UNIVERSITE KASDI MERBAH OUARGLA
Faculté des Sciences et Sciences de l'Ingénieur
Département de Génie des Procédés

N⁰ d'ordre : ……
Série : ……

Mémoire
Présenté en vue de l'obtention du diplôme de magister
en Génie des Procédés
Option : Chimie organique et physicochimie moléculaire

Présenté par :
Thair Bouziane

Thème

Contribution à l'étude de la relation structure chimique-odeur : Utilisation des réseaux de neurones et la logique floue. (Application à la famille des pyrazines)

Soutenu publiquement le : 19 /11 /2008

Devant le jury composé de :

DADA MOUSSA Belkheir	Professeur. UKM - Ouargla	Président
GHERRAF Noureddine	MC. U.Larbi ben m'hidi- Oum El-bouagui	Examinateur
DOUNIT Salah	MC. UKM - Ouargla	Examinateur
BEN HELLAL Belkheir	MA. UKM - Ouargla	Invité
LADJEL Segni	MC. UKM - Ouargla	Rapporteur
KORICHI Mourad	MA. UKM - Ouargla	Encadreur

Année Universitaire

2008/2009

Dédicace

Mes chère parents, sachez que vous êtes les meilleurs parents qu'une personne peut avoir, si j'avais à choisir mes parents je vous aurais choisis. Vous m'avez donné tout votre amour, votre temps, vous étiez toujours à mes côtés pour me réconforter, me soutenir, et m'épauler.

Vous avez fait de votre mieux pour me donner une bonne éducation, pour me rendre heureux, et pour que je sois le meilleur de tous.

Je voulais réussir pour vous, rendre heureux. Je remercie Dieu jour et nuit parce qui me vous a donnée.

Je dédie ce travail à :

En premier lieu, à mon père

Ma mère

Mes frères, mes sœurs, spécialement à Med IKBAL et KHADIDJA avec toute ma reconnaissance

Mon Cher Med ARBAOUI, Med DABABI, THAMER et RAMZI

Toute la promotion Magister chimie organique moléculaire 2005

A tous ceux qui me sont chers

THAIR

Remerciements

Avant tout, louange au bon dieu qui nous a aidés à terminer ce travail :

Je tiens à remercier vivement ma famille, pour son soutien et ses encouragements durant toutes mes années d'études.

En présentant ce travail, je tiens à remercier à **Dr. KORICHI Mourad,** mon encadreur, pour sa serviabilité, sa disponibilité et ses remarques constructives qui m'ont servis tout au long de mon projet, et pour son aide précieuse que ça soit en matière de documentation ou bien aides morales qu'il m'a témoigné tout au long de ce travail.

Je remercie le président de jury **Pr.DADA MOUSSA Belkheir**, qui a accepté de juger ce travail.

Mes vifs remerciements aussi à **M.A. BEN HELAL Belkheir,** qui a dirigé ce travail, merci pour votre aide disponibilité et pour vos conseils judicieux.

Mes très grands remerciements à toute l'équipe de laboratoire de simulation :

- Le responsable M.A**. SAIDAT Mustapha ;**
- Mlle ZIGHMI Souad;
- Mlle CHAOUCH Noura;
- Mes collègues au laboratoire de simulation.

Je remercie particulièrement pour l'équipe et les responsables de laboratoire VPRS. Je ne saurais oublier mes enseignants et mes camarades qui trouvent ici l'expression de mon profond respect

Un grand merci encore est adressé à tous ceux qui d'une façon ou d'une autre ont fait part de leurs aide, m'ont encouragé et participé de près ou de loin à la réalisation de ce travail.

Sans oublier personne, je ne pourrais citer tous, merci à tous, que dieu vous protège et vous garde.

Thair

Liste des tableaux

Liste des Figures

Introduction générale :

Les molécules organiques odorantes sont utilisées dans la majorité des produits de consommation pour inciter les consommateurs à associer des impressions favorables a un produit donné.

Selon **REACH** (la nouvelle législation chimique européenne), qui a pour objectif d'offrir au public une meilleure protection vis-à-vis les substances chimiques, certaines d'entre elles (molécules odorantes) doivent être remplacées (toxicité, allergie,…)

La recherche de molécules odorantes nouvelles proches de molécules commerciales existantes nécessite la connaissance de la relation structure moléculaire-odeur. Cette dernière est difficile à modéliser, en raison de la subjectivisée de l'odeur (Amboni et al., 2000). La recherche scientifique dans ce domaine est orienté vers l'utilisation des paramètres qui représentent la structure moléculaire (description moléculaires) tels que : paramètres structuraux, indices topologique, indices géométrique, facteurs électronique et propriétés physico chimiques pour développer des modèle capable de prédire/classifier l'odeur des molécules. Un certain nombre de méthodes et techniques ont été employées avec succès dans ce domaine, à savoir les réseaux de neurones (chastrette et al. (1995) et cherqaoui et al. (1998)).

L'objectif du présent travail est d'étudier la classification de l'odeur pour la famille des pyrazines, nous utilisons une approche basée sur les descripteurs moléculaires pour la prédiction et la classification de l'odeur en appliquant les techniques d'analyses des données et des méthodes de régression/ classification, spécialement les réseaux de neurones et la logique floue.

Le plan de travail de cette recherche est divisé sur cinq (5) chapitres comme suite :

Chapitre 1 : les odeurs et molécules odorantes.

Chapitre 2 : construction de la base des données

Chapitre 3 : application de la méthode analyse discriminante sur molécule de pyrazine.

Chapitre 4 : applications méthode de logique floue sur la molécule de pyrazine.

Chapitre 5 : application méthode de réseaux neurones sur la molécule de pyrazine.

Conclusion générale

Problématiques :

Le rapport entre qualité de l'odeur et les propriétés des moléculaire sont la question la plus importante d'olfaction. En début de sophistication dans la caractérisation chimique des molécules, accompagnées pour la caractérisation perceptrice, ont peu d'utilité quantitative, compter sur la description de l'énumérative principalement.

Cette relation entre la qualité de l'odeur et les propriétés moléculaires sont définies par un nouveau terme pour l'odeur, c'est **la classification des odeurs.**

Dans cette étude nous présentons la classification de la famille pyrazine avec base des données pour cette famille (125 molécules) sur six (6) odeurs différentes correspondant à cette famille (**GREEN, NUTTY, BELL-PEPPER, EARTHY, PECASY ET SWEET).**

Pour l'étude de cette classification sur la famille pyrazine, on utilise trois (3) méthodes d'analyse et de classification qui sont : analyses discriminantes, logiques floues et réseaux des neurones. On utilise les propriétés moléculaires comme descripteurs qui représentent le modèle de classification pour la famille pyrazine à partir d'utilisation des trois méthodes d'analyse précédentes.

Le but de ce travail est d'obtenir meilleur modèle de classification pour la famille pyrazine après l'utilisation des trois méthodes précédentes (analyses discriminantes, logiques floues et réseaux des neurones). Ce modèle de classification est utilisé sur les molécules de la même famille (pyrazine) pour obtenir les classifications de ces molécules qui ne connait pas ses classifications correspondante.

I.1. Les odeurs

I.1.1. Introduction

Jusqu'à la montée en puissance de la chimie organique à la fin du XIX^{e} siècle, les parfums sont composés principalement à partir de produits d'origine végétale auxquels s'ajoutent quelques composants d'origine animale comme le musc ou l'ambre gris. Le traitement des plantes fournit des hydrolats (eau de rose, eau de fleur d'oranger…), des huiles essentielles et des alcoolats (alcoolat de romarin, base de l'Eau de la Reine de Hongrie ou alcoolat de mélisse, à l'origine de l'Eau de mélisse des Carmes).

Il existe aussi depuis très longtemps des compositions parfumées, faites de substances solides. Ainsi, l'encens, qui est probablement l'un des plus anciens parfums connus, utilise la technique de la combustion pour produire les substances odorantes. De même, la myrrhe, qui servait à la momification en Egypte, était utilisée dans la composition des parfums.

L'addition à des compositions classiques de produits chimiques de synthèse va révolutionner l'art de la parfumerie en permettant des accords nouveaux. La coumarine dont la synthèse est due à Perkin - par ailleurs auteur du premier colorant de synthèse, la mauvéine est l'un des premiers produits à entrer dans un parfum célèbre de Guerlain.

I.1.2. Définition d'une odeur

Une odeur est une émanation transmise par un l'air et perçue par l'appareil olfactif. Les récepteurs olfactifs sont situés dans les fosses nasales et sont reliés au cerveau par le nerf olfactif. Il existe sept odeurs primaires qui correspondent aux sept types de récepteurs sensoriels situés sur les cils des cellules olfactives. Des substances ayant des odeurs semblables ont des molécules de forme similaire. La forme d'une molécule détermine la nature de son odeur. Ces molécules se fixent sur les récepteurs. Une même molécule peut se fixer sur plusieurs sites si elle possède la " clé " de récepteurs différents. Ce phénomène est le début d'une série d'événements : transmission de l'influx par le nerf olfactif et perception d'une odeur par le cerveau [1].

On peut distinguer sept odeurs de base :

Camphrée, musquée, florale, mentholée, éthérée, piquante, putride

I.1.3. La **perception des odeurs**

I.1.3.1. Sensibilité **olfactive et seuils de perception**

On peut distinguer deux valeurs de seuils: le seuil de détection, lorsque le sujet a une sensation olfactive sans pour autant affecter de label à cette odeur, et le seuil de reconnaissance, lorsque la qualité de l'odeur est accessible au sujet. Plusieurs méthodes sont utilisées pour mesurer la sensibilité olfactive chez l'homme. La plupart de ces procédures utilisent des présentations par paires, et elles tirent leur spécificité par le mode de détermination des seuils. La détermination des seuils, se fait généralement en utilisant une gamme de concentrations croissantes d'un odorant mis en solution dans un solvant [2]. Trois grandes méthodes peuvent être définies [3] :

- La première technique consiste à présenter de manière croissante des concentrations, jusqu'à ce que le sujet détecte correctement le bon flacon. La valeur seuil est prise comme la première concentration des 3 ou 4 successivement correctes.
- Dans la seconde technique, les concentrations sont présentées de manière croissante et décroissante, jusqu'à ce que le sujet stabilise ses réponses selon un procédé bien défini.
- Enfin, la troisième technique, développée par Doty et al. (1986). Les seuils sont déterminés grâce à une procédure de choix forcés. L'expérience comporte plusieurs essais. Un essai consiste en la présentation successive rapide de 2 flacons : un contenant du solvant et l'autre contenant l'odorant étudié, dilué dans le solvant. En fonction de l'odorant étudié, 10 ou 12 niveaux de concentrations sont préparés, soit 10 ou 12 paires. Les deux bouteilles d'une même paire sont ouvertes et immédiatement placées sous le nez du sujet. La tâche du sujet est de déterminer parmi les deux flacons celui qui évoque la plus forte sensation olfactive. Même s'il ne perçoit aucune sensation ou s'il ne détecte aucune différence, le sujet doit choisir l'un des 2 flacons. On n'indique pas au sujet s'il a fait le bon ou le mauvais choix. Dans le cas d'un choix incorrect, on lui présente la paire de flacons correspondant à la concentration de 2 niveaux supérieurs. Par contre, si le choix du sujet est correct, alors on lui présente la même paire de flacons (même niveau de concentration de l'odorant). On prend comme valeur seuil initiale le niveau de concentration pour lequel le sujet indique 4 fois consécutivement le flacon contenant l'odorant. Une fois la valeur du seuil initial obtenue, on inverse la progression en présentant au sujet la paire de flacons correspondant au niveau de concentration directement inférieur à celui du seuil initial.

S'il répond correctement, on représente le même niveau de concentration, et s'il répond correctement deux fois de suite, on passe alors au niveau de concentration directement inférieur. On effectue au total 7 renversements de niveaux de concentration. Le seuil final est pris comme la moyenne des 4 derniers renversements.

I.1.3.2. Solubilité des odeurs

Pour être odorante, une substance doit être légèrement soluble dans l'eau, donc posséder une partie hydrophile pour se dissoudre au niveau du mucus sécrété par l'épithélium olfactif recouvrant les cellules nerveuses sensorielles. Afin d'atteindre les récepteurs olfactifs, la substance doit ensuite pouvoir traverser la membrane cellulaire constituée de lipides (graisses). La solubilité dans les lipides doit être suffisante : elle doit donc présenter une partie lipophile [4].

I.1.3.3. Volatilité des odeurs

Un corps qui ne possède pas de molécules volatiles n'émet pas d'odeur (ex : le verre).Pour qu'un corps soit odorant, il doit d'abord être assez volatil pour être facilement vaporisé à température ordinaire et atteindre les récepteurs olfactifs.

La volatilité des constituants d'un parfum est une contrainte chimique qui doit être maîtrisée par l'industrie des parfums. Ainsi, si les composés sont trop volatils, l'odeur va rapidement se dissiper. C'est pourquoi, un parfum contient des molécules lourdes inodores. Non volatiles, elles exercent un effet fixateur du parfum lui conférant une plus grande ténacité [2].

I.1.4. Structures moléculaires odoriférantes

Les composés odorants peuvent être classés selon la principale fonction qu'ils possèdent ; comme [5] :

- Les hydrocarbures ou essences, surtout, des alcènes (ex : limonène)
- les alcools, dont les chaînes carbonées comportent huit à douze carbones (ex : le menthol ou autre que l'on trouve dans les fleurs)

- les phénols, composés dans lesquels le groupe hydroxyle -OH est porté par un atome de carbone trigonal (ex : le thymol pour le thym, l'eugénol pour le clou de girofle)
- les éther-oxydes, composés dans lesquels un atome d'oxygène est lié à deux chaînes carbonées (ex : anéthol de l'anis)
- les aldéhydes dont les chaînes carbonées comportent de huit à douze atomes de carbones (ex : citral de la citronnelle)

- -les cétones comme irone de l'iris

- les esters, composés oxygénés présents dans presque toutes les huiles d'origine végétale (ex : acétate de linalyle de la lavande)

- les composés azotés tels que le musc xylol.

I.1. 5. Les sources d'odeurs

Les activités susceptibles de provoquer des problèmes d'odeurs sont relativement nombreuses. On peut citer par exemple [6] :

I.1.5.1. Secteur agricole :

- Sources étendues : épandage en surface (lisiers, boues, produits de traitement…)
- Sources ponctuelles : élevages (bovins, volailles…)

I.1.5.2. Secteur industriel :

- Industries agroalimentaires,
- Raffineries de pétrole.
- Industries chimiques,
- Industries des matières plastiques,
- Métallurgie,
- Épuration des eaux usées : stations d'épuration
- Traitement des déchets.

Pour une même unité industrielle, les sources sont diverses :

- effluents canalisés (cheminée)
- sources ponctuelles génératrices d'odeurs très intenses à proximité immédiate (évents, puisards…)
- sources d'odeurs peu intenses mais qui peuvent représenter des nuisances importantes du fait de la surface d'échange (décanteurs, bassins d'épandage…)

I.1.6. La mesure des odeurs

I.1.6.1. L'olfactométrie

Le terme 'olfactométrie' désigne à la fois la **mesure des odeurs** et la **mesure des capacités olfactives** d'un sujet. Dans les deux cas un capteur est mis en présence d'une odeur (le stimulus). Un stimulus connu permet de caractériser le capteur, un capteur étalonné permet d'objectiver le stimulus [2].

Les paramètres d'une odeur sont [6]:

- **Quantitatif :** son intensité, sa force
- **Qualitatif :** sa description comme la référence à un objet odorant (ex. la rose) ou la constitution chimique du mélange odorant
- **Temporel :** l'évolution dans le temps de son intensité et/ou de sa qualité (ex. note de tête d'un parfum)

Dans de nombreux domaines il s'avère nécessaire d'évaluer une odeur par exemple dans la parfumerie et les cosmétiques, le contrôle de qualité et les tests de préférence dans l'agro-alimentaire, le traitement des nuisances olfactives, etc. Les méthodes utilisées pour mesurer les odeurs sont l'analyse physico-chimique et l'utilisation d'un 'jury de nez' entraîné. Dans le cas de l'évaluation de nuisance olfactive on utilise aussi les enquêtes auprès des populations concernées. Enfin des nez artificiels ou nez électroniques commencent à apparaître sur le marché.

La caractérisation des sources émettrices d'odeurs ; soit par [2,5] :

- **l'analyse olfactométriques** : résultats que ce type d'analyse peut fournir est le niveau d'odeur, le débit d'odeur, l'intensité de l'odeur ou encore l'aire de persistance de la nuisance. Les principes de mesures de niveau d'odeur sont décrits dans la norme AFNOR NFX 43-101.

- **l'analyse physico-chimique** : permet l'identification et la quantification des composés incriminés (analyse complexe car plusieurs centaines de composés dans les effluents).

Une odeur est liée à la présence de composés chimiques de l'air. La méthode physico-chimique consiste à rechercher et quantifier les éléments chimiques présents dans l'atmosphère puis a se reporter à une table des propriétés olfactives de chaque corps pour caractériser l'odeur résultante. Cette méthode a des limites :

- la concentration des produits odorants est souvent si faible qu'ils ne sont pas détectables, même par les analyseurs les plus performants,
- les propriétés olfactives des mélanges sont différentes de celles des constituants pris séparément, et souvent varient avec la concentration
- enfin, tous les composés chimiques ne sont pas odorants, mais ils peuvent influencer la sensation perçue avec le mélange.

I.1.6.2. Grandeurs olfactométriques (quelques définitions) [6]

- **Seuil olfactif** : pour chaque corps pur ou mélange odorant, on peut définir une concentration seuil pour laquelle l'effluent est ressenti comme odorant par 50 % des membres d'un jury constituant un échantillon de population. Dans le cas d'un corps pur, cette concentration est appelée par convention « seuil olfactif ».

- **Niveau d'odeur** : ce niveau est défini conventionnellement comme étant le facteur de dilution qu'il faut appliquer à un effluent pour qu'il ne soit pas ressenti comme odorant par 50 % des personnes d'un jury constituant un échantillon de population. On parle aussi de « Facteur de dilution au seuil de perception ». (Art. 29 de l'AM du 02/02/98)

- **Débit d'odeur** : le débit d'odeur est défini conventionnellement comme étant le produit du débit d'air rejeté, exprimé en Nm^3/h, par le facteur de dilution au seuil de perception (Art. 29 de l'AM du 02/02/98). Combiné à un modèle de dispersion atmosphérique, cet indicateur permet de déterminer une aire de persistance de la nuisance en fonction des conditions météorologiques.

- **Intensité d'odeur** : l'intensité d'odeur ou intensité odorante caractérise la grandeur de la sensation olfactive. Sa mesure, réalisée par un jury entraîné, consiste à comparer l'intensité du mélange gazeux à l'intensité d'échantillons de référence.

I.1.6.3. Mesure de l'intensité de l'odeur

L'intensité perçue d'un mélange odorant est une fonction de sa concentration, la mesure de la concentration requiert un olfactomètre à dilution. Il existe plusieurs méthodes permettant d'objectiver la concentration d'un mélange. La plus méthode la plus courante consiste à placer le jury et l'olfactomètre dans un local déodorisé. Le mélange odorant, après avoir été dilué par un gaz inodore, est présenté aux membres du jury.

Une autre méthode, applicable sur le terrain, utilise un olfactomètre pouvant délivrer un stimulus odorant d'intensité connue et réglable de n-butanol. La méthode consiste à demander au jury d'égaliser l'intensité du stimulus et celle de l'air ambiant. On caractérisera alors l'intensité comme étant équivalente à une concentration de n-butanol [2,6].

I.1.7. Le jury de nez

Pour un spécialiste des études de nuisances olfactives le terme 'olfactométrie' désigne la mesure des odeurs à l'aide d'un jury de nez. Par rapport à un instrument d'analyse l'être humain fait une évaluation directe de l'odeur.

Un jury est constitué de quatre à seize sujets sélectionnés sur leurs capacités olfactives, ils doivent être représentatifs de la moyenne de la population. Ces sujets sont entraînés en fonction de la tâche qui leur sera confiée : s'ils doivent décrire des odeurs complexes leur entraînement consistera à détecter et nommer les constituants d'un mélange, s'ils doivent mesurer des intensités ils classeront par ordre concentration croissante des solutions. Si ces solutions sont distribuées dans une large gamme il est alors possible de déterminer les seuils de détection des sujets pour le produit considéré.

Ce jury, une fois entraîné, est placé dans les situations olfactives pour caractériser une source de nuisance, tester l'efficacité d'un procédé de désodorisation, évaluer la dilution d'une odeur dans l'environnement [3,6].

I.1.8. Les nez artificiels

Un nez artificiel est constitué d'un ensemble de capteurs chacun faiblement sélectif à un composé chimique. Il existe des capteurs de type oxydes métalliques, de type polymères conducteurs et des biosenseurs à base de bicouches lipidiques. Ces capteurs sont associés à un dispositif de traitement des informations par réseau de neurone formel de type percepteur multicouche. Après une période d'entraînement, supervisée à l'aide d'une règle de rétro-propagation de l'erreur, ces dispositifs sont capables d'identifier certains mélanges mais surtout d'évaluer les différences d'un mélange odorant avec un prototype. Les performances de ces nez électroniques sont très inférieures à celles d'un nez humain mais ils sont infatigables, c'est pourquoi ils commencent à être utilisés dans le contrôle de qualité principalement dans l'agro-alimentaire [2,4].

I.1.9. Les nuisances

Les odeurs sont généralement dues à une multitude de molécules différentes, en concentration très faible, mélangées à l'air que nous respirons. La plupart des composés odorants sont détectés à des niveaux très faibles par rapport aux niveaux toxiques. A l'inverse, des gaz très toxiques comme le monoxyde de carbone n'ont aucune odeur.

Les nuisances sont liées à la perception des odeurs agréables ou désagréables, aimées ou détestées. L'odeur dépend de la dilution plus ou moins forte (d'où une intensité plus ou moins élevée) d'une ou d'un mélange de substances chimiques dans l'atmosphère [5].

Les substances appartiennent aux principaux composés suivants :

- composés azotés (amine, ammoniac…)
- acides gras volatils,
- aldéhydes et cétones,
- composés soufrés (hydrogène sulfuré, mercaptans, sulfures et disulfures…)
- mélange de ces composés.

Les substances sont issues de décomposition thermique ou anaérobie de composés chimiques, de produits animaux ou de déjections animales.

I.1.10. Le traitement des odeurs [6] :

- soit une réduction à la source en amont de l'émission (confinement des effluents, étude de la ventilation des locaux, agencement des aires de stockage…)
- soit un traitement de l'effluent gazeux : combustion thermique (env. 750 °C - ex pour les UIOM) ou catalytique (T° inférieure), adsorption (par charbon actif - ex pour les STEP, silos à boues, désodorisation de locaux…), l'absorption par voie humide ou lavage, la biodésodorisation (transformation de polluants par des microorganismes
- soit pour réduire la nuisance et pas forcément l'émission : utilisation de masquants (superposer une odeur à une autre) ou autres produits en pulvérisation.

I.1.11. Résumé

Les odeurs et les nuisances olfactives sont des préoccupations environnementales dont l'importance est croissante, aussi bien du côté des industriels (milieu émetteur) qui cherchent à maîtriser ces nuisances que du côté de la population riveraine (milieu récepteur) qui exige le respect de son cadre de vie.

Leurs effets sont difficiles à caractériser de manière précise mais les nuisances olfactives doivent être prises en compte en matière de qualité de l'air car leurs conséquences sur la santé au sens large sont indéniables.

Les nuisances olfactives apparaissent comme le deuxième motif de plainte après le bruit leur association à une notion de toxicité est rarement justifiée sur le plan physiologique, les odeurs étant le plus souvent perçues à des concentrations très faibles, inférieures aux limites acceptables pour la santé.

Le seuil de perception olfactive peut varier couramment d'un facteur 10 à 100 entre des personnes différentes ou pour une même personne en fonction de nombreux facteurs (humidité relative, température, présence, d'autres composés dans l'air, fatigue…)

II.1. Introduction

Dans ce chapitre nous construisons la base des données de famille pyrazine qui m'intéressent pour mes études de la classification les molécules pour cette famille avec les odeurs correspondante (Green, Nutty, Bell-pepper, Earthy, Pecasy, Sweet).cette base des données sont construites à partir des études ou des recherches (les publications) faite sur la même famille (pyrazine) pour la classification.

Après cette opération nous trouvons 125 molécules différentes de structure moléculaires de famille pyrazine qui correspondant a six (6) odeurs différentes (classification des molécules de famille pyrazine avec les odeurs correspondant différentes).

Enfin nous calculons les discripteurs coréspendant pour chaque molécule de base des données de famille pyrazine, nous utilisons pour cette opération (calcules les discripteurs) logiciel spécifique sont nommer « DRAGON » qui donnent 1664 discripteurs pour chaque molécules du base des données, cette discripteurs utilisant dans les études de classification pour optimisant meilleur modèles de classification dans les chapitres suivants.

Avants tous ça nous présentons la molécule de pyrazine, et spécifiquement la base de molécule pyrazine qui sont basé pour construire la base des données.

II.2. Construction de la base des données

Pour construire la base des données de famille pyrazine, nous basons sur l'article de : Bettina Wailzer, Johanna Klocker, Gerhard uchbauer, Gerhard Ecker, et Peter Wolschann titré « **Prediction of the Aroma Quality and the Threshold Values of Some Pyrazines »** [9], qui constitue 98 molécules de pyrazine avec trois types d'odeurs correspondantes (trois classifications). Et pour compléter la base des données, nous utilisons les articles qui sont présentés dans la page bibliographie sous la numérotation suivante : 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 et 26.

