COEFFICIENT DE DETERMINATION Les composés organiques industriels, estimés actuellement à 120 000 avec apparition annuelle sur les marchés de 1000 produits nouveaux, ne sont pas toujours sans risques pour la santé publique et l'environnement.

Les fichiers de données expérimentales, complets, homogènes et précis les concernant, s'ils sont parfois disponibles, peuvent faire défaut même pour les composés du commerce les plus courants et les plus importants.

La détermination expérimentale systématique de toutes les données manquantes qui se traduirait par une lourde charge, économiquement insupportable pour les industriels et l'autorité de régulation, dépasse les capacités de recherche disponibles, nonobstant les larges marges d'erreurs qu'elle pourrait engendrer.

Aussi, la gestion systématique et globale des risques encourus par la présence sur le marché et dans l'environnement de la grande masse de produit chimiques, ne peut reposer uniquement sur la seule disponibilité des données expérimentales. D'où l'intérêt à développer des modèles quantitatifs qui permettent la prévision rapide et précise de la toxicité et de l'évolution dans l'environnement de polluants organiques, à partir de la seule information encodée dans leurs formules structurales.

La concentration d'inhibition 50% de la croissance (CIC 50) d'une population de *protozoaires ciliés* sert souvent d'indice de toxicité, on considère que l'action des polluants se manifeste par un dysfonctionnement des membranes cellulaires et donc la toxicité éventuelle d'une molécule dépend de sa tendance à s'y accumuler. L'octanol, milieu apolaire, constitue un modèle simple des membranes, ce qui explique que de nombreuses relations structure /activité intègrent logP comme variable explicative.

L'analyse de régression est réalisée en utilisant, souvent, la méthode des moindres carrés ordinaire.

L'utilisation de la méthode des moindres carrés dans le modèle de régression linéaire nécessite certaines hypothèses, notamment sur les erreurs.

En effet, l'estimateur des moindres carrés (LS) doit sa popularité en partie au fait qu'il possède sous certaines hypothèses, la variance minimale parmi tous les estimateurs linéaires non biaisés.

Pour construire le modèle et admettre que les coefficients de la régression sont sans biais et convergents, on montre qu'il faut poser comme hypothèses:

a) Les résidus e_i ont une espérance (E) mathématique nulle:

$$E(e_i) = 0$$

b) Le modèle choisi est correct (aucune variable explicative n'a été omise).

c) Les résidus sont indépendants entre eux:

$$E(e_j, e_i) = 0 \qquad \text{si } i \neq j$$

leurs covariances sont nulles.

d) Les résidus ont tous même variance σ^2 (propriété d'homoscédascité).

Par ailleurs, l'emploi de tests statistiques pour analyser la variation expliquée par la régression conduit à admettre que:

e) Les résidus suivent une distribution normale (de Laplace- Gauss).

Il faut de plus mentionner que, même si la majorité des erreurs dans le modèle suivent une distribution normale, il arrive souvent qu'un petit nombre d'observations suivent une distribution différente. Dans ce cas, on dit que l'échantillon est contaminé par des valeurs aberrantes. Puisque les estimateurs LAD sont peu sensibles aux données aberrantes, ils sont particulièrement adaptés à ce genre de situations.

Le but de ce travail consiste à faire une comparaison entre les méthodes LAD et LS en ce qui concerne la modélisation de la toxicité CIC50 de 21 alcools et 9 amines avec l'indicateur d'hydrophobicité logP.

En anglais *least absolue déviations*, la méthode des moindres écarts en valeurs absolues, est une méthode de régression basée sur la minimisation de la somme des erreurs en valeurs

absolues $\sum |e_i|$.

I.1.2- LS :

En anglais *least squares*, la méthode des moindres carrés, est une méthode de régression basée sur la minimisation de la somme des carrés des erreurs $\sum e_i^2$.

