# Analyse des réclamations d'allocataires de la CAF : un cas d'étude en fouille de données

Sabine Loudcher*, Julien Velcin*, Vincent Forissier*, Cyril Broilliard** et Philippe Simonnot***

*Laboratoire ERIC, Université de Lyon
{sabine.loudcher, julien.velcin, vincent.forissier}@univ-lyon2.fr
**CNEDI Rhône-Alpes - CNAF
cyrille.broilliard@cnedi69.cnafmail.fr
***CAF du Rhône
philippe.simonnot@cafrhone.cnafmail.fr

**Résumé.** La gestion des réclamations est un élément fondamental dans la relation client. C'est le cas en particulier pour la Caisse Nationale des Allocations Familiales qui veut mettre en place une politique nationale pour faciliter cette gestion. Dans cet article, nous décrivons la démarche que nous avons adoptée afin de traiter automatiquement les réclamations provenant d'allocataires de la CAF du Rhône. Les données brutes mises à notre disposition nécessitent une série importante de prétraitements pour les rendre utilisables. Une fois ces données correctement nettoyées, des techniques issues de l'analyse des données et de l'apprentissage non supervisé nous permettent d'extraire à la fois une typologie des réclamations basée sur leur contenu textuel mais aussi une typologie des allocataires réclamants. Après avoir présenté ces deux typologies, nous les mettons en correspondance afin de voir comment les allocataires se distribuent selon les différents types de réclamation.

## 1 Introduction

La Caisse Nationale des Allocations Familiales (CNAF), branche "famille" de la sécurité sociale française, gère un réseau régional de Caisses d'Allocations Familiales (CAF) dont l'objectif est de venir en aide aux familles et aux personnes en difficulté financière, pour des raisons de santé, familiales ou professionnelles. A ce titre, elles versent différentes prestations à leurs allocataires dans quatre grands domaines : le logement, la naissance du jeune enfant, l'entretien de la famille et la garantie de revenus. Dans un souci d'amélioration de la qualité de service, la CNAF veut mettre en œuvre une politique nationale de gestion des réclamations.

Selon un travail préliminaire réalisé par la CNAF, une *réclamation* est définie comme "tout mécontentement exprimé à l'égard d'une décision, d'une procédure ou d'un service de la Caisse d'Allocations Familiales, quelle qu'en soit la forme, et pour lequel une réponse est explicitement ou implicitement attendue". Dans une logique marketing, la gestion des réclamations est un élément fondamental dans la gestion de la relation client (*Customer Relationship Management* ou CRM), comme le soulignent (Stauss et Seidel, 2004).

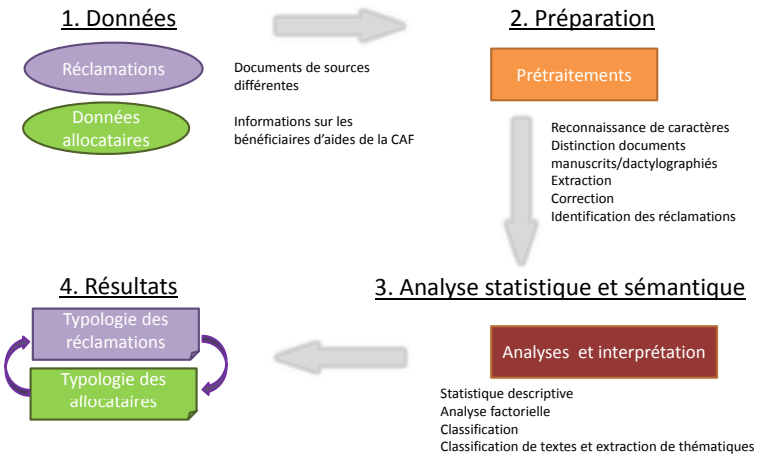


FIG. 1 – *Processus général d'analyse automatique des réclamations.*

Dans ce cadre, le travail confié au laboratoire de recherche ERIC consiste à étudier de façon exploratoire les opportunités offertes aujourd'hui par les techniques de fouille de données, et en particulier de fouille de textes, pour réaliser une analyse automatique des réclamations envoyées par les allocataires. Des travaux préalables concernant la gestion des réclamations (*complaint management* en anglais) ont été listés dans la littérature – voir à ce sujet le large panorama proposé par (Ngai et al., 2009) –, mais très peu d'efforts semblent avoir été entrepris pour traiter ce problème spécifique. Il est intéressant de noter que (Bae et al., 2005) ont déjà essayé d'utiliser des cartes auto-organisatrices (*Self-Organized Maps* ou SOM, c.f. (Kohonen, 2001)), c'est-à-dire une technique issue de l'apprentissage non supervisé.