Après les recherches que nous avons faites sur la famille pyrazine dans les articles précédents, nous trouvons 125 molécules de famille pyrazine avec 6 classifications d'odeur (6 types d'odeurs pour cette famille).

Les six classifications de famille pyrazine ou six les types d'odeurs sont :

Green, Nutty, Bell-pepper, Earthy, Pecasy, Sweet.

Elle defining comme suites:

- **GREEN (**VERT) : nous présentons 32 molécules correspondant à cette odeur pour Base des données de famille pyrazine.
- **NUTTY** (DE NOIX) : nous présentons 25 molécules correspondant à cette odeur Pour base des données de famille pyrazine.
- **BELL-PEPPER (**BELL-PEPPER) : nous présentons 44 molécules Correspondant à cette odeur pour base des données de famille pyrazine.
- **EARTHY (**TERREUX) : nous présentons 14 molécules correspondant à cette odeur Pour base des données de famille pyrazine.
- **PECASY (**PECASY) : nous présentons 06 molécules correspondant à cette odeur Pour base des données de famille pyrazine.
- **SWEET (**SWEET) : nous présentons 04 molécules correspondant à cette odeur pour Base des données de famille pyrazine.

II.2.1. Présentation de la molécule de base du pyrazine

La molécule de base du pyrazine est une composées aromatique et la base structurelle de cette molécule c'est le benzène qui constitue deux (2) atomes d'azote symétriques et quatre (4) radicaux chimique. A partir de la structure de molécule de base du pyrazine que nous définirons les indices qui se trouvent dans la structure de molécule de base du pyrazine sont comme suit [9] :

- **R1, R2, R3, R4 :** les radicaux chimiques qui présentent la structure de molécule Pyrazine

.

- **N :** l'atome d'azote.

Nous présentons le schéma de la structure de molécule de base du pyrazine comme suit :

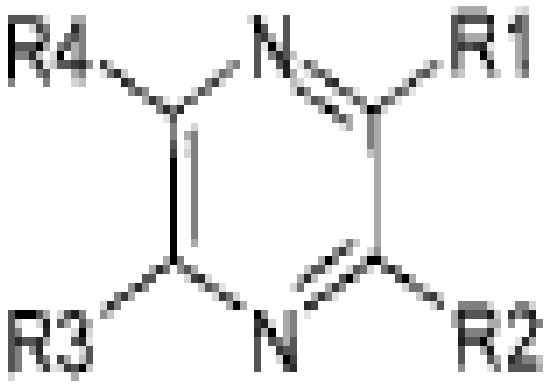


Figure N°05 : présentation de la molécule à base du pyrazine.

II.2.2. Présentation tableaux de base des données

Enfin nous représentons tableaux de base des données qui donnent la structure moléculaire pour toutes les molécules de famille pyrazine avec les classifications correspondante (6 types d'odeurs pour cette famille) à partir de la molécule de base du pyrazine.

Elle peut être définie par les termes suivants qui de trouvent dans le tableaux de base des données comme suite ; [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] :

- **R1, R2, R3, R4 :** les radicaux chimiques qui présentent la structure de molécule Pyrazine, ces radicaux que nous définirons dans le tableau de base des Données à partir de molécule de base du pyrazine.

- **Qualités :** classification des odeurs du pyrazine (1= **green**, 2 = **nutty**, 3 = **bell - pepper**, 4 = earthy, 5 = **pecasy**, 6 = **sweet**).

Nous présentons les tableaux de base des données comme suit :

N°	R1	R2	R3	R4	qualité
Obs001	$N(CH_3)_2$	H	H	$CH_2CH(CH_3)_2$	1
Obs002	OC_4H_9	H	H	H	1
Obs003	OC_6H_5	H	$CH(CH_3)_2$	H	1
Obs004	SC_2H_5	H	$(CH_2)_2CH(CH_3)$ C_2H_5	H	1
Obs005	$N(CH_3)_2$	CH_3	H	H	1
Obs006	OCH_3	CH_3	$CH_2CH(CH_3)_2$	H	1
Obs007	OCH_3	CH_3	$CH_2CH(CH_3)$ C_2H_5	H	1
Obs008	OCH_3	CH_3	$CH_2CH_2CH(CH_3)$ C_2H_5	H	1
Obs009	OC_2H_5	CH_3	$CH(CH_3)$ C_2H_5	H	1
Obs010	OC_2H_5	CH_3	$CH_2CH(CH_3)_2$	H	1
Obs011	OC_2H_4	CH_3	$CH_2CH(CH_3)$ C_2H_5	H	1
Obs012	OC_2H_5	CH_3	$(CH_2)_2CH(CH_3)$ C_2H_5	H	1
Obs013	OC_2H_5	CH_3	$CH_2CH(CH_3)_2$	H	1
Obs014	OC_2H_5	CH_3	$(CH_2)_2CH(CH_3)$ C_2H_5	H	1
Obs015	SCH_3	CH_3	$CH_2CH(CH_3)_2$	H	1
Obs016	SCH_3	CH_3	$CH_2CH(CH_3)$ C_3H_7	H	1
Obs017	SC_2H_5	CH_3	$CH_2CH(CH_3)$ C_2H_5	H	1
Obs018	OCH_3	$COH(CH_3)_2$	CH_3	H	1
Obs019	OCH_3	$COH(CH_3)_2$	H	CH_3	1
Obs020	OCH_3	$COCH_3$	H	CH_3	1
Obs021	OCH_3	$COCH_3$	OCH_3	CH_3	1
Obs022	H	C_2H_5	H	CH_3	1
Obs023	C_2H_5	C_2H_5	H	H	1
Obs024	H	$CH(CH_3)_2$	CH_3	CH_3	1
Obs025	H	C_4H_9	H	H	1
Obs026	H	$CH_2CH(CH_3)_2$	H	H	1
Obs027	SCH_3	$CH_2CH(CH_3)_2$	H	H	1
Obs028	H	C_5H_{11}	H	H	1
Obs029	H	C_5H_{11}	CH_3	CH_3	1
Obs030	OCH_3	C_5H_11	H	H	1
Obs031	CH_3	$(CH_2)_2CH(CH_3)_2$	CH_3	H	1
Obs032	OCH_3	C_7H_{15}	H	H	1
Obs033	CH_3	H	CH_3	H	2
Obs034	CH_3	H	H	CH_3	2
Obs035	C_2H_5	CH_3	H	H	2
Obs036	OCH_3	H	H	H	2
Obs037	OCH_3	H	H	CH_3	2
Obs038	OC_2H_5	H	H	H	2

Obs039	SCH_3	H	H	H	2
Obs040	SCH_3	H	H	CH_3	2
Obs041	SC_2H_5	H	H	H	2
Obs042	CH_3	CH_3	H	H	2
Obs043	CH_3	CH_3	CH_3	H	2
Obs044	CH_3	CH_3	CH_3	CH_3	2
Obs045	$NHCH_3$	CH_3	H	H	2
Obs046	OCH_3	CH_3	H	H	2
Obs047	OCH_3	CH_3	H	CH_3	2
Obs048	OC_2H_5	CH_3	H	H	2
Obs049	SCH_3	CH_3	H	H	2
Obs050	SC_2H_5	CH_3	H	H	2
Obs051	H	C_2H_5	H	H	2
Obs052	H	C_2H_5	CH_3	CH_3	2
Obs053	CH_3	C_2H_5	H	CH_3	2
Obs054	CH_3	C_2H_5	CH_3	H	2
Obs055	SC_2H_5	C_2H_5	H	H	2
Obs056	H	$CH_2CH(CH_3)_2$	CH_3	CH_3	2
Obs057	SC_6H_5	C_8H_{17}	H	H	2
Obs058	CH_3	H	H	H	3
Obs059	CH_3	C_3H_7	H	H	3
Obs060	OCH_3	C_3H_7	H	H	3
Obs061	SCH_3	C_3H_7	H	H	3
Obs062	CH_3	$CH(CH_3)_2$	H	H	3
Obs063	OCH_3	$CH(CH_3)_2$	H	H	3
Obs064	OCH_3	$CH(CH_3)_2$	H	CH_3	3
Obs065	OCH_3	$CH(CH_3)_2$	CH_3	H	3
Obs066	OCH_3	$CH(CH_3)_2$	OCH_3	CH_3	3
Obs067	OCH_3	$CH(CH_3)_2$	CH_3	OCH_3	3
Obs068	OCH_3	$CH(CH_3)_2$	OCH_3	$CH(CH_3)_2$	3
Obs069	SCH_3	$CH(CH_3)_2$	H	H	3
Obs070	OCH_3	C_4H_9	H	H	3
Obs071	SC_2H_5	C_4H_9	H	H	3
Obs072	CH_3	$CH_2CH(CH_3)_2$	H	H	3
Obs073	OCH_3	$CH_{2C}H(CH_3)_2$	H	H	3
Obs074	OCH_3	$CH_2CH(CH_3)_2$	H	CH_3	3
Obs075	OCH_3	$CH_2CH(CH_3)_2$	CH_3	H	3
Obs076	OCH_3	$CH_2CH(CH_3)_2$	CH_3	CH_3	3
Obs077	OCH_3	$CH(CH_3)\ C_2H_5$	H	H	3

Obs078	OC_2H_5	C_5H_{11}	H	H	3
Obs079	SCH_3	C_5H_{11}	H	H	3
Obs080	SC_2H_5	C_5H_{11}	H	H	3
Obs081	OCH_3	$(CH_2)_2CH(CH_3)_2$	H	H	3
Obs082	OCH_3	$CH_2CH(CH_3)\ C_2H_5$	H	H	3
Obs083	OCH_3	$(CH_2)_3CH=CH_2$	H	H	3
Obs084	OCH_3	$(CH_2)_2CH=CHCH_3$(E)	H	H	3
Obs085	OCH_3	$(CH_2)_2CH=CHCH_3$(Z)	H	H	3
Obs086	OCH_3	C_6H_{13}	H	H	3
Obs087	OCH_3	$(CH_2)_3CH(CH_3)_2$	H	H	3
Obs088	OCH_3	$CH_2CH(CH_3)\ C_3H_7$	H	H	3
Obs089	OCH_3	C_8H_{17}	H	H	3
Obs090	OC_2H_5	C_8H_{17}	H	H	3
Obs091	SCH_3	C_8H_{17}	H	H	3
Obs092	SC_2H_5	C_8H_{17}	H	H	3
Obs093	OCH_3	$C_{10}H_{21}$	H	H	3
Obs094	OC_2H_5	$C_{10}H_{21}$	H	H	3
Obs095	OCH_3	CH_3	OCH_3	CH_3	3
Obs096	OCH_3	C_2H_5	H	H	3
Obs097	OCH_3	$CH(CH_3)\ C_3H_7$	H	H	3
Obs098	OCH_3	$(CH_2)_6CH(CH_3)_2$	H	H	3
Obs099	OCH_3	$CH_2CH(CH_3)\ C_6H_{13}$	H	H	3
Obs100	OCH_3	$CH_2CH(CH_3)_2$	H	$CH_2CH(CH_3)_2$	3
Obs101	OC_2H_5	$CH_2CH(CH_3)_2$	H	H	3
Obs102	C_2H_5	C_2H_5	C_2H_5	H	4
Obs103	$CH=CH_2$	CH_3	H	H	4
Obs104	CH_3	$CH=CH_2$	CH_3	H	4
Obs105	$CH=CH_2$	CH_3	CH_3	H	4
Obs106	C_3H_7	CH_3	CH_3	H	4
Obs107	$CH=CHCH_3$ (E)	CH_3	CH_3	H	4
Obs108	$CH=CHCH_3$ (Z)	CH_3	CH_3	H	4
Obs109	CH_2-$CH=CH_2$	CH_3	CH_3	H	4
Obs110	$CH(CH_3)_2$	CH_3	CH_3	H	4
Obs111	C_4H_9	CH_3	CH_3	H	4
Obs112	C_2H_5	C_2H_5	CH_3	H	4
Obs113	$CH=CH_2$	C_2H_5	CH_3	H	4
Obs114	C_2H_5	$CH=CH_2$	CH_3	H	4
Obs115	CH_3	C_2H_5	H	H	4
Obs116	C_5H_{11}	CH_3	CH_3	H	5

Obs117	CH_3	H	$CH(CH_3)_2$	H	5
Obs118	CH_3	H	H	$CH(CH_3)_2$	5
Obs119	C_2H_5	$CH(CH_3)_2$	H	H	5
Obs120	C_2H_5	H	$CH(CH_3)_2$	H	5
Obs121	C_2H_5	H	H	$CH(CH_3)_2$	5
Obs122	C_2H_5	H	C_2H_5	H	6
Obs123	C_2H_5	H	H	C_2H_5	6
Obs124	CH_3	H	C_3H_7	H	6
Obs125	CH_3	H	H	C_3H_7	6

Tableaux N°02 : Présentation de la base des données de la famille pyrazine

II.3. Calcul des descripteurs des molécules pyrazine

Dans cette partie nous présentons les structures des molécules de la famille pyrazines qui se trouvent dans la base des données qui forment la structure dans l'espace des cette molécules, nous utilisons dans cette présentations (la structure dans l'espace) le logiciel Hyperchem.

II.3.1. Logiciel Hyperchem 07 : [27]

Hyperchem version 7 est un logiciel de modélisation moléculaire, ce logiciel offre la possibilité de plusieurs types de calculs de modélisation moléculaire, et a été conçu afin de permettre des extensions importantes. Nous utilisons le logiciel Hyperchem 7 pour présenter la forme des structures des molécules pyrazine dans l'espace (vide ou trois dimensions).

II.3.2. Conception générale du logiciel : [28]

Hyperchem, regroupe des icônes usuels que l'on retrouve dans tous les logiciels tels que : File et Edit. De plus ce logiciel contient des icônes spécifiques, citons par exemple : Display, Select et compte, chaque icône à pour objectif bien défini. Ainsi qu'on retrouve des icônes qui nous permettent de crée, de sélectionner, ou de tourner les molécules.

Les composantes principales sont comme indiquées ci-dessous :

II.3.2.1. Base de données [28]

La partie la plus importante du système et sans doute la base de données, qui sert de référence pour plusieurs corps chimiques.

En effet, on peut employer les commandes de bases de données pour créer des polypeptides et des polynucleotides d'une bibliothèque des acides aminés (20 acides aminés existent dans cette base standard) et des acides nucléiques. La base de données offre la possibilité de relier ces molécules, afin de préparer des macromolécules [58].

Soulignons, quand peut également lire dedans, des structures dans les dossiers des bases de données standards, comme la banque de données de protéine de Brookhaven et celle cristallographique de Cambridge.

La base de données cristallographique Cambridge (C.S.D.S : Combridge Structural Database System) mise en place et maintenue par le C.C.D.C (Combridge Cristallographic Data Center) est la plus importante des bases de données concernant les structures cristallines moléculaires en trois dimensions, obtenus par la technique de diffraction des rayons X. Son avantage principal est sa mise à disposition sous une forme programme permettant la détermination des données géométriques associées aux structures cristallines.

II.3.2.2. Système d'unité [28]

Il est évident que les unités font partie du programme de ce logiciel. En effet, le système d'unité employée est un système standard, a titre d'exemple, quelques grandeurs de ces unités prédéfinies dans ce logiciel sont citées dans ce qui suit :

- La température T en K ;
- L'énergie E en kcal / mol ;
- Le temps t en ps (pico-seconde) ;
- La distance en A° ;

II.3.2.3. Système de calcul [28]

Hyperchem est capable de simuler n'importe quel module (poly) moléculaire (respectant les contraintes de logiciel et la capacité de l'ordinateur). Bien que, le choix d'un type de calcul adapté au système étudié est très important, car il conditionne en grande partie la fiabilité des calculs effectués. Les différentes méthodes de calculs prédéfinis dans ce logiciel sont les suivantes :

- **Méthodes quantiques pures**

  HyperChem renferme un certain nombre des méthodes quantique par voie ab initial à savoir : STO-3G, 3-21G, 6-31G* et 6-31G**,, et dont les options : charges, UHF / RHF et MP2. En plus il comporte les méthodes dites DFT

- **Méthodes semi-empiriques**

  Le logiciel dont on dispose, incluse un certain nombre des méthodes de calcul quantique de type semi-empirique, qui sont : La méthode de Hückel étendue, et les méthodes AM1, PM3, INDO, CNDO, MNDO, MINDO/3, MNDO/d, TNDO, ZINDO/1 et ZINDO /s.

- **Méthodes empiriques**

  Mécanique moléculaire et dynamique moléculaire

II.3.2.4. Système d'affichage des résultats [28]

Le logiciel dont on dispose, est doté d'un système automatique d'affichage des résultats, ces derniers sont listés au bas gauche de l'écran, à la fin de l'exécution de l'opération de simulation précédemment choisi.

Il est à noté que, dans le cas de calcul par le dynamique moléculaire ou le dynamique de Langevin, on plus des résultats de bas gauche de l'écran, un graphique spécial de résultats va s'ouvrir automatiquement, dans ce cas on obtient un spectre représentatif de calcul effectué. Cependant, il y a la commande Start log (sous file), qui sert au stockage de l'information numérique lors de calculs, il faut donc activer cette fonction avant la soumission d'un travail, et en donne le même nom de fichier, il se crée un fichier.log que l'on peut examiner avec un traitement de texte.

II.3.3. Etapes de modélisation [27,28]

Pour la modélisation d'une structure moléculaire par le logiciel hyperChem, il est impératif de suivre des étapes bien définis, tous dépend de divers facteurs, notamment : la nature chimique de molécule ou plus précisément des atomes, les propriétés à calculées, le type de calcul à utiliser et les options choisies.

De manière générale, il est préférable de suivre les étapes indiquées dans ce qui suit :

1-Acitver la commande Start log (menu file) avant la soumission d'un calcul, sous le même nom de fichier ;
2-Dessiner la molécule, en choisissant les atomes d'après la commande Defult Élément (menu build) ou crée la molécule d'après la base de données (Database) ;
3-Choisissez la méthode de calcul à utiliser ;
4-Choisissez les options ;
5-Introduir les conditions si c'est nécessaire ;
6-Enregistrement des données précédentes ;
7-Lancement de la modélisation ;
8-Impression des résultats.

II.3.4. Utilisation logicielle pour la base des données pyrazine [27,28]

Après cette utilisation de logiciel Hyperchem 07 pour la base des données précédant (dessiner tous les molécules (125 molécules de famille pyrazine) dans l'espace (trois dimensions)), on utilise ce logiciel pour dessiner et pour optimiser chaque molécule pour construire molécule plus stable dans l'espace.

Enfin on obtient des molécules (la base des données précédant de famille pyrazine) qui possèdent la plus stabilité dans l'espace et engestrie cette dernière se forme Hin (text.hin) pour l'utiliser dans logiciel DRAGON qui calcule les descripteurs des molécules de bases des données.

II.4. Descripteurs [29]

Un descripteur est un mot ou un groupe de mots choisis pour caractériser les informations contenues dans un document et pour faciliter les recherches documentaires. Le descripteur d'un fichier est un enregistrement indiquant la méthode de stockage du fichier et/ou la structure de son contenu. Le descripteur peut aussi être son sélecteur, aussi appelé handle.

Au sens le plus général, mot caractérisant l'information contenue dans un élément afin d'en faciliter l'utilisation ; le sélecteur proprement dit concerne la mémoire vive ; plus généralement, un descripteur est un type de pointeur contenant, en interne, davantage d'informations sur l'objet auquel il fait référence.

II.4.1. Logiciel utilisé pour calculer les descripteurs

Pour calculer les discripteurs des molécules du base des données nous utilisons le logiciel DRAGON qui donnant 1664 paramètres ou discripteurs distribuait a 20 block, chaque block présenté caractère spécifique (caractères chimiques ; caractères physiques ; …ect) nous représentant et définie dans le titre suivant.

II.4.2. Logiciel DRAGON [29]

Le DRAGON est originairement une application pour le calcul de descripteurs moléculaires développé par le Milano Chemisettes et Groupe de recherche QSAR. Ces descripteurs peuvent être utilisés pour évaluer moléculaire structure activité ou rapports de la structure - propriété, aussi bien que pour analyse de la ressemblance et haut débit qui masque de bases de données de la molécule.

La première parution de dates du DRAGON en arrière à 1997. Mises à jour et éléments à inclure de nouveaux descripteurs moléculaires est fait régulièrement pour avancer recherche dans QSAR.

II.4.3. Descripteurs moléculaires

Le DRAGON fournit 1664 descripteurs moléculaires qui sont divisés en 20 blocs logiques (13 blocks pour deux dimension 2D et 7 blocks pour trois dimension 3D) [30]:

N°	les blocks	Nombres des Discripteurs
01	descripteurs Constitutionnels	48
02	descripteurs Topologiques	119
03	Walk et le chemin compte	47
04	indice des Connectivités	33
05	indice des Informations	47
06	2D Autocorrélations	96
07	Indices d'Edge pour la contiguïté	107
08	descripteurs Burden de l'eigenvalue	64
09	indices de la charge Topologiques	21
10	indices Eigenvalue-Basés	44
11	Randic profils moléculaires	41
12	descripteurs Géométriques	74
13	descripteurs RDF	150
14	descripteurs de 3D morses	160
15	descripteurs de CAPRICE	99
16	descripteurs de FUITE	197
17	le groupe de compte fonctionnel	154
18	fragments Atome - Centrés	120
19	descripteurs des Charges	14
20	propriétés Moléculaires	29
	Somme totale de discripteurs	1664

Tableaux N°03 : Présentation les blocks de descripteurs

II.4.4. Fichiers d'entrée [29]

Courir le DRAGON les besoins de l'utilisateur la structure moléculaire classe précédemment obtenu par autre spécifique logiciel de la modélisation moléculaire. Les formats de fichier moléculaires les plus communs sont acceptés sont :

1. Sybyl © MOL2 classe (.mol.ml2, mol2) par Tripos, Inc.
2. Sybyl © Molfiles (.sm2) comme fourni par ChemOffice, Corp CambridgeSoft.
3. Sybyl © Molfiles multiple (.mol. ml2) par Tripos, Inc.
4. Molfiles (.mol) par Dessin Moléculaire Ltd. (MDL).
5. SD multiple classe (.sdf) par Dessin Moléculaire Ltd. (MDL).
6. HyperChem © classe (.hin) par Hypercube, Inc.

7. Notations des SOURIRES (.smi).

8. MacroModel © classe (.dat. ehors) par Schrodinger.

Pour faire usage plein de calculs du DRAGON, 3D structures optimisées avec les hydrogènes devraient être utilisées.

Cependant, le DRAGON peut traiter aussi de molécules H - Épuisées et 2D structures; dans ce cas, c'est apparent que quelques restrictions à calcul du descripteur sont appliquées.

II.4.5. Fichiers de sortie [29]

Le fichier de sortie du DRAGON standard a été organisé afin qu'en premier il y ait deux colonnes fixes, un contenir le nombre séquentiel de molécules (No.) et l'autre l'identificateur de la molécule (MolID) choisi par l'utilisateur dans le 'dossier Choisi' forme.

Alors, les variables de la ficelle, au plus 3, chargés par finalement, l'utilisateur par le menu de réponses de la Charge, suit. Finalement, les descripteurs moléculaires sélectionnés, leurs PCs (si vérifié) et variables programmées par l'utilisateur numériques (si a chargé et a sélectionné) suivez. La possibilité d'ajouter cordez des variables dans le fichier de sortie est très important si information supplémentaire sur les molécules par exemple, CAS comptez, le produit code ou noms spécifiques, a besoin d'être entreposé.

II.4.6. Analyse du descripteur

Les menus du graphique interactifs sont disponibles dans DRAGON pour une analyse du descripteur préliminaire. Histogramme les graphiques, intrigues Pareto et statistiques de l'uni - variante sont disponibles dans le menu consacré à l'analyse de chacun descripteur seul. Intrigues de la ligne, 2D et 3D intrigues de la dispersion sont disponibles dans le nouveau menu 'molécules de la vue dans les descripteur/responsive espacent'; ces graphiques autorisent une analyse préliminaire de distribution de la molécule dans le descripteur ou directeur espace composant, aussi bien qu'une analyse de la corrélation préliminaire quand les réponses programmées par l'utilisateur ont déjà été chargées [30].

II.5. Les descripteurs utilisées pour la méthode « analyses discriminantes » :

Finalement nous trouvons les descripteurs de la base des données (1664 paramètres ou descripteurs distribues sur 20 block) mais les descripteurs sont constants dans tous les molécules (tous les molécules dans la base des données possèdent la même valeur numérique avec même descripteurs) sont éliminé, aussi les descripteurs qui varient avec des autres descripteurs en fonction linière sont éliminé, (choix d’un seul descripteur entre tous les descripteurs qui possèdent la même linéarité).

Après ce traitement sur les descripteurs de base des données, nous trouvons 350 paramètres ou descripteurs distribuait sur 20 blocks.

