualité des données.

 Decision-makers are showing a growing interest for data-driven decision making. At the same time, the administration has access to vast amounts of data, infrastructures and skilled human resources to leverage them. These recent developments allow inspectorates, such as the General Inspectorate of Finances (IGF), to strengthen their evaluation and advisory missions through specific studies using already available data. The IGF uses this data to quantify its observations and strengthen its analyses. Unlike the official statistics system, however, it neither produces data nor monitors it over time.

One example is a study conducted by the IGF's data science team about the insurability of municipalities' real estate assets. It demonstrates how combining data from the National Institute of Geographic and Forestry Information (IGN) and the Directorate General of Public Finances (DGFIP) with data science techniques can provide highly detailed insights to address policy questions submitted to the IGF.

Although demand for tailored studies is increasing and other inspectorates have been replicating the model of a data science dedicated hub, this model needs to fulfil conditions relating both to the culture of quantitative analysis and the quality of the data.

* Data scientist, inspection générale des finances.
quentin.bolliet@igf.finances.gouv.fr

** Responsable adjoint du pôle science des données, inspection générale des finances.
aymeric.floyrac@igf.finances.gouv.fr

*** Responsable du pôle science des données, inspection générale des finances.
sophie.maillard@igf.finances.gouv.fr

**** Data scientist, inspection générale des finances.
agathe.rosenzweig@igf.finances.gouv.fr



L'inspection générale des finances (IGF), forte de sa tradition d'objectivation et d'évaluation des politiques publiques, s'est dotée en 2019 d'un pôle science des données, qui compte aujourd'hui une dizaine d'agents.



La prise de décision, qu'elle soit le fait des administrations, des entreprises ou des individus, repose de plus en plus sur l'exploitation des données. En parallèle, la quantité de données disponibles a explosé au cours des deux dernières décennies, tandis que les techniques de traitement, tant statistiques qu'informatiques, se sont démocratisées et sont aujourd'hui bien plus accessibles qu'auparavant. Cette évolution a été soutenue par le développement d'infrastructures robustes, capables de gérer des traitements parfois massifs : les plateformes de *data science*.

Sous réserve d'investir dans les ressources matérielles et surtout humaines nécessaires, ces transformations ouvrent la possibilité aux acteurs publics et privés de réaliser des analyses quantitatives mieux ciblées sur leurs sujets d'intérêt et questionnements spécifiques. En particulier, elles offrent aux acteurs publics une opportunité unique d'élaborer des politiques mieux informées, fondées sur des preuves tangibles, afin de répondre aux besoins des citoyens de manière plus efficace et équitable.

L'inspection générale des finances (IGF), forte de sa tradition d'objectivation et d'évaluation des politiques publiques (Prada, 2012), s'est ainsi dotée en 2019 d'un pôle science des données, qui compte aujourd'hui une dizaine d'agents, capable de mener à bien des travaux à forte composante quantitative au service de ses missions d'évaluation et de conseil.

► La science des données à l'IGF : une réponse moderne à un besoin ancien

Les institutions publiques s'appuient depuis longtemps sur le système statistique public (SSP), qui offre des productions générales de haute qualité pour éclairer les politiques publiques. Cependant, ces productions ne permettent pas de répondre à toutes les questions que se pose le décideur : les données existantes, souvent conçues pour des objectifs larges, peuvent ne pas répondre précisément aux besoins spécifiques et parfois étroits d'une mission d'évaluation ou de conseil. Par ailleurs, les délais imposés à la décision publique – et par conséquent à certaines missions, souvent de quelques mois – rendent impossible la création de nouvelles données adaptées.

Depuis la loi pour une République numérique de 2016¹ et le rapport Bothorel (Bothorel, 2020), l'augmentation du volume de données administratives disponibles a ouvert de nouvelles perspectives pour une exploitation ad hoc. Néanmoins, pour en tirer pleinement parti, les institutions doivent disposer d'équipes dédiées et compétentes en analyse de données. C'est la raison d'être du pôle science des données de l'IGF, comme des pôles équivalents créés dans d'autres institutions, notamment à la Cour des comptes et dans

¹ Voir les références juridiques en fin d'article.

d'autres inspections générales (voir dernière partie). Ces pôles travaillent en interaction permanente avec des profils généralistes (que sont par exemple les inspecteurs des finances), pour répondre à des questions variées : objectiver une situation à partir de données existantes, évaluer les effets de politiques publiques, ou simuler les impacts potentiels d'une réforme. On peut par exemple citer la mission sur les ressources humaines de l'État dans le numérique, le comité d'évaluation du plan France Relance ou encore la mission sur le coût du travail en France, portée par Antoine Bozio et Étienne Wasmer.

Mobiliser les données pour répondre à des questions de politique publique

Les pôles science des données de l'IGF et des autres instances d'évaluation sont composés de statisticiens et de *data scientists* (Comte et al., 2022). Ils se distinguent de par leur rôle des structures du système statistique public. Contrairement à l'Insee et aux services statistiques ministériels, ils n'ont pas vocation à produire des statistiques générales, périodiques et comparables. Ainsi, ils ne participent pas à « la production statistique d'informations économiques ou sociales », domaine qui délimite selon Michel Volle la « statistique » telle qu'entendue dans son acception courante (Volle, 1984). Ils se concentrent sur des analyses ponctuelles et ciblées, adaptées aux besoins à court terme des décideurs. À ce titre, leur hiérarchie des valeurs diffère de celle du système statistique public :

1. Réactivité : contraints par les délais des missions, les pôles doivent souvent adapter au temps qui leur est imparti le degré de finesse de leurs analyses : un résultat produit hors délai ne sera jamais communiqué au commanditaire de la mission.

2. Adaptabilité : ces pôles travaillent avec des données issues de multiples sources administratives, souvent hétérogènes. Cela nécessite des infrastructures partagées et sécurisées pour simplifier l'accès aux données et garantir leur confidentialité, tout en réduisant le temps consacré à leur préparation.

3. Transparence et comparabilité : ces pôles n'ont pas pour objectif de produire des données publiques ou comparables entre pays, contrairement à la statistique publique (de Peretti et Touchelay, 2024). Leur priorité est l'aide à la décision, souvent dans un cadre confidentiel. Ainsi, dans le cas de l'IGF, il appartient au commanditaire de la mission de décider du caractère public ou non du rapport².

Les analyses produites par les pôles de science des données apportent une réponse moderne à un besoin ancien : produire des analyses sur mesure, rapidement mobilisables, pour répondre aux défis immédiats des politiques publiques.

Ces nouvelles pratiques rappellent les origines mêmes de la statistique, pensée initialement comme un outil au service de l'État pour éclairer et guider l'action publique. Elles répondent au besoin très ancien des « pratiques de mise en chiffre », dont les objectifs sont « de coordonner les activités humaines, de trouver des accords entre parties prenantes, de prendre et de justifier des décisions fiscales, militaires ou sociales » (Martin, 2023). Ainsi, à l'image des enquêtes

² Le commanditaire de la mission est généralement un membre du gouvernement, le plus souvent le Premier ministre ou le ministre de l'Économie et des Finances.

ponctuelles de Colbert ou des dénombrements de Vauban, les analyses produites par les pôles de science des données s'inscrivent dans une tradition d'investigation spécifique et orientée.

Cependant, ces nouvelles pratiques s'écartent des standards actuels de la statistique publique, qui met particulièrement en avant les valeurs de transparence et de comparabilité. En contrepartie, elles apportent une réponse moderne à un besoin ancien : produire des analyses sur mesure, rapidement mobilisables, pour répondre aux défis immédiats des politiques publiques (*figure 1*).

