وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR- ANNABA UNIVERSITY

UNIVERSITE BADJI MOKHTAR-ANNABA

جامعة باجي مختار-عنابة

FACULTE DES SCIENCES

DEPARTEMENT DE CHIMIE

MEMOIRE

Présenté en vue de l'obtention du diplôme de ***MAGISTER***

Option : Chimie analytique

Thème

Etude structure-activité d'une série de benzènes substitués comportant divers groupements

Présenté par : BOUCHAMA Fateh

RAPPORTEUR MESSADI Djelloul Pr UBMA

DEVANT LE JURY

PRESIDENTE	M^{me} LARKEM Hamama	MC	UBMA
EXAMINATEURS	M. GUERSI Nour-eddine	MC	UBMA
	M^{me} BIDJOU-HAIOUR Chahra	MC	UBMA
	M^{me} NOUAR Leila	MC	Université 8 mai 1945 GUELMA

Année : 2008/2009

Le travail présenté dans ce mémoire a été réalisé au Laboratoire de Sécurité Environnementale et Alimentaire de l'Université de Annaba- BADJI Mokhtar, sous la direction du Professeur Djelloul MESSADI. Je tiens à lui exprimer toute ma gratitude pour l'aide désintéressée et constante qu'il m'a apportée pour l'exécution de ce projet.

J'exprime également ma profonde gratitude au Docteur Nour-eddine GUERSI qui m'a initié aux « réseaux de neurones » avec patience, et une disponibilité permanente. Je le remercie sincèrement pour avoir accepté de juger le travail final et de participer à mon jury.

M^me^ le docteur Hamama LARKEM a pris sur son temps pour examiner le mémoire final et présider mon jury. Je veux lui dire tout mon respect.
Les Docteurs, mesdames Leila NOUAR, Chahra BIDJOU-HAIOUR ont accepté le travail ingrat de critiquer le travail présenté. Sans arrière pensée, elles ont droit à toute ma sympathie.

Je ne peux pas terminer sans évoquer le rôle de mes camarades de Laboratoire pour mener à bonne fin ce travail. Je leur souhaite beaucoup de courage.

Toutes mes pensées vont aussi à ceux qui m'ont encouragé par des gestes d'amitié dont je leur serais toujours reconnaissant. Merci à tous.

A ma famille,

Qui m'a toujours encouragé et supporté,

À ceux et à celles,

Qui ont cru en moi,

Je dédie ce travail.

Sommaire

ملخص :

ان النموذج QSAR الهجين (GA/RNA) طور من اجل التنبؤ بالتسمم المائي, المعطيات الخاصة ب: 138 مستبدل بنزيني, قسمت الى مجموعتين مختلفتين تحتويان على : 110 عناصر لحساب و اختبار النموذج, و 28 عنصر من أجل تثبيته الخارجي.

لقد قمنا بحساب المواصفات الجزئية النظرية باستعمال برامج المحاكاة الجزئية التجارية, وحددنا حجم النموذج باستعمال معادلة FIT de KUBINYI, و تم اختيار المواصفات باستعمال الخوارزميات الوراثية.

قيم المعايير الاحصائية (R^2, Q^2, Q^2ext, SDEC, SDEP, SDEPext) المتحصل عليها تثبت صلابة النموذج المطور.

كلمات مفتاحية: مستبدلات بنزينية, التسمم المائي, مواصفات جزئية نظرية, نموذج هجين(GA/RNA).

Résumé :

Un modèle QSAR hybride (GA/RNA) a été développé pour la prédiction de la toxicité aqueuse. Les données concernant 138 benzènes substitués, ont été éclatés en deux sous-ensembles disjoints comprenant respectivement 110 éléments pour le calcul et le test du modèle, et 28 éléments pour sa validation externe.

Des descripteurs moléculaires théoriques ont été calculés en utilisant des logiciels de modélisation moléculaire du commerce. La taille du modèle a été déterminée en optimisant le FIT de Kubinyi, et la sélection des descripteurs réalisée par algorithme génétique.

Les valeurs des paramètres statistique (R^2, Q^2, Q^2_{ext}, SDEC, SDEP, $SDEP_{ext}$) obtenues attestent de la pertinence du modèle développé.

Mots-clés : Benzènes substitués – Toxicité aquatique – Descripteurs moléculaires théoriques – Modèle hybride AG/RNA.

Summary:

An hybrid (GA/ANN) QSAR model was developed for the prediction of aqueuse toxicity. The data concerning 138 substituted benzenes were separated into two disjoined subsets including, respectively 110 elements for calculus and test of the model, and 28 elements for its external validation.

Theoretical molecular descriptors were calculated by using commercially available software of molecular modeling. The size of the model was determined by optimizing the FIT of Kubinyi, and the selection of the descriptors realized by genetic algorithm.

The values of the statistical parameters (R^2, Q^2, Q^2ext, SDEC, SDEP, SDEPext) obtained attest relevance of the developed model.

Key words: Substituted benzenes – Aqueous toxicity – Theoretical molecular descriptors – Hybrid GA/ANN model.

Liste des figures

Liste des Tableaux

-INTRODUCTION GENERALE :

Durant les deux dernières décennies des modèles QSAR et SAR ont été développés et appliqués dans l'estimation d'une large gamme de propriétés chimiques et d'activités biologiques. La première étape consiste à sélectionner un ensemble d'apprentissage (ou d'essai) significatif avec des données bien établies. Par la suite, des paramètres structuraux appropriés (descripteurs ou propriétés moléculaires) sont choisis de façon à établir la meilleure corrélation QSAR possible. L'analyse statistique de l'ensemble des résultats obtenus permet alors d'aboutir au modèle ou à la corrélation requis.

Before applying this model to the estimate of the properties or activities wished to a great scale, one must validate it on representative a test unit.

The choice of the whole of training plays a pivot role in the derivation of the awaited model, insofar as the representativeness and the size of this unit, as well as the quality of the selected data, affect the later stages of the construction of the model.

A range of molecular descriptors were used successfully in the construction of significant models QSAR. The physico-chemical descriptors as the theoretical descriptors have advantages and disadvantages. Thus, one of the advantages of the theoretical descriptors resides in their availability, in the sense that they can be easily calculated for all the types of organic compounds, however that their difficulty of interpretation often constitutes their disadvantage. On the other hand, the limitation of the physicochemical descriptors and the electronic parameters arises in their occasional availability which can restrict their application considerably. Like, in addition, these descriptors have a physico-chemical base good established and easily perceptible, their interpretation is easier and more direct.

A central stage in the derivation of the models structure/activity relates to the statistical analysis whose constraints differ according to whether L ' one seeks a quantitative model (QSAR) or qualitative (SAR).

The last phase of develops lies of the model relates to the test and the validation using adapted a test unit. L be limiting of application of the model will depend largely on this final stage; these limits will be all the more wide as the whole of test is broad, and the condi tions experimental varied.

For the estimate of the risk of toxicity that can present the organic pollutants, particularly with respect to the organizations watery S and mammals, the data of acute toxicity are necessary.

In spite of the data banks which relates to, or which includes, of the files of toxicity as in the "Registry of Toxic Effects of Chemical Substances" (RTECS), "Aquatic Toxicity

Information Retrieval" (AQUIRE), "Environmental Chemicals Data and Information Network" (ECDIN), it remains less true about it than a small fraction, only, great number of polluting and/or toxic compounds is covered E by reliable and confirmed experimental data of toxicity. Thus, the derivation and the estimate of the data of "acute" toxicity using validated models QSAR, constitute more and more, a significant complementary tool.

A gence A méricaine of Environmental protection (D U lute, Minnes O your) established a program for the generation of the data of high quality on the toxicity of the fish, which were published in a series of volumes. An E started from these data will be used as a basis for study QSAR presented in this work.

In addition to one general introduction and of a conclusion, our memory comprises three distinct parts:

A bibliographical study (CHAP.I) which specifies the risks incurred by an exposure to benzene and its derivatives of substitution; we also define the amount lethal 50 in it (DL 50) and its use. In the CHAP II, we present the collected data.

In a second part (CHAP.III and IV), we developed all that milked with the pretreatment of the molecules for the calculation of the theoretical molecular descriptors. We also developed to with it the basic theoretical knowledge used throughout this work for the development and the evaluation of the model.

Lastly, in a last part (CHAP. V), we present and discuss the calculated model.

Etude bibliographique.

I-1 Le benzène et ses dérivés:

I-1-1Introduction et aperçu historique:

Le benzène, monocyclic aromatic hydrocarbon, is a carcinogenic liquid. Before the years 1920, the benzene was frequently used like industrial solvent, particulièr E lies to degrease metals. When its toxicity became obvious, it is replaced by A U very solvents for the applications requiring a direct exposure of the user. Be N zene is used in major part like intermediary in the synthesis of others made up chim I ques.

A great number of very significant chemical compounds in industry are obtained by replacing one or more hydrogen atoms of benzene by other groupings fonctio N nels.

are used in the manufacture of polymers and the plastics; the phenol, intervenes in that of the resins and the adhesives.Less quantities of benzene are announced in the manufacture of tires, lubricants, dyes, detergents, méd I caments, explosives or pesticides.

Toluene is used to raise the octane number in the fuels. It is also used as solvent for paintings. One makes use of it like product of departure for various industrial processes: the synthesis of rubber, phenol, the TNT, the toluene diisocyanate necessary for obt E to nir the polyurethane foam. One also makes use of it in printing works, the adhesives, the L have ques, and the tanning of leather. In years 1980, ethylbenzene, intermediary in the prép ration of styrene has, represented the principal derivative of benzene.

I-1-3 Influences benzene and its derivatives of substitution on the environment:

The carcinogenic properties of benzene come from what it behaves as an intercalating agent (i.e. it slips between the nucleotidic bases of the acids nuclé I ques, of which the ADN, causing replication and/or misreadings). There are other agents intercalating (like bromide of éthidium, or Study Bureau, used in experimental biology to mark the ADN, in particular during electrophoreses). All the plane compounds are not however carcinogenic. The benzoic acid, for example, very near to the Be N zene, and whose combined base is absolutely plane, is not carcinogenic (it is used as conservative in various types of sodas). In the same way, phenylalanine, an amino acid which comprises a grouping phenyl (a benzene cycle), is not carcinogenic.

The intoxication by benzene alone bears the name of benzenism; that by benzene or its derivatives (toluene, xylene…) bears the name of benzolism. The inhalation of a very high benzene rate can cause death, while high rates can cause somn O lences, giddinesses, an acceleration of the rate of heartbeat, headaches, trembl E ments, confusion or the loss of consciousness. An exposure from five to ten minutes toa benzene rate in the air of 2 % approximately is enough to result in death. *The lethal amount by I N management is 50 mg/kg*. The ingestion

of foods or drinks containing of the rates él E vés of benzene can cause vomiting, an irritation of the stomach, giddinesses, somnolences, convulsions, an acceleration of the rate of heart-beat, even death. The E F fet principal of a chronic exposure to benzene is a damage of the marrow O S seuse, which can cause a decrease of the rate of red globules in blood and an anaemia. It can also cause bleedings and a weakening of system I m munitaire. The effect of benzene on the fertility of the man or the good development of the foetus is not known. Lastly, the benzene is recognized as being a carcinogenic substance.

Xylene has a harmful effect on the brain. Levels of exposures raised for perished O of very short can involve headaches, a defect of coordination of driven S cles, giddinesses, the confusion and losses of the direction of balance. Exposures torates él E vés for short periods of time can also cause an irritation of the skin, eyes, nose and throat, difficulties of breathing, problems pulmona I LMBO, an increase in the reaction times, losses of memory, irritations of E S tomac and deteriorations of the operation of the liver and kidneys. Very high rates of exposure can involve the loss of conscience even death. Studies on animals showed that high xylene concentrations increase the number of year I evils still-born children, as well as delays of growth and development. In much of case, these same conce N trations also has negative effects on health of the mothers. The effect of expos I tions of the mother to weak xylene concentrations on the foetus is not known per hour has C tuelle.

Aniline is a very toxic substance which must be handled with precaution. An E X position with high concentrations can be mortal. She can be absorbed by inhal has tion, ingestion and contact with the skin, *including in form vapor*.

Chlorobenzenes are toxic substances which must be handled with precautions. The L I mite of professional exposure is fixed in France at 10 ppm, that is to say 46 mg/m ^{3} of air.

Nitrobenzines can cause serious poisonings by ingestion, inhalation or contact with the skin. They react with the haemoglobin of blood and prevent it from reacting with oxygen. They can also involve disorders of the central nervous system, Ca U sant a feeling of weakness, headaches and vomiting. A high rate of nitr O benzene can result in death in less than one hour; moreover, its toxic effect is exacerbated by the alcohol catch.

I-2- Proportions lethal 50(DL $_{50}$):

I-2-1 Definition:

The lethal amount 50 or DL50 (***English LD50 for Lethal Proportions 50***) or CL50 (conce N lethal tration 50) are a quantitative indicator of the toxicity of a substance.

This indicator measures the amount of substance causing the death of 50 % of one animal population data (often of mouse or of rats) under precise conditions of experimentation.

The amount mortal minimum in the animal, or proportions lethal, is always delicate with déte R to undermine in a precise way. One prefers to establish the definite DL 50 like "the statistical estimate of a single amount of product supposed to kill 50 % of the animals" in experimentation.

The test is usually practised on 5 or 6 batches of animals, generally the rat. Each animal of the same batch receives an identical amount (single amount) of the substance to be tested, but the administered dose is different from one batch to another, so that the percentage in mortality varies between 0 and 100. The route of administration is that which will be used in private clinic if it is about a m 3rd dicament (oral way, injection, etc), or that by which the substance will be able to penetrate in the organization if it is about a chemical (oral way, inhalation, transcutanée way, etc. .>.>.).

.>After the administration the animals are observed during 14 days during which the clinical examinations are frequent. The moment and the circumstances of death are carefully noted. The animals still in life at the end of the test are sacrificed. All the animals (deaths in the course of E S sai and sacrificed at the end of the test) are the subject of an autopsy.

One builds then the curve giving the percentage of mortality according to the log has rithme amount. It is a curve in "S", known as curve of Trévan, which can be linearized by suitable means. One deduces from them the DL50 (expressed out of Mg per kg of body weight), which one calculates also the standard deviation.

For a substance managed by oral way one considers that:

- if the DL 50 is < 5 mg/kg, the product is extremely toxic;

- if the DL 50 lies between 5 and 50 mg/kg, the product is very toxic;

- if the DL 50 lies between 50 and 500 mg/kg, the product is toxic;

- if the DL 50 lies between 0,5 and 5 g/kg, the product is not very toxic;

- if the DL 50 is > 5 g/kg, the product is not toxic or very little.

I-2-2 Forms of toxicity:

Toxic effects of the chemicals, and how they cause these effects, EP U wind being divided in several ways. For the studies of properties by QSAR, the effects tox I ques were di-

vided into three main categories: toxicity receiver negotiated; acute toxicity not réce p tor negotiated; and effects supposed on human health **[1].**

I-2-3 Use of the lethal amount 50:

One manages G énéralemen T the poison with animals divided into several groups and this, with sufficient increasing amounts to obtain a percentage of mortality spreading out between 0 and 100 %. the effect of a substance is, overall, inversely proportional to the mass of the animal with which it is managed, this is why it is measured in G / kg. In gén 3rd ral, if immediate toxicity is similar at all the types of animals, it will be probabl E lies similar at the human ones. When the DL_{50} are different at various species anim has them, one must make approximations and assumptions during the estimate of the amount probable mortal at the human ones.

I-2-4 Identification of toxicity:

The DL50 is used to measure all the toxicity of a substance, measures which is carried out via qualitative studies (nonmeasurable) and quantitative (measurable of which the DL_{50}).

The DL_{50} is often used as departure with the studies of toxicity because it provides a minimum of knowledge by identifying them symptoms of intoxication and the toxic amount. It should despite everything be considered with prudence because it is often a preliminary study (Pr E mière analyzes) which can be influenced by several factors, such it animal species, it sex, it age, moment of the day, etc.

It however has a limit value, because it relates to only mortality, from where appearance of values like the IC_{50}.

There are other methods of study of toxicity, for example the tests of irritation of skin and of corrosion of eyes, which generally form part of a program of evaluation toxicolog I that.

