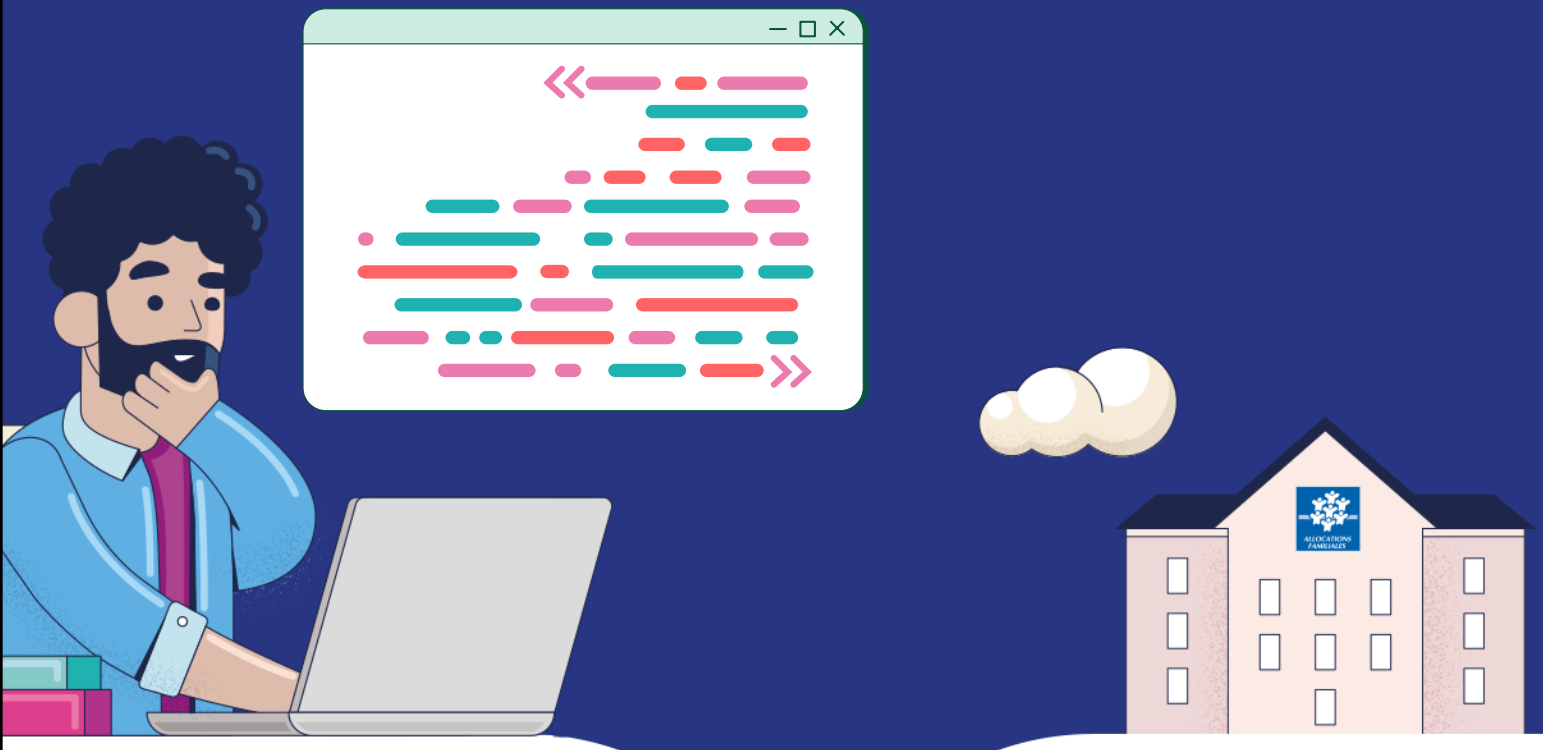
ALLOCATIONS FAMILIALES
Caisse nationale

caf·fr

Présentation détaillée du « Datamining données entrantes » 2026



Note détaillée

Janvier 2026

Présentation détaillée du « Datamining Données entrantes » 2026

Le « datamining données entrantes » (ou DMDE) est un outil d'aide au ciblage de certains contrôles réalisés par les agents des caisses d'Allocations familiales (Caf)[1]. Ce ciblage repose sur une cotation du risque d'indus. Les autres contrôles visant à limiter les erreurs déclaratives sont soit aléatoires, soit déclenchés par le repérage d'incohérences (par exemple avec des données de partenaires comme France travail ou la DGFIP).

Chaque mois, les Caf versent des prestations à 13,8 millions de foyers (prestations familiales, aides au logement, revenu de solidarité active, prime d'activité, allocation aux adultes handicapés…). Les Caf veillent à ce que ces aides soient versées au « juste droit », c'est-à-dire que le « bon » montant soit versé au « bon » moment, tels que définis par la législation, et à respecter l'égalité de traitement des allocataires se trouvant dans la même situation. Comme les informations permettant de calculer les montants à verser reposent en grande partie sur les déclarations des allocataires (sur leurs ressources, leur situation professionnelle, la composition de leur foyer, des informations sur leur logement…) et que la législation est complexe, des erreurs déclaratives des allocataires sont possibles. Cela conduit à des situations d'indu, où l'allocataire a trop perçu, et ou de rappel, où l'allocataire n'a pas assez perçu. Les Caf mènent donc une politique de contrôle pour détecter ces erreurs et les rectifier.

Place du DMDE dans la politique de contrôle de la Branche Famille

Les contrôles réalisés par les Caf sont autant que possible globaux, que ce soit dans le ciblage ou dans les modalités de réalisation. Cela veut dire qu'il n'y a pas de ciblage par prestation (sauf cas particulier), et que le protocole de contrôle balaie l'ensemble des points de contrôle qui peuvent l'être.

Les Caf ont une approche majoritairement par risque. Les dossiers à contrôler sont identifiés par les Caf selon trois grandes modalités :

- A partir d'incohérences : en cas d'anomalie détectée par le système d'information des Caf, par un gestionnaire ou à la suite d'un signalement fait par un partenaire ou un tiers
- A partir d'une cotation de risque : les contrôleurs utilisent alors un algorithme d'aide à la décision, le modèle « Datamining Données Entrantes » (DMDE), qui a pour objectif

[1] En parallèle, la branche Famille mobilise aussi les données dont elle dispose pour le ciblage de campagnes d'accès aux droits, notamment à la prime d'activité.

de les aider à cibler les contrôles sur les dossiers allocataires les plus à risque d'avoir des indus (trop-perçus de prestations sociales) élevés.

- De façon aléatoire : « l'opération de paiement à bon droit (OPBD) » est effectuée chaque année sur un échantillon aléatoire d'allocataires de 6000 dossiers, ce qui permet de modéliser la cartographie du risque d'erreur sur l'ensemble de la population allocataire.

Par ailleurs, à l'exception des cibles issues du Service National de Lutte contre la Fraude à Enjeux (SNLFE), tous les contrôles ont comme objectif la recherche d'erreurs déclaratives, et non des fraudes, quand bien même certains contrôles peuvent permettre d'en détecter – la fraude se démarquant de l'erreur par son caractère intentionnel.