MODELISATION:

Propriété d'une substance (poison) capable de tuer un être vivant, pCIC50 signifie log(1/CIC50) servira d'indicateur de toxicité (CIC50 =Concentration d'inhibition 50 % de la croissance.).

La modélisation des données est l'art d'extraire des informations utiles d'un ensemble de données obtenues par des mesures, et de condenser cette information dans un modèle exploitable.

I.1.5- REGRESSION :

Un problème de régression consiste à étudier les changements de la valeur moyenne d'une variable (aléatoire) quand une autre variable ou plusieurs autres variables prennent différentes valeurs fixes. La première variable est appelée variable dépendante ou variable expliquée, les autres variables sont appelées variables indépendantes, variables explicatives. Comme dans notre étude il y a une seule variable explicative, on dit qu'il y a une régression simple; lorsqu'il y a au moins deux variables explicatives on dit qu'il y a une régression multiple.

Il est ensuite possible de quantifier la plus ou moins bonne adaptation de la droite de la régression aux données grâce au coefficient de détermination R^2.

I.1.7- ESTIMATION :

L'estimation est une opération ou action de prédire une grandeur, elle a pour objectif de connaître, à partir de l'observation de l'échantillon, la véritable valeur d'une variable dans la population (sa fréquence, s'il s'agit d'une variable qualitative ; sa moyenne, s'il s'agit d'une variable quantitative).

Mais, du fait de l'incertitude liée aux fluctuations d'échantillonnage, il est impossible de connaître avec certitude la valeur exacte dans la population : on ne peut que l'*estimer* en calculant la probabilité que cette véritable valeur se trouve comprise dans un certain intervalle.

I.1.8- PROTOZOAIRES :

Les Protozoaires, étant unicellulaires, sont de petits organismes de moins d'un millimètre, pouvant s'associer en colonies.

Ils vivent exclusivement dans l'eau ou dans de la terre humide. Ils sont connus pour être responsables de nombreuses maladies telle que la malaria.

I.2- PROBLEMATIQUE :

La méthode la plus utilisée pour estimer les paramètres d'un modèle de régression linéaire simple est sans doute la méthode des moindres carrés (LS) mais cette dernière présente moins de robustesse aux valeurs aberrantes, qui sont assez fréquentes dans la recherche d'un modèle, qui prédit la toxicité pCIC50 de 21alcools et 9 amines en fonction du coefficient de partage (Octanol/Eau) logP, d'où le nécessaire recours à une méthode alternative robuste aux valeurs aberrantes.

I.3- SOLUTION PROPOSEE :

Comme il n'est pas rare de rencontrer le problème des données aberrantes, nous allons modéliser nos données via la méthode LAD, cette méthode consiste à utiliser diverses approches robustes, parmi lesquelles, les trois suivantes exploitées dans cette étude :

(1) La méthode des moindres carrés re-pondérés itérativement, désignée par LAD_1.
(2) L'approche Itérative de base, désignée par LAD_2.
(3) La méthode de la descente directe, désignée par LAD_3.

I.4- ESQUISSE DE LA SOLUTION :

La solution que nous avons proposée passe par les étapes suivantes :
(1) Collecte des données.
(2) Modélisation des données par la méthode des moindres carrés (LS).
(3) Programmation en langage Pascal des trois algorithmes de la méthode des moindres écarts en valeur absolue :
i La méthode des moindres carrés re-pondérés itérativement.
ii L'approche itérative de base.
iii La méthode de la descente directe.
(4) Modélisation des données par les trois approches LAD.
(5) Comparaison des modèles obtenus par les trois approches LAD et le modèle LS.

I.5- PLAN DU MEMOIRE :

Ce mémoire comporte six chapitres, le premier chapitre est une introduction visant à présenter le cadre de notre travail, dans lequel la problématique de cette étude a été exposée. Le deuxième chapitre intitulé état de l'art développe la méthodologie du travail.
Le troisième chapitre parle de l'histoire des deux méthodes LS et LAD, puis la statistique de ces deux méthodes sera détaillée dans le quatrième chapitre. Les résultats seront exposés et discutés dans le cinquième chapitre, et nous achèverons le mémoire par une conclusion générale dans le sixième chapitre.
A la fin de ce mémoire se trouve une bibliographie suivie par une annexe comportant le code source des programmes en langage Pascal.

