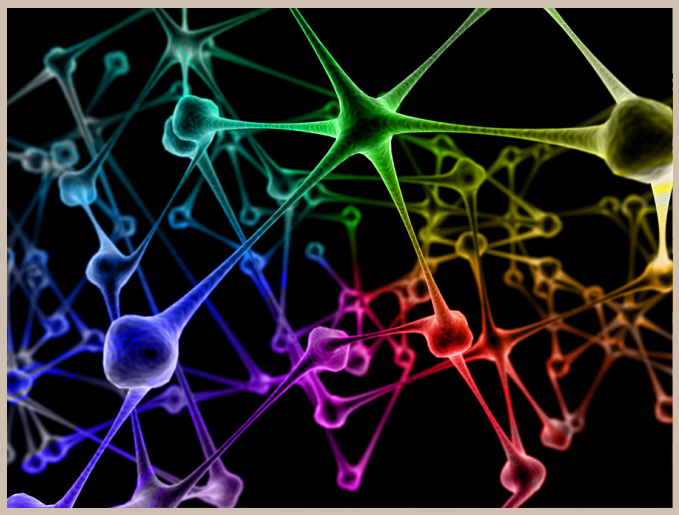
École Nationale de la Statistique et de l'Administration Économique

Mastère Spécialisé Gestion des Risques et Finance de marché : année 2015-2016

GROUPE BPCE

Prévision du passage au contentieux des contrats sains par méthodes neuronales



Quentin JAMMES

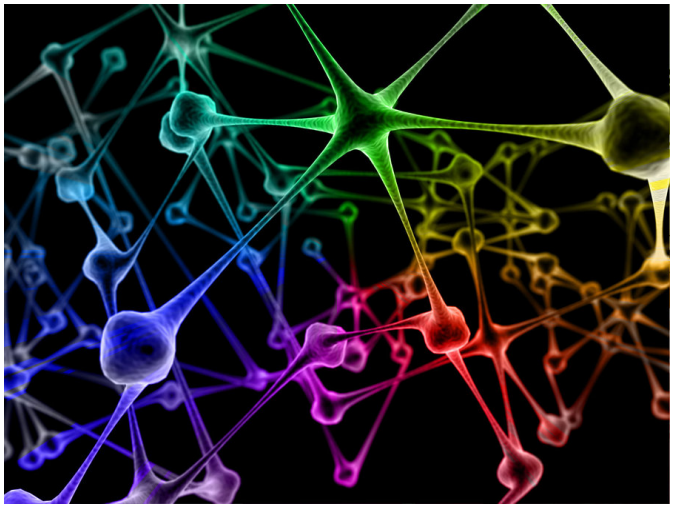
Mots clés : Réseaux de neurones, cartes de Kohonen, analyse de données, dynamic time warping

École Nationale de la Statistique et de l'Administration Économique

Mastère Spécialisé Gestion des Risques et Finance de marché : année 2015-2016

GROUPE BPCE

Prévision du passage au contentieux des contrats sains par méthodes neuronales



Quentin JAMMES

Mots clés : Réseaux de neurones, cartes de Kohonen, analyse de données, dynamic time warping

Remerciements

Je souhaite adresser des remerciements spéciaux à Smaïl IBBOU et Stéphanie Moré pour m'avoir donné l'opportunité d'effectuer ce stage au sein de BPCE, et à Sarah GAUDIN pour son encadrement et sa promptitude à répondre à mes questions.

J'adresse également de chaleureux remerciements à mes collègues Julien, Samir, Philippe, Hajar, Nicolas, Angela et Huan pour leur aide, leur intérêt et leur implication dans mon travail.

Table des matières

1 Présentation du groupe BPCE

Le groupe BPCE est le deuxième groupe bancaire de France, opérant au travers de ses deux entités majeures Banque Populaire et Caisse d'Épargne. Le groupe compte 108 000 employés et environ 35 millions de clients, et se positionne sur tous secteurs bancaires – solutions d'épargne et d'investissement, services de cash management, financement, assurance et gestion d'actifs – grâce à plusieurs filiales telles que Natixis, Banque Palatine, Crédit foncier de France ou BPCE International.

Ce groupe résulte de la fusion du groupe Caisse d'Épargne et du groupe Banques Populaires en 2009 dans une optique de robustesse face à la crise des subprimes. Ces deux réseaux fondateurs gardent leur enseigne et leur indépendance mais mettent en commun des services de back-office tels que la direction des risques au sein de laquelle j'ai réalisé mon stage, dans l'unité Validation du pôle Analyses consolidées et Modèles.

2 Introduction

Après la banqueroute de la banque allemande Herstatt en 1974, la régulation bancaire devient un enjeu principal au niveau international. Dans cette optique, le comité de Bâle est créé par les banques centrales du G10 en vue de recommander les meilleures pratiques en matière de gestion des risques, et afin de faire converger les régulations nationales. L'une des décisions majeures de ce comité est la mise en place du ratio de Cooke qui oblige les banques à avoir un ratio de solvabilité minimum. La régulation Bâle II étend le champ d'application de ce ratio en 1999 aux risques de crédit, de marché, et opérationnels. Les trois piliers de cette régulation sont les suivants :

— Exigence de fonds propres : utilisation de méthodes quantitatives pour modéliser le risque ;

— Surveillance individuelle de la gestion des risques des banques ;

— Transparence : les banques doivent communiquer à propos de la composition de leur capital, de leurs risques, de la composition et de la notation de leurs portefeuilles de crédits.

Afin de quantifier le capital requis par la régulation vis-à-vis du risque de crédit, la banque peut adopter différentes approches. La première et la plus simple est l'approche standard, qui consiste en l'adoption des modèles fournis par le régulateur. Ces modèles sont simples, largement conservateurs en termes de risque et intègrent des paramètres estimés en dehors de la banque. Au contraire, la banque peut choisir l'approche interne, utilisant ainsi ses propres modèles – validés par le régulateur – pour estimer ses risques et ses paramètres (notamment probabilité de défaut, pertes sachant défaut et exposition en cas de défaut). Cette approche est plus complexe du fait qu'elle nécessite des départements de modélisation et de validation, une gestion saine de données, et la formation de son personnel à l'utilisation de ces modèles. Cependant, des modèles plus performants induisent une baisse de l'exigence en fonds propres, ce qui en justifie les coûts.

3 Les paramètres bâlois

Les trois principaux paramètres balois sont l'exposition au défaut EAD, la probabilité de défaut PD et la perte sachant défaut LGD.

L'exposition au défaut d'un contrat désigne le montant susceptible d'être sujet au défaut pour ce contrat. C'est-à-dire l'intégralité de son bilan (B) si le contrat plus le taux de croissance de l'utilisation de ce contrat jusqu'au défaut (CCF) multiplié par son hors-bilan (HB) ($EAD = B + CCF * HB$) où

$$CCF = \frac{max(U(d) - U(t), 0)}{Umax - U(t)}$$

avec $U(t)$ l'utilisation (en montant) du contrat à l'instant t et U_{max} le montant maximal d'utilisation.

La probabilité de défaut est la probabilité qu'un contrat sain tombe en défaut, défini dans le cadre bâlois comme 3 impayés ou 90 jours de retard sur un impayé.

La perte sachant défaut correspond à la perte effective liée à un contrat tombant en défaut, c'est-à-dire à son bilan diminué des recouvrements actualisés et augmenté des coûts de recouvrements (coûts structurels c_s et coûts actualisés de gestion C) :

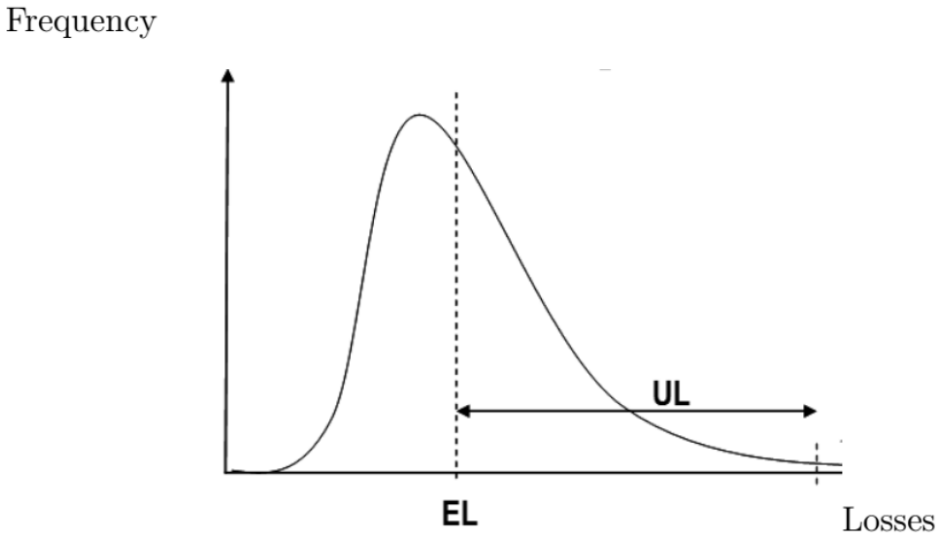
$$LGD = 1 - \frac{\sum_{j \geq 0} \frac{R_j - C_j}{(1+r)^j}}{EAD} + c_s$$

À partir de ces paramètres, on calcule la perte attendue

$$EL = PD * LGD * EAD$$

que l'on considère comme un coût à couvrir par la marge commerciale. La perte inattendue UL,

en revanche, est définie comme le montant d'une perte intervenant dans un cas sur mille (ie. une survie de la banque à un niveau de 99.9%).



C'est pour couvrir ce risque que sont mis en place les fonds propres, dont le besoin est défini par $0.08RWA$, le RWA étant les expositions au défaut pondérées $EAD * RW$.

La pondération utilisée est définie par

$$RW = \left(LGD * \Phi \left(\frac{1}{\sqrt{1-\rho}} \Phi^{-1}(PD) + \frac{\rho}{\sqrt{1-\rho}} \Phi^{-1}(0.999) \right) - PD * LGD \right) * 12.5 * 1.06$$

où $\Phi(.)$ est la fonction de répartition d'une loi normale centrée réduite et où ρ est un paramètre de corrélation inhérent aux modèles bâlois dont le but est de capturer la corrélation entre les défauts dus à un facteur systémique commun. Cette corrélation est fixée, dans l'approche interne, à 0.15 pour de l'immobilier, à 0.04 pour des crédits renouvelables, et supposé être une fonction décroissante de la probabilité de défaut, de 0.16 à 0.03 :

$$\rho = 0.03 \left(\frac{1 - e^{-35PD}}{1 - e^{-35}} \right) + 0.16 \left(\frac{1 - e^{-35PD}}{1 - e^{-35}} \right)$$

4 Présentation du sujet

Dans le cadre d'un chantier sur le calcul de l'ELBE (Expected Loss Best Estimate) pour les contrats de crédits aux particuliers, les modèles de LGD doivent être réétudiés. Empiriquement, la LGD a une densité quasiment bimodale : très proche de 0 lorsque les contrats parviennent à effectuer un retour en sain, et très proche de 1 (voire même supérieure à 1) lorsque le contrat passe au contentieux. Être en mesure de prévoir si un contrat risque de passer au contentieux en cas de défaut est alors un outil puissant d'optimisation du provisionnement et de sélection à l'octroi. L'objectif est donc ici de se détacher de la formule classique des pertes attendues (EL), $EL = EAD*PD*LGD$ pour une formule proche de $EL = EAD*PD*P_{CTX}*LGD_{CTX}$ (négligeant les coûts structurels du défaut). Nous pouvons en effet distribuer $EL = EAD * PD * LGD$ en $EL = EAD*PD*(P_{CTX}*LGD_{CTX} + P_{NCTX}*LGD_{NCTX})$. En raison du caractère quasiment bimodal de la LGD, on a $LGD_{NCTX} \approx 0$, d'où la simplification $EL = EAD*PD*P_{CTX}*LGD_{CTX}$; CTX et $NCTX$ signifiant respectivement contentieux et non contentieux.

En plus de l'intérêt évident en termes de gestion des provisions, les modèles d'octroi de crédit sont aujourd'hui centrés autour de la notion de défaut. Il est toutefois possible que certaines populations soient susceptibles de passer en défaut, mais également susceptible de revenir en sain, là où d'autres populations peuvent avoir une probabilité de défaut plus faible mais une prababilité de passage au contentieux conditionnelle au défaut bien plus importante. Est donc également en jeu une meilleure allocation des crédits par une meilleure identification des risques.

Cette problématique a de plus pour objectif de démontrer la performance des méthodes neuronales pour ce type de tâches, et de former ainsi un précédent à BPCE. Il s'agira de réaliser une

carte de Kohonen des contrats vis-à-vis du contentieux dans la continuité des travaux de IBBOU [1]. Il s'agira d'un côté d'explorer l'espace des contrats et de l'autre d'en développer un aspect prédictif pour les contrats sains.