Nous résumons cette traitant pour les descripteurs dans le tableau suivent :

N°	les blocks	NB Dis avant Le traitant	NB Dis après le traitant
01	descripteurs Constitutionnels	48	11
02	descripteurs Topologiques	119	22
03	Walk et le chemin compte	47	06
04	indice des Connectivités	33	08
05	indice des Informations	47	13
06	2D Autocorrélations	96	25
07	Indices d'Edge pour la contiguïté	107	13
08	descripteurs Burden de l'eigenvalue	64	07
09	indices de la charge Topologiques	21	13
10	indices Eigenvalue-Basés	44	07
11	Randic profils moléculaires	41	02
12	descripteurs Géométriques	74	17
13	descripteurs RDF	150	22
14	descripteurs de 3D morses	160	61
15	descripteurs de CAPRICE	99	17
16	descripteurs de FUITE	197	54
17	le groupe de compte fonctionnel	154	12
18	fragments Atome - Centrés	120	17
19	descripteurs des Charges	14	09
20	propriétés Moléculaires	29	14
	Somme totale de discripteurs	1664	350

Tableaux N°04 : Présentation les descripteurs utilisées pour la méthode « A.F.D »

Enfin, nous présentons cette base des données (125 molécules avec 6 types d'odeurs différentes de la famille pyrazine) avec les descripteurs pour chaque molécules (350 distribuée sur 20 block) sous formes de matrice (20 matrice) dans logiciel d'EXCEL (les ligne présentent les 125 molécules avec type (classification) d'odeur pour chaque molécule (6 classification ou 6 type d'odeur),et les colonnes présentent les 350 descripteurs distribués sur 20 block pour chaque molécule) (les 20 matrices correspondants aux 20 blocks de descripteurs qui sont définis précédemment sont nommes « **matrice(I)** »).

III.1. Introduction

Dans cette partie nous optimisons les meilleurs descripteurs qui se trouvent dans la **matrice (I)** avec la méthode « analyse discriminante » ces descripteurs donnent de meilleurs résultats (meilleurs pourcentages d'estimation, validation et validation croisées qui présentent les meilleurs modèle de classification de la famille pyrazine), pour appliquer cette méthode « analyse discriminante » nous utilisons le logiciel de calcul et statistique nommer **« XLSTAT »** que nous présenterons dans les titres suivants.

III.2. Méthode d'analyse discriminante

III.2.1 Définition [33]

L'analyse discriminante connue dans la pratique marketing comme une des techniques de **"scoring"** essaye de déterminer la contribution des variables qui expliquent l'appartenance des individus à des groupes.

Deux ou plusieurs groupes sont comparés, sur plusieurs variables pour déterminer s'ils différent et pour comprendre la nature de ces différences.
On peut, en marketing, distinguer différents types d'utilisateurs d'un produit :

- utilisateurs permanents et occasionnels d'un produit.
- acheteurs d'une marque et les acheteurs de marques concurrentes.
- clients fidèles et infidèles.
- vendeurs bons, médiocres et mauvais.

L'analyse discriminante étudie des données provenant de groupes connus à priori. Elle vise deux buts principaux [34]:

- **Description** : Parmi les groupes connus, quelles sont les principales différences que L'on peut déterminer à l'aide des variables mesurées.

- **Classement** : Peut- on déterminer le groupe d'appartenance d'une nouvelle Observation uniquement à partir des variables mesurées.

Les domaines d'application de l'analyse discriminante sont nombreux en géologie: définition d'indices de prospection géochimique, analyse d'images, caractérisation géochimique de types de roches, etc. L'analyse discriminante se rattache au champ plus vaste de la reconnaissance des formes. Par ses objectifs, elle s'apparente également aux réseaux neuronaux, sujet très à la mode en recherche informatique.

III.2.3. Aspect théorique de la méthode

III.2.3.1. Aspect descriptif [34 ,35]

Soit un vecteur u_1. On choisira u_1 de telle sorte que les projections des moyennes des groupes sur u_1 soient le plus espacées possible et que, simultanément, les projections des observations d'un même groupe soient le plus rapprochées possible de la projection de la moyenne du groupe. Bref, sur ce vecteur u_1 on cherche à observer des groupes compacts et distants les uns des autres.

La matrice X centrée par rapport aux moyennes calculées avec toutes les observations (sans tenir compte du groupe) est donnée par :

$$Xc = X - 11'X/n \ldots. (\mathbf{III.1})$$

De même, la matrice C centrée (i.e. la matrice contenant les moyennes de chaque groupe centrées par rapport à la moyenne globale) s'écrit :

$$Cc = C - 11'X/n \ldots. (\mathbf{III.2})$$

On pourrait également centrer chaque observation de la matrice X par rapport à la moyenne du groupe correspondant :

$$Xg = X - C \ldots. (\mathbf{III.3})$$

Bien sûr, on a :

$$Xc = Cc + Xg \ldots. (\mathbf{III.4})$$

La matrice de variabilité (totale) s'écrit alors :

$$T = Xc'Xc \ldots. (\textbf{III.5})$$

$$T = Cc'Cc + Xg'Xg \ldots. (\textbf{III.6})$$

Car ;

$$Xg'Cc = 0\ldots. (\textbf{III.7})$$

$$T = E + D\ldots. (\textbf{III.8})$$

Le premier membre de droite représente la matrice de variabilité entre les centres des groupes (E pour "Entre"). Le second membre représente la matrice de variabilité à l'intérieur des groupes (D pour "Dans").

Les groupes seront d'autant plus faciles à discriminer (à séparer) que E sera grand par rapport à D (où à T). En effet, si E est grand, ceci signifie que les centres des groupes sont éloignés. Si D est petit, ceci signifie que les observations d'un même groupe sont proches de leur centre. Si on a simultanément E grand et D petit alors les groupes sont éloignés les uns des autres et compacts, la situation idéale.

III.2.3.1.1. Recherche du vecteur séparant le mieux possible les groupes [35]

Soit un vecteur u sur lequel seront effectués les projections des observations. Effectuons les "projections" sur u : Xcu

La variabilité de ces projections est donnée par : u'Tu

On a :

$$u'Tu = u'Du + u'Eu\ldots. (\textbf{III.9})$$

Le vecteur u recherché est le vecteur qui maximise le rapport :

$$\frac{u'Eu}{u'Du} ou \frac{u'Eu}{u'Tu} \quad \ldots. (\textbf{III.10})$$

Nous choisirons le premier rapport parce qu'il est utilisé plus souvent (les deux sont admissibles et donnent des résultats identiques).

Il est équivalent de maximiser u'Eu/u'Du ou de maximiser u'Eu sujet à u'Du = 1 (en effet soit u le vecteur obtenu en solutionnant directement le rapport ; si u'Du = c ≠ 1 on n'a qu'à poser u* = 1/√c u et on a le même maximum avec la contrainte respectée).

Comme déjà vu en ACP, on a un problème de maximisation sous contrainte qui est résolu par la technique de Lagrange.

On trouve u est solution de :

$$D^{-1}Eu = \lambda u \ldots. (\textbf{III.11})$$

$$u'Du = 1\ldots. (\textbf{III.12})$$

On reconnaît un problème de vecteurs propres et de valeurs propres. Le vecteur recherché est le vecteur propre associé à la plus grande valeur propre de $D^{-1}E$. Les autres vecteurs propres de cette matrice seront successivement les vecteurs, orthogonaux aux précédents (i.e. u'iDuj = 0) donnant la meilleure séparation entre les groupes. On aura, aux plus, k-1 valeurs propres non-nulles car le rang de la matrice E est de k-1 (k groupes centrés). Ainsi deux groupes centrés définissent une droite passant par l'origine (dimension 1), trois groupes définissent un plan (dimension 2), etc.

III.2.3.1.2 Cas particulier de deux groupes [34]

C'est un cas qui se présente très fréquemment et pour lequel la solution est particulièrement simple puisqu'on a alors un seul vecteur discriminant (vecteur propre de $D^{-1}E$). On peut montrer que le vecteur propre est donné par :

$$u = \sqrt{\frac{n_1 n_2}{n\lambda}} D^{-1}(y_1 - y_2) \ldots. (\textbf{III.13})$$

La valeur propre associée est :

$$\lambda = (y_1-y_2)'D^{-1}(y_1-y_2)\ n_1n_2/n=u'Eu \ldots. (\textbf{III.14})$$

Dans le cas de deux groupes, on n'a donc aucune recherche de valeurs propres et vecteurs propres à effectuer.

III.2.3.2. Aspect classement [33,34]

On a de nouvelles observations que l'on veut classer dans un des groupes connus uniquement à partir des valeurs mesurées.

Exemple

- Vous prélevez un certain nombre de roches volcaniques en Abitibi pour lesquelles vous analysez les éléments majeurs. Vous formez deux groupes selon qu'il existe ou non un gisement connu situé à proximité de l'observation. Dans une nouvelle zone d'exploration, vous mesurez les mêmes variables et vous classez l'observation. Si celle-ci est classée dans le groupe "proximal", alors c'est que cette roche présente une signature géochimique plus similaire aux roches rencontrées à proximité des gisements qu'aux roches "distales". Il s'agit donc d'une zone favorable.

- Vous disposez d'images satellites dans plusieurs bandes de fréquences. Vous voulez utiliser cette information pour identifier les types de roches sur l'image. En quelques endroits (pixels), vous connaissez le type de roche pour l'avoir identifié sur le terrain. Vous formez des groupes avec ces pixels connus et vous cherchez à classer les autres pixels de l'image.

Remarque

Le classement est particulièrement indiqué lorsque les groupes sont difficiles à déterminer pour une raison ou une autre (coût, inaccessibilité,...).

Nous traiterons de deux approches différentes ; une approche géométrique et une approche probabiliste (simplifiée).

III.2.3.2.1. Approche géométrique du classement [34]

L'idée de base est très simple. Il s'agit de calculer la distance (définie par D^{-1}) entre la nouvelle observation et le centre de chacun des groupes. On classera la nouvelle observation dans le groupe pour lequel cette distance est minimale.

La distance entre une observation x (p x 1) et un groupe i s'écrit :

$$d^2(x,yi) = (x - y_i)' D^{-1} (x - y_i) \ldots. (\textbf{III.15})$$

Où y_i est le vecteur p x 1 des moyennes des p variables pour le groupe i. Développant le produit on trouve :

$$d^2(x,y_i) = x'D^{-1}x - 2x'D^{-1}y_i + y_i'D^{-1}y_i \ldots. (\textbf{III.16})$$

Le terme $x' D^{-1} x$ ne dépend pas du groupe considéré. On veut classer dans le groupe pour lequel la distance est minimale. On peut tout aussi bien classer dans le groupe pour lequel g_i est maximal avec :

$$g_i =[x'D^{-1}y_i - 1/2\ y'_iD^{-1}y_i] * (n-k) = [x'D^{*-1}y_i - 1/2\ y'_iD^{*-1}y_i] \ldots. (\textbf{III.17})$$

Les gi sont ce que l'on appelle des "fonctions de classification" ou encore des "fonctions linéaires discriminantes". On en possède autant qu'il y a de groupes et on affecte la nouvelle observation au groupe pour lequel sa fonction de classification est maximale. Le facteur (n-k) est introduit pour pouvoir utiliser D* au lieu de D. En effet, D* est la matrice de covariances nécessaires pour pouvoir calculer les probabilités d'appartenance à chaque groupe.

III.2.3.2.2. Cas de deux groupes [35]

On affecte l'observation au groupe 1 si $g_1 > g_2$ ou $g_1-g_2 > 0$

Or $g_1 - g_2$ s'écrit :

$$x'D^{*-1}y_1 - 1/2y_1'D^{*-1}y1 - x'D^{*-1}y_2 + 1/2y_2'D^{*-1}y_2 > 0 \ldots. (\textbf{III.18})$$

Ceci devient :

$$(x' - 1/2(y_1+y_2)')\ D^{*-1}\ (y_1-y_2) > 0 \ldots. (\textbf{III.19})$$

Ou :

$$x'D^{*-1}\ (y_1-y_2) > 1/2(y_1+y_2)'\ D^{*-1}\ (y_1-y_2) \ldots. (\textbf{III.20})$$

Comparant ces résultats au vecteur propre trouvé dans l'approche descriptive, on constate que le résultat du classement s'observe directement sur le premier vecteur propre. L'observation est classée dans le groupe dont le centre se projette du même côté par rapport au point milieu séparant les deux groupes.

III.2.4. Approche probabiliste (simplifiée) [34]

L'idée est de classer une observation dans le groupe pour lequel la probabilité conditionnelle d'appartenir à ce groupe étant données les valeurs observées est maximale. En pratique on ne peut calculer ces probabilités que si les observations proviennent d'une loi multinormale. Si tel n'est pas le cas on devra au préalable transformer les données pour s'en rapprocher le plus possible. (La pratique a toutefois prouvée que l'AD était très robuste face à l'hypothèse de multinormalité).

La fonction de densité multinormale est :

$$f(x) = (2\pi)^{-p/2}\ |\Sigma|^{-1/2} \exp(-1/2(x - y)^1\ \Sigma^{-1}\ (x - y)) \ldots. (\textbf{III.21})$$

Si x provient du groupe i alors sa fonction de densité est estimée par : $N(y_i, D^*_i)$.

De la définition de probabilité conditionnelle, si l'observation appartient nécessairement à un des k groupes, et si l'on suppose qu'à priori chaque groupe a une probabilité égale d'être observé, on a:

$$p(groupei / x) = \frac{f_i(x)}{\sum_{j=1}^{k} f_j(x)} \quad \ldots. (\textbf{III.22})$$

Si l'on suppose de plus que les k groupes ont même matrice de covariances D alors on a:

$$p(groupei/x) = \frac{\exp[-\frac{1}{2}(x-y_i)'D^{*-1}(x-y_i)]}{\sum_{j=1}^{k}\exp[-\frac{1}{2}(x-y_j)'D^{*-1}(x-y_j)]} \quad \text{…. (III.23)}$$

Après quelques manipulations, cette expression peut s'écrire :

$$p(groupei/x) = [\ \sum_{j=1}^{k}\exp(g_j - g_i)\ ]^{-1} \text{…. (III.24)}$$

Où les gi sont les fonctions de classification décrites à la section précédente. Cette probabilité est maximale quand gi est maximale (ou quand la distance d'un point au centre du groupe est minimale).

Conclusion

Les approches géométriques et probabilistes sont strictement équivalentes lorsque l'on a k populations multinormales avec mêmes matrices de covariances.

Remarque

- Dans l'approche probabiliste, on peut inclure des probabilités à priori de rencontrer chaque groupe. Dans le cas de deux groupes, ceci revient d'un point de vue géométrique à déplacer le point milieu de façon à favoriser le groupe ayant la plus grande probabilité à priori d'être rencontrée. Également, on peut inclure des pénalités reliées au mauvais classement d'une observation. Toutefois, tous ces résultats ne sont valides que si l'hypothèse de multinormalité est respectée.

- Lorsqu'on permet que les matrices de variances-covariances Di* varient d'un groupe à l'autre, on se trouve alors à effectuer une discrimination non-linéaire. Les zones attachées à chaque groupe ne sont plus délimitées par des plans (hyperplans) comme c'était le cas précédemment, mais plutôt par des surfaces courbes. On donne le nom de discrimination quadratique à cette approche. Elle est rarement utilisée.
- D'autres variantes existent encore pour l'AD. L'étude de celles-ci dépasse toutefois le cadre de ce cours.

III.2.5. Évaluation de la qualité de l'analyse discriminante [35]

Il existe plusieurs façons de vérifier la qualité d'une analyse discriminante ; certaines font appel à des hypothèses probabilistes, d'autres non. Les résultats présentés dans les sections suivantes le sont principalement pour référence car ces statistiques sont fréquemment utilisées dans les logiciels commerciaux.

III.2.5.1. Pourcentage du bon classement [35]

C'est la statistique la plus utilisée et aussi la plus "parlante" tout en étant la plus simple. L'idée est la suivante : on a une procédure de classement, alors pourquoi ne pas l'appliquer aux observations dont on connaît le véritable groupe et vérifier ainsi si l'on effectue un bon classement.

Exemple :

Groupe AD			
		1	2
Groupe véritable	1	50	10
	2	30	110

Tableaux N°05 : exemples présentés d'évaluation de qualité l'analyse discriminante

Ici on aurait 160/200 = 80% des observations de bien classés. C'est un fort pourcentage si l'on considère qu'un classement fait entièrement de façon aléatoire donnerait en moyenne 50% de bien classés. De plus on note que les observations du groupe 1 sont bien classées dans une proportion de 83% alors que les observations du groupe 2 sont bien classées dans une proportion de 78%. Le groupe 1 est donc légèrement plus homogène que le groupe 2.

Notons que ce pourcentage de bien classés est trop optimiste, surtout lorsque le nombre d'observations est faible. En effet, si l'on forme deux groupes provenant d'une même population et que l'on applique l'analyse discriminante, on devrait trouver un pourcentage légèrement supérieur à 50% car les fonctions de classification s'ajustent aux variations échantillonneuses. Une façon d'obtenir un estimé plus réaliste consiste à mettre de côté une certaine proportion des observations initiales de chaque groupe, de trouver les fonctions de

classification avec les autres observations puis d'effectuer le classement des observations mises de côté (échantillon test). Une autre variante consiste à mettre de côté une observation à la fois et de répéter l'analyse et le classement n fois.

Remarque

Puisque le tableau de classement (appelé aussi matrice de confusion) est une forme de tableau de contingences, on peut tester le caractère significatif du classement à l'aide d'un test d'indépendance du Khi2.

III.2.5.2. Lambda de Wilks [35]

Cette statistique est définie comme étant le rapport des déterminants des matrices D et T.

$$L = |D| / |T| = |T^{-1}D \quad \text{…. (III.25)}$$

$$L = \prod_{i=1}^{p} \gamma_i \quad \text{…. (III.26)}$$

Où γi est une valeur propre de $T^{-1}D$.

La relation suivante relie les valeurs propres λ et γ :

$$\gamma = \frac{1}{\lambda + 1} \quad \text{…. (III.27)}$$

Sous hypothèse de multinormalité et d'égalité des matrices de covariances, on peut montrer que

$$-[n - (p+k)/2 - 1] \text{ Ln } L \quad \text{…. (III.28)}$$

Où :

n est le nombre total d'observations.

p est le nombre de variables.

k est le nombre de groupes.

est approximativement distribuée suivant une loi Khi^2 avec p(k-1) degrés de liberté.

Lorsque l'on a plusieurs groupes (k > 2) et que l'on veut vérifier le caractère significatif des vecteurs propres qui restent après en avoir accepté q, on peut formuler le test suivant :

H_0 : les vecteurs propres q+1, q+2 .... k-1 n'ajoutent rien à la discrimination des k groupes.

H_1 : non H_0.

Alors :

$$-[n - (p+k)/2 - 1] \text{ Ln } L^* \text{ ….} \quad \textbf{(III.29)}$$

Où : L* est donné par :

$$L^* = \prod_{i=q+1}^{k-1} \gamma_i \text{ ….} \quad \textbf{(III.30)}$$

Est approximativement distribué selon une loi Khi^2 avec (p-q) (k-q-1) degrés de liberté.

Un autre test similaire à ce dernier utilise le fait que (n-k) λq est approximativement distribué suivant une loi Khi^2 avec (p+k-2q) degrés de liberté. On vérifie successivement si la $1^{ère}$ (q = 1) valeur propre est significative, puis la 2^{eme} (q = 2), et ainsi de suite.

Remarque

Ces deux derniers tests sont utiles surtout pour des fins de description. Ces résultats ne peuvent pas être incorporés dans l'étape classement.

III.2.5.3. Le "V" de Rao [34]

La statistique V mesure la somme des distances entre les centres des groupes et la moyenne globale. La distance est normalisée par la matrice D^{*-1} (généralisation de la distance de Mahalanobis). Elle est définie comme étant :

$$V = \sum_{i=1}^{k} n_i (y_i - y)' D^{*-1} (y_i - y) \text{ ….} \quad \textbf{(III.31)}$$

Où :

y_i: vecteur moyenne du groupe i (px1).

y : vecteur moyenne totale.

D* : matrice de variance-covariance intra-groupe (i.e. D/ (n-k) ou k est le nombre de groupes

Et D est la matrice des produits croisés intra-groupes).

n_i : est le nombre d'observations dans le groupe i, n est le nombre total d'observations.

On peut démontrer que sous hypothèse de multinormalité et d'égalité des matrices de covariances, V est distribuée suivant une Khi^2 avec p(k-1) degrés de liberté.

Également si on effectue la discrimination avec p variables puis avec p+1 variables, on peut vérifier le caractère significatif de l'ajout de la variable. En effet, le changement de V (i.e. V_{fin}-V_{ini}) est alors distribué suivant une Khi^2 avec (k-1) degrés de liberté.

III.2.5.4. Corrélation canonique ou pouvoir discriminant d'un vecteur propre [33,35]

Soit le rapport :

$$\alpha = \frac{u'Eu}{u'Tu} \quad \text{…. (III.32)}$$

Par un développement similaire à ce qui a été vu précédemment, on montre que α est valeur propre de $T^{-1}E$. Cette valeur propre est reliée aux valeurs propres λ de $D^{-1}E$ par:

.

$$\alpha = \frac{\lambda}{\lambda+1} \quad \text{…. (III.33)}$$

Ce rapport α exprime la proportion de la variabilité totale imputable aux différences entre les centres des groupes. Cette quantité est donc analogue au R^2 en régression. Pour cette raison, on définit $\alpha^{1/2}$ comme le coefficient de corrélation canonique ou pouvoir discriminant.

Remarque

Le nom corrélation canonique fait référence à une méthode appelée analyse canonique. Cette méthode étudie deux ensembles de variables mesurées sur un même ensemble d'observations. Elle cherche les combinaisons linéaires des deux ensembles de variables qui seront le plus corrélées entre elles. En AD, les deux ensembles de variables sont d'une part les p variables mesurés et d'autre part, les (k-1) variables indicatrices permettant d'identifier les groupes. La corrélation maximale que l'on peut obtenir entre ces deux ensembles de variables est précisément $\alpha^{1/2}$

III.2.5.5. Test d'égalité des matrices de covariances intra-groupes [34,35]

Le calcul des probabilités ainsi que les différents tests présentés précédemment pour le V de Rao et le Lambda de Wilks nécessitent la multinormalité des observations et l'égalité des matrices de covariances à l'intérieur de chaque groupe. On peut tester cette dernière hypothèse par le test approximatif suivant (test de Kullback1 (1959)) nécessitant aussi la multinormalité des observations :

$$x^2 = \sum_{i=1}^{k} \frac{n_i - 1}{2} Ln \frac{\left|D^*\right|}{\left|D_i^*\right|} \quad \text{.... (III.34)}$$

est approximativement distribué suivant une loi Khi^2 avec (k-1)*n*(n+1)/2 d.l.

On rejette l'hypothèse d'égalité des matrices de variance-covariance lorsque la statistique excède le seuil lu dans une table Khi^2.

D* est la matrice de variance-covariance intra-groupes

D_i* est la matrice de variance-covariance pour le groupe i

n_i est le nombre d'observations dans le groupe i, n est le nombre total d'observations

| | Signifie le déterminant

III.2.5.6. Procédures de sélection des variables [35]

On est souvent intéressé à obtenir la meilleure discrimination possible avec le minimum de variables, possiblement pour des raisons d'interprétation, de robustesse des résultats, de fiabilité, sûrement pour des raisons économiques. En effet avec des analyses géochimiques, par exemple, si on obtient une aussi bonne (et parfois meilleure) discrimination avec trois variables qu'avec huit, on vient d'économiser un coût considérable.

Les mêmes procédures vues en régression peuvent être utilisées ici, i.e. sélection avant, élimination arrière et "stepwise". La section précédente a été consacrée à la définition de statistiques qui peuvent toutes servir de critère d'inclusion ou d'élimination. D'autres critères sont présentés dans certains programmes d'analyse discriminante (ex. SPSS, XLSTAT).

Certains de ces critères permettent de vérifier si l'ajout d'une variable supplémentaire est significatif (ex. V de Rao) d'autres ne le permettent pas (Lambda de Wilks, pourcentage de bien classés, corrélation canonique).

Malgré la diversité des méthodes, la pratique montre que le sous-ensemble de variables retenues est relativement robuste au choix du critère d'inclusion. De plus, même si deux sous-ensembles diffèrent quant aux variables retenues, très souvent l'interprétation est identique et les performances (classement) très comparables. Tout ceci est finalement rassurant pour l'utilisateur.

III.2.5.7. Remarques et résumé

L'analyse discriminante peut être vue comme un cas spécial d'analyse factorielle. Mais le but diffère : il s'agit de faire ressortir au maximum les différences entre des groupes mesurés dans un espace multidimensionnel, en projetant chaque cas dans l'espace unidimensionnel d'un petit nombre de fonctions linéaires orthogonales.

Cette opération fait suite habituellement à celle de l'analyse de variance multivariée où, en présence d'une situation où plusieurs groupes sont mesurés sur plusieurs variables, on s'intéresse d'abord à déterminer s'il y a différence significative entre les groupes. Dans le cas de résultats positifs, il devient intéressant de déterminer, parmi les variables, celles qui sont responsables dans un ordre décroissant d'importance des différences entre les groupes: c'est le but de l'analyse discriminante.

Une exploitation plus poussée des résultats conduit à leur utilisation dans le but de classifier (en se donnant comme objectif une probabilité minimum d'erreurs) des nouveaux sujets dans les divers groupes.

Le rôle de l'analyse discriminante peut être envisagé de deux façons quant à l'attribution des qualificatifs d'indépendance et de dépendance, aux variables mesurées sur les populations visées et aux fonctions discriminantes. En sciences d'exploration, en général, les populations sont considérées comme variables indépendantes (predictors) et les fonctions discriminantes comme variables dépendantes (critères). En sciences expérimentales, ces rôles se trouvent renversés.