CHAPITRE II

ETAT DE L'ART

II.1- TOXICITE :

Il est généralement admis que la toxicité de nombreuses substances, particulièrement les produits chimiques organiques industriels, est la conséquence de leur solubilité dans les lipides, alors que leurs caractéristiques moléculaires spécifiques ont peu ou pas d'influence. Leur mode d'action consisterait en la destruction des processus physiologiques associés aux membranes cellulaires.

Les *protozoaires* sont souvent utilisés pour l'évaluation de la toxicité, les méthodes mises en œuvre sont basées sur des critères morphologiques, ultra-structuraux, éthologiques et métaboliques

L'inhibition de la croissance d'une population est un indicateur très en vogue, parce qu'il peut être déterminé directement ou indirectement à l'aide d'un équipement électronique, ce qui permet l'acquisition rapide des observations nécessaires pour les analyses de régression. Nous considérerons la concentration d'inhibition 50% de la croissance (CIC50), dont le logarithme de l'inverse, soit $pCIC50 = \log (CIC50)^{-1}$, servira d'indicateur de toxicité.

II.2- COLLECTE DES DONNEES :

Les tests de toxicité ont été réalisés (Schultz, 1990) en examinant la croissance d'une population de *Tetrahymena pyriformis*. La température a été fixée à 27 ± 1°C et les essais ont été menés dans des erlenmeyers de 250 ml, contenant 50 ml d'un milieu dont la composition est précisée ci après :

Eau distillée	1000 mL
Protéose peptone	20 g
D-glucose	5 g
extrait de levure	1 g
FeEDTA	1 mL d'une solution à3 % (masse/v)
pH	7,35

Ce milieu est inoculé avec 0,25 ml d'une culture contenant approximativement 36000 cellules par ml. La croissance des ciliés est suivie par spectrophotométrie, en mesurant la densité optique (absorbance) à 540 nm après 48 heures d'incubation.

Plusieurs critères ont guidé au choix des composés toxiques examinés. Tous sont disponibles dans le commerce avec une pureté suffisante (95 % et plus), ce qui ne nécessite pas une re-

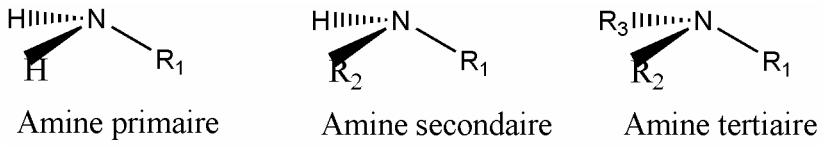
purification préalablement au test. Des précautions ont été observées afin d'assurer une diversité concernant, à la fois, les propriétés physico-chimiques et la position des substituants. Les solutions stocks des divers composés toxiques, ont été préparées dans le diméthylsulfoxide (DMSO) à des concentrations de 5, 10, 25 et 50 grammes par litre. Dans chaque cas, le volume de solution stock ajouté à chaque fiole est limité par la concentration finale de DMSO qui ne doit pas excéder 0,75 % (350 ml par fiole), quantité qui n'altère pas la reproduction de *Tétrahymena*.

II.3- COEFFICIENT DE PARTAGE (OCTANOL/EAU) LOGP:

Le coefficient de partage appelé aussi *Log Kow*, est une mesure de la solubilité différentielle de composés chimiques dans deux solvants (coefficient de partage Octanol/Eau), logP est égal au logarithme du rapport des concentrations de la substance étudiée dans l'octanol et dans l'eau, logP =Log(C_{oct} /C_{eau}), Cette valeur permet d'appréhender le caractère hydrophile ou hydrophobe (lipophile) d'une molécule. En effet, si logP est positif et très élevé,cela exprime le fait que la molécule considérée est bien plus soluble dans l'octanol que dans l'eau,ce qui reflète son caractère lipophile,et inversement. Une valeur de logP=0 signifie que la molécule se répartit de manière égale entre les deux phases et C_{oct} =C_{eau}.

