RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

SDES
Statistique publique
de l'énergie, des transports,
du logement et de l'environnement

La méthode d'estimation à date réelle (DR++) des logements autorisés et mis en chantier

Notice méthodologique

(dernière mise à jour : septembre 2025)

Introduction

Le suivi conjoncturel du marché immobilier est essentiel pour les politiques publiques du logement et les nombreux acteurs intéressés par ce domaine. Dans ce contexte, il est indispensable de disposer d'estimations mensuelles de qualité sur le volume total de logements autorisés et mis en chantier. Compte-tenu des délais parfois importants de remontées des informations sur les autorisations et mises en chantier relatives à un mois donné, il est nécessaire pour de disposer d'une estimation avancée des valeurs manquantes. C'est l'objet de la méthodologie d'estimations « à date réelle ». Plusieurs méthodologies d'estimation se sont succédées par le passé, la dernière datant de 2016-2017 avec la mise en place de DR+, avec des aménagements apportés en 2020 dans le contexte de la crise sanitaire. Bien que plus robuste que la méthode antérieure, cette méthode présentait certaines fragilités, en raison notamment d'hypothèses fondées sur l'observation passée pour estimer les valeurs manquantes récentes, hypothèses se révélant assez fortes depuis la crise sanitaire, mais aussi d'une estimation très sensible au taux de collecte depuis 2020.

À compter du mois de septembre 2025, pour la diffusion des données portant jusqu'au mois d'août 2025, une nouvelle méthodologie a été développée. Celle-ci repose sur une meilleure exploitation des données à l'aide d'outils de « machine learning » et sur des modèles plus précis. Cette nouvelle méthodologie vise à améliorer la fiabilité des estimations, notamment pour les périodes récentes où les incertitudes sont les plus fortes.

Cette note présente cette nouvelle méthodologie et en compare les résultats à ceux de la méthode précédente, mettant notamment en lumière les gains en précision et en stabilité.

Description de la méthodologie

Objectifs et méthodologie actuelle

L'estimation à date réelle est nécessaire, car chaque mois, seulement 80 à 85% des autorisations et surtout 45 à 50 % des mises en chantier sont enregistrées dans Sitadel. Ce taux augmente progressivement au fil du temps grâce à la collecte de données retardataires. Pour disposer d'une estimation fiable des résultats des derniers mois, il est indispensable d'estimer les données manquantes et ainsi de produire une estimation dite « à date réelle ».

Depuis 2016, une méthodologie d'estimation « à date réelle » a ainsi été développée et mise en œuvre par le SDES. A compter de l'année 2020, cette méthodologie, appelée DR+, a connu un aménagement pour tenir compte des modifications de comportements induites par la crise sanitaire. Initialement, l'estimation « en date réelle » s'appuyait sur les permis autorisés et des hypothèses relatives au taux d'annulation des projets et aux délais moyens entre l'autorisation et la mise en chantier. Concrètement, le modèle estimait d'abord un taux d'annulation en s'appuyant sur l'enquête "anciens permis" pour la période ancienne, puis prolongeait ces taux pour la période récente. Dans un second temps, il établissait une grille de délais à partir des délais effectivement observés sur le passé. Compte-tenu des modifications de comportements intervenues sur les autorisations et mises en chantier dans le contexte de la crise sanitaire, la méthode a été aménagée pour l'estimation au niveau national. Depuis la mi-2020, elle repose sur la collecte des permis effectivement mis en chantier sur le mois concerné, auquel s'applique un redressement fondé sur l'analyse des taux de non-réponse au fil des mois. Les résultats au niveau départemental s'appuient encore sur la méthode initiale mais en calant les résultats sur les estimations nationales. Les données sont ensuite révisées une fois par an grâce à l'enquête sur les anciens permis, qui consolide les informations à N-4 et améliore ainsi les estimations des années suivantes.

Cette méthode, fondée sur des données assez anciennes, repose sur des hypothèses de permanence des comportements assez fortes et peut générer des révisions relativement importantes sur les points les plus récents de la série, notamment en cas de changement de tendance comme cela a été constaté lors de la crise sanitaire. L'aménagement fait en 2020 permet de mieux tenir compte des mouvements récent mais dépend des taux de collecte qui n'informent pas toujours sur l'activité de construction.

