Chapitre 2

Représentation des données

Nous choisissons de suivre deux exemples de TIPE. Le premier a pour sujet d'étude les arbres de la forêt du Gâvre et leurs dispositions géographiques, le second s'intéresse aux débits de la Loire.

2.1 Etude d'une parcelle de la forêt du Gâvre

On suppose qu'une parcelle d'un hectare a été choisie comme représentative d'une zone composée exclusivement de conifères. L'ensemble a été inventorié et 156 arbres ont été recencés.
Dans un premier temps, tous les arbres sont numérotés. On s'intéresse ensuite à deux grandeurs que sont leur diamètre pris à 1 mètre du sol, mesuré en millimètres, et la distance au plus proche voisin, mesurée elle en centimètres.

Les figures 2.1 et 2.2 qui suivent présentent les résultats obtenus en associant à chaque arbre une croix dont l'abscisse est le rang de l'arbre dans la numérotation retenue et l'ordonnée respectivement son diamètre et la distance au plus proche voisin.

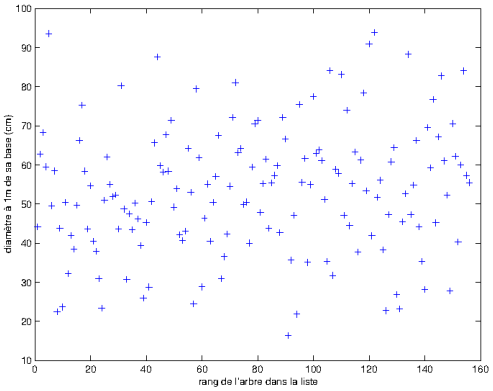


FIGURE 2.1 – nuage de points associés aux diamètres

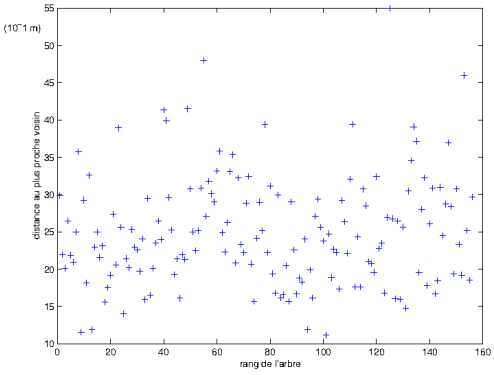


FIGURE 2.2 – nuage de points associés aux distances

C'est loin d'être clair !

Une idée possible pour aider au commentaire pourrait être de relier les points par une courbe ou ligne polygonale. Les figures 2.3 et 2.4 prouvent que le bénéfice est bien mince !

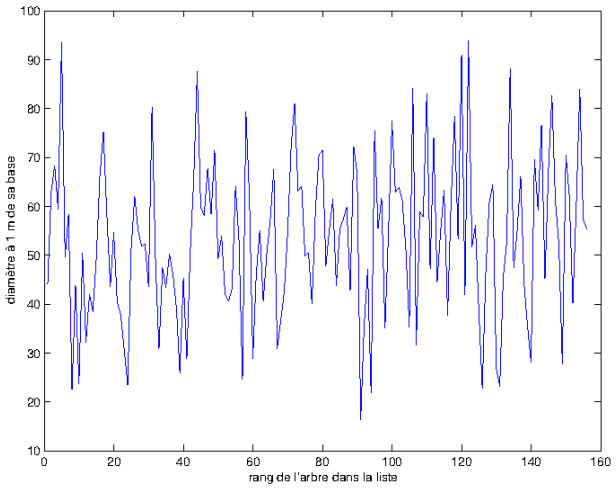


FIGURE 2.3 – Courbe obtenue en joignant les points de la figure 2.1

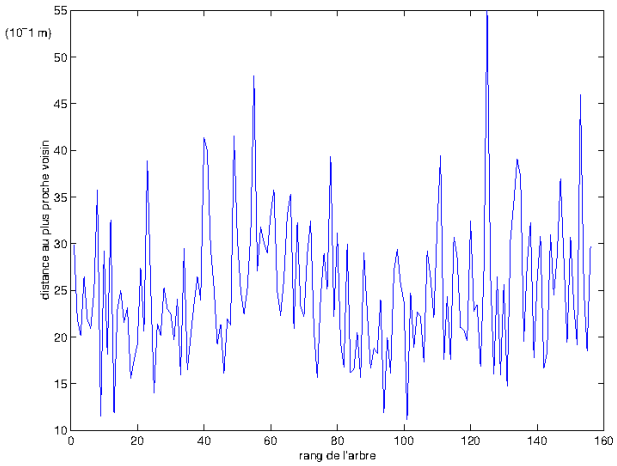


FIGURE 2.4 – Courbe obtenue en joignant les points de la figure 2.2

Nous pouvons conclure que les représentations précédentes ne sont pas adaptées à notre étude, ne serait-ce que parce que les courbes ou les nuages de points obtenues dépendent du choix du rang des arbres, autrement dit du parcours effectué au sein de la parcelle...
Retenons peut-être pour leur défense qu'à leur vue, il semble que la dispersion des valeurs est plus grande dans le cas des diamètres que dans celui des distances au plus proche voisin... c'est mince mais c'est un début !