II.4- AMINES :

Découvertes en 1849, par Wurtz les amines furent initialement appelées alcaloïdes artificiels. Une **amine** est un composé organique dérivé de l'ammoniac dont certains hydrogènes ont été remplacés par un groupement carboné. Si l'un des carbones liés à l'atome d'azote fait partie d'un groupement carbonyle, la molécule appartient à la
famille des amides,On parle d'amine primaire, secondaire ou tertiaire selon que l'on a un, deux ou trois hydrogènes substitués.



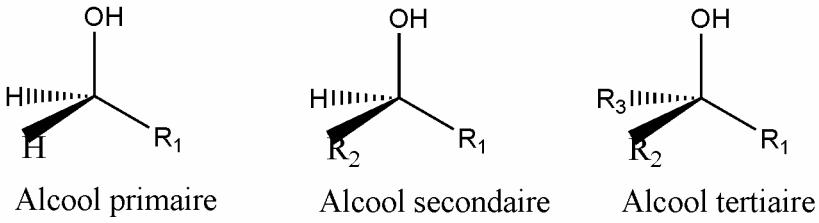
Amine primaire Amine secondaire Amine tertiaire

II.5- ALCOOLS :

En chimie organique, un **alcool** est un composé organique dont l'un des carbones (celui-ci étant tétragonal) est lié à un groupement hydroxyle (-OH). L'éthanol (ou alcool éthylique) entrant dans la composition des boissons alcoolisées est un cas particulier d'alcool, mais tous

les alcools ne sont pas propres à la consommation. En particulier, le méthanol est toxique et mortel à haute dose.

Nous distinguons trois types, l'alcool primaire, secondaire ou tertiaire selon que l'on a un, deux ou trois hydrogènes substitués.



Alcool primaire Alcool secondaire Alcool tertiaire

L'éthanol est une substance psychotrope toxique voire mortelle en grande quantité, même en quantité modérée en cas de consommation régulière, les autres alcools sont généralement beaucoup plus toxiques.

II.6- MODELISATION DES DONNEES :

Les relations LAD et LS ont été déterminées en prenant pour variable dépendante le logarithme de l'inverse de CIC50 en (mmol / litre)**,** et pour variable explicative logP.

Les données ont été modélisées en utilisant la régression par la méthode des moindres carrés de Minitab, et pour la méthode des moindres écarts en valeurs absolues (LAD), à l'aide d'applications programmées en langage Pascal et exécutables sous Windows.

Les données utilisées dans ce travail sont condensées dans le tableau1:

Tableau1 : Données.

N°	Composé	logP	pCIC50
1	Méthanol	-0.77	-2,77
2	Ethanol	-0.31	-2,41
3	Propan-1-ol	0.25	-1,84
4	Butan-1-ol	0.88	-1,52
5	Pentan-1-ol	1.56	-1,12
6	Hexan-1-ol	2.03	-0,47
7	Heptan-1-ol	2.57	0,02
8	Octan-1-ol	3.15	0,5
9	Nonan-1-ol	3.69	0,77
10	Decan-1-ol	4.23	1,1
11	Undecan-1-ol	4.77	1,87
12	Dodecan-1-ol	5.13	2,07
13	Tridecan-1-ol	5.67	2,28
14	Propan-2-ol	0.05	-1,99
15	Pentan-2-ol	1.21	-1,25
16	Pentan-3-ol	1.21	-1,33
17	2-methylbutan-1-ol	1.42	-1,13
18	3-methylbutan-1-ol	1.42	-1,13
19	3-methylbutan-2-ol	1.28	-1,08
20	(ter) pentanol	1.21	-1,27
21	(neo) pentanol	1.57	-0,96
22	1-propylamine	0.48	-0,85
23	1-butylamine	0.97	-0,7
24	1-amylamine	1.47	-0,61
25	1-hexylamine	2.06	-0,34
26	1-heptylamine	2.57	0,1
27	1-octylamine	3.04	0,51
28	1-nonylamine	3.57	1,59
29	1-decylamine	4.1	1,95
30	1-undecylamine	4.63	2,26