Encadré - La place du DMDE dans la politique de contrôle en 2024

En 2024, les Caf ont réalisé 31,5 millions de contrôles auprès de 6,4 millions d'allocataires, soit près d'un allocataire sur deux. Ceux-ci ont permis de détecter au total 1,68 milliard d'euros d'indus (trop-perçus de prestations) et de rappels (moins-perçus, c'est-à-dire des régularisations en faveur de l'allocataire). Un quart des montants régularisés sont des rappels. 92% de ces contrôles sont des régularisations automatiques issues d'échanges de données avec d'autres organismes, notamment France Travail. 8% sont des contrôles réalisés par des agents des Caf. Ils sont réalisés en très grande majorité sur pièces (2,5 millions de contrôles en 2024) ; cependant 91 000 contrôles sont réalisés sur place par l'un des 700 agents assermentés et agréés des Caf, qui bénéficient de prérogatives spécifiques permettant des contrôles plus approfondis.

166 000 contrôles réalisés en 2024 sont issus du DMDE, soit moins de 1% de l'ensemble des contrôles réalisés par les Caf. Un tiers de ces contrôles sont réalisés sur place et les deux-tiers restant sur pièce. Ces contrôles issus du DMDE représentent un peu moins de la moitié (46%) des contrôles sur place et moins de 5% des contrôles sur pièce. 7% des contrôles DMDE donnent lieu à une qualification en fraude, à l'issue d'un protocole formalisé. 26% des fraudes sont détectées par le DMDE, la première catégorie de contrôles permettant de lutter contre la fraude étant ceux menés par le Service national de lutte contre la fraude à enjeux (30%).

La réforme de la « solidarité à la source » change radicalement le paysage des erreurs déclaratives pour le revenu de solidarité active (RSA) et la prime d'activité et, conséquemment, de la politique de contrôle. Le premier volet de cette réforme, mis en œuvre à partir de juillet 2023, est le déploiement du montant net social comme montant de référence à déclarer. Inscrit sur les fiches de paie des salariés et sur les relevés de prestations sociales, ce montant correspond aux ressources à déclarer pour le calcul du RSA et de la prime d'activité. Les allocataires ont l'obligation de reporter ce montant dans leur déclaration trimestrielle de ressources (DTR), à compter des revenus de janvier 2024. Le second volet de la réforme est le pré-remplissage des DTR avec les données récupérées directement des déclarations des employeurs et des organismes de protection sociale. Déployé en avance de phase dans cinq Caf

de « présérie » entre octobre et décembre 2024 (les Alpes-Maritimes, l'Aube, l'Hérault, les Pyrénées-Atlantiques et la Vendée), il a été généralisé aux autres départements à partir de mars 2025. Le calcul des ressources à déclarer, souvent complexe, a ainsi été simplifié pour les salariés et les bénéficiaires de revenus de remplacement (allocations chômage, indemnités journalières d'assurance maladie, rentes d'accident du travail…), permettant d'éviter des erreurs déclaratives ou des omissions. La solidarité à la source devrait donc permettre de réduire considérablement le risque d'indus et de rappels de prime d'activité et de RSA.

Les contrôles portant par nature sur le passé, avec un recul de deux ans (hors fraude), les effets de la solidarité à la source sur les actions de contrôles ne seront pleinement montés en charge qu'en mars 2027. Cette réforme constitue en outre une opportunité pour la Branche famille de mieux articuler les contrôles et limiter la réitération. Il est en particulier prévu d'allonger la durée minimale entre deux contrôles, de 12 ou 18 mois aujourd'hui à 24 mois à partir de mars 2026. Compte tenu de la disparition de nombreux contrôles devenus obsolètes avec la solidarité à la source, les capacités de contrôle des Caf devraient être redéployées notamment vers les contrôles datamining et la lutte contre la fraude à enjeux (fraudes à résidence, déclaration des ressources issues de l'économie des plateformes, schémas d'usurpations…).

Une refonte importante du DMDE en 2025

Le datamining « données entrantes » a été rénové en 2025. Il est déployé dans les Caf en janvier 2026. Cette refonte du modèle était nécessaire d'une part pour adapter la modélisation aux évolutions réglementaires récentes. En effet, le modèle DMDE précédent (« DMDE 2018 ») avait été déployé dans les Caf depuis fin 2019 ; ce DMDE avait été conçu en 2018 sur la base de données portant sur la période allant d'octobre 2015 à mars 2017. Il ne prend donc pas en compte les réformes de la prime d'activité de 2019[2] et des aides aux logements de 2021 ni la solidarité à la source, expérimentée par 5 Caf depuis octobre 2024 et généralisée depuis mars 2025.

D'autre part, cette refonte d'ampleur repose sur la mise en place d'une « démarche éthique dès la conception » dans laquelle la Cnaf s'est engagée pour la production de ses algorithmes. Celle-ci est mise en œuvre pour la première fois pour la construction de ce nouveau DMDE 2026[3]. Elle s'inscrit dans le cadre d'un plan d'actions de la Cnaf relatif au datamining, décliné selon deux axes :

1. Le déploiement d'un cadre de gouvernance de nos usages des données et des algorithmes, couvrant un périmètre plus large que les seuls modèles de datamining, et qui pourra intégrer à terme les projets faisant appel aux techniques d'intelligence artificielle ;

[2] La création de la prime d'activité en 2016 n'est donc que partiellement prise en compte.

[3] Dans cette note technique, le nouveau DMDE déployé début 2026 dans les Caf est parfois nommé « DMDE 2025 », 2025 correspondant à l'année de sa construction (tout comme 2018 correspondant à l'année de construction du modèle précédent).

2. Le renforcement de la transparence et de l'information. Cet effort de transparence concerne à la fois la relation à l'allocataire dans le cadre de la gestion de son dossier mais aussi un niveau d'information plus général à destination du grand public.

Dans le cadre du premier axe, la Cnaf a mis en place un comité d'éthique des usages des données, des algorithmes et de l'intelligence artificielle. Installé en mars 2025, il appuie la Cnaf dans sa démarché d'identification des risques de nature éthique et des moyens pour les atténuer, afin de garantir la protection des droits des personnes (CAF - Le comite d'éthique). Les différentes étapes de construction du DMDE 2025 ont fait l'objet d'échanges avec le comité d'éthique, sur la base d'une charte éthique encadrant le développement et l'usage des outils algorithmiques et d'intelligence artificielle, dont la Cnaf s'est dotée.

Dans un souci de transparence et pour permettre des débats informés, la Cnaf met à disposition du grand public début 2026 une présentation du nouveau modèle DMDE.

Le déploiement de ce modèle a fait l'objet d'actions d'accompagnement menées auprès des directions des Caf, des contrôleurs et de leurs managers pour l'utilisation du modèle, qui seront notamment sensibilisés aux potentiels biais discriminatoires du modèle. Par ailleurs, le modèle sera mobilisé pour bâtir une feuille de route en appui aux actions de prévention des indus.

1 Méthode de construction de l'algorithme

1.1 Les données utilisées pour la construction de l'algorithme

Les données utilisées sont issues du système d'information de la CNAF, c'est-à-dire des données produites dans le cadre de l'activité de gestion des prestations par les Caf (composition familiale et âge des enfants, ressources, prestations reçues, etc), à une exception près.

