Health-policyholder clustering using health consumption

Romain Gauchon, Stéphane Loisel, Jean-Louis Rullière

Noname manuscript No.
(will be inserted by the editor)

Health-policyholder clustering using health consumption

A useful tool for targeting prevention plans

Romain GAUCHON[1,2] · Stéphane LOISEL[1] · Jean-Louis RULLIERE[1]

Received: date / Accepted: date

Abstract On paper, prevention appears to be a good complement to health insurance. However, its implementation is often costly. To maximize the impact and efficiency of prevention plans these should target particular groups of policyholders. In this article, we propose a way of clustering policyholders that could be a starting point for the targeting of prevention plans. This two-step method mainly classifies using policyholder health consumption. This dimension is first reduced using a Nonnegative matrix factorization algorithm, producing intermediate health-product clusters. We then cluster using Kohonen's map algorithm. This leads to a natural visualization of the results, allowing the simple comparison of results from different databases. We apply our method to two real health-insurer datasets. We carry out a number of tests (including tests on a text-mining database) of method stability and clustering ability. The method is shown to be stable, easily-understandable, and able to cluster most policyholders efficiently.

Keywords Clustering · Health insurance · Kohonen's map · Nonnegative Matrix Factorization · Prevention.

1 Introduction

Prevention, as it reduces risk, would appear to be a useful tool for insurers. However, until recently European insurance companies have been wary of prevention, which was generally seen as a marketing product. The first ambitious prevention plan was initiated in 1997 by Discovery in South Africa. It then took until 2016 for a major prevention plan to appear in Europe: the Vitality program (arising from the merger of Discovery and Generali) allows policyholders to win points by adopting healthy lifestyles that can be converted into gifts or discount coupons. First

Romain Gauchon
romain.gauchon@addactis.com

[1]Université de Lyon, Université de Lyon 1, Laboratoire de Sciences Actuarielle et Financière, Institut de Science Financière et d'Assurances (50 Avenue Tony Garnier, F-69007 Lyon, France)

[2]addactis in France

launched in Germany, this program arrived in France in 2017, creating controversy about the use of health data by a private insurance company.

There are a number of reasons why private companies might be reluctant to develop prevention in France. It is first difficult for insurance companies to achieve high participation rates in prevention plans. And even if they do, they could produce severe adverse selection: individuals who are the most interested in prevention plans are those with specific health risks ([5]). Moreover, policyholders can easily switch between health-insurance providers, rendering prevention investment less efficient for insurance companies ([19]).

It is in addition difficult to choose who will benefit from the plan. Data Protection (regarding health data in particular) has always been a touchy subject in France. For example, mutual health-insurance companies cannot ask for any health information when a policyholder takes out a new contract. Therefore, a useful algorithm to predict health outcomes has to be unsupervised.

The new European General Data Protection Regulation (GDPR) has limited the possible uses of data. In particular, health data are explicitly considered as very sensitive. Article 22 of the GDPR states that private companies cannot take decisions that significantly affect an individual only based on an automated process. As such, the targeting of prevention plans by insurance companies is complicated.

There are nevertheless some exceptions to the GDPR, allowing insurance companies to use data in order to target prevention. First, the GDPR has introduced the notion of individual agreement. If an individual has agreed to a certain use of his personal data for a clearly-defined purpose, the use of these data is allowed. If she does not agree, it is still possible to use these data for regulatory compliance (such as creating financial reserves) or the production of aggregate statistics.

The clustering method we propose here purports to meet all of these requirements. Clustering is frequently used in insurance. For example, credibility theory, which leads to the bonus-malus system, is based on the idea that there exist hidden policyholder clusters (e.g. [10]). Clustering is also used to identify fraud (Derrig and Ostaszewski [15]) and classify risks (Yeo et al. [53]). It has been used to improve general linear models by Verrall and Yaboukov [49]. In the health-insurance context, Ghoreyshi and Hosseinkhani classify policyholders using the k-means algorithm [18]; Peng et al. apply clustering algorithms to detect fraud in a health-insurance dataset [43]; and Kuo et al. cluster Taiwanese respondents in a health-insurance dataset matched to medical information in order to identify the relationships between illnesses [29].

However, to the best of our knowledge, the clustering of policyholders based on the relationship between claims has not been addressed. Clustering methods in insurance do not thus fully exploit the available data, as they do not use factors like the individual's treatment programme when ill. Analyses that do exploit this kind of information can be found in the fields of text mining and natural language processing, where the clustering of texts is very common (e.g.[47], [6] and [23]).

To cluster texts, we need to capture the meaning of words. A computer can guess this meaning by looking at their context: for example whether "Doctor" is usually combined with "medication", "apple" or "eagle" (e.g. [35], [34]). In the

health field, the question becomes "Do individuals undertaking kinesitherapy also undertake radiography?", which is very similar.

Text-clustering methods often start in the same way: they first pre-process the data to obtain a frequency matrix. This matrix contains one line by text, and one column for each word found in the body of the text. However, this matrix is of high dimension, notably affecting the quality of classic clustering methods via the dimension curse (e.g. [1]).

The analysis of high-dimension data has been a prolific research area (e.g. [54], [20], [24]). One classic method of dealing with this consists in reducing the dimension before clustering (e.g. [2], or [12]). When the dimension is reduced using a singular-value decomposition (SVD), the method is called Latent Semantic Analysis [14]. Of course, other dimension-reduction algorithms can be used. For example, Mote et al. use a Nonnegative Matrix Factorization algorithm (NMF) to reduce the dimension and a self-organizing map to cluster brain-tumor segmentation [36]. To the best of our knowledge, these kinds of methods have not yet been used for the clustering of health policyholders.

How can we apply these methods in an insurance context? In the end, a text is similar to a policyholder, with policyholder health consumption playing the role of words.

The goal of this paper is thus to present the first clustering method based only on policyholder health consumption, in order to help prevention targeting. The resulting clusters capture particular health profiles. In order to do so, we use a similar process to that in Mote et al., [36]. The dimension is reduced by the NMF and policyholders are clustered using a self-organizing map. This method is then applied to two real insurance databases. We demonstrate how to carry out each algorithm step via the analysis of a database covering 20 000 women. As this technique has never before been applied to insurance data, and very little elsewhere, we also apply it to a text-mining data set, in order to set out the limits of and potential issues with this clustering process.

The remainder of the paper is organized as follows: Section 2 defines the notation we use and Section 3 sets out the main algorithms used for clustering policyholders and the databases to which we apply this method. Section 4 describes the clustering process and the tests that are carried out, as well as the final results. Last, Section 5 discusses the method and proposes some possible extensions.

2 Notation and definitions

This section briefly describes the notation, definitions and main characteristics of the mathematical tools and algorithms used in this article.

Let $n, m \in \mathcal{N}$. In the following, we consider $x, y \in \mathbb{R}_+^n$. We first need to define three distances, which are useful to calculate the score function.

Euclidean distance: We denote the Euclidean distance by $d(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2}$ and the associated norm by $\|\|_2$. We also denote the L_1 norm by $\|\|_1$.

Kullback-Leibler similarity: We denote Kullback-Leibler similarity (or relative Entropy) by $L(x,y) = \sum_{i=1}^n x_i ln(\frac{x_i}{y_i})$.

Cosine similarity: We denote cosine similarity by $c(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$. This dissimilarity is often used in text mining to compare two documents.

We also define inertia and the R^2 coefficient, which is a well-known information-quality measure, as follows.

Inertia: Let $x^1, ..., x^k \in \mathbb{R}_+^n$ be a cluster C. Let $\overline{x} = \frac{\sum_{i=1}^{n} x^i}{n}$ be the gravity center. Let d be a distance. We call the inertia of the cluster the quantity $I_C = \frac{\sum_{i=1}^{n} d(x^i - \overline{x})}{n}$.

Given $C_1, ..., C_k$ a partition of C, we define $R^2 = 1 - \frac{\sum_{i=1}^{k} I_{C_i}}{I_C}$.

The R^2 coefficient measures the proportion of information that is captured by a cluster. This depends strongly on the distance chosen. This latter is usually the Euclidean distance but, as shown by Huang [23], this may be inappropriate for some kinds of data such as text. We here use the cosine similarity, as this is more appropriate for our data, which are similar to those from text mining.

We last define two supervised classification-quality measures, which we use to check our method on a supervised dataset. The following formulae are those used by Huang ([23]).