L'analyse discriminante consiste donc à projeter dans un sous-espace approprié des échantillons de mesures multidimensionnelles. L'interprétation de cette opération peut être faite en termes soit du nombre et de l'importance relative des fonctions discriminantes retenues, soit de la localisation dans l'espace discriminant des populations étudiées.

Ses applications en marketing sont multiples, certains l'ont proposé comme une méthode de positionnement perceptuel, elle a été largement utilisée en marketing pour score des fichiers même si aujourd'hui elle doive laisser le pas à la régression logistique. Elle peut aussi avantageusement servir à dresser les profils d'une typologie.

III.3. Logiciel utilisé

Dans cette partie nous utilisons logiciel de calcul et statistique nommé **«XLSTAT** » pour permettre l'utilisation de la méthode « analyse discriminante » qui nous présenterons dans le titre suivant.

III.3.1. Présentation du logiciel « XLSTAT » [36]

XLSTAT est une addition pour le programme Microsoft Excel. Il vous permet de calculer une large gamme d'analyses statistiques des données contenues dans un tableur. Les résultats et les graphiques sont insérés dans le tableur pour un traitement ou une impression ultérieures. Cela vous permet d'appliquer des techniques multivariables telles que l'analyse discriminante, l'analyse (factorielle) de correspondance, l'analyse par la théorie des grappes, ainsi que toute une série de techniques de régression, tests de qualité de l'ajustement et tri tabulaire. Afin de vous familiariser avec l'utilisation d'**XLSTAT**, des tutoriels sont à votre disposition. **XLSTAT** offre de très nombreuses fonctionnalités qui font d'Excel un outil performant et facile d'accès pour répondre à la majorité de vos besoins en analyse de données et modélisation.

III.3.1.1. Analyse statistique pour Excel [37]

- **XLSTAT** est l'outil d'analyse de données et de statistiques pour Microsoft Excel le plus complet et le plus utilisé.
- **XLSTAT** offre de très nombreuses fonctionnalités qui font d'Excel un outil performant et facile d'accès pour répondre à la majorité de vos besoins en analyse de données et modélisation. Toutes les fonctions **XLSTAT** sont accessibles à partir d'une icône qui est ajoutée à la barre de menus d'Excel. L'utilisation d'Excel comme interface rend le produit très convivial, simple d'utilisation et efficace.
- La qualité des calculs est quant à elle identique à celle des logiciels scientifiques les plus renommés, et **XLSTAT** couvre l'essentiel des besoins du statisticien qu'il soit expert ou débutant. Des modules optionnels répondent à des besoins plus spécifiques (séries chronologiques, analyse de survie, contrôle statistique des procédés, analyse des effets de doses en chimie et pharmacologie...). Cet ajout au logiciel de Microsoft permet aux utilisateurs de réaliser de l'analyse de données et de la modélisation.

III.3.1.2. Utilisation de la méthode « analyse discriminante » par *XL*STAT [36,37]

L'analyse factorielle discriminante **(AFD)** est une méthode permettant de modéliser l'appartenance à un groupe d'individus en fonction des valeurs prises par plusieurs variables, puis de déterminer le groupe le plus probable pour un individu, connaissant uniquement les valeurs des variables qui le caractérisent. Dans **XLSTAT**, les variables qui décrivent les individus sont forcément des variables quantitatives, les groupes étant spécifiés par une variable qualitative. L'**AFD** peut être considérée comme une extension de la régression multiple dans le cas où la variable à expliquer est une variable qualitative décrivant des groupes.

Utilise l'analyse factorielle discriminante pour classer de nouveaux individus décrits par plusieurs variables quantitatives, connaissant un échantillon d'individus décrits par les mêmes variables, dont les groupes sont connus, et pour analyser la façon dont les variables descriptives contribue à la constitution des différents groupes.

III.3.2. Présentation La matrice (I) dans logiciel XLSTAT

Les 20 blocks de descripteurs qui présenté les 20 matrices se forme EXCEL, et chaque matrice ou block contenue une base des données 125 molécules avec vous classification (types d'odeur) coréspendant qui présente dans les colonnes de cette matrice, et les lignes de même matrice présente des discripteurs pour chaque molécules qui spécifié cette block.

Dans cette opération les 20 blocks de **la matrice (I)** diviser sur 3 types de dimensionnement sont représentés comme suit :

- **Les blocks 2D :** représentent les descripteurs qui correspondent aux caractères du la molécule pyrazine dans deux dimensions (globale 13 blocks).
- **Les blocks 3D :** représentent les descripteurs qui correspondent aux caractères de la molécule pyrazine dans trois dimensions (globale 07 blocks).
- **Les blocks ALL :** représentent les descripteurs qui correspondent aux caractères de la molécule pyrazine dans deux dimensions et trois dimensions (globale 20 blocks).

Nous utilisons **la matrice (I)** dans le logiciel XLSTAT qui nous permet de présenter **la matrice (I)** sous forme d'**EXCEL**, et après ça nous utilisons la méthode « analyse discriminante » dans le logiciel pour calculer la corrélation des descripteurs entre luis pour tous les molécules (chaque bloque ou matrice correspondant et calculer la corrélation seulement et individuellement). L'utilisation de méthode « analyse discriminante » dans le logiciel XLSTAT conditionnent les applications suivantes :

III.3.2.1. Les entrées

Les entrées pour **la matrice (I)** sont les descripteurs des molécules **(**Pour chaque block les entrées sont les descripteurs correspondants de ce block et pour tous les molécules).

III.3.2.2. Les sorties

Les sorties pour **la matrice (I)** sont les classifications des molécules (les six types ou classification des odeurs différentes), (Pour chaque block les sorties sont les six types ou classification des odeurs différentes).

III.3.2.3. Pourcentages des tests d'estimation (trining)

Les pourcentages de test d'estimation est 70% de la base des données (ci t'a dire 87 molécules qui présente dans la partie d'estimation (la base des données sont 125 molécules) pour réalisation la méthode « analyse discriminante »).

III.3.2.4. Pourcentages des tests de validation (test)

Les pourcentages de test de validation est 30% de la base des données (ci t'a dire 38 molécules qui présente dans la partie de validation (la base des données sont 125 molécules) pour réalisation la méthode « analyse discriminante ».

III.3.2.5. Application de la méthode « analyse discriminante »

Après la réalisation des conditions précédentes, on laisse le logiciel XLSTAT calculer la méthode « analyse discriminante » pour obtenir les résultats. Dans cette opération les 20 blocks de **la matrice (I)** diviser sur 3 types de dimensions (2D, 3D et ALL).

III.4. présentations des Résultats

On résumé les résultats de 2D, 3D et ALL dans les tableaux suivants :

Remarque

Nous avons négligé les pourcentages d'estimation et de validations pour des molécules qui possédant la classification ou types odeurs **Pecasy** et **Sweet** (c'est t'à dire que les pourcentages d'estimation et des validations pour des classifications **Pecasy** et **Sweet** non importantes) parce que le nombre des molécules est petit pour cette classification correspondant (10 molécules), alors que ces molécules n'influencent pas sur les pourcentages globale de tests d'estimation et des validations pour chaque bloque et pour **la matrice (I) ,** aussi que les pourcentages d'estimations et des validations de cette classification (**Pecasy** et **Sweet)** pour ces molécules très éloignées à la réalité (ne donne pas la corrélation exacte ou minime corrélation entre les descripteurs pour **la matrice (I)** (ou pour les 20 block).

III.5. Analyses des résultats

Après ces résultats nous remarquons que :

1. Pour les résultats 2D :
    - Les meilleurs blocks dans **la matrice (I)** qui donnent les pourcentages plus élevés (meilleur %) dans les classifications **GREEN, NUTTY, BELL-PEPPER et EARTHY** (types les odeurs) pour les tests d'estimations, validations et validations croisées sont suivants :

GREEN : meilleur block est **2D_B06 (**pourcentage d'estimations est **88.24%**, pourcentage de validations est **72.5%** et pourcentage de validations croisées est **85.88%).**

NUTTY : meilleur block est **2D_B07 (**pourcentage d'estimations est **91.76%**, pourcentage de validations est **85%** et pourcentage de validations croisées est **91.76%).**

BELL-PEPPER : meilleur block est **2D_B02 (**pourcentage d'estimations est **81.18%**, pourcentage de validations est **77.5%** et pourcentage de validations croisées est **81.18%).**

EARTHY : meilleur block est **2D_B02 (**pourcentage d'estimations est **94.12%**, pourcentage de validations est **87.5%** et pourcentage de validations croisées est **94.12%).**

2. Pour les résultats 3D :
    - Les meilleurs blocks dans **la matrice (I)** qui donnent les pourcentages plus élevés (meilleur %) dans les classifications **GREEN, NUTTY, BELL-PEPPER et EARTHY** (types les odeurs) pour les tests d'estimation, validations et validations croisées sont suivants :

GREEN : meilleur block est **3D_DISCRIPTEUR (**pourcentage d'estimations est **95.29%**, pourcentage de validations est **72.5%** et pourcentage de validations croisées est 91.76**%).**

NUTTY : meilleur block est **3D_DISCRIPTEUR (**pourcentage d'estimations est **92.94%**, pourcentage de validations est **82.5%** et pourcentage de validations croisées est **91.76%).**

BELL-PEPPER : meilleur block est **3D_DISCRIPTEUR (**pourcentage d'estimations est **90.59%**, pourcentage de validations est **80%** et pourcentage de validations croisées est **87%).**

EARTHY : meilleur block est **3D_DISCRIPTEUR (**pourcentage d'estimations est **92.94%**, pourcentage de validations est **92.5%** et pourcentage de validations croisées est **89.11%).**

3. Pour les résultats ALL :
    - Les meilleurs blocks dans **la matrice (I)** qui donnent les pourcentages plus élevés (meilleur %) dans les classifications **GREEN, NUTTY, BELL-PEPPER et EARTHY** (types les odeurs) pour les tests d'estimations, validations et validations croisées sont suivants :

GREEN : meilleur block est **ALL_B06 (**pourcentage d'estimations est **85.88%**, pourcentage de validations est **75%** et pourcentage de validations croisées est 81.18**%).**

NUTTY : meilleur block est **ALL _B07 (**pourcentage d'estimations est **90.59%**, pourcentage de validations est **87.5%** et pourcentage de validations croisées est **90.59%).**

BELL-PEPPER : meilleur block est **ALL _B17 (**pourcentage d'estimations est **81.18%**, pourcentage de validations est **82.5%** et pourcentage de validations croisées est **81.18%).**

EARTHY : meilleur block est **ALL _B05 (**pourcentage d'estimations est **92.94%**, pourcentage de validations est **87.5%** et pourcentage de validations croisées est **92.94%).**

- Nous remarquons que les meilleurs résultats sont correspondants les blocks suivants :

  (**2D_B06, 2D_B07 et 2D_B02)** pour **2D**

  (3D_DISCRIPTEURS) pour **3D**

  (ALL_B06, ALL_B07, ALL_B17, ALL_B05) pour **ALL**

III.6. Interprétations des résultats

Après ces résultats, nous remarquons que :

- **Pour la classification GREEN :**
  1. Les molécules sont mal classés dans cette classe sont 12 molécules (note classé GREEN par la méthode « analyse discriminante » et pourtant leur classification réelle est GREEN), ces molécules possèdent le radical R2 (suivre la base des données dans le tableau précédant) qui spécifie de ramification du sous – radicalaires pour le radical R2.

2. les molécules qui possèdent les composés S, N et O dans leurs structures moléculaires sont bien classées dans la classe GREEN, pourtant le classement réel pour quelque molécules note GREEN, mais classées dans le classement GREEN parce que elles possèdent ces caractère dans la formule structurelle (construire des composées S, N et o dans la structure moléculaire) par exemple les molécules 65 et 66 (suivre la base des données dans le tableau précédent).
3. généralement la classification des molécules par la méthode « analyse discriminante » qui correspond à la classe GREEN de mal résultats de classification para pour des autres classifications (NUTTY, BELL-PEPPER et EARTHY) (test de validation est entre (60-88 %)).

- **<u>Pour la classification NUTTY :</u>**
    1. Les molécules sont mal classées dans cette classe sont 02 molécules (note classées NUTTY par la méthode « analyse discriminante » et pourtant leur classification réelle est NUTTY).
    2. les molécules qui possèdent le radical R3 et R4 hydrogénisations (H) sont bien classées dans la classe NUTTY, mêmes les molécules qui ne sont pas classées NUTTY, sont classées dans classe NUTTY parce que elles possèdent ces caractère dans la formule structurelle (R3 et R4 sont H) par exemple les molécules 2, 5, 22,96 dans la base des données (suivre la base des données dans le tableau précédant).
    3. généralement la classification des molécules par la méthode « analyse discriminante » qui correspondent à la classe NUTTY à de bons résultats de classification para pour des autres classifications (GREEN, BELL-PEPPER et EARTHY) (test de validation est entre (85-98 %)).

- **Pour la classification BELL-PEPPER :**
    1. Les molécules sont mauvaise classées dans cette classe sont 04 molécules (note classé BELL-PEPPER par la méthode « analyse discriminante » et pourtant leur classification réelle est BELL-PEPPER).
    2. les molécules qui possèdent les radicales R3 et R4 sont différent (types des radicaux R3 ≠ R4) sont bien classé dans la classe BELL-PEPPER, mêmes les molécules qui ne sont classées BELL-PEPPER, sont classées dans la classe BELL-PEPPER parce que elles possèdent à caractère dans la formule structurelle (R3 ≠ R4) par exemples les molécules 1, 3, 13,14, 18, 21, 47, 3 , 19 , 20, 124 dans la base des données (suivre la base des données dans le tableau précédent).
    3. généralement la classification des molécules par la méthode « analyse discriminante » qui correspond à de la classe BELL-PEPPER a un bon résultat de classification para pour les autres classifications (NUTTY, GREEN et EARTHY) (teste de validation est entre (80-94 %)).

- **Pour la classification EARTHY :**
    1. Les molécules sont mal classées dans cette classe est seule molécule (note classée EARTHY par la méthode « analyse discriminante » et pourtant leur classification réelle est EARTHY).
    2. les molécules qui possèdent deux radicaux sont symétrique dans l'espace (mêmes types des radicales R3 = R4 ,R1 = R2, R1 = R3, R1 = R4, R2 = R3,R2 = R4 par exemple (H,H) ou (CH_3, CH_3)…..) sont bien classées dans la classe EARTHY, mêmes les molécules qui ne sont pas classées EARTHY, sont classées dans la classe EARTHY parce que elles possèdent a caractère dans la formule structurelle (deux radicales sont symétrique) par exemples les molécules 22, 24, 26,28, 40, 42, 43, 44 , 51 , 52, 53,…..ect dans la base des données (suivre la base des données dans le tableau précédent).

3. généralement la classification des molécules par la méthode « analyse discriminante » qui correspond à la classe EARTHY a un bon résultat de classification para pour des autres classifications (NUTTY, GREEN et BELL-PEPPER) (test de validation est entre (65-94 %)).

III.7. Sélection des descripteurs

A partir des résultats de la méthode « analyse discriminante » par logiciel XLSTAT, nous remarquons que chaque résultat d'analyse discriminante pour chaque bloque par logiciel XLSTAT, elle sélectionne les meilleurs descripteurs automatiques qui donnent meilleurs corrélations avec des autres descripteurs c'est à dire les descripteurs des molécules qui donnent des meilleurs résultats des cœfficients des corrélations suivantes :

- R^2 partiel : corrélation canonique
- F : facteur discriminante
- Lambda de Wilks : Lambda de Wilks

III.7.1. Affichage des descripteurs sélectionnés :

Après avoir sélectionné les meilleurs descripteurs correspondants les blocks qui donnent les meilleurs résultats procèdent (pourcentages des estimations, des validations et des validations croisées), ces descripteurs des blocks sélectionnés à partir de meilleurs résultats des cœfficients des corrélations (R^2 partiel, F, Lambda de Wilks) de ces descripteurs de blocks.

- Nous présentons cette sélection des descripteurs qui correspondent aux meilleurs descripteurs qui donnent des meilleurs résultats des blocks précédents :

  (**2D_B06, 2D_B07 et 2D_B02)** pour **2D**
  (3D_DISCRIPTEURS) pour **3D**
  (ALL_B06, ALL _B07, ALL _B17, ALL _B05) pour **ALL**

Comme suit dans les tableaux suivants :

III.7.2. Chois des descripteurs

Ce choix des descripteurs, est basé sur les meilleurs descripteurs qui donnent de meilleurs paramètres ou cœfficients de corrélations (R^2 partiel, F, Lambda de Wilks) pour tous les blocks précédent qui donnent des meilleurs résultats statistiquement.

Cette opération (choix des discripteurs), nous appliquons sur tous les descripteurs correspondant des blocks sélectionnées précédent :

2D_B06, 2D_B07 et 2D_B02) pour **2D**
(3D_DISCRIPTEURS) pour **3D**
(ALL_B06, ALL _B07, ALL _B17, ALL _B05) pour **ALL**

III.8. Sélectionnement de matrice (II)

Après ce choix sélection des descripteurs, nous trouvons 36 discripteurs sont donnés descripteurs plus corrélées avec des autres descripteurs.

Elle est basé sur la validation des cœfficients (paramètres) des corrélations existent dans l'intervalle suivants :

- $0.508 \leq R^2$ partiel ≤ 0.050
- $85.572 \leq F \leq 4.099$
- $0.782 \leq$ Lambda de Wilks ≤ 0.207

Enfin, nous présentons nouveau matrice (même **matrice (I)**) mais elle contient du 36 descripteurs sélectionnés dernière l'opération de choix les descripteurs, nous nommons cette matrice **« matrice (II) »** que nous utilisons dans les deux méthodes des classifications reste (logique floue et réseaux de neurones).

Nous présentons les 36 descripteurs sélectionnées dans les tableaux suivants :

IV.1. Introduction

Dans cette partie nous utilisons les descripteurs qui se trouvent dans la **matrice (II)** avec la méthode « **logique floue** » pour étudier la classification des molécules de famille pyrazine avec six types d'odeurs (Green, Nutty, Bell-pepper, Earthy, Pecasy, Sweet), l'objectif de cette utilisation est d'obtenir un meilleur modèle (réseaux) de classification pour la famille pyrazine qui donne de meilleurs résultats de test d'estimation et validation , ce modèle est utilisé pour confirmer ou trouver la classification de quelques molécules de famille pyrazines qui ne connaissent pas de type de classification.

Pour cette application de la méthode « **logique floue** », nous utilisons le logiciel **« MATLAB 7.1 »** que présenterons dans les titres suivants.

IV.2. Méthode logique floue

IV.2.1. Introduction

La logique floue, les réseaux de neurones et les algorithmes génétiques constituent des approches qui tout compte fait, ne sont pas nouvelles. Leur développement se fait à travers les méthodes par lesquelles l'homme essaye de copier la nature et de reproduire des modes de raisonnement et de comportement qui lui sont propre. Bien que ces approches paraissent "naturelles", et si elles se sont imposées dans des domaines allant de traitement de l'image à la gestion financière. Elles commencent à peine à être utilisées dans les domaines de l'électrotechnique et de l'industrie afin de résoudre les problèmes, d'identification de régulation, de processus d'optimisation, de classification, de détection de défauts, ou de prise de décision.

Le terme d'ensemble flou apparaît pour la première fois en 1965 lorsque le professeur Lotfi A.Zedeh, de l'université de Berkeley aux USA publie un article intitule "ensemble flou" il a réalisé depuis de nombreuse avancées théoriques majores dans le domaine et a été rapidement accompagné par de nombreux chercheurs développant des travaux théoriques.

Parallèlement, certain chercheurs se sont penchés sur la résolution par logique floue des problèmes réputés difficiles ; Ainsi en 1975, le professeur Mamdani a Londres développe une stratégie pour le contrôle des procédés et présente les résultats très encourageants qu'il a obtenus sur la conduite d'un moteur a vapeur.

En 1978 la société danoise F.L Smidth réalise le contrôle d'un four à ciment ; c'est là, la première véritable application industrielle de la logique floue.

IV.2.2. Aspect théorique de méthode

IV.2.2.1. Théorie des ensembles floues

IV.2.2.1.1. Notion de l'appartenance partielle [38,39]

Dans la théorie des ensembles, un élément appartient ou n'appartient pas à un ensemble.

La notion d'ensemble est à l'origine de nombreuses théories mathématiques. Cette notion essentielle ne permet cependant pas de rendre compte des situations pourtant simples et rencontrés fréquemment. Il est facile de définir l'ensemble des pommes ; Par contre il sera plus difficile de définir l'ensemble des pommes mures. En constate bien que la mûrit est progressive. La notion de pommes murs est donc graduelle.

C'est pour prendre en compte de telles situations qu'a été créée la notion d'ensemble floue. La théorie des ensembles flous repose sur la notion d'appartenance partielle ; chaque élément appartient partiellement ou graduellement aux ensembles flous qui ont été définis. Les contours de chaque ensemble flou (Fig.N°07) ne sont pas «nets», mais « flous » ou « graduels ».

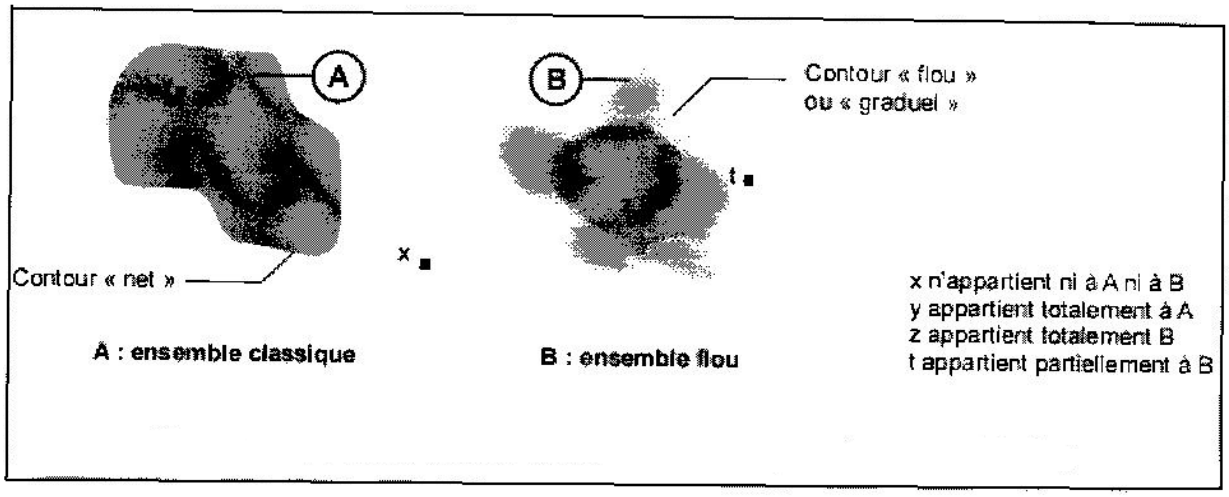


Figure N°06 : Comparaison d'un ensemble classique et un ensemble flou

IV.2.2.1.2. Fonction d'appartenance [38]

Un ensemble flou est défini par ça « fonction d'appartenance », qui correspond à la notion de « fonction caractéristique » ou logique classique. On peut définir le degré d'appartenance de la variable température à f ensemble « faible » comme « le degré de vérité » de la proposition «la température est faible ».