1. IBBOU, 1992, Classification, analyse des correspondances et méthodes neuronales, Thèse

5 Environment de travail

Pour réaliser ce stage, j'ai eu à ma disposition

— Un ordinateur opérant sous Windows 7 avec SAS et R, relié au réseau BPCE et à un serveur SAS

— Un ordinateur opérant sous Windows 7 avec R relié au réseau BPCE

— Un ordinateur opérant sous Windows 7 spécifiquement dédié au logiciel R prêté par l'unité Modélisation

— L'entrepôt Risques de BPCE contenant les informations sur les contrats, les événements les affectant, les clients et leurs garanties

— Logiciels utilisés : SAS, R, WinSCP, Putty, pack Microsoft Office

6 Présentation des données

Les informations concernant les contrats et les clients sont stockées dans la base de pertes de l'entrepôt Risques de BPCE. En effet, bien que l'objectif soit de modéliser la probabilité de passage au contentieux pour les contrats sains, cette information n'est disponible que pour les contrats tombés en défaut. Ces contrats constitueront donc la base d'apprentissage ; ce qui reste cohérent avec notre objectif tant que seules les variables observables sur les contrats sains sont utilisées.

6.1 Alimentation de la base des pertes

L'organisation du groupe BPCE rend l'alimentation de la base des pertes complexe. Pour chaque contrat tombé en défaut, les informations concernant le client concerné, les produits qu'il détient, et ses garanties sont extraites et stockées dans une base des pertes locales, propre à chaque banque du groupe. Ensuite, des données historiques portant sur l'évolution de la situation financière du contrat l'année précédent le passage en défaut est également extraite et ajoutée à la base des pertes locale. Ces données passent ensuite par différents filtres automatiques controlant leur intégrité, et sont périodiquement remontées à la base des pertes nationale. Malgré ces filtres et plusieurs étapes de contrôle de la qualité des données, la multiplicité des bases locales et des systèmes d'information implique une multiplicité des sources d'erreurs potentielles et une fiabilité des données amoindrie. Nous travaillerons donc avec beaucoup de valeurs manquantes, et parfois des données peu fiables. En outre, certaines variables ne sont observables que sur une portion spécifique des contrats – typiquement les informations portant sur les entreprises lorsque le crédit permet au client de les créer ou de les développer –.

6.2 Variables disponibles

Notre variable d'intérêt est l'indicatrice du passage au contentieux. Au total, 26.38% des contrats en défaut passent au contentieux pour une LGD d'approximativement 100%.

Pour tenter d'identifier ces contrats, nous disposons de variables concernant :

— le client : les produits qu'il détient – nécessaire en raison du principe de contagion –, les évenements les affectant, des informations personnelles tels que son age, le nombre d'enfants à charge, son adresse, sa catégorie socio-professionnelle, la durée de la relation avec client ;

— le contrat : montant, date de signature, taux, nombre de passage en défaut, type de contrat, nombre de co-emrunteurs ;

— le portefeuille de garanties.

À partir de ces variables nous en construisons de nouvelles permettant de capturer de l'information additionnelle : client étant déjà passé au contentieux, durée depuis le dernier emménagement en tant qu'indicateur de stabilité du client, temps de survie du contrat.

Pour le modèle final, nous disposons donc de 52 variables qualitatives, de 420 variables binaires – pour le modèle utilisé, les variables qualitatives doivent être mise sous forme de tableau disjonctif complet –, et de 9 séries temporelles de 13 points retraçant l'évolution des états financiers du contrats l'année précédant son passage en défaut –. Nous avons donc au total 589 variables.

7 Démarche suivie

7.1 Forme du modèle souhaité : carte de Kohonen

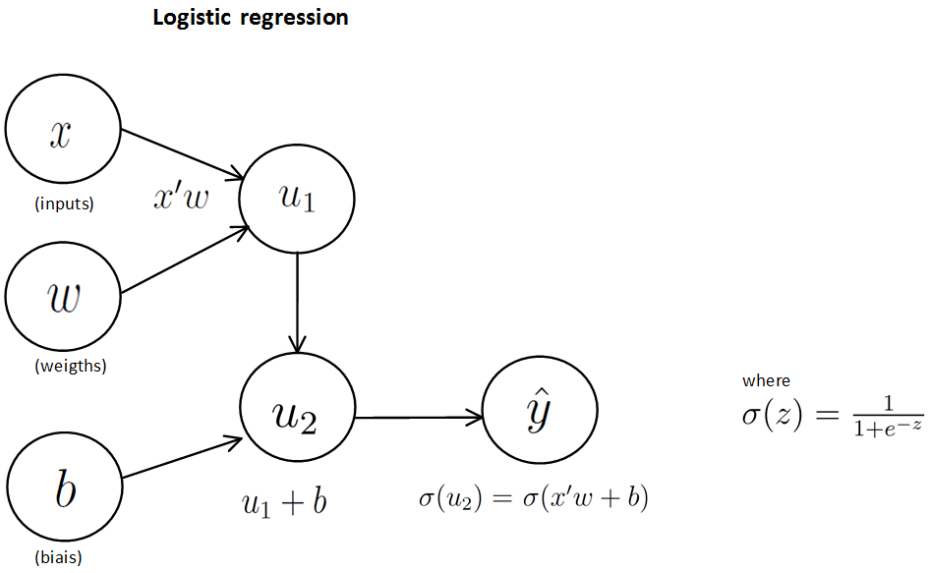
L'utilisation d'une carte de Kohonen pour modéliser la probabilité de passage au contentieux est une requête explicite de l'unité au sein de laquelle j'ai effectué mon stage. La motivation pour cette demande est de pouvoir simultanément cartographier les zones de contentieux, tout en ayant un pouvoir prédictif sur cette variable.

Contrairement aux méthodes factorielles telles que l'analyse en composantes principales ou l'analyse de correspondances multiples qui réalisent des projections sur des sous-espaces linéaires de l'espace initial des variables, les méthodes neuronales réalisent une cartographie non linéaire de l'espace. Bien que moins détaillées que la visualisation par plans factoriels successifs, ces méthodes s'efforcent de capturer les structures des données dans leur espace de départ. Ces méthodes sont de plus très robustes aux données manquantes (voir IBBOU [2]).

7.1.1 Principe de la carte de Kohonen

On appelle neurone tout composant d'un modèle qui réalise une tâche simple sur une entrée et qui renvoie une sortie. Tous les modèles peuvent être représentés sous forme d'un réseau de neurones, bien que la phase d'apprentissage – i.e. l'estimation des paramètres – puisse être différente. Par exemple, une régression logistique peut se mettre sous forme neuronale de la façon suivante :

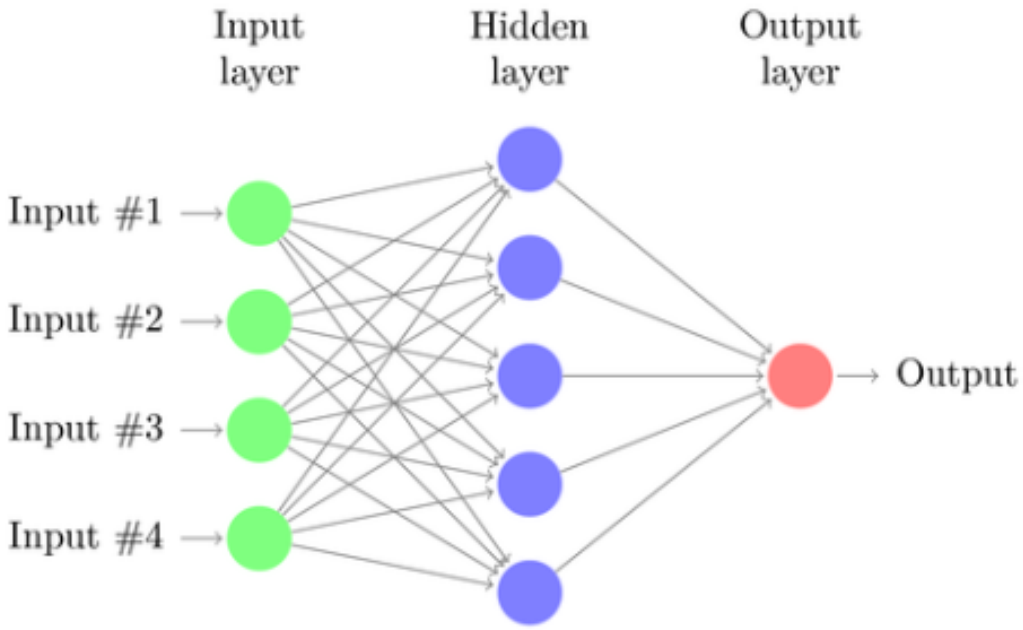
2. IBBOU, 1992, Classification, analyse des correspondances et méthodes neuronales, Thèse



Dans cet exemple, les unités u_1 et u_2 sont deux neurones cachés qui réalisent une multiplication et une addition, et qui en envoient le résultat au neurone de la couche de sortie $\hat{y}$ réalisant la transformation logistique.

Les cartes de Kohonen constituent une méthode statistique de classification, de représentation et de reconnaissance des structures des données s'appuyant sur l'algorithme de Kohonen. Contrairement à d'autres méthodes neuronales, il s'agit d'une méthode non supervisée. Elle est capable de s'organiser seule dans l'espace des données grâce à une forme compétitive d'apprentissage ; elle est formée d'une couche de neurones en compétition les uns avec les autres pour la représentation des observations qui lui sont présentées. Ces neurones sont caractérisés par un vecteur poids qui détermine leur position dans l'espace des variables, permettant de découper cet espace en une partition de Voronoi.

À la fin de la phase d'apprentissage, chaque neurone sera spécialisé dans la reconnaissance d'un certain type d'observation. Pour illustrer ce point, voici comment s'organise la carte :



Chaque neurone w_i regarde chaque élément du vecteur d'entrée (l'observation), et calcule la distance totale $d(w_i, x)$ entre ce neurone et cette observation. Cette distance est alors envoyée à un neurone supérieur qui recherche le neurone minimisant la distance avec l'observation présentée pour le désigner comme vainqueur. Cette information est alors renvoyée à travers le réseau pour mettre à jour ses vecteurs poids.

Il s'agit d'un outil puissant en raison de ses quatre caractéristiques principales ; il s'agit d'un algorithme non supervisé qui permet de :

— réaliser une classification des entrées en associant à chaque observation un neurone ;

— réaliser une quantification vectorielle grâce à l'observation des vecteurs codes des neurones ;

— capturer des structures non linéaires ;

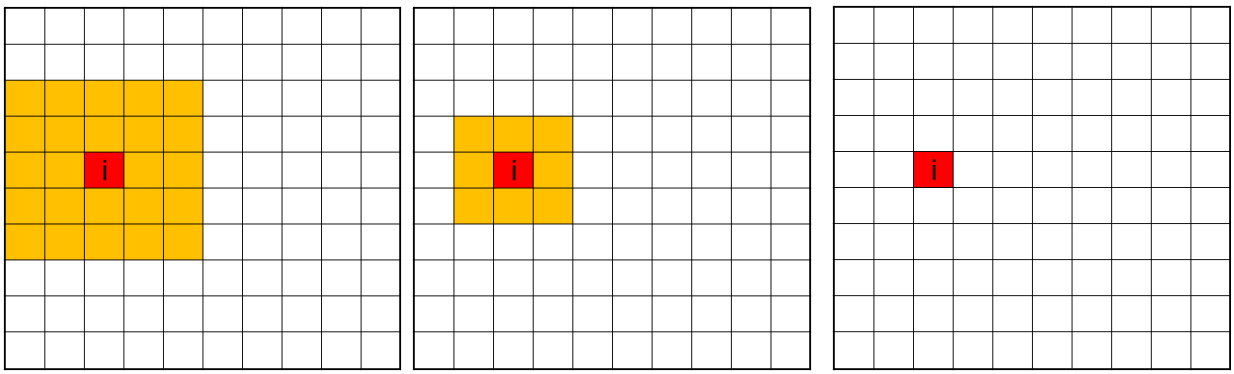
— tout en conservant la topologie de l'espace d'entrée grâce à la notion de voisinage.

Soit $\Xi \in \mathbb{R}^d$ l'espace des n observations d'apprentissage observées sur d variables. Le réseau $\mathcal{N}$ est formé de m unités organisées en grille de dimension $n_r * n_c$ ($m = n_r * n_c$) où n_r et n_c désignent

respectivement les nombres de lignes et de colonnes de la carte, et où chaque unité est caractérisée par un un vecteur de $\mathbb{R}^d$, son code,- $W_1(t), ..., W_m(t)$ pour tout instant t. Il s'agit en réalité de m variables aléatoires qui réalisent $w_1(t), ...w_m(t)$ en t. Ainsi, à t, le réseau est défini par un état W_t tel que $W_t = (W_1(t), ..., W_m(t))$.

Pour entrainer le réseau, une fonction de voisinnage $V_{r(t)}(i)$ est nécessaire afin de décrire les liens entre l'unité i et les autres unités du réseau situées dans un rayon $r(t)$ autour de i. Dans une grille, le rayon est défini par rapport à la norme infinie. En pratique, le rayon $r(t)$ est une fonction décroissante du temps, ce qui permet de tirer la carte de façon assez grossière dans un premier temps, puis d'affiner l'apprentissage plus subtilement. Si le rayon est strictement inférieur à 1, nous sommes dans le cas où les unités n'ont pas de voisin ; seul le vecteur code du neurone vainqueur est modifié. Ce cas est identique à l'algorithme K-means. La décroissance du rayon d'apprentissage implique que l'amplitude de la mise à jour des vecteurs codes doit également dépendre d'un paramètre d'apprentissage $\alpha(t)$ décroissant avec le temps.