Purity and entropy: Let $x^1, ..., x^k$ be k observations. There exists a partition $\mathcal{C}^1$ of $x^1, ..., x^k$ into l subsets. We want to reproduce the partition $\mathcal{C}^1$. Using a clustering algorithm, we obtain a new partition $\mathcal{C}^2$ into l different clusters. We define the confusion matrix $CM \in \mathcal{M}_{l,l}(\mathbb{N})$ such that $CM_{i,j} = n_{i,j}$, the number of observations in the initial class i now clustered into the new class j.

We define purity as $Purity = \frac{1}{k}\sum_{i=1}^{l} max_j n_{i,j}$.

We define the entropy of class i as $E_i = \frac{1}{log(N)} \sum \frac{n_{i,j}}{n_{i,.}} log(\frac{n_{i,j}}{n_{i,.}})$.

Finally, we define entropy as $E = \sum_{i=1}^{l} \frac{n_{i,.}}{N} E_i$.

3 Algorithms

This section sets out the main algorithms used in this article: the Nonnegative Matrix Factorization and Kohonen's map. It also presents the health-insurer data.

3.1 NMF algorithm:

This algorithm is a dimension-reduction method. First introduced by Paatero and Tapper ([39]) and Lee and Seung [32], it is almost unknown in the insurance sector. The only application of which we are aware is Nesvijevskaia and Taudau, who announced that they had used this method at the 17th Rencontre MutRé [38].

However, this method is widely used in Medicine to analyze the human genome (e.g. [9], [31]) and in text mining to extract features (e.g. [32], [42], [25]). The NMF can also be used as a clustering method (e.g. [28])

For all $n, m \in \mathbb{N}$, we denote by $\mathcal{NM}(n, m)$ the set of nonnegative matrices with n rows and m columns. Let $n, m, k \in \mathbb{N}$, $V \in \mathcal{NM}(n, m)$. The purpose of the NMF is to find $W \in \mathcal{NM}(n, k)$, $H \in \mathcal{NM}(k, m)$ such that $V \approx WH$.

To do so, a number of algorithms have been proposed (e.g. Lee and Seung ([32], [33]), Brunet et al. ([9]), Pascual-Montano et al. ([40]), Badea ([4]), Kim and Park ([26]) and Pauca et al. ([41])). Except for the latter, all of these have been tested. We will only present here the algorithm that performs the best: the "snmf/l" algorithm created by Kim and Park.[1]

This method combines the cost function proposed by Pauca et al. ([41]) with the concept of sparseness, first introduced in an NMF algorithm by Hoyer ([22]).

Kim and Park propose to find $\min_{W,H}(\frac{1}{2}[\|V - WH\|_2^2 + \alpha \|H\|_2^2 + \beta \sum_{i=1}^{n} \|W(i,.)\|_1^2])$, with β being a coefficient controlling for the sparseness of W and α a coefficient reflecting H's smoothness, as suggested by Pauca et al. .

It is important to note that the cost function is not convex. Therefore, the snmf/l algorithm is only able to find local optima and is sensitive to the initial values of W and H. The classic way to deal with this is to randomly initialize the two matrices a number of times and then compare the resulting local minima. Boutsidis and Gallopoulos address this by initializing both matrixes with a nonnegative SVD (nSVD, see Boutsidis et al. [8]). Other initialization methods have been suggested by Langville et al. ([30]).

Starting from a random value, Kim and Park propose the following update rules to find a local minimum:

$$H_{n+1} = \min_{H \geq 0} \left\| \begin{pmatrix} W_n \\ \sqrt{\beta} e_{1\times k} \end{pmatrix} H - \begin{pmatrix} V \\ 0_{1\times n} \end{pmatrix} \right\|_2^2,$$

$$W_{n+1} = \min_{W \geq 0} \left\| \begin{pmatrix} H_{n+1}^T \\ \sqrt{\alpha} I_k \end{pmatrix} W - \begin{pmatrix} V \\ 0_{k\times m} \end{pmatrix} \right\|_2^2,$$

with n, m being the dimensions of A, k the final dimension, $e_{1\times k}$ a vector of height k containing only 1, and I_k the identity matrix. This method, from Van Banthem and Keenan [48], is called ANLS (Alternative Nonnegativity constrained Least Squares) and guarantees convergence. The stopping criterion is based on the optimality criterion of Karush Kuhn Tucker.

3.2 Kohonen's map algorithm:

Kohonen's map, also called a self-organizing map (SOM), is a clustering method based on neural networks [27].[2]. In this network, every neuron is arranged according to a given topology, usually a two-dimensional grid with a hexagonal disposition. This way, each neuron has neighboring neurons.

[1] The snmf/l implementation of the R package "NMF", developed by Gaujoux and Seoighe [17], was used in the analysis presented here.

[2] The R package "Kohonen", developed by Wehrens et al. ([51]), was used in the analysis presented here.

Say that we wish to classify N policyholders using a a n-neuron SOM. To start with, each neuron is assigned a random weight m_i (a well-known alternative is to choose the starting points via a PCA, however Akinduko et al. show that this is not suitable for non-linear datasets ([3]). For each learning iteration t, a random policyholder is chosen. It is then possible to determine the neuron that best represents this policyholder, by solving $c = \min_{j \in [\![1,n]\!]} \|x - m_j(t)\|$. The neuron c is called the best machine unit (BMU).

Once the BMU is determined, its weight is adjusted in order to improve policyholder representativeness: $m_c(t+1) = \alpha(t)(x - m_c(t))$. The coefficient $\alpha(t)$ is the learning rate. This falls over time, so that learning is fast at the beginning and meticulous at the end.

In order to create an influence zone, the BMU's neighbors' weights are also changed. We define the neighborhood function $h(c, i, t)$. This function falls with the distance between the neuron i and the BMU c. It also falls over time: at the beginning many neurons are adjusted, while at the end only few are. Last, the weight of neuron i becomes $m_i(t+1) = m_i(t) = \alpha(t)h(c, i, t)(x - m_c(t))$. This process is repeated several times.

3.3 Data

Health insurers usually possess two different databases. The first is called the policyholder database, and contains all the information the insurer possesses about the policyholder: age, sex, contract details, contact information and so on. This information is typically used to analyze insurance-porfolio profitability.

This database is systematically matched to a second one: the health-consumption database. This latter contains all the information the health insurer needs to reimburse the policyholder when she buys a health product:[3] the date, amount and nature of the expense, sometimes called the medical act. The elements in this database, and the nature of the consumption in particular, can vary from one insurer to another: some reimburse a product but others not, and some insurers have more detailed information than others. This database can be large: from one million entries for a small mutual health-insurance company to over one billion for national health-insurance systems. This base may also depend on the national health system. For example, in France, medication for long-term diseases (such as cancer) are fully reimbursed by the national public insurer, and do not necessarily appear in these databases.

We set out the results in detail in Section 4.4, but illustrate the algorithm used here in a small database of around 20 000 women,[4] aged 62[5] or over, and observed for one year. This includes over 500 000 different health reimbursements, and 160 different health products. Most of the entries concern extra charges (such

[3] We use the term health product for every item of health expenditure that may be refunded by the insurer (such as GP visits, nights at the hospital, medication and glasses).

[4] There are actually 19 727 women: we say 20 000 as shorthand.

[5] The legal retirement age in France is 62.

as additional fees for night consultations), apart from extra charges for home-care services that are dropped. We merge identical health products that appeared separately due to spelling mistakes. Last, glasses and contact lenses are merged into a single optical product, although this does not change the results. After these changes, the database contains 80 different health products.

Transforming the health-consumption database into a frequency matrix does not suffice. It is well-known in text mining that some words, such as "are", "the" and "a" are so common that they carry no useful information for clustering purposes. It is thus useful to assign less weight to these kinds of words and more weight to more uncommon words. One classic way to do so is the tf-idf method (see [45] for a general presentation).

In the health-insurance context, pharmacy, and to a lesser extent general and specialist practitioner, consultation is so common that it is not really useful for policyholder clustering. To improve cluster quality, we apply the tf-idf method and logarithm function to our database. In our case, the clusters of policyholders obtained from the logarithm treatment are more homogenous.

As the combined use of the NMF and Kohonen map method has almost never been carried out, the 20-newsgroups dataset is also used to test the method. This is a text-mining dataset. We use the training test detaset, which covers 11 293 texts from 20 different newsgroups. As our goal is not text mining, we consider the pre-treated dataset of Cardoso Cachopo [11] (the "no-short" dataset).

4 The clustering process

4.1 Dimension reduction using the NMF algorithm

The clustering method here contains two steps: the dimension is first reduced and then policyholders are clustered. This sub-section discusses dimension reduction.