La démarche adoptée dans cette étude exploratoire est illustrée dans la figure 1. Elle se décompose en quatre étapes :

*Etapes 1 et 2 pour*

– identifier les documents contenant les réclamations et les rendre exploitables en vue d'analyses automatiques,
– récupérer les données disponibles sur les allocataires ayant rédigé ces réclamations.

*Etapes 3 et 4 pour*

– à partir des informations sur les allocataires réclamants, établir une typologie des réclamants en utilisant des techniques classiques d'analyse des données,
– à partir du contenu textuel des documents, établir une typologie des réclamations sur la base d'analyses statistiques et sémantiques,
– croiser les types de réclamations et les caractéristiques des allocataires afin de définir

des groupes d'individus au comportement homogène en matière de réclamation.

Cet article s'organise comme suit. Tout d'abord, nous présentons en détail dans la section 2 les données issues d'un échantillon des réclamations envoyées par des allocataires à leur CAF. Nous donnons à cette occasion les différents prétraitements qui ont rendu les analyses ultérieures possibles. Dans la section 3, nous présentons les techniques de fouille de données qui ont été choisies pour construire les deux typologies. La section 4 donne les résultats des deux typologies obtenues, ainsi que la mise en correspondance réalisées entre la typologie des allocataires réclamants et celle des réclamations. Enfin, la section 5 propose une conclusion à ce travail, ainsi que des pistes d'études futures.

# 2 Préparation du jeu de données

## 2.1 Données brutes de la CAF

Les allocataires peuvent adresser des réclamations à leur CAF par différents canaux, allant de l'appel téléphonique à la lettre manuscrite en passant par le mail, le(s) site(s) internet ou la lettre dactylographiée. L'un des objectifs de ce travail étant l'automatisation des traitements, il a fallu déterminer quels formats de données rendaient possible une analyse automatique. Concernant les courriers reçus par voie postale, il a été décidé d'identifier automatiquement les courriers dactylographiés pour ne retenir que ceux-ci et d'écarter les courriers manuscrits. Tous les courriers électroniques via différents sites Web ont été retenus. Une partie des documents fournis n'ayant pas été identifiée comme des réclamations, il a fallu mettre en place une procédure pour discriminer automatiquement les réclamations à partir de leur contenu textuel. Cette procédure est brièvement présentée dans la section qui suit.

Les données transmises par la CAF du Rhône pour ce travail sont des réclamations réceptionnées entre janvier 2010 et mars 2012. Parmi les 174000 documents parvenus à la CAF pendant cette période, seul un échantillon de 12534 documents a pu être traité dans la durée de l'étude. Parmi ces 12534 documents, 2385 documents ont été identifiés automatiquement comme étant des réclamations ; ils ont donc été retenus pour les analyses ultérieures. Chaque document contient un texte, en général plutôt court et aisément identifiable à l'aide de techniques automatiques. De plus, il peut être associé aux informations ou données qui décrivent l'allocataire *au moment de la réclamation*, comme l'état civil (civilité, sexe, âge) et des variables construites par les agents de la CAF. Au final, 39 variables descriptives ont été sélectionnées initialement pour l'étude. Un allocataire qui réclame plusieurs fois voit son profil dédoublé comme s'il s'agissait de plusieurs personnes différentes, le profil pouvant en effet varier au fil du temps.

## 2.2 Prétraitements des données

Dans cette partie, nous détaillons les traitements nécessaires pour transformer les données brutes brièvement décrites dans la section précédente dans un format propice à l'emploi d'algorithmes de fouille de données. Ces traitements représentent au moins 70% de l'effort investi dans ce travail, ce qui justifie la place qui leur est dédiée dans cet article.

Concernant les informations liées à la description des allocataires réclamants, nous avons sélectionné 19 variables parmi les 39 initiales mises à notre disposition. Cette sélection permet de limiter la redondance de l'information (par exemple, plusieurs variables traitent du nombre d'enfants dans le foyer) et d'éviter à ce que le nombre de valeurs manquantes soient trop élevées (des nouvelles variable descriptives ont été créées à partir de 2011). Afin de permettre l'emploi des méthodes d'analyse des données présentées dans la section suivante, certaines variables qualitatives ont été recodées et les variables continues ont été discrétisées. La plupart du temps, cette discrétisation se base sur des intervalles définis antérieurement par la CAF.

Concernant les réclamations proprement dites, le travail de prétraitement des documents a été plus important. Il se décompose en six grandes étapes que nous listons ci-dessous.