Principe de la nouvelle modélisation

La nouvelle modélisation vise à limiter les révisions grâce à une intégration plus complète des données prises en compte. Si cette méthode utilise globalement les mêmes informations que la précédente, elle utilise un algorithme d'apprentissage automatique (XGBoost, basé sur la technique de gradient boosting) qui combine plusieurs modèles simples pour améliorer progressivement la précision des estimations des valeurs manquantes récentes. Elle est particulièrement efficace en raison de sa puissance de calcul et de sa gestion des relations complexes entre les variables, des interactions non linéaires et des données partielles ou bruitées.

Plus précisément, les données passées utilisées par le modèle d'entraînement sont calculées au niveau départemental et portent sur des mois passés où quasiment toutes les autorisations et les mises en chantiers sont connues. Pour chaque observation (département*mois), sont ainsi connus le nombre de logements effectivement mis en chantier (ou autorisés), considéré comme la valeur réelle et fiable reconstituée ex-post une fois toutes les informations remontées, ainsi que la chronique du nombre de mises en chantiers (ou autorisations) relatives à cette date d'observation qui ont été reçues au fil des mois. Ces informations permettent au modèle d'apprendre la relation entre les caractéristiques

disponibles — telles que le flux de logements reçus, la date, le département et les variables macroéconomiques (telles que l'indice du coût de la construction) — et le nombre réel de chantiers démarrés. Grâce à cette base d'apprentissage de données réelles, XGBoost peut ensuite appliquer ces relations aux périodes récentes pour estimer les mises en chantier ou les autorisations.

Méthodologie pour les autorisations

Pour estimer les autorisations, les modèles ont été entraînés sur les données consolidées jusqu'à l'année N-2, car au bout de deux ans, près de 100% des autorisations sont reçus.

Les « hyperparamètres » utilisés pour configurer le modèle de machine learning sur les autorisations ont été définis en fonction de leur performance dans l'estimation des autorisations de l'année 2023. Pour cette estimation, les modèles ont été entraînés sur les données consolidées jusqu'à 2021 et la période 2021 à 2023 a été utilisée comme jeu de test afin de vérifier si le modèle était capable de prédire correctement le nombre de logements autorisés réels à partir des informations disponibles à l'époque. Une fois les hyperparamètres réglés, le modèle a également été entraîné pour prédire les autorisations des années 2020, 2021, 2022 et 2024.

Pour estimer les autorisations, non encore connues de 2025, le modèle s'appuiera sur des données d'entraînement consolidées antérieures à 2023.

Méthodologie pour les mises en chantier

Pour estimer les mises en chantier, les modèles ont été entraînés sur les données consolidées jusqu'à l'année N-4, car c'est seulement au bout de quatre ans, avec l'enquête « ancien permis » que l'on dispose d'observations fiables.

Les « hyperparamètres » utilisés pour configurer le modèle de machine learning sur les mises en chantier ont été définis en fonction de leur performance dans l'estimation des mises en chantier de l'année 2020. Les modèles ont été entraînés sur les données consolidées jusqu'à 2016, et la période 2016 2020 a été utilisée comme jeu de test afin de vérifier si le modèle était capable de prédire correctement le nombre de logements mis en chantier réels à partir des informations disponibles à l'époque. Une fois les hyperparamètres réglés, le modèle a également été entraîné pour prédire les mises en chantier relatives à 2021 et 2024.

Pour estimer les mises en chantier non encore connues de 2025, le modèle s'appuiera sur des données d'entraînement consolidées antérieures à 2021.