2.1.1 Diagrammes en bâtons et histogrammes

Les figures précédentes ont permis de représenter l'ensemble $S = \{a_1, a_2, ..., a_N\}$ des résultats recueillis (dans notre cas : $N = 156$). Ininterprétables, elles nous obligent à nous pencher sur d'autres formes de représentations :
La première possibilité est de tracer **le diagramme en bâtons**. Sa construction se développe en trois étapes :

- Trier par ordre croissant les différentes valeurs obtenues, prises dans S.
- Calculer le nombre de fois où chaque valeur est observée, autrement dit leur *effectif absolu*
- Tracer sous forme de lignes verticales les résultats précédents en plaçant en abscisse les valeurs observées et en ordonnée leur effectif.

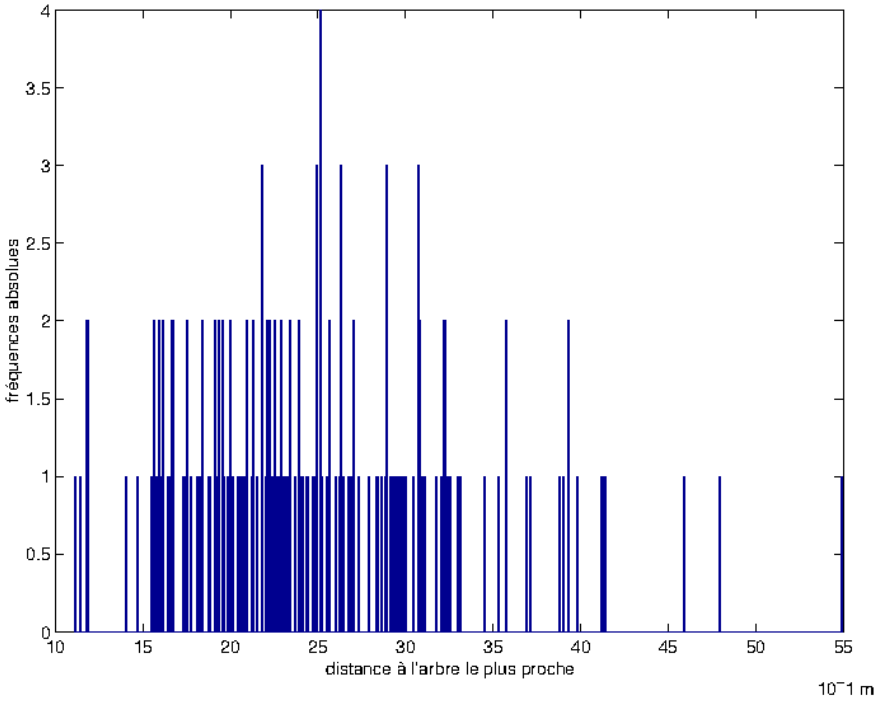


FIGURE 2.5 – Diagramme en bâton représentant la distributions des fréquences dans les distances entres arbres voisins au pas de 1 cm.

L'irrégularité persiste et l'interprétation des résultats est loin d'être évidente. Pour pallier ces défauts, on peut construire dans ce cas un ***histogramme***. On rassemble pour cela les données dans des intervalles dont la longueur commune est appelée *le pas* de l'histogramme.

La question du choix de ce *pas* se pose d'emblée. Nous l'évoquerons un peu plus loin. Choisissons dans un premier temps un pas de 5 cm dans le cas des diamètres et un pas de 10 cm dans celui des distances au plus proche voisin.

On le voit en 2.6 et en 2.7, des informations utiles au commentaire apparaissent, en particulier une certaine symétrie des diamètres des arbres autour d'une valeur comprise entre 40 et 60 cm et une forte assymétrie à gauche de la répartion des distances au plus proche voisin.