II.7- TROIS APPROCHES LAD EXPLOITEES :

Nous présentons dans ce paragraphe les trois approches LAD qui ont été mises en œuvre au cours de ce travail. Tous les algorithmes utilisent les paramètres du modèle trouvés par la méthode LS comme paramètres d'initialisation, pour ensuite calculer les paramètres du modèle LAD, ce qui permet une économie en temps de calcul.

II.7.1- METHODE DES MOINDRES CARRES RE-PONDERES ITERATIVEMENT :

Le principe de la méthode des moindres carrés pondérés consiste à calculer la régression et les résidus, à assigner à chaque individu un poids, d'autant plus faible que le résidu est grand, et à faire une nouvelle régression.

Avec la méthode des moindres carrés re-pondérés itérativement, on fait pareil, mais on itère jusqu'à ce que les paramètres se stabilisent.

L'algorithme effectue une minimisation pour trouver m et b:

(i) $\sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i)^2$, où $w_i = \frac{1}{|Y_i - b_{(n-1)} - m_{(n-1)} X_i|}$.

$b_{(n-1)}$, $m_{(n-1)}$ représentent les paramètres trouvés par l'itération précédente (n-1), et $b_{(n)}$,$m_{(n)}$ sont les paramètres à trouver lors de la présente itération (n).

Prendre les dérivées partielles de (i) par rapport à b, et m.

(ii) $\frac{d}{db} \sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i)^2 = -2 * \sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i)$

(iii) $\frac{d}{dm} \sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i)^2 = -2 * \sum_{i=0}^{N-1} w_i X_i (Y_i - b_{(n)} - m_{(n)} X_i)$

Poser (ii) et (iii) égal à 0, pour trouver le minimum de chaque équation.

(iv) $-2 * \sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i) = 0$

(v) $-2 * \sum_{i=0}^{N-1} w_i X_i (Y_i - b_{(n)} - m_{(n)} X_i) = 0$

(iv) et (v) peuvent être simplifié.

(vi) $\sum_{i=0}^{N-1} w_i Y_i = b_{(n)} \sum_{i=0}^{N-1} w_i + m_{(n)} \sum_{i=0}^{N-1} w_i X_i$

(vii) $\sum_{i=0}^{N-1} w_i X_i Y_i = b_{(n)} \sum_{i=0}^{N-1} w_i X_i + m_{(n)} \sum_{i=0}^{N-1} w_i (X_i)^2$

Dans (vi) et (vii) les seules inconnues sont b et m en utilisant la méthode préférée pour résoudre le système de deux équations de deux inconnues on trouvera (viii) et (ix) :

(viii) $$b_{(n)} = \frac{\sum_{i=0}^{N-1} w_i (X_i)^2 \sum_{i=0}^{N-1} w_i Y_i - \sum_{i=0}^{N-1} w_i X_i \sum_{i=0}^{N-1} w_i X_i Y_i}{\sum_{i=0}^{N-1} w_i \sum_{i=0}^{N-1} w_i (X_i)^2 - \left(\sum_{i=0}^{N-1} w_i X_i \right)^2}$$

(ix) $$m_{(n)} = \frac{-\sum_{i=0}^{N-1} w_i X_i \sum_{i=0}^{N-1} w_i Y_i + \sum_{i=0}^{N-1} w_i \sum_{i=0}^{N-1} w_i X_i Y_i}{\sum_{i=0}^{N-1} w_i \sum_{i=0}^{N-1} w_i (X_i)^2 - \left(\sum_{i=0}^{N-1} w_i X_i \right)^2}$$

Après un nombre suffisant d'itérations n_0; (viii) et (ix) vont définir les paramètres de la méthode des moindres carrés re-pondérés itérativement (LAD_1), La **figure1** reproduit l'algorigramme correspondant.