**Reconnaissance des caractères.** La diversité dans la forme des documents fournis (dactylographiés, électroniques...) implique d'utiliser un OCR (*Optical Character Recognition*) afin de transformer les images de certains documents fournis au format `.pdf` ou `.tiff` dans un format texte. Après une revue de plusieurs solutions logicielles existantes sur le marché, nous avons opté pour le logiciel ABBYY FineReader [1]. Celui-ci nous a permis d'obtenir des taux de bonne reconnaissance très corrects lorsque les documents sont dactylographiés : le logiciel commet entre 2,5 et 10% d'erreur dans la reconnaissance des caractères ; erreur estimée en comparant (sur la base d'un échantillon de 50 documents) le texte extrait manuellement par nos soins au texte automatiquement reconnu.

**Distinction entre écriture manuscrite et dactylographiée.** La deuxième étape consiste à ne conserver que les documents dactylographiés afin d'assurer une qualité minimale aux textes reconnus à partir des images. Cette distinction a été simple à réaliser en fixant un seuil de 70% au nombre minimum de chiffres ou de lettres contenus dans un document. Lorsque ce seuil est dépassé, c'est-à-dire lorsque d'autres symboles sont trop fréquents statistiquement (comme ' ?', ' !' ou #'), il est clair que la reconnaissance a échoué. Le taux de reconnaissance des documents dactylographiés est alors de 100% sur notre échantillon.

**Extraction du texte des réclamations.** Les fichiers textes issus de l'étape d'OCR contiennent tous les éléments présents dans le document d'origine. Le texte de la réclamation est donc situé au milieu des informations personnelles de l'allocataire. Le but de cette troisième étape est de réussir à ne conserver que le texte ou le corps de la réclamation. Pour ce faire, nous utilisons des expressions régulières car elles permettent de rechercher des chaînes de caractères caractéristiques. Pour les réclamations issues du site web national de la CAF, par exemple, cet exercice s'est avéré aisé grâce à la présence dans le document d'une balise "*Message :*".

**Correction du texte.** Des erreurs peuvent être introduites dans le texte, imputables à l'allocataire lui-même ou à l'algorithme d'OCR. Afin de corriger certaines de ces erreurs, nous avons mis en place une étape de correction automatique. Celle-ci se base sur deux mécanismes. Tout d'abord, nous avons recours à un lexique de mots de la langue française (le dictionnaire

1. http://france.abbyy.com/

Gutemberg [2]), à une liste de mots utilisés à la CAF (par exemple "*APL*" et "*AAH*" qui correspondent à des prestations spécifiques), et à une liste de néologismes détectés manuellement (par exemple "*sms*" ou "*tweet*"). Les mots de la réclamation sont comparés à ce vocabulaire et directement indexés s'ils s'y trouvent : nous les appelons les mots reconnus. Le deuxième mécanisme consiste à comparer les mots non reconnus avec le vocabulaire en calculant une distance entre ces mots. Dans notre cas, nous avons employé la distance de Levenshtein (Levenshtcin, 1966), normalisée par la taille du mot à corriger afin de ne pas pénaliser les mots trop longs (Cohen et al., 2003), puis la distance de Jaro-Winkler (Corston-Oliver et Gamon, 2004) en cas d'ex-aequo. Après plusieurs tentatives, nous avons fixé un seuil de 0,2 à cette distance, ce qui permet par exemple de corriger "*prochainemant*" en "*prochainement*" (score de 0,077), mais pas "*prme*" en "*prime*" (score de 0,25) ou "*priscilla*" en "*principal*" (score de 0,444). Même s'il arrive que certaines corrections amènent à introduire des erreurs, les corrections effectuées avec le seuil de 0,2 sont la plupart du temps pertinentes.

**Identification des réclamations.** Les documents transmis par la CAF du Rhône ne sont pas tous des réclamations et une étape d'identification est nécessaire. Habituellement, ce travail est réalisé manuellement par un technicien de la CAF à l'aide d'une note de service qui définit la notion de réclamation ainsi que des critères ou des expressions pour les repérer. Ainsi sont considérés comme réclamations les textes contenant des tournures de phrases exprimant une incompréhension, une protestation ou une contestation. Il a été nécessaire d'automatiser cette étape en nous basant sur la note de service. Brièvement, des expressions comme "*je ne comprends pas votre décision*" ont été remplacées par l'association des mots "*comprend*", "*pas*" et "*décision*". C'est la présence d'une majorité de mots clefs associés à l'une des expressions typiques qui permet d'indiquer automatiquement qu'un texte relève ou non d'une réclamation. Cette méthode a permis d'automatiser le processus de discrimination afin de ne retenir que des réclamations. Bien sûr, une partie des textes a été retenue à tort. Pour estimer cette proportion, 100 documents ont été tirés au hasard. Un expert humain a lu les documents, a vérifié pour chaque document s'il s'agissait ou non d'une réclamation et a comparé avec l'identification faite automatiquement par la machine. 76% des réclamations sont retrouvées par la machine (score de rappel) et 73% des réclamations retenues par la machine en sont bien (score de précision). Sachant qu'un agent de la CAF est également susceptible de retenir ou d'écarter à tort des documents, le taux de reconnaissance automatique des réclamations a été jugé acceptable mais des travaux plus poussés devraient permettre d'améliorer les performances (voir la section 5).