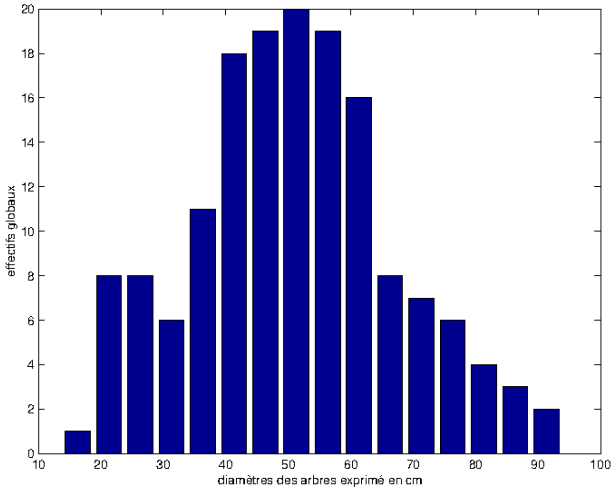


FIGURE 2.6 – diamètres (dm) ; pas de 5 cm

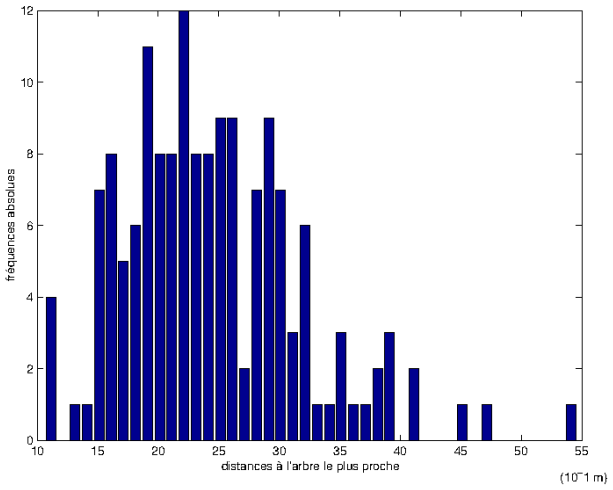


FIGURE 2.7 – distances (dm) ; pas de 10 cm

Supposons que par excès de zèle, nous voulions dans le cas des distances entre les arbres, obtenir des données plus précises, utilisant pour cela le laser d'un oncle maçon. L'histogramme des données est fourni à la figure 2.8 :

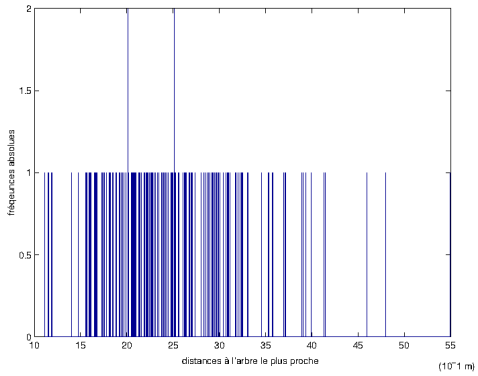


FIGURE 2.8 – Histogramme des 156 distances exprimées en dm, avec deux chiffres après la virgule ; le pas est de 1 mm.

L'information apportée ne semble pas d'une grande pertinence. Le pas n'est tout simplement pas le bon... Le plus simple est de regrouper vos données exprimées en millimètres dans un histogramme au pas de 1 cm. Autrement dit, vous avez très certainement gagné du temps en utilisant le laser de votre oncle mais l'information résiduelle sera sensiblement la même que dans la figure 2.5 issues de vos mesures manuelles, à moins que vous n'en profitiez pour augmenter le nombre de mesures, ce qui sera utile pour estimer les paramètres de la population dont est issue votre échantillon... !

✎ Dans tous les cas, notez qu'il est intéressant de chercher à "lisser" votre histogramme en cherchant des classes qui permettent de discuter son allure générale.
A titre d'exemple, vous trouverez ci-dessous les histogrammes des distances entre arbres voisins en considérant successivement en figure 2.9 un pas de 5 cm, en figure 2.10 un pas de 10 cm, en figure 2.11 un pas de 40 cm et en figure 2.12 un pas de 100 cm.

Nul doute que l'histogramme retenu serait celui de la figure 2.11 qui "subjectivement" semble le plus régulier.

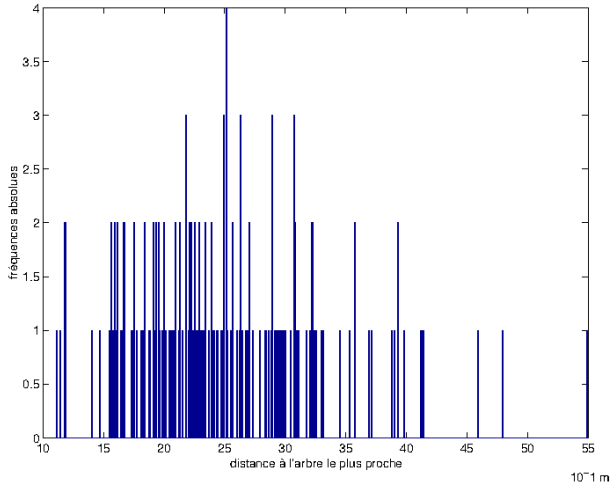


FIGURE 2.9 – Histogramme des distances en dm (un chiffre après la virgule, pas = 5 cm)

