

De : Cnaf

A: Commission d'accès aux documents administratifs
A l'attention de Monsieur Jean-Claude Cluzel
TSA 50730
75334 Paris cedex 07

A Paris, le 7 décembre 2022

V. Réf. 20225787 et 20226179

Objet : opposition à demande d'accès

Cher Monsieur,

La présente lettre a pour objet de répondre aux questions complémentaires adressées par courriel du 30 novembre 2022 de la Commission d'accès aux documents administratifs (Cada) à la suite de la réunion du 18 novembre 2022, à savoir :

- préciser les raisons pour lesquelles la demande présentée dans le dossier 20226179 qui en son point 1) vise « *la nature des variables utilisées (âge, niveau de revenu, situation professionnelle, nombre d'enfants, ...)* et les coefficients associés », ne pourrait être satisfaite qu'en adressant au demandeur la formule de calcul du score de risque, ainsi que l'ensemble du code source aboutissant à ce calcul ;
- rappeler brièvement la méthode de construction et le contenu de l'algorithme ;
- préciser la liste des variables qui ont été occultées et expliciter, pour chacune d'entre elles, les raisons conduisant la Cnaf à opposer les réserves tenant, d'une part, à l'atteinte à la recherche et à la prévention, par les services compétents, d'infractions de toute nature et ou, d'autre part, à la sécurité publique, ainsi que, le cas échéant, à la sécurité des systèmes d'information.

1. Propos introductifs

Le versement des prestations par les Caisses d'allocations familiales (Caf) repose sur un système déclaratif permettant un versement rapide des prestations aux allocataires. La Cnaf est garante du bon versement des fonds publics et vérifie dès lors l'exactitude des déclarations des allocataires.

Le modèle de Datamining Données Entrantes (ci-après : « *DMDE* » ou « *modèle de Datamining* ») s'inscrit dans l'objectif de paiement à bon droit et de lutte contre la fraude, en détectant les situations susceptibles de générer des risques d'indus. Un indu est un trop perçu par l'allocataire, à l'inverse d'un rappel. Il s'agit d'un outil de détection des situations à risques, au service du contrôle des données déclarées.

Grâce au datamining, différents dossiers avec un score de risque d'indu considéré comme important sont identifiés et des contrôles peuvent alors être diligentés. En effet, plus le dossier d'allocataire est risqué au titre du datamining, plus le risque d'indu ou de rappel est élevé. Les contrôles peuvent être faits « sur pièces » réclamées aux allocataires ou sur place au domicile des allocataires. En tous les cas, ils sont réalisés par des humains.

a) Impact du datamining sur le contrôle

En 2021, les Caf ont réalisé 35,6 millions de contrôles auprès de 7,1 millions d'allocataires. Ces contrôles ont permis la détection de **894,4 millions d'euros d'indus en 2021, et 328,6 millions d'euros de rappels**.

Sur ces 35,6 millions de contrôles, 31,5 millions étaient automatiques (fiabilisation des données déclarées par acquisition de données de partenaires), et 4,01 millions réalisés par un agent en charge du contrôle, soit sur place (déplacement au domicile de l'allocataire) ou sur pièces (exploitation de pièces demandées à l'allocataire).

Sur ces 4,01 millions de contrôles, 254 158 avaient comme origine le modèle de Datamining et ont permis de détecter 305,4 millions d'euros d'impact financier, dont 219,8 millions d'euros d'indus.

Les contrôles réalisés sur la base du datamining représentaient donc 6 % de l'ensemble des contrôles réalisés sur pièce ou sur place, mais soit presque 25 % des indus totaux détectés

	Volume	Montant d'indus	Montant de rappels	Impact financier total
Sur place	88 196	143 820 196	45 158 407	188 978 603
Sur pièces	165 962	75 967 047	40 487 253	116 454 300
Total	254 158	219 787 243	85 645 660	305 432 903

b) Impact du datamining sur la lutte contre la fraude

Les 4,01 millions de contrôles réalisés par un agent ont permis de suspecter 46 491 fraudes, soit un 1% de ces contrôles.

Les 254 158 contrôles réalisés sur la base du datamining ont permis de suspecter 13 968 fraudes.

30 % des suspicions de fraudes détectées en 2021 l'ont été dans le cadre d'un contrôle réalisé sur la base du datamining, soit 94 millions des 309 millions d'euros de fraude détectées.

Les contrôles réalisés au titre du datamining participent donc largement à la politique de contrôle et de lutte contre la fraude de la Cnaf.

2. La méthode de construction et le contenu de l'algorithme

Enquête PBDF. Chaque année, une enquête d'évaluation du paiement à bon droit, c'est-à-dire de contrôle du versement au bon moment des sommes exactes des prestations auxquelles l'allocataire a droit, ainsi que de la fraude, est réalisée par la Cnaf.

A partir d'un échantillon aléatoire et représentatif de l'ensemble des allocataires percevant une prestation, un contrôle est réalisé par des contrôleurs assermentés chargés de vérifier les dossiers sur un historique de 2 ans.