En plus des données des Caf, le modèle 2026 mobilise en effet les données du dispositif de ressources mensuelles (DRM) . Il s'agit des données utilisées pour le pré-remplissage des déclarations trimestrielles de ressources du RSA et de la prime d'activité depuis la réforme de la solidarité à la source ; elles servent aussi au calcul des aides au logement depuis 2021. Les données du DRM proviennent des données de paie déclarées par les employeurs (déclaration sociale nominative) et des données des organismes de protection sociale (pour les revenus de remplacement comme les allocations chômage par exemple).

Dans le DMDE, les données du DRM sont utilisées pour mesurer les effets de la variable sur les « signalements » (corrections par l'allocataire de la donnée préremplie avec le DRM dans sa DTR) et calculer le coefficient correspondant de la régression logistique[2]. Elles ne sont pas mobilisées ensuite pour calculer le score de risque d'indus.

Hors DRM donc, aucune collecte supplémentaire spécifique n'est mise en place et le modèle ne repose pas sur des données externes, ce qui réduit le risque d'introduire un biais dans les données.

La construction du modèle DMDE repose donc sur des données fiables et maitrisées par la Cnaf, puisque ce sont les données utilisées pour calculer les prestations. Elle repose sur des contrôles aléatoires d'un échantillon représentatif des allocataires. Plus précisément, le

modèle DMDE repose sur les données de l'enquête « Opération paiement à bon droit » (OPBD) 2022 et 2023, dernières données disponibles au début des travaux sur le modèle, soit un échantillon représentatif de 12 000 dossiers couvrant la période allant d'octobre 2021 à mars 2023. L'enquête OPBD est menée chaque année sur un échantillon représentatif des allocataires des Caf, constitué de 6000 dossiers Caf qui sont donc contrôlés de façon aléatoire. Il n'y a donc pas de biais lié au fait que les dossiers contrôlés seraient justement des dossiers « à risque d'indus ». La représentativité de l'échantillon de l'enquête OPBD fait l'objet d'un examen minutieux de la Cour des comptes dans le cadre de la certification des comptes de la branche Famille. La Cour est notamment particulièrement attentive au respect du protocole de contrôle pour cet échantillon.

L'ensemble de ces éléments contribue à limiter les biais algorithmiques liées aux données.

1.2 Choix de la modélisation statistique

Une modélisation sous forme de régression logistique a été choisie pour favoriser la compréhension et la lisibilité du modèle, dans un objectif de transparence et d'explicabilité mais aussi de faible consommation de ressources, conformément aux principes énoncés dans la charte éthique. Une fois construit, le modèle est figé pour plusieurs années ; il n'est pas actualisé au fil des contrôles, ce qui écarte le risque de boucles de rétroaction qui peuvent créer des biais ou les renforcer.

1.3 Choix de la « cible » du modèle

La « cible » du modèle est ce qu'il cherche à prédire, c'est-à-dire le type de dossiers que le modèle va chercher à caractériser pour pouvoir les repérer parmi l'ensemble des dossiers des allocataires et les contrôler en priorité. Dans le DMDE 2018, la cible était le fait que le dossier de l'allocataire comporte un indu d'un montant total d'au moins 600 euros et courant sur une période d'au moins 6 mois, soit un indu long et important.

Le choix d'une cible portant sur les seuls indus pour orienter le choix des dossiers à contrôler se justifie du fait de la prédominance des indus sur les rappels et parce qu'il reste beaucoup moins de rappels non corrigés au bout de deux ans (10%) que d'indus non corrigés (60%)[4]. Cela conduit à privilégier un ciblage centré sur les indus pour orienter la priorisation des contrôles. À noter toutefois que 25 % des montants régularisés lors des contrôles correspondent malgré tout à des rappels de prestations. L'introduction d'une durée minimale de 6 mois vise à éviter de contrôler des dossiers dont les indus se seraient résorbés sans l'intervention des services de contrôle des Caf.

Après analyse de différentes alternatives, la même « cible » a été retenue pour le DMDE 2025. A noter cependant que le seuil de 600 euros représente moins en valeur réelle (en euros constants) en 2025 qu'en 2018 puisque les prestations ont augmenté sur la période, en lien avec l'inflation : la population potentiellement ciblée sera plus large que dans le modèle précédent.

[4] Ces chiffres sont calculés par extrapolation à partir de l'enquête OPBD.

Sur l'échantillon des OPBD 2022 et 2023, 18% des dossiers allocataires sont dans la cible car ils présentent un indu « long et important ».

1.4 Sélection des variables du modèle

Le nouveau modèle est déployé dans les Caf à partir de janvier 2026. La période contrôlée pouvant remonter jusqu'à deux ans en arrière (davantage en cas de fraude), cela implique la construction de deux modèles, le premier s'appliquant à la période avant préremplissage des déclarations trimestrielles de ressources (entré en vigueur le 1er octobre 2024 dans 5 Caf et le 1er mars 2025 pour le reste des départements), le second couvrant les prestations versées après la réforme.

La méthode de sélection des variables est identique pour les 2 modèles.

Une première étape a consisté à repérer et questionner l'introduction dans le modèle de variables pouvant conduire à des biais discriminatoires ou de données sensibles au sens du RGPD.

En effet, dans le cadre de son activité, la branche Famille de la Sécurité sociale est amenée à traiter des données ou informations sur le fondement desquelles une personne pourrait être victime d'une discrimination pouvant relever des catégories suivantes, visées dans l'article 1 de la loi n°2008-496 du 27 mai 2008 portant diverses dispositions d'adaptation au droit communautaire dans le domaine de la lutte contre les discriminations :

- La situation familiale
- Le sexe
- La nationalité
- Le lieu de résidence
- La « particulière vulnérabilité résultant de la situation économique »
- L'âge
- Le handicap

Pour autant, le recours à ces critères ne constitue pas une discrimination au sens de la loi précitée, lorsque le recours à ces critères est objectivement justifié par un but légitime et lorsque les moyens mis en œuvre pour réaliser ce but sont nécessaires et appropriés. Par ailleurs, le RGPD encadre strictement le traitement de certaines données, qualifiées de « sensibles » (article 9 du RGPD). La loi n°78-17 du 6 janvier 1978 prévoit en outre qu'aucune décision individuelle fondée exclusivement sur un traitement automatisé ne peut porter sur de telles données sensibles.

Le principe retenu pour conserver une variable sensible est d'identifier s'il y a un lien entre la variable et la cible explicable par la réglementation. La variable portant sur le genre du responsable du dossier allocataire a par exemple été écartée car aucun élément de législation des prestations ne repose sur cette caractéristique. De même, l'information sur l'adresse de l'allocataire n'a pas été conservée car il n'y a pas de raison de nature réglementaire qui pourrait justifier que, toutes choses égales par ailleurs, une variable liée à l'adresse puisse expliquer des

risques d'indus plus élevés. En revanche, certains critères liés à l'âge d'un membre du foyer allocataire ont été conservés, uniquement lorsque cela peut s'expliquer par la législation (par exemple, l'âge des enfants…). Les variables liées à la perception d'une allocation liée au handicap (AAH ou AEEH) n'ont pas été écartées d'emblée[5] : c'est une caractéristique sensible, mais son introduction dans le modèle est justifiable du fait de la réglementation. Enfin, en raison de sa sensibilité particulière, la variable sur la nationalité du responsable de dossier et de son conjoint a été retirée d'emblée, même si la réglementation prévoit des conditions spécifiques pour les personnes de nationalité étrangère, qui peuvent être différentes pour les ressortissants de l'Union européenne et les autres, ce qui peut donc conduire à des indus de prestations.