We obtain a matrix with around 100 dimensions after pre-processing the health database. This is too large for traditional clustering methods, such as k-means, which perform less well when there are over 15 dimensions or so (e.g. [7]).[6]

We use the sNMF/l algorithm to reduce dimensions, applied to the pre-treated frequency matrix V. After calibrating the algorithm by examining the silhouette, cophenetic and dispersion, the final space covers 20 dimensions. We thus obtain two matrices, W and H (see Section 3.1), respectively of dimensions 20 000x20 and 20x80. It is important to note that both of these can be interpreted.

The matrix H contains one column for each health product, and 20 lines. There are two ways to make this matrix easier to understand. We can first normalize each row, by dividing each coefficient by the sum of all the other row coefficients. This normalized matrix can be used to create a heat map[7] (see Figure 1).

[6] In the datasets used here, dimension reduction dramatically improves clustering.

[7] The health product "Legal copayment" may be unfamiliar to the reader. In the French health system, many health products are partially reimbursed by the public insurer, "l'Assurance Maladie". The price of health products is fixed by Law (for example, a GP con-

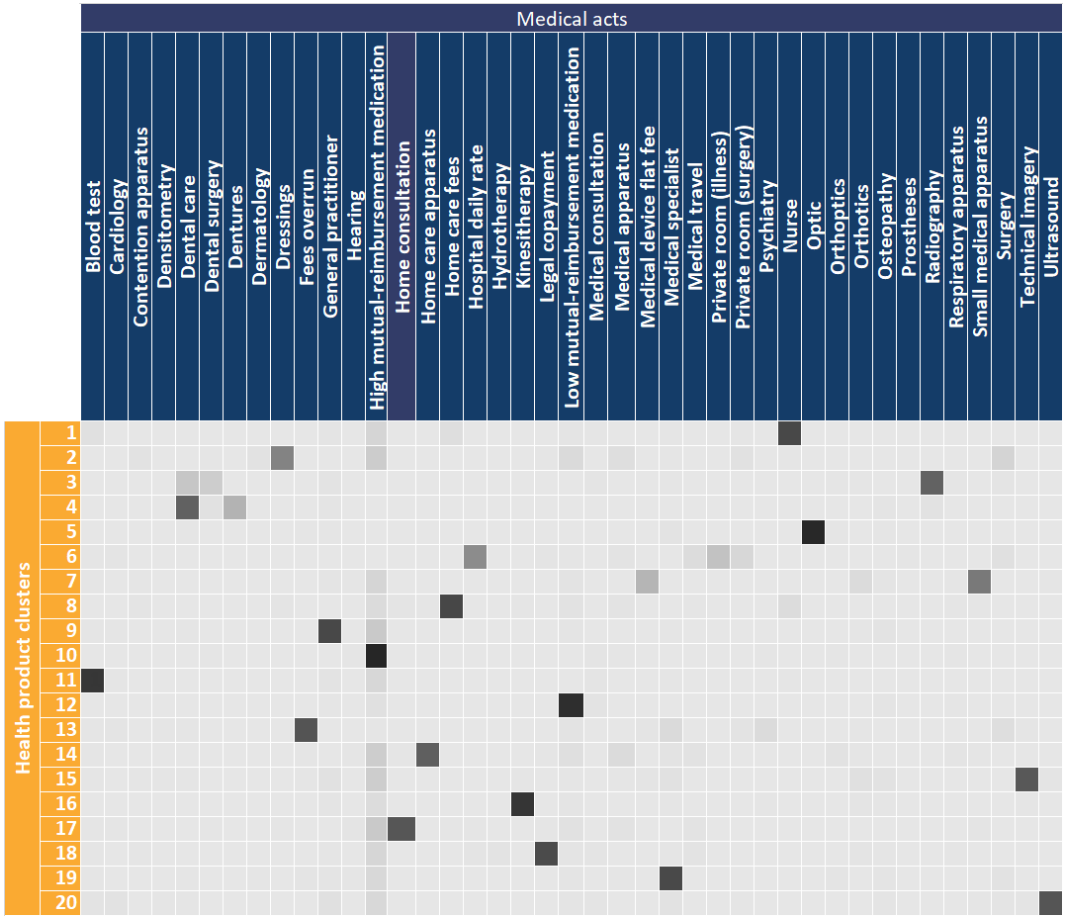

Fig. 1 The horizontal normalization of H. Only the most common health products are shown for clarity reasons.

This matrix shows that each of the new dimensions can be understood as a health-product cluster. For example, Dimension 7, covering "Medical apparatus flat fees", "Orthotics" and "Small medical apparatus", is a medical-apparatus cluster. We call each of these new dimensions a health-product cluster (HPC). Each HPC can be easily interpreted.

We should emphasize that, even though "Technical imagery" looks similar to "Radiology", they do not appear in the same HPC. Whereas a human may have made the mistake of merging these two health products, the algorithm distinguishes between them. This shows that the merging applied before using this

sultation costs 25 Euros). However, the public insurer does not refund all of this amount (only 16.5 Euros for GPs) in order to limit health consumption. We here call the 25-16.5 gap the "legal copayment". Moreover, GPs are allowed to charge higher fees that are not covered by the public insurer. The reimbursement of the legal copayment is usually covered by private insurance.

algorithm can substantially affect the final results, and thus should be carried out with caution.

Figure 1 only illustrates the most common health products: others, such as "medical travel", are not shown. As shown in sub-Section 4.4, the most important health products are not necessarily the most common: less-frequent health products (such as "orthoptics") can be significant.

In order to bring out the role of less-common products, and thus to specify the meaning of each HPC, it is also possible to normalize H by columns (see Figure 2). If each HPC is seen as a cluster, this normalization shows to which cluster each health product belongs.

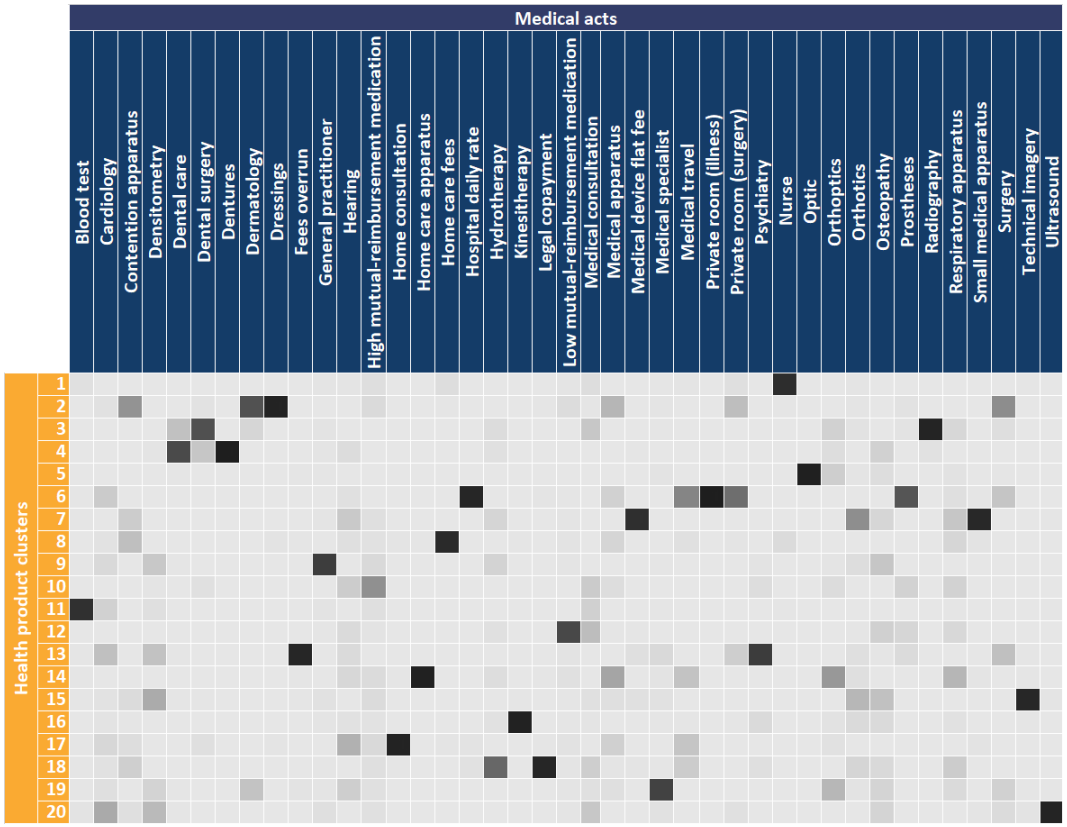

Fig. 2 The vertical normalization of H. This heatmap helps to interpret the meaning of each HPC.