I-2-5 Identification of the pathogenic capacity:

The DL_{50} is one of the two data being used to measure the capacity pathogenic of a germ. The second data being the Infecting Minimal Amount (DM

I-1-2 Utilisation du benzène et de ses dérivés:

Les dérivés du benzène, monocyclic aromatic hydrocarbon, is a carcinogenic liquid. Before the years 1920, the benzene was frequently used like industrial solvent, particulièr E lies to degrease metals. When its toxicity became obvious, it is replaced by A U very solvents for

the applications requiring a direct exposure of the user. Be N zene is used in major part like intermediary in the synthesis of others made up chim I ques.

A great number of very significant chemical compounds in industry are obtained by replacing one or more hydrogen atoms of benzene by other groupings fonctio N nels.

are used in the manufacture of polymers and the plastics; the phenol, intervenes in that of the resins and the adhesives.Less quantities of benzene are announced in the manufacture of tires, lubricants, dyes, detergents, méd I caments, explosives or pesticides.

Toluene is used to raise the octane number in the fuels. It is also used as solvent for paintings. One makes use of it like product of departure for various industrial processes: the synthesis of rubber, phenol, the TNT, the toluene diisocyanate necessary for obt E to nir the polyurethane foam. One also makes use of it in printing works, the adhesives, the L have ques, and the tanning of leather. In years 1980, ethylbenzene, intermediary in the prép ration of styrene has, represented the principal derivative of benzene.

I-1-3 Influences benzene and its derivatives of substitution on the environment:

The carcinogenic properties of benzene come from what it behaves as an intercalating agent (i.e. it slips between the nucleotidic bases of the acids nuclé I ques, of which the ADN, causing replication and/or misreadings). There are other agents intercalating (like bromide of éthidium, or Study Bureau, used in experimental biology to mark the ADN, in particular during electrophoreses). All the plane compounds are not however carcinogenic. The benzoic acid, for example, very near to the Be N zene, and whose combined base is absolutely plane, is not carcinogenic (it is used as conservative in various types of sodas). In the same way, phenylalanine, an amino acid which comprises a grouping phenyl (a benzene cycle), is not carcinogenic.

The intoxication by benzene alone bears the name of benzenism; that by benzene or its derivatives (toluene, xylene…) bears the name of benzolism. The inhalation of a very high benzene rate can cause death, while high rates can cause somn O lences, giddinesses, an acceleration of the rate of heartbeat, headaches, trembl E ments, confusion or the loss of consciousness. An exposure from five to ten minutes toa benzene rate in the air of 2 % approximately is enough to result in death. *The lethal amount by I N management is 50 mg/kg*. The ingestion of foods or drinks containing of the rates él E vés of benzene can cause vomiting, an irritation of the stomach, giddinesses, somnolences, convulsions, an acceleration of the rate of heartbeat, even death. The E F fet principal of a chronic exposure to benzene is a damage of the marrow O S seuse, which can cause a decrease of the rate of red globules in blood and an anaemia. It can also cause bleedings and a weakening of system I m munitaire. The effect of

benzene on the fertility of the man or the good development of the foetus is not known. Lastly, the benzene is recognized as being a carcinogenic substance.

Xylene has a harmful effect on the brain. Levels of exposures raised for perished O of very short can involve headaches, a defect of coordination of driven S cles, giddinesses, the confusion and losses of the direction of balance. Exposures torates él E vés for short periods of time can also cause an irritation of the skin, eyes, nose and throat, difficulties of breathing, problems pulmona I LMBO, an increase in the reaction times, losses of memory, irritations of E S tomac and deteriorations of the operation of the liver and kidneys. Very high rates of exposure can involve the loss of conscience even death. Studies on animals showed that high xylene concentrations increase the number of year I evils still-born children, as well as delays of growth and development. In much of case, these same conce N trations also has negative effects on health of the mothers. The effect of expos I tions of the mother to weak xylene concentrations on the foetus is not known per hour has C tuelle.

Aniline is a very toxic substance which must be handled with precaution. An E X position with high concentrations can be mortal. She can be absorbed by inhal has tion, ingestion and contact with the skin, *including in form vapor*.

Chlorobenzenes are toxic substances which must be handled with precautions. The L I mite of professional exposure is fixed in France at 10 ppm, that is to say 46 mg/m ^{3} of air.

Nitrobenzines can cause serious poisonings by ingestion, inhalation or contact with the skin. They react with the haemoglobin of blood and prevent it from reacting with oxygen. They can also involve disorders of the central nervous system, Ca U sant a feeling of weakness, headaches and vomiting. A high rate of nitr O benzene can result in death in less than one hour; moreover, its toxic effect is exacerbated by the alcohol catch.

I-2- Proportions lethal 50(DL $_{50}$):

I-2-1 Definition:

The lethal amount 50 or DL50 (***English LD50 for Lethal Proportions 50***) or CL50 (conce N lethal tration 50) are a quantitative indicator of the toxicity of a substance.
This indicator measures the amount of substance causing the death of 50 % of one animal population data (often of mouse or of rats) under precise conditions of experimentation.

The amount mortal minimum in the animal, or proportions lethal, is always delicate with déte R to undermine in a precise way. One prefers to establish the definite DL 50 like

"the statistical estimate of a single amount of product supposed to kill 50 % of the animals" in experimentation.

The test is usually practised on 5 or 6 batches of animals, generally the rat. Each animal of the same batch receives an identical amount (single amount) of the substance to be tested, but the administered dose is different from one batch to another, so that the percentage in mortality varies between 0 and 100. The route of administration is that which will be used in private clinic if it is about a m 3rd dicament (oral way, injection, etc), or that by which the substance will be able to penetrate in the organization if it is about a chemical (oral way, inhalation, transcutanée way, etc. .>.>.).

.>After the administration the animals are observed during 14 days during which the clinical examinations are frequent. The moment and the circumstances of death are carefully noted. The animals still in life at the end of the test are sacrificed. All the animals (deaths in the course of E S sai and sacrificed at the end of the test) are the subject of an autopsy.

One builds then the curve giving the percentage of mortality according to the log has rithme amount. It is a curve in "S", known as curve of Trévan, which can be linearized by suitable means. One deduces from them the DL50 (expressed out of Mg per kg of body weight), which one calculates also the standard deviation.

For a substance managed by oral way one considers that:

- if the DL 50 is < 5 mg/kg, the product is extremely toxic;

- if the DL 50 lies between 5 and 50 mg/kg, the product is very toxic;

- if the DL 50 lies between 50 and 500 mg/kg, the product is toxic;

- if the DL 50 lies between 0,5 and 5 g/kg, the product is not very toxic;

- if the DL 50 is > 5 g/kg, the product is not toxic or very little.

I-2-2 Forms of toxicity:

Toxic effects of the chemicals, and how they cause these effects, EP U wind being divided in several ways. For the studies of properties by QSAR, the effects tox I ques were divided into three main categories: toxicity receiver negotiated; acute toxicity not réce p tor negotiated; and effects supposed on human health **[1].**

I-2-3 Use of the lethal amount 50:

One manages G énéralemen T the poison with animals divided into several groups and this, with sufficient increasing amounts to obtain a percentage of mortality spreading out between 0 and 100 %. the effect of a substance is, overall, inversely proportional to the mass of the animal with which it is managed, this is why it is measured in G / kg. In gén 3rd ral, if immediate toxicity is similar at all the types of animals, it will be probabl E lies similar at the human ones. When the DL_{50} are different at various species anim has them, one must make approximations and assumptions during the estimate of the amount probable mortal at the human ones.

I-2-4 Identification of toxicity:

The DL50 is used to measure all the toxicity of a substance, measures which is carried out via qualitative studies (nonmeasurable) and quantitative (measurable of which the DL_{50}).

The DL_{50} is often used as departure with the studies of toxicity because it provides a minimum of knowledge by identifying them symptoms of intoxication and the toxic amount.
It should despite everything be considered with prudence because it is often a preliminary study (Pr E mière analyzes) which can be influenced by several factors, such it animal species, it sex, it age, moment of the day, etc.
It however has a limit value, because it relates to only mortality, from where appearance of values like the IC_{50}.
There are other methods of study of toxicity, for example the tests of irritation of skin and of corrosion of eyes, which generally form part of a program of evaluation toxicolog I that.

I-2-5 Identification of the pathogenic capacity:

The DL_{50} is one of the two data being used to measure the capacity pathogenic of a germ. The second data being the Infecting Minimal Amount (DM

monocyclic aromatic hydrocarbon, is a carcinogenic liquid. Before the years 1920, the benzene was frequently used like industrial solvent, particulièr E lies to degrease metals. When its toxicity became obvious, it is replaced by A U very solvents for the applications requiring a direct exposure of the user. Be N zene is used in major part like intermediary in the synthesis of others made up chim I ques.

A great number of very significant chemical compounds in industry are obtained by replacing one or more hydrogen atoms of benzene by other groupings fonctio N nels.

are used in the manufacture of polymers and the plastics; the phenol, intervenes in that of the resins and the adhesives.Less quantities of benzene are announced in the manufacture of tires, lubricants, dyes, detergents, méd I caments, explosives or pesticides.

Toluene is used to raise the octane number in the fuels. It is also used as solvent for paintings. One makes use of it like product of departure for various industrial processes: the synthesis of rubber, phenol, the TNT, the toluene diisocyanate necessary for obt E to nir the polyurethane foam. One also makes use of it in printing works, the adhesives, the L have ques, and the tanning of leather. In years 1980, ethylbenzene, intermediary in the prép ration of styrene has, represented the principal derivative of benzene.

I-1-3 Influences benzene and its derivatives of substitution on the environment:

The carcinogenic properties of benzene come from what it behaves as an intercalating agent (i.e. it slips between the nucleotidic bases of the acids nuclé I ques, of which the ADN, causing replication and/or misreadings). There are other agents intercalating (like bromide of éthidium, or Study Bureau, used in experimental biology to mark the ADN, in particular during electrophoreses). All the plane compounds are not however carcinogenic. The benzoic acid, for example, very near to the Be N zene, and whose combined base is absolutely plane, is not carcinogenic (it is used as conservative in various types of sodas). In the same way, phenylalanine, an amino acid which comprises a grouping phenyl (a benzene cycle), is not carcinogenic.

The intoxication by benzene alone bears the name of benzenism; that by benzene or its derivatives (toluene, xylene…) bears the name of benzolism. The inhalation of a very high benzene rate can cause death, while high rates can cause somn O lences, giddinesses, an acceleration of the rate of heartbeat, headaches, trembl E ments, confusion or the loss of consciousness. An exposure from five to ten minutes toa benzene rate in the air of 2 % approximately is enough to result in death. *The lethal amount by I N management is 50 mg/kg*. The ingestion of foods or drinks containing of the rates él E vés of benzene can cause vomiting, an irritation of the stomach, giddinesses, somnolences, convulsions, an acceleration of the rate of heartbeat, even death. The E F fet principal of a chronic exposure to benzene is a damage of the marrow O S seuse, which can cause a decrease of the rate of red globules in blood and an anaemia. It can also cause bleedings and a weakening of system I m munitaire. The effect of benzene on the fertility of the man or the good development of the foetus is not known. Lastly, the benzene is recognized as being a carcinogenic substance.

Xylene has a harmful effect on the brain. Levels of exposures raised for perished O of very short can involve headaches, a defect of coordination of driven S cles, giddinesses, the

confusion and losses of the direction of balance. Exposures torates él E vés for short periods of time can also cause an irritation of the skin, eyes, nose and throat, difficulties of breathing, problems pulmona I LMBO, an increase in the reaction times, losses of memory, irritations of E S tomac and deteriorations of the operation of the liver and kidneys. Very high rates of exposure can involve the loss of conscience even death. Studies on animals showed that high xylene concentrations increase the number of year I evils still-born children, as well as delays of growth and development. In much of case, these same conce N trations also has negative effects on health of the mothers. The effect of expos I tions of the mother to weak xylene concentrations on the foetus is not known per hour has C tuelle.

Aniline is a very toxic substance which must be handled with precaution. An E X position with high concentrations can be mortal. She can be absorbed by inhal has tion, ingestion and contact with the skin, *including in form vapor*.

Chlorobenzenes are toxic substances which must be handled with precautions. The L I mite of professional exposure is fixed in France at 10 ppm, that is to say 46 mg/m ^{3} of air.

Nitrobenzines can cause serious poisonings by ingestion, inhalation or contact with the skin. They react with the haemoglobin of blood and prevent it from reacting with oxygen. They can also involve disorders of the central nervous system, Ca U sant a feeling of weakness, headaches and vomiting. A high rate of nitr O benzene can result in death in less than one hour; moreover, its toxic effect is exacerbated by the alcohol catch.

I-2- Proportions lethal 50(DL $_{50}$):

I-2-1 Definition:

The lethal amount 50 or DL50 (***English LD50 for Lethal Proportions 50***) or CL50 (conce N lethal tration 50) are a quantitative indicator of the toxicity of a substance.

This indicator measures the amount of substance causing the death of 50 % of one animal population data (often of mouse or of rats) under precise conditions of experimentation.

The amount mortal minimum in the animal, or proportions lethal, is always delicate with déte R to undermine in a precise way. One prefers to establish the definite DL 50 like "the statistical estimate of a single amount of product supposed to kill 50 % of the animals" in experimentation.

The test is usually practised on 5 or 6 batches of animals, generally the rat. Each animal of the same batch receives an identical amount (single amount) of the substance to be tested, but the administered dose is different from one batch to another, so that the percentage in mortality

varies between 0 and 100. The route of administration is that which will be used in private clinic if it is about a m 3rd dicament (oral way, injection, etc), or that by which the substance will be able to penetrate in the organization if it is about a chemical (oral way, inhalation, transcutanée way, etc. .>.>.).

.>After the administration the animals are observed during 14 days during which the clinical examinations are frequent. The moment and the circumstances of death are carefully noted. The animals still in life at the end of the test are sacrificed. All the animals (deaths in the course of E S sai and sacrificed at the end of the test) are the subject of an autopsy.

One builds then the curve giving the percentage of mortality according to the log has rithme amount. It is a curve in "S", known as curve of Trévan, which can be linearized by suitable means. One deduces from them the DL50 (expressed out of Mg per kg of body weight), which one calculates also the standard deviation.

For a substance managed by oral way one considers that:

- if the DL 50 is < 5 mg/kg, the product is extremely toxic;

- if the DL 50 lies between 5 and 50 mg/kg, the product is very toxic;

- if the DL 50 lies between 50 and 500 mg/kg, the product is toxic;

- if the DL 50 lies between 0,5 and 5 g/kg, the product is not very toxic;

- if the DL 50 is > 5 g/kg, the product is not toxic or very little.

I-2-2 Forms of toxicity:

Toxic effects of the chemicals, and how they cause these effects, EP U wind being divided in several ways. For the studies of properties by QSAR, the effects tox I ques were divided into three main categories: toxicity receiver negotiated; acute toxicity not réce p tor negotiated; and effects supposed on human health **[1].**

I-2-3 Use of the lethal amount 50:

One manages G énéralemen T the poison with animals divided into several groups and this, with sufficient increasing amounts to obtain a percentage of mortality spreading out between 0 and 100 %. the effect of a substance is, overall, inversely proportional to the mass of the animal with which it is managed, this is why it is measured in G / kg. In gén 3rd ral, if immediate toxicity is similar at all the types of animals, it will be probabl E lies similar at the

human ones. When the DL_{50} are different at various species anim has them, one must make approximations and assumptions during the estimate of the amount probable mortal at the human ones.

I-2-4 Identification of toxicity:

The DL50 is used to measure all the toxicity of a substance, measures which is carried out via qualitative studies (nonmeasurable) and quantitative (measurable of which the DL_{50}).

The DL_{50} is often used as departure with the studies of toxicity because it provides a minimum of knowledge by identifying them symptoms of intoxication and the toxic amount. It should despite everything be considered with prudence because it is often a preliminary study (Pr E mière analyzes) which can be influenced by several factors, such it animal species, it sex, it age, moment of the day, etc.

It however has a limit value, because it relates to only mortality, from where appearance of values like the IC_{50}.

There are other methods of study of toxicity, for example the tests of irritation of skin and of corrosion of eyes, which generally form part of a program of evaluation toxicolog I that.

I-2-5 Identification of the pathogenic capacity:

The DL_{50} is one of the two data being used to measure the capacity pathogenic of a germ. The second data being the Infecting Minimal Amount (DM I).

Développement et évaluation de la qualité d'un modèle.