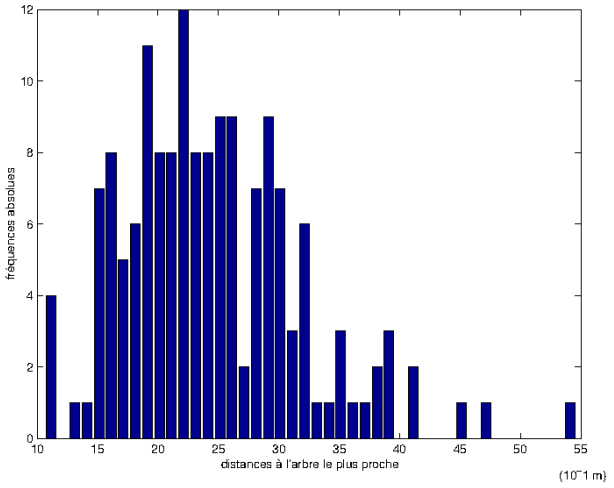


FIGURE 2.10 – Histogramme des distances en dm (un chiffre après la virgule, pas de 10 cm)

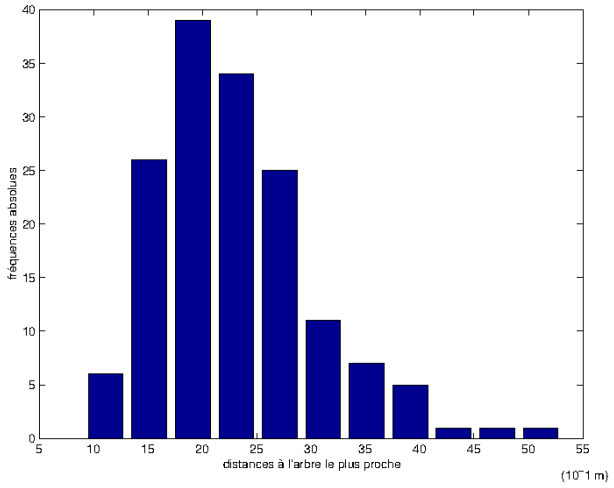


FIGURE 2.11 – Histogramme des distances en dm, (pas de 40 cm)

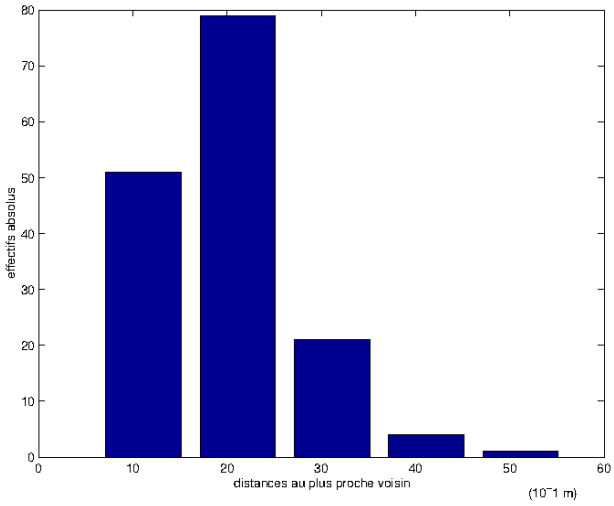


FIGURE 2.12 – Histogramme des distances en dm (pas de 100 cm)

2.1.2 courbes de fréquences cumulées

Une fois l'histogramme tracé, il peut être intéressant de faire le calcul des fréquences cumulées croissantes qui font correspondre à chaque valeur de la série (ou à chaque classe de la série) le nombre d'observations qui lui sont inférieures. Deux cas se présentent :

cas de distributions non groupées

Les distributions de fréquences cumulées sont représentées dans ce cas par un polygone de fréquences. Ce dernier est construit « en escalier »et est représenté par des segments de droites de longueurs proportionnelles aux fréquences absolues ou relatives en les décalant pour chaque valeur de la série vers les haut de telle sorte que l'origine de chacun d'eux soit situé à hauteur de l'extrémité du précédent. Les segments verticaux sont alors rejoints par des segments horizontaux.

Exemple : Voici 18 mesures de diamètres (en cm) de troncs de conifères prélevés à un mètre du sol sur une parcelle de la forêt du Gâvre.