L'exploitation de cet échantillon apporte une connaissance objective du risque d'indus et des allocataires qui le portent. Sur cette base est définie une cible de modélisation statistique. Le modèle mis en œuvre dans les Caf depuis 2020 cible le risque d'indus de plus 6 mois et d'un montant supérieur à 600€. L'objectif de la modélisation statistique est de calculer un risque, pour chacun des allocataires, d'être l'objet d'un tel indu si on contrôle sa situation. Le modèle permet donc de passer d'une connaissance sur un échantillon à une extrapolation, via une approche probabiliste, à l'ensemble des allocataires. Pour construire ce modèle, les résultats de l'enquête PBDF sont adossées aux informations disponibles dans le système d'information des Caf, au sujet de la situation de ces allocataires et de la vie de leurs dossiers.

Les experts statisticiens du Centre national d'appui au datamining (Cnad) mènent alors une démarche d'étude datamining destinée à faire émerger les caractéristiques/variables des dossiers d'allocataires, propres à identifier les dossiers ciblés.

Une méthode statistique. La méthode de Datamining est une méthode scientifique d'étude purement statistique, sans intervention métier. Elle suit des étapes de modélisation itératives, mises en œuvre par un expert statisticien.

Le score de risque. La modélisation statistique permet d'obtenir une pondération pour chacune des variables. Leur combinaison aboutit au calcul d'une probabilité, autrement appelée « Score de Risque données entrantes », comprise entre 0 et 1. Ce calcul appliqué à l'ensemble des allocataires permet de les classer par score de risque décroissant et ainsi de cibler les contrôles dits « datamining » auprès des allocataires présentant les dossiers les plus à même de comporter des indus importants lorsque le contrôle est dispensé.

La nature des variables. Pour le modèle actuellement utilisé, ce sont 41 variables qui ont émergé comme étant discriminantes du risque d'indu important (une autre modélisation aboutirait à des choix différents). Les variables de l'algorithme de calcul du score, sont de natures très variées. Il s'agit notamment de caractéristiques déclarées par l'allocataire sur sa situation familiale, sa situation professionnelle, sa situation financière et sa résidence.

Il s'agit également de données internes aux Caf relatives à la gestion des dossiers des allocataires, notamment les données relatives aux prestations reçues, les données concernant la gestion du dossier, les éléments sur l'historique du dossier, les déclarations de changement de situation, l'existence d'éventuels contentieux et des caractéristiques sociaux-économiques sur la commune de résidence de l'allocataire.

La liste des variables. Le modèle de Datamining cherche à identifier les situations présentant un risque d'indu important lorsque la situation d'un allocataire est contrôlée.

La divulgation de la liste des variables du modèle de Datamining et de leur pondération dans le calcul du score aurait pour conséquence d'inciter les fraudeurs à déclarer les caractéristiques qui leur permettrait de réduire le risque de voir leur situation contrôlée.

La liste des variables est annexée au présent courrier au format Excel.

A chacune des variables est associé un coefficient multiplicateur qui permet, en fonction de l'occurrence de la situation rencontrée, de majorer ou de minorer le niveau de risque d'indu.

3. Le refus de communication de la nature des variables utilisées et les coefficients associés

Interdépendance entre les parties du modèle datamining

Donner la liste des variables utilisées et les pondérations attribuées à chacune de leurs modalités, dans le calcul du score datamining, revient à fournir la formule de calcul du score de risque. Celui-ci n'est autre que la combinaison des variables.

Quant au code source de calcul du score, il récupère dans le SI de la Caf les données permettant de calculer ces variables et leurs modalités, puis finalement de calculer le score en les combinant. Elles ne peuvent donc pas être communicables.

L'organisation de la fraude

La communication du modèle Datamining aurait pour conséquence d'ouvrir la voie à de possibles fraudes : dans un contexte d'ouverture de droits sur du déclaratif, en identifiant les critères constituant des facteurs de ciblage, des fraudeurs pourraient organiser et monter des dossiers frauduleux, constitués d'éléments leur permettant d'éviter le contrôle en cherchant à minimiser l'ampleur de leur score.

Le tableau suivant illustre, à titre d'exemple, les conséquences de la communication de la nature des variables utilisées et des coefficients associés sur le comportement des fraudeurs :

	(
	(
	(

				I

En synthèse, la communication de la nature des variables, de leur coefficient ou la formule de calcul ou du code source, reviendrait à :

- Réduire l'efficacité du modèle
- Réduire l'efficacité de la politique de lutte contre la fraude de la branche famille
- Par effet ciseaux, encourager les mécanismes de fraude organisée, et donc de nouveaux schémas de fraude.

Le principe de proportionnalité empêche la communication de ce modèle.

La Cnaf demeure à la disposition de la Cada pour toute précision complémentaire.

Je vous prie d'agréer, Cher Monsieur, l'expression de ma considération respectueuse.

Paule-Marie GRÉGOIRE

Personne Responsable de l'Accès aux Documents
Administratifs et à des questions relatives à la
réutilisation des informations publiques (PRADA)