III-1 Sélection d'un sous-ensemble de descripteurs :

Specialized software allows the calculation of more than 3000 molecular descriptors belonging to various classes. Rather than to seek to explain the dependent variable (size of interest) by all the régresseurs (molecular descriptors), one can seek only one reduced unit régresseurs who gives such a satisfactory reconstitution of the variable to be explained. Among the put strategies opens some for the selection of a reduced whole of explanatory variables, one can quote: methods of step by step (downward method; ascending method, and method known as stepwise), as well as the evolutionary and genetic algorithms.

In general, the comparison is done with the advantage genetic algorithms (GA) which we applied in this work, and that we point out succintement.

Iii-1 -1 Principle:

In the terminology of the genetic algorithms, the binary vector $\underset{\sim}{I}$, called "chromosome", is a vector of dimension p where each position (a "gene") corresponds to a value (1 if it appears in the model, 0 if not). Each chromosome represents a model based on a whole of explanatory variables.

One starts by defining the statistical parameter to optimize (for example to maximize Q^2 by using the validation crossed by "leave one E - out"; cf will infra), with the size P of the population of the model (for example, P = 100), and the maximum number of variables L allowed for the model (for example, L = 10); the minimum of allowed variables is generally supposed to be equal to 1. Moreover, one probability of crossing p_C (usually high p_C = 0,9), and a probability of change p_M (usually weak, p_M = 0,1) must be also defined.

After definition of the principal parameters, the implementation of the genetic algorithm is started, its evolution includes/understands three principal stages.

I II –1-2 Initialization random of the model:

The population is made up at the beginning of random models with variables ranging between 1 and L, then the models are ordered have regard to the statistical parameter selected – the quality of the model – (the best model is in the first position, worst in position P);

I II 1-1-3 Stage of crossing:

From the population, one selects pairs of models (by chance, or with a probability proportional to their quality). Then, for each pair of models one preserves the common characteristics, i.e. the variables excluded in the 2 models remain excluded, and the variables integrated in the 2 models are preserved. For the variables selected in a model and eliminated in the other, one tests of it a certain number randomly which one compares with the probability of crossing p_C: if the random random number is lower than the probability of crossing, the excluded variable is integrated into the model and vice versa. Finally, the statistical parameter of the new model is calculated: if the value of this parameter is better than worst for the population, the model is integrated into the population, in the place corresponding to its row; in the contrary case, it is not taken into account. This procedure is repeated for many pairs (100 times for example).

I II –1-4 Stage of change:

For each model of the population (i.e. for each chromosome) p random random numbers are tested, and, one at the same time, each one are compared with the probability of change, p_M, defined: each gene remains unchanged if the random random number correspondent exceeds the probability of change, in the contrary case, one changes it 0 to 1 or vice versa. The low values of p_M allow only few changes, leading to new chromosomes not very different from the generating chromosomes.

After obtaining the transformed model, one calculates the statistical parameter of it: if this value is better than worst of the population, the model is integrated into the population, in the place corresponding to its row; in the contrary case it is not taken into account.

This procedure is repeated for the chromosomes, i.e. P time.

I II –1-5 Conditions of stop:

The stages of crossing and change are repeated until the meeting of a condition of stop (for example a maximum number of iterations defined by the user), or which it is put an end arbitrarily to the process.

A significant characteristic of the selection of a reduced whole of variables per genetic algorithm is that one does not obtain necessarily a single model, but the result usually consists of a population of acceptable models; this characteristic, sometimes considered as a disadvantage, provides an advisability to carry out an evaluation of the relationships to the response to various points of view.

Let us note that size of the model is fixed by the optimal value of the function the FIT of KUBINYI **[9]**, calculated according to:

$$FIT = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{(n+p^2)} \qquad (1)$$

p: indicating the number of variables of the model and R^2 the coefficient of determination. This criterion makes it possible to compare between models built on the same number N of data, but with a number of variables p differ.

I Ii-2 Développement D be m odèles:

The most current techniques to establish models QSAR use the analysis of regression (linear regression multiple: MLR; projection of the latent structures by partial least squares: PLS), neural networks, and methods of classification.

We used the MLR and the networks of artificial neurons (ANN). By imposing linear transformations between molecular descriptors and studied properties, the MLR can negatively influence the predictive capacities of the model. On the other hand, with the networks of neurons it is not necessary no to postulate a model. The networks of neurons have the capacity to represent any functional dependence which they discover by themselves. Thus, the discovery and the exploitation of the high level non-linear dependences can improve the capacity of prediction of the variable of interest.

Iii-2-1 Parameters of evaluation of the quality of the adjustment:

Two parameters are usually used:

The multiple coefficient of determination:

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \qquad (2)$$

where $\hat{y}_i$ is the estimated value of the physical parameter, and $\bar{y}$ the average of the actual values.

The root of the average quadratic error of prediction (also indicated by SDEP; Cf will infra):

$$\sigma_N = \sqrt{\frac{1}{n} \sum_1^n (y_i - \hat{y}_{(i)})^2} \qquad (3)$$

Iii-2-2 Robustness of the model:

The stability of the model was explored by using the "validation crossed by omission of an observation" (LOO: cross-country race - validation by leave-one-out) **[10]**. It consists in recomputing the model on (N – 1) drifts benzene, the model obtained E being then used to consider the activity biological of the compound eliminated noted. $\hat{y}_{(i)}$ One repeats the process for each N derived benzene.

"the sum of the squares of the errors of prediction", indicated by acronym PRESS (for: Predictive Residual Sum of Squares):

$$PRESS = \sum_1^n \left(y_i - \hat{y}_{(i)}\right)^2 \qquad (4)$$

is a measurement of the dispersion of these estimates. Is used it to define the coefficient of prediction:

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \qquad (5)$$

Contrary to R^2, which increases with the number of parameters of the regression, the factor Q^2_{LOO} posts a curve with maximum (or stage), obtained for a certain number of explanatory variables, then decrease thereafter in a monotonous way. This fact confers a great importance on the coefficient. Q^2_{LOO} A value from $Q^2_{LOO} > 0{,}5$, generally, is regarded as satisfactory, and a value known périeure with 0,9 is excellent **[11]**.

Iii-2-3 Test of randomization:

This test makes it possible to highlight randomly which had correlations. It consists in generating a vector "property considered" by random permutation of the components of the real vector. One calculates then on the vector obtained (regarded as real experimental vector) a model QSPR, according to the usual method. This process is repeated several times (100 in our case).

External Iii-2 -4 Validation:

In addition to the test of randomization, it is interesting **[12]**, to judge quality of the model, to consider the root of the average standard deviation (RMSE for Root Mean Squared Error, calculee on various sets:

- Together of estimate (called SDEC)
- Together of cross validation (also called SDEP)

- Together of external prediction (indicated by SDEPext).

These values RMSE are adapted better, to judge quality of a model that the values of R^2 and Q^2 only, which constitute good tests only for data distributed regularly.

Les réseaux de neurones artificiels.

IV–1 Les réseaux de neurones artificiels :

Les réseaux de neurones ont été étudiés depuis les années 40 **[13]**. Les idées de base de cette technique viennent de la recherche cognitive, d'où vient le nom 'réseaux de neurones'.

The technique inspired many researchers at that time, but much from the interest disappears after an article from Minsky and Papert **[14]**, finally started again at the beginning of the Eighties after a quasi-lapse of memory of a score of years. The cause of the sudden interest was the appearance of new network architectures of neurons.

Iv-1 the -1 artificial neuron:

The basic element of a network of neurons is, of course, the artificial neuron. A neuron (figure 1) contains two principal elements:

- A whole of weights ass ociés with the connex ions of the neuron, and
- A function of activation (Figure 2).

The values of entry are multiplied by their weight corresponding and are added to obtain the sum S.

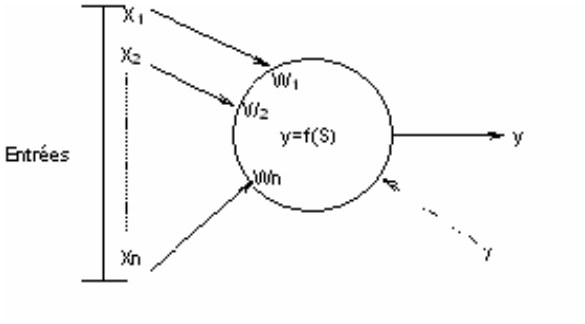
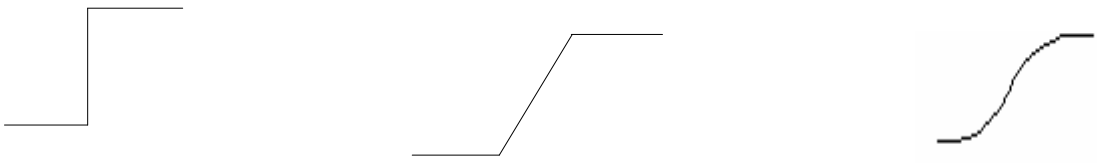


Figure – 1 *Le neurone artificiel générique.*

This sum becomes the argument of the function of activation, which is generally of one of the forms presented in figure 2. A function of significant activation is the simple multiplication with one, i.e. the exit is simply a balanced sum.



Fonction à seuil **fonction à saturation** **fonction sigmoïde**

Figure – 2 *Fonctions d'activation.*

The choice of the function of activation depends on the application. If it is necessary to have binary exits it is the first function which one usually chooses.

A special entry is practically always introduced for each neuron. This entry, normally called skew (English bias), is used for to move the step of the function of activation on the axis S. the value of this entry is always 1 and displacement depends then only on the weight of this special entry.

IV –1 –2 Properties of S networks of neurons:

A network of neurons is composed of neurons which are inter-connected so that the exit of a neuron can be the entry of one or more other neurons. Then there are entries of outside and exits towards outside **[15]**.

Rumelbart et al.. **[15]** eight principal components of a network of neurons give:

- A whole of neurons.
- A state of activation for each neuron (active, inactive…).
- A function of exit for each neuron (f(S)).
- A model of connectivity between the neurons (each neuron is connected to all the others, for example).
- A rule of propagation to propagate the values of entry through the network towards the exits.
- A rule of activation to combine the entries of a neuron (very often a balanced sum).
- A rule of training.
- An environment of operation (the operating system, for example).

The behavior of a network and the possibilities of application depend completely on these eight factors and the change of only one of them can change the behavior of network completely.

The networks of neu rones are often called "limp black " because the mathematical function which east représent 3rd E quickly becomes too complex to analyze it and include/understand it directly. That is in particular the case if the network develops distributed representations **[15]**, i.e. several neurons are more or less active and contribute to a decision. Another possibility is to have located representations, which makes it possible to identify the role of each neuron more easily. The networks of neurons nevertheless have a tendency to produce distributed presentations.

IV –2 various types of networks of neurons:

Several types of networks of neurons were developed which have very varied applicability often. In particular four types of networks are well-known:

- The network of Hopfield (and its version including the training, the machine of Boltzmann).
- The charts car - organizing of Kohonen.
- Networks with radial function which one names also RBF (for "Radial BASIC Functions ").
- Multi-layer networks or perceptron multi-layer PMC

The network of Hopfield **[16]** is a network with exits binary S where all the neurons are inter-connected with symmetrical weights. I.e. the weight of neuron N_I with neuron N_J is equal to the weight of neuron N_J has U neuron N_I. The weights are given by the user. The weights and the states of the neurons make it possible to define the "énergie" network.

It is this energy which the network tries to minimize to find a solution. The machine of Boltzmann is in theory a network of Hopfield, but which allows the training thanks to the minimization of this energy.

The charts car - organizing of Kohonen **[17]** are used to make automatic classifications of the vectors of entries.

The networks with radial function are network X multi-layer, with a chée layer Ca. However, contrary to will perceptrons multi-layer, the transfer transfer functions of the hidden layer depend on the distance between the vector of entry and the vector centers.

The multi-layer networks (PMC) are the most powerful networks of the networks of neurons which use the supervised training.

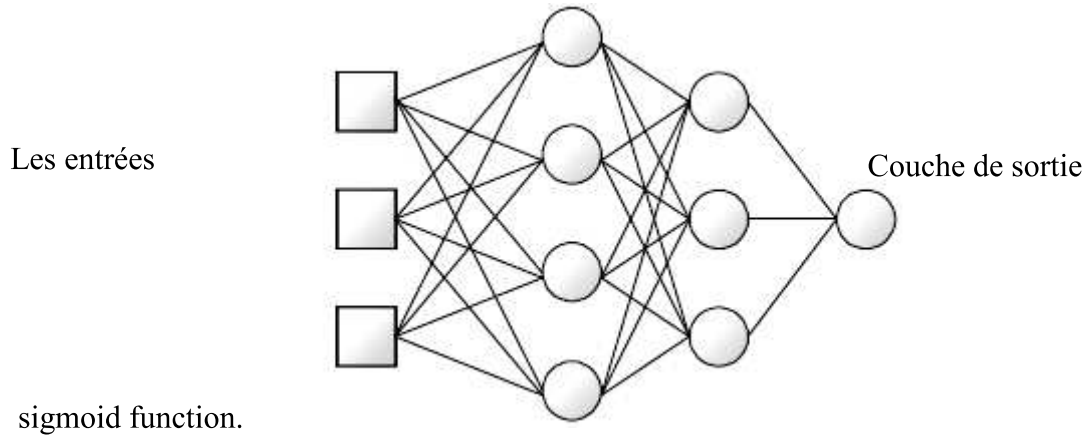
IV –2 -1 the multi-layer networks or will perceptrons multi-layer (PMC):

Multi-layer networks (PMC) (figure 3) is composed NT D entered S, a layer of exit and zero or several layer S hidden **[15]**. Connections are allowed only to one sub-base (nearer to S entered S) worm S a roadbase (nearer to the layer of exit). It is also interdict to have connections between neurons of the same layer.

L entered S ser ven T to distribute the values of entry to the neurons of the layers superior E S, possibly multiplied or modified in a way or another.

The layer of exit is composed normally of the linear neurons which calculate only one balanced sum of all its entries.

The hidden layers contain neurons with nonlinear functions of activation, normally the



sigmoid function.

Figure – 3 *Structure générale du perceptron multicouches*

It was proven **[18]** that there is always a network of neurons of this type with three layers only (the entry S, layer of exit and a hidden layer) which can approximate a function F: [0.1] $^{N} \Rightarrow R^{N}$ with any precision $\varepsilon > 0$ wished. A problem consists in finding how much hidden neurons are necessary to obtain this precision. Another problem is to make sure a priori that it is possible to learn this function.

Initially all the weights can have random values, which are normally very small S before beginning the training.

IV –3 Training:

The training of a network of neuron S means that it changes its behavior in order to enable him to approach a definite goal. This goal is normally the approximation of a whole of examples or the optimization of the state of the network according to its weights to reach the optimum of an economic function fixed a priori.

There are three types of principal trainings. They are the supervised training, the not-supervised training, and the training by attempt (graded training in English) **[18]**.

One speaks about supervised training when the network is fed with the good response for the examples of entries given. The network has then like drank to approximate these examples as well as possible and to develop at the same time the good mathematical representation which enables him to generalize these examples for then treating new situations (which étaie NT not has a presentiment of I are in the examples).

In the case of the training not-supervised the network itself decides which are the good exits. This decision guided by an internal goal with the network which expresses an ideal configuration to reach compared to the introduced examples. The charts car - organizing of KB H onen are an example of this type of network **[17]**.

` Graded learning' is a training of the test-error type where the network gives a solution E N being only fed with information indicating if the R 3rd p O nse were correct or if it were at least better than the last time.

There are several rules for each type of training. The supervised training is the type more used. For this type of training the most used rule is that of Widrow-Ho F F Of other rules of training are for example the rule of Hebb, the rule of perceptron, the rule of Grossberg etc… **[15, 18, 19]**.

IV –3 -1. The training of Widrow-Hof F:

The rule of training of Widrow-Hof is a rule which makes it possible to adjust the weights of a network of neurons to decrease with cha that stage the error made by C E network of neurons (provided that the factor of training is quite selected).

A weight is modified E N using the following formula:

$$w_{k+1} = w_k - \alpha\delta_k x_k \qquad (9)$$

Where:

w_k is the weight at the moment K;

the weight at the moment k-1;

α is the factor of training;

characterize the difference between the awaited exit and the tive exit effec of one neuron at the moment K;

x_k the value of entered the E which the weight W is associated the moment K.