L'illustration suivante permet de visualiser 3 exemples de voisinnage, de tailles respectives 25 ($r = 2$), 9 ($r = 1$) et 1 ($r = 0$) :



7.1.2 Algorithme de la carte de Kohonen

En pratique, l'algorithme peut se décrire en trois étapes consécutives répétées jusqu'à convergence de la carte.

La première étape consiste à présenter au réseau une observation choisie au hasard selon une distribution de probabilité $\mathbb{P} = (P_1, ..., P_n)$ sur les éléments de Ξ. En règle générale, la distribution uniforme $\forall j \in [1, n], P_j = \frac{1}{n}$ est préférée car on ne sait pas quelles sont les observations qui vont le plus contribuer à l'apprentissage. La seconde étape consiste ensuite en la détermination du neurone vainqueur, celui réalisant la plus faible distance vis-à-vis de l'observation. Cette distance est généralement la distance Euclidienne lorsque les données sont centrées réduites, mais peut être adaptée aux données comme nous le verront par la suite. La troisième et dernière étape est la mise à jour des vecteurs codes du vainqueur et des neurones avoisinnant pour prendre en compte l'information portée par l'observation présentée. Ces trois étapes sont répétées jusqu'à convergence.

L'algorithme :

— I. Initialisation des vecteurs codes

— II. À t, une observation $x(t+1) \in \Xi \subset \mathbb{R}^d$ est choisie aléatoirement et présentée au réseau

— III. On recherche l'unité gagnante i_0 définie par

$$i_0 = argmin_{1 \leq i \leq m} d(x(t+1) - w_i(t)) = argmin_{1 \leq i \leq m} \|x(t+1) - w_i(t)\|^2$$

pour la distance Euclidienne

— IV. Le vecteur code vainqueur w_{t0}^i et ses voisins sont mis à jour par

$$\begin{cases} w_i(t+1) = w_i(t) + \varepsilon(t)(x(t+1) - w_i(t)) & \forall i \in V_{r(t)}(i_0) \\ w_i(t+1) = w_i(t) & \forall i \notin V_{r(t)}(i_0) \end{cases} \quad (1)$$

où $V_{r(t)}(i_0)$ est le voisinnage de rayon $r(t)$ autour de l'unité i_0 et où $\varepsilon(t)$ satisfait les conditions de Robins-Monroe :

$$\sum_t \varepsilon(t) = \infty, \quad \sum_t \varepsilon(t)^2 < \infty$$

À la fin de l'apprentissage, chaque neurone (ou groupe de neurones) est spécialisé dans un type d'observations, de la même manière que chaque partie du cerveau est dédié à des tâches spécifiques. On peut alors associer chaque observation au neurone le plus proche pour réaliser une classification. Les vecteurs codes constituent une quantification vectorielle de l'espace d'entrée : si l'observation x_j est associée à l'unité i, alors

$$x_j \to^q q(x_j) = w_i(T)$$

où $w_i(T)$ est la valeur final du vecteur code de l'unité i. On peut alors définir les classes C_i par

$$\forall i \in [1, m], \quad C_i = q^{-1}(w_T^i) = \{x_j \in \Xi \mid argmin_{1 \leq i \leq m} d(x_j - w_i(T)) = i\}$$

Nous avons de cette façon au plus m classes (certaines peuvent être vides). La partition $\mathcal{C} = C_1, ..., C_m$ est une partition de Voronoi de l'espace Ξ : cet hyperplan est partitioné en régions définies par leur distance aux vecteurs codes du réseau. La structure de voisinnage des neurones peut être utilisée pour définir un voisinnage des classes et pour créer des super-classes. La carte ainsi obtenue respecte la topologie de l'espace d'entrée Ξ : des observations proches dans cet espace seront assignés au même noeud ou a des noeuds voisins sur la carte.

7.2 Initialisation de la carte de Kohonen

Si nous avons décrit l'algorithme de Kohonen dans la section précédente, nous n'avons pas évoqué l'initilisation de la carte. Bien que la convergence de cette algorithme ait été démontré par

Ritter et al.[3] en 1992, l'initialisation conditionne la facilité avec laquelle la convergence doit se faire. Il s'agit donc d'une étape permettant d'optimiser le chemin de convergence et de l'influencer pour l'accorder au problème considéré.

Aujourd'hui deux méthodes d'initialisation existent. La première consiste à initialiser de façon aléatoire la carte avec des vecteurs codes dans l'espace des entrées ou dans les valeurs possibles des variables utilisées. La seconde consiste à réaliser une analyse factorielle, en général une analyse en composantes principales, et à initialiser la carte sur le plan principal mis en évidence. Si la première méthode est très intéressante d'un point de vue théorique puisqu'elle permet de démontrer la convergence de l'algorithme, la seconde en augmente considérablement la vitesse de convergence. Néanmoins, puisque nous avons un objectif de prédiction du passage au contentieux, nous devons tenter d'inclure cet objectif de supervision dès l'initialisation de la carte. Pour ce faire, nous développons une méthode inédite d'initialisation supervisée.

Nous cherchons à initialiser la carte le long des axes les plus discriminant en termes de contentieux afin d'obtenir dans la carte finale des zones de contentieux clairement identifiées. Pour cela, nous reprenons l'idée d'ACP supervisée développée par BAIR, HASTIE, PAUL et TIBSHIRANI[4] en réalisant une analyse factorielle sur le sous-espace engendré pas les variables les plus corrélées avec le passage au contentieux. On définit le rapport de corrélation entre une variable quantitative

3. voir Ritter, H., Martinetz, T. Shulten, K., (1992) : Neural computation and Self-Organizing Maps, an Introduction, Addison-Wesley, Reading.

4. voir BAIR, HASTIE, PAUL et TIBSHIRANI, (2006) : Prediction by supervised principal components, Journal of the American Statistical Association, Vol.101, No. 473

X et une variable qualitative Y – la variable de contentieux étant binaire – par

$$\eta_{X|Y} = \frac{\mathbb{V}[\mathbb{E}[X|Y]]}{\mathbb{V}[X]}$$

où $\mathbb{V}[.]$ désigne la variance, et entre deux variables binaires – les variables quantitatives étant sous forme de tableau disjonctif complet – par le phi de Pearson ϕ^2 défini par

$$\phi^2 = \sqrt{\frac{\chi^2}{n}}$$

Nous avons ainsi pour chaque variable une mesure de la force de son association avec la variable du contentieux. En sélectionnant les variables les plus pertinentes, on se positionne dans un sous-espace sur lequel réaliser une analyse factorielle sera discriminante vis-à-vis du contentieux. Toutefois, contrairement à l'article de BAIR et al., nous disposons de données mixtes – quantitatives et qualitatives –. Pour les traiter, nous utilisons la méthode décrite par PAGES [5] en 2004. Pour réaliser cette analyse factorielle sur données mixtes, les variables quantitatives sont centrées et réduites, et les variables qualitatives sous mises sous la forme d'un tableau disjonctif complet corrigé. Supposons que nous observons Q variables qualitatives à m_q modalités – pour un total de M modalités – sur n individus, alors le tableau disjonctif complet est la matrice $K \in \mathcal{M}_{N*M}$ où $\forall i, \forall j, k_{ij_q} = 1$ si l'individu i prend la modalité q de la variable j et 0 sinon. On suppose que chaque individu ne peut prendre qu'une modalité par variable. On corrige ensuite la matrice en divisant les $k_{ij_q} = 1$ par $\sqrt{f_{.j_q}}$ où $f_{.j_q}$ est la fréquence d'apparition de la modalité j_q ; on nomme la nouvelle matrice K^c.

En réalisant une analyse en composantes principales simultanément sur les variables quantitatives centrées réduites et sur le tableau disjonctif complet, et en ajoutant les variables qualitatives

5. voir PAGES, J., (2004) : Analyse factorielle de données mixtes, Revue de statistique appliquée, Vol.52, No. 4, pp.93-111

"brutes" en tant que variables supplémentaires on réalise l'analyse factorielle sur données mixtes. Cependant, nos variables comportent des données manquantes, ce qui est prohibitif pour une analyse en composantes principales classique. Une imputation des données manquantes par forêts aléatoires ou analyses factorielles n'est pas possible du fait que certaines variables ne sont pertinentes – et observables – que sur certains types de contrats ou de clients. Par ailleurs, nous ne sommes intéressés que par le plan principal de l'espace des entrées ; le calcul de l'intégralité des axes principaux peut être couteux en temps au regard de la dimension de notre matrice.

L'algorithme NIPALS (Nonlinear Iterative Partial Least Squares) permet d'approximer les axes principaux d'une analyse en composantes principales en évitant de calculer la matrice de covariance des entrées. On évite donc sa dégénération liée aux variables manquantes. De plus, cet algorithme permet de calculer les axes principaux les uns après les autres, par ordre décroissant de contribution à la variance. Appelons X la matrice comportant à la fois les variables qualitatives centrées réduites le tableau disjonctif complet corrigé K^c, que l'on centre également. L'algorithme NIPALS s'appuie sur les constats suivants. Si les composantes principales existent, alors on peut décomposer X (de rang r) en

$$X = \sum_{h=1}^{r} t_h p_h'$$

où les $t_h = (t_{h1}, ..., t_{hn})$ sont les composantes principales et où les $p_h = (p_{h1}, ..., p_{hn})$ sont les vecteurs directeurs de ces axes. Nous pouvons donc exprimer les variables x_j et les individus x_i par

$$x_j = \sum_{h=1}^{r} p_{hj} t_h \quad x_i = \sum_{h=1}^{r} t_{hi} p_h$$

On peut ainsi interpréter les p_{hj} comme les coefficients de la régression de x_j sur t_h et les t_{hi} comme ceux de la régression de x_i sur p_h. De plus, l'équation de la décomposition de X peut se

comprendre comme un modèle où les $t_h = (t_{h1}, ..., t_{hn})$ et les $p_h = (p_{h1}, ..., p_{hn})$ sont à estimer, en utilisant itérativement sur h les régressions de x_j sur t_h et de x_i sur p_h. En présence de données manquantes, il suffit de n'utiliser que les informations existantes à chaque étape de l'algorithme :

L'algorithme :

- I. $X_0 = X$
- II. Pour h allant de 1 au nombre d'axes choisis (r au maximum) :
  - t_h=première colonne de X_{h-1}
  - Jusqu'à convergence de p_h répéter
    - $\forall j, p_{hj} = \frac{\sum_{i:\exists(x_{ji},t_{hi})} x_{h-1,ji} t_{hi}}{\sum_{i:\exists(x_{ji},t_{hi})} t_{hi}^2}$
    - Normer p_h à 1
    - $\forall i, t_{hi} = \frac{\sum_{i:\exists x_{ji}} x_{h-1,ji} p_{hj}}{\sum_{i:\exists x_{ji}} p_{hj}^2}$
- III. $X_h = X_{h-1} - t_h p_h'$

L'étape II. permet donc de calculer les estimateurs des moindres carrés (sans constante) sur les données disponibles. À la fin de l'algorithme, on peut, pour tout h, calculer la pseudo-valeur propre associée au h-ième axe principal par le calcul de la variance $\frac{1}{n-1} t_h' t_h$ de la composante t_h.

Une fois le plan principal estimé, on estime les coordonnées des modalités variables qualitatives ayant été utilisées dans l'analyse factorielle sur ce plan, et pour chacune d'entre elles on réalise une partition de Voronoi du plan. Les vecteurs codes sont ainsi initialisés en prenant compte des modalités corrélées avec la variable d'intérêt. Les éléments des vecteurs codes correspondant à des variables quantitatives ou à des modalités non utilisées dans l'analyse factorielle sont initialisés à 0. Notre carte initiale forme ainsi bien un plan orienté en fonction des zones de contentieux.

7.3 Supervision par fusion

Afin de continuer à prendre en compte l'information sur le passage au contentieux que nous possédons effectivement, nous adaptons le principe de la carte de Kohonen pour obtenir un algorithme supervisé. Nous choisissons pour cela la méthode de fusion (X-Y fused network) développée en 2005 [6]. Cette méthode consister à créer une seconde carte de Kohonen de mêmes dimensions que la première, mais dont les vecteurs codes ne caractérisent que la variable d'intérêt (ici, le taux de passage au contentieux). La recherche du vainqueur ne se fait alors plus en minimisant la distance entre l'observation d'entrée et les vecteurs codes de la première carte de Kohonen, mais en minimisant une combinaison linéaire de cette distance et de celle par rapport à la seconde carte. Soit $d(i,k)$ la distance entre l'observation i et le noeud k, on a alors $d(i,k) = \beta(t)d(i,k_X)+(1-\beta(t))d(i,k_Y)$ où X représente la carte des variables explicatives, et Y celle de la variable d'intérêt. Le paramètre $\beta(.)$ dépend du temps et décroit avec le temps. Les distances $d(i,k_X)$ et $d(i,k_Y)$ doivent être normalisée pour être comparables – pour cela, on les divise par $max_k d(i,k_X)$ et $max_k d(i,k_Y)$ respectivement –.