Toujours dans le cadre de notre démarche éthique, les données dites « de comportement » (exemples : contacts des allocataires avec la Caf à travers les connexions à l'espace « Mon Compte » sur le site caf.fr ou par téléphone, échanges de pièces justificatives…) ont été exclues[6], tout comme les informations sur des contrôles ou des contentieux antérieurs.

Après cette étape d'exclusion de variables pouvant conduire à des biais discriminatoires ou de données sensibles au sens du RGPD, la sélection des variables du modèle et de leurs modalités a été réalisée en fonction de critères statistiques (corrélation avec la cible, paramètres de performance statistique du modèle…), et avec un objectif de minimisation du nombre de variables.

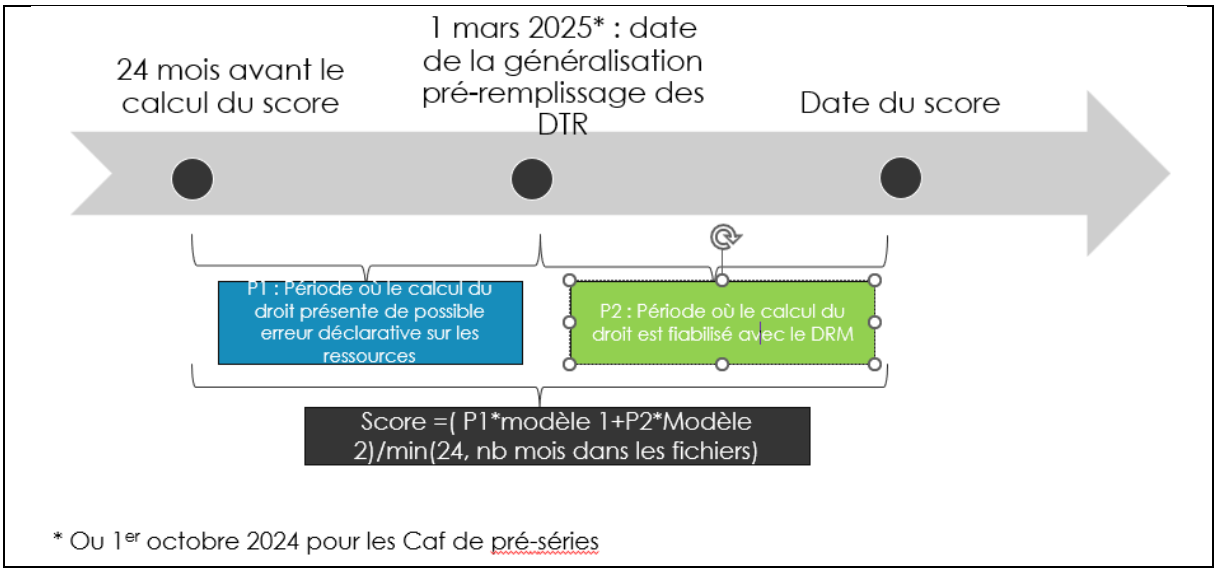
Concernant le second modèle, destiné à couvrir la période concernée par la réforme de la solidarité à la source, une variable supplémentaire a été introduite pour tenir compte des « signalements », c'est-à-dire des corrections par l'allocataire des ressources préremplies dans sa déclaration trimestrielle de ressources (DTR), lors que ces corrections n'ont pas été vérifiées par la cellule dédiée des services de la Cnav et des Urssaf.

2 Description de l'algorithme

Deux modèles ont été construits. Le premier modèle (modèle 1) ne prend pas en compte la réduction des erreurs en lien avec la mise en place du préremplissage. Le second modèle 2 prend en compte la réduction des erreurs en lien avec le préremplissage. Les deux modèles vont cohabiter le temps de la montée en charge du pré-remplissage. Le calcul score de risque associé au dossier allocataire pour les 24 derniers mois sera une combinaison au *prorata temporis* des scores calculés avant solidarité à la source avec le modèle 1 et des scores calculés après solidarité à la source avec le modèle 2 (schéma). Le poids du modèle 2 va augmenter au fil des mois, jusqu'à la disparition complète du modèle 1 en mars 2027.

[5] Elles n'ont finalement pas été retenues dans le modèle final car elles ne ressortaient pas après la phase de sélection sur critères statistiques.

[6] A l'exception de la variable sur les « signalements » dans le modèle 2 (voir *infra*).



L'annexe présente le détail des variables et des coefficients des 2 modèles, ainsi que l'interprétation des liens entre les variables et la « cible » des modèles. Il est notable qu'au final, aucune variable sur la perception de prestations liées au handicap n'a été retenue à l'issue de la phase de sélection sur critères statistiques.

2.1 Le modèle 1 (avant pré-remplissage)

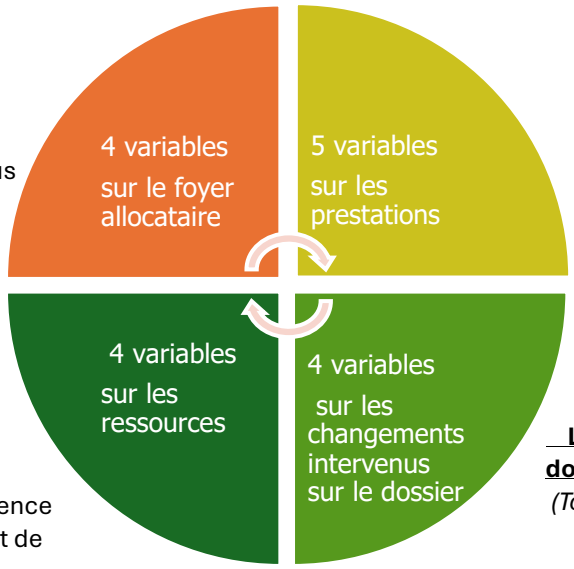
Le modèle 1 ne prend pas en compte la réduction des erreurs déclaratives du fait du préremplissage des déclarations trimestrielles de ressources pour le RSA et la prime d'activité ; il permet d'estimer correctement le risque d'indus sur la période antérieure à la solidarité à la source.

Le modèle 1 comporte 17 variables.