D (cm)	25	31	16	13	13	19	11	32	16	9	26	29	20	27	24	14	12	16

La première opération consiste à trier les valeurs de la série, calculer leur fréquence absolue avant de produire leur fréquence cumulée. Voici le tableau ainsi obtenu [1] :

n	9	11	12	13	14	16	19	20	24	25	26	27	29	31	32
f	1	1	1	2	1	3	1	1	1	1	1	1	1	1	1
F	1	2	3	5	6	9	10	11	12	13	14	15	16	17	18

On obtient alors :

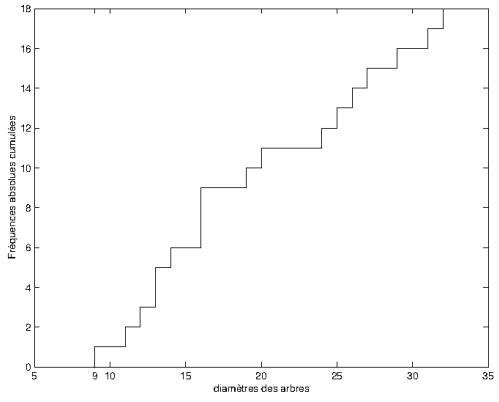


FIGURE 2.13 – courbe des fréquences cumulées pour les diamètres

On apprend, entre autres choses, que la moitié des individus de cette zone ont un diamètre inférieur ou égale à 16 cm puisque la valeur du polygone au point d'abscisse 16 est égale à 9 sur 18 individus...

1. On trouvera en annexe A.3 une fonction Matlab permettant d'obtenir ces résultats

cas de distributions groupées

Si vos données sont regroupées par classes, c'est une ligne brisée que vous tracerez. La méthode consiste à relier les points dont les abscisses sont les limites supérieures des classes et les ordonnées sont égales aux fréquences cumulées croissantes (absolues ou relatives) correspondant aux classes. Par convention, le premier point a pour coordonnées $(x_1, 0)$ ou x_1 désigne la borne inférieure de la première classe.
A titre d'exemple, supposons que nous regroupions les diamètres des arbres par classes d'amplitude 10 cm. Etant donnée les résultats du tableau précédent, nous obtenons les chiffres suivants :

	[0-10[	[10,20[	[20,30[	[30,40[
effectifs absolus (relatifs)	1 (5,56%)	10 (55,55%)	5 (27,77%)	2 (11,11%)
effectifs relatifs cumulés	(0,0556)	(0,6111)	(0,8889)	(1)

Ce qui peut se représenter sous cette forme :

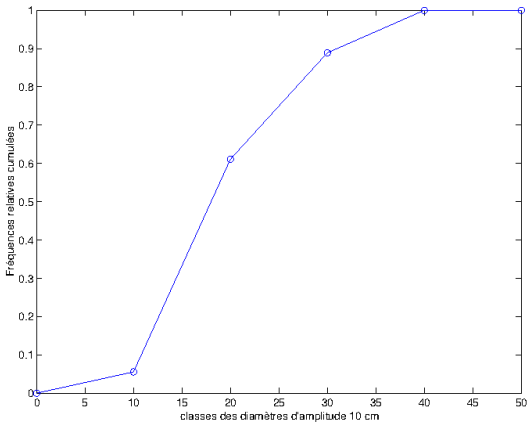


FIGURE 2.14 – courbe des fréquences relatives cumulées

2.1.3 diagrammes permettant la comparaison de données

Supposons qu'une autre parcelle soit choisie et recensée et qu'on souhaite comparer les résultats issus du recensement de ces deux parcelles. Sur la deuxième parcelle, 25% des arbres font moins de 10 cm de diamètre, 50% font entre 10 et 20 cm de diamètre et les 25% restant font entre 20 et 30 cm de diamètre.
Pour comparer ce type de résultat, on retiendra parmi bien d'autres deux types de diagrammes :

les diagrammes à secteurs ou « en fromage »

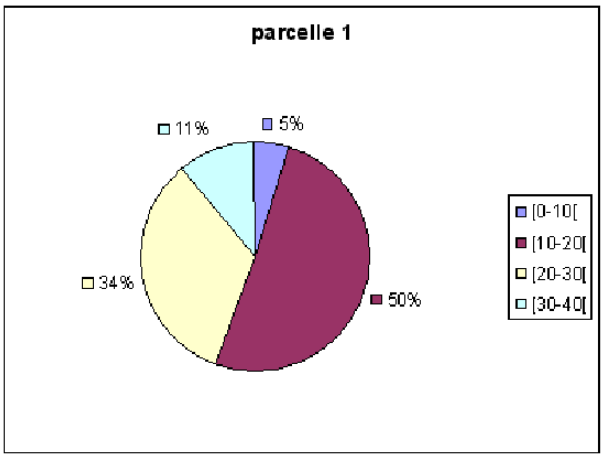


FIGURE 2.15 – diamètres, parcelle 1

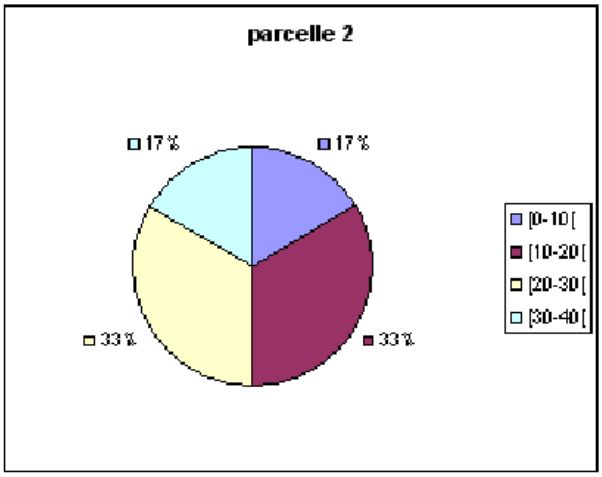


FIGURE 2.16 – diamètres, parcelle 2

les diagrammes linéaires ou « en bâtons »