Exploiter une variété de sources issues, entre autres, de la statistique publique

Le pôle science des données de l'IGF bénéficie pour la conduite de ses travaux du mouvement général de décloisonnement des données publiques et de la diffusion des outils permettant de les exploiter. Les questions spécifiques traitées par les missions le conduisent à manipuler une grande diversité de données :

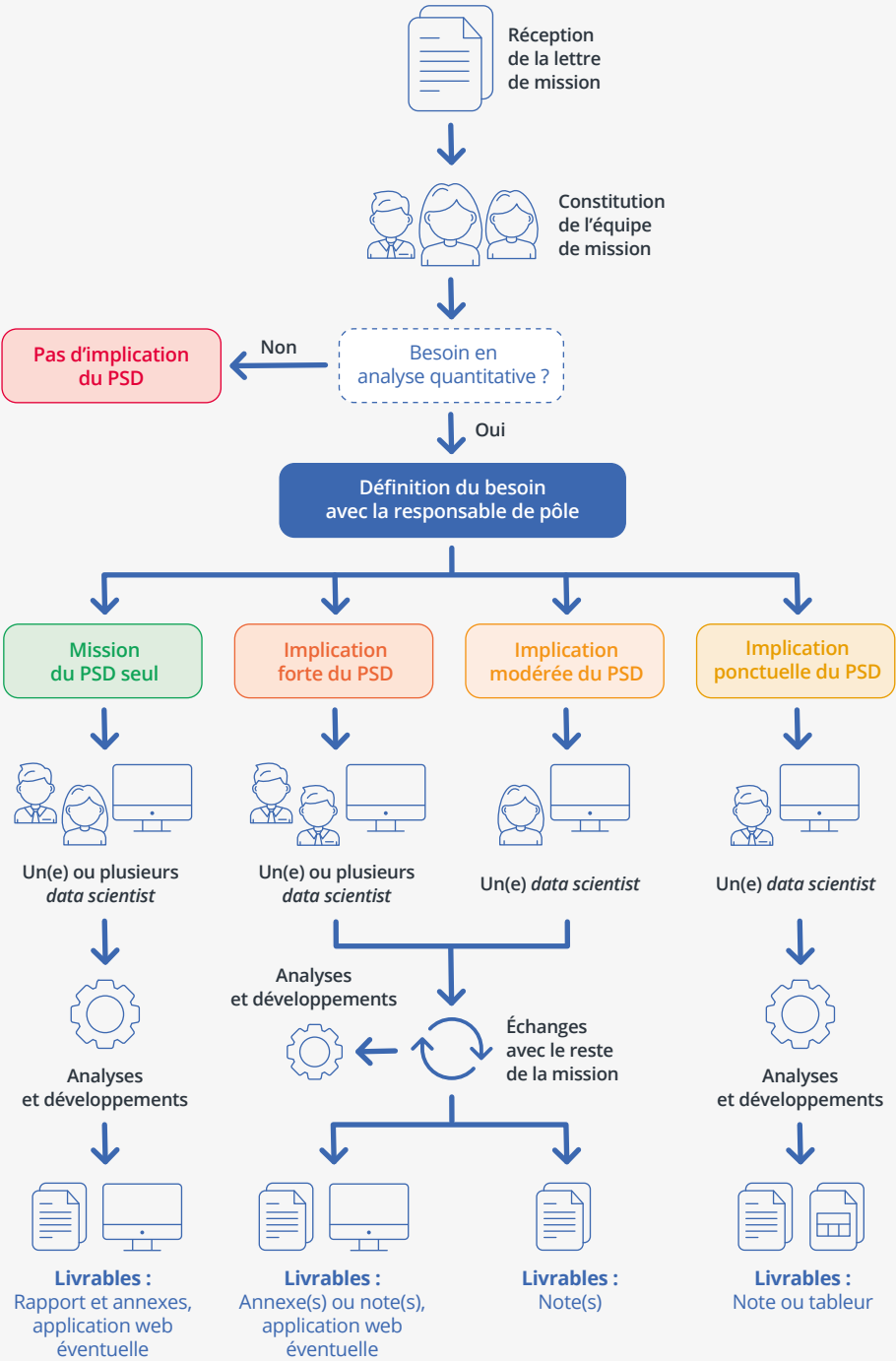
- **des données agrégées produites par la statistique publique** : comme tout citoyen, les membres du pôle ont accès à ces données dont la qualité est garantie ;
- **des données individuelles produites par la statistique publique** : les membres du pôle sont habilités par le comité du secret statistique à accéder à un large catalogue de ces sources via le centre d'accès sécurisé aux données (CASD) [Gadouche, 2019], dont notamment les données d'imposition des individus (fichier POTE³) et des entreprises (fichiers FARE⁴ et BIC-IS⁵), ainsi que celles issues des déclarations sociales nominatives (DSN) des employeurs. Ces accès s'accompagnent d'obligations strictes en matière de respect du secret statistique, avec une distinction claire entre les prérogatives des statisticiens du pôle, habilités à utiliser ces données, et celles des inspecteurs des finances, qui ne peuvent accéder qu'aux résultats agrégés des analyses ;
- **des données d'autres sources que celles de la statistique publique** : au gré des thématiques abordées par les missions, les membres du pôle peuvent aussi mobiliser, au-delà de cet ensemble déjà riche, d'autres données souvent moins structurées, ou se prêtant moins naturellement à des traitements statistiques :
 - des données textuelles, par exemple issues des bulletins officiels ;
 - des données géographiques, notamment produites par l'Institut national de l'information géographique et forestière (IGN), comme la BD Forêt ou la BD TOPO, mais aussi les données cadastrales ;
 - et, naturellement, des données de gestion des administrations ou autres organismes rencontrés par les inspecteurs. La manipulation de celles-ci demande une certaine prudence ; il est en effet nécessaire de s'assurer de leur qualité et de leur périmètre, mais aussi de la cohérence des concepts qui les sous-tendent avec ceux nécessaires à la mission d'inspection. Il peut s'agir par exemple de données issues de logiciels internes, qui seraient remplies à la main (avec le risque d'erreur et d'inconsistance que

³ Le fichier POTE rassemble les informations recueillies dans le cadre des déclarations des foyers fiscaux à l'impôt sur le revenu des personnes physiques (IRPP).

⁴ Le fichier FARE rassemble les données individuelles comptables des entreprises.

⁵ Le fichier BIC-IS rassemble les informations recueillies dans le cadre des déclarations de bénéfices industriels et commerciaux (BIC), pour différents régimes d'imposition, et des déclarations à l'impôt sur les sociétés (IS).

► **Figure 1 - Fonctionnement du pôle science des données de l'inspection générale des finances**



PSD : Pôle science des données.

cela entraîne), sous un format potentiellement propriétaire, sans identifiant commun avec d'autres bases accessibles par ailleurs, et donc sans possibilité d'appariement (Koumarianos et al., 2024).

Parfois, il arrive que le premier niveau de données soit suffisant pour la mission, auquel cas le rôle du pôle se borne à accompagner les inspecteurs dans la compréhension des concepts statistiques sous-jacents aux données en question, en expliquant par exemple le fonctionnement de la nomenclature d'activité française (NAF)⁶, ce qu'est le **numéro Siren**⁷, etc.

Souvent, les données existantes ne permettent de répondre aux questions que se posent les missions que de façon très imparfaite : elles peuvent être trop anciennes, ne couvrir que partiellement le champ d'intérêt ou, pour certaines données de gestion, souffrir de défauts dans leur conception tels que les résultats obtenus doivent être considérés avec grande précaution. Ce dernier type de données « externes non maîtrisées » met aussi le statisticien-*data scientist* au défi d'entrer dans la logique de constitution des fichiers afin de « reconstituer a posteriori un pseudo-appareil d'observation, à partir de données

qui n'ont pas été conçues à cette fin » (Rivière, 2020). Cela passe par des échanges étroits avec les producteurs, gestionnaires ou utilisateurs de ces fichiers tout au long des travaux.



La démarche initiale du data scientist est systématiquement la même : comprendre le sens des variables, les clés entre les fichiers, puis essayer de reproduire, quand ils existent, des résultats agrégés produits par une autre administration.



Quelle que soit la catégorie des données mobilisées, la démarche initiale du *data scientist* est systématiquement la même : comprendre le sens des variables, les clés entre les fichiers, puis essayer de reproduire, quand ils existent, des résultats agrégés produits par une autre administration. Il s'agira par exemple de retrouver le montant total d'exonération de cotisations sociales, ou bien le nombre total de patients en affection de longue durée en France

hexagonale. Cette étape est cruciale et permet de s'assurer que le *data scientist* a compris la signification des variables et qu'il applique les bons filtres et transformations à sa base de données. Une fois les faits stylisés établis, le *data scientist* peut sereinement mener son analyse pour répondre aux questions réelles de la mission. Bien sûr, dans le cas de données issues du CASD, le pôle s'enrichit de ses bonnes pratiques acquises au fur et à mesure des missions, si bien qu'il a développé des traitements de données standardisés (ou *pipelines*) pour certaines bases fréquemment utilisées, ce qui permet de limiter le risque d'erreur.