Résultats obtenus

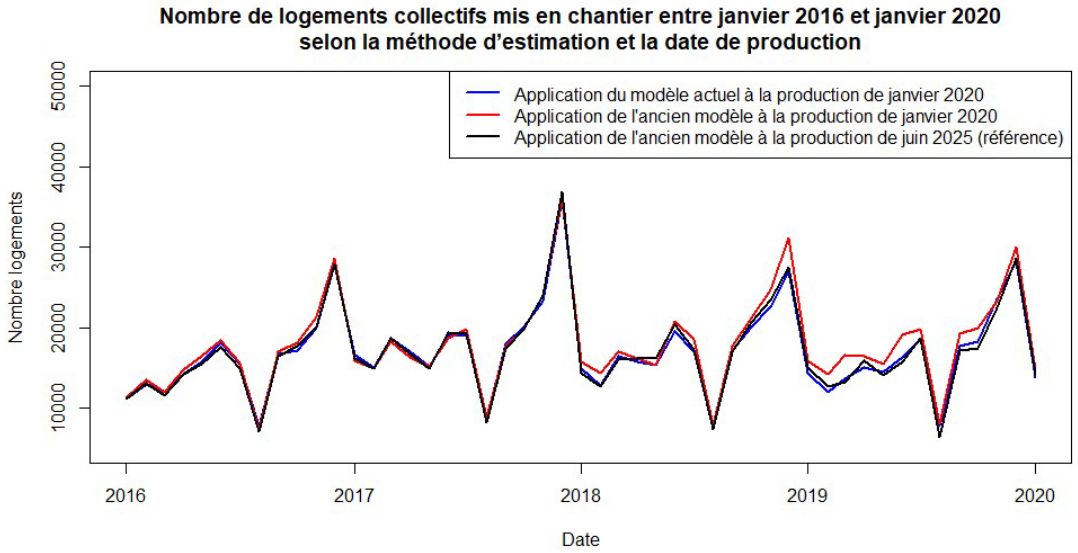
Résultats pour les mises en chantier

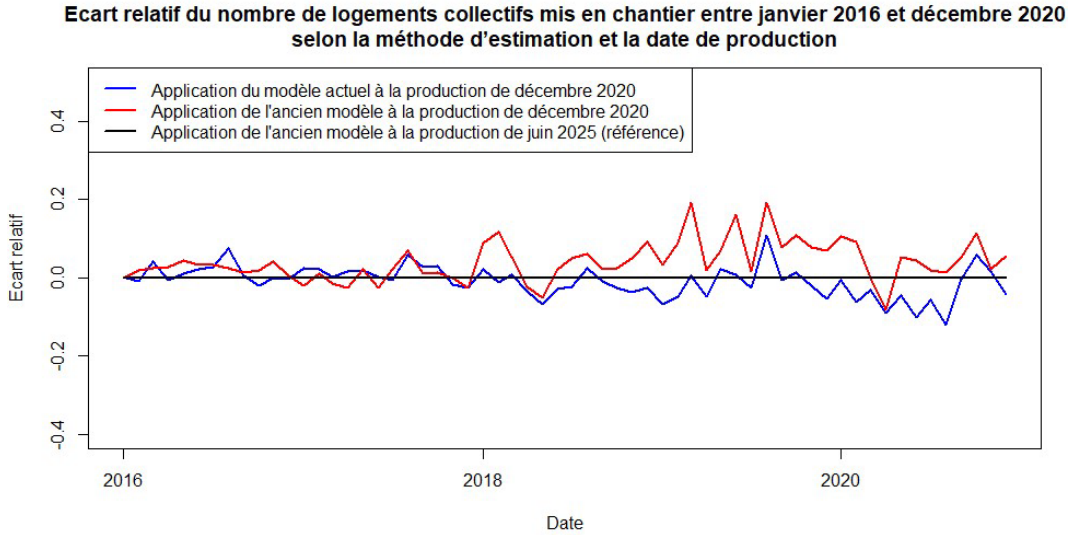
Les résultats présentés ci-dessous portent sur les logements collectifs, qui représentent une part importante de l'ensemble des logements créés. L'analyse se concentre sur la période de janvier 2016 à janvier 2020, choisie car elle offre suffisamment de recul pour considérer les données de référence comme stabilisées et représentatives de la réalité. Ces données servent donc de base de comparaison pour évaluer les performances de la méthode actuelle à partir de XGBoost. Les résultats de 2024 et 2025

ont néanmoins été ajoutés afin d'évaluer la robustesse de la modélisation, en comparant les révisions d'un mois de production à l'autre.