Thus, if δ_k and x_k are positive both, then the weight must be increased. The size of the change depends above all on the size on δ_k but also on that on. x_k The coefficient α is used to decrease the changes to prevent that they become too large, which can involve oscillations of the weight.

Two versions improved E S of this training exist, the version ` by lois' and the version ` by inertie' (English momentum) **[18]**, of which one uses several examples to calculate the

average of the necessary changes before modifying the weight and the other prevents that the change of the weight at the time K becomes much larger only at the time k-1.

IV –3 the -2 training p has R R 3rd tro - propag has tion gradient (L evenberg-Marquardt backpropagation):

The algorithm of training by rét ro-propagation of the gradient (figure 4) is an iterative algorithm which aims to find the weight of connections minimizing the variation made by the network on the unit of training. This minimization by a method of the gradient led to the algorithm of training of retro-propagation.

The procedure of training breaks up into two stages. To start, the values of entries are presented at the network, which propagates then C are values to the layer of exit and gives the response to the network thus. At the second stage the good corresponding exits are presented at the neurons of the layer of exit which calculate the variation, modify their weights and retro - propagate the error until entées to allow the neurons hidden to modify their weights in the same way. The principle of modification of the weights is normally the training of Widrow-Hoff.

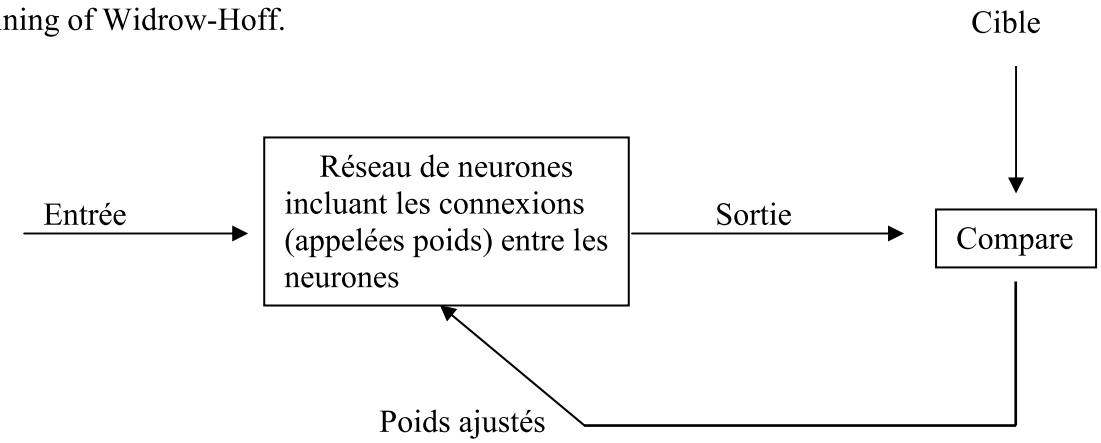


Figure – 4 *Apprentissage par un algorithme de rétro-propagation*

Generally for the calculation of the variation one uses quadratic error average *MSE* (*Mean Square Error*) defined by the relation:

$$MSE = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} \qquad (10)$$

y_i est la valeur observée , $\hat{y}_i$ est la valeur estimée , et n le nombre d'observations.

IV – 4 Criteria of stop:

Several criteria of stop can be utilis 3rd S with the algorithm D ' training. The first criterion consists in fixing a preliminary number of cycles or iterations, but it is difficult to know a priori how much iterations would be adapted to arrive at the fixed goal.

A deuxi 2nd me crit the 2nd Re consists in fixing a terminal inf 3rd rieure on the average quadratic error (MSE), it is sometimes possible to lay down an objective a priori to be reached. When the selected index of performance decreases below this objective, one consid the 2nd Re simply that the network learned its donn sufficiently well 3rd be and one stops the training. The disadvantage of this criterion is that it can generate a pH enomene D E on-training in D 3rd sirable in practice.

The third criterion is "the early stop", which consists in following the evolution of the performances of the network of generalization during the course of the training and stopping this one just before these performances are not put to be degraded, i.e. as soon as the index of performance calculated on given are of validation ceases improving. This method, the most used to avoid the on-training, is that which we chose in this work. The following graph illustrates this criterion:

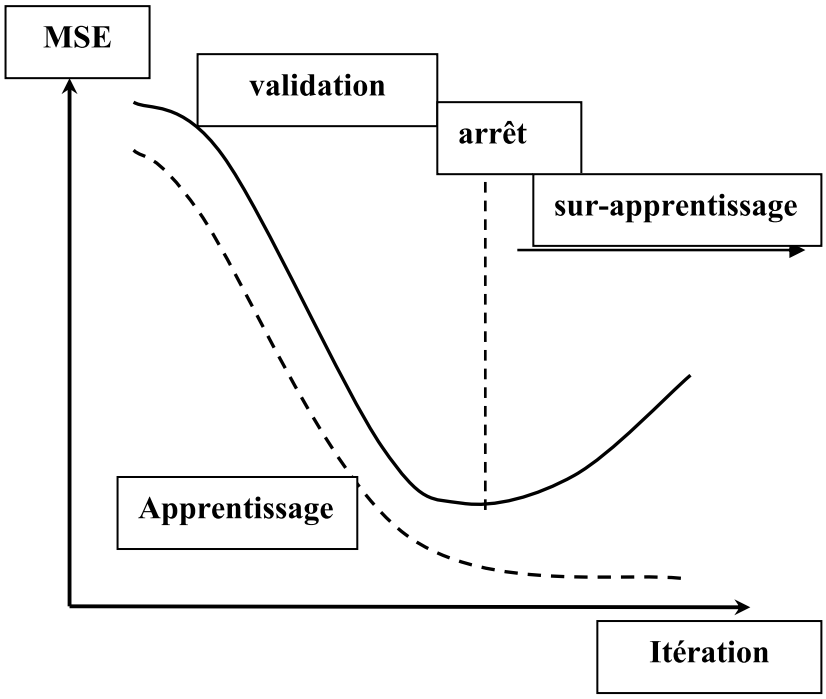


Figure – 5 *Illustration de l'arrêt précoce*

IV–5 Construction d'un modèle :

La construction d'un modèle implique dans un premier temps le choix des échantillons des données d'apprentissage, de test et de validation. Le choix du type de réseau intervient dans un second temps.

Les quatre grandes étapes de la création d'un réseau de neurones sont détaillées comme suit :

IV–5-1 Construction de la base de données :

Le processus d'élaboration d'un réseau de neurones commence par la construction d'une base de données.

Afin de développer une application à base de réseaux de neurones, il est nécessaire de disposer de deux bases de données, une pour effectuer l'apprentissage et l'autre pour tester le réseau obtenu et déterminer ses performances. Pour cette raison nous avons partagé notre base des données (**tableau I**) aléatoirement en trois sous-ensembles comme suit :

- Un ensemble de 110 composés pour l'apprentissage du réseau de neurones.
- Un deuxième de 28 composés pour la validation externe.
- Et un troisième de 28 composés choisis aléatoirement de l'ensemble d'apprentissage pour le test.

Généralement, les bases de données subis **IV –5 Construction of a model:**

The construction of a model initially implies the choice of the samples of the data of training, test and validation. The choice of the type of network intervenes in the second time.

The four great stages of the creation of a network of neurons are detailed as follows:

IV –5 1 Idiot struction of the data base:

The development process of a network of neurons starts with the construction of a data base.

In order to develop an application containing networks of neurons, it is necessary to have two data bases, to carry out the training and the other to test the network obtained and to determine its performances. For this reason we have partag 3rd our base of the data (**table I**) by chance in three subsets as follows:

- A whole of 110 composed for the training of the network of neurons.
- a second of 2 8 composed for the external validation.
- And a third of 2 8 by chance selected compounds of the whole of training for the test.

Generally, data bases undergone sent un prétraitement qui consiste à effectuer une normalisation appropriée tenant compte de l'amplitude des valeurs acceptées par le réseau.

Les valeurs d'entrées et de sortie sont normalisées dans un intervalle spécifique afin de donner à chaque paramètre la même influence statistique. Les valeurs d'apprentissage et de test ont été normalisées dans la marge [- 1, 1], au moyen de l'équation

$$x_{norm} = 2 \times \frac{\left(x_j - x_{\max}\right)}{\left(x_{\max} - x_{\min}\right)} - 1 \qquad (11)$$

où x_{norm} est la valeur normalisée ; x_j est la j^{ième} valeur ; $x_{\max}$ est la valeur maximale ; $x_{\min}$ est la valeur minimale

VI–5-2 Définition de la structure du réseau :

Nous avons retenu le Perceptron Multicouches comme base du modèle. Nous structurons ce réseau en précisant le nombre de couches et de neurones cachés pour que le réseau soit en mesure de reproduire ce qui est déterministe dans les données.

VI –5 - 3 Nombre de couches et de neurones cachés :

Mis à part les entrées et la couche de sortie, il faut décider du nombre de couches cachées. Sans couche cachée, le réseau n'offre que de faibles possibilités d'adaptation. Néanmoins, il a été démontré qu'un Perceptron Multicouches avec une seule couche cachée

pourvue d'un nombre suffisant de neurones, peut approximer n'importe quelle fonction avec la précision souhaitée **[20]**.

Chaque neurone peut prendre en compte des profils spécifiques de neurones d'entrée. Un nombre plus important permet donc de mieux "coller" aux données présentées mais diminue la capacité de généralisation du réseau. Il faut alors trouver le nombre adéquat de neurones cachés nécessaire pour obtenir une approximation satisfaisante.

VI –5 - 4 Présentation de l'environnement utilisé :

Accordingly, software MATLAB **[21]**, which contains a module devoted to the développ E lies of abstr. water of neurons, was retained; a PC Dell P4 with a RAM D E 512 and one speed of 3.4 GHz was used.

The network of neurons stores information in a chain of neuronal interconnections, by calling upon the concept of weight (weight C ouche of entry = *IW* - I *nitial weights*, weight C ouch E of exit = *Lw-last weights*).

A capacity of training is necessary to adjust the weights of the networks of neurons during the phase of training during which all the data are presented at the RNA on several occasions.

The sigmoid function of transfer, hyperbolic tangent, was adopted like function of A C tivation for the hidden layers and the function flax field for the layer of exit.

We present in the following page the algorithm of the network of neurons used:

Algorithme du réseau de neurones utilisé :

```
P = [ descriptors ];
T = [ the physical property studied ];
N =  138; % all compounds
N1 =  110; % Made up of training
N2 =  28; % Made up of validation
P0 = (P) '; % Transposition of the matrix P
T0 = (T) ';% Transposition of the matrix T
[ pn, minP, maxP, tn, minT, maxT ] = premnmx(P0, T0); % standardization between [ - 1, +1 ]
P1n = (pn) ';
T1n = (tn) ';
% Training
P1=Pn(1:N1,:)'; % Descriptors standardized of training
T1=Tn(1:N1,:)'; % Physical property standardized of training
T10=T(1:N1,:)';
% Test
[ R, Q ] = size (P1);
iitst = [ 4:4:Q 2: 108:Q ]; % Choices random of 2 8  composed of the test
test.P = P (:, iitst);
T20=T10 (:, iitst);
% Validation
val.P = Pn(N1+1:N,:)'; % Descriptors standardized of validation
val.T = Tn(N1+1:N,:)'; % Physical property standardized of validation
T30=T (N1+1:N,:) ';
Net = newff(minmax(P), [ S1 S2], {TF1 TF2}, BTF); % Creation of a network
 % S1: neurons of the hidden layer  – S2: the exit (= 1)
 % TF1, TF2: Transfer transfer functions  – BTF: transfer transfer function of retro-propagation
net.trainParam.epochs = 500; % iteration count
net.trainParam.goal = 0.0000001; % the desired error
Net = init(net); % Initialization of the network
[ Net, tr]=train (Net, P1, T1, [ ], [ ], valley); % Drive of the network
plotperf(tr)
a1n=sim(net, P1); % Simulation of the network for the data of training
[ a1]=postmnmx(a1n, minT0, maxT0);%Remettre results of training to their actual values
E1 = T10-a1; % analysis error
a2n=sim(net, test.P); % Simulation of the network for the data of the test
[ a2]=postmnmx(a2n, minT0, maxT0);%Remettre results of the test to their actual values
E2 = T20-a2; % analysis error
a3n=sim(net, val.P);    % Simulation of the network for the data of validation
[ a3]=postmnmx(a3n, minT0, maxT0);%Remettre results of validation to their actual values
E3 = T30-a3; % analysis error
```

Résultats et discussion

V-1 Calcul et choix des descripteurs :

The structures of the molecules were obtained using the molecular software of modeling Hyperchem 7.5 **[22]**, and the final geometries using method semi empirical AM1 of the same software. All calculations were carried out within the framework of formalism RHF without interaction of configuration. The molecular structures were optimized using the Polak-Ribiere algorithm with for criterion a root of the average square of the gradient equalizes with 0.001 kcal.mol $^{-1}$. The geometries obtained were transferred in the software data processing **[22 23]** used for calculation of more than 1700 descriptors belonging to 20 different classes.

The software DRAGON Professional **[23]** that we used, is a developed molecular software of modeling, at the origin, by the group of research in chimometry and QSAR of Milan. These descriptors can be employed to develop models of relations qualitative molecular Structure-Activités/Propriétés.

Selection of a minimal whole of significant descriptors at summer realized by genetic algorithm, in version MOBYDIGS of Todeschini **[24]** by maximizing R^2, Q^2_{LOO} and $Q^2_{ext.}$.

The size of the model is fixed by optimizing the value of the function the FIT of KUBINYI **[9]**, calculated according to (1):

$$FIT = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{(n+p^2)} \qquad (1)$$

Nous avons réuni dans la tableau II les 10 descripteurs entrant dans la construction du modèle ainsi que leur signification :

Tableau – II Descripteurs moléculaires intervenant dans la modélisation de la dose létale 50.

These descriptors, which belong to 7 different classes, all are calculated using the software DRAGON Professional **[23]**.

In what follows we specify their relations of definitions, and we present a summary of their properties.

V-1-1 Descriptors of constitution:

They are the simplest descriptors usually used;ils reflect the molecular composition of a compound, without informing about the geometry of its molecule.

These descriptors are insensitive to a change of conformation and do not distinguish between isomers;

- No is the number of atom S of oxygen.

- ARR is the report/ratio of aromaticity, equal to the number of aromatic connections on the total number of connection non-H.

V-1 - 2 topological Descriptors:

The topological descriptors are based on the graph representative of the deprived molecule, in general, of the hydrogen atoms. They are quantifiers of the topology of the molecule obtained by application of algebraic operators to the matrices images of the molecular graphs, and of which the values are independent of the classification or the labelling of the tops (atoms). They can be sensitive to one or more characteristics structurale(s) of the molecule telle(s) that: cut, form, symmetry, ramification and cyclicity, and also can encoder of information chemical concerning the type of atom and the multiplicity of the connections.

The matrix outdistances (D) obtained starting from the molecular graph G summarizes information relating to the topological distance between all the pairs of atom **[25]**. The topological distance D_{ij} is the number of edges (connections) of the shortest way $^{m}p_{ij}$ between tops v_I and v_J, it is with saying the length of geodetic between v_I and v_J.

$$d_{ij} = \begin{cases} \left|\min pij\right| & si\ \ i \neq j \\ 0 & \text{si}\ \ i = j \end{cases} \qquad (12)$$

The nondiagonal elements dij of the matrix distance are equal to 1 if tops VI and vj are adjacent (it is with saying if atoms I and J are dependent), and higher than 1 in the contrary case. Obviously, the diagonal elements are null. The matrix outdistances which is

symmetrical and from dimension A*A (A indicating the number of atoms) is obtained, in general, starting from the graph of the molecule deprived of the hydrogen atoms.

The element having the maximum value in line I, it is with saying the maximum distance between top I and any other top is called eccentricity of the atom (or eccentricity of the top).

$$\eta_i = \max_j (d_{ij}) \qquad (13)$$

From the definition of the eccentricity, it is possible to characterize a graph G by 2 molecular descriptors which are: the topological ray R and the topological diameter D; the topological ray of a molecule is defined by the minimal value of the eccentricity of the atom, and the topological diameter by the value maximum E of the eccentricity of the atom, S E lon:

$$\text{and } D = \max_i (\eta_i) \qquad (14)$$

Other simple molecular descriptors are calculated starting from the eccentricity of the atom **[2 6]**.