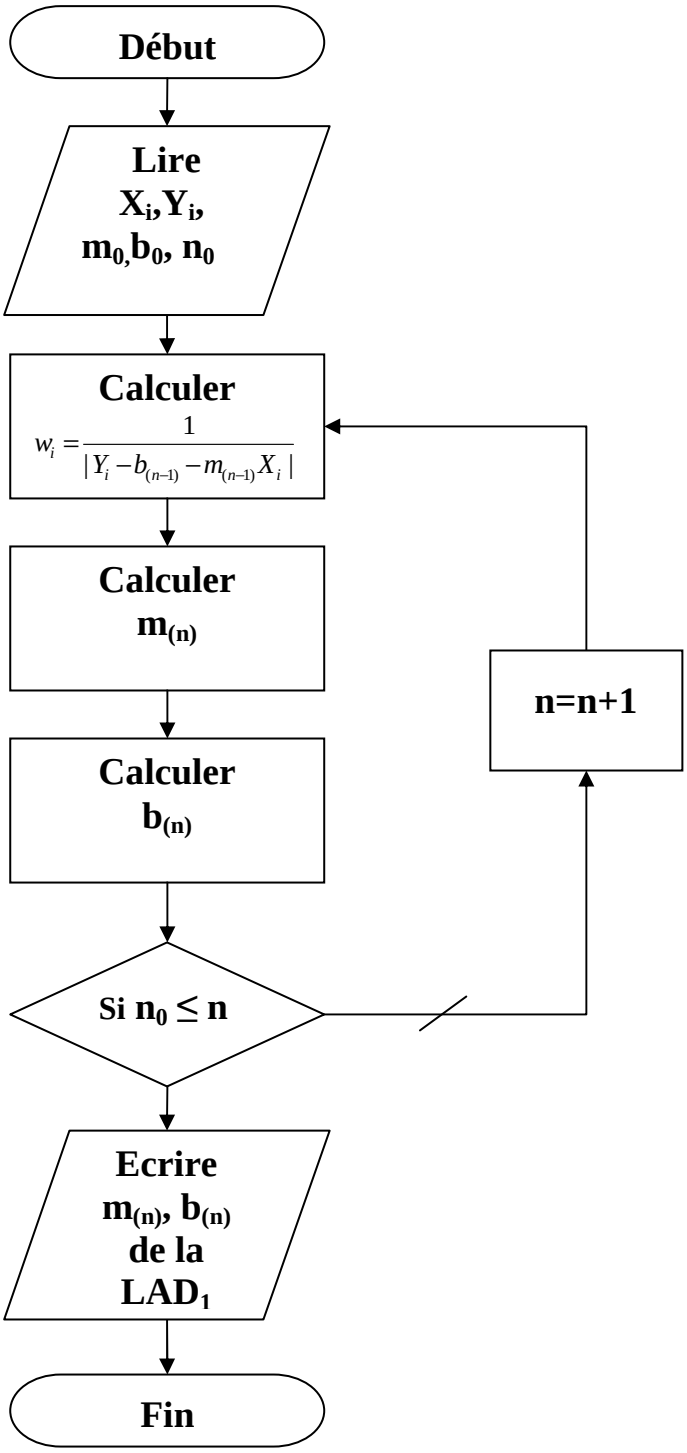


Figure1 : Algorigramme de la méthode des moindres carrés re-pondérés.

II.7.2- APPROCHE ITERATIVE DE BASE :

Cet algorithme se base sur le principe de la médiane pondérée, introduite par Laplace (voir le chapitreIII).

(i) Poser k=0, et trouver m_0 initial trouvé par la méthode des moindres carrés.