6. voir MELSSEN, W., WEHRENS, R., BUDENS, L., (2006) : Supervised Kohonen networks for classification problems, Chemometrics and Intelligent Laboratory Systems, Vol.83, pp.99-113

8 Application

8.1 Traitement des variables

8.1.1 Variables quantitatives

Pour la représentation de variables quantitatives avec cet algorithme, nous devons dans un premier temps les normaliser afin de s'affanchir de leurs différences d'échelle. La méthode KACP développée par S. IBBOU en 1992 est la version de Kohonen de l'analyse en composantes principales (ACP). En un sens, une carte de Kohonen est une projection des données sur la carte, dans le même esprit d'une ACP où l'on effectue des projections des données successivement sur les plans principaux. Toutefois, avec la méthode KACP, la projection se fait sur une classe et non sur un plan. Cela permet de résumer les données en une représentation à 2 dimensions, et ne nécessite donc pas l'utilisation simultanée de plusieurs plans comme en ACP. Par ailleurs, la classification ainsi opérée peut être directement exploitée, notamment par une méthode de classification hiérarchique afin de définir des super-classes. De plus, la carte fournit également le vecteur code du parangon de chaque classe. Un exemple simple d'utilisation de la méthode KACP est détaillé en annexe A.

8.1.2 Variables qualitatives

Pour les variables qualitatives, les cartes de Kohonen peuvent également être utilisées à l'image d'une analyse des correspondances multiples (ACM) grâce à la méthode KACM. Cette méthode permet de cartographier simultanément les individus et les modalités des variables considérées. La carte KACM permet donc d'apprécier la proximité entre les individus, entre les modalités et entre individus et modalités.

Supposons que l'on observe Q variables qualitatives avec m_q modalités pour un total de M

modalités observées sur n individus. Pour le traitement de ces variables, on utilise la table de Burt B et le tableau disjonctif complet K. Comme défini dans la section précédente, on a $K \in \mathcal{M}_{N*M}$ où $\forall i, \forall j, k_{ij_q} = 1$ si l'individu i prend la modalité q de la variable j et 0 sinon. On suppose que chaque individu ne peut prendre qu'une modalité par variable. On a alors, $\forall i \in [1, N], k_{i.} = \sum_{j=1}^{M} k_{ij} = Q$. La table de Burt $B \in \mathcal{M}_{M*M}$ est une matrice carrée symétrique définie par $B = K^t K$.

Dans une ACM classique, les axes maximisant l'inertie de la projection des individus sur ceux-ci sont portés par les valeurs propres de la matrice

$$\frac{1}{Q} K^t K W^{-1} = \frac{1}{Q} B W^{-1}$$

où $W = diag(k_{.1}, ..., k_{.M})$. L'idée est donc de corriger la matrice de Burt et le tableau disjonctif complet afin de pouvoir utiliser des poids uniformes et la distance Euclidienne. Bien que l'algorithme fonctionne pour toutes entrées et distance, le choix des hyperparamètres pour ce type de réseaux de neurones est bien plus étudié et compris pour le cas des poids uniforme et de la distance Euclidienne.

On corrige ensuite la matrice en divisant les $k_{ij_q} = 1$ par $\sqrt{f_{.j_q}}$ où $f_{.j_q}$ est la fréquence d'apparition de la modalité j_q ; on nomme la nouvelle matrice K^c. Pour cartographier les modalités, on utilise la matrice de Burt corrigée B^c dont les lignes formeront des entrées soumise à la carte de Kohonen. On définit $B^c = (b^c_{ij_q})_{ij_q}$ où

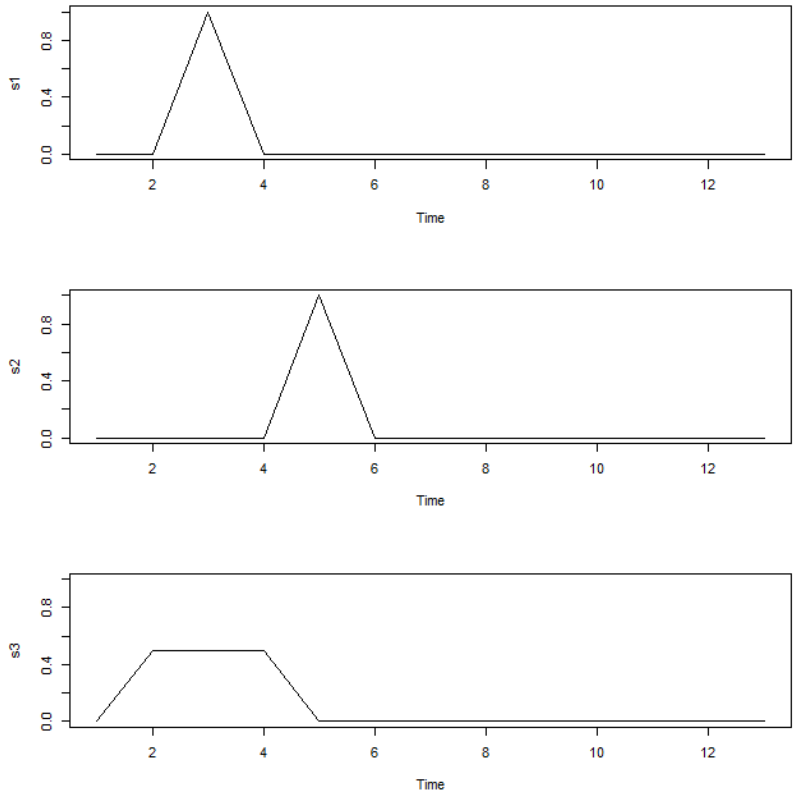
$$b^c_{ij_q} = \frac{b_{ij_q}}{\sqrt{b_{i.}}\sqrt{b_{.j_q}}}$$

Le tableau disjonctif complet corrigé, quant à lui, sera présenté au réseau et permettra d'y placer les individus.

Un exemple simple d'utilisation de la méthode KACM est détaillé en annexe B.

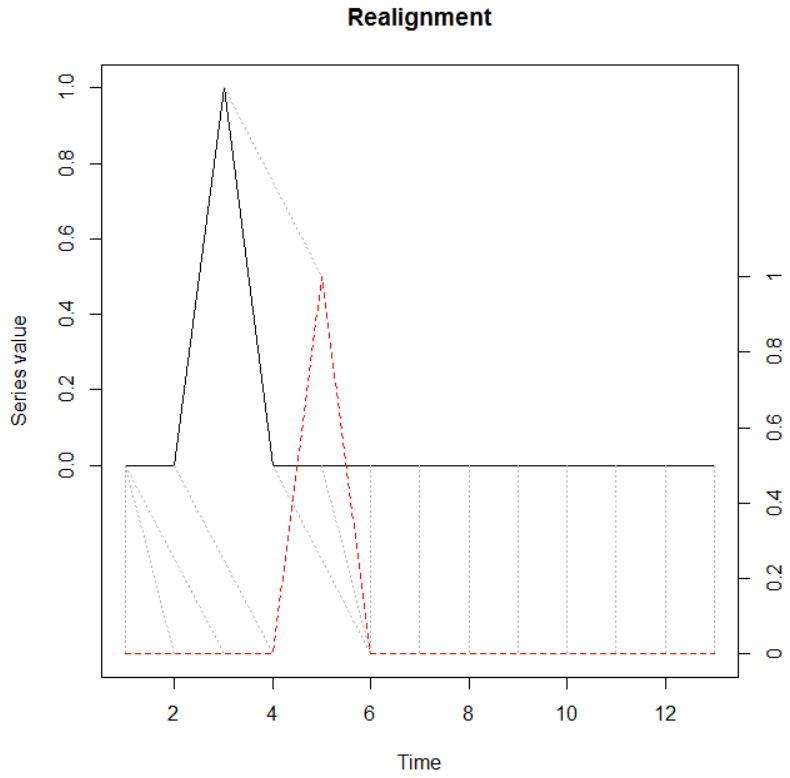
8.1.3 Séries temporelles

Comme nous avons pu l'évoquer précédemment, parmi les informations dont nous disposons figurent des séries temporelles retraçant les états financiers des contrats pendant l'année précédent leur entrée en défaut, au ryhtme d'une image par mois. À présent, cette information n'est pas exploitée, hormis une seule photo spécifique dans le cadre des prévisions de pertes. Cependant, au-delà des montants d'expositions ou d'engagements, l'évolution de ces montants dans le temps recèle beaucoup d'information. En effet, certains profils peuvent être révélateurs d'une mauvaise gestion ou d'une dégradation permanente de la contrepartie là où d'autres pourraient s'interpréter comme un événement ponctuel dont la contrepartie pourra se remettre. En considérant ces variables comme de simples variables quantitatives, et utilisant la distance Euclidienne classique, notre carte pourrait ne pas réussir à capter des comportements distincts. En guise d'exemple, considérons les trois séries temporelles S1, S2 et S3 :



Visuellement, les séries S1 et S2 semblent similaires modulo un décalage temporel. En revanche,

une utilisation naive de la distance Euclidienne conduit à une plus grande proximité entre les séries S1 et S3 ($d = 0.86$) qu'entre les séries S1 et S2 ($d = 1$). Afin de pouvoir comparer la proximité entre les formes de ces séries, nous devons réaligner leurs axes temporels, ce que nous faisons à l'aide d'une méthode appelée "Dynamic Time Warping". En utilisant le package "dtw" de R, on réaligne les séries temporelles :



Nous pouvons voir en noir la série S1, en rouge la série S2, et en pointillés la correspondance entre les axes temporels des deux séries. Ici, les pics sont alignés, ce qui conduit à une distance dtw de 0, tandis que la distance dtw entre les séries S1 et S3 est de 1.4.

D'un point de vue plus rigoureux, considérons les deux séries R et S de tailles respectives m et n. Afin de les aligner, nous devons construire une matrice C de taille $n * m$ où $C_{ij} = d(r_i, s_j)$, $d(r_i, s_j)$ étant la distance Euclidienne entre r_i et s_j. Cette matrice représente alors le coût d'alignement entre ces deux points : si l'on décide de faire correspondre l'instant i de la série R avec l'instant

j de la série S, le coût de ce réajustement est $C_{ij} = d(r_i, s_j)$. Un chemin de déformation W est un chemin contigu (nous définissons cette notion ci-dessous) au sein de la matrice de coût C permettant de lier les éléments de chacun des deux axes temporels. On a alors $W = w_1, ...w_k, ..., w_K$ où $w_k = (i, j)_k$ avec $max(m, n) \leq K < m + n - 1$. On voit ainsi que chaque instant de chaque série doit être lié à un instant de l'autre série, et que le chemin de déformation ne peut pas être plus long que la somme des longueurs des deux séries temporelles.

Un tel chemin doit respecter plusieurs contraintes :

— Limites :

En général, on souhaite faire coincider les deux premiers instants des séries, et leurs deux derniers : $w_1 = (1, 1)$ et $w_K = (m, n)$

— Continuité :

Si $w_k = (i, j)$ alors $w_{k+1} = (i', j')$ doit être de telle sorte que $i' - i \leq 1$ et $j' - j \leq 1$ afin d'avoir un chemin contigu

— Monotonicité :

Si $w_k = (i, j)$ alors $w_{k+1} = (i', j')$ doit être de telle sorte que $i' - i \geq 0$ et $j' - j \geq 0$, afin de continuer d'avancer dans le sens des axes temporels.