They are the eccentricity ή, the average atomic eccentricity $\overline{\eta}$ and the eccentric $\Delta\eta$ defined, respectively, as follows:

$$\eta = \sum_{i=1}^{A} \eta_i \; ; \; \overline{\eta} = \frac{1}{A} \sum_{i=1}^{A} \eta_i \; ; \quad \Delta\eta = \frac{1}{A} \sum_{i=1}^{A} \left| \eta_i - \overline{\eta} \right| \qquad (15)$$

* The index of form-2d of Petitjean **[27]** is a topological descriptor of anisometry defined by:

$$PJi2 = \frac{D - R}{R}; \qquad (16)$$

where R and D are: respectively, the topological ray and the topological diameter.

V-1 - 3 Descriptors based on the eigenvalues:

Descriptors calculated starting from the eigenvalues of a matrix square, (generally) symmetrical representing a molecular graph.

These descriptors can correspond to selected eigenvalues (usually the greatest eigenvalue), or functions of several or all the eigenvalues of the matrix considered **[28,29]**. The descriptors "eigenvalues of Burden", are definite molecular descriptors like eigenvalues S of a matrix of modified adjacency, indicated by matrix B of Burden **[30]**. This matrix B,

which repré feels the graph of a molécul E including the hydrogen atoms, is defined as follows: the diagonal elements Bii represent atomic properties; the nondiagonal elements Bij corresponding to pairs of dependent atoms are equal to the square roots of the conventional orders of connection, i.e.: 0.1 - 0.2 - 0.3 and 0.15 respectively for the connections simple, double, triple, and aromatic. All the other matric elements are posed equal to 0.001.

The software DRAGON **[23]** calculates 4 matrix of Burden different whose diagonal elements correspond to

1/masses atomiques(m).

2/atomic volumes of Van DER W has als(v).

3/Eléctronégativités atomic of Sanderson (E).

4/Polarizabilities (p).

For each one of these matrices the descriptors of Burden are calculated like sequences highest and of the lowest eigenvalues, of which it was shown that they reflect aspect in connection with the molecular structure, and are thus useful for the chemical search for similarite/diversity to which were proposed, in the beginning, these descriptors.

The Software DRAGON provides the first 8 highest values BEHwk and the first eight values lowest (in absolute value) BELwk for each matrix, W representing the atomic property and K the row of the eigenvalue.

V-1 - 4 molecular Profiles:

The molecular profiles constitute molecular sequences of descriptors proposed by Randic **[31,32]** starting from the interatomic geometrical distances from a molecule.

The software DRAGON provides two molecular profiles. One is much more connected to the total molecular structure 3d:

(DP 01, DP02.....,DPk,DP20)

And the other with the molecular form

(SP01, SP02…, SPk…, SP20)

Each DPk descriptor of profile DP is calculated according to:

$$DPk = \frac{1}{k!} \frac{\sum_{i=1}^{A} \sum_{j=1}^{A} r_{ij}^{k}}{A} \qquad (17)$$

où r_{ij} est la distance géométrique entre les atomes i et j, A le nombre d'atomes de la molécule et k l'ordre du descripteur (k=1,…, 20).

For the great values of K, the values of DP tend towards 0 in consequence of the effect of the factorial term of standardization.

C haque SPk descriptor of the profile of form is calculated in the same way that descriptors DP, but by holding account only atoms of the molecular periphery (i.e. atoms of connectivity 1 or 2, by not taking account of the hydrogen atoms).

The profile of form is a local profile in the sense that it is attached to local molecular characteristics.

The molecular profiles of Randic can be used in models QSAR, but are particularly adapted in the molecular analysis similarite/diversity insofar as each profile characterizes a molecule well.

V-1 - 5 Descriptors Morse-3d (3d-Molecule Representation of Structures based one Electron diffraction; Representation of the Structures of the Molecules in 3d based on diffraction of the Electrons):

These descriptors are based on the idea to obtain information starting from the atomic co-ordinates 3d by using the transformation put into practice in the studies of electronic diffraction for the preparation of the curves of dispersion **[33].**

The descriptors Morse-3d are calculated starting from the relation:

$$Morsw = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i w_j \frac{\sin(s.r_{ij})}{s.r_{ij}} \qquad (18)$$

où Morsw est l'intensité électronique dispersée, w une propriété atomique, les r_{ij} sont les distances interatomiques et A le nombre d'atomes. Le terme s représente la dispersion dans différentes directions par une collection de A atomes.

Dans le but d'obtenir des descripteurs tout à fait uniformes, la répartition de l'intensité est rendue discrète, en calculant sa valeur comme une séquence de valeurs régulièrement distribuées; dans le logiciel DRAGON, en particulier, s est supposé prendre les valeurs entières de 0 à 31 (pour s = 0 le rapport de dispersion est posé égal à 1).

32 descripteurs MoRSE-3D sont calculés pour 5 propriétés atomiques différentes w: le cas non pondéré (u); la masse atomique m; le volume de Van Der Waals (v); l'électronégativité de Sanderson (e) et la polarisabilité atomique (p).

V-1-6 Descripteurs WHIM (Weighted Holistic Invariant Molecular descriptors descripteurs Moléculaires à Invariant Holistique Pondéré):

Les descripteurs WHIM constituent une classe de descripteurs moléculaires géométriques; ils permettent de saisir dans le détail les informations relatives à la taille; la forme, la symétrie et la distribution des atomes d'une molécule par rapport à des cadres de référence fixes. Le calcul des descripteurs WHIM repose sur l'analyse en composantes principales de la matrice de covariance des coordonnées atomiques pondérées, dont les éléments sont définis par:

$$s_{jk} = \frac{\sum_{i=1}^{n} w_i (q_{ij} - \overline{q_j})(q_{ij} - \overline{q_k})}{\sum_{i=1}^{n} w_i} \qquad (19)$$

où n représente le nombre d'atomes de la molécule, w_i le poids du i ème atome, q_{ij} la j ème coordonnée cartésienne de l'atome i (j =1, 2, 3), alors que $\overline{q_j}$ est la moyenne de cette j ème coordonnée.

Six modes de pondération sont proposés, et selon le mode adapté on obtient différentes matrices de covariance et différents axes principaux pour la molécule.

On distingue les descripteurs WHIM dirigés, calculés individuellement selon les directions des composantes principales et les descripteurs WHIM non dirigés, ou globaux, calculés pour la molécule entière à partir des combinaisons des premiers.

Le groupe de descripteurs de forme WHIM **[34-36]** dirigés (ou indice d- WSHA) θ_1, θ_2 et θ_3 reliés à la forme moléculaire, sont calculés comme rapport de valeurs propres:

$$\theta_k = \frac{\lambda_k}{\sum_k \lambda_k}, \quad k=1, 2, 3 \qquad (20)$$

A cause de la condition de fermeture ($\theta_1+\theta_2+\theta_3=1$), uniquement deux de ces descripteurs sont indépendants.

Le symbole Pkw est utilisé pour désigner ces descripteurs. Ainsi P1p correspond à k=1, pour une pondération par les polarisabilités atomiques (p).

V-1-7 Descripteurs GETAWAY (Geometry, Topology, and Atom-weights Assembly):

[37,38] Récemment proposés comme descripteurs de structure chimique a partir d'une nouvelle représentation de la structure moléculaire, la Matrice d'Influence Moléculaire (MIM) notée $\underset{\sim}{H}$ et définie comme suit :

$$\underset{\sim}{H} = \underset{\sim}{M}.(\underset{\sim}{M}^{T}.\underset{\sim}{M})^{-1}.\underset{\sim}{M}^{T} \qquad (21)$$

Où $\underset{\sim}{M}$ est la matrice moléculaire composée des coordonnées cartésiennes centrées x, y, z des atomes (y compris les hydrogènes) de la molécule dans une configuration déterminée, l'exposant T désigne la matrice transposée. Les coordonnées atomiques sont supposées rapportées au centre de gravité de la molécule dans le but d'assurer l'invariance par translation. La matrice d'information moléculaire est une matrice symétrique qui présente une invariance rotationnelle relative aux coordonnées de la molécule.

Les éléments diagonaux h_{ii} de la matrice d'influence moléculaire dénommés leviers, se distribuent de 0 à 1 et encodent une information atomique liée à « l'influence » de chaque atome de la molécule à déterminer la forme globale de celle-ci ; en effet, les atomes périphériques présentent toujours des valeurs h_{ii} plus élevées que celles des atomes proches du centre de la molécule.

De plus, l'amplitude du levier maximum d'une molécule dépend de la taille et de la forme de la molécule, comme il ressort de la géométrie de la molécule, les valeurs des leviers sont effectivement sensibles à des changements conformationnels significatifs et aux longueurs de liaison qui tiennent compte du type d'atomes et de la multiplicité de la liaison.

Chaque élément non diagonal h_{ij} représente le degré d'accessibilité du j eme atome quant à des interactions avec le i eme atome, l'aptitude des deux atomes considérés à interagir entre eux. Un signe négatif pour les éléments non-diagonaux signifie que les deux atomes occupent des régions moléculaires opposées par rapport au centre de la molécule, ce qui fait que le degré de leur accessibilité mutuelle sera faible.

Le rayon maximal d'autocorrection de distance k est calculé selon la relation :

$$Rkw+ = \max_{ij}\left(-\frac{\sqrt{h_{ii}h_{jj}}.}{r_{ij}}.w_i.w_j.\delta(k.d_{ij})\right) \quad i \neq j \text{ et } k = 1, 2, \ldots, 8 \qquad (22)$$

Où : d_{ij} est la distance topologique entre les atomes i et j ; w_i désigne une pondération atomique physicochimique ; $\delta(k;d_{ij})$ est la fonction delta de Dirac (δ=1 si d_{ij}=k ; zéro par tout ailleurs).

Les propriétés atomiques w utilisées dans le calcul des descripteurs GETAWAY sont : la masse atomique (m), la polarisabilité, l'électronégativité (e), le volume atomique de Van Der Wals (v), en plus du poids unité (u).

Nous avons réuni en annexe, pour les composés étudiés, les valeurs calculées des différents descripteurs.

V-2 Calcul du modèle par réseau de neurone:

Le graphe suivant montre que la valeur maximale du FIT qui correspond à 10 descripteurs, ce qui permet de définir la dimension du modèle à 10 descripteurs.

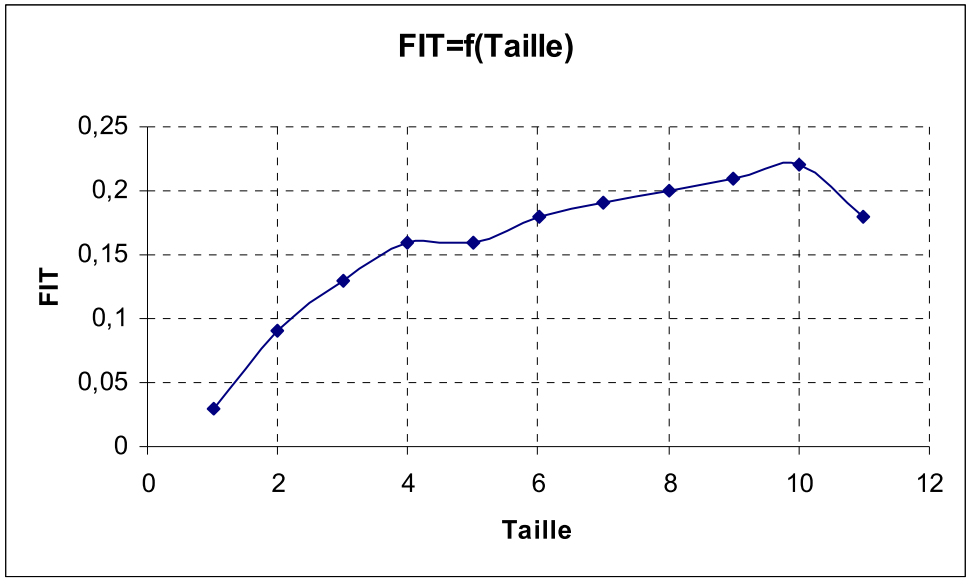


Figure – 6 FIT en fonction du nombre de descripteurs.

V-3 Optimisation par les réseaux de neurones artificiels :

En utilisant ces descripteurs pour l'optimisation neuronale, et après avoir cherche le nombre de neurones dans la couche cachée (figure 3), l'algorithme d'apprentissage est composé suit:

Tableau III Structure optimale du réseau de neurones

Nombre d'entrées	10(les descripteurs)
Nombre de sorties	01 (La dose létale 50)
Nombre de couches cachées	01
Nombre de neurones dans la couche cachée	08
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonctions d'apprentissage	Tangente hyperbolique

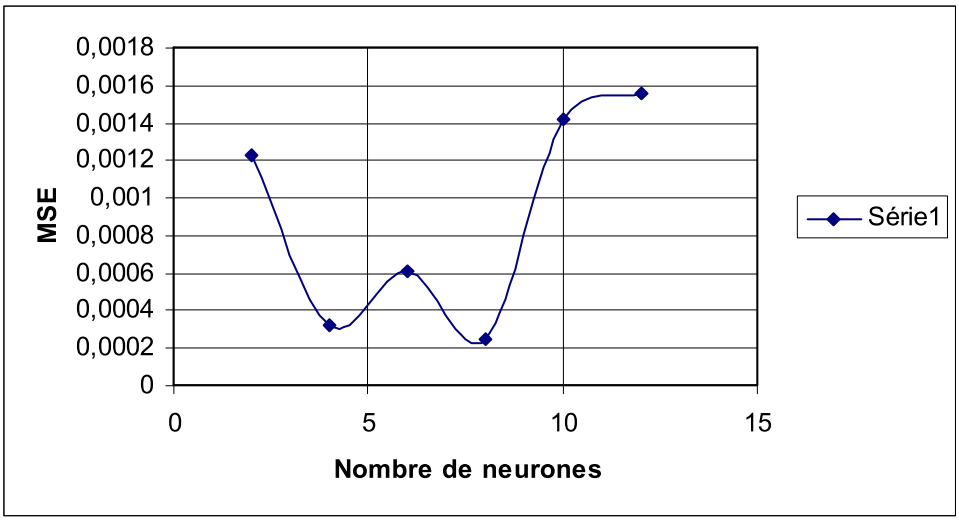


***Figure** – 7 Le choix du nombre de neurones de la couche cachée.*

Parmi les modèles a dix descripteurs, nous avons retenu celui qui présente les valeurs maximales pour Q^2, R^2 et $Q^2_{ext,}$ les descripteurs correspondants sont (AAR ; nO; PJi2; DECC; BEHp6; DP15; SP09; Mor04p; P1p; R1u+), dont les paramètres statistiques du modèle sont :

$$n = 110 \quad ; \quad \sigma_N = 0.0174078; \quad R^2 = 99.94\ \% \quad ; \quad Q^2 = 0.9993$$

Les valeurs des paramètres statistiques montrent que les dix descripteurs (AAR ; nO; PJi2; DECC; BEHp6; DP15; SP09; Mor04p; P1p; R1u+) permettent de corréler la dose létale 50 des dérivés benzéniques.

En effet, la valeur du coefficient de détermination (R^2) signifie 99.94% de la variabilité de pDL50 peut être expliquée par ces dix descripteurs, alors que la racine de l'erreur quadratique moyenne de prédiction est de l'ordre de 0.0174078 (σ_N=0.0174078); en outre ce modèle est très hautement significatif (grande valeur du paramètre de Fisher: F=16842.2496).

La commande « régression » de MINITAB fournit les valeurs des résidus caractéristiques réunis dans le tableau (IV), ainsi que les valeurs h_{ii}, Di et DFITSI qui permettent d'établir des diagnostics d'influence.

Tous les résidus standardisés d_i de la colonne (6) sont compris entre les limites ± 2, à l'exception des points déjà signalés 50 et 107.