(ii) Poser k=k+1et obtenir une nouvelle estimation de b en fixant m_{k-1}.

$$b_k = MED\left(Y_i - m_{k-1} X_i \Big|_{i=1}^{N} \right)$$

(iii) Obtenir une nouvelle estimation de m en fixant b_k.

$$m_k = MED\left(|X_i| \lozenge \frac{Y_i - b_k}{X_i} \Big|_{i=1}^{N} \right)$$

(vi) Une fois que m_k et b_k ne dérivent pas au dessus de l'ordre de tolérance arrêter, m_k et b_k vont définir les paramètres de l'approche itérative de base (LAD_2).

Sinon aller à l'étape (ii).

La **figure2** reproduit l'algorigramme correspondant à l'algorithme itératif de base.

REMARQUE :

◊ : c'est l'opérateur de réplication

Exemple**: A** ◊ **B** produit B répliquer A fois.

MED: c'est la médiane pondérée des données, un poids positif est associé à chaque donnée, elle est déterminée avec la procédure suivante:

$$Y = MED\left(W_i \lozenge X_i \Big|_{i=1}^{N} \right)$$

(i) Calculer $W_0 = \frac{1}{2} \sum_{i=1}^{N} W_i$.

(ii) Associer à chaque donnée son poids correspondant, et classer les X_i par ordre croissant.

(iii) Sommer les poids ordonnés jusqu'à ce que $\sum W_i \geq W_0$

(iv) prendre le premier X_j qui satisfait $\sum W_i \geq W_0$

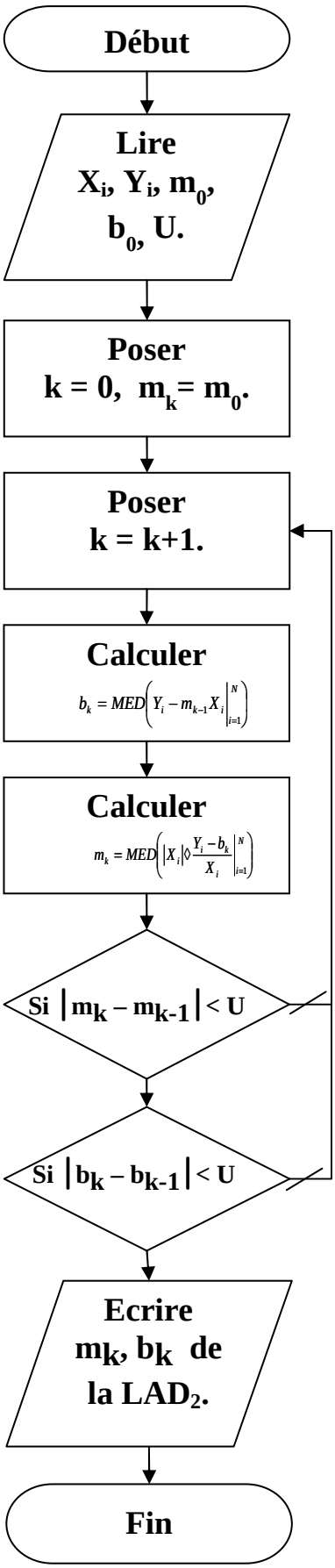


Figure2 : Algorigramme de l'approche itérative de base.

II.7.3- METHODE DE LA DESCENTE DIRECTE:

Cet algorithme se base sur le principe de la médiane pondérée, introduite par Laplace (voir le chapitreIII).

(i) poser k=0, choisir m_0, et b_0 qui sont la solution de la méthode des moindres carrés, et choisir J qui donne $\left|Y_j - m_0 X_j - b_0\right|$ minimale.

(ii) poser k=k+1 et utiliser la médiane pondérée pour trouver *b*:

$$b_k = MED\left(\left|1 - \frac{X_i}{X_j}\right| \Diamond \frac{Y_i - \frac{Y_j X_i}{X_j}}{1 - \frac{X_i}{X_j}} \Bigg|_{i=1}^{N} \right)$$

(iii) Si b_k- b_{k-1} est au dessous de l'ordre de tolérance aller à l'étape (iv), sinon poser J=i et aller à l'étape 2.