Utiliser des infrastructures de calcul développées par l'écosystème

Le pôle est un utilisateur enthousiaste du CASD qui, en sus d'héberger les données sensibles, offre également des serveurs de calcul permettant de réaliser des analyses dans une bulle sécurisée. Pour traiter les données hors CASD, en revanche, il est nécessaire

⁶ <https://www.insee.fr/fr/information/2406147>.

⁷ <https://www.insee.fr/fr/metadonnees/definition/c2047>.

de disposer d'un environnement adapté : les ordinateurs du ministère n'ont encore pas la puissance de calcul nécessaire pour traiter des données volumineuses, par exemple des données géographiques fines, et les politiques de sécurité informatique brident largement l'installation de logiciels nécessaires à ces traitements.

Pour contourner ces difficultés, le pôle a recours à deux plateformes de *data science* offrant un vaste catalogue de services : le SSPCloud et Nubonyxia. Ces deux plateformes sont des instances d'Onyxia (Comte et al., 2022), une solution de *data science* intégralement développée par l'Insee et accessible en *open source*, qui présente une souplesse, une puissance et une ergonomie parfaitement adaptées aux besoins du pôle. Le SSPCloud, développé et hébergé par l'Insee, est la plateforme la plus mature, mais sa principale limite est de ne pas permettre d'héberger des données confidentielles. Nubonyxia, de son côté, est un déploiement d'Onyxia sur les serveurs Nubo de la Direction générale des Finances publiques (DGFIP) réalisé par le Bercy Hub. Ce projet est à un stade de développement plus précoce, mais il offre des garanties de sécurité plus importantes et est connecté au réseau interministériel de l'État. Ces deux infrastructures disposent par ailleurs de tous les services nécessaires à la plupart des projets de *data science* : environnement de développement intégré (VSCode, Rstudio, JupyterLab), système de gestion de base de données (PostgreSQL), suivi d'entraînement de modèle de *machine learning* (MLFlow). Elles permettent également de déployer des applications web développées en Python ou en R, et de les mettre à disposition des autres membres des missions ou des commanditaires.

Malgré les qualités de ces plateformes, il peut cependant arriver que, dans un arbitrage entre la qualité du rendu et les délais que le pôle doit tenir, les exploitations de données n'aillent pas jusqu'à leur terme théorique. Ainsi, dans le cadre d'une mission sur la gestion de la forêt privée, le *data scientist* a dû se limiter à travailler sur un échantillon de dix départements forestiers pour produire ses résultats, les volumes de données à traiter étant trop importants (de l'ordre d'une dizaine de gigaoctets par département).

Implémenter des modèles de *data science* pour objectiver les faits

Les missions de l'IGF auxquelles le pôle est associé sont essentiellement des missions d'évaluation des politiques publiques. Dans ce contexte, la plus-value du pôle science des données est sa polyvalence, c'est-à-dire sa capacité à mobiliser des techniques

quantitatives variées, et notamment aussi bien des méthodes économétriques que des méthodes de *machine learning*.

La plus-value du pôle science des données est sa polyvalence, c'est-à-dire sa capacité à mobiliser des techniques quantitatives variées, et notamment aussi bien des méthodes économétriques que des méthodes de machine learning.

Ces missions comprennent généralement deux phases distinctes : une phase de diagnostic et une phase de simulation de réforme. Pour chacune d'elles, il est crucial de procéder avec des méthodes scientifiques et notamment de **s'assurer de raisonner « toutes choses égales par ailleurs »**. Les délais courts (2 mois en général) dans lesquels les conclusions des travaux du pôle sont attendues exigent de procéder avec beaucoup d'efficacité, d'abord

pour comprendre et pré-traiter les données (nettoyage, changements de nomenclatures, traitement des valeurs manquantes), puis pour construire un contrefactuel convaincant, écarter les variables confondantes⁸, estimer le modèle lui-même et, enfin, l'interpréter (Givord, 2014). Quand la tâche impose de construire un modèle de microsimulation (Blanchet, 2020), la même efficacité est requise. La capitalisation sur les codes produits d'une mission à l'autre se révèle ainsi essentielle pour livrer les productions à temps et harmoniser les réalisations du pôle.

Les méthodes de *machine learning* ont des usages plus variés : prédire des valorisations, dégager des regroupements (ou *clusters*) d'individus, faire du codage automatique, etc. Au-delà des distinctions méthodologiques entre les tâches confiées au pôle, un point commun à celles-ci est que les *data scientists* ont vocation à produire une méthodologie ou une application qui reste à l'état de « preuve de concept ». En effet, contrairement au système statistique public, l'IGF n'a pas vocation à réaliser des productions périodiques : les modélisations, évaluations et chiffrages sont faits dans le cadre d'une mission délimitée dans le temps et ne sont pas destinés à être répliqués d'une année à l'autre. Si toutefois cette réplique s'avère utile pour le décideur, il appartient aux administrations compétentes de reprendre à leur compte les méthodologies développées par le pôle, de les améliorer et de les intégrer à leur production statistique. En particulier, selon les cas et le statut de confidentialité de la mission, les codes produits peuvent être directement transmis aux services concernés, ou mis en ligne pour contribuer à des projets en code source ouvert.

► Un exemple de mission : l'assurabilité des collectivités territoriales

Afin d'illustrer de façon plus concrète le type de travaux conduits, nous présentons le soutien apporté par l'équipe du pôle science des données à une mission portant sur l'assurabilité des biens des collectivités locales, mission à laquelle l'IGF et l'inspection générale de l'administration ont apporté une assistance technique⁹.

Attendu : objectiver les difficultés des collectivités locales à assurer leurs biens

La mission a été lancée à la suite de difficultés financières rencontrées par certains acteurs de l'assurance des biens des collectivités territoriales et d'obstacles, signalés par certaines communes, au renouvellement de leur contrat d'assurance. Ces difficultés et obstacles s'inscrivent dans un contexte d'accroissement des risques climatiques et après les épisodes de violences urbaines à l'été 2023, ayant engendré plus d'un milliard d'euros de dommages. Dans le cadre de cette mission, les membres du pôle science des données ont contribué à quantifier l'évolution des dépenses d'assurance des collectivités territoriales, à mettre en évidence leurs déterminants et à estimer la valeur du patrimoine exposé à un ou plusieurs risques naturels.

⁸ Variables qui influencent à la fois une variable dépendante et une variable indépendante.

⁹ Mission d'assistance auprès des personnalités qualifiées Alain Chrétien, maire de Vesoul, et Jean-Yves Dagès, ancien président de Groupama.



Les membres du pôle science des données ont contribué à quantifier l'évolution des dépenses d'assurance des collectivités territoriales, à mettre en évidence leurs déterminants et à estimer la valeur du patrimoine exposé à un ou plusieurs risques naturels.



Ces estimations relatives à l'exposition du patrimoine des collectivités territoriales font écho à des travaux menés par des acteurs de la statistique publique sur le patrimoine immobilier des ménages à partir du répertoire Fidéli¹⁰ (André et Meslin, 2022 et 2025). Cependant, ces travaux diffèrent par le champ sur lequel ils portent : les logements possédés par les ménages dans le cas des travaux conduits par l'Insee, les biens des communes et communautés de communes pour les travaux de la mission. Ces derniers ont aussi été menés dans un délai serré de quelques mois, ce qui induit des différences importantes dans le retraitement des données et les méthodes mises en œuvre.