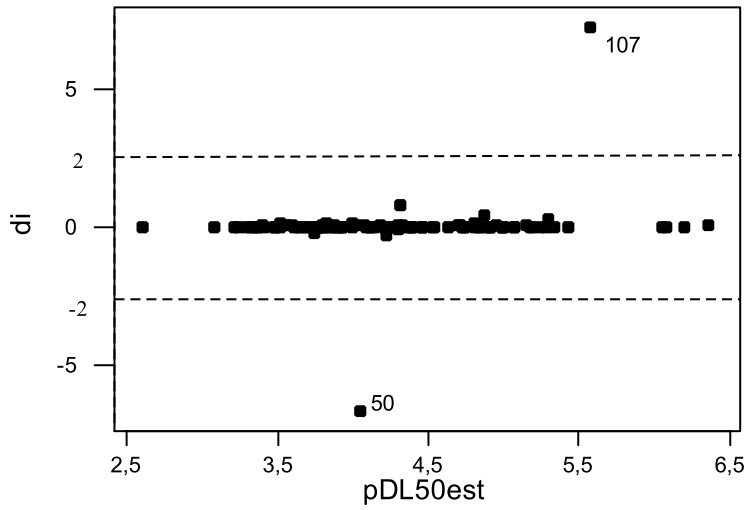


Figure – 8 Graphe des résidus standardisés en fonction des valeurs estimées

La colonne (4) donne les valeurs de h_{ii}, $i^{ème}$ terme diagonal de la matrice de projection : $\underset{\sim}{H} = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\ \underset{\sim}{X}'$ où $\underset{\sim}{X}$ est la matrice des valeurs observées des variables explicatives et $\underset{\sim}{X}'$ sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques. .La valeur critique pour déterminer les points leviers correspond à $h^* == \frac{3p}{n}\ \frac{3\times10}{110} = 0{,}272$. On constate que tous les h_{ii} sont inférieures à cette valeur critique 0.272, à l'exception des composés 50 et 107.

Le diagramme de Williams (di en fonction de h_{ii}) de la figure (9) montre l'influence potentielle des points 1, 50, 68, 106, et 107, qui nécessitent un examen plus approfondi.

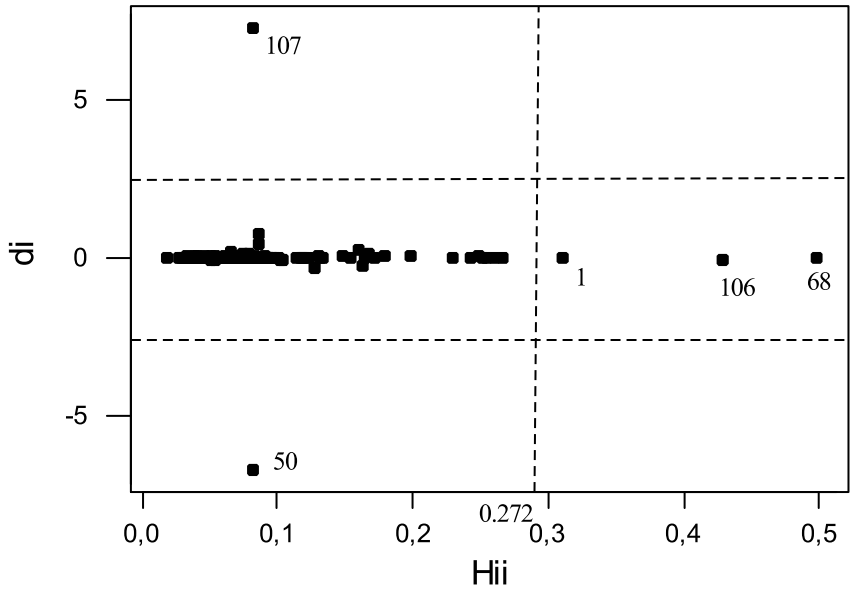


Figure – 9 ***Diagramme de Williams.***

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par « leave –one –out ». La figure (10), qui reproduit les valeurs prédites pDL50 en fonction de celles observées, fait ressortir une faible dispersion caractéristique d'une bon ajustement, d'ailleurs confirmé par la grande valeur de Q^2 (=0.9993).

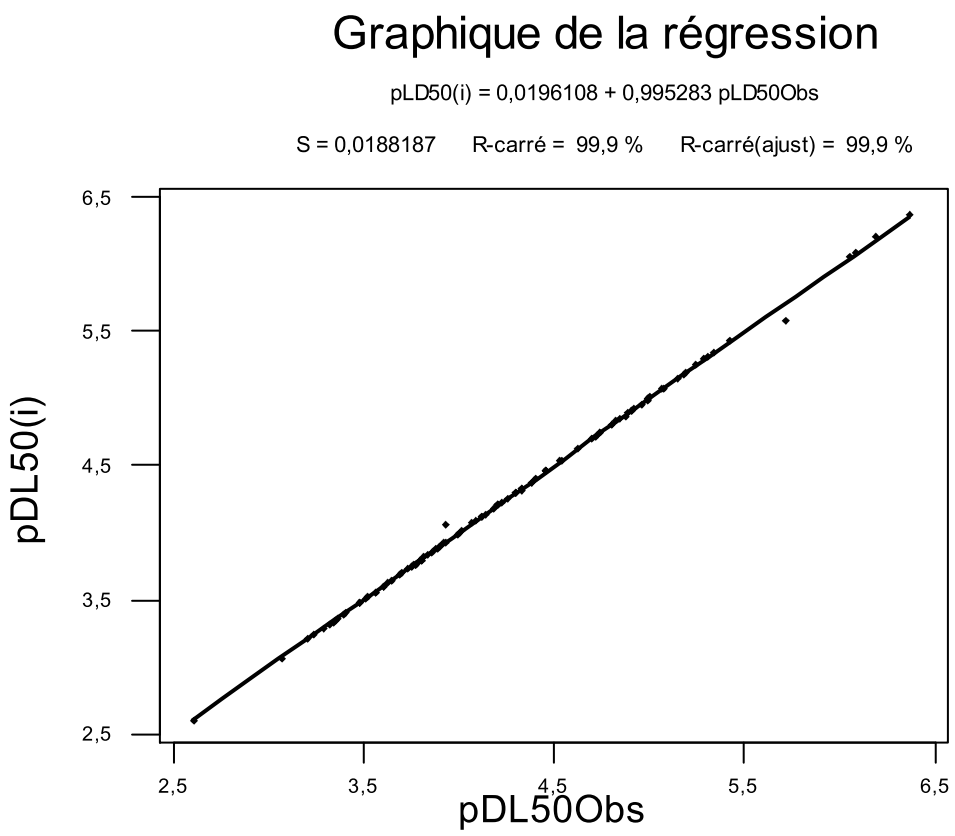


Figure – 10 Graphe des valeurs prédites pDL50 $_{(i)}$ en fonction des valeurs observées.

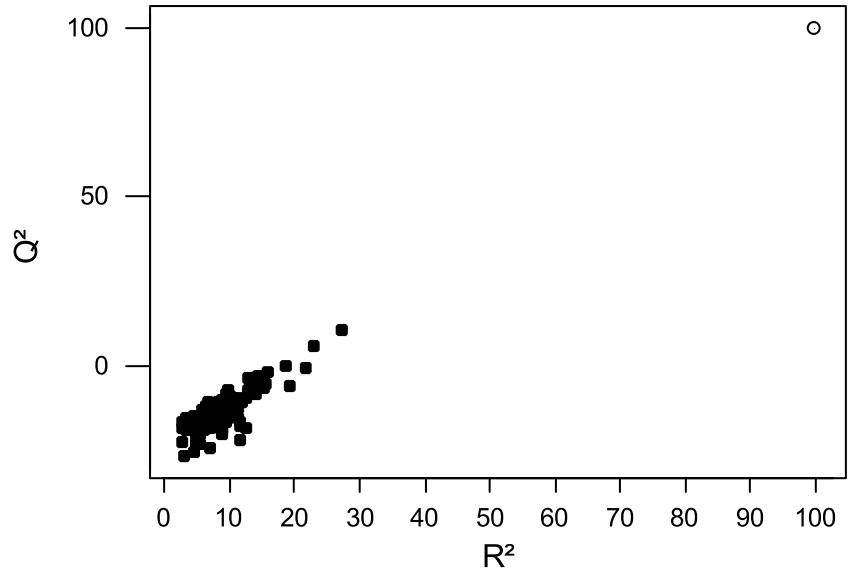
Dans le but d'établir que le modèle obtenu n'est pas dû au hasard ou à une sur spécification, nous avons appliqué le test de randomisation.

Ainsi 100 nouveaux vecteurs de la dose létale 50 ont été générés par permutation des positions des composantes du vecteur réel:

$$y = (y_1, y_2, \ldots, y_{27})' \xrightarrow{RND} y_{RND} = (y_8, y_5, \ldots, y_2)'$$

et utilisés comme sources d'observations pour des modèles QSAR dans les conditions optimales établies (10 paramètres) .

La figure (11) qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (cercles noircis) au modèle réel de départ (astérisque).



Figure–11 Test de randomisation associé au modèle QSPR. Les cercles noircis représentent les doses létales 50 ordonnées de façon aléatoire, et l'astérisque correspond au modèle réel.

Il est clair que les statistiques obtenues pour les vecteurs modifiés de la dose létale 50 sont plus petites que celles du modèle QSAR réel, et pour la majeure partie on obtient un $Q^2 < 0$. Ceci permet d'assurer qu'une relation structure/ dose létale réelle a été établie.

V-4 Validation externe:

To predictive ensure itself of the good effective capacity of the model, we operated by external validation one the whole of 2 8 compounds chosen by chance and which C not make share of the whole of test (NT number 3 carries while exposing in L E counts I).

With rigorous validation of the produced model results in A significant proportion of exact predictions given one the whole of the validation. The performance of the model is then measured by the coefficient of regression R^2.

The results obtained show that the predicted been worth (table III) are closed to the actual been worth. Been worth The of R^2 (figure-1 2) is equal to 0,999, which confirms that the neuronal model described in year adequate way the relation between L are amount S lethal 50 predicted and observed.

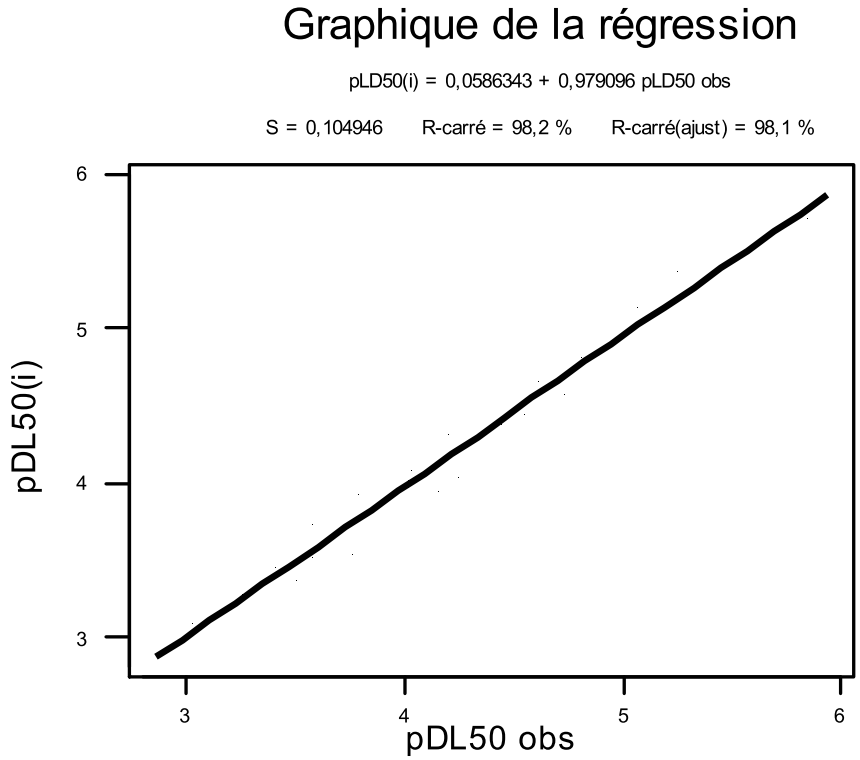


Figure – 12: Graphe des $pDL50_{(i)}$ prédites en fonction des valeurs observées pour validation.

Les résultats confirment ainsi la faisabilité de l'approche neuronale comme technique de modélisation de la dose létale 50.

Tableau - IV Valeurs des pDL50 observées, prédites, erreurs et pourcentages d'erreurs pour l'ensemble de validation.

	N°	composés	pDL50	$pDL50_{(i)}$	e_i	e%
1	55	1-methyl-3,5-dinitrobenzene	5,85	5,716	0,134	2,2906
2	20	1,4-dimethoxybenzene	3,57	3,532	0,038	1,06443
3	73	2-chloro-4-methylaniline	4,02	4,02	0	0
4	69	4-bromoaniline	3,42	3,445	-0,025	-0,73099
5	15	Aniline	3,58	3,727	-0,147	-4,10615
6	03	Chlorobenzène	3,77	3,55	0,22	5,83554
7	08	1,4-dichlorobenzene	4,02	3,975	0,045	1,1194
8	116	1,5-dimethyl-2,4-dinitrobenzene	4,27	4,238	0,032	0,74941
9	89	4-nonylphenol	2,87	2,899	-0,029	-1,01045
10	87	2-nitrobenzaldehyde	4,56	4,451	0,109	2,39035
11	101	4-fluoroaniline	4,26	4,03	0,23	5,39906
12	113	4-methyl-3-nitroaniline	5,07	5,141	-0,071	-1,40039
13	133	Toluène	4,82	4,814	0,006	0,12448
14	127	1-(2,4-dichlorophenyl)ethanone	4,47	4,398	0,072	1,61074
15	11	2-methyl-1,3,4-trinitrobenzene	3,04	3,091	-0,051	-1,67763
16	33	3-bromo-4-hydroxy-5-methoxybenzaldehyde	4,74	4,573	0,167	3,52321
17	27	1-methyl-4-nitrobenzene	3,80	3,924	-0,124	-3,26316
18	123	1,3,5-trinitrobenzene	3,36	3,357	0,003	0,08929
19	07	1,3-dichlorobenzene	4,62	4,671	-0,051	-1,1039
20	122	1-naphthol	3,59	3,525	0,065	1,81058
21	04	Phénol	3,51	3,372	0,138	3,93162
22	135	1,4-dinitrobenzene	5,26	5,379	-0,119	-2,26236
23	134	1,2-dinitrobenzene	4,45	4,386	0,064	1,4382
24	65	1,2,3,4-tetrachlorobenzene	5,93	5,878	0,052	0,8769
25	108	4-nitrobenzonitrile	4,16	3,956	0,204	4,90385
26	124	2-phenoxyethanol	3,26	3,29	-0,03	-0,92025
27	40	1,3,5-trichlorobenzene	4,04	4,076	-0,036	-0,89109
28	19	4-nitrophenol	4,21	4,312	-0,102	-2,4228

If small values of Q^2_{LOO} indicate not very robust models, characterized by low predictive capacities intern, the opposite is not necessarily true. In fact, if a strong value of Q^2 is a condition necessary of robustness and a possible high predictive capacity of a model, this condition alone is not sufficient, and can lead to an over-estimate of the predictive capacity of the model, when it is applied to really external compounds.

In addition to the test of randomization, L be values RMSE are adapted better, to judge quality of a model, that the values of R^2 and Q^2 only, which constitute good tests only for data regularly distributed.

Les valeurs de tous ces paramètres statistiques, réunies ci-après,

SDEC	= 0,0174	(110 OBJETS)	R^2	= 99,94
SDEP	= 0,0189	(110 objets)	Q^2	= 99,93
SDEP (ext)	= 0,1062	(28objets)	Q^2 (ext)	= 98,03

Suggèrent, tout à la fois, une bonne capacité prédictive (faibles valeurs des RMSE) et une possibilité d'extension suffisante (valeurs proches ou similaires) du modèle.