En logique booléenne, le degré d'appartenance (μ) ne peut prendre que deux valeurs (0 ou 1). La température peut être :

- Faible : $\mu_{faible}(T) = 1$, $\mu_{moyenne}(T) = 0$, $\mu_{eleveé}(T) = 0$.
- Moyenne : $\mu_{faible}(T) = 0$, $\mu_{moyenne}(T) = 1$, $\mu_{eleveé}(T) = 0$.
- Elevé : $\mu_{faible}(T) = 0$, $\mu_{moyenne}(T) = 0$, $\mu_{eleveé}(T) = 1$

On peut présenter un ensemble flou par plusieurs types d'univers de discours :

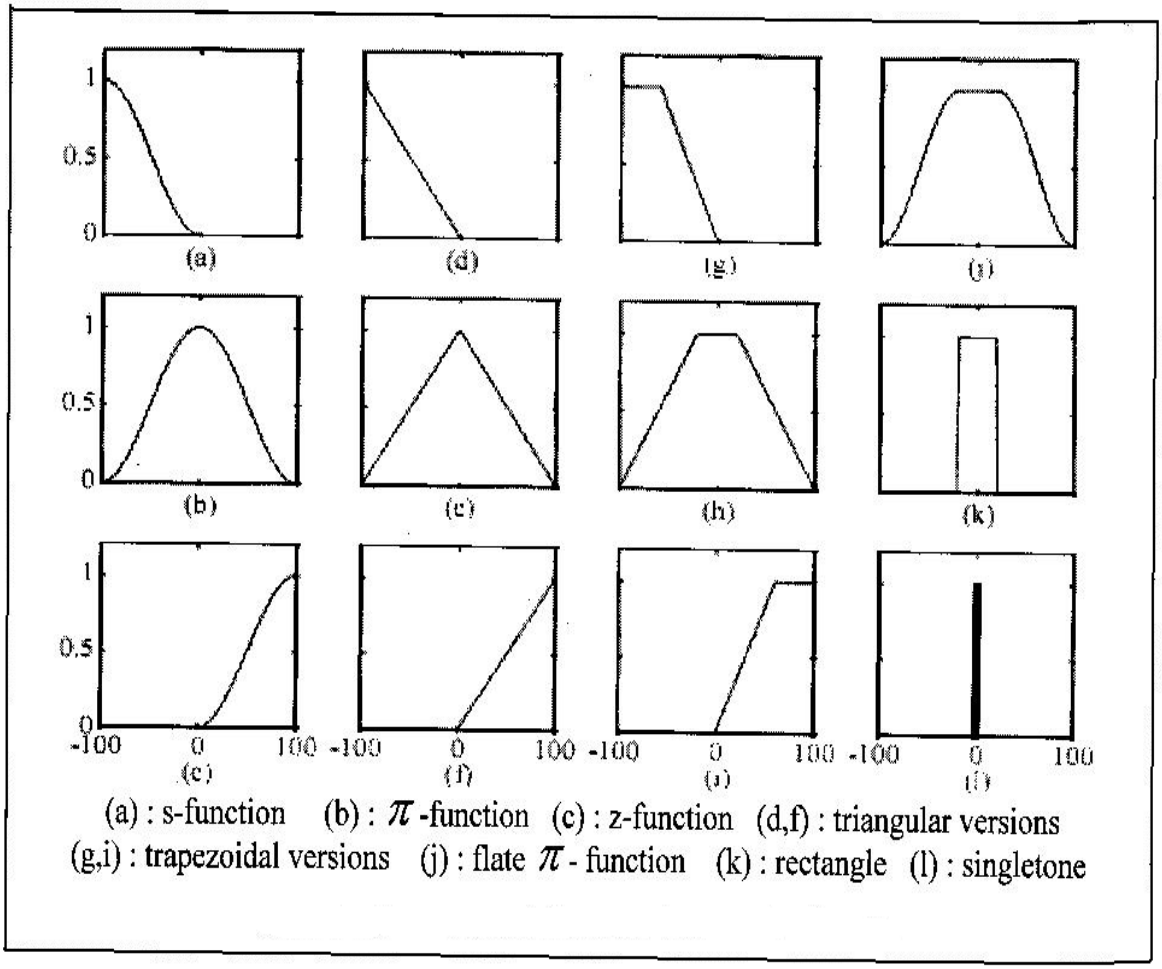


Figure N°07 **:** représentations de l'univers de discours

Dans notre exemple, la variable flou est la température, l'univers de discours est l'ensemble des réelles de l'intervalle [0, 40].On attribue a cette variable trois sous-ensembles lions : Faible, moyenne, et élevée. Chacun est caractérisée par sa fonction degré d'appartenance : $\mu_{faible}(T)$, $\mu_{moyenne}(T)$, $\mu_{eleveé}(T)$

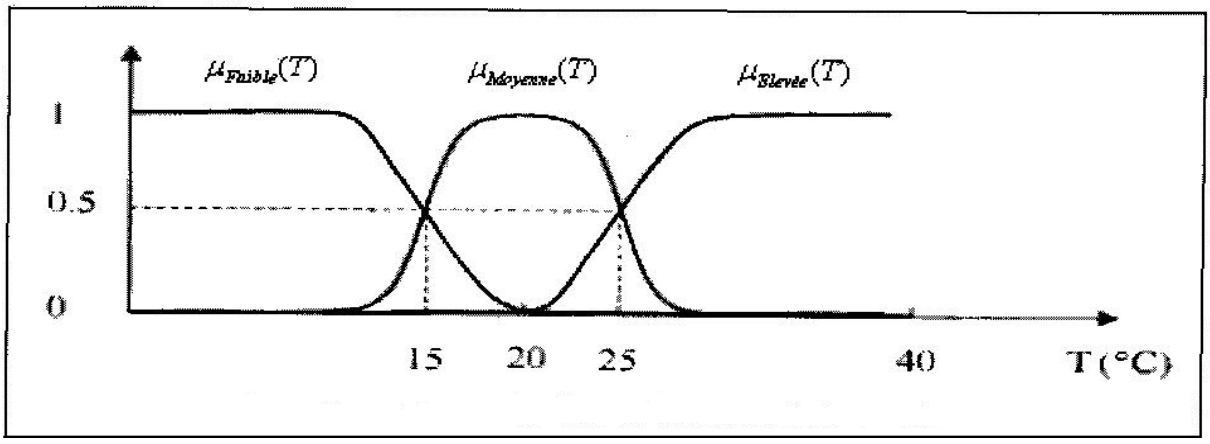


Figure N°08 : ensembles flous de la variable température

IV.2.2.2. Le contrôleur flou [39]

Un contrôleur flou est un système à base de connaissance particulier utilisant un raisonnement en profondeur limite, dans une procédure de chaînage avant des règles (activation des règles par les prémisses) ; Un schéma représentatif peut être le suivant :

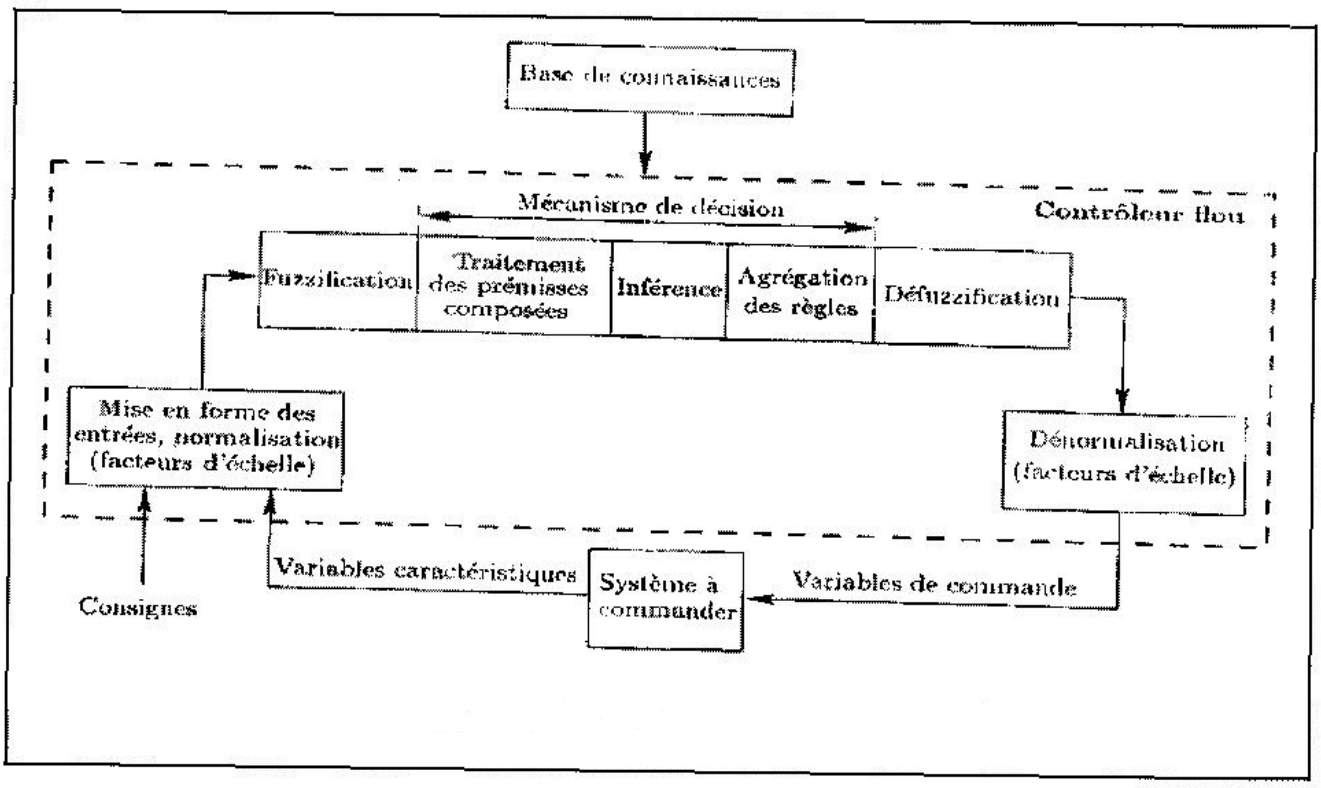


Figure N°09 : Structure de base d'un contrôleur flou

IV.2.2.3. Normalisation [40]

Cette première étape permet le traitement des variables d'entrée du contrôleur flou. Par exemple, calcul d'erreurs (Différence entre grandeurs mesurée et consignes) et variation d'erreurs.

L'utilisation de domaine normalisée (Univers de discours compris entre [-1, l]) nécessite une transformation d'échelle, celle-ci est réalisée par l'intermédiaire de facteurs d'échelle de transformation des grandeurs physiques des entrées en des valeurs normalisées appartenant à l'intervalle [-l, 1].

IV.2.2.4. Fuzzification[38]

C'est l'opération de projection des variables physiques réelles sur des ensembles flous caractérisés par les valeurs linguistiques prises par ces variables. Deux cas peuvent se présenté selon que la mesure d'une variable physique réelle est précise (valeur numérique) ou pas.

Le choix de la forme de fonction d'appartenance (triangulaires, trapézoïdales,...) est arbitraire. Quant au nombre de fonction d'appartenance, il est généralement impaire car elles se répartissent autour de zéro (3, 5 et 7 sont des valeurs courant). Un exemple de fonctions d'appartenance triangulaires est donné dans la figure suivante :

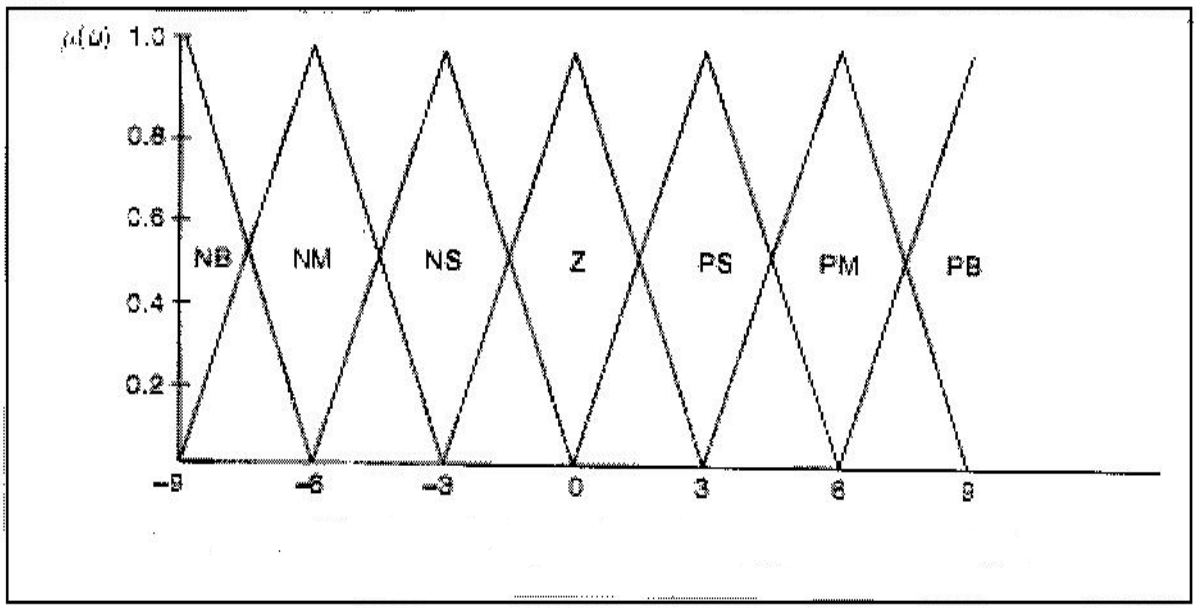


Figure N°10 : Exemple de fonction d'appartenance triangulaire

A' A/,..., PG sont des valeurs linguistique, avec :

- NB : Négative big (négative grand)
- NM : Négative middle (négative moyen)
- NS : Négative small (négative petit)
- AS' : Positive small (positive petit)
- PM : Positive middle (positive moyen)
- PB : Positive big (positive grand)

Les figures suivantes illustrent la fuzzifïcation des entrées et des sorties dans un univers de discours normalisé, comme elle indique les fonctions d'appartenance utilisées pour le contrôleur.

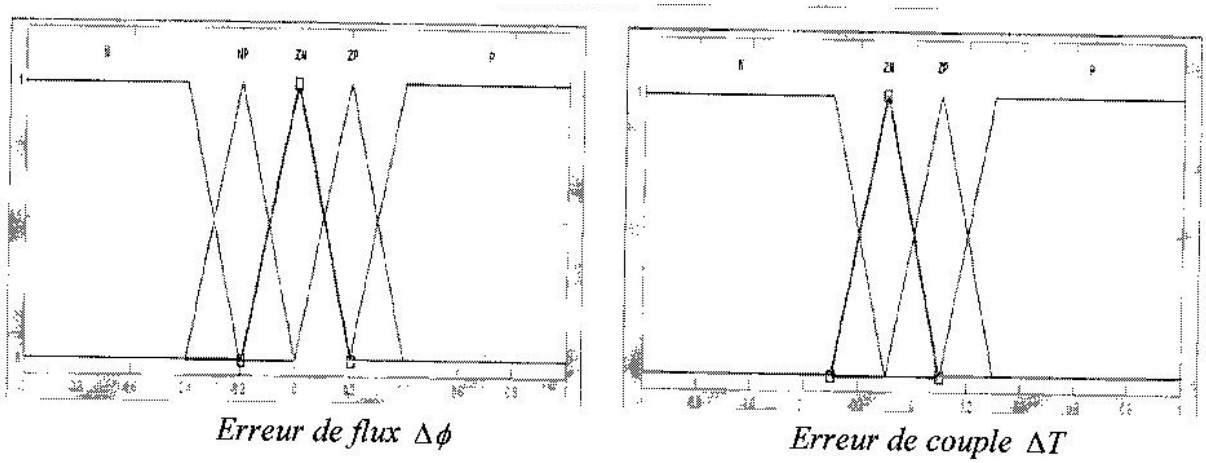


Figure N°11 : Fuzzification des entrées

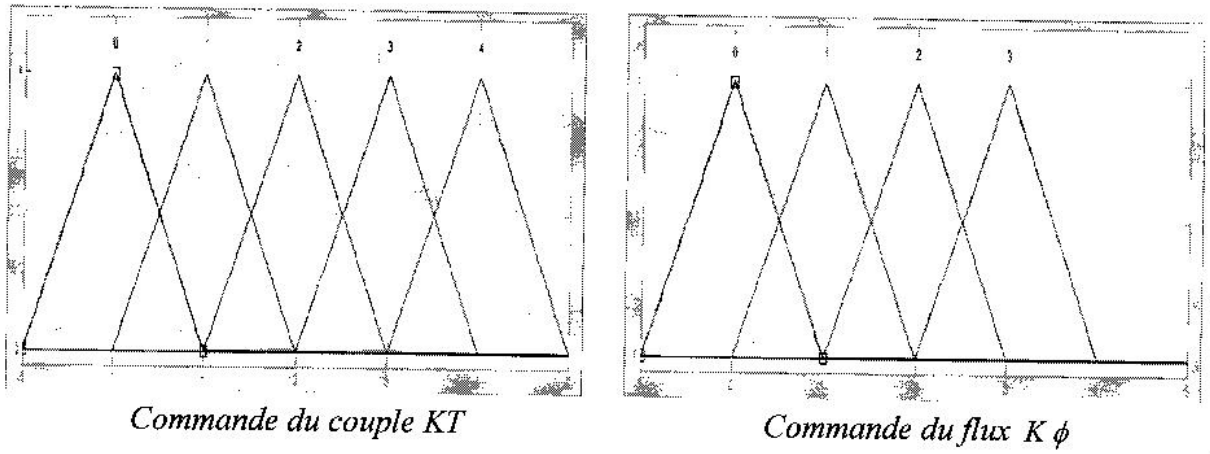


Figure N°12 : Fuzzification des sorties

IV.2.2.5. L'inférence [40]

En logique classique, la règle de raisonnement du modus ponens permet, a partir des deux assertions,

- X est A
- Si x est A alors y est B, de conclure que Y est B.

En logique Houe, la règle s'appelle *modus ponens* généralisé et permet à partir des assertions.

- X est A '
- Si x est A alors y est B, de conclure que Y est B '.

L'inférence est l'opération d'agrégation des règles

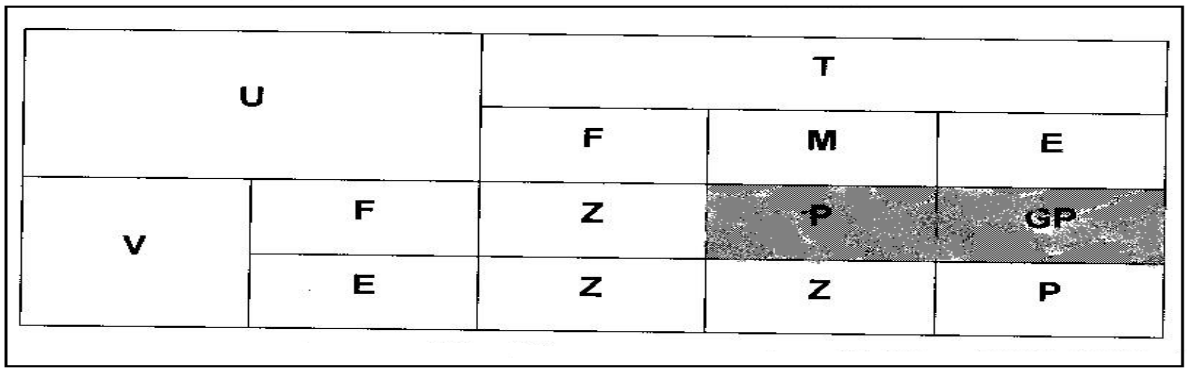


Tableau N°14: Exemple d'inférence des règles

Les règles que décrit ce tableau sont (sous forme symbolique) :

Dans l'exemple ci-dessus on a représenté les règles qui se sont activées à un instant donné par des cases sombres :

SI (**T** est **M** ET **V** est **F**) **ALORS U** = **P** …. (**IV.1**)

SI (T est **E** ET**V** est **F**) **ALORS U** = **GP** …. (**IV.2**)

II arrive que toutes les cases du tableau ne soient pas remplies, on parle alors de règles d'inférences incomplètes. Cela ne signifie pas que la sortie n'existe pas, mais plutôt que le degré d'appartenance n'est nul pour la règle en question.

Il s'agit maintenant de définir les degrés d'appartenance de la variable de sortie à ses sous ensembles flous.

Il existe plusieurs méthodes d'inférence comme «Max-Min», «Max-Produit». «Max-Somme» qui permet d'y arriver.

Ces méthodes se différencient essentiellement par la manière dont vont être réalisés les opérateurs («ET» et «OU») utilisés dans les règles d'inférence**.**

IV.2.2.6. Défuzzification : [38, 39,40]

Consiste à transformer l'ensemble flou résultant en une grandeur de commande précise. Là aussi existe plusieurs méthodes (DHR98), parmi lesquelles :

- La méthode de la hauteur
- La premier maxima.
- La dernier maxima.
- La moyenne maxima.
- Le centre de gravité.
- Le centre des aires.
- Le centre de maxima.
- Le centre de la plus grand surface.

La méthode de défuzzification les plus utilisées en commande floue sont : le centre de gravité, le centre des aires et le centre de maxima

On utilisé dans cette mémoire la méthode de centre de gravité.

Pour la défuzzification on illustre un exemple par la figure suivante :

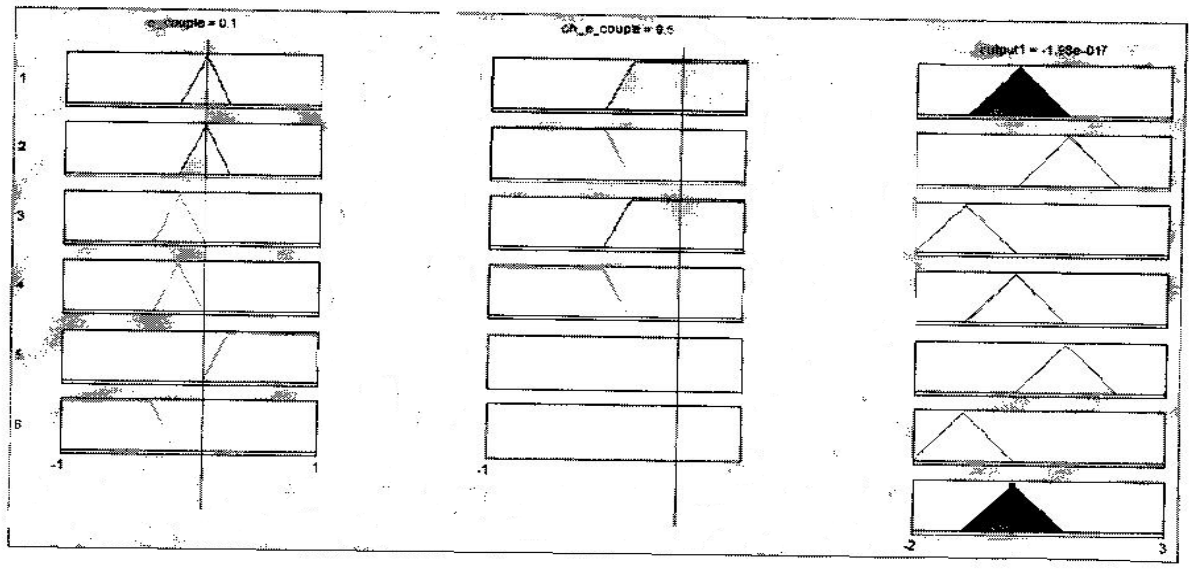


Figure N°13 : Exemple de défuzzification du couple

IV.2.2.7. La méthode du centre de gravité [38]

C'est la méthode de défuzzification la plus connus en commande floue. cette méthode fournit intuitivement la valeur le plus représentative de l'ensemble floue issu de l'agrégation des règle. C'est aussi la méthode la plus coûteuse en temps de calcul.

$$u = \frac{\int_y \mu_{B'}(y).y.dy}{\int_y \mu_{B'}(y).dy}$$

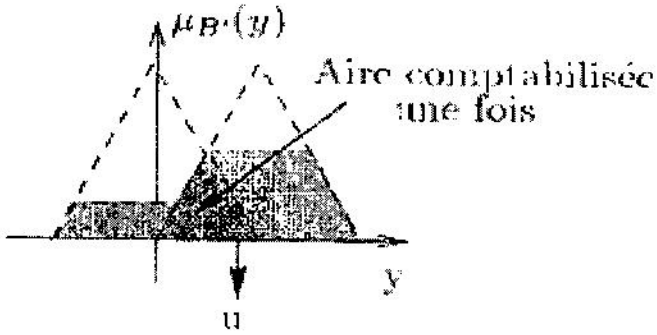


Figure N°14 : Exemple de défuzzification du couple **IV.2.2.8. Dénormalisation**

Cette dernière étape transforme les valeurs normalisées des variables de commande des valeurs appartenant a leur domaine physique respectif.

IV.2.2.9. Développement des contrôleurs flous

Les contrôleurs flous développés utilisent :

- Des fonctions d'appartenances triangulaires et trapézoïdales.
- Un univers de discours normalisé.
- L'implication de **mamdani** pour l'interface [41].
- La méthode du centre de gravité pour la défuzzification.

IV.2.2.10. Résumé et remarque sur la logique floue

La logique floue, ou plus généralement le traitement des incertitudes, a pour objet d'étude la représentation des connaissances imprécises et le raisonnement approché. On peut donc la situer à côté des heuristiques de résolutions de problèmes, des systèmes experts, de l'apprentissage, de l'intelligence artificielle distribuée et même du traitement de la langue naturelle, domaines qui composent les techniques d'intelligence artificielle au sein des sciences cognitives. C'est de cet aspect «intelligence artificielle», où s'établissent des modèles du comportement intellectuel humain, que traite ce livre à travers les applications concrètes qui y sont détaillées et du lien avec des techniques d'apprentissages telles que les réseaux de neurones et les algorithmes génétiques.

Cependant la logique floue peut être intégrée, à côté d'autres extensions, à la logique, qui, en toute généralité peut être vue comme la grammaire des mathématiques (la logique floue a d'ailleurs donné lieu à toute une généralisation des mathématiques classiques fondée sur le concept très simple de sous-ensemble flou). C'est pourquoi on trouvera en annexe d'autres points de vue sur la logique et les fondements du raisonnement.

Enfin, par ses nombreuses applications industrielles en commande, la logique floue est aussi associée à l'automatique. Ces thèmes n'étant pas le propos du livre, figurent néanmoins en annexe.

Dans les problèmes de prise de décision, d'aide au diagnostic et plus généralement dans tous les systèmes à base de connaissances, on souhaite, à partir d'observations, parvenir à une conclusion qui peut être la détermination d'un objet ou une action à prendre. Or lors du fonctionnement de ces systèmes, interviennent des connaissances mal définies, mal décrites et imparfaitement connues, puis au niveau des règles d'inférence, intervient un traitement imparfait et incomplet du déroulement de la déduction, enfin survient le problème du traitement des contradictions et de la fusion (agrégation) de données voisines.

Tous les problèmes concrets sont, en fait, confrontés aux notions d'incertitude et d'imprécision. Ces deux notions sont habituellement mêlées et c'est essentiellement l'observation statistique qui induisait, jusqu'à présent dans la pratique, la mesure probabiliste des incertitudes.