Données : inventorier le patrimoine des communes à l'aide du cadastre

Le premier chantier entrepris afin de caractériser le patrimoine immobilier et son exposition aux risques naturels a consisté en l'inventaire des biens immobiliers possédés par les communes. En effet, si un tel inventaire existe pour les ménages (Fidélismo¹¹ côté Insee ou Majic¹² côté DGFIP), ce n'est pas le cas pour les collectivités territoriales¹³. Or, il était indispensable de pouvoir localiser finement tous les bâtiments possédés par les communes (dont les écoles, les églises, etc.), afin de mesurer leur exposition aux risques naturels. Il fallait aussi pouvoir valoriser chaque bâtiment et donc recueillir des caractéristiques sur lesquelles s'appuyer pour estimer une valeur vénale. En lien avec la sous-direction de la gestion comptable et financière des collectivités territoriales et le bureau du cadastre de la DGFIP, une base de données inventoriant et valorisant le patrimoine bâti des collectivités territoriales a ainsi été construite.

Les parcelles appartenant à des communes ou des établissements publics de coopération intercommunale (EPCI) sont recensées à partir des fichiers des locaux et des parcelles des personnes morales, dérivés des données cadastrales et publiés par la DGFIP. Ainsi, 356 000 parcelles bâties et 5,4 millions de parcelles non bâties appartenant à une collectivité du bloc communal¹⁴ en 2023 ont pu être approchées à partir de ces données. Les infrastructures telles que les ouvrages d'art et de voirie n'y sont cependant pas référencées.

¹⁰ <https://www.insee.fr/fr/metadonnees/source/serie/s1019>.

¹¹ Fidélismo est une extension du répertoire Fidéli qui vise à mieux analyser le patrimoine immobilier des ménages.

¹² Le fichier Majic rassemble les informations cadastrales sur les parcelles, bâties ou non bâties, et les locaux (propriétés bâties).

¹³ Les données comptables renseignent sur la valorisation du patrimoine détenu par les collectivités, mais ne sont pas détaillées bien par bien.

¹⁴ On désigne par là une commune, une communauté de communes, une communauté de villes, une communauté d'agglomérations, une communauté urbaine, un syndicat intercommunal à vocation multiple (SIVOM) ou un syndicat intercommunal à vocation unique (SIVU).

Cet inventaire a ensuite été enrichi à partir d'autres sources de données géocodées :

- La table « Bâti – Parcelle » de la BAN PLUS, produite par l'IGN, associe à chaque parcelle référencée les bâtiments qu'elle abrite : 535 000 bâtiments ont ainsi pu être identifiés sur 224 000 des 356 000 parcelles bâties (en ne retenant que la parcelle principale de chaque bâtiment) ;
- La table « Bâtiment » de la BD TOPO, produite par l'IGN, fournit les caractéristiques physiques des bâtiments (surface au sol, hauteur, état) ;
- La table « ERP » de la BD TOPO renseigne sur la nature et la capacité d'accueil des établissements recevant du public (ERP) comme les mairies, églises, écoles, etc. ;
- La **base permanente des équipements**¹⁵ (BPE), produite par l'Insee et géocodée, répertorie en outre un large éventail d'équipements et de services, marchands ou non, accessibles au public au 1^{er} janvier de chaque année. Elle renseigne ainsi sur la fonction des bâtiments publics dont les collectivités sont propriétaires. En 2021, elle porte sur 188 types de services et équipements différents, répartis en sept grands domaines : services aux particuliers, commerces, enseignement, santé-social, transports-déplacements, sports-loisirs-culture et tourisme (Helfenstein, 2022).

L'appariement entre les données cadastrales et la table « Bâtiment » est réalisé en utilisant l'identifiant cadastral de la parcelle. Les informations contenues dans la table « ERP » et dans la BPE sont ajoutées à la base de données en superposant les informations géocodées dans ces deux sources avec la géométrie des bâtiments et des parcelles (**figure 2**).

Modélisation : estimer la valeur du patrimoine par apprentissage

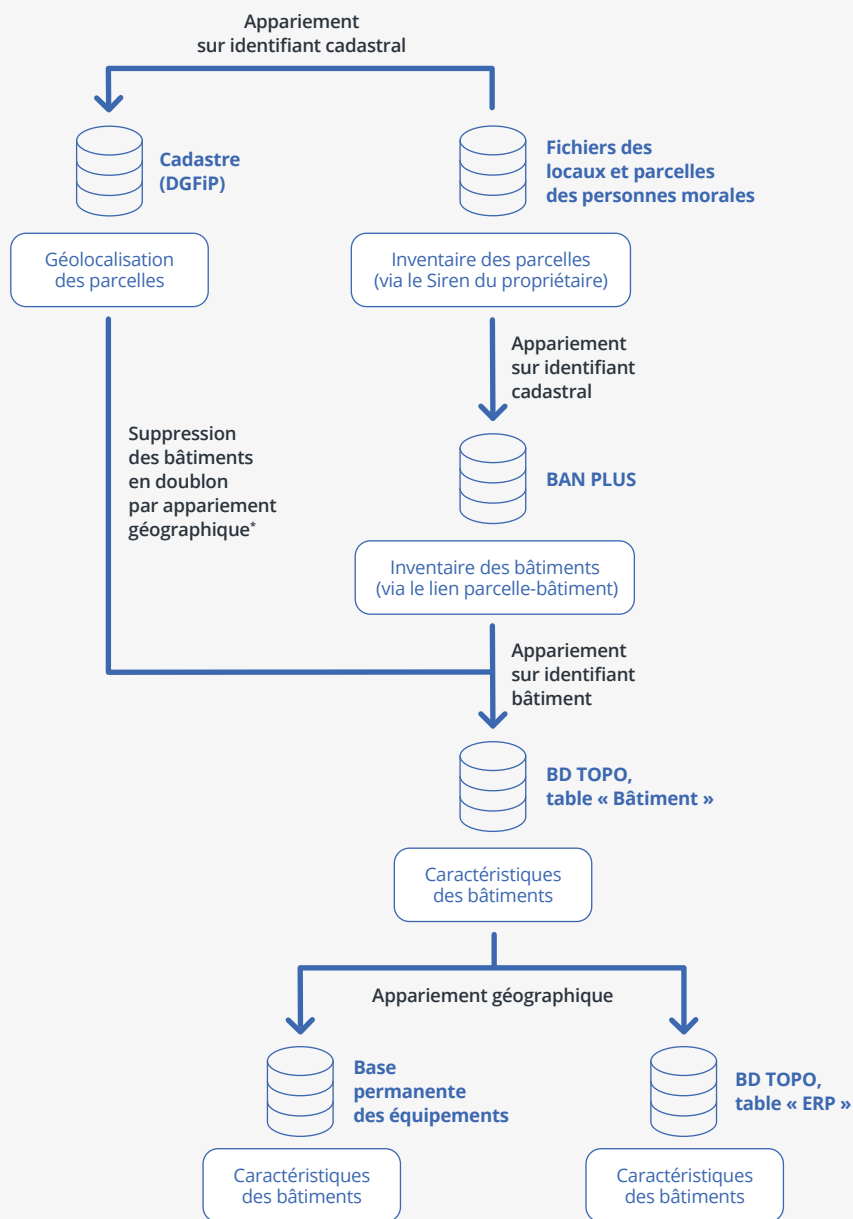
Pour mesurer l'exposition du patrimoine des collectivités aux risques, il est nécessaire de quantifier ce patrimoine, en l'occurrence via une valeur vénale. Cette valeur n'est naturellement pas présente dans les données citées ci-dessus. Il faut donc trouver une façon d'en construire une estimation, typiquement à l'aide d'un modèle de *machine learning* qui apprendrait la valeur en fonction d'un certain nombre de caractéristiques. Cela suppose d'avoir une base de référence qui présente à la fois les caractéristiques en question et une valeur vénale, c'est-à-dire un prix de vente.

L'inférence de la valeur vénale des biens des collectivités a donc été réalisée à partir du fichier des demandes de valeurs foncières (DVF) publié par la DGFIP. Ce dernier recense l'ensemble des ventes de biens fonciers réalisées entre 2018 et 2023 en France, sauf à Mayotte et en Alsace-Moselle, soit 7,7 millions de transactions sur cinq ans. La base contient en particulier les prix de mutation enregistrés par les notaires. Toutefois, il y a peu de ventes de biens publics dans ces données. L'estimation de la valeur des biens des collectivités a donc été conduite en deux étapes : une première qui permet d'obtenir le prix du bien s'il était initialement détenu par un acteur privé et une deuxième qui adapte ce prix au fait que le bien serait hypothétiquement vendu par un acteur public :

- Un premier modèle XGBoost (**encadré 1**) est entraîné sur les 3,1 millions de transactions de biens n'appartenant pas à des collectivités en 2023 pour lesquelles les prix et les

¹⁵ <https://www.insee.fr/fr/metadonnees/source/serie/s1161>.