Tableau V ***Diagnostics d'influence*****:**

Obs	composés	pDL50	pDL50est	ei	di	Hi1	ti	DFITS
1	Benzene	3,4	3.40034969	-0.00034969	-0.01905726	0.311188	-0.02284578	-0.01535561
2	bromobenzene	3,89	3.89025971	-0.00025971	-0.01415357	0.133083	-0.01512422	-0.00592578
3	1.3.5-trichloro-2-hydroxybenzene	4,4	4.39977196	0.00022804	0.01242763	0.090654	0.0129664	0.004094
4	1.2-dichlorobenzene	4,3	4.29995785	4.21E-05	0.00229686	0.09003	0.00239299	0.0007527
5	1-chloro-2-hydroxybenzene	3,84	3.84027832	-0.00027832	-0.01516777	0.083488	-0.01576335	-0.00475764
6	1-chloro-3-methylbenzene	4,33	4.33003572	-3.57E-05	-0.00194671	0.071106	-0.00200843	-0.00055568
7	1-chloro-4-methylbenzene	3,21	3.21011313	-0.00011313	-0.00616531	0.092463	-0.006439	-0.00205528
8	1.3-dihydroxybenzene	3,77	3.77034102	-0.00034102	-0.01858477	0.092531	-0.01941052	-0.00619819
9	1-hydroxy-3-methoxybenzene	3,29	3.290028	-2.80E-05	-0.00152588	0.114035	-0.00161296	-0.00057867
10	1-hydroxy-2-methylbenzene	3,36	3.36015517	-0.00015517	-0.00845639	0.083855	-0.00879019	-0.00265938
11	1-hydroxy-3-methylbenzene	3,07	3.06985145	0.00014855	0.00809562	0.083748	0.00841469	0.002544
12	1-hydroxy-4-methylbenzene	3,48	3.4801671	-0.0001671	-0.00910655	0.066621	-0.00937822	-0.00250551
13	1-amino-2-chlorobezene	3,63	3.62986225	0.00013775	0.00750704	0.075849	0.0077695	0.00222585
14	1.2-dimethylbenzene	3,76	3.76000448	-4.48E-06	-0.00024428	0.083642	-0.00025376	-7.6665E-05
15	1.4-dimethylbenzene	4,38	4.38018691	-0.00018691	-0.01018615	0.081276	-0.01057335	-0.00314486
16	1-methyl-2-nitrobenzene	3,48	3.47987816	0.00012184	0.00663999	0.042608	0.00675177	0.00142436
17	1-methyl-3-nitrobenzene	3,24	3.23990306	9.69E-05	0.00528321	0.054443	0.00540322	0.00129652
18	1-aldehydo-2-nitro-5-hydroxybenzene	3,35	3.35075472	-0.00075472	-0.04113042	0.050358	-0.04199349	-0.00967022
19	1.3-dinitrobenzene	3,8	3.80057315	-0.00057315	-0.0312353	0.118147	-0.03309365	-0.01211317
20	1-amino-2-methyl-3-nitrobenzene	3,79	3.7895796	0.0004204	0.02291079	0.059922	0.0235101	0.00593561
21	1-amino-2-methyl-4-nitrobenzene	3,77	3.76956105	0.00043895	0.02392172	0.044215	0.02434494	0.00523617
22	1-amino-2-methyl-5-nitrobenzene	4,89	4.88969579	0.00030421	0.01657871	0.033251	0.01677605	0.00311125

La suite du **tableau V** ***Diagnostics d'influence*****:**

Obs	ei	di	Hi1	ti	DFITS
23	0.00030922	0.01685175	0.037174	0.01708704	0.00335747
24	0.00091795	0.05002607	0.061236	0.05137106	0.0131203
25	0.00029449	0.016049	0.046833	0.01635533	0.00362536
26	3.95E-05	0.00215396	0.041059	0.00218713	0.00045257
27	0.0001988	0.01083412	0.09074	0.01130434	0.00357109
28	2.11E-05	0.00114968	0.05849	0.00117908	0.00029388
29	6.53E-05	0.00355842	0.051037	0.00363464	0.00084291
30	5.88E-06	0.00032024	0.033994	0.00032438	6.0851E-05
31	3.82E-05	0.00208197	0.060032	0.00213638	0.0005399
32	-3.60E-05	-0.001963	0.053448	-0.00200633	-0.00047676
33	0.0002051	0.01117746	0.076035	0.0115694	0.00331887
34	0.00023501	0.01280748	0.029084	0.01293208	0.00223823
35	-0.00030161	-0.01643702	0.083916	-0.01708643	-0.00517137
36	0.00013871	0.00755936	0.031373	0.00764191	0.00137531
37	-0.00013237	-0.00721385	0.100162	-0.00756624	-0.00252435
38	-0.00018004	-0.00981175	0.116439	-0.01038541	-0.00377011
39	-0.0054661	-0.2978893	0.127713	-0.31746095	-0.12147255
40	0.00095582	0.05208989	0.053941	0.0532838	0.01272318
41	0.00040022	0.02181103	0.049622	0.02225995	0.00508643
42	0.00065918	0.03592372	0.082687	0.03731818	0.01120419
43	0.00040409	0.02202193	0.073434	0.02276216	0.00640802

La suite du **tableau V** ***Diagnostics d'influence*****:**

Obs	ei	di	Hi1	ti	DFITS
44	0.0081199	0.44251501	0.085331	0.46077033	0.14073624
45	9.73E-06	0.00053031	0.089466	0.00055289	0.00017331
46	0.0016708	0.09105458	0.148353	0.09817089	0.04097331
47	0.000929	0.05062826	0.045356	0.05155526	0.0112375
48	0.0006871	0.0374453	0.040572	0.03803551	0.00782161
49	-0.00064153	-0.03496184	0.102547	-0.0367186	-0.012412
50	-0.12302	-6.70429399	0.081626	-9.11565379	-2.71764228
51	0.00042692	0.02326611	0.119215	0.02466523	0.00907435
52	0.00020867	0.01137201	0.080048	0.01179642	0.00347971
53	-0.00018252	-0.0099469	0.12234	-0.0105638	-0.00394404
54	-0.0046895	-0.25556647	0.162068	-0.27785283	-0.12219658
55	0.0022937	0.12500113	0.167626	0.13632593	0.06117725
56	0.00070414	0.03837394	0.198872	0.04265631	0.02125294
57	0.00011102	0.00605032	0.036718	0.00613334	0.00119746
58	0.002506	0.13657097	0.077698	0.14149947	0.04106985
59	-0.00062813	-0.03423157	0.05327	-0.0350035	-0.00830309
60	-0.00046841	-0.02552722	0.251853	-0.0293634	-0.01703674
61	-3.77E-05	-0.00205363	0.088603	-0.00214121	-0.00066762
62	0.0047954	0.26133776	0.159738	0.28373701	0.12371226
63	0.00025529	0.01391269	0.09298	0.01453444	0.00465355
64	0.0015445	0.08417153	0.129696	0.08977151	0.03465503
65	-7.46E-05	-0.00406743	0.051547	-0.0041534	-0.00096827
66	6.35E-05	0.0034606	0.132248	0.00369614	0.00144293

La suite du **tableau V** ***Diagnostics d'influence*****:**

Obs	ei	di	Hi1	ti	DFITS
67	0.00012245	0.00667323	0.082629	0.00693201	0.00208043
68	0.00039827	0.02170476	0.498815	0.03050364	0.03043143
69	0.0021093	0.11495177	0.07351	0.11882763	0.03347112
70	-0.00045729	-0.0249212	0.06775	-0.02568026	-0.00690229
71	0.00027258	0.01485495	0.096536	0.01554933	0.00508277
72	0.00018342	0.00999595	0.065307	0.01028691	0.00271913
	0.00121125	0.06607833	0.067574	0.06808573	0.01832

Obs	pDL50est	ei	di	Hi1	ti	DFITS
90	3.87929373	0.00070627	0.03849002	0.032831	0.03893994	0.00717441
91	4.6295249	0.0004751	0.02589181	0.062451	0.02660492	0.00686649
92	4.90995341	4.66E-05	0.00253893	0.061667	0.00260844	0.00066869
93	3.31963189	0.00036811	0.02006111	0.266725	0.02330867	0.01405775
94	5.06981995	0.00018005	0.00981229	0.065655	0.01009979	0.00267727
95	3.64962722	0.00037278	0.02031561	0.05995	0.02084734	0.00526465
96	5.14940646	0.00059354	0.0323465	0.248248	0.03711821	0.02133009
97	4.315567	0.014433	0.78656377	0.085209	0.82056542	0.25043499
98	6.36920608	0.00079392	0.04326673	0.038618	0.04390412	0.00879938
99	3.5585273	0.0014727	0.08025861	0.049898	0.08192476	0.01877464
100	3.59908413	0.00091587	0.04991271	0.068451	0.05145276	0.01394748
101	4.9588997	0.0011003	0.0599637	0.17911	0.06584877	0.0307585
102	3.92967319	0.00032681	0.01781036	0.077097	0.0184455	0.00533128

103	4.46003179	-3.18E-05	-0.00173248	0.254599		-0.00199712			-0.00116718		
104	5.18069644	-0.00069644	-0.0379543	0.103632		-0.03988558			-0.01356187		
105	6.20039243	-0.00039243	-0.02138649	0.259934		-0.02473434			-0.01465874		
106	4.3008111	-0.0008111	-0.044203	0.428449		-0.05817307			-0.05036676		
107	5.58658	0.13342	7.27106896	0.081596		10.5750698			3.15210586		
108	4.52995636	4.36E-05	0.00237827	0.037622		0.00240983			0.00047647		
109	4.84997554	2.45E-05	0.0013329	0.038781		0.00135496			0.00027216		
110	4.90982151	0.00017849	0.00972728	0.164378		0.01058721			0.00469568		
73											
74					0.0002 4319	0.0132 5327	0.025 901	0.0133 6033	0.0021 7858		
75					0.0002 6407	0.0143 9118	0.017 672	0.0144 4654	0.0019 3766		
76					0.0001 5599	0.0085 0108	0.048 585	0.0086 7131	0.0019 5952		
77					0.0006 0846	0.0331 5961	0.090 774	0.0345 9959	0.0109 324		
78					2.45E-05	0.0013 3476	0.028 738	0.0013 4794	0.0002 3186		
79					0.0005 1613	0.0281 2784	0.069 12	0.0290 0591	0.0079 039		
80					0.0030 045	0.1637 3802	0.065 007	0.1684 9833	0.0444 2951		
81					0.0005 1504	0.0280 6844	0.241 845	0.0320 7272	0.0181 1446		
82					0.0002 4587	0.0133 9932	0.154 104	0.0144 9506	0.0061 8683		
83					0.0002 2512	0.0122 685	0.056 57	0.0125 6702	0.0030 773		
84					0.0003 3762	0.0183 9948	0.229 826	0.0208 5964	0.0113 9494		
85					-2.83E-05	-0.0015 4332	0.069 06	-0.0015 9037	-0.0004 3316		

86	-7.49E-05	-0.0040 7974	0.095 417	-0.0042 7002	-0.0013 8682
87	0.0002 089	0.0113 8455	0.087 239	0.0118 5585	0.0036 653
88	-0.0004 44	-0.0241 9693	0.172 089	-0.0264 5847	-0.0120 6284
89	0.0001 2854	0.0070 0512	0.098 299	0.0073 3972	0.0024 2339

La suite du **tableau V** ***Diagnostics d'influence*****:**

CONCLUSION GENERALE

Nous avons appliqué la méthodologie QSAR pour relier la toxicité [DL50 (96 h)] vis-à-vis du vairon (Pimephales promelas), d'une série de composés organique polluants potentiels de l'environnement aquatique.

Le mélange pris en compte consiste en 138 benzènes substitués, présentant huit groupes fonctionnels : amino, cyano, nitro, hydroxy, fluoro, chloro, bromo et aldéhydo.

The model at summer established by using the networks of standard neurons with three layers [the entries (molecular descriptors), a hidden layer and a layer of exit (DL50 (96h))], with algorithm of training by retro-propagation of the gradient (Levenberg-Marquard).

The 138 source data were burst by chance in two disjoined sets:

- a principal set of 110 elements, burst in its turn by chance in two pennies whole disjoined of 82 elements (together of training) and 28 elements (test unit);

- a whole of 28 elements for the external prediction.

The size of the model (10 molecular descriptors) at summer fixed by maximizing the function the FIT of KUBINYI, and numbers it neurons of the hidden layer (8 neurons) is obtained by minimizing the average quadratic error. The selection of the explanatory variables was carried out by genetic algorithm. by maximizing Q^2, R^2 and $Q^2_{ext.}$.

Statistics: $R^2 = 99.94$; $\sigma_N = 0.0174078$; $F = 16842.2496$ calculated establish the relevance of the developed model.

The quality of the adjustment was checked while carrying out a validation crossed by "leave-one-out "; the value $Q^2 = 99.93$, emphasizes, clearly, the quality of the adjustment.

The test of randomization (figure 10) shows, that only the real vector of the observations leads to high values of the statistics R^2 and Q^2, which proves that the model obtained is not random.

Values RMSE (SDEC = 0.0174; SDEP = 0.0189; $SDEP_{ext.} = 0.1062$; $Q^2_{ext.} = 98.03$) make it possible to be ensured, all at the same time, of the good predictive capacity of the model and its sufficient possibility of extension.

The random choice of the whole of tests which can negatively influence the predictive capacity of the model, the method of retrieval of this whole starting from the source data must be reconsidered. In the same way, the limitations of the model must be more clearly defined, and the possible existence of the aberrant points analyzed carefully.

Lastly, other methods which can prove more advantageous with regard to the precision and the interpretation of the models, and from the point of view of the capacity of generalization, must be tested.

REFERENCE BIBLIOGRAPHIQUE

[1] T E), Bull.

[6] L H. Hall, E L Maynard, L B Kier (1989a). J Approximately. Toxicol. Chem. 8, 431-436. [7] L H. Hall, E L Maynard, L Creighton, Proteins: Structures and Molecular Properties, Second ED, Freeman, New York, 1993, p. 335

[2] L Brooke, D. Call, D. Geiger, C Northcott, Eds, 1984, Acute toxicity of organic CH E micals to Fathead Minnows (Pimephales promelas). Center for hake superior Environmental Studies, University Approximately. Contam. Toxicol. 32, 354-362.

[4] L H. Hall, L B Kier of W(1986), J. In v iron. Toxicol. Chem. 5 333-337.

[5] L H. Hall, L B Kier, G Phipps (1985). J Approximately. Toxicol. Chem. 3, 333-337. isconsin-superior, Superior, WI.

[3] L H. Hall, L.B. Kier (1984B Kier (1989b). J Approximately. Toxicol. Chem. 8, 738.

[8] L B Kier, L Act. Relat., 13, 1994, 285.

[10] N.R To drape, H. Smith, Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New York, 1998.

[11] L Eriksson, J Jaworska, A.P. Worth, Perspective M.T.D., 111 (10), (2003), 1361-1375. H. Hall (1999) M olecular Consonni, Handbook of Molecular Descriptors, R. Mannhold, H. Kubinyi, H. Timmerman eds., Wiley- VCH, Verlage GMBH, Weinheim, 2000.

[13] Mc Culloch-Pitts. Bulletin At math. Biophysics Structure Description. The Electrotopological State. Chap. 9, Academic Press, New York.

[9] H. Kubinyi, As. Struct. –

[12] R. Todeschini, V.1943, Flight. 5, p.115-133.

[14] Mr. Minsky, S.

[16] J J Hopfield. Proceedings of the National Academy of S ciences. The USA. 1982. Vol.79. p. 2554-58.

[17] T Kohonen. Associative coil-organization and memory. Bulletin: Springer-Verlag.984.

[18] R. Hecht-Nielson. Neurocomputing. Addison-Wesly Publishing Company.

1990. 433 p. Papert, Perceptrons. Massachusetts: MIT P ress, 1969.

[15] D. E p.