**Suppression des mots-outils et stématisation.** Afin d'optimiser les résultats des méthodes statistiques et sémantiques que nous abordons dans la section prochaine, il est souvent préférable de retirer préalablement des textes les "mots outils" (*stopwords*), c'est-à-dire des mots fréquemment utilisés dans la langue française pour construire les phrases, comme "*et*", "*avec*", etc. Des listes préconstruites de mots outils sont disponibles pour la plupart des langues, et en particulier pour le français [3]. Une deuxième technique permet de restreindre le nombre de mots en supprimant les préfixes et les suffixes, et ce afin de se rapprocher du radical des termes. Cette technique est appelée "stématisation" (*stemming* en anglais), elle ne doit pas être confondue

2. http://www.pallier.org/ressources/dicofr/dicofr.html
3. http://snowball.tartarus.org/algorithms/french/stop.txt

avec la lemmatisation. Dans notre cas, elle nous permet de diminuer de manière significative le nombre de termes présents dans notre vocabulaire : les 10 248 termes présents initialement dans les textes du projet ont été ramenés à 7256 formes stématisées, soit une diminution de 29% de la taille du vocabulaire.

# 3 Analyse des réclamations

Dans cette partie, nous motivons le choix des techniques employées pour analyser les réclamations, puis nous donnons le protocole expérimental que nous avons suivi pour obtenir les résultats présentés dans la section suivante.

## 3.1 Choix des techniques de fouille de données

**Combiner ACM et CAH** Afin de construire une typologie des allocataires qui font une réclamation, nous proposons d'utiliser l'Analyse des Correspondances Multiples (ACM) en combinaison avec une Classification Ascendante Hiérarchique (CAH), techniques éprouvées issues de l'analyse des données (Lebart et al., 1995).

Brièvement, l'ACM est une méthode de décomposition factorielle qui fournit une représentation graphique synthétique d'une grande quantité de données décrites par des variables qualitatives. Elle synthétise l'information, met en évidence les informations intéressantes ainsi que les liens qui les caractérisent et aboutit à la création d'axes factoriels pour représenter les individus et/ou les modalités. En revanche l'ACM ne fournit pas de typologie proprement dite. Pour construire la typologie, nous appliquons la CAH avec le critère de Ward. La classification permet de faire émerger des groupes ou classes d'allocataires au profil semblable. Le nombre de classes est sélectionné à partir du plus grand "saut" constaté dans le critère de Ward et en faisant appel à un expert métier pour attester du meilleur niveau de granularité.

L'intérêt d'effectueur une classification après une méthode factorielle est double : (1) l'ACM procède à une réduction du nombre de variables. La classification est faite avec les axes factoriels issus de la méthode factorielle et non pas avec les variables d'origine ; (2) sur le graphique de l'ACM représentant les individus (ici les allocataires réclamants), il est possible de visualiser l'appartenance des individus aux classes, c'est-à-dire la typologie des allocataires.

**Extraction de thématiques** L'un des objectifs de ce projet est de construire une typologie des réclamations de manière automatique. La démarche générale consiste à travailler directement à partir du contenu textuel des documents en utilisant le moins d'information *a priori*, ce afin de faire émerger des groupes ou catégories de réclamations homogènes, que nous appellerons des thématiques (*topics* en anglais). Les textes sont placés dans la même thématique à partir du moment où leur auteur emploie un vocabulaire similaire, et dans des thématiques différentes lorsque le vocabulaire employé est différent.

Il existe plusieurs méthodes d'extraction des thématiques : par exemple les modèles à base de distance (Velcin et Ganascia, 2007), inspirés de l'algorithme classique des c-moyennes, les modèles reposant sur la factorisation de matrices de type LSA (Deerwester et al., 1990) ou NMF (Paatero et Tapper, 2006) , ou les modèles graphiques basés sur la statistique bayésienne (Steyvers et Griffiths, 2007). Relevant de ce dernier type, le modèle qui a été choisi pour cette étude est LDA (*Latent Dirichlet Allocation*), proposé par (Blei et al., 2003). Il a été choisi car il

s'agit d'un modèle aux performances reconnues qui a déjà été appliqué à de nombreux corpus de natures variées (voir par exemple (Bíró et al., 2008)).