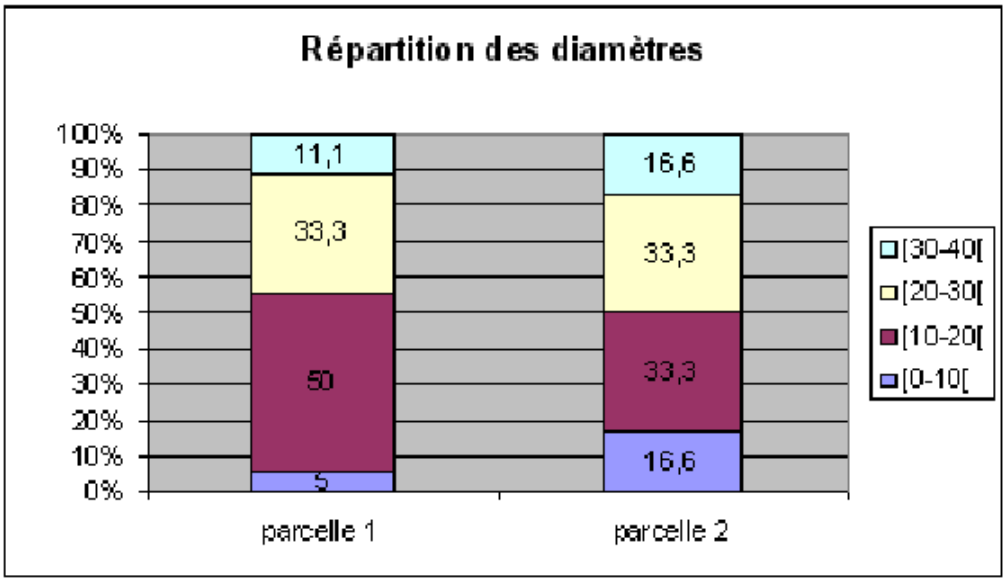


FIGURE 2.17 – diagramme linéaire

♠ ***Mise en garde*** : Attention à l'interprétation car il s'agit ici de fréquences relatives et non de fréquences... Ainsi, il semble que sur la base de nos diagrammes, sur l'exemple ci-dessus, on puisse conclure que les deux parcelles considérées ont une répartition différente des diamètres, la parcelle 2 contenant en particulier une part plus importante d'arbres jeunes.

On omet peut-être de dire que par manque de temps, la seconde parcelle a été constituée seulement de 6 arbres... la conclusion est-elle aussi pertinente ? Et même si la parcelle 2 contenait autant d'arbres que la parcelle 1 (18 arbres ont été recensés) peut-on assurer que les différences sont significatives ?

Pour répondre à cette question, je vous renvoie au chapitre 5 sur les tests statistiques et en particulier au test du Khi2...

2.2 Étude des débits de la Loire

Cette étude est une étude **rétrospective** et porte sur les **données chronologiques** de 18 années de débit. Quel mode de représentation choisir ?
Commençons par fournir un histogramme que nous lissons.