Schéma 1 : représentation des effets des variables du modèle 1 (sans préremplissage)

<u>Informations sur le dossier</u>
(Toutes choses égales par ailleurs)

- Accroissement du risque lorsque le conjoint est sans activité
- Accroissement du risque en présence d'un enfant étudiant
- Accroissement du risque en présence d'enfants âgés de plus de 19 ans
- Réduction du risque en présence de dossier géré par une tutelle



<u>Les variables sur les prestations</u>
(Toutes choses égales par ailleurs)

- Accroissement du risque lorsque le foyer bénéficie de la prime d'activité
- Accroissement du risque lorsque le foyer bénéficie du RSA,
- Accroissement du risque lorsque le foyer bénéficie du complément familial
- Plus le montant des prestations augmente plus le risque augmente
- Réduction du risque en présence d'allocations familiales

<u>Les variables sur les ressources</u>
(Toutes choses égales par ailleurs)

- Accroissement du risque en présence de neutralisation ou d'abattement de ressources
- Réduction du risque lorsque les revenus sont supérieurs à 1,5 Smic
- Accroissement du risque en cas de présence de pension alimentaire
- Accroissement du risque lors de changement dans la pratique d'une activité non salariée dans les 24 derniers mois

<u>Les variables sur les changements sur le dossier</u>
(Toutes choses égales par ailleurs)

- Accroissement du risque en présence d'un fait générateur de confirmation d'un logement étudiant
- Présence d'un fait générateur "PROAC" PROjet d'ACCompagnement
- Réduction du risque en cas de fait générateur sur les volets sociaux (en lien avec le CMG)
- Accroissement du risque en cas de changements de droit au RSA (aller-retour dans la prestation)

2.2 Le modèle 2 (après pré-remplissage)

Le modèle 2 prend en compte le pré-remplissage des déclarations de ressources avec les données du dispositif de ressources mensuelles (DRM). Grâce à la mobilisation des données de ressources issues du DRM lors de la construction du modèle 2, celui-ci pourra prédire le risque d'indus en neutralisant l'effet des écarts entre les ressources déclarées (présentes dans les données des enquêtes OPBD 2022 et 2023 qui sont antérieures à la réforme de la solidarité à la source) et les ressources préremplies (après la réforme).

Le modèle 2 comporte 11 variables, dont une variable relative aux « signalements » réalisés dans le cadre du préremplissage des DTR pour le RSA et la prime d'activité réalisé dans le cadre de la solidarité à la source. Toutes choses égales par ailleurs, le score de risque issu du DMDE 2025 sera majoré d'un facteur 2,9 pour les allocataires ayant réalisé au moins 2 DTR avec un signalement non traité par la cellule experte et dont l'écart mensuel moyen sur le trimestre entre les ressources préremplies et les ressources corrigées est supérieur à 150€ (corrections des ressources à la baisse par l'allocataire).

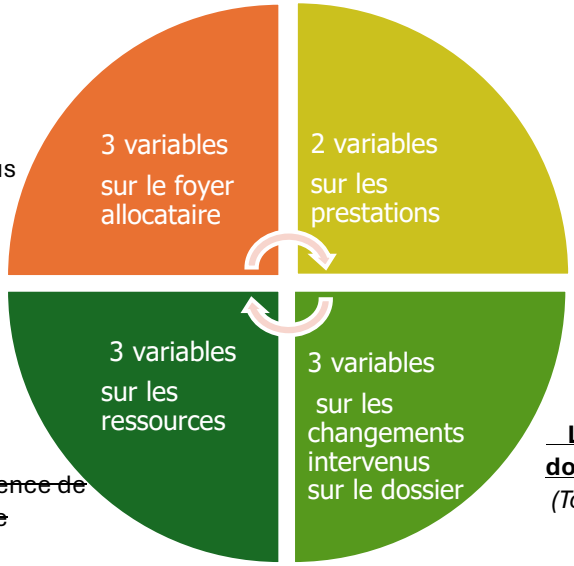
Schéma 2 : représentation des effets des variables du modèle 1 (avec préremplissage)

Informations sur le dossier
(Toutes choses égales par ailleurs)

- Accroissement du risque lorsque le conjoint est sans activité
- ~~Accroissement du risque en présence d'un enfant étudiant~~
- Accroissement du risque en présence d'enfants âgés de plus de 19 ans
- Réduction du risque en présence de dossier géré par une tutelle

Les variables sur les prestations
(Toutes choses égales par ailleurs)

- Accroissement du risque lorsque le foyer bénéficie de la prime d'activité
- ~~Accroissement du risque lorsque le foyer bénéficie du RSA,~~
- ~~Accroissement du risque lorsque le foyer bénéficie du complément familial~~
- Plus le montant des prestations augmente plus le risque augmente
- ~~Réduction du risque en présence d'allocations familiales~~



Les variables sur les ressources
(Toutes choses égales par ailleurs)

- ~~Accroissement du risque en présence de neutralisation ou d'abattement de ressources~~
- Réduction du risque lorsque les revenus sont supérieurs à 1,5 Smic
- ~~Accroissement du risque en cas de présence de pension alimentaire~~
- ~~Accroissement du risque lors de changement dans la pratique d'une activité non salariée dans les 24 derniers mois~~
- Accroissement du risque pour une personne déclarant avoir perçu des revenus d'activité non-salariés
- Accroissement du risque pour un dossier effectuant au moins deux déclarations de ressources avec une correction des montants pré-affichés (signalements sans vérification de la CEM) avec un écart moyen mensuel de sous-déclaration supérieur à 150 €

Les variables sur les changements sur le dossier
(Toutes choses égales par ailleurs)

- Accroissement du risque en présence d'un fait générateur de confirmation d'un logement étudiant
- ~~Présence d'un fait générateur "PROAC" PROjet d'ACCompagnement~~
- Réduction du risque en cas de fait générateur sur les volets sociaux (en lien avec le CMG)
- Accroissement du risque en cas de changements de droit au RSA (aller-retour dans la prestation)

Note de lecture : les variables barrées correspondent aux variables qui sont prises en comptes dans le modèle 1 et qui ne sont plus prises en compte dans le modèle 2. Les variables en violet sont les nouvelles variables prises en compte dans le modèle 2.

4 L'analyse des potentiels biais de discrimination du modèle théorique

Les actions conduites dans le cadre de la démarche éthique lors de la construction du modèle visent à limiter les risques de discrimination, notamment au regard des 26 critères pour lesquels toute discrimination directe ou indirecte est prohibée, une différence de traitement directement en raison d'un de ces critères, ou même indirectement à travers des effets différenciés en raison de ces critères, devant être justifié par un objectif légitime, et être nécessaire et approprié au regard de celui-ci.

4.1 Analyse des potentiels biais conduisant à une discrimination directe

Les variables correspondant aux 26 critères de discrimination ont toutes été écartées d'emblée lors de la construction du modèle et aucune variable ne porte de risque de discrimination directe[7], à l'exception de celles relatives au critère de situation économique. En effet, plusieurs variables du modèle liées au niveau de ressources ou à la perception des prestations reflètent la situation économique du foyer allocataire. Cela parait inévitable puisqu'une large part des prestations versées sont conditionnées aux ressources, en particulier les prestations les plus à risque d'indus. La mobilisation de ces variables apparait ainsi comme nécessaire et appropriée pour répondre à l'objectif de ciblage des contrôles pour réduire les indus de prestations et favoriser leur paiement à bon droit.

Certes le DMDE cible par nature davantage les foyers avec des revenus faibles que ceux plus aisés. Cependant, il cible moins les foyers sans aucune ressource car ils ont moins de risque de se tromper dans leur déclaration de ressources que les foyers avec des faibles ressources, qui peuvent en outre varier fréquemment d'un mois à l'autre. Les foyers avec des ressources nulles seraient davantage contrôlés en cas de sélection aléatoire des dossiers ou si l'on retenait les dossiers avec les montants de prestations les plus élevés.

4.2 Analyse des potentiels biais conduisant à une discrimination indirecte

Examen des effets du modèle théorique

Une analyse des potentiels biais du modèle théorique conduisant à de la discrimination indirecte a été réalisée. Elle repose sur un examen de la sur- ou sous-représentation – au sens statistique – de différentes populations au sein des scores de risques les plus élevés et une identification des variables du modèle jouant de façon significative à la hausse ou à la baisse sur les scores de risque, afin de mieux comprendre les mécanismes à l'œuvre.