La méthode LDA utilise les principes des modèles graphiques et de la statistique bayésienne, appliqués aux données textuelles. Elle se base sur une représentation en "sac de mots" (*bag of words*) où un document est traité comme un ensemble de mots dont la position dans le texte n'est pas prise en compte. Cette hypothèse simplificatrice entraîne une perte dans la précision des résultats de l'analyse, mais elle rend possible le traitement de corpus volumineux.

Le modèle calculé par LDA contient un ensemble de catégories (les thématiques) et précise comment les documents sont répartis sur ces catégories. Chaque catégorie correspond à une thématique décrite comme une distribution sur l'ensemble du vocabulaire de mots choisi. La thématique est souvent caractérisée par les mots clefs qui contribuent le plus à la catégorie (*top keywords*). L'étiquetage (nommage) des catégories peut s'effectuer en observant les expressions que l'on peut reconstituer à partir de ces mots clefs et en s'aidant des textes du corpus qui contiennent ces expressions. Par exemple, voici une liste de mots clefs les plus pertinents pour illustrer une thématique qui pourrait être obtenue : "*réponse*", "*allocation*", "*enfants*", "*montant*", "*logement*", "*situation*", "*été*", "*part*", "*aide*", "*allocataire*". En se basant sur ces mots et sur les textes des réclamations, on retrouve des expressions telles que : "*montant [de l'] allocation logement*", "*montant [de l'] aide [au] logement*", ou encore "*réponse [de votre] part*". Cette thématique couvre des textes tels que le suivant : "*Me mr comme je vais déménager le 1er mai 2010 je souhaite que l'aide au logement d'avril 2010 soit versée sur mon compte. Le loyer d'avril a été intégralement payé au propriétaire. L'aide au logement d'avril me permettra de compléter le paiement du loyer du nouveau logement. Je perçois toujours l'ASS, la dédite a été déjà envoyée au propriétaire je passerai pour remplir un nouveau dossier.*"

## 3.2 Protocole expérimental

Concernant la typologie des réclamations, les expérimentations ont été réalisées en deux temps. Il s'agit tout d'abord de réduire le vocabulaire en introduisant deux méthodes de filtrage des mots. Ensuite viennent les expérimentations proprement dite pour lesquelles se pose la question de la sélection du nombre de catégories thématiques.

**Première étape : réduction du vocabulaire.** Il est nécessaire de réduire encore la taille du vocabulaire qui contient encore 7256 mots après l'étape de stématisation. Pour cela, deux filtres sont employés et permettent de réduire le vocabulaire à 364 mots.

Le premier filtre consiste à éliminer du vocabulaire les mots qui apparaissent trop peu dans le corpus. En effet, ces mots apportent du "bruit" dans l'analyse alors qu'ils n'apportent rien dans la constitution des catégories thématiques. La "rareté" du mot (*sparsity* en anglais) est simplement calculée en prenant le ratio du nombre de documents contenant le mot par le nombre total de documents dans le corpus. Chaque mot est donc associé à une valeur numérique située entre 0 et 1 : plus la valeur est proche de 0, plus le mot est rare et doit être ignoré. Il faut donc définir un seuil $\theta_s$ en dessous duquel les mots sont retirés du vocabulaire et donc de l'analyse.

Le second filtre consiste à éliminer du vocabulaire les mots trop fréquents dans les textes, mais dans des proportions similaires, et qui peuvent être assimilés à des mots outils. Pour cela, nous nous sommes basés sur la mesure classique de TF.IDF qui est le produit entre la fréquence du mot dans un texte en particulier (*Term Frequency* ou TF) et le logarithme de l'inverse de la

fréquence du mot dans tous les documents du corpus (*Inverse Document Frequency* ou IDF). Cette mesure est d'autant plus proche de zéro que le mot est peu fréquent dans un texte et qu'il est au contraire présent tout au long du corpus. Là encore, il s'agit de définir un seuil $\theta_T$ en dessous duquel le mot est retiré de la description du texte. Attention, contrairement à la mesure précédente, le mot est éliminé d'un texte, et non pas nécessairement du vocabulaire dans son ensemble. En effet, il se peut qu'un mot soit exceptionnellement fréquent dans un texte (score de TF élevé), ce qui compense le fait qu'il se trouve dans tout le corpus (score de IDF faible).