[19] F Fogelmen-Soulié.. Acts of the National days on the Artificial intelligence Release 7.5 for Windows, Molecular Modelling system, 2000.
[23] R. Todeschini, V Consonni, Mr. Pa. Paris: Teknea. 1988.pp. 275-293.
[20] K HORNIK, 4 (1991. May 06, 2004.
[22] Hyperchem TM v year, DRAGON, Software for the calculation of Molecular Descriptors. Release 5.3 for Windows, Methods connexionnists for the training Milano, 2005.
[24] R. Todeschini, D. Ballabio processing vol. 1. Massachusetts: MIT Press, 1988. 547 14) The Language of Technical Computing The Research, Chem. Inf. Comput. Sci., 36, 54-57.) 251 - 257.
[21] MATLAB Version 7.0 (Release Rumelbart, J L Mcclelland et al.. Parallel Distributed Group MobyDigs MathWorks, Inc
[27] Mr. Petitjean, J Chem. Inf. Comput.Sci. 1992, 32, 331-337.
[28] L.H.Hall, L.B. Kier (1981). Eur. J Med. Chem., 16, 399-407.
[29] S. L Lee, Y. N Yeh (1993). Math. Chem., 12, 121-135.
[30] F.R. Burden (1989). J Chem. Inf. Comput. Sci., 29, 225-227.
[3 1] Mr. Randic, New J Graph Theory. CRC Press, Boca Little rat (FL), 332 pp.
[26] E.V. konstantinova, J 19, 781-791 ** Mr. Randic, J Chem. Inf. Comput. Sci. 1995, 35, 373-382.
[3 2] Mr. Randic, Mr. Razinger (1985). J Chem. Inf. Comput. Sci., 35, 594-606
[3 3] J H. Schuvr, P. Selzer, J Gasteiger Consonni, Handbook of Molecular Of the criptors, R.Manhold, H. Kubiniy H. Timmerman eds., Wiley-VCH, Verl Ag GMBH, Weinheim, 2000. 1992) Chemical
[3 4] R.Todeshini, V 1998,
[3 5] R.Todeshini, Mr. (The Netherlands),, J Am. Chem. Ploughshare, 1996, 36, 334-344. Chel. 1995, V Consonni, A. Mauri and Mr. Pavan Milano Chemometrics and QSAR Professional – Version 1.0 Lasagni, E Marengo, J Chemom.1994, 8, 263-273.
[3 6] R.Todeshini, P gramatica, 3d QSAR in Drug Design vol.2, H. Kubiniy, G Folkers, V C Martin eds., Kluwer/ESCOM, Dobrrecht – 2004.
[25] N Trinajstic (355-380.
[37] V Consonni, R. Todeschini, Mr. Pavan, J Chem. Inf. Comput. Sci. 2002, 42, 682-692.

-ANNEXE :

Valeurs des descripteurs moléculaires utilisés comme variable explicatives

[Les nombres 1, 2 3, en exposant du nom IUPAC désignent l'ensemble d'appartenance du composé considéré :

(1)- ensemble d'assai ;

(2)- ensemble test ;

(3)- ensemble de validation externe].

N°	ARR	nO	PJI2	DECC	BEHp6	DP15	SP09	Mor04p	P1p	R1u+
1	1	0	0	0	0.82	0	0.043	-0.53	0.5	0.129
2	0.857	0	0.333	0.408	0.925	0.003	0.986	-0.204	0.687	0.128
3	0.857	0	0.333	0.408	0.905	0.002	0.776	-0.362	0.63	0.128
4	0.857	1	0.333	0.408	0.936	0.001	0.466	-0.494	0.574	0.22
5	0.6	1	0.667	0.4	1.373	0.091	3.013	-0.24	0.536	0.194
6	0.75	0	0.333	0.5	0.927	0.003	1.208	-0.267	0.593	0.12
7	0.75	0	0.333	0.5	0.999	0.019	1.82	-0.284	0.608	0.127
8	0.75	0	0.667	0.75	1.121	0.133	2.64	-0.168	0.709	0.122
9	0.75	1	0.333	0.5	0.942	0.002	1.024	-0.356	0.565	0.215
10	0.75	0	0.333	0.5	1.03	0.012	1.601	-0.313	0.647	0.218
11	0.375	6	0.75	0.938	1.916	1.507	5.139	0.101	0.58	0.143
12	0.75	0	0.667	0.75	1.121	0.084	2.373	-0.191	0.745	0.213
13	0.75	2	0.333	0.5	1.03	0.003	1.102	-0.477	0.605	0.212
14	0.667	2	0.667	0.593	1.099	0.083	2.517	-0.285	0.743	0.221
15	0.857	0	0.333	0.408	0.952	0.001	0.496	-0.406	0.626	0.159
16	0.75	1	0.333	0.5	0.959	0.001	0.853	-0.398	0.608	0.216
N°	**ARR**	**nO**	**PJI2**	**DECC**	**BEHp6**	**DP15**	**SP09**	**Mor04p**	**P1p**	**R1u+**
17	0.75	1	0.333	0.5	1.041	0.005	1.239	-0.411	0.659	0.221
18	0.75	1	0.667	0.75	1.121	0.037	1.934	-0.36	0.703	0.214
19	0.6	3	1	1	1.121	0.291	3.344	-0.3	0.681	0.182
20	0.6	2	0.75	1.04	1.121	1.041	4.285	-0.186	0.832	0.16

21	0.75	0	0.333	0.5	0.954	0.002	1.036	-0.218	0.57	0.156
22	0.75	0	0.333	0.5	0.962	0.001	0.918	-0.199	0.574	0.156
23	0.75	0	0.667	0.75	1.121	0.051	2.099	-0.215	0.762	0.163
24	0.6	2	0.667	0.6	0.963	0.015	2.017	-0.194	0.575	0.205
25	0.6	2	0.667	0.64	1.05	0.101	2.783	-0.25	0.642	0.223
26	0.5	4	1	0.792	1.438	0.287	3.539	-0.189	0.598	0.182
27	0.6	2	1	1	1.121	0.377	3.508	-0.152	0.744	0.208
28	0.5	4	1	0.778	1.051	0.722	4.084	-0.182	0.63	0.137
29	0.545	2	0.667	0.595	1.721	0.09	2.832	-0.103	0.509	0.183
30	0.545	2	1	0.843	1.675	0.347	3.662	-0.086	0.631	0.213
31	0.545	2	1	0.843	1.721	0.403	3.743	-0.044	0.674	0.2
32	0.545	2	0.667	0.595	1.67	0.12	2.97	-0.208	0.587	0.202
33	0.5	3	0.5	0.611	1.797	0.509	4.168	0.256	0.599	0.227
34	0.545	2	1	0.959	1.735	0.369	3.595	-0.057	0.68	0.214
35	0.545	2	0.667	0.661	1.675	0.12	3.055	-0.088	0.664	0.215
36	0.545	2	0.667	0.661	1.72	0.108	3.005	-0.023	0.653	0.197
37	0.667	0	0.333	0.444	1.017	0.02	2.029	-0.248	0.513	0.119
38	0.667	0	0.667	0.593	1.161	0.134	2.892	-0.153	0.647	0.12
N°	**ARR**	**nO**	**PJI2**	**DECC**	**BEHp6**	**DP15**	**SP09**	**Mor04p**	**P1p**	**R1u+**
39	0.667	1	0.667	0.593	0.962	0.012	1.753	-0.246	0.577	0.205
40	0.667	0	0.333	0.444	1.347	0.046	2.564	-0.291	0.5	0.108
41	0.667	1	0.667	0.593	1.224	0.07	2.577	-0.264	0.602	0.203
42	0.667	0	0.667	0.593	1.237	0.084	2.637	-0.164	0.698	0.213
43	0.667	0	0.667	0.593	1.327	0.09	2.692	-0.098	0.67	0.21
44	0.545	3	1	0.777	1.427	0.411	3.813	-0.159	0.721	0.221
45	0.667	1	0.667	0.593	1.38	0.039	2.219	-0.306	0.64	0.179
46	0.667	1	0.333	0.444	1.359	0.008	1.55	-0.304	0.625	0.172
47	0.667	1	0.667	0.593	1.432	0.037	2.175	-0.285	0.608	0.186
48	0.462	5	1	0.769	1.387	0.802	4.363	-0.135	0.621	0.157
49	0.667	0	0.667	0.593	1.829	0.051	2.339	-0.152	0.669	0.151
50	0.462	4	0.667	0.663	1.915	0.094	3.052	-0.136	0.556	0.152

51	0.545	2	0.667	0.579	1.733	0.305	3.758	-0.06	0.523	0.16
52	0.462	4	1	0.769	1.865	0.846	4.414	-0.008	0.657	0.2
53	0.462	4	1	0.805	1.892	0.769	4.235	-0.089	0.522	0.182
54	0.462	4	1	0.746	1.88	0.401	3.86	-0.024	0.664	0.213
55	0.462	4	0.5	0.615	2.13	0.751	4.328	-0.106	0.528	0.224
56	0.429	4	0.5	0.663	1.986	0.856	4.583	0.126	0.589	0.172
57	0.667	1	0.667	0.593	0.959	0.012	1.717	-0.417	0.658	0.14
58	0.429	4	0.75	1.061	1.881	1.545	5.05	0.036	0.586	0.183
59	0.429	4	1	0.806	1.893	0.867	4.519	0.006	0.578	0.18
N°	**ARR**	**nO**	**PJI2**	**DECC**	**BEHp6**	**DP15**	**SP09**	**Mor04p**	**P1p**	**R1u+**
60	0.429	4	1	0.806	1.866	0.942	4.598	-0.027	0.62	0.199
61	0.429	4	0.5	0.571	2.169	0.829	4.509	0.038	0.571	0.218
62	0.429	4	0.5	0.571	2.082	0.79	4.462	0.1	0.561	0.174
63	0.6	1	0.667	0.4	1.733	0.166	3.398	-0.09	0.522	0.187
64	0.6	1	1	0.76	1.452	0.359	3.598	-0.212	0.765	0.178
65	0.6	0	0.667	0.4	1.162	0.136	3.113	-0.185	0.58	0.113
66	0.6	0	0.667	0.4	1.37	0.147	3.222	-0.211	0.568	0.107
67	0.375	6	0.5	0.656	2.169	1.327	5.136	0.193	0.575	0.171
68	0.5	1	0.25	0.5	1.576	0.263	3.94	-0.149	0.528	0.189
69	0.75	0	0.667	0.75	1.121	0.104	2.499	-0.003	0.76	0.149
70	0.545	1	1	1.025	1.893	3.589	6.001	0.458	0.848	0.149
71	0.667	0	0.667	0.593	1.23	0.069	2.53	-0.135	0.67	0.151
72	0.462	4	1	0.769	1.685	0.809	4.37	-0.044	0.642	0.13
73	0.667	0	0.667	0.593	1.361	0.046	2.352	-0.118	0.681	0.211
74	0.545	2	1	0.843	1.361	0.385	3.759	-0.042	0.658	0.138
75	0.545	1	1	1.025	2.447	0.437	3.726	0.447	0.689	0.167
76	0.6	0	0.667	0.4	1.301	0.077	2.829	-0.083	0.598	0.146
77	0.4	4	1	0.8	1.432	1.189	5.013	0.338	0.571	0.128
78	0.545	0	0.667	0.529	1.407	0.226	3.633	0.037	0.556	0.145
79	0.545	1	0.5	0.661	1.734	0.543	4.149	-0.194	0.517	0.173
80	0.6	0	0.667	0.64	1.324	0.166	3.192	-0.271	0.638	0.168

N°	ARR	nO	PJI2	DECC	BEHp6	DP15	SP09	Mor04p	P1p	R1u+
81	0.6	0	0.667	0.54	1.324	0.03	2.355	-0.45	0.54	0.206
82	0.667	0	0.667	0.593	0.964	0.023	2.024	-0.477	0.581	0.209
83	0.5	1	0.8	1.097	2.169	5.682	7.25	0.795	0.885	0.136
84	0.5	3	1	0.778	1.289	0.979	4.506	-0.071	0.645	0.138
85	0.6	1	1	0.88	1.329	0.47	3.687	-0.194	0.659	0.132
86	0.667	1	1	0.938	1.121	0.421	3.488	-0.206	0.736	0.136
87	0.545	3	0.667	0.661	0.963	0.025	2.361	-0.294	0.566	0.132
88	0.75	1	0.667	0.656	0.959	0.011	1.584	-0.409	0.657	0.142
89	0.375	1	0.857	1.57	2.66	11.614	10.911	1.602	0.947	0.107
90	0.5	3	0.75	0.889	2.111	1.008	4.38	0.081	0.591	0.172
91	0.6	2	0.667	0.64	1.323	0.114	2.98	-0.131	0.653	0.173
92	0.6	2	0.667	0.64	1.238	0.079	2.792	-0.27	0.614	0.176
93	0.667	2	0.667	0.593	0.96	0.011	1.717	-0.386	0.596	0.186
94	0.545	3	1	0.76	1.438	0.207	3.306	-0.167	0.543	0.225
95	0.667	1	0.667	0.593	1.224	0.07	2.578	-0.207	0.633	0.2
96	0.545	3	1	0.777	1.427	0.212	3.33	-0.154	0.584	0.23
97	0.462	4	0.75	0.84	2.136	1.037	4.599	-0.091	0.655	0.174
98	0.5	0	0.25	0.5	0.909	0.068	3.103	-0.369	0.624	0.15
99	0.6	2	1	1	0.941	0.286	3.332	-0.332	0.637	0.134
100	0.545	1	0.667	0.529	1.449	0.177	3.502	-0.141	0.569	0.195
N°	**ARR**	**nO**	**PJI2**	**DECC**	**BEHp6**	**DP15**	**SP09**	**Mor04p**	**P1p**	**R1u+**
101	0.75	0	0.667	0.75	0.932	0.029	1.804	-0.386	0.623	0.156
102	0.462	1	0.5	0.592	0.916	0.263	3.816	-0.378	0.656	0.135
103	0.6	1	0.667	0.54	0.96	0.02	2.165	-0.311	0.582	0.136
104	0.462	3	1	0.769	1.802	1.182	4.724	0.227	0.74	0.206
105	0.545	1	1	0.959	1.416	0.655	4.041	0.123	0.698	0.208
106	0.5	1	0.75	0.917	2.447	1.774	4.858	0.586	0.724	0.156
107	0.545	2	1	0.777	1.051	0.685	3.989	-0.439	0.647	0.188
108	0.545	2	0.75	1.124	1.121	1.73	4.895	-0.416	0.751	0.161
109	0.545	2	1	0.843	1.721	0.401	3.742	-0.007	0.678	0.195

110	0.545	2	0.667	0.595	1.721	0.091	2.855	-0.145	0.506	0.183
111	0.545	2	1	0.959	1.735	0.369	3.595	-0.07	0.68	0.214
112	0.857	1	0.75	0.805	2.326	1.082	4.734	-0.272	0.777	0.127
113	0.545	2	0.667	0.661	1.72	0.108	3.012	-0.096	0.646	0.202
114	0.545	3	0.667	0.545	1.42	0.019	2.274	-0.29	0.561	0.204
115	0.545	3	1	0.959	1.45	0.362	3.58	-0.134	0.699	0.212
116	0.429	4	1	0.796	2.215	0.958	4.687	0.205	0.625	0.156
117	0.8	1	0.8	1	2.556	4.457	6.73	-0.58	0.85	0.124
118	0.429	4	0.5	0.663	1.986	0.856	4.582	0.125	0.589	0.172
119	0.429	4	1	0.806	1.866	0.93	4.589	0.062	0.624	0.196
120	0.429	4	1	0.806	1.893	0.875	4.526	0.023	0.566	0.182
121	0.429	5	1	0.796	1.922	0.912	4.638	0.042	0.65	0.202
122	0.917	1	0.667	0.661	1.532	0.025	2.347	-0.448	0.656	0.167
N°	**ARR**	**nO**	**PJI2**	**DECC**	**BEHp6**	**DP15**	**SP09**	**Mor04p**	**P1p**	**R1u+**
123	0.4	6	0.5	0.64	2.169	1.255	4.965	-0.006	0.5	0.13
124	0.6	2	0.75	0.9	1.198	1.4	4.717	-0.293	0.835	0.147
125	0.8	1	1	0.959	2.536	2.725	5.926	-0.498	0.833	0.101
126	0.5	1	1	0.833	1.66	0.741	4.344	0.237	0.702	0.205
127	0.545	1	1	0.959	1.416	0.655	4.041	0.123	0.698	0.208
128	0.917	0	0.667	0.661	1.499	0.037	2.555	-0.236	0.623	0.111
129	0.545	2	1	0.76	2.113	0.741	4.011	0.048	0.705	0.141
130	0.857	1	1	0.947	2.536	2.473	5.707	-0.771	0.84	0.102
131	0.706	3	1	1.336	2.652	6.109	7.793	-0.555	0.863	0.102
132	0.429	4	0.5	0.571	2.082	0.79	4.461	0.1	0.561	0.174
133	0.857	0	0.333	0.408	0.958	0.001	0.571	-0.404	0.665	0.219
134	0.5	4	0.667	0.667	0.964	0.045	2.71	-0.227	0.557	0.127
135	0.5	4	0.75	1.167	1.121	1.533	4.856	-0.108	0.725	0.121
136	0.545	3	0.667	0.661	1.367	0.1	2.956	-0.203	0.61	0.161
137	0.667	1	0.667	0.593	1.219	0.059	2.4	-0.226	0.616	0.217
138	0.462	4	0.75	1.112	1.881	1.508	4.935	0.027	0.639	0.208