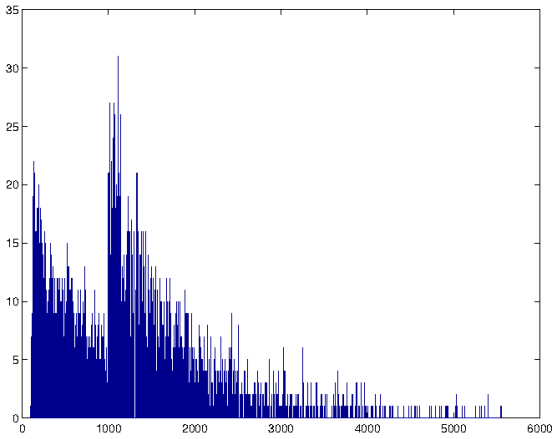


FIGURE 2.18 – débits au pas de 1 m^3/s

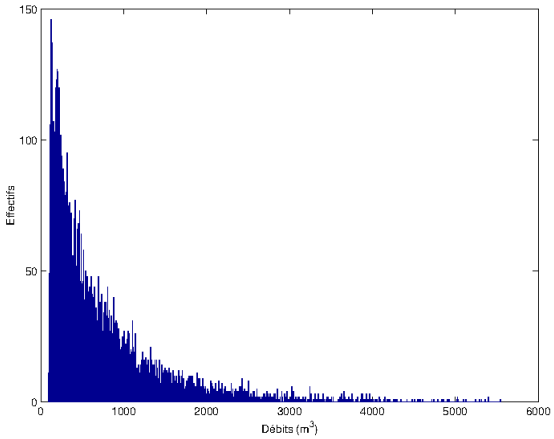


FIGURE 2.19 – débits au pas de 10 m^3/s

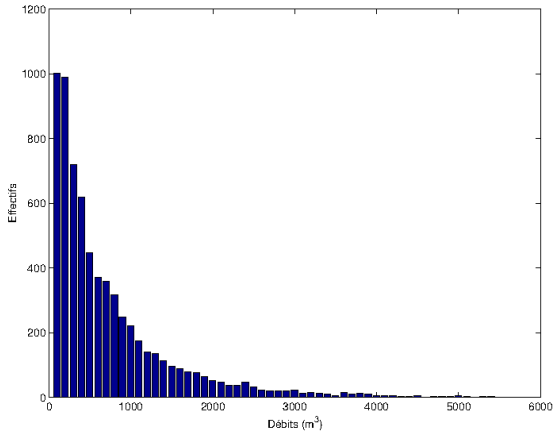


FIGURE 2.20 – débits, pas de 100 m^3/s

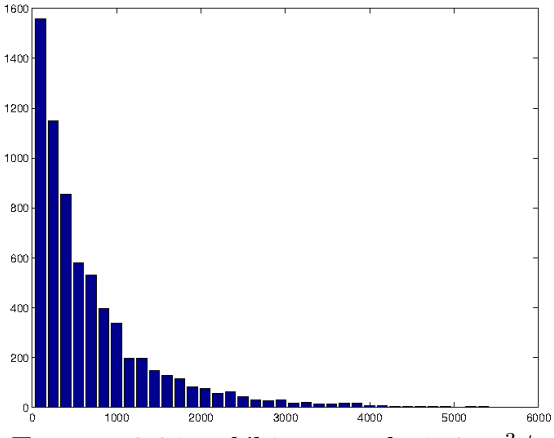


FIGURE 2.21 – débits, pas de 150 m^3/s

considérations pratiques sur les hauteurs des rectangles : Pour l'instant nous ne nous sommes pas posé trop de questions. Les classes étaient d'égale amplitude et la hauteur de chaque rectangle était donnée par l'effectif absolu (ou relatif si on indique un pourcentage). L'histogramme a permis ainsi de visualiser d'un coup d'oeil l'allure générale de la série de donnée.
Imaginez maintenant qu'on souhaite regrouper les débits de crue dans une seule et même classe, le débit de crue étant fixé à $2795m^3/s$ (seuil au delà duquel la fréquence cumulée est supérieure à 0,9726). Le pas désormais n'est plus constant : il est de 50 entre 148 et $2795m^3/s$ et de 2750 ensuite ! ($2,74\%$ des débits sont supérieurs à $2795m^3/s$). Dans ces conditions, si on décide de conserver la même convention pour l'échelle des ordonnées et de représenter un effectif de 208 pour cette classe (soit 208 jours de crue en 18 années) voici ce qu'on obtient :