► **Figure 2 - Appariement de diverses sources de données permettant d'inventorier et caractériser le patrimoine des collectivités territoriales**



* Certains bâtiments sont à cheval sur plusieurs parcelles et apparaissent donc plusieurs fois dans la base, liés à plusieurs parcelles. Un bâtiment dont 75 % de la géométrie se trouve dans la parcelle A et 25 % dans la parcelle B sera rattaché à la parcelle A.
 BAN PLUS : Base Adresse Nationale PLUS.
 BD TOPO : Base de données topographiques.
 ERP : Établissements recevant du public.
 DGFIP : Direction générale des Finances publiques.

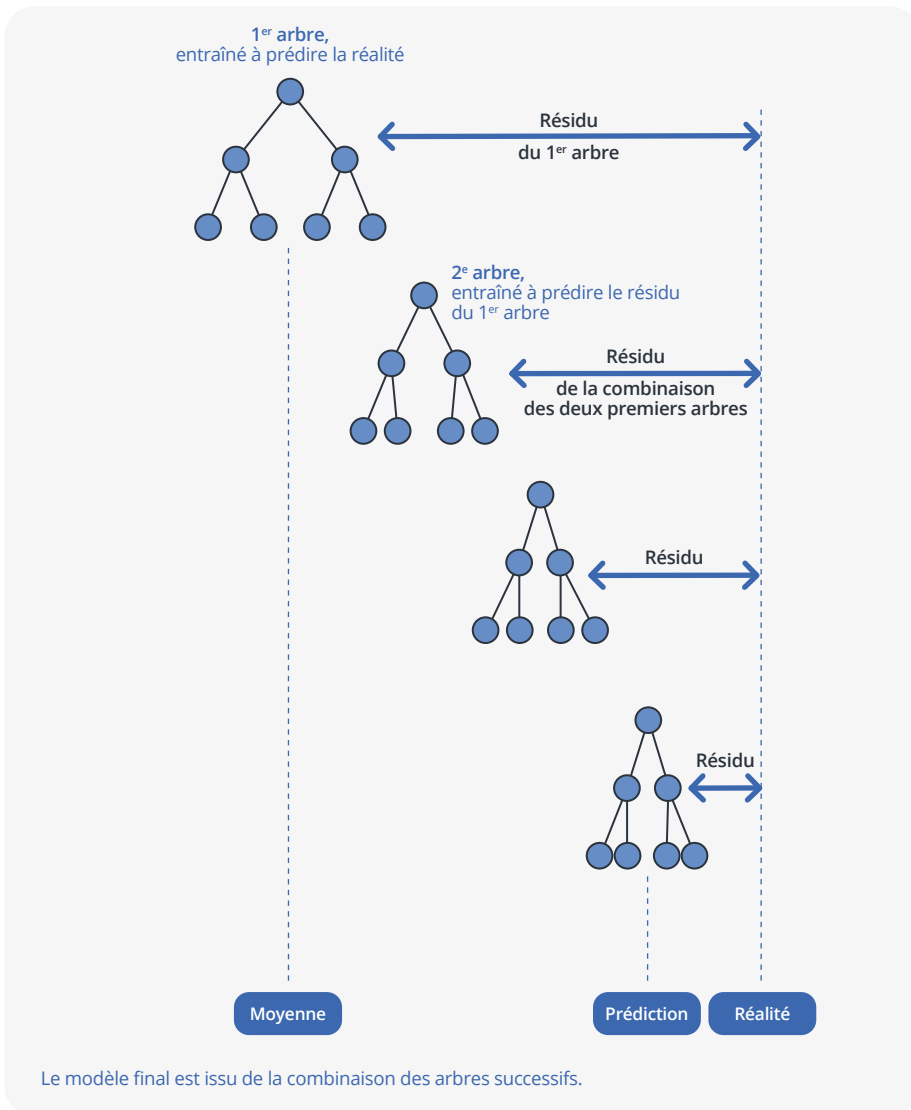
► Encadré 1. Fonctionnement de l'algorithme XGBoost

L'algorithme XGBoost (*Extreme Gradient Boosting Tree*) est un modèle d'apprentissage basé sur des arbres de régression ou de classification (Chen et Guestrin, 2016). La statistique publique l'utilise dans certains travaux (André et Meslin, 2025).

L'algorithme d'entraînement combine les prédictions de nombreux modèles simples pour créer un modèle plus puissant. Pour réaliser cette combinaison, il emploie des méthodes dites de *boosting* (Freund et Schapire, 1996), c'est-à-dire que le modèle apprend en construisant itérativement des arbres de décision, chaque arbre étant entraîné à prédire les erreurs de prédiction des arbres précédents combinés (*figure encadré*).

Ainsi, le premier arbre est entraîné à prédire la réalité. Puis, le deuxième arbre est entraîné à prédire les résidus du premier arbre. Les prédictions du deuxième arbre sont alors multipliées par un facteur η inférieur à 1 et ajoutées aux prédictions du premier arbre. Un troisième arbre est alors entraîné à prédire les résidus de la combinaison des deux premiers arbres, écart entre la réalité et la combinaison des prédictions, etc.

L'algorithme XGBoost est optimisé, notamment pour le calcul parallèle, afin de pouvoir être entraîné rapidement, et régularisé pour empêcher que le modèle devienne trop complexe, diminuant ainsi les risques de surapprentissage.



caractéristiques des bâtiments sont correctement renseignés. Le modèle s'appuie sur les caractéristiques physiques du bâtiment (nature, surface, hauteur, nombre d'étages, superficie de la parcelle) et sur des informations afférentes à son environnement (coordonnées, revenu médian de l'Iris¹⁶, caractéristiques de la commune) pour estimer un prix de vente.

- Un second modèle XGBoost est ensuite entraîné sur les 43 000 ventes de parcelles appartenant en 2023 aux communes et aux EPCI à partir des mêmes caractéristiques que celles utilisées en première approche, enrichies des prédictions du premier modèle. Cette deuxième étape permet de raffiner le modèle sur des données plus spécifiques à notre cas d'usage, tout en exploitant les résultats du premier modèle.

Suivant cette approche, le patrimoine immobilier des communes et EPCI s'élèverait à 428 milliards d'euros (Md€), soit une valeur de même ordre mais supérieure aux estimations tirées des données comptables (350 Md€) ou de la comptabilité nationale (275 Md€). Ces trois décomptes souffrent de limites qui leur sont propres. Pour celui mis en œuvre ici, la difficulté est d'extrapoler la valeur des biens vendus, qui peuvent être assez spécifiques, aux biens possédés par les collectivités n'ayant souvent pas vocation à être cédés (écoles, églises, piscines, etc.). Si les valeurs de patrimoine agrégées diffèrent, la répartition géographique du patrimoine est cependant assez cohérente dans notre approche et dans les données comptables.

Dernière étape : croiser l'estimation du patrimoine et les risques naturels

Une fois les bâtiments inventoriés, caractérisés et associés à une valeur vénale approchée, l'objectif est d'identifier les différents risques naturels pesant sur chacun d'eux (**encadré 2**). Le choix des bases de données permettant de qualifier le niveau des risques encourus a été effectué en suivant les recommandations de différents producteurs de données rencontrés par la mission, notamment le Service des données et études statistiques (SDES)¹⁷, le Cerema¹⁸ et la Direction générale de la prévision des risques (DGPR). Les cartes de risques au niveau infracommunal ont été retenues lorsque les données existaient, ce qui est le cas pour le retrait-gonflement des sols argileux, les inondations, via l'enveloppe approchée du risque d'inondation potentiel (EAIP), et les feux de forêt. Dans le cas contraire, des données mises en ligne par l'Observatoire national sur les effets du réchauffement climatique (ONERC) et produites par le SDES ont été utilisées pour mesurer les risques à l'échelle communale (mouvements de terrain, séismes).