Etant donné un nombre de catégories thématiques choisi à l'avance, il faut être en mesure de déterminer les meilleures valeurs pour les deux seuils $\theta_s$ et $\theta_T$. Pour cela, nous avons fait varier ces valeurs, ainsi que le nombre de catégories, et nous avons calculé à chaque fois le modèle LDA associé sur un échantillon d'apprentissage en relançant 20 fois l'algorithme afin de faire varier son initialisation (et donc réduire la chance de tomber dans un optimum local). L'échantillon d'apprentissage est constitué des deux tiers des 2385 réclamations de l'analyse. La mesure de perplexité, que l'on trouve dans l'article de (Blei et al., 2003), permet de confronter les modèles obtenus à un échantillon de test en se basant sur un calcul de vraisemblance. Pour ne pas être biaisée par la taille du vocabulaire qui varie en fonction des seuils utilisés, cette mesure doit être normalisée par la taille du vocabulaire (Corston-Oliver et Gamon, 2004). Il est ainsi possible de sélectionner automatiquement les seuils qui mènent aux résultats qui optimisent cette mesure, sans avoir recours à une expertise humaine.

**Deuxième étape : calcul du modèle.** Une fois déterminées les valeurs optimales pour les seuils permettant de filtrer le vocabulaire, il reste le problème de trouver le nombre "optimal" de catégories thématiques. Ce problème est lié à celui, très classique en apprentissage non supervisé, de déterminer le nombre optimal de classes (*clusters*). Pour ce faire, nous avons tout d'abord calculé la mesure de perplexité en faisant varier le nombre de catégories de 2 à 20, avec une relance de 20 fois pour chaque valeur. La courbe obtenue présentant une forme en "U" caractéristique en apprentissage automatique, nous en avons conclu que les valeurs trop faibles ou trop élevées étaient à écarter. Les valeurs trop élevées dégradent les résultats de la mesure de la même manière que pour le phénomène de sur-apprentissage (*overfitting*) constaté en apprentissage automatique supervisé. Parmi les valeurs intermédiaires restantes, et comme nous sommes dans le cadre d'un travail exploratoire mené sur un cas d'étude réel, nous avons finalement choisi de déterminer le nombre exact de thématiques en ayant recours à des experts métier.

Toutes les étapes de prétraitements et les analyses statistiques sont réalisées avec le logiciel de statistique R sous Windows [4].

# 4 Résultats des analyses

Pour des raisons de confidentialité, les résultats ne peuvent pas être présentés exhaustivement dans cet article. Nous ne présentons qu'une partie d'entre eux afin d'illustrer le processus.

4. http://cran.r-project.org/

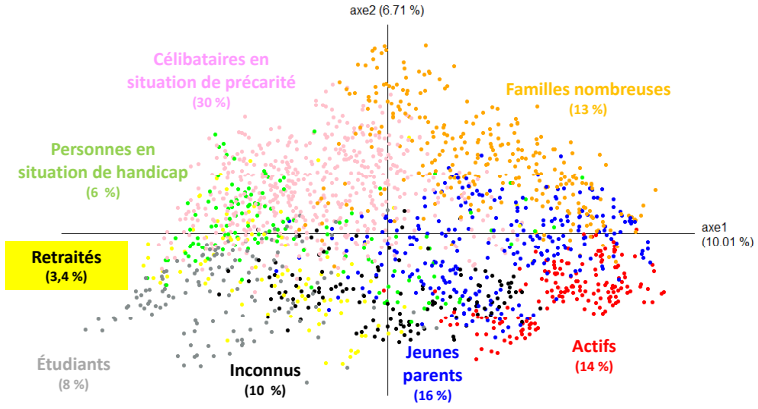


FIG. 2 – *Typologie des allocataires sur les deux premiers axes de l'ACM.*

## 4.1 Typologie des allocataires réclamants

Pour construire la typologie des allocataires réclamants, les dix premiers axes factoriels (qui expliquent 46 % de l'intertie) sont retenus comme variables pour la CAH. Pour la CAH, le critère du "saut" dans la mesure est maximum pour des valeurs de 3, 8 et 5 classes. L'évaluation des experts nous a permis de conclure que la typologie la plus informative est obtenue pour 8 classes d'allocataires. Il est possible de visualiser la typologie des allocataires sur le graphique synthétique de l'ACM en colorant chaque individu selon la catégorie à laquelle il appartient (cf. figure 2). A partir des caractéritiques communes des allocataires constituant chacune des classes, une étiquette est donnée à chaque classe.

A titre d'exemple, Les individus en gris se démarquent des autres par une forte proportion de personnes qui ont entre 20 et 25 ans, qui sont célibataires et qui perçoivent l'allocation au logement sociale. 73 % d'entre eux ont un quotient familial inférieur à 300 et tous ont un revenu brut annuel inférieur à 20 000€. Enfin, aucun d'entre eux n'a d'enfant et, en toute logique, aucun ne touche les prestations familiales. Nous interprétons cette classe comme correspondant à des "étudiants".