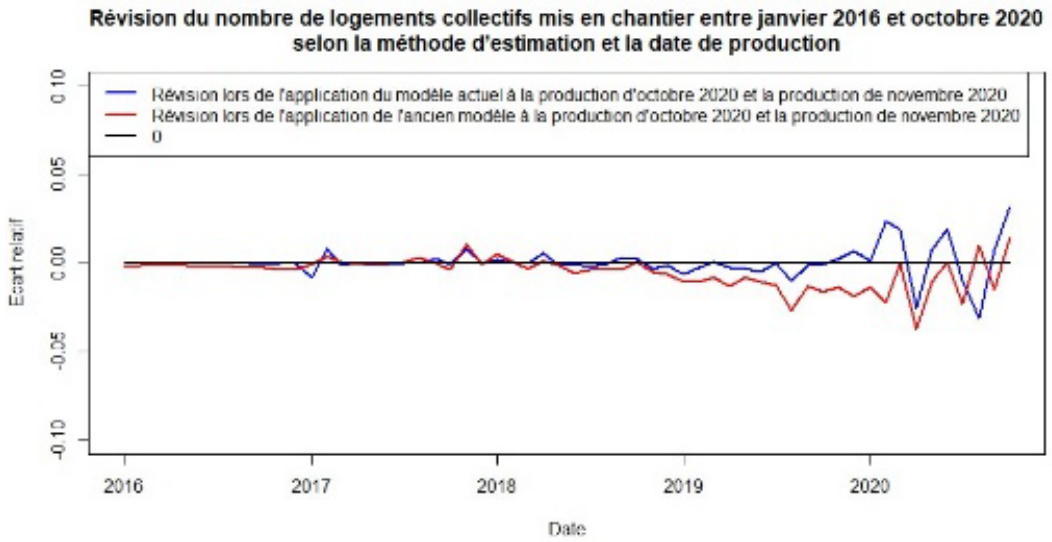
Le premier graphique présente le nombre de logements mis en chantier sur la période 2016-2020, avec les données consolidées connues en juin 2025 (données retardataires et résultats de l'enquête anciens permis intégrés, en noir) et avec l'ancienne et la nouvelle méthodologie d'estimation à partir des seules données connues en janvier 2020. L'objectif est de voir si la nouvelle méthode (en bleu) permet de mieux s'approcher des données réelles (en noir) que l'ancienne méthode (en rouge). Le graphique montre que la nouvelle méthode s'approche davantage des résultats réels, ceci pour l'ensemble de la production.

Le deuxième graphique montre les écarts à valeur réelle entre la nouvelle et l'ancienne méthodologie. Les écarts sont plus faibles avec la nouvelle méthode qu'avec l'ancienne à date de production du dernier point.





Concernant les révisions, XGBoost reste stable dans ses prédictions tout au long de la période de production, à l'exception de la fin de la période mais il est normal d'avoir des révisions sur les derniers mois.



Comparaison des révisions entre l'ancien et le nouveau modèle

Les tableaux ci-dessous comparent les révisions du nombre de logements totaux mis en chantier au moment de l'intégration de l'enquête anciens permis en février 2025 (sur les données de janvier 2025) entre la nouvelle méthodologie (DR++) et l'ancienne (DR+) pour la période de 2020 à 2024. Les révisions sont stables et relativement faibles pour DR++, tandis qu'elles sont plus importantes pour DR+ sur les deux dernières années.

DR++ : révision du nombre de logements mis en chantier sur la période 2020-2024

Année de mise en chantier	Résultats à fin janvier 2025	Résultats à fin février 2025	Révision en % (1)
2020	366 000	370 200	+1,1%
2021	407 600	414 100	+1,6%
2022	382 200	387 900	+1,5%
2023	286 000	293 300	+2,6%
2024	265 700	270 800	+1,9%

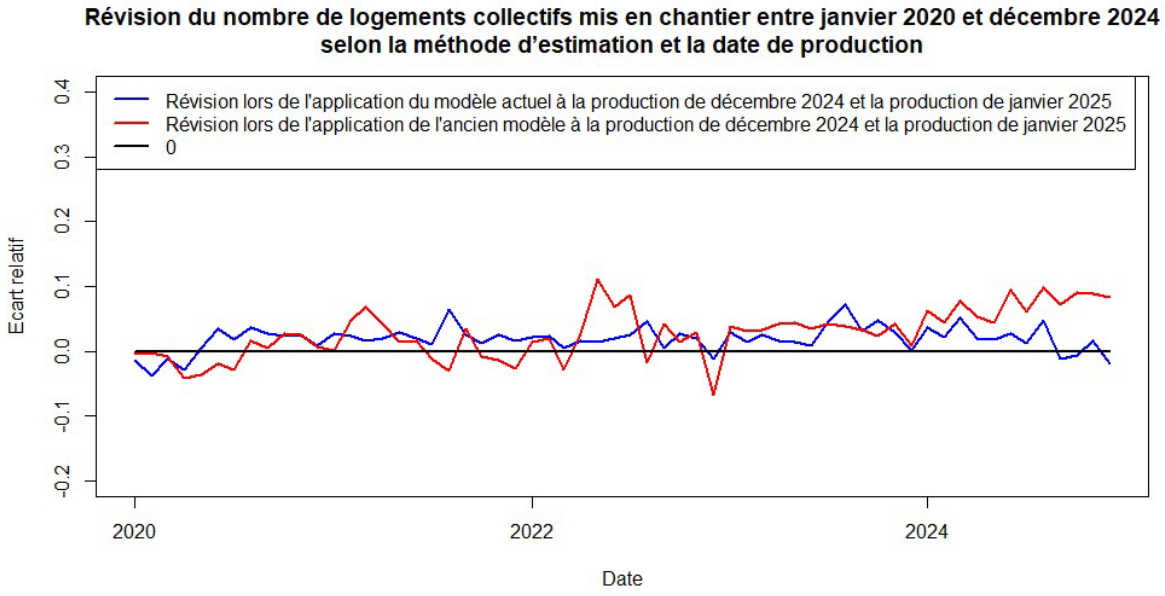
(1) Estimation à fin février 2025 par rapport à l'estimation à fin janvier 2025