Mais la théorie des probabilités reste assez rigide et il existent d'autres types d'incertitudes liées à la difficulté des observations, aux imprécisions linguistiques, à la fiabilité tant des observateurs humains que des capteurs et instruments de mesure, à l'utilisation de connaissances empiriques et à l'imprécision du raisonnement humain. Toutes ces questions de l'utilisation de catégories linguistiques habituelles et du raisonnement humain vont nous amener à distinguer certains concepts.

IV.3. Logiciel utilisé pour la méthode logique floue

Dans cette partie nous utilisons le logiciel de calcul nommé « MATLAB 7.1 » pour permettre l'utilisation de la méthode « logique floue » que présenterons dans le titre suivant.

IV.3.1. Présentation du logiciel « MATLAB 7.1 »

IV.3.1.1. Introduction [41]

Matlab signifie Matrix laboratory. C'est un logiciel de calcul numéricienne. Il est destine `a traiter des applications `a partir des outils de l'analyse numéricienne matricielle. Matlab possède aussi tout un ensemble de fonctionnalités graphiques permettant de visualiser les résultats numériquement. Il possède des boıtes `a outils, c'est `a dire des fonctionnalité ´es supplémentaires, dédies `a des domaines particuliers du calcul scientifique, comme la résolution d'équations aux dérives partielles, l'optimisation, l'analyse de données, etc. Matlab est aussi un langage de programmation avec des possibilités d'interfaces vers des programmes ´écrits en C ou en Fortran.

En Matlab les calculs sont effectues avec une arithmétique a précision finie. Ceci le différencier des logiciels de calcul symbolique tel que Matlab, mais la comparaison n'a pas lieu d'être. Calcul numéricienne et calcul symbolique sont des outils complémentaires du calcul scientifique.

IV.3.1.2. Définition [41]

Matlab est un système interactif de programmation scientifique, pour le calcul numérique et la visualisation graphique. Développé à l'origine pour le calcul matriciel (le nom Matlab est dérivée de cette représentation Matlab = Matrix Laboratory), il offre aujourd'hui bien d'autres possibilités, dont certaines seront décrites dans la suite. Il contient des bibliothèques spécialisées (toolbox) qui répondent à des besoins spécifiques : analyse numérique, traitement du signal, traitement de l'image, etc.

Matlab est un logiciel qui permet de faire des calculs mathématiques et numérique, et non un logiciel de calcul formel et symbolique comme Maple. Matlab connaît un grand nombre d'opérations ou de fonctions mathématiques : fonctions usuelles, calcul matriciel, fonctions plus spécifiques du signal.

IV.3.2. Présentation de la matrice (II)

Nous présentons la nouvelle matrice « matrice (II) » (même matrice (I)) mais elle contient 36 descripteurs sélectionnés de la dernière opération de choix des descripteurs dans le chapitre précédent, nous utilisons cette « matrice (II) » dans la méthode logique floue pour la classification dans la famille pyrazine.

Alors que « la matrice (II) » contient 125molécules (base des données) ou 125 lignes avec les six (6) classifications correspondantes de cette molécules et 36 descripteurs ou 36 colonnes, nous pouvons utiliser cette méthode (logique floue).

IV.3.2.1. Traité de La matrice (II)

Dans cette étape nous divisons **la matrice (II)** sur 7 matrices avec même base des données (125 molécules) pour toutes les 7 matrices, c'est à dire ne varie pas les nombre des lignes pour toutes les 7 matrices mais chaque matrice qui est divisée contribue aux nombre des colonnes (ou nombres des descripteurs) différents comme suit : 5, 10, 15, 20, 25, 30 et 36 descripteurs ou colonnes.

IV.3.2.2. Optimisé la meilleure matrice qui est divisée

Après l'étude des classifications sur les 7 matrices par la méthode « **logique floue** » (nous utilisons le logiciel **« MATLAB 7.1 »** pour réaliser cette méthode) , nous obtenons que la matrice qui contient 15 descripteurs (15 colonnes) avec 125 molécules de base des données (125 lignes) donnent de meilleurs résultats pour ce modèle de classification (meilleurs résultats pour des estimations et des validations de ce modèle) , mais nous trouvons des contraires dans cette matrice optimisée parce que la base des données contient des molécules qui possèdent même la configuration dans la « **logique floue** » avec des classifications différentes (c'est à dire les molécules possèdent les mêmes règles dans la boite des règles avec des classifications différentes dans la base des données).

Alors, la solution de ce problème est éliminée de quelques molécules dans la base des données qui fait ce problème, c'est à dire éliminé les molécules possèdent les mêmes règles dans la boite des règles du logique floue avec des classifications différentes dans la base des données, et comme résultat nous trouvons 5 molécules dans la base des données qui font cette problème (les molécules possèdent les mêmes règles dans la boite des règles avec des classifications différentes dans la base des données).

IV.3.2.3. Les molécules sont éliminées de la base des données

Les cinq (05) molécules sont éliminées de la base des données sont classées dans la classe GREEN et les numéros de ces molécules suit dans le tableau de base des données sont : 05, 22 ,23 ,31 et 32, nous présentons dans le tableau suivant :

N°	R1	R2	R3	R4	qualité
5	$N(CH_3)_2$	CH_3	H	H	1
22	H	C_2H_5	H	CH_3	1
23	C_2H_5	C_2H_5	H	H	1
31	CH_3	$(CH_2)_2CH(CH_3)_2$	CH_3	H	1
32	OCH_3	C_7H_{15}	H	H	1

Tableaux N°15 : Présentation des molécules qui sont éliminées de la base des données

Enfin nous trouvons une nouvelle matrice qui contient 15 descripteurs (15 colonnes) avec 120 molécules de base des données (120 lignes), que nous nommons **La matrice (III).**

IV.4. La matrice (III)

Dans cette partie nous utilisons **La matrice (III)** contenue dans une base des données 120 molécules avec sa classification (types d'odeurs) correspondant qui est présentée dans les colonnes de cette matrice, et les lignes de la même matrice présente les 15 descripteurs pour chaque molécule.

IV.5. Application de la méthode logique floue sur La matrice (III)

Avants cette application nous présentons **La matrice (III)** dans le tableau suivant :

IV.6. Présentation du système flou

Le système flou est une boite noire qui constitue des règles floues, chaque règle floue présente la réalisation la relation entre les descripteurs de molécule et la classification de cette molécule, chaque molécule admis dans le système floue c'est à dire qui possède une règle ou plus qui réalise ou donné la classification de cette molécule.

L'analyse pour le système flou est
composée de trois (3) parties importantes :

- **Fusion des entrées :** reçoivent les valeurs des descripteurs qui sont normalisés.
- **Système floue** : la boite des règles qui optimise les valeurs de descripteurs qui est normalise pour donner les différents types de classification.
- **Defusion des sorties** : vont donner les valeurs normalisées qui optimisent dans le système flou a quelle classification appartient la molécule qui possèdent ces descripteurs.

Avants l'analyse de système flou il y a une partie très importante c'est la normalisation pour tous les descripteurs sélecte dans un intervalle choisi (par exemple : entre [0,1] ou [-1,1]), et après l'analyse de système flou il y a une autre partie c'est la dénormalisation des sorties pour donner une classification bien définie.

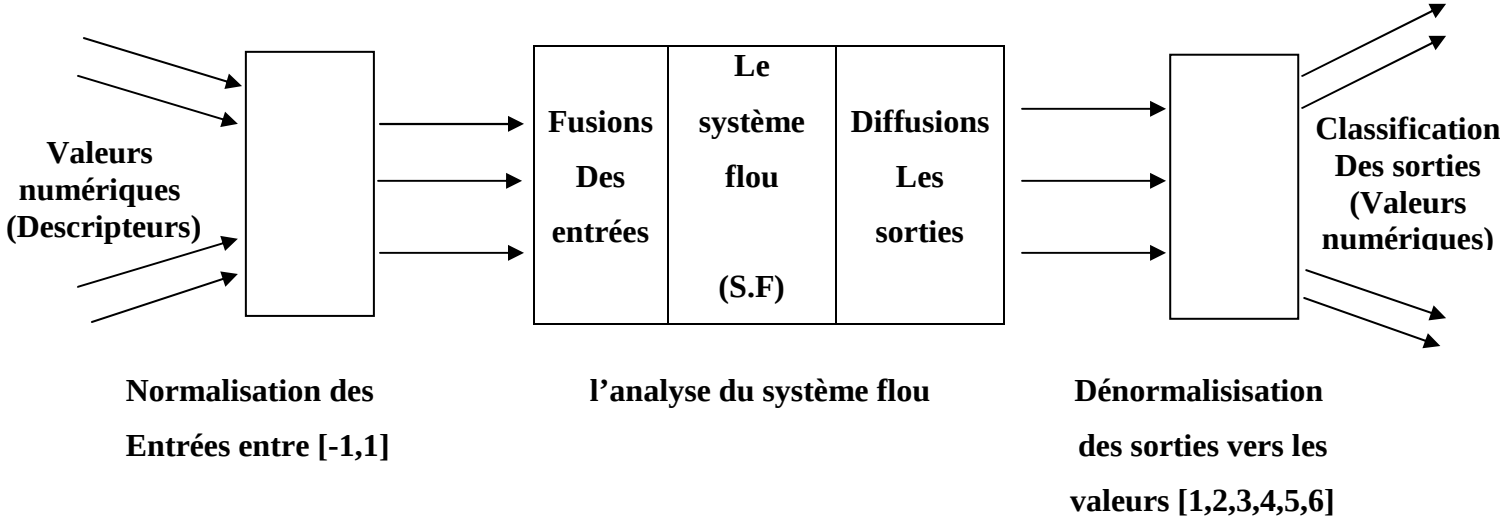


Figure N°15 : Schéma général représentant le système flou

IV.7. Normalisation de la matrice (III) dans le système flou

Normalisation des 15 descripteurs de matrice (III) pour 120 molécules de base des données entre l'intervalle avec présentation de la loi de normalisation correspondant à l'intervalle [-1,1].

IV.7.1. Loi de normalisation

Pour normaliser tous les valeurs numériques pour 15 descripteurs en l'intervalle [-1,1], nous présentons la loi (la fonction) qui normalise tous les valeurs numériques pour 15 descripteurs entre [-1,1] comme suit :

$$x_n = \frac{x - x_{moy}}{\frac{x_{\max} - x_{\min}}{2}} \quad \text{…. (IV.3)}$$

Ou :

X_n : la valeur normalisée entre l'intervalle [-1,1] (l'image).

X_{max}: la valeur numérique maximale entre toutes les valeurs numériques pour toutes les molécules et pour chaque descripteurs.

X_{min} : la valeur numérique minimale entre toutes les valeurs numériques pour toutes les molécules et pour chaque descripteurs.

X_{moy} : la valeur moyenne (numérique) entre la valeur minimale et la valeur maximale pour chaque descripteur qui est calculé avec la relation suivant :

$$x_{moy} = \frac{x_{\min} + x_{\max}}{2} \quad \text{…. (IV.4)}$$

X : la valeur numérique qui peut normaliser (la variable).

Après cette opération, nous représentons le tableau qui donne la loi de normalisation pour chaque descripteur (15 descripteurs) dans le tableau suivant :

Descripteurs	**Intervalle [x_{min}, x_{max}]**	**x_{moy}**	$\frac{x_{max} - x_{min}}{2}$	$\frac{x - x_{moy}}{\frac{x_{max} - x_{min}}{2}}$
GATS7m	[0, 4.399]	2.1995	2.1995	$\frac{x - 2.1995}{2.1995}$
PW5	[0.06, 0.107]	0.0835	0.0235	$\frac{x - 0.0835}{0.0235}$
Jhetv	[1491, 2795]	2143	652	$\frac{x - 2143}{652}$
nArOR	[0, 2]	1	1	$\frac{x - 1}{1}$
EEig11x	[-0.833, 1146]	572.5835	573.4165	$\frac{x - 572.5835}{573.4165}$
EEig04d	[-.132, 2370]	1184.934	1185.066	$\frac{x - 1184.934}{1185.066}$
Lop	[0.802, 2814]	1407.401	1406.599	$\frac{x - 1407.401}{1406.599}$
J	[1631, 2747]	2189	558	$\frac{x - 2189}{558}$
T (N, S)	[0, 3]	1.5	1.5	$\frac{x - 1.5}{1.5}$
PW4	[0.126, 0.190]	0.158	0.0072	$\frac{x - 0.158}{0.032}$
Mor03m	[-2832, -0.505]	-1416.2525	1415.7475	$\frac{x + 1416.2525}{1415.7475}$
Du	[0.354, 0.571]	0.4625	0.1085	$\frac{x - 0.4625}{0.1085}$
R5p	[0.107, 0.498]	0.3025	0.1955	$\frac{x - 0.3025}{1955}$
EEig12x	[-0.786, 0.806]	0.01	0.796	$\frac{x - 0.01}{0.796}$
IDDE	[0.971, 4071]	2035.9855	2035.0145	$\frac{x - 2035.9855}{2035.0145}$

Tableaux N°17 : présentant la loi de normalisation pour chaque descripteur (15 descripteurs)

IV.8. Transformation de la matrice normalisée vers la matrice floue

Transformations des valeurs numériques qui normalisent la matrice (III) entre l'intervalle [-1,1] vers des valeurs floues, les valeurs floues sont présentées P1, P2, M1, M2, G1, G2 qui correspondent aux intervalles suivantes :

- [-1.3, -0.7] correspondant à la valeur floue P1
- [-0.9, -0.3] correspondant à la valeur floue P2
- [-0.5, 0.1] correspondant à la valeur floue M1
- [-0.1, 0.5] correspondant à la valeur floue M2
- [0.3, 0.9] correspondant à la valeur floue G1
- [0.7, 1.3] correspondant à la valeur floue G2

Et les valeurs exactes pour les valeurs floues sont :

- P1 exacte donne la valeur normalisée qui est : -1
- P2 exacte donne la valeur normalisée qui est : -0.6
- M1 exacte donne la valeur normalisée qui est : -0.2
- M2 exacte donne la valeur normalisée qui est : 0.2
- G1 exacte donne la valeur normalisée qui est : 0.6
- G2 exacte donne la valeur normalisée qui est : 1

Exemple de fonctionnement

Exemple de fonction de la règle dans le système flou pour réaliser la méthode de logique floue :

Molécule Obs 001 : la règle de cette molécule dans le système flou est comme suite :

Si la molécule **Obs 001** réalise (SI ; GATS7m est P1 ,et PW5 est M1 ,et Jhetv est M2 ,et nArOR est P1 ,et EEig11x est P1 ,et EEig04d est M2 ,et Lop est M2 ,et J est M2 ,et T (N, S) est P1 ,et PW4 est M1 ,et Mor03m est M1 ,et Du est G1 ,et R5p est M2 ,et EEig12x est P1 ,et IDDE est G1 ALLORS QUE la classification est GREEN).

Enfin nous transformons tous les descripteurs et les valeurs numériques de matrice (III) vers la forme des valeurs floues, et nous représentons la nouvelle matrice que nous nommons matrice floue.

IV.9. Présentation de la matrice floue

Nous présentons la matrice floue qui est transformée dans le tableau suivant :

IV.10. Etude de la classification de la famille pyrazine par la méthode logique floue (En utilisant la matrice floue précédente)

Dans cette étude nous divisons la base des données en deux parties : partie d'estimation (70% de base des données ou 84 molécules) et partie de validation (30% de base des données ou 36 molécules).

IV.10.1. Partie d'estimation

Dans cette partie, nous utilisons 84 molécules qui définissent 70% de base des données pour le test d'estimation [Green (19 molécules), Nutty (17 molécules), Bell-pepper (31 molécules), Earthy (10 molécules), Pecasy (04 molécules), Sweet (03 molécules)].

IV.10.2. Partie de validation

Dans cette partie, nous utilisons 36 molécules qui définissent 30% de base des données pour le test de validation [Green (08 molécules), Nutty (08 molécules), Bell-pepper (13 molécules), Earthy (04 molécules), Pecasy (02 molécules), Sweet (01 molécule)].

IV.10.3. Les tableaux de statistiques qui représentent les statistiques simples pour les deux Parties (Estimation et validation) :

Nous présentons les statistiques simples pour les parties d'estimations et de validations dans les tableaux suivants :

IV.10.4. Partie de simulation et modélisation

Dans cette partie on utilise le logiciel MATLAB 7.1 pour simuler la partie d'estimation et la partie de validation pour le système flou ou pour utiliser la méthode logique floue.

La modèle qui est choisi pour appliquer cette méthode est le modèle de MAMDANI (il y a deux (2) modèles pour utiliser la méthode logique floue : modèle de MAMDANI et le modèle de SURGENO).

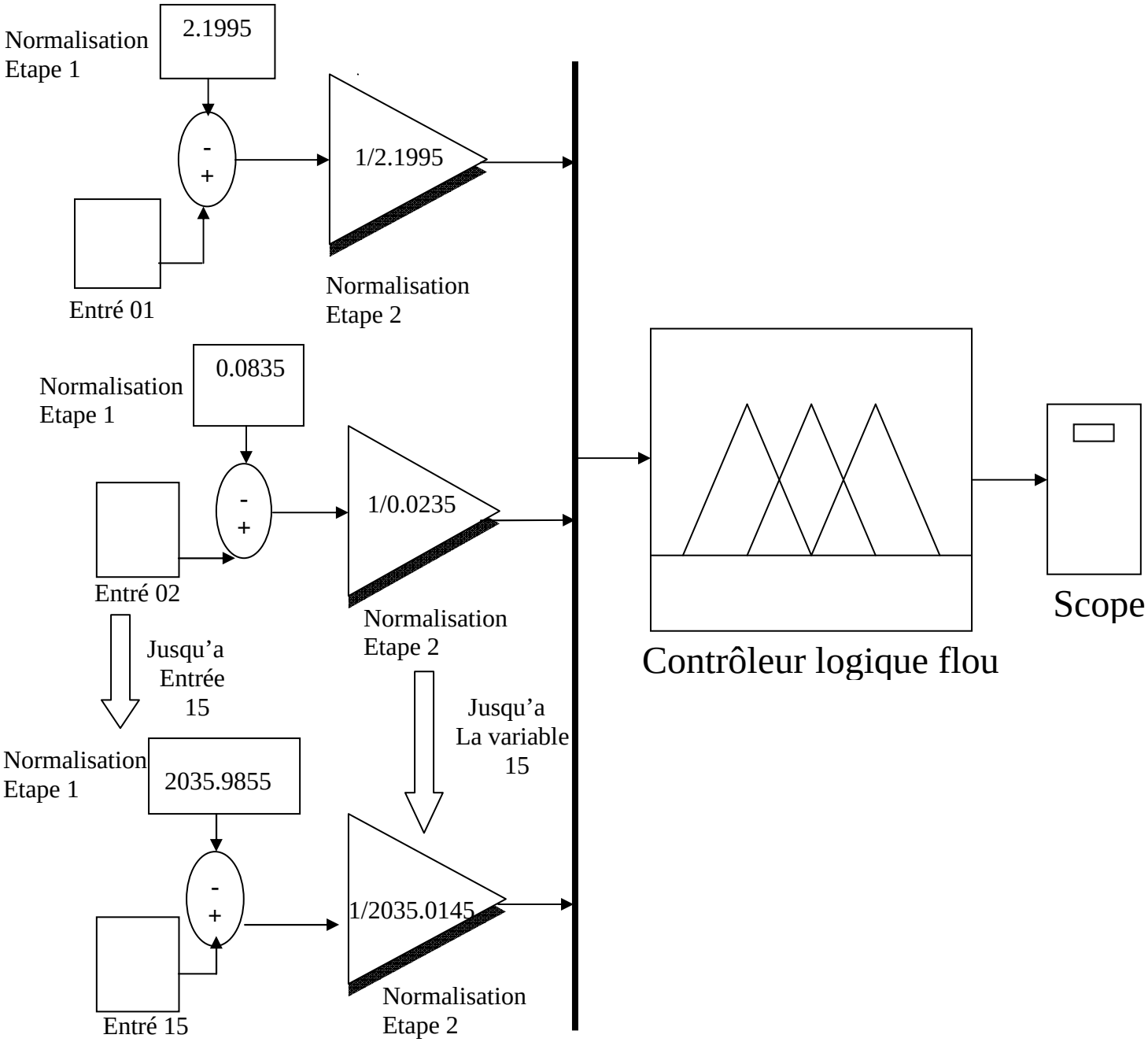


Figure N°16 : Schéma représentant le modèle utilisé pour la simulation (logique floue)

IV.11. Résultats et discutions

Nous présentons les résultats des classifications pour le test d'estimation et de validation et de classification a priori et a posteriori pour la partie de validation

Dans les tableaux suivants :

IV.12. Discussions et analyses des résultats

- Les molécules de la partie d'estimation (contient 84 molécules ou 70% de base des données) sont construites de modèles de classification (la boite noir ou système floue) , ces modèle de classification donne le pourcentage d'estimation de ces molécules est 100% (84 molécules ou 70% de base des données) parce que les 84 molécules sont présentes tous dans le modèle de classification (système floue), cette présentation dans le système floue se forme de règles floues (84 règles) et ces caractère sont importants est très positifs pour l'utilisation de logique floue dans la classification
  (L'étude de la classification a donné le pourcentage d'estimation 10%).
- Dans la partie de validation le pourcentage de classification totale est 61.11%, et pour détailler cette discussion des résultats on résume comme suit :

1. les molécules qui possèdent la classe NUUTY donnent de meilleurs résultats de tests de validation (75% ou 6 molécules sont bien classées par apport à 8 molécules), les deux (2) molécules qui sont numéroté 38 et 46 sont mal classées (molécules numéro 38 classées dans BELL-PEPPER, et molécules numéro 46 classées dans GREEN).
2. les molécules qui possèdent la classe BELL-PEPPER donnent le pourcentage de test de validation (69.23% ou 9 molécules sont bien classées par apport à 13 molécules), les quatre (4) molécules qui sont numéroté 62, 58, 69 et 73 sont mal classées (les molécules numéro 62 et 58 classées dans NUUTY, et molécule numéro 69 classée dans EARTHY, et molécule numéro 73 classée dans GREEN).
3. les molécules qui possèdent la classe GREEN donnent le pourcentage de test de validation (50% ou 4 molécules sont bien classées par apport à 8 molécules), les quatre (4) molécules qui sont numéroté 4, 8, 9 et 11 sont mal classées (les molécules numéro 4 et 8 classées dans BELL-PEPPER, et molécule numéro 9 classée dans NUUTY, et molécule numéro 11 classée dans EARTHY).
4. les molécules qui possèdent la classe EARTHY donnent le pourcentage de test de validation (50% ou 2 molécules sont bien classées par apport à 4 molécules), les deux (2) molécules qui sont numéroté 103 et 105 sont mal classées (la molécule numéro 103 classée dans NUUTY, et la molécule numéro 105 classée dans SWEET).

5. les molécules qui possèdent la classe PECASY donnent le pourcentage de test de validation (50% ou 1 molécules sont bien classées par apport à 2 molécules), la molécule qui est numéroté 117 est mal classées (la molécule numéro 117 classée dans BELL-PEPPER).
6. les molécules qui possèdent la classe SWEET donnent le pourcentage de test de validation (0% ou 0 molécules sont bien classées par apport à un molécule), la molécule numéroté 123 est mal classée (la molécule numéro 123 classée dans PECASY).

IV.13. Interprétations des résultats

Généralement le modèle de classification est optimiste par la méthode de logique floue c'est un bon modèle de classification de famille pyrazine, parce que le pourcentage de test de validation totale (toutes les classes) est 69.11% et ce pourcentage est considéré un très bon résultat.

Les molécules qui sont mal classées a partir de tableau de test de validation et tableau de classifications a priori et a posteriori sont interprétées comme suit :

- les molécules qui sont numérotées : 4, 8, 38 et 117 sont classées dans la classe BELL-PEPPER parce que ces molécules possèdent dans la forme structurelle des radicaux R3 et R4 différents (R3 ≠ R4) (suivre la base des données dans le tableau précédent).

- les molécules qui sont numérotées : 9, 58, 62 et 103 sont classées dans la classe NUUTY parce que ces molécules possèdent dans la forme structurelle R3 = H et R4 = H c'est à dire R3 et R4 ont le même radical hydrogéné (suivre la base des données dans le tableau précédent).

- les molécules qui sont numérotées : 46 et 73 sont classées dans la classe GREEN parce que l'atome d'oxygène (O) existe dans la forme structurelle de ces molécules (suivre la base des données dans le tableau précédent).

- les molécules qui sont numérotées : 11 et 69 sont classées dans la classe EARTHY parce qu'il existe deux radicaux symétriques et non différents dans la forme structurelle pour chaque molécule (R3 ≠ R4) (suivre la base des données dans le tableau précédent).

Remarque importante

Nous remarquons que les interprétations pour les deux méthodes (analyse discriminante et logique floue) sont conservées pour les mêmes causes d'interprétations pour les molécules mal classées (interprétation qui utilisé la forme structurelle de molécule et la base des données dans le tableau précédent).

V.1. Introduction

Dans cette partie nous utilisons les descripteurs qui se trouvent dans la **matrice (II)** avec la méthode « **réseaux de neurones** » pour étudier la classification des molécules de famille pyrazine avec six types d'odeurs (Green, Nutty, Bell-pepper, Earthy, Pecasy, Sweet), l'objectif de cette utilisation est obtenir un meilleur fonction (modèle, réseaux) de classification pour la famille pyrazine qui donne de meilleurs résultats de test d'estimations et de validations , ce fonction est utilisé pour confirmer ou trouver la classification des molécules de la famille pyrazines qui ne connaissent pas de type de classification (type d'odeurs).