On cherche le chemin qui minimise le coût global de réalignement, c'est-à-dire la somme des éléments de la matrice C appartenant au chemin :

$$DTW_{R,S} = min\left(\sqrt{\sum_{k=1}^{K} w_k}\right)$$

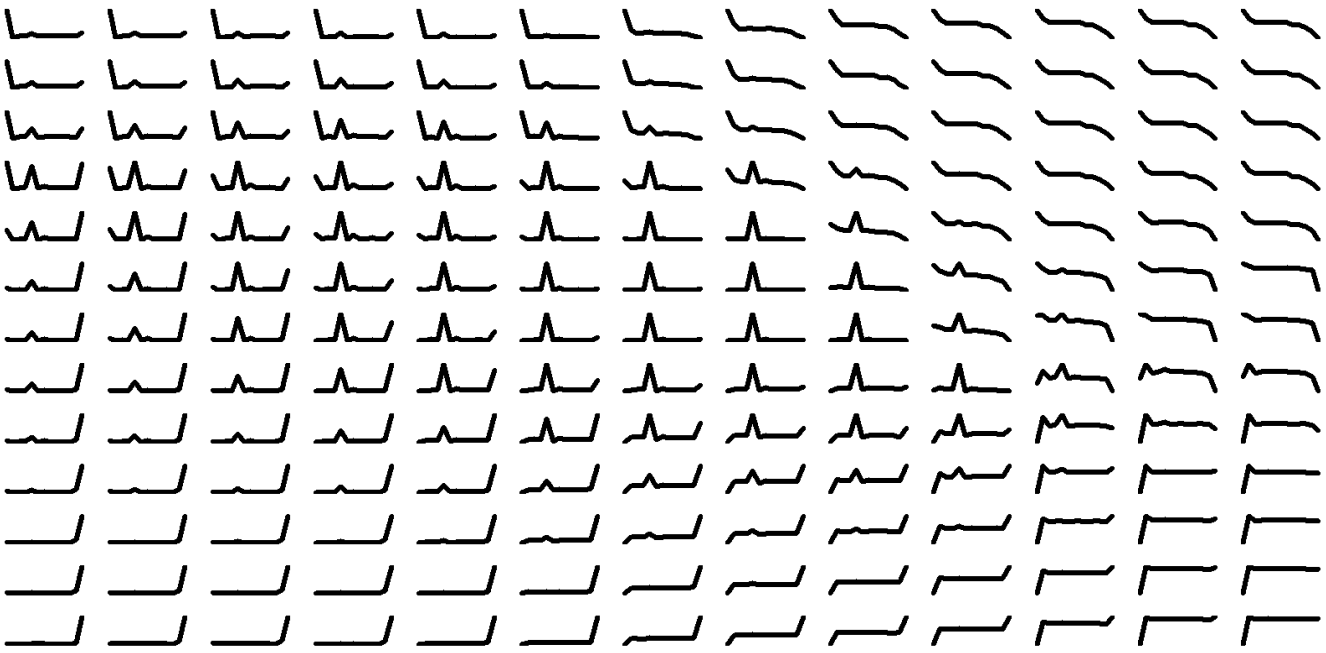
Cette optimisation se fait par la recherche pas à pas du minimum de cette distance cumulée $\gamma(i, j)$:

$$\gamma(i, j) = c_{ij} + min(\gamma(i - 1, j - 1), \gamma(i - 1, j), \gamma(i, j - 1))$$

À noter que l'optimisation démarre à la fin du chemin $w_K = (m, n)$, et calcule les distances cumulées à partir de $w_1 = (1, 1)$: pour chaque élément contigu, l'algorithme calcule sa valeur minimum par

rapport à tous les chemins possibles pour l'atteindre en commençant en $w_1 = (1, 1)$. Afin d'alléger le coût de calcul induit par cet algorithme, plusieurs types de contraintes existent. Cependant, en raison du faible nombre de points dont nous disposons, ajouter des contraintes risque d'empêcher totalement la convergence et de rendre ainsi ces variables inutilisables. De plus amples explications sur ces contraintes sont données en annexe C.

Pour illustrer la façon dont s'organise ces dynamiques dans une carte de Kohonen, on en créé une sur une série temporelle, en utilisant la distance DTW. Pour cela, on centre et on réduit chaque série individuellement, puis on les compare selon la distance DTW à des vecteurs codes de 13 points initialisés aléatoirement. On obtient ainsi :



La carte distingue bien des dynamiques différentes et s'organise en fonction. Elle prend également en compte l'échelle des séries temporelles, ce qui ne se voit pas sur cette représentation.

8.2 Choix des hyperparamètres

On appelle hyperparamètres les paramètres qui conditionnent l'apprentissage de la carte mais qui ne jouent plus de rôle losque que celle-ci est entrainée. Le terme d'hyperparamètre fait ici référence en particulier aux dimensions du réseau de neurones, au taux d'apprentissage, au rayon d'apprentissage, au paramètre de fusion et au nombre de répétition des données. Le choix de ces paramètres doit normalement être effectué par validation croisée, en surveillant la performance du modèle sur un échantillon de validation indépendant de l'échantillon d'apprentissage tout en prenant en compte le coût de calcul.

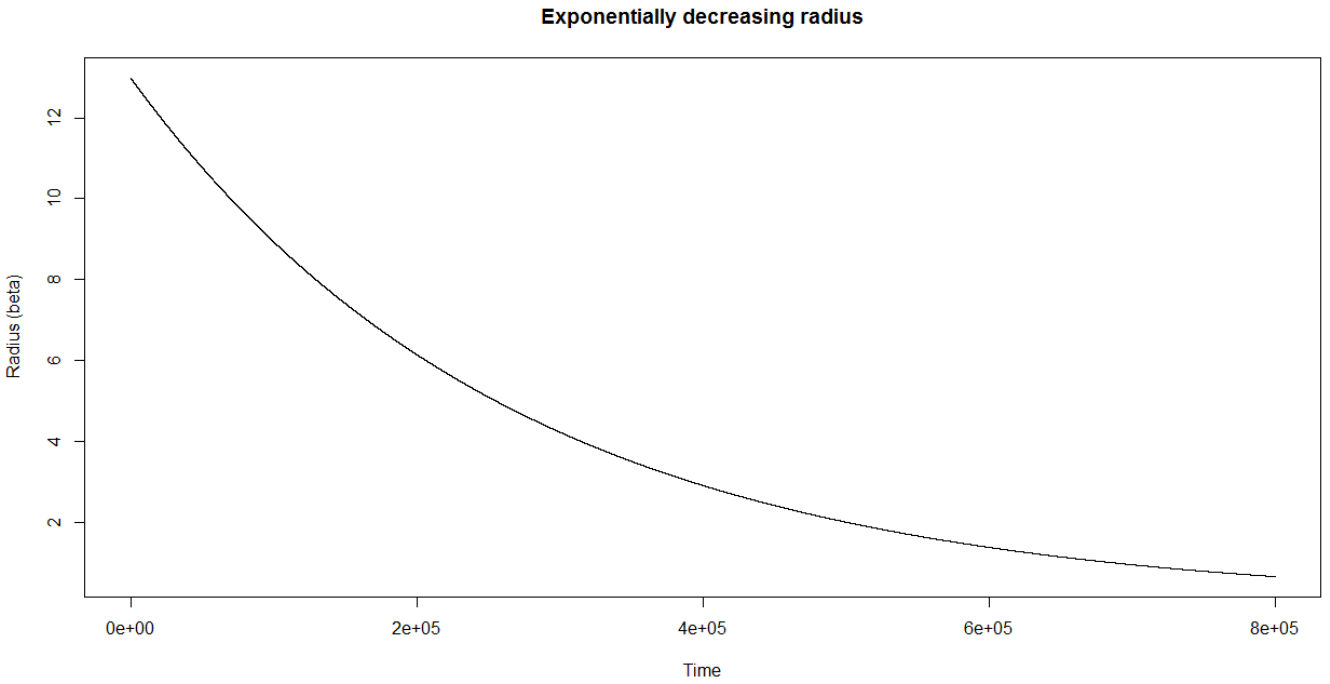
Le groupe BPCE devait initialement être équipée d'un serveur R dans les mois suivant le début du stage, mais différents problèmes techniques ont grandement retardé sa mise en place, si bien que le serveur n'est aujourd'hui pas installé. Ceci nous contraint à n'utiliser qu'un petit nombre d'observations en vue de démontrer l'intérêt de la méthode, sans toutefois être en mesure de choisir les paramètres optimaux par validation croisée, ni d'exploiter l'intégralité de l'information disponible.

Pour cette première carte, nous utilisons une carte de dimensions 10x10 entrainée sur un échantillon de 40 000 contrats répété 2 fois. On définit la décroissance exponentielle du taux d'apprentissage $\alpha(.)$ – la force avec laquelle l'information est intégrée à la carte – et du rayon d'apprentissage $r(.)$ – le voisinnage autour du neurone vainqueur dans lequel les poids sont modifiés – par les fonctions

$$\alpha(t) = \alpha_0(\frac{\alpha_T}{\alpha_0})^{\frac{T}{t}} \quad r(t) = r_0(\frac{r_T}{r_0})^{\frac{T}{t}}.$$

avec $t \in [0, T]$ On choisit un taux d'apprentissage relativement faible afin de ne pas avoir de

problème de sur-apprentissage : de 0.07 à 0.01. Quant on rayon d'apprentissage, on choisit de le faire varier de 4.99 à 0.65. De cette façon, les permières observations tirent des pans entiers de la carte, ce qui permet de mieux couvrir l'espace des entrée. Un rayon final de 0.65 permet de finaliser l'apprentissage sur environ 15% des données qui ne modifieront que le vecteur code du neurone vainqueur. Ainsi, la carte apprend rapidement mais grossièrement au début puis apprend très subtilement à la fin. Le faible taux d'apprentissage compense alors le risque de sur-apprentissage.



La pondération entre la carte des variables et la carte spécifique au taux de contentieux qui régissent la fusion suit également une fonction de décroissance exponentielle ; le poids de la carte des variables explicatives est de 1 pour la première observation, et décroit exponentiellement jusqu'à 0.5.

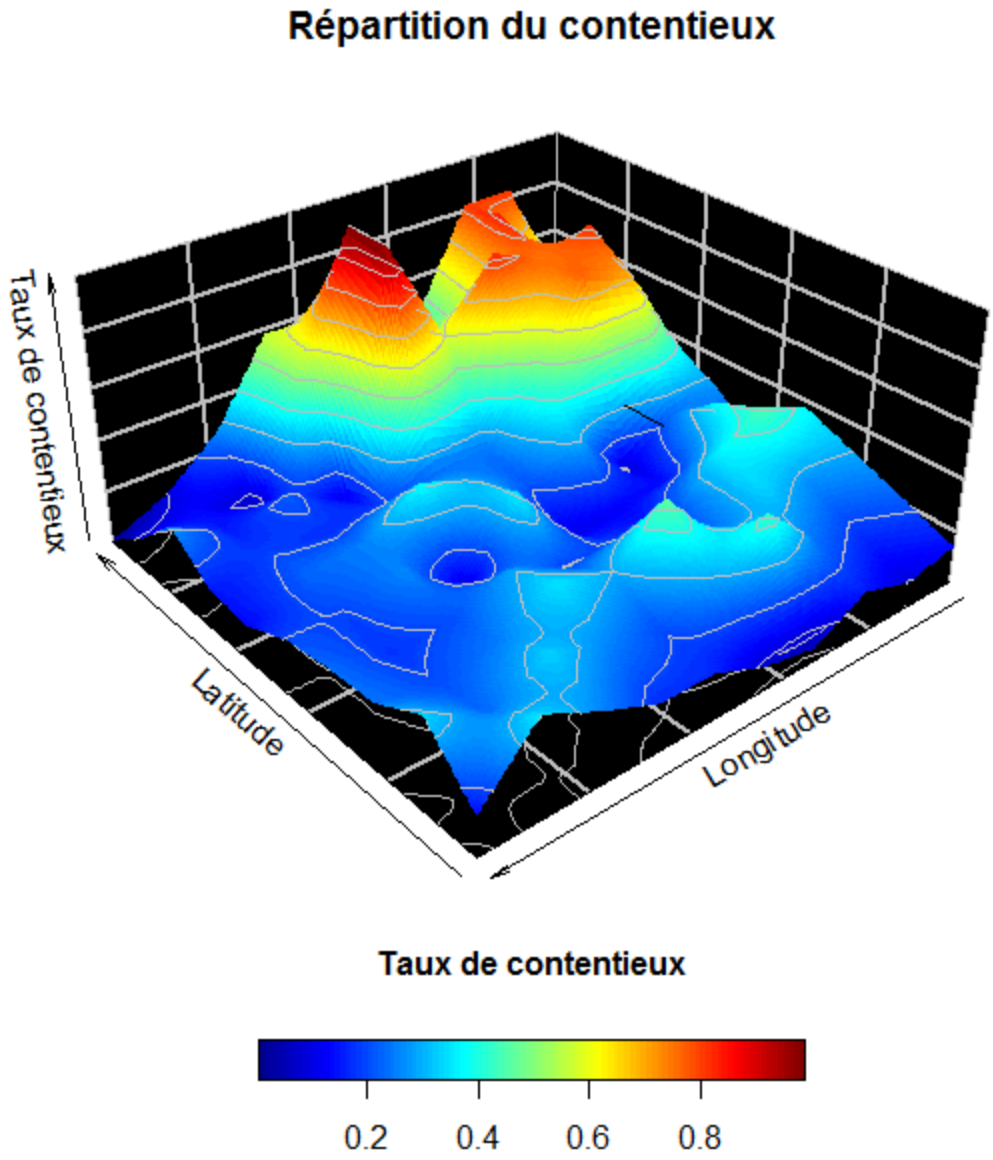
Le choix de la taille de la carte est également un hyperparamètre à déterminer. Il s'agit d'avoir une carte suffisament large pour pouvoir identifier des comportements spécifiques, mais compor-

tant suffisament d'observations dans les noeuds pour être significative. Le choix d'une carte 10x10 paraît pour le moment être un bon compromis. Il est également possible de détailler plus amplement la carte dans certaines régions d'intérêt, ou de réaliser des regroupements en super-classes dans d'autres régions.

Concernant le nombre d'itération du jeu de données, il est courant de répéter plusieurs fois voire plusieurs centaines de fois les entrées pour avoir une carte très robuste. Cependant, pour les crédits aux particuliers de nombreux contrats sont très similaires et peuvent être considérés comme une réplication de la même observation, si bien qu'il peut ne pas être nécessaire d'utiliser l'intégralité du jeu de données pour observer une convergence de la carte. Nous choisissons toutefois d'utiliser 40000 observations répétées 2 fois, car c'est là le maximum que l'ordinateur dont nous disposons peut traiter.