DR+ : révision du nombre de logements mis en chantier sur la période 2020-2024

Année de mise en chantier	Résultats à fin janvier 2025	Résultats à fin février 2025	Révision en % (1)
2020	368 700	370 200	+0,4%
2021	410 200	415 400	+1,3%
2022	393 800	398 900	+1,3%
2023	296 000	307 700	+4,0%
2024	263 100	290 800	+10,5%

(1) Estimation à fin février 2025 par rapport à l'estimation à fin janvier 2025

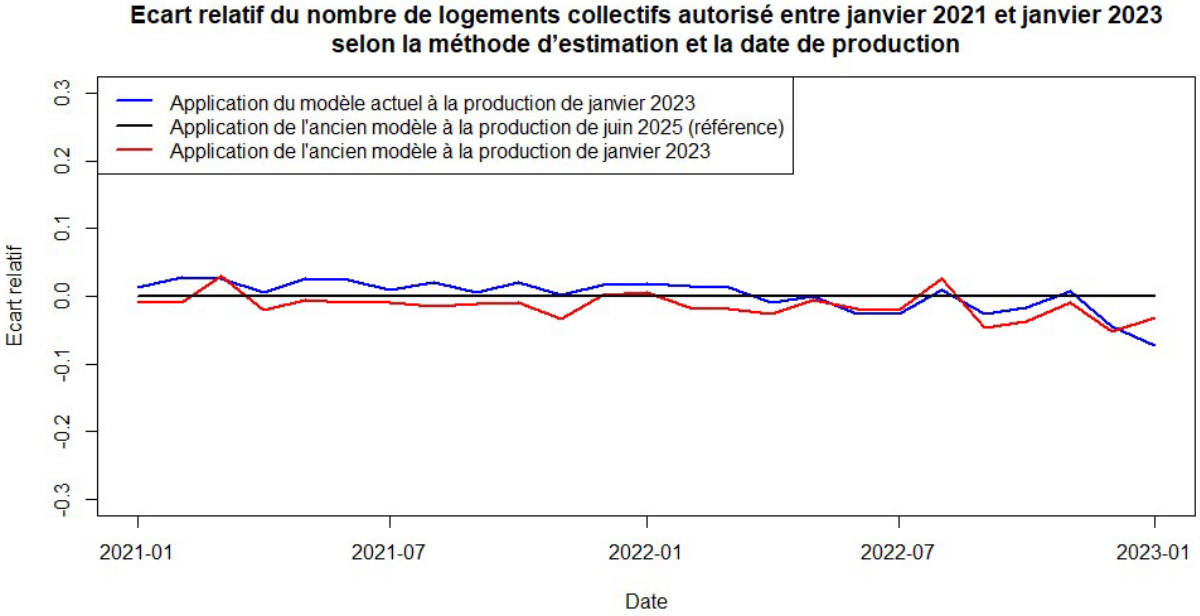
Le graphique ci-dessous détaille les effets de l'intégration de l'enquête "anciens permis" de 2020 sur les estimations en date réelle, au mois le mois, tant pour l'ancienne méthodologie (DR+) que pour la nouvelle (DR++). Ce graphique confirme que l'impact des révisions est nettement plus marqué dans le cas de DR+, où l'intégration de l'enquête dans la modélisation amplifie significativement les ajustements, tandis que pour DR++, les révisions demeurent plus limitées.



Résultats pour les autorisations

De façon analogue, pour les autorisations, le premier graphique montre l'écart entre la réalité et les deux modélisations (en bleu le nouveau modèle et en rouge l'ancien modèle) pour la période 2021-2023, avec des estimations fondées sur les dernières données connues en janvier 2023.

Les performances entre les deux modélisations apparaissent assez équivalentes.



Commissariat général au développement durable

Service des données et études statistiques
Sous-direction des statistiques du logement et de la construction
Tour Séquoia - 92055 La Défense cedex
Courriel : diffusion.cgdd@developpement-durable.gouv.fr
www.statistiques.developpement-durable.gouv.fr