Using this new heat map allows for a better understanding of each HPC. For example, HPC 3 can now be seen to cover dental surgery, so that "Radiography" means "Dental radiography". HPC 15, containing "Densitometry", "Technical imagery", "Orthotics" and "Osteopathy", can be seen as a fracture HPC. As these

results come from data on older respondents, this HPC reflects those who have had falls.

Once we have understood the H matrix, matrix W is easy to read: this contains one line for each policyholder and 20 columns, one for each new dimension. As we interpret the latter as HPCs, W shows HPC consumption for each policyholder. We will call H the HPC matrix and W the policyholders matrix.

We compare the snmf/l[8] method to two other standard reduction methods: the SVD and the PCA algorithms. By fixing the final dimension at 20, and using a Kohonen map for clustering (see Section 4.2), we compare the R^2 coefficients from the PCA and NMF classifications; tf-idf pre-processing performs very poorly, whereas NMF with the logarithm treatment performs better than SVD. The former is also easier to understand, but more fickle. We decide to retain the NMF method with random initialization and logarithm pre-processing.

4.2 Clustering using Kohonen's map

Once the dimension has been reduced, it is possible to cluster policyholders using the policyholder matrix W.

We use the classic Kohonen map method. The neighborhood function is linear and the learning rate decreases linearly. The neurons are disposed on a hexagonal grid with 20 lines (25 neurons by lines). To reduce the number of classes, Hierarchical Agglomerative Clustering (HAC) is carried out [37]. After examining the dendrogram, 17 final classes are retained. The matrix W is first scaled by lines (which produces better results than scaling by columns).

To interpret these classes, it is possible to calculate the centroid of each class, using the policyholder matrix W. Since each new dimension is interpreted as an HPC, it is easy to interpret each cluster (see Figure 3). The final Kohonen map appears in Figure 4. Note that a cluster "fracture" is obtained, which can be read as individuals who are likely to suffer from falls, as the database here covers older respondents.

We first remark that some of the neurons are empty (the neurons with a 0 in Figure 3). We try to avoid this in traditional approaches, as it means that some clusters are empty so that there are too many clusters. However, it is acceptable when self-organizing maps are combined with a HAC: the number of neurons is not the number of clusters. Here, empty neurons mainly mark the edge between clusters and low-density areas. To remove empty neurons, we would need to drastically reduce the number of neurons, and cluster quality would suffer.

The visual size of the clusters in the figure is closely correlated with their real size. Cluster 5, "everyday care", is thus the most-populated.

[8] As noted above, there exist many NMF algorithms. We here test six of them: those proposed by Lee and Seung (Lee, [32]), Brunet et al. (Brunet, [9]), Pascual-Montano et al. (nsNMF, [40]), Badea (Offset, [4]), and the two of Kim and Park (snmf/l and snmf/r, [26]). The "snmf/l" algorithm yields one of the best results, while being significantly faster. This is the method that we use.

Policyholders clusters	Blood test	Dental care / denture	Dental surgery	Fees overrun	General Practitioner	Home care apparatus	Home care fees	Home consultation	Hospitalization	Kinesitherapy	Legal copayment	High mutual reimbursement medication	Medical device	Medical specialist (with prescription)	Nurse	Optic	Radiography	Surgery	Ultrasound	Low mutual reimbursement medication	
	Health product clusters																				
1	0.4	0.1	0	0.1	0.3	0.2	1.8	0.6	0.3	0.3	0.3	0.3	0.2	0.2	1.7	0.1	0.1	0.4	0.1	0.4	Home care
2	0.5	0.1	0.1	0.2	0.4	0	1.1	0.1	0.2	0.2	0.2	0.3	0.2	0.3	1.1	0.1	0.2	0.4	0.2	0.4	Nurse home care
3	0.2	0.1	0.1	0.2	0.3	0	0	0	0.1	0.1	0.1	0.4	0	0.4	0.1	0.2	0.1	0.7	0.1	0.4	Surgery
4	0.3	0.1	0.1	0.2	0.3	0.9	0	0.1	0	0.1	0.1	0.4	0.1	0.3	0.1	0.2	0.1	0.1	0.1	0.4	Home care apparatus
5	0.2	0.1	0	0	0.3	0	0	0	0	0	0	0.5	0	0.1	0.1	0	0	0	0	0.2	Everyday care
6	0.2	0.1	0	0.1	0.2	0	0.1	0.8	0.1	0	0.1	0.5	0.1	0.2	0.1	0	0.1	0.1	0.1	0.4	Home consultation
7	0.3	0.1	0.1	0.3	0.2	0.2	1.4	0.4	0.3	1.4	0.2	0.3	0.2	0.3	0.3	0.1	0.2	0.2	0.1	0.4	Home kinesitherapy
8	0.3	0.1	0.1	0.1	0.4	0	0	0	0	0	0.6	0.4	0	0.2	0.1	0	0.1	0	0.1	0.3	Legal copayment
9	0.3	0.1	0	0.3	0.3	0	0	0	0	0	0	0.5	0	0.3	0.1	0	0	0	0.2	0.3	Everyday care with fees overrun
10	0.2	0.1	0	0.1	0.4	0	0	0	0	0	0.1	0.4	0.6	0.2	0.1	0.2	0.1	0.1	0.1	0.3	Medical apparatus
11	0.2	0.2	0.1	0.2	0.4	0	0	0	0	1.3	0.1	0.4	0.1	0.3	0.1	0.1	0.2	0	0.1	0.4	Kinesitherapy
12	0.1	0	0	0.1	0.1	0	0	0	0	0	0	0.1	0	0.1	0	0	0	0	0	0.3	Occasional consumer
13	0.3	0.1	0	0.2	0.4	0	0	0	0	0	0	0.4	0	0.3	0.1	0	0.5	0	0.2	0.3	Radiography
14	0.3	0.1	0	0.1	0.3	0	0	0	0	0	0	0.4	0	0.3	0.1	0	0	0	0.5	0.3	Ultrasound
15	0.2	0.1	0	0.2	0.3	0	0	0	0	0	0.1	0.4	0	0.3	0.1	1.1	0.1	0	0.1	0.3	Optic
16	0.2	0.5	0.6	0.1	0.3	0	0	0	0	0	0	0.4	0.1	0.2	0.1	0.1	0.1	0	0.1	0.3	Dental care
17	0.2	0.1	0	0.1	0.1	0	0	0	0.8	0	0.2	0.3	0.1	0.1	0	0	0	0	0	0.3	Hospitalization

Fig. 3 The centre of gravity of the 17 final clusters, calculated using policyholder-matrix coefficients.

This map can also illustrate the correlations between clusters. For example, clusters 1, 2 and 7 are "home-care" clusters.

We can test the consistency of the results by looking at policyholder profiles in each cluster in terms of variables that were not used in the clustering process, such as age and total medical expenses (see Figure 5). As expected, individuals in the "occasional consumers" cluster do not cost insurers much, whereas those in the "home-care" clusters do. We also check that the average age for comfort-care clusters, such as "Optic" and "Dental", is below that in the "home-care" clusters (e.g. [13]). As such, the method appears to produce consistent results.

This approach can also help to choose prevention plans. For example, cluster 17 ("Hospitalization") is expensive for private insurers. It may be of interest to analyze more closely the profile of individuals in this class in order to target prevention plans.

One simple approach would be to target all those with one specific type of health consumption, which we refer as the statistical method. For example, we could target all those with psychiatric expenditure in a prevention plan for psychiatric illness. However, this method leads to quite different classes (see Appendix 1 for the detailed results). Most of the time, the NMF method produces clusters with one main health consumption, whereas the clusters obtained from the statistical method are less centered on one type of consumption. Moreover, the NMF clusters are obtained in an unsupervised way, whereas the statistical method requires an arbitrary edge to cluster policyholders.