8.3 Résultats

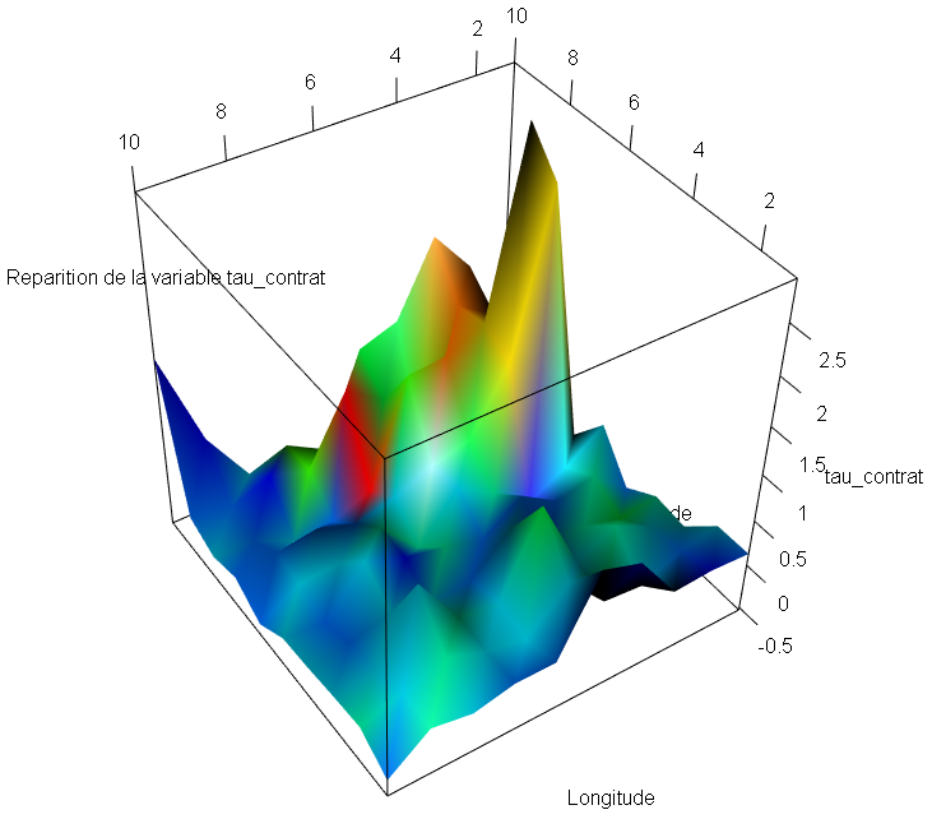
En utilisant les algorithmes et hyper-paramètres définis dans les sections précédentes, on obtient la carte suivante :



Il s'agit d'une carte 10x10 sur laquelle le taux de passage en contentieux est mis en relief. Comme nous pouvons le voir, alors qu'aujourd'hui on ne pratique pas ou peu de discrimination vis-à-vis du passage des contrats au contentieux au sein de BPCE, nous avons sur cette carte des zones avec plus de 85% de passage au contentieux (en orange et rouge sur la carte), et d'autres avec exactement 0% (en bleu foncé). Malgré la faible proportion d'observations utilisées pour l'apprentissage, nous parvenons déjà à avoir un pouvoir prédictif sur certaines parties de la carte.

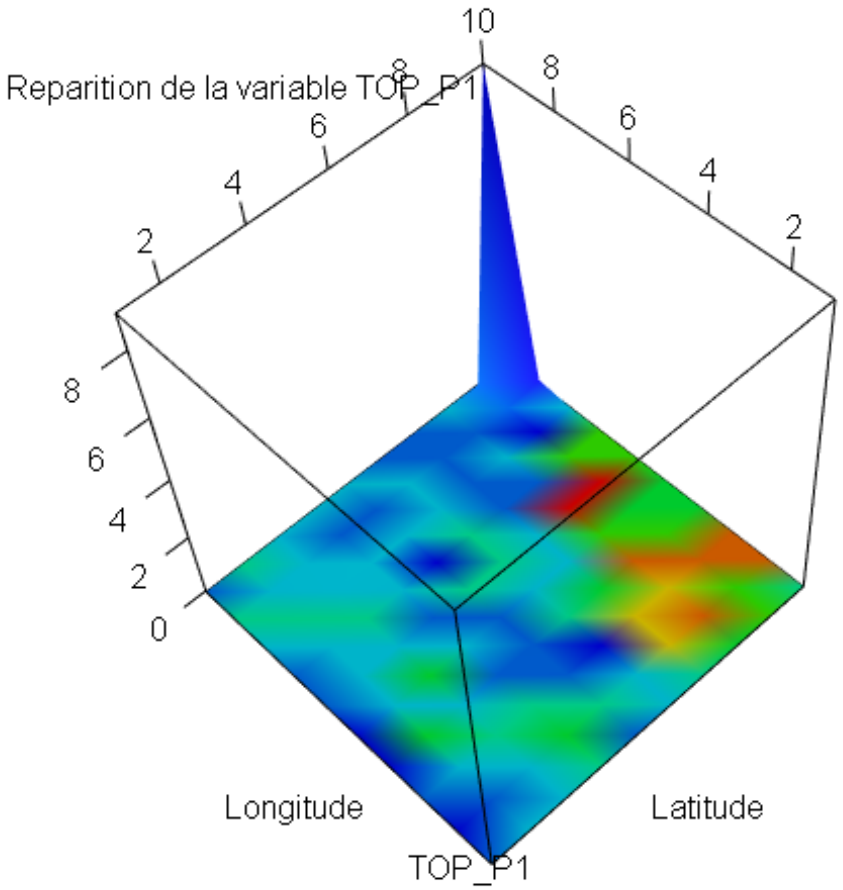
Interpréter la carte est un travail de longue haleine puisqu'il nécessite de regarder pour chaque

neurone sa position dans l'espace de nos 589 variables. En effet, puisque les neurones sont les parangons de chaque classe, comprendre leur vecteur code permet de caractériser les contrats affectés à cette classe, et donc de saisir l'impact de cette combinaison de variables spécifique. Une autre méthode consiste à ne pas s'intéresser aux noeuds mais d'observer la distribution des variables sur la carte. Désormais, on cartographie des variables explicatives en côte en conservant les couleurs de la carte ci-dessus ; pour chaque noeud de la carte, on calcul la moyenne des valeurs prises par cette variable. Nous pourrons ainsi aisément comparer la distribution des variables considérées avec les zones de contentieux ou de non-contentieux. En exemple, voici le taux initial du contrat :



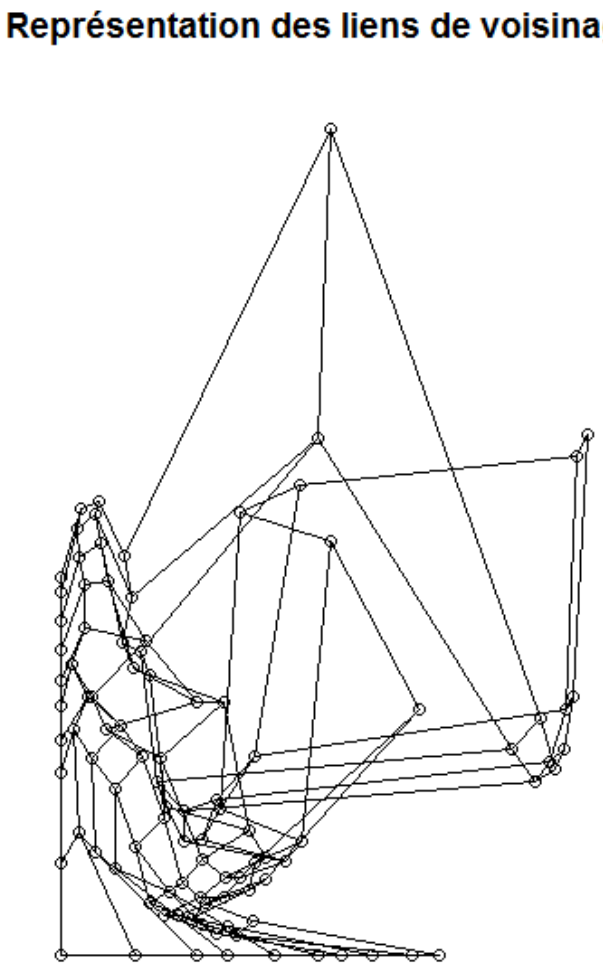
Cette variable du taux initial du contrat prend des valeurs élévées dans les zones chaudes (jaune et orange) de la carte, et des valeurs faibles dans les autres. Nous pouvons donc conclure qu'elle est – comme nous pouvions nous y attendre – explicative du risque de passage au contentieux. Toutefois, elle prend également des valeurs faibles dans les quelques zones rouges de la carte (taux

de passage au contentieux maximum) et une valeur élevée en (10,10), où le taux de contentieux est de 0. Différentes interprétations sont possibles et devront être inspectées. Le point aux coordonnées (10,10) peut correspondre à un produit spécifique dont le taux est plus élevé, mais peut également correspondre à une classe de contrats que l'on identifie aujourd'hui comme risqués sans raison effective. Les faibles valeurs de taux de crédit sur les zones rouges peuvent également se voir comme un échec de la détection du risque à l'octroi de crédit. L'hypothèse de produits spécifiques se confirme lorsque l'on regarde la distribution de classes de produits :

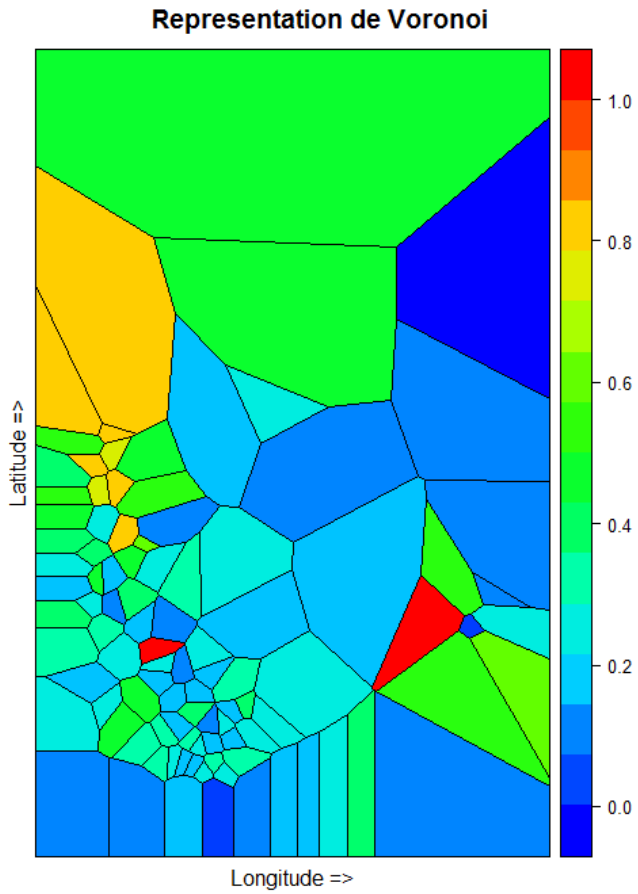


La répartition des produits de type P1 sur la carte illustre bien d'une part la capacité de spécialisation des neurones de la carte de Kohonen, et d'autre part le fait que c'est ce produit qui détermine le comportement des contrats de cette classe par rapport au contentieux.

Nous pouvons également montrer la façon dont la carte déforme l'espace en regardant les distances entre les neurones voisins. Pour cela, on considère un plan dans lequel la position de chaque neurone dépend de ses neurones voisins en latitude et longitude.



On voit bien la déformation de la carte qui ne ressemble plus du tout à une grille. On voit également qu'en augmentant les dimensions de la carte, cette représentation risque d'être inexploitable. Nous pouvons donc introduire une nouvelle représentation du réseau, sous forme d'une partition de Voronoi :



Parce qu'il s'agit d'une représentation en 2 dimensions d'un réseau en 472 dimensions (on exclut les variables de type séries temporelles), les voisins sur la carte ne sont pas nécessairement des voisins sur la grille. En revanche, cette représentation permet de bien appréhender la largeur des classes en termes d'espace occupé. On note pour cela la multitude de petites classes vers le centre gauche de la carte, avec parfois la découverte de niche de contentieux, qui se distingue des très larges classes partant vers la partie supérieure droite de la carte.

9 Conclusion et développements futurs

En raison de l'indisponibilité du serveur R sur lequel je devais initiallement pouvoir travailler, les résultats obtenus dans ce rapport ne prennent en compte qu'un nombre très limités d'observations. Néanmoins, nous avons pu voir que la discrimination entre contrats contentieux et non contentieux s'opère. Avec le temps de stage restant il sera possible de prendre en compte une plus grande quantité d'observations afin de gagner en robustesse. Cette indisponibilité du serveur a été la source majeure des difficultés rencontrées lors de ce travail, et la question de l'arbitrage entre performance et temps de calcul a été structurante.

En outre, une meilleure puissance de calcul ou une optimisation des programmes permettrait d'effectuer véritablement une validation croisée pour les hyper-paramètres du notre modèle, permettrait de tester des modèles alternatifs, et d'en comparer les performances. Par ailleurs, la comparaison avec les modèles linéaires aujourd'hui utilisés au sein de BPCE devrait pouvoir dénoter les intérêts de la carte de Kohonen supervisée. L'utilisation de différentes distances lors de l'apprentissage de la carte en particulier (Euclidienne, DTW) devra également faire l'objet d'une étude plus poussée pour assurer une contribution équivalente des variables.