Pour appliquons la méthode « **réseaux de neurones** », nous utilisons logiciel **« SPSS 16.0 »** qui présenter dans les titres suivant**.**

V.2. Méthode réseaux de neurones [42]

Les réseaux de neurones artificiels (formels) intervenir dans plusieurs domaines comme la médecine, biologie, chimie et statistique. On développés pour résoudre des problèmes de contrôle.

Dans le domaine chimique on base sur les composés organiques, ce qui après l'utilisation de la méthode de contributions de groupes qui fragmente le composé. On distinguer la relation par réseaux de neurone (NeuroOne) pour calculer les propriétés des composés organiques.

Ce chapitre est consacre à la présentation de réseaux de neurones et on va aborder successivement: l'historique, la définition, la comparaison avec le cerveau humain, les différents types ainsi que l'apprentissage du réseau.

V.2.1. Historique [42]

En 1943 par W.MCCulloch et W.Pitts du neurone formel qui est une abstraction du neurone physiologique. Le retentissement va être énorme. Par cette présentation, ils veulent démontrer que le cerveau est équivalent à une machine de Turing, la pensée devient alors purement des mécanismes matériels et logiques. Ils déclarèrent en 1955 [27] "Plus nous apprenons de choses au sujet des organismes, plus nous sommes amenés à conclure qu'ils ne sont pas simplement analogues aux machines, mais qu'ils sont machine."[Mysterium Iniquitatis of Sinful Man Aspiring into the Place of God, repris in Embodiments of mind]". La démonstration de McCulloch et Pitts sera un des facteurs importants de la création de la cybernétique.

En 1949, D. Hebb présente dans son ouvrage "The Organization of Behavior" une règle d'apprentissage, de nombreux modèles de réseaux aujourd'hui s'impirent encore de la règle de Hebb.

En 1958, F. Rosenblatt développe le modèle du Perceptron. C'est un réseau de neurones inspiré du système visuel. Il possède deux couches de neurones : une autre de perception et une couche lié à la prise de décision. C'est le premier système artificiel capable d'apprendre par expérience.

Dans la même période, Le modèle de L'ADALINE (ADAptive LINar Elément) a été présenté par B. Widrow, chercheur américain à Stanford. Ce modèe sera par la suite le modèle de base des réseaux multicouches.

En 1969, M. Minsky et S. Papert publient une critique des propriétés du Perceptron. Cela va avoir une grande incidence sur la recherche dans ce domaine. Elle va fortement diminuer jusqu'en 1972, où T. Kohonen présente ses travaux sur les mémoires associatives, et propose des applications à la reconnaissance de formes.

C'est en 1982 que J. Hopfield présente son étude d'un réseau complètement rebouclé, dont il analyse la dynamique.

Aujourd'hui, les réseaux neuronaux sont utilisés dans de nombreux domaines (entre autres, vie artificielle et intelligence artificielle) à cause de leur propriété en particulier, leur capacité d'apprentissage, et qu'ils soient des systèmes dynamiques.

V.2.2. Définition [42,43]

Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau.

V.2.3. Comparaison avec le cerveau humain [43]

Les modèles de réseaux de neurones artificiels sont, à l'origine, une imitation du fonctionnement du cerveau. Il contient, chez l'homme, environ 10 milliards de neurones, et chacun est connecté à environ 10.000 autres neurones. On voit ainsi sa complexité étonnante.

Les connexions permettent le transfert d'informations sous forme d'impulsions électriques entre les neurones (fig. N°17).

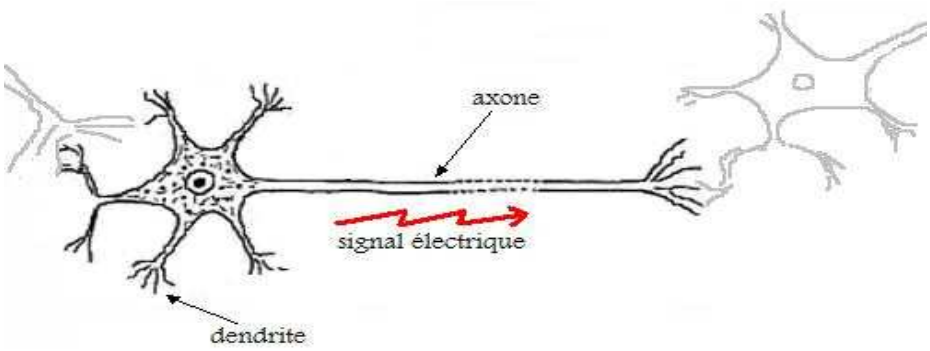


Figure N°17 : Réseau de neurones de cerveau humain.

Un neurone reçoit des impulsions de ses voisins par l'intermédiaire des "dendrites". Si la somme des signaux dépasse un certain seuil, il renvoie un signal vers d'autres neurones, par l'intermédiaire de son "axone". Ce mécanisme complexifie la façon dont les informations sont transmises : un neurone ne se borne pas à faire passer l'information, il la filtre.

Pour résumer, un neurone peut être schématisé ainsi (fig. N°18) il fait la somme de toutes les informations qu'il reçoit et il émet un signal à condition que la somme soit suffisamment élevée.

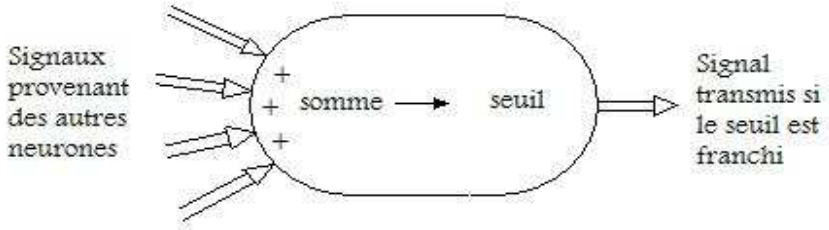


Figure N°18 : Réseaux de neurones artificiels.

Si l'on ramène la contribution d'un neurone au cerveau tout entier, on se rend compte que chacun effectue un travail très simple par rapport au résultat obtenu. En effet, les neurones réalisent des opérations basiques, et pourtant, lorsque l'on en met 10 milliards ensemble, on peut créer une entité pensante.

Cependant, cela n'est pas suffisant : un cerveau ne peut rien faire s'il n'a pas de quoi apprendre. Il a besoin d'informations venant de l'extérieur. C'est pour cela qu'il est relié aux différents organes du corps. Par exemple, il reçoit les images provenant des yeux, les sons, les douleurs...

Grâce à ces informations il est capable de faire son apprentissage : lorsqu'une action a provoqué une douleur, il doit changer l'organisation des neurones afin de ne pas répéter la même erreur.

L'objectif des réseaux de neurones artificiels est donc de modéliser le fonctionnement des neurones réels, mais aussi de permettre un apprentissage.

V.2.4. Les neurones formels

Un "neurone formel" (ou simplement "neurone") est une fonction algébrique non linéaire et bornée, dont la valeur dépend de paramètres appelés coefficients ou poids. Les variables de cette fonction sont habituellement appelées "entrées" du neurone, et la valeur de la fonction est appelée sa "sortie" [44].

Un neurone est donc avant tout un opérateur mathématique, dont on peut calculer la valeur numérique par quelques lignes de logiciel. On a pris l'habitude de représenter graphiquement un neurone comme indiqué sur la Figure N°19 [42].

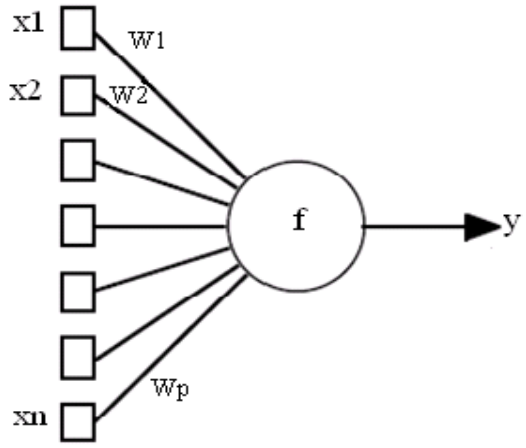


Figure N°19 : Un neurone réalise une fonction non linéaire bornée.

Un neurone réalise une fonction non linéaire bornée $y = f(x_1, x_2, \ldots x_n; w_1, w_2, \ldots w_p)$ où les $\{x_i\}$ sont les variables et les $\{w_i\}$ sont des paramètres.

Pour des raisons que nous expliquerons plus loin, les neurones les plus fréquemment utilisés sont ceux pour lesquels la fonction f est une fonction non linéaire (généralement une tangente hyperbolique) d'une combinaison linéaire des entrées :

$$y = th\left[w_0 + \sum_{i=1}^{n-1} w_i x_i\right] \quad \ldots . \textbf{(V.1)}$$

Les paramètres sont attachés aux entrées du neurone : la sortie du neurone est une fonction non linéaire d'une combinaison des entrées $\{x_i\}$ pondérées par les paramètres $\{w_i\}$, qui sont alors souvent désignés sous le nom de « poids » ou, en raison de l'inspiration biologique des réseaux de neurones, « poids synaptiques », $\{w_0\}$ le poids constant d'entrée x_0 «biais» [43].

Un neurone formel ne réalise donc rien d'autre qu'une somme pondérée suivie d'un non linéarité. C'est l'association de tels éléments simples sous la forme de réseaux qui permet de réaliser des fonctions utiles pour des applications industrielles.

V.2.5. Modélisation d'un neurone formel [42]

Les réseaux de neurones formels sont à l'origine d'une tentative de modélisation mathématique du cerveau humain. Les premiers travaux datent de 1943 et sont l'œuvre de MM. Mac Culloch et Pitts, présentent un modèle assez simple et explorent les possibilités de ce modèle.

La modélisation consiste à mettre en œuvre un système de réseaux neuronaux sous un aspect non pas biologique mais artificiel. Cela suppose que d'après le principe biologique on aura une correspondance pour chaque élément composant le neurone biologique, donc une modélisation pour chacun d'entre eux.

V.2.6. Les entrées [42]

Elles peuvent être :

- Booléennes ;
- Binaires (0, 1) ou bipolaires (-1, 1) ;
- Réelles.

V.2.7. Fonction d'activation [42,44]

La fonction d'activation (ou fonction de seuillage, ou encore fonction de transfert) permet de définir l'état interne du neurone en fonction de son entrée totale, citons à titre d'exemple quelques fonctions souvent utilisées sont citées, dans le tableau N°24 :

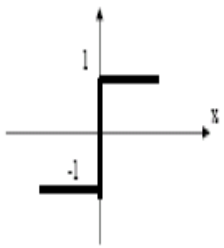
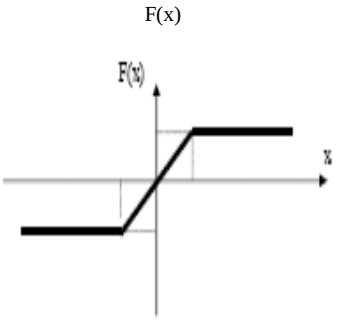
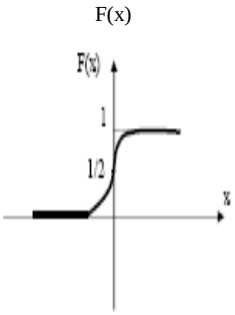
Fonction binaire a seuil	**Fonction linéaire à seuil ou multi-seuils**	**Fonction sigmoïde**
Sgn (x) 	F(x) 	F(x) 
$Sgn(x)=\begin{cases}+1 & \text{si } x \geq 0 \\ -1 & \text{sinon}\end{cases}$	$F(x)=\begin{cases}x & x \in [u, v] \\ v & \text{si } x \geq v \\ u & \text{si } x \leq u\end{cases}$	$f(x)=\frac{1}{1+e^{-x}}$
(1)	(2)	(3)

Tableau N°24 : Différents types de fonctions de transfert pour le neurone artificiel.

V.2.8. Fonction de sortie [42]

Elle calcule la sortie d'un neurone en fonction de son état d'activation. En général, cette fonction est considérée comme la fonction identité. Elle peut être :

- Binaire (0, 1) ou bipolaire (-1, 1) ;
- Réelle.

V.2.9. Architecture des réseaux neuronaux [42]

La figure 3 dans le tableau N°24 présente une taxonomie possible en termes d'architecture de réseaux. La différence majeure porte sur la possibilité d'avoir des boucles dans le réseau (cycle ou circuit), ce qui permet au système d'avoir accès (dans une certaine mesure) au passé.

Par ailleurs on notera la possibilité d'avoir des couches de cellules, c'est-à-dire des groupements de faire transiter le flot de données dans le réseau de manière séquentielle (entre les couches) et parallèle (au sein d'une mémé couche).

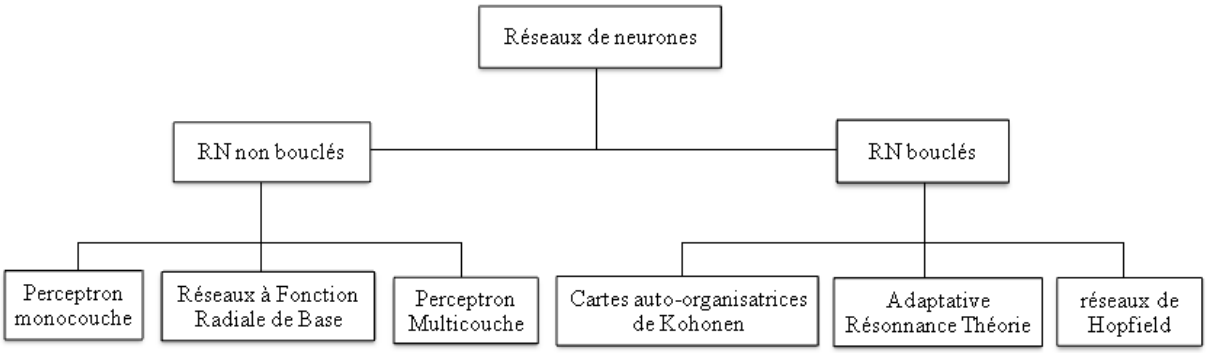


Figure N°20 : Une taxonomie possible.

V.2.10. Différents types des RN [42]

On distingue deux types de réseaux de neurones : les réseaux non bouclés (FeedForward) et les réseaux bouclés ou récurrent (Feedback).

V.2.11. Les réseaux de neurones non bouclés [42]

Un réseau de neurones non bouclé réalise une (ou plusieurs) fonctions de ses entrées, par composition des fonctions réalisées par chacun des neurones.

V.2.11. Les Perceptrons [42]

V.2.11.1. Perceptrons monocouche

C'est un réseau simple, puisque il ne se compose que d'une couche d'entrée et d'une couche de sortie. Il est copie, sur le système visuel et de ce fait il a été connu dans un premier but de reconnaissance des formes. Cependant, il peut aussi être utilisé pour faire de la classification et pour résoudre des opérations logiques simples. Sa principale limite est qu'il ne peut résoudre que des problèmes linéairement séparables. Il suit généralement un apprentissage supervisé selon la règle de correction de l'erreur (ou selon la règle de Hebb).

V.2.11.2. Perceptrons multicouche (PMC)

C'est une extension du précédent, avec une ou plusieurs couches entre l'entrée et la sortie. Chaque neurone dans une couche est connecté à tous les neurones de la couche précédente et de la couche suivante (excepté pour la couche d'entré et de sortie) et il n'y a pas de connexions entre les cellules d'une même couche. Les fonctions d'activation utilisées dans ce type de réseaux sont principalement les fonctions à seuil ou sigmoïdes. Il peut résoudre des problèmes non linéairement séparables et des problèmes logiques plus compliqués. Il suit aussi un apprentissage supervisé selon la règle de correction de l'erreur.

V.2.12. Réseaux à Fonction Radiale de Base [43]

Ce sont les réseaux que l'on nomme aussi RBF (pour «Radial Basic Function ») comportes deux couches de neurones. Les cellules de sortie effectuent une combinaison linéaire de fonctions de base non linéaires, fournies par les neurones de la couche cachée. Ces fonctions de base produisent une réponse différente de zéro seulement lorsque l'entrée se situe dans une petite région bien localisée de l'espace des variables. Bien que plusieurs modèles de fonctions de base existent, le plus courant est de type Gaussien :

$$y_{1,i}(x) = \exp\left[-\frac{(X-W_{1,i})^T (X-W_{1,i})}{2\sigma_i^2}\right] \quad \text{…. (V.2)}$$

Où:

- X est le vecteur d'entrée du réseau ;
- $y_{1,i}$ est la sortie du neurone *i* de la première couche ;
- W et σ_i^2 sont respectivement le vecteur de poids synaptiques et le paramètre de normalisation de ce neurone (W correspond ici aux coordonnées du centre de la Gaussienne).

La sortie d'un neurone de la seconde couche est simplement donnée par :

$$y_{2,i} = W_{2,i}^T Y_1 \quad \text{…. (V.3)}$$

Où Y_1 est le vecteur des sorties des neurones de la première couche.

V.2.13. Les réseaux de neurones bouclés (ou récurrents) [44]

Un réseau de neurones bouclé à temps discret réalise une (ou plusieurs) équations aux différences non linéaires, par composition des fonctions réalisées par chacun des neurones et des retards associés à chacune des connexions.

Tout cycle du graphe des connexions d'un réseau de neurones bouclé doit comprendre au moins une connexion de retard non nul.

V.2.14. Les cartes auto-organisatrices de Kohonen [43,44]

Ce type de réseaux aussi appelé SOM (Self Organised Maps), est un réseau à apprentissage non-supervisé qui établit une carte discrète, ordonnée typologiquement en fonction de patterns d'entrée. Le réseau forme ainsi une sorte de treillis où chaque nœud du treillis est neurone associé à un vecteur de poids. La correspondance entre chaque vecteur de poids est calculée pour chaque entrée. Par la suit, le vecteur de poids ayant la meilleur corrélation, ainsi que certains de ses voisin, vont être modifiés afin d'augmenter cette corrélation.

Les LVQ (pour « Learning Vector Quantization ») sont un cas particulier des SOM, où seul le nœud ayant la meilleure corrélation est adapté. On parle, pour ce type de réseau où seul le « vainqueur » est sélectionné, de réseau à compétition.

V.2.15. Les réseaux de Hopfield [42]

Les réseaux de Hopfield sont donc des réseaux récurrents et entièrement connectés. Dans ce type de réseau. Chaque neurone est connecté à chaque autre neurone et il n'y a aucune différenciation entre les neurones d'entrée et de sortie. Ils fonctionnent comme une mémoire associative non linéaire et sont capables de trouver un objet stocké en fonction de représentations partielles ou bruitées. L'application principale des réseaux de Hopfield est l'entrepôt de connaissances mais aussi la résolution de problèmes d'optimisation. Le mode d'apprentissage utilisé ici est le mode non-supervisé.

V.2.16. Les ART [42]

Les réseaux ART (Adaptative Résonances Théorie) sont des réseaux à apprentissage par compétition. Le problème majeur qu'il se pose dans ce type de réseau est le dilemme « stabilité/plasticité ». En effet, dans un apprentissage par compétition, rien ne garantit que les catégories formées vont reste stables. La seule possibilité, pour assurer la stabilité, serait que le coefficient d'apprentissage tende vers zéro, mais le réseau perdrait alors sa plasticité. Les ART ont été connus spécifiquement pour contourner ce problème dans ce genre de réseau, les vecteurs de poids ne seront adaptés que si l'entrée fournie est suffisamment proche, d'un prototype déjà connu par le réseau, on parlera alors de résonance. A l'inverse, si l'entrée s'éloigne trop des prototypes existants, une nouvelle catégorie va alors se créer, avec prototype, l'entrée qui a engendré sa création. Il est à noter qu'il existe deus principaux types de réseaux ART : les ART-1 pour des entrées binaires et les ART-2 pour des entrées continues. Le mode d'apprentissage des ART peu être supervisé ou non.

V.2.17. L'apprentissage des RNA [42]

V.2.17.1. Type d'apprentissage

On appelle « apprentissage » des réseaux de neurones la procédure qui consiste à estimer les paramètres des neurones du réseau, afin que celui-ci remplisse au mieux la tâche qui lui est affectée.

Dans le cadre de cette définition, on peut distinguer trois types d'apprentissages :

L'apprentissage « supervisé », l'apprentissage « non supervisé » et l'apprentissage «hybride».

V.2.17.2. Apprentissage supervisé

Dans ce cas fournit au réseau la donnée à traiter aussi la réponse attendue. Le réseau effectue une évaluation de la donnée, puis compare la valeur obtenue avec la valeur désirée, il va ensuite modifier ses paramètres internes afin de minimiser l'erreur constatée.

L'apprentissage par renforcement est une variante de l'approche supervisée, dans ce cadre on fourni au réseau une critique qui qualifie la réponse calculée.

V.2.17.3. Apprentissage non supervisé (auto-organisationnel)

Dans ce paradigme aucune information (en plus des données à apprendre) n'est fournie au système. Celui-ci est amené à découvrir la structure sous-jacente des données afin de les organiser en clusters.

V.2.17.4. Apprentissage hybride

Plus rare (et encore mal explorée), cette mode reprend en fait les deux autres approches, puisque une partie des poids va être déterminée par apprentissage supervisé et l'autre partie par apprentissage non supervisé.

V.2.17.5. Les règle d'apprentissage

Un réseau de neurones artificiel, comme le cerveau animal, apprendre à réagir correctement à un stimulus provenant de l'extérieur. Le principe de l'apprentissage consiste à soumettre le réseau à un stimulus dont on connaît la réponse souhaitée, autant de fois qu'il lui est nécessaire à la modification des poids des connexions, jusqu'à obtention de la bonne réponse.

Il existe plusieurs règles de modification des poids, les principales sont :

- La règle de Hebb ;
- Règle de correction d'erreurs ;
- Apprentissage de Boltzmann ;
- Règle d'apprentissage par compétitions.

V.2.17.6. Règle de Hebb [42, 43,44]

Cette règle, basée sur des données biologiques, modélise le fait que si des neurones, de part et d'autre d'une synapse, sont activées de faon synchrone et répétée, la force de la connexion synaptique va aller croissant. Il est à noter ici que l'apprentissage est localisé, c'est-à-dire que la modification d'un poids synaptique w_{ij} ne dépend que de l'activation d'un neurone i et d'un autre neurone.

V.2.17.7. Règle de correction d'erreurs [42, 43,44]

Cette règle s'inscrit dans le paradigme d'apprentissage supervisé, c'est-à-dire, dans le cas où on fournit au réseau, une entrée et la sortie correspondante. Si on considère y, la sortie calculée par le réseau et d, la sortie désirée, le principe de cette règle est d'utiliser l'erreur (d-y), afin de modifier les connexions et de diminuer ainsi l'erreur globale du système. Le réseau va donc s'adapter jusqu'à ce qu'y soit égale.

V.2.17.8. Apprentissage de Boltzmann [42]

Ce qu'il faut savoir tout d'abord, c'est que les réseaux de Boltzmann sont des réseaux symétriques récurrents et qu'ils possèdent deux sous-groupe de cellules, le premier étant relié à l'environnement (cellules dites visibles) et le second ne l'étant pas (cellules dites cachées). Cette règle d'apprentissage est de type stochastique (= qui relève partiellement du hasard) et elle consiste à ajuster les poids des connexions, de telle sorte que l'état des cellules visibles satisfasse une distribution probabiliste souhaitée.

V.2.17.9. Règle d'apprentissage par compétitions [42,43]

La particularité de cette règle, c'est qu'ici l'apprentissage ne concerne qu'un seul neurone. Le principe de cet apprentissage est de regrouper les données en catégories, les patrons similaires vont donc être rangs dans une même classe, en se basant sur les corrélations des données, et seront représentés par un seul neurone, on parle de « winner-take-all ».

Dans un réseau à compétition simple, aux autres cellules de la couche de sortie (connexions inhibitrices) et à elle-même (connexion excitatrice), la sortie va donc dépendre de la compétition entre les connexions inhibitrices et excitatrices.

V.2.17.10. Les algorithmes d'apprentissage [42]

L'algorithme Levenberg-Marquardt a aujourd'hui souvent la faveur des spécialistes, car il converge plus vite que l'algorithme de rétropropagation du gradient et vers une solution meilleure. Mais il exige une grande capacité de mémoire de l'ordinateur sur lequel il tourne, proportionnelle au carré du nombre de nœud. De ce fait, il est limité à des petits réseaux, avec peu de variables, il est aussi limité à un nœud de sortie.

L'algorithme de rétropropagation du gradient est le plus ancien et le plus répandu, surtout sur la grande valeur de données. Mais il manque de fiabilité, en raison de sa sensibilité aux minima locaux.

L'algorithme de la descente du gradient conjugué est un bon compromis, puisque ses performances rapprochent de celles de Levenberg-Marquardt en termes de convergence, mais qu'il est applicable à des réseaux plus complexes, avec éventuellement plusieurs sorties.

V.2.18. Normalisation des données [42,44]

V.2.18.1. Variables contenues

Rappelons que les données utilisées dans un réseau de neurones doivent être numériques et leurs modalités comprises dans l'intervalle [0,1], ce qui implique, quand ce n'est pas le cas, une normalisation des données. Pour que le travail de normalisation décrit ci-dessous soit correct, il faut, bien entendu, que le jeu de données d'apprentissage couvre toutes les valeurs rencontrées dans la population tout entière, et, en particulier, les valeurs extrêmes des variables continues.