(iv) Laisser b* =b_k, m*=$\frac{y_j}{x_j} - \frac{b^*}{x_j}$ ou b*, m* vont définir les paramètres de la méthode de la descente directe (LAD_3).

La **figure3** reproduit l'algorigramme correspondant à l'algorithme de la descente directe.

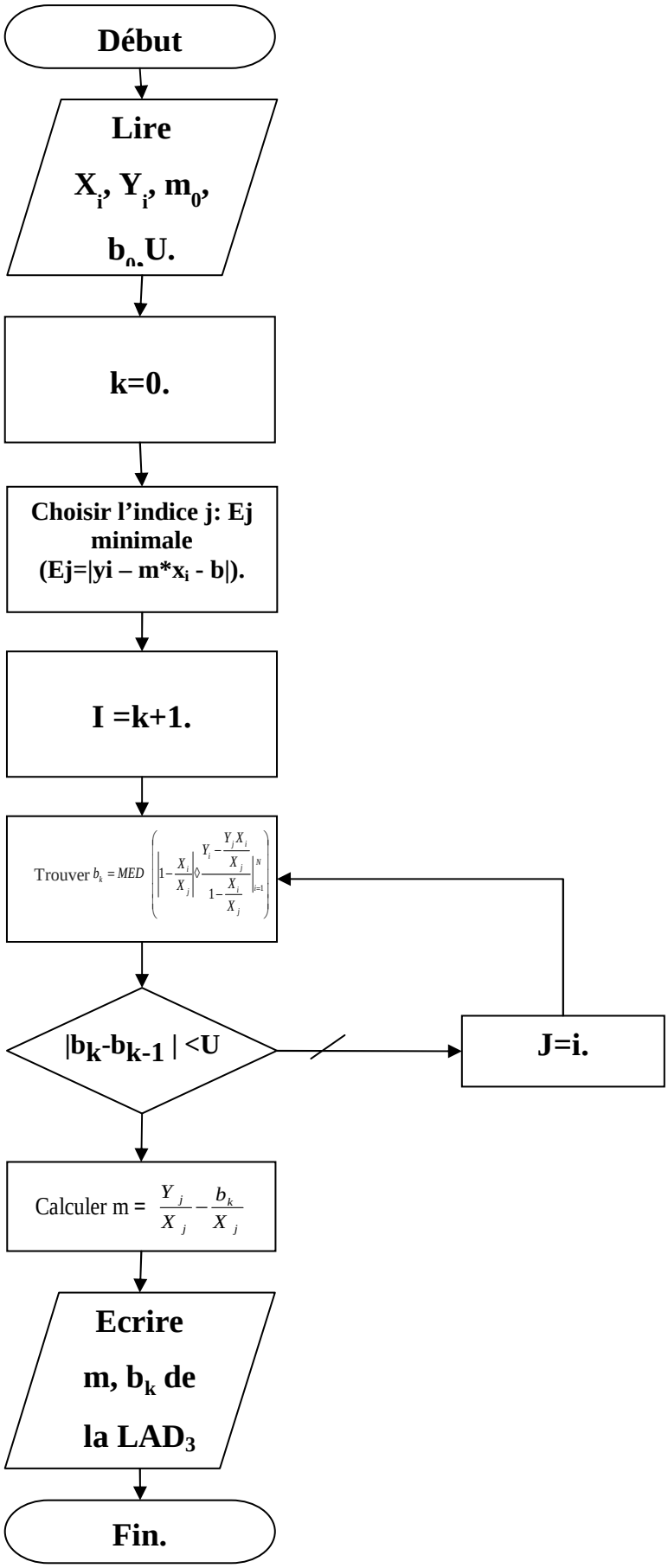


Figure3 : Algorigramme de la méthode de la descente directe.