Par la suite, des réflexions sur l'adaptation du modèle aux contrats ayant moins de profondeur historique devront être menées. Enfin, l'utilisation de réseaux de neurones plus profonds pourrait être testée en vue d'une amélioration du pouvoir pédictif de notre modèle.

10 Bibliographie

Références

[1] BAIR, HASTIE, PAUL et TIBSHIRANI, (2006) : Prediction by supervised principal components, Journal of the American Statistical Association, Vol.101, No. 473.

[2] BERGLUND, Erik et al. The Parameter-Less Self-Organizing Map algorithm, IEEE Transactions on Neural Networks v.17, n.2, pp.305-316, 2006 (updated in 2013).

[3] COTTRELL, Marie. Nouvelles techniques neuronales en analyse des données : applications à la classification, à la recherche de typologie et à la prévision, Université Paris 1, 1998.

[4] COTTRELL, Marie et al. SOM-based algorithms for qualitative variables, Université Paris 1, 2005.

[5] FREEMAN, James et al. Neural Networks : Algorithms, Applications, and Programming Techniques, Computation and neural systems series, 1991.

[6] GOODFELLOW, Ian et al. Deep Learning, MIT Press, 2016.

[7] IBBOU, Smail. Classification, analyse des correspondances et méthodes neuronales, Thèse, Université Paris 1, 1992.

[8] KEOGH, Eamonn et al. Exact indexing of dynamic time warping, Knowledge and Information Systems, 2004.

[9] KOHONEN, Teuvo. Self-Organized Formation of Topologically Correct Feature Maps, Biological Cybernetics, vol. 46, pp. 59–69, 1982.

[10] KOHONEN, Teuvo. Essentials of the self-organizing maps, Neural networks, vol. 37, pp. 52–65, 2013.

[11] MELSSEN, Willem et al. Supervised Kohonen networks for classification problems, Chemometrics and Intelligent Laboratory Systems 83 99–113, 2006.

[12] PAGES, J. Analyse factorielle de données mixtes, Revue de statistique appliquée, 52 (4) : 93–111, 2004.

[13] PEARSON, Karl. On Lines and Planes of Closest Fit to Systems of Points in Space, Philosophical Magazine, 2 (11) : 559–572, 1901.

[14] RITTER, Helge. Neural computation and Self-Organizing Maps, an Introduction, Addison-Wesley, Reading, 1992.

[15] SEVERIN, Eric et al. Self organizing maps in corporate finance : Quantitative and qualitative analysis of debt and leasing, Neurocomputing Volume 73, Issues 10–12 p. 2061–2067, 2010.

[16] TENENHAUS, Michel. La régression PLS - Théorie et pratique , Technip, 2009.

[17] WOLD, Svante. Principal component analysis, Chenometrics and intelligent laboratory systems Volume 2, p. 37–52, 1987.

A – Utilisation de la méthode KACP

En guise d'exemple, nous créons un petit jeu de données contenant 4 variables macro-économiques (PIB par tête, tuax de mortality infantile, taux de croissance de la population et taux d'inflation) observées sur 272 pays et régions du monde. Les données sont disponibles dans la World Data-Bank [7]. On prend une carte de dimensions 6x6 (36 neurones). Après 1000 itérations, on obtient la carte suivante

KUWAIT LEBANON OMAN	IRAQ JORDAN WEST BANK AND GAZA	MALDIVES SAUDI ARABIA	BRUNEI DARUSSALAM CYPRUS ISRAEL NEW ZEALAND	AUSTRALIA MACAO SINGAPORE SWEDEN	BRITISH VIRGIN ISLAND LUXEMBOURG NEW CALEDONIA NORWAY QATAR SWITZERLAND TURKS AND CAICOS ISLANDS
ETHIOPIA KENYA MADAGASCAR TANZANIA UGANDA ZAMBIA	ALGERIA BELIZE EGYPT GUATEMALA KYRGYZ MOGOLIA SOLOMON ISLANDS VANUATU	KAZAKHSTAN MALAYSIA MEXICO PANAMA PERU SEYCHELLES TURKEY	ANITGUA AND BARBUDA BAHAMAS BAHRAIN CHILE FRENCH POLYNESIA MALTA ST. KITTS AND NEVIS	BELGIUM FINLAND FRANCE GERMANY HONG KING ICELAND JAPAN SAN MARINO UNITED ARAB EMIRATES UNITED KINGDOM	AUSTRIA CANADA DENMARK GIBRALTAR IRELAND ISLE OF MAN LIECHTENSTEIN NETHERLANDS ST MARTIN(FRENCH) UNITED STATES
GHANA LIBERIA MALAWI PAPUA NEW GUINEA SUDAN	BOTSWANA CONGO REP ERITREA GABON NAMIBIA RWANDA SAO TOME AND PRINCIPE TAJIKISTAN UZBEKISTAN YEMEN	CABO VERDE CAMBODIA CURACAO ECUADOR GUAM HONDURAS MOROCCO NICARAGUA PARAGUAY PHILIPPINES VIETNAM	ARGENTINA COSTA RICA NORTHERN MARIANA ISLANDS SRI LANKA ST. LUCIA	BARBADOS KOREAN DEM PPL REP TRINIDAD AND TOBAGO	AMERICAN SAMOA BERMUDA CUBA CZECH REPUBLIC ITALY KOREAN REP MONACO SLOVENIA SPAIN
GAMBIA SENEGAL TIMOR-LESTE TOGO ZIMBABWE	DJIBOUTI HAITI KIRIBATI LAO LESOTHO SWAZILAND TURKMENISTAN	AZERBAIDJAN BHUTAN BOLIVIA CAYMAN ISLANDS DOMINICAN REPUBLIC INDONESIA SOUTH AFRICA	BRAZIL COLOMBIA IRAN PALAU SURINAME TUNISIA	ARUBA CHANNEL ISLANDS CHINA DOMINICA EL SALVADOR FIJI GRENADA SAMOA ST VINCENT AND THE GRENADINES THAILAND TONGA	CROATIA ESTONIA FAROE ISLANDS HUNGARY POLAND SLOVAK REPUBLIC
BURUNDI NIGER SOUTH SUDAN	AFGHANISTAN BENIN BURKINA FASO CAMEROON COMOROS COTE IVOIRE GUINEA GUINEA-BISSAU MAURITANIA MOZAMBIQUE PAKISTAN	BANGLADESH INDIA MYANMAR NEPAL	BELARUS LIBYA MOLDOVA RUSSIA URUGUAY	ALBANIA ARMENIA JAMAICA MACEDONIA MAURITIUS MONTENEGRO	GREECE LATVIA LITHUANIA PORTUGA
ANGOLA CENTRAL AFRICAN REP CHAD SIERRA LEONE SOMALIA	CONGO DEM REP EQUATORIA GUINEA MALI NIGERIA SINT MAARTEN (DUTCH)	GUYANA MARSHALL ISLANDS MICRONESIA FED STS NAURU TUVALU VENEZUELA	BOSNIA AND HERZEGOVINA GREENLAND UKRAINE	BULGARIA KOSOVO ROMANIA SERBIA VIRGIN ISLANDS (US)	ANDORRA GEORGIA PUERTO RICO SYRIA

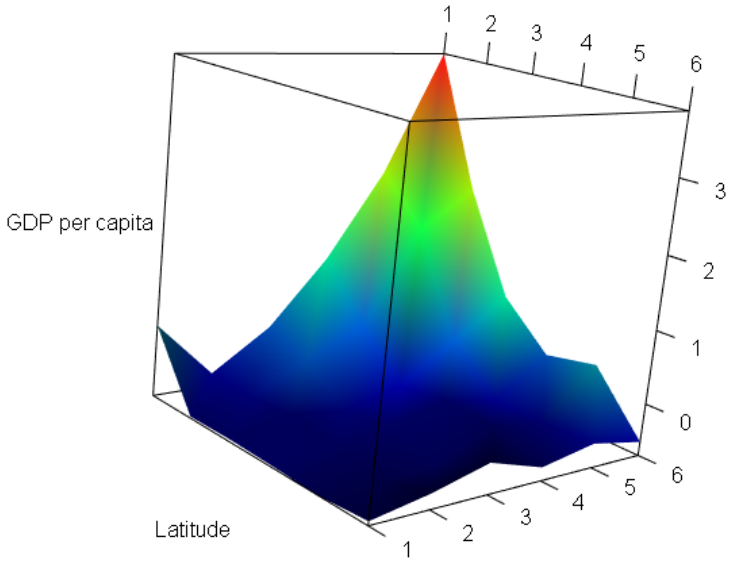
On reconnaît aisément en regardant cette carte quelques groupes que nous aurions pu constituer a piori. En effet, les pays d'Afrique sub-saharienne sont concentrés en bas à gauche de la carte, tandis que les pays du Moyen-Orient se trouvent en haut à gauche et que les pays d'Amérique du Sud sont au centre de la carte. Les pays européens sont quant à eux plus en haut à droite. Bien

7. http ://databank.worldbank.org/data/

entendu, l'idée n'est pas ici de produire une carte en s'appuyant sur la proximité géographique des pays, mais d'avoir proches sur la carte les pays ayant des caractéristiques communes selon les variables considérées. Notre satisfaction devant le reconnaissance de groupes de pays que nous aurions a priori mis ensemble vient du fait que les pays géographiquement proches partagent des éléments de culture, d'histoire et de ressources, et sont donc plus susceptibles d'avoir de caractéristiques similaires.

Regardons maintenant comment sont sitribuées les variables sur la carte :

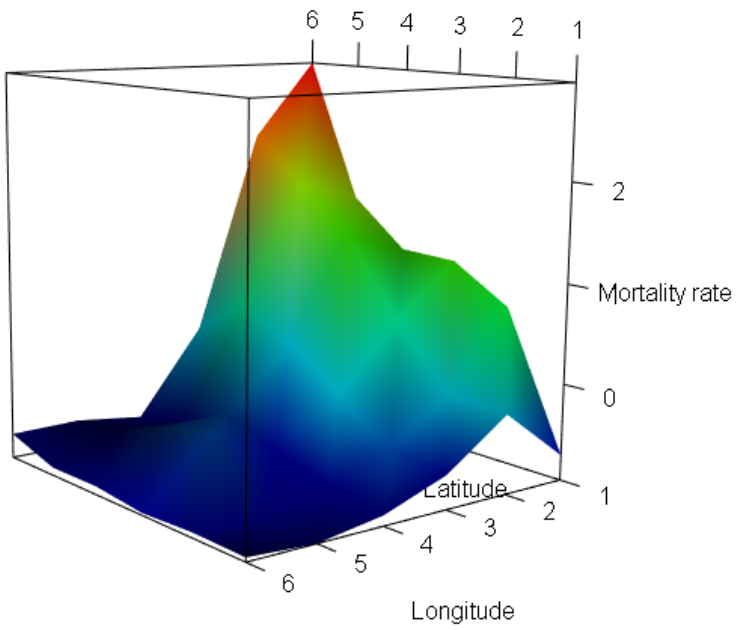
— PIB par tête



Cette carte est obtenue en représentant la valeur moyenne de la variable PIB par tête de chaque noeud de la carte. Nous pouvons voir que les valeurs les plus hautes sont concentrées autour des coordonnées (latitude = 1, longitude = 6), ce qui signifie que le coin en haut à droite de la carte est caractérisé par des fortes valeurs de PIB par tête, ce qui est confirmé par les pays classés dans cette région (Iles vierges, Luxembourg, Nouvelle Calédonie, Norvège, Qatar, Suisse).