Cette approche présente cependant des limites inhérentes aux données disponibles pour documenter l'exposition aux différents risques du territoire français. Ainsi, les mesures peuvent se révéler imprécises. En outre, ces données ne permettent pas de caractériser la sinistralité induite par la matérialisation d'un risque donné.

L'exposition du patrimoine des communes est finalement qualifiée par la superposition de l'inventaire du patrimoine total et des cartes des risques naturels ainsi construites (**figure 3**). 240 000 bâtiments des communes sont situés dans des zones exposées à un

¹⁶ <https://www.insee.fr/fr/metadonnees/definition/c1523>.

¹⁷ Service statistique ministériel de l'énergie, du logement, du transport et de l'environnement.

¹⁸ Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement.

► Encadré 2. Les risques naturels retenus dans l'analyse

Certains risques naturels ont été écartés de l'analyse : le recul du trait de côte, qui est jugé trop déterministe pour être pris en compte dans un mécanisme assurantiel, et le risque de grêle, dont l'emplacement géographique est trop difficile à prédire. Les risques pris en compte sont les suivants :

- **Le risque d'inondation** est mesuré par l'enveloppe approchée du risque d'inondation potentiel (EAIP). Définie à l'échelle infracommunale, l'EAIP permet de tenir compte de deux aléas hydrologiques distincts : les inondations par submersion marine et les débordements de cours d'eau ou de crues. Elle renseigne sur le risque potentiel, sans détailler l'intensité du risque afférent aux différentes zones concernées. Son approche est maximaliste au regard de l'étendue des zones de risques.

- **Le risque de retrait-gonflement des sols argileux** (RGA) est documenté par une cartographie départementale réalisée entre 1997 et 2010 par le Bureau de recherches géologiques et minières (BRGM). Ce risque, lié aux conditions météorologiques, n'est pas dangereux

pour l'humain, mais est potentiellement très dommageable pour le bâti.

- **Les risques de mouvement de terrain et de séisme** sont cartographiés à la maille communale à partir de données mises en ligne par l'Observatoire national sur les effets du réchauffement climatique (ONERC). Ils incluent notamment les glissements de terrain, les coulées boueuses de diverses intensités, les affaissements et effondrements dus à des cavités souterraines.

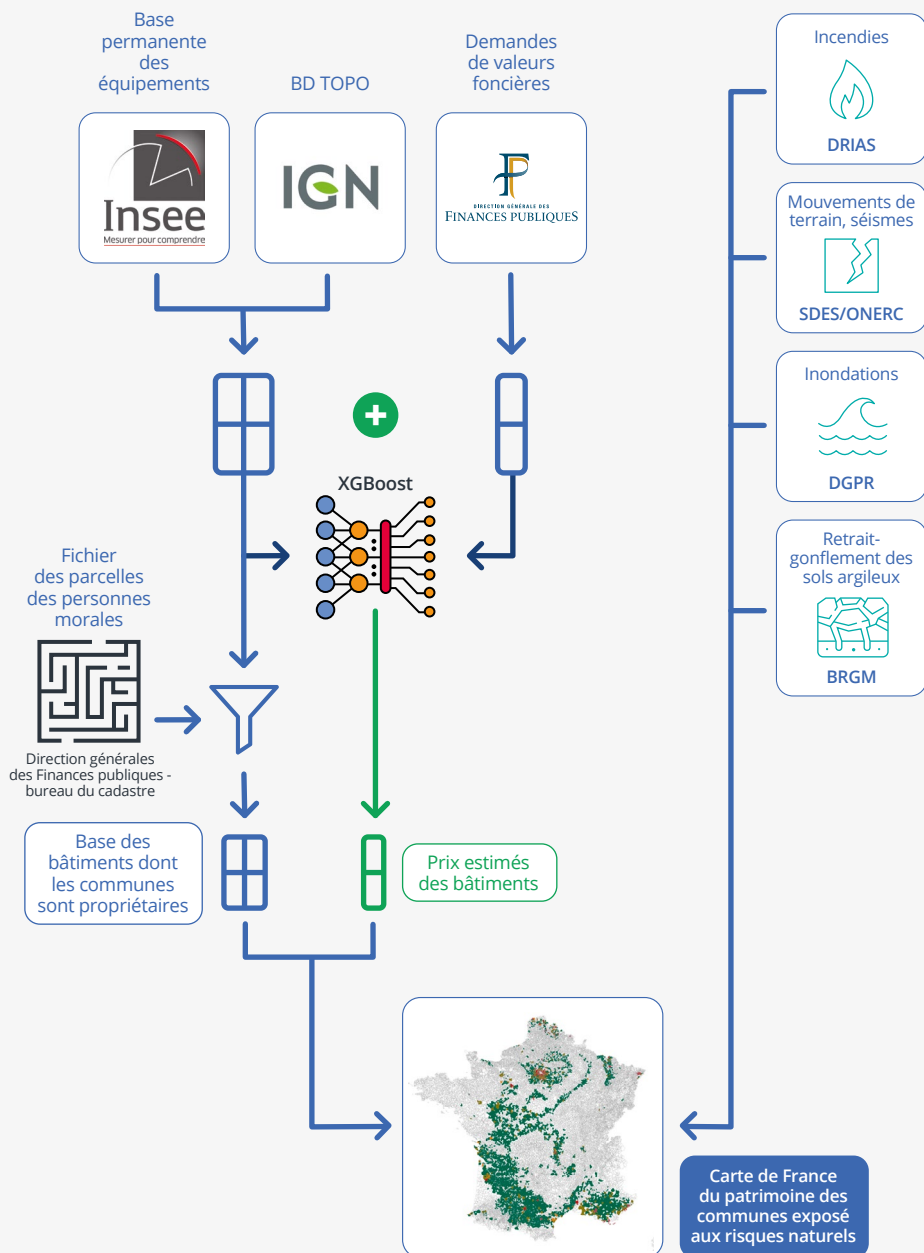
- **Le risque de feux de forêt** est quantifié par l'indice Feu Météo (IFM), qui rend compte de la propension à l'éclosion et à la propagation des feux. Cet indice peut être calculé à des horizons temporels éloignés, en s'appuyant sur les scénarios climatiques du deuxième Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC2). Le scénario retenu est le plus proche de la trajectoire de référence pour l'adaptation au changement climatique (TRACC), soit le scénario RCP 4.5, à partir duquel est calculé un indice Feu météo journalier. Le nombre de jours tels que l'indice Feu Météo est supérieur au seuil de 40 est retenu comme indicateur de risque de feux de forêt à l'horizon 2100.

risque sévère et géolocalisé précisément, soit 49 % des bâtiments des communes et 51 % de leur patrimoine bâti estimé par les données de transactions immobilières. Le facteur de risque sévère précisément appréhendé par les données et qui touche le plus de biens correspond aux inondations provoquées par les débordements de cours d'eau (31 % des bâtiments et 40 % du patrimoine bâti). La part des bâtiments des communes exposée aux risques est proche de celle des personnes morales, pour lesquelles l'analyse a été répliquée à titre de comparaison (49 %, contre 46 %). La part du patrimoine bâti des communes exposée aux risques d'inondation par débordement de cours d'eau (31 %) ou submersion marine (3 %) est également comparable à celle des ménages, champ pour lequel une mesure comparable est produite par le SDES (SDES et ONRN, 2024), s'établissant à 28 % et 3,5 % respectivement.

La méthode permet de s'intéresser spécifiquement à l'exposition à certains risques (*figures 4 et 5*). Par exemple, 15 % des bâtiments des collectivités territoriales sont à risque fort de retrait-gonflement des sols argileux, pour une valeur vénale de 61 Md€ (soit 14 % du patrimoine bâti) ; les bâtiments les plus exposés à ce risque et les plus chers sont concentrés en Île-de-France et autour de la Méditerranée. De même, 31 % des bâtiments se trouvent dans des zones à risque d'inondation par remontée de cours d'eau au sens de l'EAIP (162 Md€, soit 38 % du patrimoine bâti).