Les individus colorés en orange se caractérisent quant à eux par une absence de personnes sans enfants. Plus de 90% d'entre eux ont au moins trois enfants (67% avec trois enfants et 25% avec quatre enfants ou plus). Parallèlement 90% d'entre eux touchent le complément familial et 99% les prestations familiales. Ces allocataires sont âgés de 40 à 50 ans. On trouve très peu de célibataires et de personnes avec un fort quotient familial dans cette classe qui correspond aux "familles nombreuses".

## 4.2 Typologie des réclamations

Le protocole expérimental permet de sélectionner le meilleur modèle en 14 catégories et avec des valeurs de seuil respectivement de 0,98 et 0,03 pour $\theta_s$ (rareté) et $\theta_T$ (TF.IDF). L'un des avantages de la méthode LDA est de pouvoir identifier les mots qui caractérisent ces différentes thématiques (mots clefs principaux ou *top keywords* en anglais) et d'analyser plus qualitativement les thématiques : pour chaque thématique, nous précisons les dix premiers mots clefs, nous proposons une étiquette, et nous donnons une réclamation type dans laquelle nous soulignons la présence d'un ou plusieurs mots clefs.

A titre d'exemple, dans la première catégorie se retrouvent les mots-clefs suivants : **dossier, demande, droit, familiales, réponse, allocation, compte, été, prestations, pourriez**. Cette catégorie regroupe donc des documents qui abordent le thème des allocations familiales et on retrouve dans les textes des expressions comme "Allocation familiales" ou "prestations familiales". La deuxième catégorie est aussi relativement marquée ; elle contient des réclamations qui font suite à un changement de situation de l'allocataire avec les mots-clefs suivants : **mois, situation, été, caf, informations, allocataire, changement, dossier, reçu, compte**. Les réclamations de cette thématique abordent un "changement de situation" et/ou la prise en "compte [d']informations". Voici un exemple de texte de cette catégorie : "*madame monsieur suite au mail que j'ai recu le 20 mars 2010 de la part de la caf de lyon dont la copie se trouve en piece jointe je souhaiterai contester la decision qui a ete prise en effet le fait que je sois en collocation n'a pas ete enregistre par les services de la caf de lyon ce qui a entraine une retenue de mon aide au logement pour les mois de janvier 2010 hauteur de 239 62 euros et de fevrier 2010 hauteur de 45 euros de plus j'ai recu une notification de dette de 807 84 euros suite cette erreur je souhaiterai que ma situation ainsi que mon dossier soient regularise*".

## 4.3 Mise en correspondance des typologies

La typologie des allocataires a montré que ces derniers pouvaient être classés en huit groupes avec des caractéristiques différentes. La typologie des réclamations, elle, a fait émerger quatorze thématiques. Se pose alors la question suivante : certains de ces thèmes sont-ils le fait d'une catégorie particulière d'allocataires ? Les réclamations sur les prestations familiales ne sont-elles écrites que par les familles nombreuses, par exemple ? Pour répondre à ces questions il est possible de croiser les deux typologies.

Des analyses statistiques basées sur un test d'indépendance (p-value=0,0005) et une analyse factorielle des correspondances simples indiquent qu'il n'y a pas d'indépendance entre le type de réclamations et le type d'allocataires. Il existe en effet des thèmes de réclamations privilégiés selon les catégories d'allocataires. Par exemple, les familles nombreuses semblent réclamer davantage sur les thèmes du changement de situation, de l'allocation logement et du RSA. Les retraités réclament quant à eux principalement sur l'allocation logement (mais dans des proportions semblables aux autres allocataires) ainsi que sur le RSA et le montant des droits. Comme on peut s'y attendre, ils sont sous-représentés dans le thème des allocations familiales, leurs enfants étant des adultes.

En revanche, les analyses statistiques ne permettent pas de conclure qu'une catégorie de réclamation est imputable à une classe particulière d'allocataires. En effet, les personnes en situation de handicap et les célibataires apparaissent presque toujours en première position, mais ceci est dû au fait qu'ils sont plus nombreux que tous les autres allocataires réclamants.

## 5 Conclusion et perspectives

L'objet de l'étude présentée dans cet article est d'étudier l'opportunité d'utiliser des techniques de fouille de données pour réaliser une analyse sémantique des réclamations faites par les allocataires. L'étude a été menée dans une optique exploratoire.