Last, this approach is consistent with the European data regulation GDPR. The GDPR distinguishes between two cases. When the policyholder gives her consent to her data being treated for prevention purposes, individuals in each class can

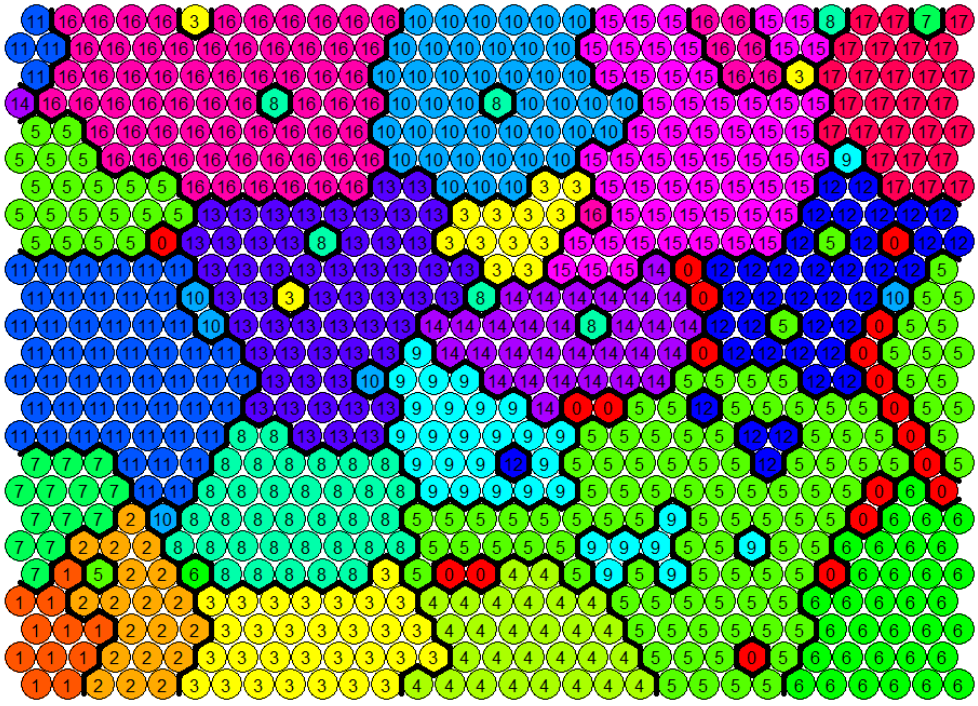

Fig. 4 An example of one of the Kohonen maps. The associated cluster meanings can be found in Figure 3. Each color represents a different cluster. The red neurons are empty.

be targeted (individuals in cluster 1 can be proposed a specific prevention plan). Insurance companies do not necessarily have the consent of all policyholders. However, it is still possible to aggregate data in order to obtain cluster characteristics, and thus obtain a general objective characterization (for example, people in home-care clusters are in general over 80, so that we can target tertiary prevention plans at those over 80, or primary prevention plans towards those aged 70 to 80). In our databases, we only know policyholder age, sex and family situation: with more complete information we could expand the statistical analyses of each cluster.

Another possibility from this method when the database includes covers a long-enough time period is to calculate clusters for a number of periods and see how cluster characteristics change over time.

The Kohonen map algorithm may be sensitive to initialization and input data order, due to the multi-label context (for example, Appendix 2 shows a map obtained in the same way as Figure 4, changing only the original seed). However, the cluster meanings are very similar between the different maps. One way to tackle this issue would be to construct a number of different Kohonen maps based on the same NMF result. It is then possible to consider policyholders who appear at least once, or twice, or every time in a given cluster. We thus obtain for each policyholder the empirical likelihood of belonging to this cluster.

Cluster	Meaning	Age	Private Insurer refund	Public insurer refund	Number of policyholders
15	Optic	69	497	284	1419
9	Everyday care with fees overrun	70	145	301	870
12	Occasional consumer	70	62	52	1541
13	Radiography	70	161	407	1393
14	Ultrasound	70	164	350	802
16	Dental care	70	413	448	1771
5	Everyday-day care	71	97	202	4926
8	Legal copayment	71	220	359	970
11	Kinesitherapy	71	307	603	1265
10	Medical apparatus	73	220	399	1069
2	Nurse home care	74	550	794	295
3	Surgery	74	355	532	959
4	Home care apparatus	75	315	533	577
17	Hospitalization	77	1126	277	679
1	Home care	81	792	1389	169
6	Home consultation	81	260	452	826
7	Home kinesitherapy	82	937	1014	196

Fig. 5 Cluster statistics.

4.3 Other tests

We carry out a number of additional tests to better understand the results produced by this method.

We first test the model's capacity to identify strong but infrequent correlations. We do so by adding a randomly-chosen number between 3 and 10 to the "Keratotomy", "Hydrotherapy accommodation" and "Orthoptics" consumption of n random policyholders. "Keratotomy" and "Hydrotherapy accommodation" are consumed only infequently, whereas "Orthoptics" is much more common. We then re-run the NMF algorithm. The goal is to identify the minimum n for which the NMF algorithm detects the new correlation. It turns out that only 60 modified policyholders are necessary (out of 20 000 individuals in the database) in order to detect this correlation.

We also test result consistency via the analysis of a text-mining dataset, the 20-newsgroups dataset. This contains 18 821 documents from 20 different newsgroups. These are typically split into a training dataset of 11 293 documents and a test dataset of 7528 documents. To reduce computing time, we here only use

the training dataset. This dataset is labeled, which allows us to calculate objective error measures. It has been extensively studied in the literature regarding all of the classic text-mining tasks, such as word embedding (e.g. [21], [50]), unsupervised clustering (e.g. [16], [23]) and supervised classification (e.g. [46], [44]). We downloaded the "no-short" dataset from Ana Cardoso Cachopo's website ([11]).

As text mining is not one of the goals of this paper, the results presented below come from the first run of the algorithm, without trying to calibrate the model or improve the results. The dimension is first reduced to 60 before clustering, and the frequency matrix is pre-processed using the tf-idf method.

We already know that the 20-newsgroups dataset contains 20 different clusters. In the NMF / Kohonen method, the number of classes is established by analyzing a dendrogram. It is of interest to note that from the dendrogram we would have chosen 3 or 19 clusters.

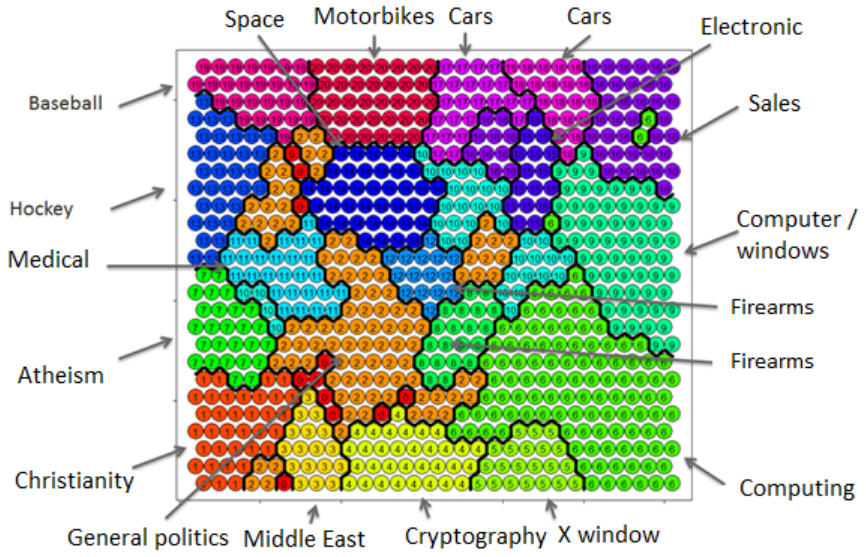

Fig. 6 Kohonen's map using the 20-Newsgroups Dataset. Cluster 10 cannot be construe.

From the Kohonen map (Figure 6), we see that clusters 2 and 10 are spread out. Moreover, clusters 6 and 9 seem significantly larger than the others. Their purity score confirms that they are less homogeneous than the other clusters (purity is shown in Figure 8). Apart from these four clusters and cluster 18, purity is acceptable. Global purity is 62% and total entropy is 0.4, which is significantly better than the results obtained by Huang from the same dataset [23], even though we do not aim to achieve a good score.

Comparing Figures 7 and 8, even though the algorithm does not identify all of the documents in a given cluster, the resulting clusters are still reliable. This means

that if we want to identify all of the policyholders with psychiatric medication, this algorithm is not very appropriate. However, if we identify a psychiatric class, this is reliable enough to justify the targeting of a prevention plan.