Les estimations de la valeur et de l'exposition aux risques naturels de chaque bien détenu par les communes et les EPCI sont enfin combinées et utilisées pour expliquer les différences de niveau des dépenses d'assurance, telles qu'enregistrées dans les données comptables des collectivités. Toutes choses égales par ailleurs, une commune dont l'intégralité des biens est à risque d'inondation par remontée des cours d'eau s'acquitte d'une prime plus élevée de 8 % par rapport à celles dont aucun bien n'est sujet à ce risque. La surprime est de 5 % en cas de risque de retrait-gonflement des sols argileux.

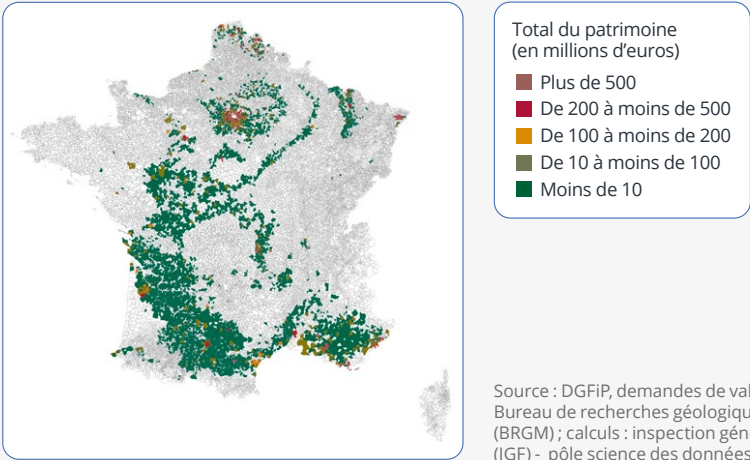
► **Figure 3 - Les données mobilisées pour l'analyse des dépenses d'assurance des collectivités**



DRIAS : Donner accès aux scénarios climatiques Régionalisés français pour l'Impact et l'Adaptation de nos Sociétés et environnement.
 SDES/ONERC : Service des données et études statistiques / Observatoire national sur les effets du réchauffement climatique.
 DGPR : Direction générale de la prévision des risques.
 BRGM : Bureau de recherches géologiques et minières.

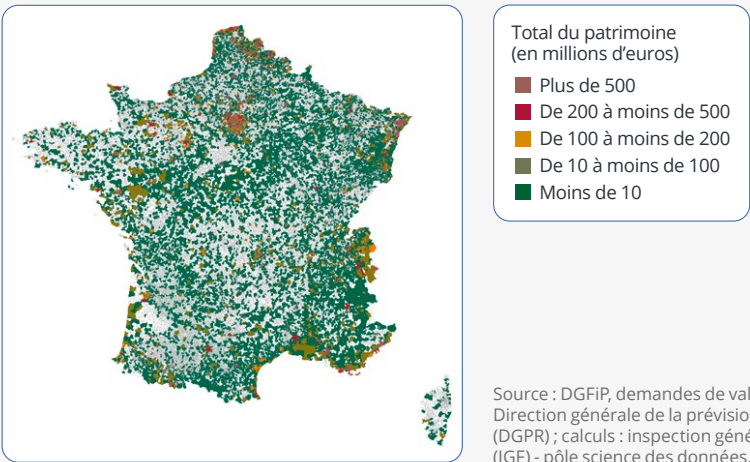
Les risques définis à une maille moins fine (séismes ou mouvements de terrain) ne sont pas significativement associés à des dépenses d'assurance plus élevées. Ces résultats ne valent que lorsque l'analyse est pondérée par la taille des communes : la relation positive entre exposition aux risques et dépenses d'assurance concerne les communes les plus peuplées. Ce résultat pourrait indiquer que les communes les plus peuplées et leurs assureurs connaissent davantage leur exposition au risque, et que ces derniers sont plus à même de la retranscrire dans le niveau des primes d'assurance.

► **Figure 4 - Valeur vénale du patrimoine exposé au risque de retrait-gonflement des sols argileux**



Source : DGFIP, demandes de valeurs foncières ; Bureau de recherches géologiques et minières (BRGM) ; calculs : inspection générale des finances (IGF) - pôle science des données.

► **Figure 5 - Valeur vénale du patrimoine exposé au risque d'inondation par remontée de cours d'eau**



Source : DGFIP, demandes de valeurs foncières ; Direction générale de la prévision des risques (DGPR) ; calculs : inspection générale des finances (IGF) - pôle science des données.

► La contribution de la science des données à l'IGF : une demande en forte croissance

Un levier de transformation, présentant néanmoins des limites

Des équipes comme le pôle science des données de l'IGF répondent à une demande croissante d'analyses quantitatives, dans des délais souvent contraints, et contribuent à la solidité des constats établis par les missions de conseil au décideur.

La demande en analyse de données est forte de la part des décideurs publics. Elle l'est en conséquence aussi pour les inspections. Grâce à leur flexibilité, des équipes comme le pôle science des données de l'IGF répondent à une demande croissante d'analyses quantitatives, dans des délais souvent contraints, et contribuent à la solidité des constats établis par les missions de conseil au décideur.

Ces équipes jouent un rôle d'agrégateur de la connaissance et des données produites par le système statistique public. La quantification existait auparavant à l'IGF, mais les missions

devaient se débrouiller avec les données, en extrapolant des chiffrages existants ou en mobilisant elles-mêmes des méthodes qu'elles maîtrisaient. Les *data scientists*, à travers l'agrégation et la mise en perspective des données, apportent une dimension quantitative et systématique qui complète les approches généralement plus qualitatives des inspecteurs. Cela conduit en pratique à une forte participation des *data scientists* aux missions de conseil de l'inspection. Une trentaine de missions ont ainsi été accompagnées par le pôle science des données de l'IGF en 2023, dont quelques-unes portées uniquement par le pôle, soit une mission sur deux (IGF, 2024).

Par ailleurs, ce développement contribue fortement à l'acculturation des décideurs et des inspecteurs aux méthodes d'analyse quantitative, en se plaçant en intermédiaire, pour ainsi dire en pont, entre la statistique publique et les décideurs.

Ce mode de fonctionnement présente néanmoins certaines limites : s'il est utile de pouvoir mobiliser rapidement des données administratives ou des productions du système statistique public, il ne faut toutefois pas imaginer qu'il est possible de réaliser n'importe quelle étude sur cette base. Certaines études se heurtent à l'absence de données adaptées, les bases disponibles étant parfois trop anciennes, ne couvrant pas le périmètre nécessaire ou requérant un temps d'acculturation et de retraitement incompatible avec les délais de la mission. Il est également fréquent, mais c'est un problème très général en évaluation des politiques publiques, qu'une politique dans sa définition ne permette pas d'établir un contrefactuel, c'est-à-dire de trouver un point de comparaison pour apprécier l'impact de la mesure. Par exemple, si l'on cherche à évaluer un plan de soutien à un secteur entier de l'industrie, il est délicat de discerner un effet sur les entreprises traitées. Il faut alors recourir à des techniques plus élaborées comme le contrôle synthétique (Abadie et al., 2010), dont la mise en œuvre est susceptible de ne pas tenir dans les délais impartis, ou bien qui risque in fine de manquer de puissance statistique.

Les conditions d'une intégration réussie de la donnée dans la décision publique

L'expérience acquise dans le déploiement de la science des données au sein de l'IGF permet de dégager plusieurs enseignements pour maximiser son impact et favoriser son adoption.

Tout d'abord, il est très important de toujours faire preuve d'une grande pédagogie auprès des décideurs et des inspecteurs pour accompagner à chaque étape la remise des conclusions des analyses quantitatives. En effet, leur portée, leurs limites et leurs implications ne sont pas toujours évidentes pour les non-spécialistes. Bien expliquer les hypothèses sous-jacentes, les choix méthodologiques et les résultats est essentiel pour garantir une adoption éclairée et éviter tout malentendu. L'acculturation va d'ailleurs dans les deux sens : les *data scientists* eux-mêmes bénéficient des retours des inspecteurs et décideurs, qui apportent un regard critique et des connaissances opérationnelles indispensables. Cette interaction favorise une meilleure intégration des résultats dans les recommandations, en les rendant à la fois plus accessibles et plus actionnables.



La fluidité dans les échanges de données et de documentation entre administrations apparaît comme un levier crucial.