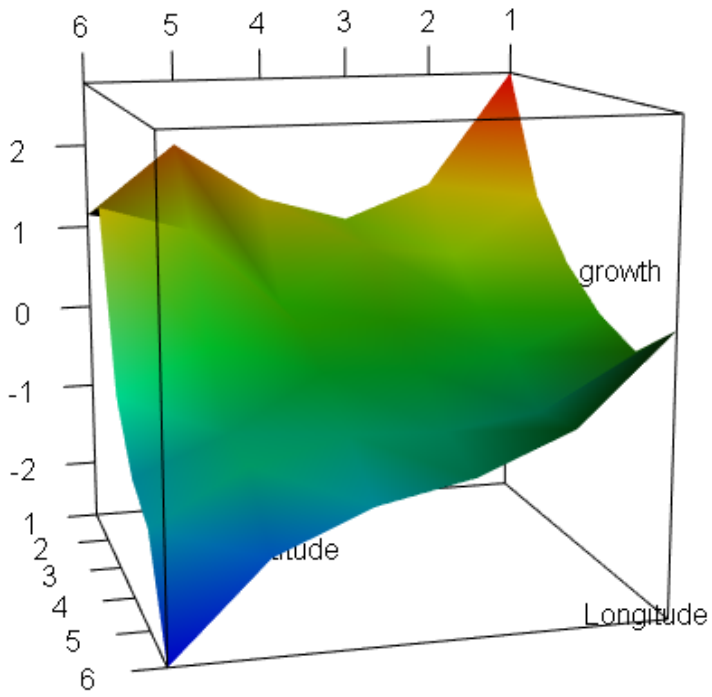
De la mêeme façon on représente les autres variables

— Taux de mortalité infantile (pour mille)



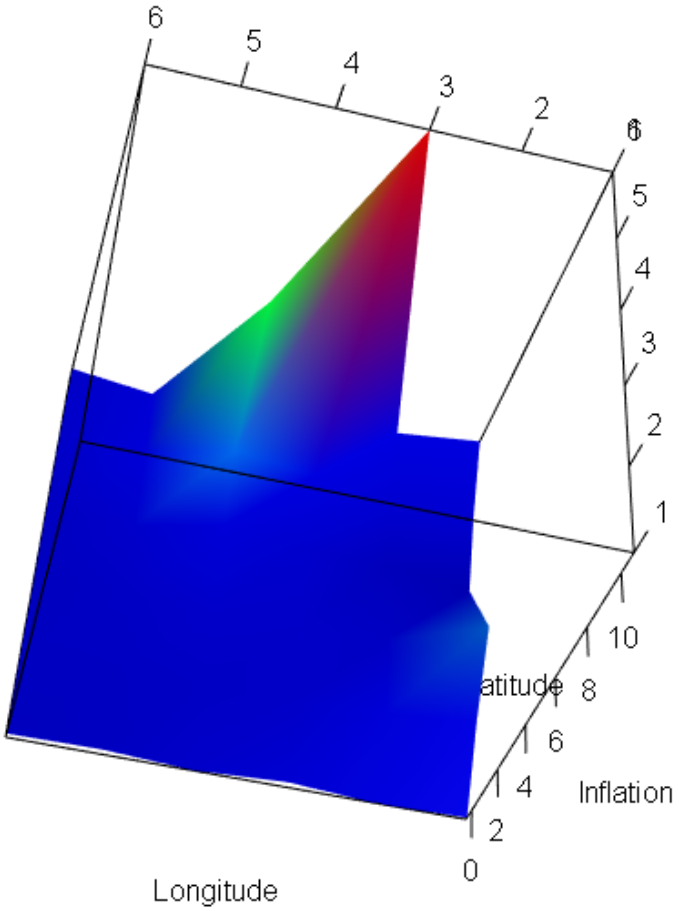
La variable de mortalité infantile tire tout un côté de la carte. Les pays aux coordonnées (6,1), l'Angola, la République d'Afrique Centrale, le Tchad, la Sierra Leone et la Somalie sont caractérisés presque intégralement par cette variable.

— Taux de croissance de la population



On voit que le taux de croissance de la population décroit à la fois le long de la lattitude et de la longitude, atteignant un maximum en (1,1) et un minimum en (6,6). On peut également voir que le côté de longitude égale à 1 est tiré par cette variable. Par comparaison avec la carte précédente, on en conclut que les pays avec le plus fort taux de mortalité infantile sont également les plus susceptibles d'avoir un fort taux de croissance de leur population.

— Taux d'inflation



La dernière variable considérée fait preuve de bien moins de linéarité. Les hauts niveaux d'inflation sont concentrés sur quelques neurones (6,3) et (6,4). C'est une bonne illustration de la capacité de la carte de Kohonen à avoir des neurones très spécialisés dans la classification de comportements spécifiques.

B – Utilisation de la méthode KACM

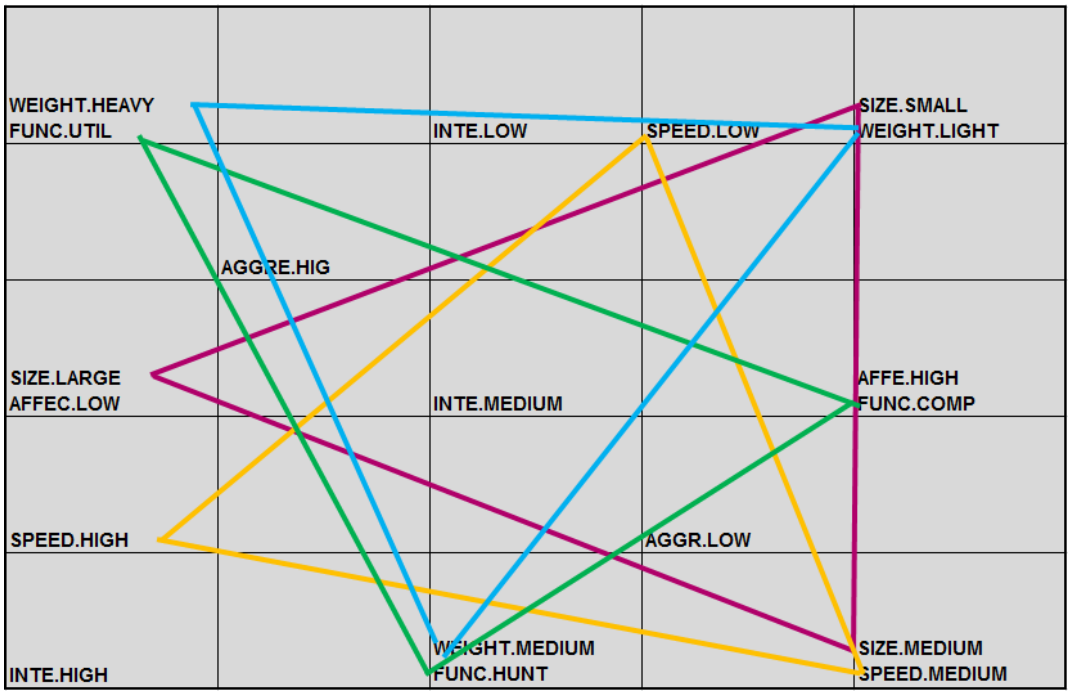
Pour produire un exemple simple de la méthode KACM, nous l'utilisons à célèbre jeu de données "BreedsDogs" disponible dans le package R "FactoClass". Ce jeu de données contient l'observation de 7 variables (taille, poids, vitesse, fonction, affection, agressivité et intelligence) sur 27 races de chien.

```
> BreedsDogs
     SIZE WEIG SPEE INTE AFFE AGGR FUNC
bass  sma  lig  low  low  low  hig  hun
beau  lar  med  hig  med  hig  hig  uti
boxe  med  med  med  med  hig  hig  com
buld  sma  lig  low  med  hig  low  com
bulm  lar  hea  low  hig  low  hig  uti
cani  sma  lig  med  hig  hig  low  com
chih  sma  lig  low  low  hig  low  com
cock  med  lig  low  med  hig  hig  com
coll  lar  med  hig  med  hig  low  com
dalm  med  med  med  med  hig  low  com
dobe  lar  med  hig  hig  low  hig  uti
dogo  lar  hea  hig  low  low  hig  uti
foxh  lar  med  hig  low  low  hig  hun
foxt  sma  lig  med  med  hig  hig  com
galg  lar  med  hig  low  low  low  hun
gasc  lar  med  med  low  low  hig  hun
labr  med  med  med  med  hig  low  hun
masa  lar  med  hig  hig  hig  hig  uti
mast  lar  hea  low  low  low  hig  uti
peki  sma  lig  low  low  hig  low  com
podb  med  med  med  hig  hig  low  hun
podf  lar  med  med  med  low  low  hun
poin  lar  med  hig  hig  low  low  hun
sett  lar  med  hig  med  low  low  hun
stbe  lar  hea  low  med  low  hig  uti
teck  sma  lig  low  med  hig  low  com
tern  lar  hea  low  med  low  low  uti
```

Nous utilisons une carte de dimensions 5x5 avec un taux d'apprentissage décroissant exponentiellement de 0.1 à 0.01 et un rayon d'apprentissage décroissant exponentiellement également de 2.99 à 0.65. Voici la distribution des modalités après 300 répétitions :

WEIGHT.HEAVY FUNC.UTIL		INTE.LOW	SPEED.LOW	SIZE.SMALL WEIGHT.LIGHT
	AGGRE.HIG			
SIZE.LARGE AFFEC.LOW		INTE.MEDIUM		AFFE.HIGH FUNC.COMP
SPEED.HIGH			AGGR.LOW	
INTE.HIGH		WEIGHT.MEDIUM FUNC.HUNT		SIZE.MEDIUM SPEED.MEDIUM

Cette carte semble très cohérente avec ce que nous aurions pu imaginer. En effet, certaines modalités sont très proches les unes des autres, voire assignées au même neurone : petit et léger, taille moyenne et vitesse moyenne, poids moyen et chasse, poids lourd et fonction utilitaire. Au contraire, certains modalités sont opposées l'une à l'autre, ce qui peut être illustré par exemple pour les variables à 3 modalités :



En outre, la proximité des triangles met en exergue une corrélation entre deux variables, tandis que la proximité entre deux modalités ne met en évidence qu'une corrélation entre ces modalités.

On peut le voir aisément en regardant les triangles bleu et vert, représentant respectivement les variables du poids et de la fonction. On observe que la fonction des races de chien et très fortement corrélée à leur taille ; la fonction compagnie étant également attirée par une forte affection et une faible agressivité.

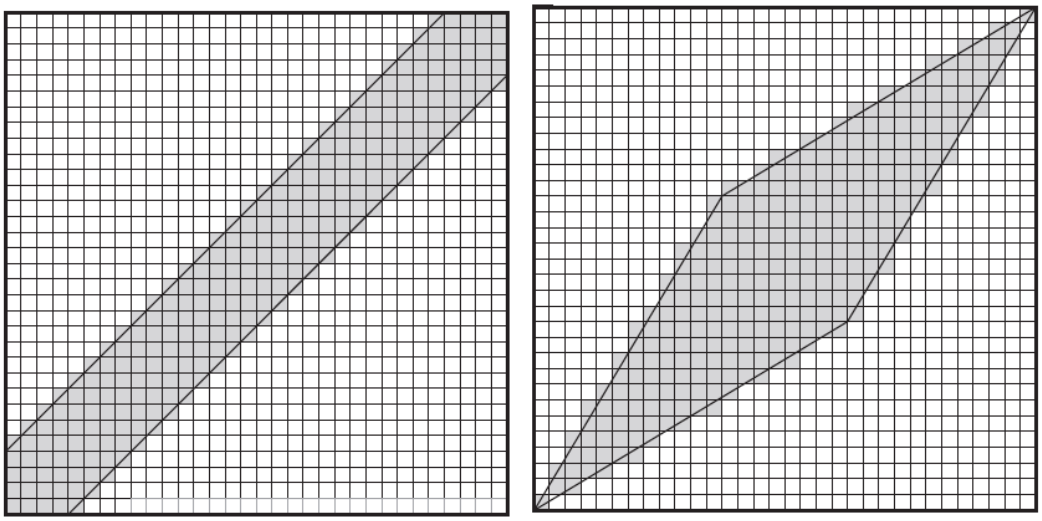
Pour cartographier les races de chien, on utilise le tableau disjonctif complet corrigé K^c, ce qui va nous permettre de d'avoir simultanément sur une même carte les races de chien et les modalités des variables. Lorsque la carte est entrainée sur les modalités, on utilise K^c en entrée, et les races de chien sont assignées au noeud leur correspondant.

WEIGHT.HEAVY **FUNC.UTIL** bulm - dogo- mast - stbe - tern		**INTE.LOW** bass	**SPEED.LOW**	**SIZE.SMALL** **WEIGHT.LIGHT** buld - cani - chih - foxt - peki - teck
	AGGRE.HIG			
SIZE.LARGE **AFFEC.LOW**		**INTE.MEDIUM**		**AFFE.HIGH** **FUNC.COMP** cock
SPEED.HIGH beau - dobe - foxh		coll	**AGGR.LOW**	
INTE.HIGH masa	poin	**WEIGHT.MEDIUM** **FUNC.HUNT** galg - gasc - podf - sett		**SIZE.MEDIUM** **SPEED.MEDIUM** boxe - dalm - labr - podb

Les races de chien sont bien classées au regard des modalités. Les races Saint-Bernard et Terre-neuve par exemple sont bien dans le même noeud que les modalités poids lourd et fonction utilitaire. Petite taille et poids léger sont également dans le même noeud que les races bulldog, caniche, chihuahua, foxterrier, pekinois et teckel.

C – Contraintes de la distance “dynamic time warping”

Comme évoqué plus haut, l’utilisation de contraintes permet d’accélérer le temps de calcul de l’algorithme. Ces contraintes consistent à réduire le nombre de chemine en supprimant les possibilités trop éloignées de la diagonale de la matrice de coût – la diagonale correspondant au cas où les deux axes temporels sont déjà alignés – :



Dans ces figures, deux contraintes globales sont illustrées. Celle de gauche est la bande de Sakoe-Chiba ; qui force le chemin de déformation à rester à moins d’un certain nombre de points de la diagonale. À droite, le parallélogramme d’Itakura qui donne plus de liberté à l’algorithme au milieu de la ligne temporelles des séries, mais qui les force à s’aligner dans les premiers et derniers instants.

D’autres contraintes, locales, concernent le choix de pas pour l’algorithme, la notion de contiguïté. En regardant l’équation initiale sans contrainte,

$$\gamma(i,j) = c_{ij} + min(\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)),$$

on voit qu’un pas en diagonale est équivalent à un pas selon chaque axe temporel. Cependant,

cette équation peut être modifiée de cette façon

$$\gamma(i,j) = c_{ij} + min(\gamma(i-1,j-1), \gamma(i-1,j-2), \gamma(i-2,j-1)),$$

et devient telle que le choix ne se fait pas entre éléments adjacents mais entre un pas en diagonal ou un pas le long d'un axe plus un pas en diagonal.