Newsgroup	Clusters best representing the newsgroup	Document % in the most representative cluster	Document % in the second most representative cluster
Christianity	1	76%	10%
Various politics	2	72%	8%
Middle East politics	2, 3	49%	45%
Cryptography	4	85%	7%
X window system	5, 6	55%	37%
Windows	6, 9	69%	21%
Atheism	7, 1	62%	20%
Firearms	8, 12, 2	38%	32%
IBM computers	10, 6	63%	23%
Digital graphics	6	66%	11%
Medical	12, 2	49%	25%
Various religion	1	44%	30%
Hockey	13	88%	7%
Space	14	77%	8%
Electronic	15, 6	24%	20%
General sales	16	82%	10%
Cars	17, 18	52%	15%
Mac computers	10, 6, 15	50%	22%
Baseball	19	75%	8%
Motorbike	20	87%	4%

Fig. 7 Newsgroup reconstitution capacity

Clusters	Mainly represented newsgroup	Cluster purity	% of total documents clustered in the class
1	Christianity, atheism, various religion	74%	6%
2	Various politics, Middle East politics, Medical, Firearms	27%	11%
3	Middle East politics	96%	3%
4	Cryptography	94%	5%
5	X window system	90%	3%
6	Digital graphics, Windows, X window system, IBM computers, Mac computers, Electronic	24%	15%
7	Atheism	65%	4%
8	Firearms	76%	2%
9	IBM computers, Mac computers, Windows	39%	9%
10	None	20%	3%
11	Medical	91%	3%
12	Firearms	88%	2%
13	Hockey	93%	5%
14	Space	92%	4%
15	Electronic	75%	2%
16	General sales, Mac computers	57%	8%
17	Cars	88%	3%
18	Cars	33%	2%
19	Baseball	96%	4%
20	Motorbikes	93%	5%

Fig. 8 Cluster purity

To summarize, this method produces acceptable results in the 20-newsgroup dataset. Most of the clusters represent a specific newsgroup. However, this method cannot differentiate between very similar newsgroups, such as IBM and Mac computers. This produces large clusters containing most of the documents between which the method cannot differentiate.

This clustering method is thus able to construct meaningful policyholder clusters. However, large classes (such as the everyday-care cluster) are heterogeneous and should not be used to target prevention plans: they contain policyholders who cannot be differentiated by the algorithm.

4.4 Discussion of the results

The method has been applied to databases from two different insurance companies: a collective database (CB) and an individual database (IB). For each database, four different splits are carried out (women over 62, women between 16 and 62, men between 16 and 62 and men over 62), producing a total of eight different databases (and eight different clusterings). Splitting the database in this way helps to reduce the heterogeneity between populations, as well as speed up the process. The CB and IB contain policyholders with top-range and mid-range market contracts respectively. Moreover, policyholders with zero health consumption are removed. Figure 9 contains descriptive figures for the eight populations.

		Individual database		Collective database	
		Headcount	Average age	Headcount	Average age
Men	-62 years	17153	41	15415	43.7
	+62 years	7949	73.9	12691	71.8
Women	-62 years	38386	42.8	15365	43.4
	+62 years	19727	71.8	13269	74.1

Fig. 9 Database statistics

As the two databases come from different companies there are some small differences. For example, the CB does not distinguish medical specialists with or without prescriptions. The CB also does not separate fee overruns and the legal copayment from the price of health consumption. On the other hand, the IB does not separate biological from blood tests, for example.

Taking these differences into account, it is possible to analyze the HPCs from the eight databases (see Figure 10). 22 HPCs are found in each of the CB populations, with an analogous figure of 20 for the IB database. The HPCs obtained are essentially the same, with the main differences being due to database construction. The method is thus resistant to a change in the database.

However, some other differences do merit discussion. In the CB, the HPCs "Respiratory apparatus" and "Home care apparatus" are more important than

in the IB. This can be due to care refusal: these are expensive products that are better-covered by the collective top-range market contract. Also, the "Hospitalization" HPC only concerns older people in the individual base but everybody in the CB. This is due to the "Legal copayment" HPC, which also contains "Hospitalization" health consumption for younger people in the IB. As the "Legal copayment" health product does not exist in the CB, the HPC becomes "Hospitalization" there.

In both databases, the "Osteopathy", "Orthoptics" and "Psychiatry" HPCs do not concern older people. Osteopathy is a modern practice of which the older are less aware, who prefer going to a kinesiologist instead. It is notable that "Orthoptics"[9] is an HPC on its own. This is a quite narrow health product (in both bases it is consumed by fewer than 2% of policyholders and represents under 0.12% of total expenditure). Orthoptics mainly covers children, which explains why it is an HPC for younger people, with parents paying for their children. Finally, to understand why "Psychiatry" does not appear as an HPC for older people, it is important to underline that dementia in France is usually treated by neurologists rather than psychiatrists. However, burnout, mainly affecting working people, can be treated by psychiatrists. Breakdowns of this kind amongst seniors are often not diagnosed.

Some home-care HPCs are particular to older people, due to dependency. However, it is of interest to note that the "nurse" HPC typically contains some home-care consumption and exists for both younger and older people.

Last, we can also see that the "respiratory apparatus" HPC concerns only men, as they are more subject to sleep apnea than women. The "Ultrasound" clusters are found for both men and women, as this is not only used for pregnancy but also for heart, blood and musculoskeletal-system radiography.

Once the HPCs have been discussed, we can cluster using Kohonen's map. The clustering is carried out using the classic Kohonen's map with a linear neighborhood function. Gaussian neighborhoods usually produce maps with more empty neurons, although this is not always the case.

We first analyze the differences between cluster interpretation (see Figure 11 for a short summary).

"Everyday care" and "occasional consumer" are both clusters with no specific consumption and are joint in the CB dataset. Individuals in these clusters are mostly in good health or refuse treatment.

The HPC cluster-meaning analyses are similar (see the comments on "Psychiatry", "Osteopathy", "Orthoptics" and "Respiratory apparatus"). Note that the "home care" cluster appears in all eight databases. "Orthotics / medical apparatus" does not form a cluster for women and "home care apparatus" does not form a cluster for men in the CB. For the younger men in the CB, a "surgery" cluster in addition to the "dressings" cluster is found, producing two surgery-like clusters. Surprisingly, in both the IB and CB, younger men also have a "blood test" cluster,

[9] According to Wikipedia, *"Orthoptics is a profession allied to eye care professions whose primary emphasis is the diagnosis and non-surgical management of strabismus (wandering eyes), amblyopia (lazy eye) and eye movement disorders".*

	Individual database				Collective database			
	Men		Women		Men		Women	
<62years old		Medical specialist (without prescription) Orthoptics Osteopathy Psychiatry			Surgery	Dental care Osteopathy Psychiatry		Orthoptic
<62years old		Blood test Dental care / Denture / Surgery Drugs Fees overrun General practitioner Kinesitherapy Legal copayment Medical apparatus Medical specialist Nurse Optic Radiography Surgery Ultrasound			Dental care Respiratory apparatus	Blood / Biology test Dental preventive care / Radiology / Denture Drugs Dressings General practitioner Kinesitherapy Hospitalization Medical specialist Nurse Optic Orthotics / medical apparatus Radiography Technical medical procedures Ultrasound		Home care apparatus
>62 years old								
>62 years old	Respiratory apparatus	Home care fees Home consultation Hospitalization		Home care apparatus		Home general practitioner Home care apparatus Personal medical room		Home nurse Orthopedic kinesitherapy

Fig. 10 Summary of the HPCs

which is not the case in the other databases. Last, "dentures" appears as a cluster only in the CB, mainly due to differences in contract quality.

The analysis of average age and expenditure by cluster also provides considerable information. For example, the average age in the "Dental care" and "Optic" clusters is lower than the average age for older people in both databases. For the younger, the average age in the "ultrasound" clusters is lower and contains more policyholders for women than for men, due to motherhood. The "Psychiatry" clusters cover more women than men, which is also a well-known medical fact (see [52]).

Other findings are more difficult to explain. For young women the "Kinesitherapy" clusters have higher average age than for men, and both are higher than the overall average age. For the IB, the cluster "Medical specialist without prescription" has a lower average age than "Medical specialist with prescription". Last, the average age in the global "Hospitalization" clusters for the older are above the overall average age, but below the average age in the "Home care" clusters.

From a more global point of view, the "everyday care" and "occasional consumer" clusters are very large. On the contrary, there usually exist some very narrow clusters with fewer than 100 policyholders. These usually contain very archetypal consumers and thus can be used to target small prevention plans.

5 Conclusion

We have here presented a method for clustering policyholders based on their health consumption. This first reduces the dimension problem by carrying out a Nonnegative Matrix Factorization. This stage improves the results and helps to interpret the clustering, by identifying meaningful health-product clusters.