Ensuite, il est fondamental de pouvoir s'appuyer sur des infrastructures communes, sécurisées et robustes pour l'accès aux données. La diversité des sources mobilisées et la complexité des systèmes administratifs peuvent rendre ce travail particulièrement fastidieux et chronophage. Des infrastructures partagées comme le CASD permettent de standardiser les accès, garantir la sécurité et réduire le temps

consacré à la collecte et à la préparation des données, laissant ainsi davantage de place à l'analyse et à la réflexion stratégique. Ces gains d'efficacité sont particulièrement précieux dans le cadre de missions souvent soumises à des délais contraints : il s'agit ici de lever le plus de freins possibles à la mobilisation de la donnée pour la décision publique.

Enfin, toujours dans cet objectif, la fluidité dans les échanges de données et de documentation entre administrations apparaît comme un levier crucial. Les travaux de science des données mobilisent souvent de multiples sources administratives. Or, la cartographie des bases de données et leur documentation sont parfois inexistantes, si bien que leur accès peut se révéler complexe. Harmoniser ces échanges, que ce soit par des protocoles simplifiés ou par des plateformes communes, permettrait non seulement de fluidifier les travaux, mais aussi de mieux exploiter le potentiel des données publiques, comme l'exemple du CASD le montre avec brio.

Le développement de la démarche

Le développement d'un pôle dédié à la science des données a été permis par l'ouverture et la mise à disposition des données publiques. Dans le contexte récent de réinternalisation des missions des cabinets privés au sein de l'État, l'IGF est ainsi en mesure de répondre à un panel plus large de missions, en mobilisant à plein les capacités de ses *data scientists*.

Comme évoqué, d'autres inspections ou instances d'évaluation suivent une démarche similaire afin d'appuyer ou d'approfondir leurs constats à l'aide des données. C'est notamment le cas de l'inspection générale des affaires sociales (Igas)¹⁹, de l'inspection générale de l'environnement et du développement durable (IGEDD) et de l'inspection générale de l'éducation, du sport et de la recherche (IGESR), mais aussi de la Cour des comptes, preuve de l'intérêt suscité au sein de ces services par la perspective de conduire en interne une partie des analyses quantitatives dont les missions ont besoin.

Ces développements s'inscrivent dans une démarche plus large de renforcement de l'écosystème public de la science des données et de l'évaluation des politiques publiques, incluant notamment le monde académique (via des institutions comme l'Institut des politiques publiques). Cette démarche renforce par là même la sensibilisation des décideurs aux enjeux de l'évaluation et du recours aux données pour réaliser leurs arbitrages.

¹⁹ Voir l'article de Juliette Berthe dans ce même numéro.

► Fondements juridiques

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance*. [en ligne]. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746>.

► Bibliographie

- ABADIE, Alberto, DIAMOND Alexis et HAINMUELLER, Jens, 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. In : *Journal of the American Statistical Association*. [en ligne]. Juin 2010, Vol. 105, N° 490, pp. 493-505. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://www.tandfonline.com/doi/abs/10.1198/jasa.2009.ap08746>.
- ANDRÉ, Mathias et MESLIN, Olivier, 2022. Patrimoine immobilier des ménages : enseignements d'une exploitation de sources administratives exhaustives. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 107-125. [Consulté le 10 mars 2025]. Disponible à l'adresse : <https://www.insee.fr/fr/information/6035944?sommaire=6035950>.
- ANDRÉ, Mathias et MESLIN, Olivier, 2025. Le bonheur est dans le prix : estimation du patrimoine immobilier brut des ménages sur données administratives exhaustives. In : *Documents de travail*. [en ligne]. 10 février 2025. Insee. N° 2025-04. [Consulté le 14 février 2025]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/8351238>.
- BLANCHET, Didier, 2020. Des modèles de microsimulation dans un institut de statistique - Pourquoi, comment, jusqu'où ? In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee. N° N4, pp. 6-22. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://www.insee.fr/fr/information/4497054?sommaire=4497095>.
- BOTHOREL, Éric, 2020. Pour une politique de la donnée. In : *Rapport de la mission sur la politique publique de la donnée, des algorithmes et des codes sources*. [en ligne]. 23 décembre 2020. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://www.info.gouv.fr/rapport/11979-rapport-sur-la-politique-publique-de-la-donnee-des-algorithmes-et-des-codes-sources>.
- CHEN, Tianqi et GUESTRIN, Carlos, 2016. XGBoost: A Scalable Tree Boosting System. In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [en ligne]. 13 août 2016. Association for Computing Machinery. Pp. 785-794. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://dl.acm.org/doi/10.1145/2939672.2939785>.
- COMTE, Frédéric, DEGORRE, Arnaud et LESUR, Romain, 2022. Le SSPCloud : une fabrique créative pour accompagner les expérimentations des statisticiens publics. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 68-87. [Consulté le 10 mars 2025]. Disponible à l'adresse : <https://www.insee.fr/fr/information/6035940?sommaire=6035950>.
- DE PERETTI, Gaël et TOUCHELAY, Béatrice, 2024. Statistiques publiques et débat démocratique : de nouvelles attentes et de nouveaux enjeux (1988-2016). In : *Courrier des statistiques*. [en ligne]. 8 juillet 2024. Insee. N° N11, pp 11-30. [Consulté le 10 mars 2025]. Disponible à l'adresse : <https://www.insee.fr/fr/information/8203034?sommaire=8203072>.

- FREUND, Yoav et SCHAPIRE, Robert E., 1996. *Experiments with a New Boosting Algorithm*. In : *site de AT&T*. [en ligne]. 22 janvier 1996. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d186abec952c4348870a73640bf849af9727f5a4>.
- GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la *data science* et la recherche scientifique. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://www.insee.fr/fr/information/4254227?sommaire=4254170>.
- GIVORD, Pauline, 2014. Méthodes économétriques pour l'évaluation de politiques publiques. In : *Économie et Prévision*. [en ligne]. Direction générale du Trésor. N° 204-205, pp. 1-28. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://doi.org/10.3917/ecop.204.0002>.
- HELFENSTEIN Xavier, 2022. La base permanente des équipements (BPE). Une source statistique singulière et constamment en mouvement. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 131-148. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://www.insee.fr/fr/information/6665194?sommaire=6665196>.
- INSPECTION GÉNÉRALE DES FINANCES, 2024. *Rapport d'activité de l'année 2023*. [en ligne]. Mai 2024. [Consulté le 10 mars 2025]. Disponible à l'adresse : <https://www.igf.finances.gouv.fr/sites/igf/accueil/nos-activites/nos-rapports-dactivite.html>.
- KOUMARIANOS, Heïdi, LEFEBVRE, Olivier et MALHERBE, Lucas, 2024. Les appariements : finalités, pratiques et enjeux de qualité. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2024. Insee. N° N11, pp 117-139. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://www.insee.fr/fr/information/8203044?sommaire=8203072>.
- MARTIN, Olivier, 2023. *Chiffre*. Anamosa, collection le mot est faible. ISBN 978-2-38191-054-3. 11 mars 2023.
- PRADA, Michel 2012. Le métier d'inspecteur : de la vérification à l'évaluation, 1946-2009. In : *Dictionnaire historique des inspecteurs des Finances 1801-2009*. [en ligne]. IGPDE. Pp. 121-126. [Consulté le 10 mars 2025]. Disponible à l'adresse : <https://books.openedition.org/igpde/3623>.
- RIVIÈRE, Pascal, 2020. Qu'est-ce qu'une donnée ? – Impact des données externes sur la statistique publique. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 114-131. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://www.insee.fr/fr/information/5008707?sommaire=5008710>.
- SERVICE DES DONNÉES ET DES ÉTUDES STATISTIQUES et OBSERVATOIRE NATIONAL DES RISQUES NATURELS, 2024. *Chiffres clés des risques naturels - Édition 2023*. [en ligne]. 30 janvier 2024. [Consulté le 18 décembre 2024]. Disponible à l'adresse : <https://www.statistiques.developpement-durable.gouv.fr/chiffres-cles-des-risques-naturels-edition-2023>.
- VOLLE, Michel, 1984. *Le métier de statisticien*. Economica. ISBN 2-7178-0824-8.