Il résulte de ces travaux d'analyse que le recours aux critères et variables retenus apparait objectivement justifié par un but légitime – la détection des situations d'indus et le versement au juste droit. Les moyens pour atteindre ce but, à savoir l'optimisation du ciblage des contrôles s'appuyant sur l'établissement automatisé d'un score de risque de percevoir un indu important

[7] En particulier, l'analyse de la variable *d'Activité du responsable de dossier et de son conjoint* (réalisée à la demande d'un membre du comité d'éthique) conduit à conclure à l'absence de biais de discrimination directe liée à la configuration familiale qui serait introduite par cette variable.

sur la base de ces critères et variables, sont nécessaires dans un contexte de ressources humaines limitées.

Enfin, les moyens mis en œuvre sur la base de travaux statistiques approfondis, d'une méthodologie maîtrisée et explicable (la régression logistique) appliquée à des données représentatives et de qualité, et encadrés par une démarche éthique d'analyse des effets de sous- ou surreprésentation sur des populations déterminées, sont appropriés.

En outre, la mise à jour du modèle en 2025 a fait l'objet d'une inscription au registre des traitements de la Cnaf et d'une étude de conformité au règlement général sur la protection des données (RGPD), avec réalisation d'une analyse d'impact relative à la protection des données (AIPD).

5 Le suivi des effets du modèle en conditions réelles

L'évaluation complète des effets du modèle nécessite de disposer, en plus des statistiques produites sur le modèle « théorique », des données observées en conditions réelles, une fois le modèle déployé dans les Caf, pour tenir compte de l'impact de la façon dont le modèle est utilisé par les contrôleurs.

En effet, les scores de risque sont calculés tous les mois pour tous les allocataires. Pour décider des dossiers allocataires à contrôler, les contrôleurs s'appuient sur une liste des dossiers allocataires classés par ordre décroissant de score de risque, mais aussi sur d'autres informations. Leur choix est aussi guidé par des contraintes d'ordre logistique pour les contrôles sur place : la tournée est planifiée en optimisant les déplacements. Les contrôleurs peuvent donc s'écarter de la liste des plus hauts scores pour décider d'un contrôle, qui ne dépend donc pas du score de façon « automatique ». Par ailleurs, un délai minimal est fixé entre deux contrôles DMDE (18 ou 12 mois, selon la présence ou non d'un impact financier du contrôle ; 24 mois à partir de mars 2026).

Le suivi du modèle portera d'une part sur sa performance afin de vérifier sa bonne adéquation au risque, en lien avec les évolutions de la législation. Plusieurs indicateurs sont utilisés pour évaluer et suivre les performances du modèle : la proportion de dossiers dans la « cible » du modèle 2025 ; la proportion de dossiers contrôlés présentant un indu ; les montants moyens et médians d'indus et de rappels détectés à l'issue des contrôles. En outre, pour évaluer la valeur ajoutée du modèle, ces mêmes indicateurs seront produits pour deux requêtes plus simples : un tirage aléatoire, consistant à sélectionner aléatoirement des individus parmi l'ensemble des allocataires ; une sélection en priorité des dossiers allocataires présentant les montants cumulés de prestations les plus élevés sur une période de six mois.

En parallèle du suivi des performances du modèle, la surveillance en continu du déploiement du modèle du point de vue éthique repose sur trois axes :

- **La mesure de la sur- ou sous-représentation de différentes populations dans les contrôles réalisés sur la base du DMDE.** Les populations faisant l'objet d'une attention particulière sont celles correspondant aux 7 des 26 critères de discrimination pertinents dans le cadre des données traitées pour construire le DMDE (situation familiale ; genre ; nationalité ; lieu de résidence ; situation économique ; âge ; handicap via la perception de prestations spécifiques) et deux autres populations aussi identifiées lors des débats autour du DMDE 2018 (étudiants ; travailleurs indépendants).
- **La volumétrie des contrôles et le taux de contrôle**, sur pièces et sur place, issus du DMDE. Cette information permet d'une part de bien situer la portée de l'outil et donc de ses effets ; d'autre part, plus le nombre de contrôles datamining est important, plus on « descendra bas » dans la liste des scores et moins les effets seront concentrés sur les populations les plus à risque.
- **Caractère déterminant du score dans la prise de décision humaine.** Des indicateurs de part des dossiers non contrôlés après plusieurs mois seront produits et déclinés en fonction du score de risque et par population.

Ces 3 axes de suivi ont été définis à l'issue de la consultation du comité d'éthique. Ces indicateurs de suivi seront partagés avec le comité et pourront donner lieu à des actions correctives.

Annexe - Liste des variables du DMDE 2025 (modèles 1 et 2) et interprétation des liens avec la « cible » du modèle

information dossier : activité déclaré des membres du foyer, des enfants, présence de tutelle et âge des enfants		odd-ratio modèle 1 sans préremplissage	odd-ratio modèle 2 avec préremplissage	Part des dossiers de l'échantillon dans cette situation	Part des dossiers de l'échantillon dans la cible	Commentaires
Activité du responsable de dossier et conjoint	Autre situation (personne seule avec ou sans activité, couple mono-actif où le conjoint travaille...)	**1**	**1**	62%	21%	
	Présence d'un conjoint sans activité	1,3	1,3	8%	29%	La majoration du risque d'être dans la cible est due au risque d'avoir oublié de déclarer l'activité ou la reprise d'activité du conjoint.
	Couple bi-actif	0,8	0,8	24%	8%	Un foyer dont les deux conjoints occupent une activité perçoit généralement des montants de prestation plus faibles, ce qui réduit le risque d'être dans la cible.
	Responsable du dossier ou conjoint retraité	0,8	0,6	7%	8%	Un allocataire retraité possède généralement une situation relativement stable (pas d'enfant à charge, revenu constant, logement stable, ...), le risque déclaratif est donc plus faible d'où la minoration du risque d'être dans la cible.
Présence d'au moins un enfant âgé de plus de 19 ans dans le foyer	non	**1**	**1**	94%	17%	
	oui	1,5	2,1	6%	32%	Pour qu'un enfant soit considéré à charge, ses revenus nets mensuels (montant net social) ne doivent pas dépasser 55% du Smic. Un foyer où vit au moins un enfant de plus de 19 ans a donc mécaniquement un risque plus élevé d'être dans la cible car il est plus fréquent que l'enfant travaille à ces âges que plus jeune et du fait du risque d'oubli de déclaration des ressources de l'enfant. 19 ans est aussi la dernière année avant la limite d'âge pour le bénéfice des prestations familiales (allocations familiales, et allocation de soutien familial) ; le versement du complément familial ainsi que des aides au logement s'arrête aux 21 ans de l'enfant. Cela augmente le risque de percevoir indûment ces prestations. Au moment des étapes de sélection de variables, plusieurs critères d'âge ont été testés (18,19 et 20 ans), tous significatifs. Le seuil de 19 ans était le plus corrélé à la cible.
Activités des enfants	pas d'enfant en âge d'avoir une activité déclarée	**1**		83%	16%	
	au moins un enfant étudiant	1,5		11%	24%	Quand un foyer comprend au moins un enfant étudiant, le risque d'être dans la cible augmente car les étudiants sont souvent dans des situations instables (changement de domicile fréquents, changement de statut, emploi saisonnier, ...) qui peuvent influer sur le droit et sur les montants perçus.
	au moins un enfant salarié	1		2%	24%	Quand un foyer compte au moins un enfant exerçant une activité salariée, le risque est identique à celui de ne pas avoir d'enfant en âge d'avoir une activité.
	autres types d'activité	1,2		3%	21%	Cette catégorie comprend les enfants stagiaires, en apprentissage, chômeurs, scolarisés, ou ceux dont l'activité n'est pas renseignée. Les enfants stagiaires, en apprentissage ou chômeurs (indemnisés) font l'objet d'une majoration de leur risque d'indus en raison d'un risque plus élevé d'oublier ou de se tromper dans le montant de revenu déclaré pour les prestations qui sont calculées à partir des ressources de l'ensemble du foyer. Ceux dont l'activité n'a pas été renseignée peuvent faire l'objet d'un oubli de déclaration d'un changement de statut de l'enfant. Ces situations impliquent toutes deux un risque majoré de percevoir des montants indus de prestation.
	au moins un enfant sans activité	1,6		2%	36%	Un enfant sans activité (et donc probablement sans ressources) peut faire l'objet de changements de situation qui doivent être déclarés à la CAF (indemnisation par France Travail, déménagement, activité dont les revenus nets mensuels dépassent