	Individual database			Collective database		
	Men		Women	Men		Women
<62years old	Blood test	Medical specialist (without prescription) Orthoptics Osteopathy Psychiatry		Blood test Surgery	Dental preventive care / Radiography Hospitalization (with and without personal room) Orthotics / Medical apparatus Osteopathy Psychiatry	Heavy home apparatus Hospitalization Orthoptic
		Dental care Everyday care Home care Kinesitherapy Medical apparatus Legal copayment Optic Radiography Surgery Ultrasound		Dental preventive care / radiography Orthotics / medical apparatus Respiratory apparatus	Biology test Dental care Denture Dressings Everyday care Home care Kinesitherapy Optic Ultrasound	Home care apparatus
>62 years old	Respiratory apparatus	Home care / Kinesitherapy Home consultation Hospitalization Occasional consumer Nurse home care	Everyday care with fees overrun Home care apparatus		Home care device Home general practitioner Hospitalization with personal room Hospitalization without personal room	Orthopedic kinesitherapy Home nurse Hospitalization / Home care

Fig. 11 Summary of the final clusters

In the second stage, policyholders are clustered using Kohonen's map. The Kohonen map algorithm offers a readable visualization of the results. This allows the simple comparison of clusterings carried out on different databases. Moreover, these can be interpreted as different risk clusters. The method has been subjected to a number of tests, revealing its reliability and the quality of the results. Except for the "everyday care" clusters (composed of occasional consumers or very particular policyholders), most clusters are pure and can be used in practice. These clusters are established using common insurance data, as possessed by every health insurer. By constituting clusters, we aggregate the data and so respect the legislation. The method applied here can thus be used to target prevention plans aimed at policyholders, and our tests have shown that this process is accurate when we do not have a clear idea of the prevention plans to be instigated.

There are a number of ways in which the method can be improved. First, this is a mono-label clustering, and multi-label clustering would usually be more appropriate for health-risk profiles. This multi-label clustering can be carried out via fuzzy clustering (such as fuzzy c-means) instead of Kohonen's map.

Moreover, the NMF method has considerable advantages, but one main disadvantage: due to the initialization requirements, it can be quite slow. In order to accelerate dimension reduction, other methods may be more appropriate (e.g. word-embedding methods).

The method that we have applied here does not take into account the temporality of health consumption. Knowing that a policyholder consumes a great

deal over a short period or regularly throughout the year could be useful. We are currently working on these last two points.

The results from this method are very dense, and are sometimes difficult to interpret. Medical advice on these kinds of results would be useful in understanding the potential scope of this method.

Last, this method is not particular to health-policyholder clustering, and could for example also be applied to customer clustering.

Acknowledgements The authors would like to thank Alexandra Barral for useful comments over the duration of this research, and Nabil Rachdi for technical advice. They are also grateful to Addactis in France for providing the data, and Jean-Pascal Hermet, Louis Bachaud, Steve Briand, Astrid Servajean and Andrew Clark for re-reading the paper. This research was carried out in the framework of the Chair Prevent'Horizon, supported by the risk foundation Louis Bachelier and in partnership with Claude Bernard Lyon 1 University, Addactis in France, AG2R La Mondiale, G2S, Covea, Groupama Gan Vie, Groupe Pasteur Mutualité, Harmonie Mutuelle, Humanis Prévoyance and La Mutuelle Générale.

References

1. Aggarwal, C.C., Yu, P.S.: Finding generalized projected clusters in high dimensional spaces, vol. 29. ACM (2000)
2. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications, vol. 27. ACM (1998)
3. Akinduko, A.A., Mirkes, E.M., Gorban, A.N.: Som: Stochastic initialization versus principal components. Information Sciences **364**, 213–221 (2016)
4. Badea, L.: Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. In: Biocomputing 2008, pp. 267–278. World Scientific (2008)
5. Beaulieu, N., Cutler, D.M., Ho, K., Isham, G., Lindquist, T., Nelson, A., O'Connor, P.: The business case for diabetes disease management for managed care organizations. In: Forum for Health Economics & Policy, vol. 9. De Gruyter (2006)
6. Berkhin, P.: A survey of clustering data mining techniques. In: Grouping multidimensional data, pp. 25–71. Springer (2006)
7. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is âĂIJnearest neighborâĂİ meaningful? In: International conference on database theory, pp. 217–235. Springer (1999)
8. Boutsidis, C., Gallopoulos, E.: Svd based initialization: A head start for nonnegative matrix factorization. Pattern Recognition **41**(4), 1350–1362 (2008)
9. Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the national academy of sciences **101**(12), 4164–4169 (2004)
10. Bühlmann, H., Gisler, A.: A course in credibility theory and its applications. Springer Science & Business Media (2006)
11. Cardoso-Cachopo, A.: Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa (2007)
12. Cheng, C.H., Fu, A.W., Zhang, Y.: Entropy-based subspace clustering for mining numerical data. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 84–93. ACM (1999)
13. Darblade, M.: Analyse de profils de consommation et tarification des futures garanties sur-complémentaire santé. Master's thesis, ISFA (2015)
14. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
15. Derrig, R.A., Ostaszewski, K.M.: Fuzzy techniques of pattern recognition in risk and claim classification. Journal of Risk and Insurance **62**(3), 447–482 (1995)
16. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on Machine learning, p. 29. ACM (2004)

17. Gaujoux, R., Seoighe, C.: A flexible r package for nonnegative matrix factorization. BMC bioinformatics **11**(1), 367 (2010)
18. Ghoreyshi, S., Hosseinkhani, J.: Developing a clustering model based on k-means algorithm in order to creating different policies for policyholders in insurance industry. International Journal of Advanced Computer Science and Information Technology (IJACSIT) **4**(2), 46–53 (2015)
19. Herring, B.: Suboptimal provision of preventive healthcare due to expected enrollee turnover among private insurers. Health Economics **19**(4), 438–448 (2010)
20. Hinneburg, A., Keim, D.A.: Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. pp. 506–517. 25 th International Conference on Very Large Databases (1999)
21. Hinton, G.E., Salakhutdinov, R.R.: Replicated softmax: an undirected topic model. In: Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, A. Culotta (eds.) Advances in Neural Information Processing Systems 22, pp. 1607–1614. Curran Associates, Inc. (2009)
22. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of machine learning research **5**(Nov), 1457–1469 (2004)
23. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, pp. 49–56 (2008)
24. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing, pp. 604–613. ACM (1998)
25. Jones, B.W., Chung, W.: Topic modeling of small sequential documents: Proposed experiments for detecting terror attacks. In: Intelligence and Security Informatics (ISI), 2016 IEEE Conference on, pp. 310–312. IEEE (2016)
26. Kim, H., Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**(12), 1495–1502 (2007)
27. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78**(9), 1464–1480 (1990)
28. Kuang, D., Choo, J., Park, H.: Nonnegative matrix factorization for interactive topic modeling and document clustering. In: Partitional Clustering Algorithms, pp. 215–243. Springer (2015)
29. Kuo, R., Lin, S., Shih, C.: Mining association rules through integration of clustering analysis and ant colony system for health insurance database in taiwan. Expert Systems with Applications **33**(3), 794–808 (2007)
30. Langville, A.N., Meyer, C.D., Albright, R., Cox, J., Duling, D.: Initializations for the non-negative matrix factorization. In: Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining, pp. 23–26. Citeseer (2006)
31. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al.: Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature **499**(7457), 214 (2013)
32. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788 (1999)
33. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp. 556–562 (2001)
34. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
35. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013)
36. Mote, S.R., Baid, U.R., Talbar, S.N.: Non-negative matrix factorization and self-organizing map for brain tumor segmentation. In: Wireless Communications, Signal Processing and Networking (WiSPNET), 2017 International Conference on, pp. 1133–1137. IEEE (2017)
37. Murtagh, F.: Interpreting the kohonen self-organizing feature map using contiguity-constrained clustering. Pattern Recognition Letters **16**(4), 399–408 (1995)
38. Nesvijevskaia, A., Taudou, B.: La data science au service de la prévention santé et prévoyance : nouveaux paradigmes - 17eme rencontre mutré, 14-15 november - nantes. Tech. rep., Malakoff Mederic (2016)
39. Paatero, P., Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (1994)

40. Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehmann, D., Pascual-Marqui, R.D.: Nonsmooth nonnegative matrix factorization (nsnmf). IEEE transactions on pattern analysis and machine intelligence **28**(3), 403–415 (2006)
41. Pauca, V.P., Piper, J., Plemmons, R.J.: Nonnegative matrix factorization for spectral data analysis. Linear algebra and its applications **416**(1), 29–47 (2006)
42. Pauca, V.P., Shahnaz, F., Berry, M.W., Plemmons, R.J.: Text mining using non-negative matrix factorizations. In: Proceedings of the 2004 SIAM International Conference on Data Mining, pp. 452–456. SIAM (2004)
43. Peng, Y., Kou, G., Sabatka, A., Chen, Z., Khazanchi, D., Shi, Y.: Application of clustering methods to health insurance fraud detection. In: Service Systems and Service Management, 2006 International Conference on, vol. 1, pp. 116–120. IEEE (2006)
44. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th international conference on machine learning (ICML-03), pp. 616–623 (2003)
45. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation **60**(5), 503–520 (2004)
46. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: Advances in neural information processing systems, pp. 1289–1296 (2008)
47. Steinbach, M., Karypis, G., Kumar, V., et al.: A comparison of document clustering techniques. In: KDD workshop on text mining, vol. 400, pp. 525–526. Boston (2000)
48. Van Benthem, M.H., Keenan, M.R.: Fast algorithm for the solution of large-scale nonnegativity-constrained least squares problems. Journal of Chemometrics: A Journal of the Chemometrics Society **18**(10), 441–450 (2004)
49. Verrall, R.J., Yakoubov, Y.H.: A fuzzy approach to grouping by policyholder age in general insurance. Journal of Actuarial Practice **7**, 181–204 (1999)
50. Wang, D., Cui, P., Zhu, W.: Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1225–1234. ACM (2016)
51. Wehrens, R., Buydens, L.M., et al.: Self-and super-organizing maps in r: the kohonen package. Journal of Statistical Software **21**(5), 1–19 (2007)
52. W.H.O., et al.: Depression and other common mental disorders: global health estimates (2017)
53. Yeo, A.C., Smith, K.A., Willis, R.J., Brooks, M.: Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry. Intelligent Systems in Accounting, Finance and Management **10**(1), 39–50 (2001)
54. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: an efficient data clustering method for very large databases. In: ACM Sigmod Record, vol. 25, pp. 103–114. ACM (1996)

6 Appendix

6.1 Appendix 1: Comparing the clusters obtained from the proposed method to those from a very basic approach

Health consumption type	Class 1 - Home care	2 - Nurse home care	3 - Surgery	4 - Home care apparatus	5 - Everyday care	6 - Home consultation	7 - Home kinesitherapy	8 - Legal copayment	9 - Everyday care - fees overrun	10 - Medical apparatus	11 - Kinesitherapy	12 - Occasional consumer	13 - Radiography	14 - Ultrasound	15 - Optic	16 - Dental care	17 - Hospitalization
Drugs	409	375	281	300	187	288	319	253	236	252	262	65	253	213	194	199	163
Blood test	74	105	43	47	27	46	52	49	44	42	44	17	52	50	36	38	36
Medical apparatus	155	120	96	184	47	66	207	55	75	154	95	34	64	57	52	84	105
Other	76	31	24	22	7	15	38	56	24	17	26	8	14	14	17	14	23
Nurse / medical auxiliaries	423	60	4	13	2	7	25	2	3	2	3	1	2	2	2	2	1
Surgery	200	235	223	50	11	27	265	22	28	26	34	13	38	28	32	25	167
Dental care	21	88	62	89	67	42	80	86	25	57	107	36	67	81	62	369	41
Home care	319	37	6	11	1	93	112	3	2	3	4	1	2	2	2	2	5
General practitioner	69	124	86	81	46	45	77	84	91	82	101	28	97	83	71	77	33
Hospitalization	347	220	122	102	29	126	495	70	50	51	61	31	68	58	55	44	983
Kinesitherapy	77	48	12	22	3	11	391	8	5	6	257	3	7	5	5	7	8
Optic	113	146	162	181	20	42	109	46	21	170	142	32	32	35	985	138	38
Denture	19	108	74	129	88	60	127	108	29	70	141	28	86	119	70	492	46
Radiography	44	70	44	34	5	24	59	34	9	43	57	6	102	83	33	53	13
Medical specialist	39	58	56	47	11	20	61	30	70	30	46	13	45	38	39	33	18
Total	2386	1824	1296	1311	553	913	2417	905	712	1006	1380	315	931	868	1655	1580	1679
Size	169	295	959	577	4926	826	196	970	870	1069	1265	1541	1393	802	1419	1771	679

Fig. 12 Detailed statistics for all classes.

Health consumption type	At least one consumption of : Drugs	Blood test	Medical apparatus	Surgery	Dental care	Home care	General practitioner	Hospitalization	Kinesitherapy	Optic	Radiography	Medical specialist	Global mean
Drugs	227	274	302	311	241	332	239	270	290	241	266	265	216
Blood test	40	78	53	60	46	65	45	53	52	45	54	53	39
Medical apparatus	77	92	266	124	88	123	79	118	112	92	98	94	74
Other	18	23	27	29	22	28	20	24	31	23	23	25	18
Nurse / medical auxiliaries	8	10	16	24	6	38	7	11	13	6	8	9	7
Surgery	47	66	81	272	51	111	54	97	85	67	71	85	45
Dental care	92	104	103	106	284	87	103	101	116	113	168	107	92
Home care	11	14	20	23	8	65	9	14	23	8	11	11	11
General practitioner	70	91	90	101	86	84	89	91	105	84	98	97	68
Hospitalization	94	115	167	204	86	196	85	228	144	98	110	110	96
Kinesitherapy	28	34	45	43	33	57	31	37	245	33	40	39	27
Optic	137	153	165	183	169	123	153	192	159	843	169	204	136
Denture	120	139	135	134	367	118	134	131	158	140	219	139	118
Radiography	35	48	50	55	49	45	41	52	63	43	84	53	34
Medical specialist	32	42	45	63	40	40	38	48	52	48	48	72	31
Total	1036	1281	1564	1730	1575	1513	1127	1466	1647	1884	1467	1362	1013
Size	18770	9761	5516	3277	6363	3245	15099	8339	2213	3193	8072	8441	19727

Fig. 13 Consumption if consuming at least some of a particular health product, and overall consumption.

6.2 Appendix 2: An example of another map obtained from the same data

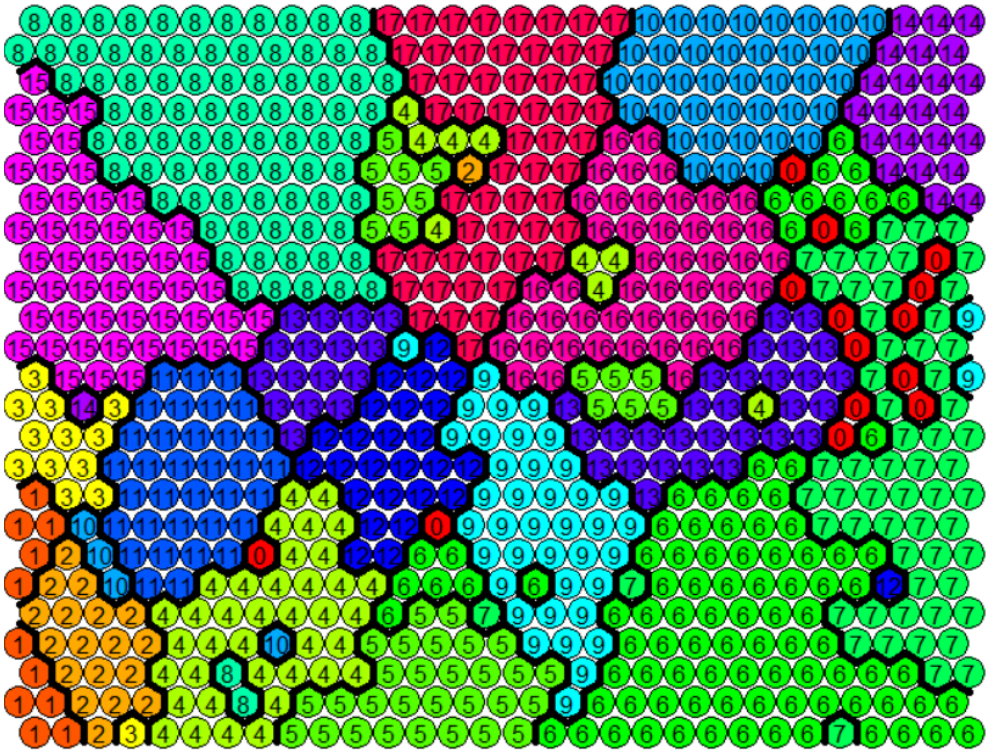

Fig. 14 Self-organizing map obtained from the same data as in Figure 4, using a different seed.