Après avoir identifié les différentes sources de données, l'étude a mis en évidence l'importance et la difficulté de la phase de préparation des documents et des données. Nous avons proposé et réalisé une succession d'étapes pour traiter des problèmes de la reconnaissance de caractères, de la sélection des documents dactylographiés, de l'extraction du texte contenant le corps du document, de la correction des fautes d'orthographe, de l'identification des textes qui sont proprement dits des réclamations, et enfin de la construction automatique du vocabulaire de mots pour décrire ces réclamations. En se plaçant dans une optique nationale de traitement automatique des réclamations, nous avons cherché à automatiser ces différentes étapes afin que les agents de la CAF interviennent le moins possible dans l'opération. Nous avons aussi montré que grâce à des techniques de fouille de données, et en particulier de fouille de textes, il est possible de savoir à la fois qui sont les allocataires qui réclament et de savoir sur quelles thématiques portent les réclamations.

S'agissant d'une étude exploratoire réalisée dans un temps assez court, nous pouvons dès à présent lister des améliorations possibles de ce travail avec deux horizons.

A court terme, une amélioration consisterait à compléter les résultats obtenus dans la typologie des réclamations en lançant les algorithmes sur l'intégralité du volume de documents fournis, contenant ou non une réclamation. Une typologie de l'ensemble des documents permettrait de voir si les réclamations se distinguent significativement des autres documents et se retrouvent dans les mêmes catégories ou non. De plus, à l'heure actuelle, la seule étape manuelle du processus proposé est l'étiquetage des thématiques extraites par la méthode LDA. En utilisant des travaux développés dans notre laboratoire, une automatisation est possible et pourrait être intégrée sans rencontrer de grandes difficultés.

A plus long terme, plusieurs extensions peuvent être étudiées. Concernant la discrimination automatique des documents qui relèvent d'une réclamation, la solution proposée s'inspire de la démarche manuelle suivie par les agents de la CAF. Mais un nombre encore trop important de textes qui ne sont pas des réclamations sont retenus à tort. L'identification automatique des réclamations pourrait certainement être améliorée en utilisant d'autres techniques de fouille de données telles que l'apprentissage automatique supervisé. Dans cette étude, les typologies des réclamants et des réclamations ont été extraites de manière synchronique, c'est-à-dire que la dimension temporelle n'a pas du tout été prise en compte. Pourtant, il semble qu'il s'agisse d'un aspect très important dans la gestion d'une réclamation. Une autre extension de cette étude serait alors d'étudier s'il est possible de replacer l'allocataire réclamant dans une chronologie. Des techniques de fouille de données pourraient être utilisées, par exemple, pour extraire des "trajectoires" d'allocataires dans lesquelles s'inscrivent les réclamations.

## Références

Bae, S., S. Ha, et S. Park (2005). A web-based system for analyzing the voices of call center customers in the service industry. *Expert Systems with Applications 28*(1), 29–41.

Bíró, I., J. Szabó, et A. Benczúr (2008). Latent dirichlet allocation in web spam filtering. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pp. 29–32. ACM.

Blei, D., A. Ng, et M. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Cohen, W., P. Ravikumar, S. Fienberg, et al. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, pp. 73–78.

Corston-Oliver, S. et M. Gamon (2004). Normalizing german and english inflectional morphology to improve statistical word alignment. *Machine Translation : From Real Users to Research*, 48–57.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, et R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American society for information science 41*(6), 391–407.

Kohonen, T. (2001). *Self-organizing maps*, Volume 30. Springer Verlag.

Lcvenshtcin, V. (1966). Binary codes capable of correcting deletions, insertions. In *Soviet Physics-Doklady*, Volume 10.

Lebart, L., A. Morineau, et M. Piron (1995). *Statistique exploratoire multidimensionnelle*, Volume 2. Dunod Paris.

Ngai, E., L. Xiu, et D. Chau (2009). Application of data mining techniques in customer relationship management : A literature review and classification. *Expert Systems with Applications 36*(2), 2592–2602.

Paatero, P. et U. Tapper (2006). Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics 5*(2), 111–126.

Stauss, B. et W. Seidel (2004). *Complaint management : the heart of CRM*. Thompson/South-Western.

Steyvers, M. et T. Griffiths (2007). Probabilistic topic models. *Handbook of latent semantic analysis 427*(7), 424–440.

Velcin, J. et J. Ganascia (2007). Topic extraction with AGAPE. In *Proceedings of the conference Advanced Data Mining and Applications*, pp. 377–388. Springer.

# Summary

The management of the complaints is a fundamental element in the customer relationship management. It is the case for the French national family allowance fund (Caisse Nationale des Allocations Familiales) that wants to set up a national policy to facilitate this management. In this paper, we describe a process that automatically deals with the complaints coming from beneficiaries of the office of the Rhone Department. The available raw data require an important series of pretreatments in order to handle these data with a computer. Once the data cleaned, we use methods coming from both the multidimensional data analysis and the unsupervised machine learning in order to extract a typology of the complaints, based on their textual contents, and a typology of the claiming beneficiaries. Finally, we study the link between these two typologies.