						55% du Smic ...) car elles peuvent influer sur le droit et les montants perçus. Le risque d'oublier de les déclarer est plus grand, ce qui majore le risque d'être dans la cible.
Présence d'un dossier géré par une tutelle morale	pas de tutelle morale	**1**	**1**	97%	18%	
	au moins une tutelle morale	0,3	0,3	3%	6%	Une personne dont le dossier est géré par une tutelle morale est accompagnée dans les déclarations qu'elle doit faire à sa CAF, réduisant ainsi le risque d'erreur déclarative et minorant le risque d'être dans la cible.

Prestations		odd-ratio modèle sans préremplissage	odd-ratio modèle avec préremplissage	Part des dossiers de l'échantillon dans cette situation	Part des dossiers de l'échantillon dans la cible	Commentaires
Montant moyen de prestations perçues sur les 12 derniers mois	inférieur à 200€ par mois	**1**	**1**	37%	10%	
	compris entre 200€ et 1 400€ par mois	1,5	1,8	58%	21%	Le seuil pour la cible du modèle ayant été fixé à un indu de 600 euros sur une durée d'au moins 6 mois (soit en moyenne 100 euros sur un mois), un montant de prestations perçues supérieur à 200 euros augmente nécessairement le risque d'être dans la cible car les montants d'indus ont mécaniquement plus de chance d'être supérieurs au seuil fixé.
	supérieur à 1400€ par mois	1,9	2,9	5%	33%	En suivant la même logique, un dossier percevant des montants supérieurs à 1400 euros par mois a une majoration de risque encore plus élevée. Par ailleurs, les dossiers présentant un montant de prestation plus important sont des dossiers souvent plus complexes avec de multiples prestations, ce qui démultiplie le risque d'erreur déclarative. En outre, quand le montant versé pour une des prestations dépend d'autres prestations (interactions entre prestations), une erreur peut se répercuter d'une prestation à une autre.
Montant moyen d'allocations familiales (AF) perçues sur les 12 derniers mois	0	**1**		63%	19%	
	inférieur à 155€ par mois	1		23%	13%	155€ correspond approximativement au montant d'AF pour 2 enfants (non modulées).
	compris entre 155€ et 335€ par mois	0,9		5%	19%	La fenêtre 155-335€ correspond approximativement au montant d'AF pour 2 ou 3 enfants (non modulées).
	compris entre 335€ et 525€ par mois	0,8		6%	19%	La fenêtre 335-525€ correspond approximativement au Montant d'AF pour 3 ou 4 enfants (non modulées).
	supérieur à 525€ par mois	0,8		3%	29%	Toutes choses égales par ailleurs, percevoir des prestations familiales réduit le risque d'être dans la cible. Le barème des allocations familiales est moins sensible aux erreurs déclaratives que les prestations basées sur des ressources trimestrielles car les ressources sont pour la plupart récupérées automatiquement.
Montant moyen de complément familiale (CF) perçus sur les 12 derniers mois	0	**1**		92%	17%	
	inférieur à 191€ par mois	1,9		2%	27%	Les familles bénéficiant du CF perçoivent des montants importants de prestations, ce qui augmente mécaniquement le risque d'être dans la cible. Par ailleurs, le CF est versé pour les familles avec des enfants âgés de 3 à 21 ans pour lesquels le risque d'erreur de déclaration d'activité est possible augmentant le risque d'être dans la cible (reprise d'activité d'un parent non déclaré ou activité des enfants). 191€ correspond à un montant inférieur au montant du CF (non majoré), ce qui représente les cas des familles percevant le CF différentiel. Le risque est davantage majoré pour cette catégorie car une erreur sur les ressources se répercute immédiatement sur le montant de CF.
	compris entre 191€ et 195€ par mois	1,4		2%	14%	La fenêtre 191-195€ correspond aux montants versés aux foyers percevant le montant de CF de base chaque mois.
	supérieur à 195€ par mois	1,6		3%	34%	Les montants supérieurs à 195€ correspondent aux foyers bénéficiaires du CF majoré, leur risque est accru car le montant de prestation perçu est plus important et dépend des ressources, avec un seuil plus bas que pour le CF non majoré.
Montant moyen de prime d'activité	0	**1**	**1**	60%	11%	
	inférieur à 150€ par mois	1,2	1,5	23%	22%	

(PA) perçues sur les 12 derniers mois	compris entre 150€ et 250€ par mois	1,9	2	10%	29%	La prime d'activité demeure une prestation avec une charge déclarative complexe avec des déclarations et un réexamen du droit fréquents (trimestriels), de nombreuses ressources à déclarer et un risque d'incompréhension élevé sur les ressources qu'il faut déclarer. De ce fait, le risque d'erreur déclarative est élevé et explique la majoration du risque d'être dans la cible. Plus le montant de PA perçu est élevé, plus les indus éventuels le sont d'où la corrélation entre risque d'être dans la cible et le montant perçu.
	supérieur à 250€ par mois	2,7	2,6	7%	42%	
Montant moyen de revenu de solidarité active (RSA) perçus sur les 12 derniers mois	0	**1**		83%	15%	
	inférieur à 545€ par mois	1,3		10%	33%	Le RSA est une prestation complexe, il arrive que les ressources soient mal déclarées car les règles de déclarations sont difficiles à maitriser.
	compris entre 545€ et 560€ par mois	1,2		3%	21%	Le risque est majoré pour un montant de RSA perçu entre 545 et 560 € (autour du montant forfaitaire pour une personne seule sans ressource), il est cependant moindre que pour un montant inférieur à 545€ car le risque d'erreur est moins porté sur le montant des ressources.
	supérieur à 560€ par mois	1,8		4%	37%	Le risque est également majoré pour un montant de RSA perçu supérieur à 560 € (montant forfaitaire pour une personne seule), Par ailleurs, le seuil pour la cible du modèle ayant été fixé à 600 euros sur une durée d'au moins 6 mois (soit en moyenne 100 euros sur un mois), un montant de prestations perçues supérieur à 560 euros augmente nécessairement le risque d'être dans la cible car les montants d'indus ont mécaniquement plus de chance d'être supérieurs au seuil fixé.