Même en les normalisant, les variables continues peuvent connaître le problème d'écrasement des valeurs normales par les valeurs extrêmes. Ainsi, la plupart des revenus mensuels se situent entre 0 et 10000 euros ; mais si un revenu dépasse 100000 euros, la normalisation standard de la variable « revenu », c'est-à-dire son replacement par la variable :

$$\frac{\text{Revenu - Revenu minimum}}{\text{Revenu maximum – Revenu minimum}}$$

Rendra presque indiscernable l'écart entre 5000 et 10000 euros, et le mettra sur le même plan que l'écart-beaucoup moins significatif- entre 95000et 100000 euros.

- Plusieurs moyens existent pour bien normaliser ce type de variable ;
- On peut discrétiser la variable, et la remplacer par exemple par ses quartiles ;
- On peut normaliser, non pas la variable, mais le logarithme de cette variable, qui « distend » le début de l'échelle ;
- On peut normaliser la variable linéairement, comme indiqué ci-dessus, pour ses valeurs comprises entre -3 et +3 fois l'écart-type σ autour de la moyenne μ, et envoyer les valeurs inférieures à $\mu-3\sigma$ sur 0, et les valeurs supérieures à $\mu+3\sigma$ sur 1. Dans cette variante, on peut éventuellement découper en deux l'intervalle $[\mu-3\sigma, \mu+3\sigma]$ en envoyant la moyenne μ sur le milieu 0.5 de l'intervalle, et en appliquant linéairement les deux demi-intervalles.

V.3. Logiciel utilisé pour applique méthode réseaux des neurones

Dans cette partie nous utilisons logiciel des calcules et statistiques nommer **« SPSS 16.0 »** pour permet l'utilisation méthode « **réseaux de neurones**» qui présentera dans le titre suivant.

V.3.1. Présentation logicielle « SPSS 16.0 » [45,46]

Le logiciel SPSS pour Windows constitue un système de traitement de données permettant, à partir de fichiers SPSS ou à partir d'autres types de fichiers (Excel, dBase, FoxPro, MS Access) de générer divers tableaux, graphiques et diagrammes ou encore d'effectuer divers traitements statistiques comme le dépouillement de données, le calcul de diverses mesures de tendance centrale et de dispersion, la construction de tableaux croisés, l'exécution de divers tests statistiques paramétriques et non paramétriques... Ce document n'exploite pas tout le potentiel de ce logiciel mais est néanmoins un guide d'apprentissage suffisamment complet pour la majorité des outils statistiques requis dans les cours de formation, en gestion, sciences comptables, économiques, relations industrielles, études de marché…

V.3.2. Traité La matrice (II)

Dans cette étape nous divisons la **matrice (II)** sur 7 matrices avec la même base des données (125 molécules) pour toutes les 7 matrices, c'est à dire le nombre des lignes ne varient pas pour toutes les 7 matrices mais chaque matrice qui est divisée contribue dans le nombre des colonnes (ou nombres des descripteurs) différents comme suit : 5, 10, 15, 20, 25, 30 et 36 descripteurs ou colonnes.

Nous utilisons **la matrice (II)** dans le logiciel « **SPSS 16.0 »** qui nous permet de présenter **la matrice (II)** sous forme d'EXCEL, et après ça nous utilisons la méthode « **réseaux de neurones** » dans le logiciel pour calculer la corrélation des descripteurs entre eux pour tous les molécules (chaque matrice divisées (les 7 matrices précédentes) sont calculées la corrélation seulement et indévudilement).

L'utilisation de la méthode « **réseaux de neurones** » dans le logiciel **« SPSS 16.0 »** est conditionnée les applications sont les suivantes :

V.3.2.1. Les entrées

Les entées pour **la matrice (II)** ou les 7 matrices divisées sont les descripteurs des molécules **(**Pour chaque matrice des entrées sont les descripteurs correspondants pour toutes les molécules).

V.3.2.2. Les sorties

Les sorties pour **la matrice (II)** sont les classifications des molécules (les six types ou classification d'odeurs différentes), (Pour chaque matrice les sorties sont les six types ou classification des odeurs différentes).

V.4. Etude de la classification de la famille pyrazine par la méthode réseaux de neurones (En utilisant la matrice (II) précédente) :

Dans cette étude nous divisons la base des données pour chaque matrice divisée (Les 7 matrices divisées précédemment) en deux partie : partie d'estimation (70% de base des données) et partie de validation (30% de base des données).

V.4.1. Pourcentages de test d'estimations (training)

Le pourcentage du test d'estimations est 70% de la base des données (c'est à dire 87 molécules qui présentent dans la partie d'estimation (la base des données est 125 molécules) pour réaliser de la méthode « **réseaux de neurones** »).

V.4.2. Pourcentages de test de validation (test)

Le pourcentage du test de validation est 30% de la base des données (c'est à dire 38 molécules qui présentent dans la partie de validation (la base des données est 125 molécules) pour réaliser de la méthode « **réseaux de neurones** ».

V.5. Partie de simulation et modélisation

Après la réalisation des conditions précédentes, on laisse le logiciel « **SPSS 16.0 »** calculé par la méthode « **réseaux de neurones** » pour obtenir les résultats. Dans cette opération on applique cette méthode sur les 7 matrices divisées de **la matrice (II).**

V.5.1. Les étapes d'application de la méthode

Pour utiliser cette méthode (Réseaux de neurones) a partir du logiciel **« SPSS 16.0 »,** nous cliquons sur la fenêtre « Analyze » dans la barre d'outils de ce logiciel, après ça pour choisir la méthode Réseaux de neurones on clique sur la fenêtre « Neural Network » et enfin on clique sur la fenêtre « Radial Basie Function » pour appliquer la méthode Réseaux de neurones pour une seul couche.

V.5.2. Nombre de neurones cachés

Dans cette méthode nous choisirons une seul couche pour appliquer cette méthode, le nombre de neurones cachés, logiciel pris est optimiste car elle donne de meilleurs nombres de neurones cachés qui donnent de meilleurs fonctions pour ce modèle de classification), et on trouve une meilleur matrice qui donne un meilleur modèle de modélisation (matrice 125_36) le nombre de neurones cachés sont sept (7) (figure N°21).

V.6. Résultats et discutions

Nous présentons les résulta des classifications pour test d'estimations et de validations, pour toutes les matrice 5, 10, 15, 20, 25, 30 et 36 descripteurs Dans les tableaux suivants :

Remarque 01 :

Pour plus d'informations sur les codages dans les tableaux des résultats, on résume dans le tableau suivant :

	codages	**classification**
GREEN	0	Non Green
	1	Green
NUTTY	0	Non Nutty
	1	Nutty
BELL-PEPPER	0	Non Bell-Pepper
	1	Bell-Pepper
EARTHY	0	Non Earthy
	1	Earthy
PECASY	0	Non Pecasy
	1	Pecasy
SWEET	0	Non Sweet
	1	Sweet

Tableau N°25 : Présenter les codages et les classifications correspondants

Remarque 02 :

Même Remarque pour les codages de la figure N°21.

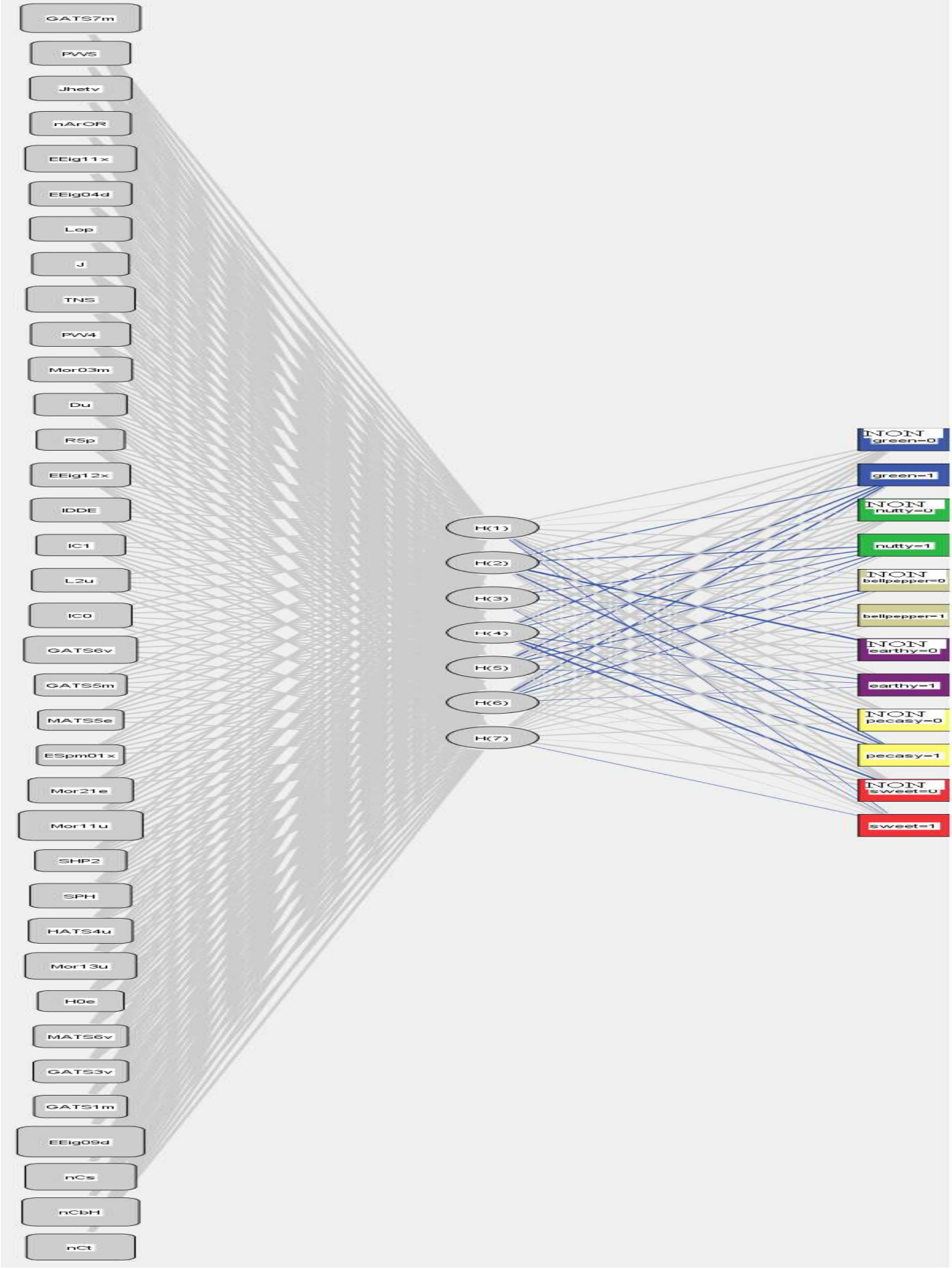


Figure N°21 : schéma représenté modèle utilisé pour simulation (Réseaux de neurones) De matrice 125_36

V.7. Analyses des résultats

Après ces résultats nous remarquons que :

- Les meilleures matrices qui donnent les pourcentages plus élevés (meilleur % d'estimations et de validations) pour les classifications **GREEN, NUTTY, BELL-PEPPER, EARTHY, PECASY ET SWEET (**types d'odeur) pour les tests d'estimations et de validations sont suivants :

GREEN : la meilleure matrice est **matrice 125_05 (**pourcentage d'estimation est **77.6%**, pourcentage de validation est **87.5%** est**).**

NUTTY : la meilleurs matrice est **matrice 125_36 (**pourcentage d'estimation est **92.5%**, pourcentage de validation est **84.4%).**

BELL-PEPPER : la meilleurs matrice est **matrice 125_30 (**pourcentage d'estimation est **88.3%**, pourcentage de validation est **87.1%).**

EARTHY : la meilleure matrice est **matrice 125_36 (**pourcentage d'estimation est **92.5%**, pourcentage de validation est **96.9%).**

PECASY : la meilleure matrice est **matrice 125_36 (**pourcentage d'estimation est **97.8%**, pourcentage de validation est **96.9%).**

SWEET : la meilleure matrice est **matrice 125_36 (**pourcentage d'estimation est **98.9%**, pourcentage de validation est **98.9%).**

- Dans les résultats totaux, nous remarquons que **la matrice 125_36** a donné des meilleurs pourcentages d'estimations et de validations (pourcentage d'estimation est **92.5%** et pourcentage de validation est **89.1%**).

Conclusions générales :

Après cette discution sur les résultats précédents, on résume les notes suivantes :

- la classification de la famille pyrazine pour trois méthodes d'analyse qui sont utilisées dans les chapitres précédents, basé sur les caractères suivants qui correspondent a chaque classification :

1. la ramification qui se trouve sur le niveau de chaque radical pour les radicaux de composée pyrazine, elle influence sur la commande de classification de cette molécules pour la classe GREEN, même l'existence des atomes O et N dans les composées pyrazine donne la classification GREEN est plus possible par exemple la ramification de radical R2 dans la famille pyrazine qui donne la classification de cette molécule est GREEN.

Ce caractère (la ramification des radicaux de la famille pyrazine) est spécifique pour la classe GREEN dans la classification de famille pyrazine.

2. l'existence d'hydrogène dans les radicaux R3 et R4 (R3 = H et R4 = H) pour les radicaux de composées de la famille pyrazine, elle influence sur la classification de cette molécule vers la classe NUTTY.

Ce caractère (R3 = H et R4 = H) est spécifique pour la classe NUTTY dans la classification de famille pyrazine.

3. l'existence de deux radicaux R3 et R4 différents (R3 = H et R4 = CH_3) dans les radicaux des composées de la famille pyrazine, elle influence sur la classification de cette molécule vers la classe BELL-PEPPER.

Ce caractère (deux radicaux R3 et R4 déférents) est spécifique pour la classe BELL-PEPPER dans la classification de famille pyrazine.

4. l'existence de deux radicaux symétriques et non différents (mêmes types des radicaux R3 = R4, R1 = R2, R1 = R3, R1 = R4, R2 = R3, R2 = R4 par exemple (H,H) ou (CH_3, CH_3)….) dans les radicaux des composées la de famille pyrazine, elle influence sur la classification de cette molécule vers la classe EARTHY.

Ce caractère (deux radicaux symétriques et non différents) est spécifique pour la classe EARTHY dans la classification de la famille pyrazine.

- Généralement nous pouvons dire que :
  1. les résultats de classification par la méthode « analyse discriminante » donnent de bons résultats et l'opération est efficace pour les classes NUTTY, BELL-PEPPER et EARTHY (80-98 %). Mais cette méthode donne de mal résultats et l'opération est non efficace pour la classe GREEN (60-85 %).
  2. les résultats de classification par la méthode « logique floue » donnent de bons résultats et l'opération est efficace pour les classes NUTTY et BELL-PEPPER (69-75 %). Mais cette méthode donne de mal résultats et l'opération est non efficace pour la classe GREEN, EARTHY et PECASY (50 %).
  3. les résultats de classification par la méthode « réseaux de neurones » donnent de bons résultats et l'opération est efficace pour les classes NUTTY, BELL-PEPPER, EARTHY et PECASY (84-97 %). Mais cette méthode donne de mal résultats et l'opération est non efficace pour la classe GREEN (77-87 %).

- La comparaison entre les trois méthodes d'analyse (analyse discriminante, logique floue, réseaux de neurones) pour la classification des molécules de famille pyrazine avec d'odeurs correspondantes différentes, en utilisant la méthode « **analyse discriminante** » on obtient un plus grand rendement et un meilleur classement de la famille pyrazine par rapport aux autres méthodes utilisées (logique floue, réseaux de neurones) pour même types d'odeurs ou même types de classification (80-98 %).
- Les commentaires statistiques sur les résultats de classification présente la correspondance et la convergence des descripteurs des composées pyrazine dans un intervalle bien défini donnons la priorité pour la classification de cette composées pyrazine c'est à dire la correspondance et la convergence de ces descripteurs dans un même intervalle pour des nombres de composées pyrazine donnons la classification de cette composées pyrazine vers la même classe et une seule classe.

Bibliographies

[1] **Enrico Riboni**, Ingénieur EPFL en mécanique et **Myriam Robert**, Ingénieur EPFL en génie rural, « **Traitement des odeurs par ozonisation dans les stations d'épuration des eaux usées** » novembre 2000.

[2] **Anne-Marie** GOURONNEC Ingénieur à l'IRSN « **Analyses olfactométriques ou mesure des odeurs par analyse sensorielle** » (Institut de radioprotection et de sûreté nucléaire) © Techniques de l'Ingénieur P 446 1 – 21.

[3] **J.-P. Dumont** « **Arômes et saveurs des aliments** ».

[4] **Karen J. Rossiter**, "**Structure - Odor Relationships",** Chem. Rev. 1996, 96, 3201-3240

[5] **Altervino** – 2006 « **Cours de dégustation** ».

[6] **Division Environnement Amiens « NOTE sur les odeurs** » 15 octobre 2004 MH/MD – 2004-1019

[7] **Masuda, H.; Mihara, S.** "**Olfactive Properties of Alkylpyrazines and 3-Substituted 2-Alkylpyrazines**." *J. Agric. Food Chem.* **1988**, *36*, 584-587.

[8] **Uwe J. Meierhenrich**, **Jérôme Golebiowski**, **Xavier Fernandez** et **Daniel Cabrol-Bass** « **De la molécule à l'odeur Les bases moléculaires des premières étapes de l'olfaction, l'actualité chimique** » - août - septembre 2005 - n° 289 p 29 – 40.

[9] **Bettina Wailzer**, **Johanna Klocker**, **Gerhard uchbauer**, **Gerhard Ecker**,and **Peter Wolschann "Prediction of the Aroma Quality and the Threshold Values of Some Pyrazines"**Using Artificial Neural Networks **,** *J. Med. Chem.* **2001,** *44*, 2805 - 2813 **2805.**

[10] **Masuda, H**.; **Mihara, S,** "**Olfactive Properties of Alkylpyrazines and 3- Substituted 2- Alkylpyrazines**". *J. Agric. Food Chem*. **1988**, *36*, 584-587.

[11] **Parliament, T. H**.; **Epstein, M. F**, "**Organoleptic Properties of Some Alkyl-Substituted Alkoxy- and Alkylthiopyrazines**". *J. Agric. Food Chem*. **1973**, *21*, 714-716.

[12] **Masuda, H**.; **Mihara, S**, "**Synthesis of Alkoxy-, (Alkylthio), Phenoxy-, and (Phenylthio) pyrazines and Their Olfactive Properties**", *J. Agric. Food Chem*. **1986**, *34*, 377-381.

[13] **Shibamoto, T**, "**Odor Threshold of Some Pyrazines**". *J. Food Sci.* **1986**, *51*, 1098-1099.

[14] **Pittet, A. O**.; **Hruza, D. E,** "**Comparative Study of Flavor Properties of Thiazole Derivatives**". *J. Agric. Food Chem.* **1974**, *22*, 264 - 269.

[15] **Mihara, S**.; **Masuda, H**, "**Structure-Odor Relationships for Disubstituted Pyrazines**". *J. Agric. Food Chem*. **1988**, *36*, 1242 - 1247.

[16] **Seifert, R. M**.; **Buttery, R. G**.; **Guadagni, D. G**.; **Black, D. R**.; **Harris; J. G**. "**Synthesis of some 2-Methoxy-3-Alkylpyrazines with StrongBellpepper-Like Odors**". *J. Agric. Food Chem.* **1970**, *18*, 246-249.

[17] **Parliment, T. H**.; **Epstein, M. F**. "**Organoleptic Properties of some alkylsubstituted alkoxy- and alkylthiopyrazines**". *J. Agric. Food Chem*. **1973**, *21*, 714-716.

[18] **Pittet, A. O**.; **Hruza, D. E**. "**Comparative Study of Flavor Properties of Thiazole Derivatives**". *J. Agric. Food Chem.* **1974**, *22*, 264-269.

[19] **Takken, H. J**.; **Van der Linde**, **M. L.; Boelens**, **M.; Van Dort, J**. "**Olfactive Properties of a Number of Polysubstituted Pyrazines**". *J. Agric. Food Chem.* **1975**, *23*, 638-642.

[20] **Masuda, H**.; **Mihara, S**. "**Synthesis of Alkoxy-, (Alkylthio)-, Phenoxy-, and (Phenylthio) pyrazines and Their Olfactive Properties**". *J. Agric. Food Chem*. **1986**, *34*, 377-381.

[21] **Shibamoto, T**. "**Odor Threshold of Some Pyrazines**". *J. Food Sci.* **1986**,

51, 1098-1099.

[22] **Masuda, H**.; **Mihara, S**. "**Olfactive Properties of Alkylpyrazines and 3-Substituted 2-Alkylpyrazines**". *J. Agric. Food Chem*. **1988**, *36*, 584 - 587.

[23] **Mihara, S**.; **Masuda, H**. "**Structure-Odor Relationships for Disubstituted Pyrazines**". *J. Agric. Food Chem*. **1988**, *36*, 1242-1247.

[24] **Mihara, S**.; **Masuda, H**.; **Tateba, H**.; **Tuda, T**. "**Olfactive Properties of 3-substituted 5-Alkyl-2-methylpyrazines**". *J. Agric. Food Chem*. **1991**, *39*, 1262-1264.

[25] **Boelens, M. H**.; **Van Gemert, L. J**. "**Structure-Activity Relationships of Natural Volatile Nitrogen Compounds**". *Perfum. Fla*V*our* **1995**, *20*, 63-76.

[26] **Grosch, W**.; **Wagner, R**.; **Czerny, M**.; **Bielohradsky, J**. "**Structure-odor activity relationships of earthy smelling alkylpyrazines**". *Z. Lebensm. Unters. Forsch.* **1999**, *208*, 308-316.

[27] Help **de logiciel Hyperchem 07.**

[28] **ZIGHMI Souad**. « **Caractérisation Physico-Chimique des conducteurs moléculaires à base de TTF-TCNQ** ». mémoire de magistaire, chapitre 2, p 46-50

[29] **Dragon**, Ishtar Gate of Babylon, Berlin Pergamon Museum Copyright, **DRAGON ersion 5 1997-2006** TALETE srl – Milano, Italy Developed by: **Viviana Consonni, Andrea Mauri and Manuela Pavan** ITALY DRAGON - © TALETE srl, 2006.

[30] **DRAGON** for Windows and Linux 2006 Copyright © 2003-2006 by **Talete srl, Milano**, Italy.

[31] **Riadh ben messaoud**, "**analyse discriminante**" institut universitaire de technologie lumière, licence C.E.STAT, Mai 2006.

[32] **Dillon, W.R**.(1979) "**the performance of the linear descriminant function in nonoptimal situations and the estimation of classification error rates: a review**

of recent findings" JMR 16, 370-381.

[33] **Giltow, H.S**.(1979)"**Descrimination procedures for the analysis of nominally scaled data sets**"JMR 16, 387-393.

[34] **Gilbert A. Churchill**, "**Marketing Research, Methodological Foundations**", 5e Ed., Dryden Press,1991.

[35] **Vedrine J.-P**. "**Le traitement des données en marketing**", Ed. Organisation, Paris, 1991.

[36] **Thiery** , "**XLSTAT**" , Brochure FR, Jan 2006.

[37] Help de **XLSTST**.

[38] **L.BAGHLI** « **contribution a la commande de la machine asynchrone, utilisation de logique floue, résaux de neurones et les algorithmes génétiques** », thèse doctorat NANCY, 2003.

[39] **F.Chenrie** et **F.Guely**, « **cahier technique N° 191 : la logique floue** », groupe Schneider, 1998.

[40] **Mohamed Kissi** · **Mohammed Ramdani Mustapha Tollabi** · **Driss Zakarya** "**Determination of fuzzy logic membership functionsusing genetic algorithms: application to structure–odor modelling**"
J Mol Model (2004) 10:335–341DOI 10.1007/s00894-004-0200-2.

[41] Help de logiciel MATLAB 7.1

[42] **G.Dreyfus**, **J-M.Martinez**, **M.Samuehids**, **M.B.Gordon**, **F.Badran**, **S.Thiria**, **B.Héranlt**, « **Réseaux des neurones, méthodologies, et application** » 2eme édition sous la direction de Gérard Dreyfus, Eyrolles.

[43] **DRISS ZAKARYA**, **DRISS CHERQAOUI**, **M'HAMED ESSEFFAR**,**DIDIER VILLEMIN3 AND JEAN-MICHEL CENSE,** « **APPLICATION OF NEURAL ETWORKS TO STRUCTURE–SANDALWOOD ODOUR RELATIONSHIPS** »
Département de Chimie, Faculté des Sciences et Techniques, JOURNAL OF PHYSICAL ORGANIC CHEMISTRY, VOL. 10, 612–622 (1997)

[44] **Johanna Klocker**, **Bettina Wailzer**, **Gerhard Buchbauer**, and **Peter Wolschann**, "**Bayesian Neural Networks for Aroma Classification**", J. Chem. Inf. Comput. Sci. 2002, 42, 1443-1449

[45] **Hubert Laforge,** "**Analyse multivariée pour les sciences sociales et biologiques avec applications des logiciels BMP, BMDP, SPSS, SAS**", Ed. Etudes Vivantes, Montréal, 1981.

.

[46] **Ludovic LE MOAL** © 2002 « **L'Analyse Discriminante sous SPSS** »

[47] **Stéphane Tufféry** , « **DATA MINING & STATISTIQUE DÉCISIONNELLE** »
18/12/2006, http://data.mining.free.fr