Ressources		odd-ratio modèle sans préremplissage	odd-ratio modèle avec préremplissage	Part des dossiers de l'échantillon dans cette situation	Part des dossiers de l'échantillon dans la cible	Commentaires
Niveau de revenu (par rapport au Smic)	0	**1**	**1**	28%	15%	
	moins de 0,4	1	1	12%	26%	Le montant des prestations sous conditions de ressources, modulées avec les ressources ou différentielles dépend du revenu des allocataires. Plus les ressources sont élevées, moins l'allocataire bénéficie de prestations ce qui entraine mécaniquement une diminution du risque. Par ailleurs, le risque d'erreur lors de la déclaration de ressources est plus important lorsque les revenus ne sont pas nuls (erreur sur le montant de revenu déclaré ; plus grande volatilité des revenus d'un mois sur l'autre).
	entre 0,4 et 0,6	0,8	0,8	6%	23%	
	entre 0,6 et 1	1,1	1	16%	26%	
	entre 1 et 1,5	1,1		11%	24%	
	entre 1,5 et 2	1	0,8	7%	15%	
	2 ou +	0,5	0,4	20%	4%	
Abattement ou neutralisation appliqué sur le dossier sur les 24 derniers mois	non concerné ou pas de présence d'abattement	**1**		76%	15%	
	au moins 1 abattement ou 1 neutralisation appliquée dans les 12 derniers mois	1,2		19%	26%	Les règles d'abattement ou de neutralisation sont complexes, ce qui peut entraîner un risque d'indus lié à la situation professionnelle de la personne et à ses évolutions. Lorsque l'abattement est récent, la situation professionnelle de l'allocataire est connue avec moins de certitude et il peut y avoir un décalage entre les informations déclarées à la CAF et la réalité de l'activité professionnelle, d'où la majoration du risque d'être dans la cible.
	au moins 1 abattement ou 1 neutralisation appliquée il y a plus de 12 mois (entre le 13 et 24 mois précédent)	1		5%	21%	En revanche, si l'abattement est ancien, le risque d'indus diminue : la situation professionnelle est mieux connue, notamment grâce aux échanges réguliers avec France Travail.
Présence de pension alimentaire dans les déclarations trimestrielles des ressources (DTR) des 24 derniers mois	non concerné	**1**		45%	7%	
	oui	1,9		5%	28%	Par rapport à un allocataire n'ayant jamais réalisé de déclarations de pensions alimentaires dans sa DTR dans les 24 derniers mois, le risque d'être dans la cible est majoré. La majoration de risque est la même lorsque l'allocataire déclare recevoir une pension alimentaire que lorsqu'il déclare ne pas en recevoir.
	non	1,9		50%	26%	
Présence de changement dans la pratique d'une activité non salariée dans les 24 derniers mois	non	**1**		99%	17%	
	oui	1,8		1%	46%	L'activité non salariée est souvent marquée par une irrégularité des revenus. Le fait d'avoir des changements fréquents dans la déclaration d'activité non salariée indique une situation professionnelle instable, source d'irrégularité qui majore le risque d'être dans la cible.
Pratique d'une activité non salariée dans les 24 derniers mois	non		**1**	98%	17%	
	oui		1,8	2%	44%	Le fait d'avoir une activité non salariée dans sa DTR majore le risque d'être dans la cible car la déclaration d'une activité non salariée est particulièrement complexe et source d'erreur déclarative.
Fréquence et montant moyen de signalements qui ne sont pas vérifiés par la Caf dans les 12 derniers mois	Autres situations		**1**	95%	16%	
	Plus de 2 déclarations de ressources avec un signalement de ressource à la baisse de plus 150€ en moyenne		2,9	5%	48%	Lorsqu'un allocataire du RSA ou de la prime d'activité corrige les ressources pré-remplies avec le DRM à la baisse d'au moins 150€ pour au moins deux DTR et que son dossier n'est pas examiné par la cellule dédiée, le risque d'être dans la cible est majoré. La condition de réitération et le montant d'au moins 150€ s'expliquent par la nature de la cible (un indu d'au moins 6 mois et d'au moins 600€). Cela peut correspondre à une incompréhension de l'allocataire sur les montants de revenu à déclarer

Faits générateurs (Fge) caractérisant les changements sur le dossier		Multiplicateur de risque odd-ratio modèle sans préremplissage	Multiplicateur de risque odd-ratio modèle avec préremplissage	Part des dossiers de l'échantillon dans cette situation	Part des dossiers de l'échantillon dans la cible	Commentaires
Présence d'un fait générateur sur une confirmation logement étudiant	aucun fait générateur CONLOGETU	**1**	**1**	96%	18%	
	1 ou plus	1,8	1,21	4%	17%	Un étudiant déclarant à sa CAF qu'il souhaite conserver son logement étudiant pendant l'été (ce qui crée un fait générateur CONLOGETU dans le dossier) mais qui en réalité quitte son logement et perçoit des aides au logement indues. Cela a pour effet de majorer son risque d'être dans la cible.
Présence d'un fait générateur "PROAC" PROjet d'ACCompagnement" permettant l'enregistrement des projets d'accompagnement des bénéficiaires de RSA soumis aux droits et devoirs, et permettant ainsi la continuité du versement de leur prestation.	aucun fait générateur PROACC	**1**		98%	17%	
	1 ou plus	0,9		2%	30%	Le projet d'accompagnement concerne les allocataires du RSA suivis à des fins de réinsertion professionnelle. Cela indique une situation où l'allocataire est accompagné et échange régulièrement avec sa CAF. Dans une telle situation, le risque d'erreur ou d'oubli de déclaration est plus faible, d'où la minoration du risque d'être dans la cible.
Présence de fait générateur de déclaration de volet sociaux (VOLSOC dans le cadre du complément de libre choix du mode de garde (CMG) assistant maternel ou garde à domicile)	aucun fait générateur VOLSOC	**1**	**1**	92%	18%	
	1 ou plus	0,7	0,6	8%	8%	Le volet social est un document que les allocataires du CMG doivent renseigner mensuellement pour pouvoir bénéficier de la prestation. Globalement, la charge déclarative et le risque d'erreur déclarative liée au bénéfice du CMG est relativement faible par rapport à d'autres prestations, d'où la minoration du risque d'indus.
Nombre de changements de droit au RSA dans les 24 derniers mois	0	**1**	**1**	91%	16%	
(aller-retour dans la prestation)	1	1,1	1,5	5%	32%	Un dossier avec beaucoup de changements de droit au RSA suppose des changements fréquents de situation (ressources, composition familiale...), sources d'irrégularités qui augmentent le risque d'erreur ou d'oubli dans les déclarations, majorant ainsi le risque d'être dans la cible. Dans le cas d'un seul changement sur la période des 24 mois précédant le contrôle, le risque n'est que très légèrement majoré.
	2 ou plus	1,3	1,8	4%	37%	Lorsque le dossier présente au moins 2 changements sur la période des 24 mois précédant le contrôle, le risque d'être dans la cible est encore plus élevé car l'instabilité des situations doit être plus forte.