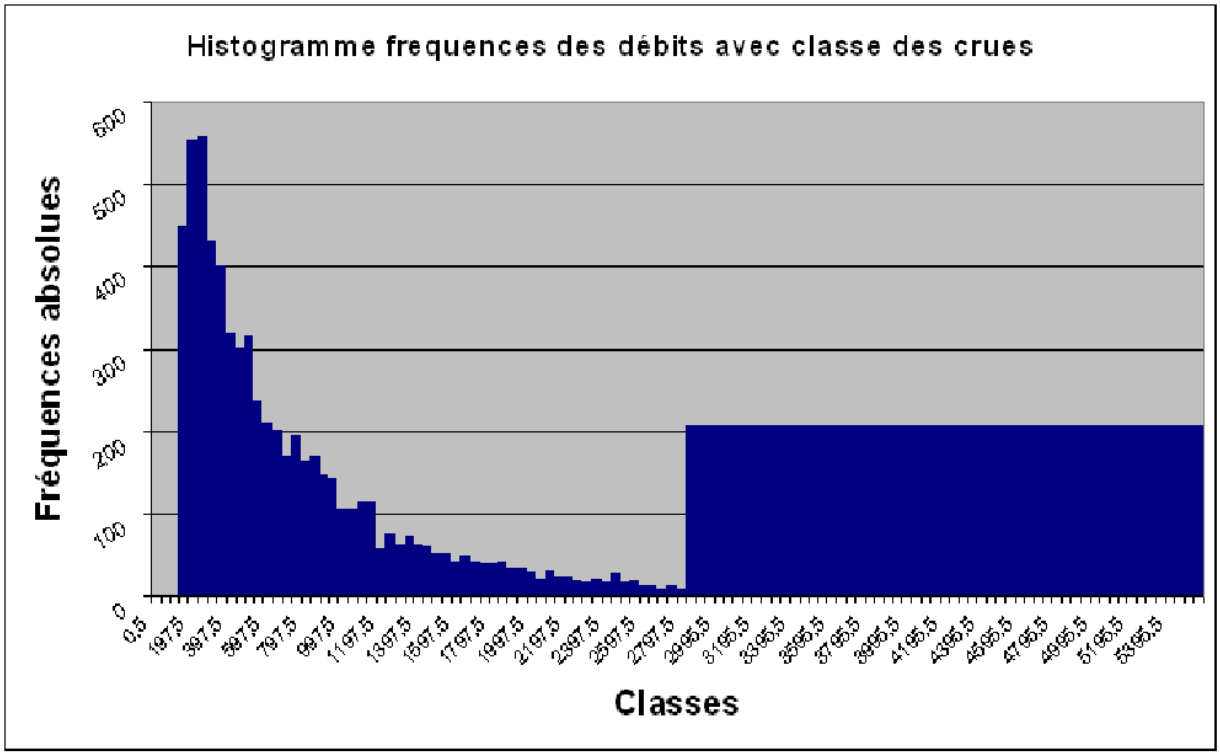


FIGURE 2.22 – débits de la Loire à pas non constant

Cet histogramme devient inexploitable visuellement ! !

✎*Méthode* : Il faut rectifier la hauteur des rectangles pour que la surface de chacun devienne proportionnelle à la fréquence. La surface de chaque rectangle est S =(amplitude de la classe)∗(hauteur du rectangle)$= 50 * h$. Or l'amplitude de la dernière classe est de 2750 soit 55 fois plus grande. Il est donc nécessaire de diviser par 55 la valeur de l'effectif de la classe. Soit ici $208/55 = 3,78$.

La représentation est désormais cohérente... L'erreur est fréquente et doit être corrigée car elle survalorise visuellement l'importance des fréquences de crue. La règle dans ce cas est de considérer la surface des rectangles de l'histogramme...

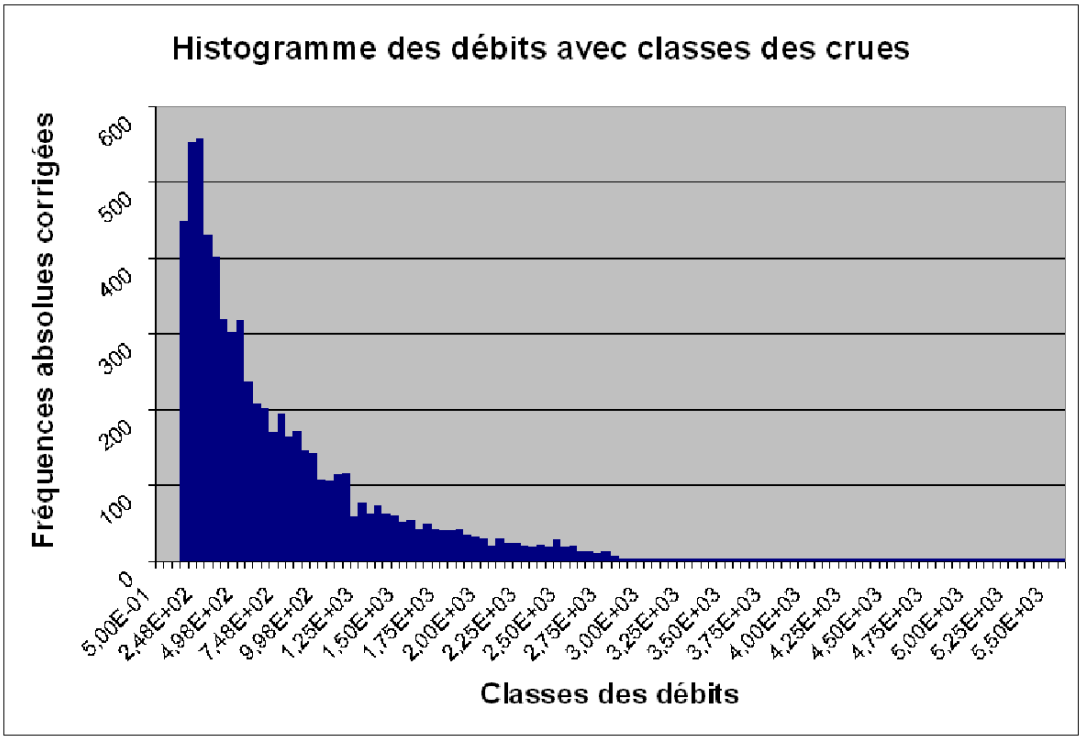
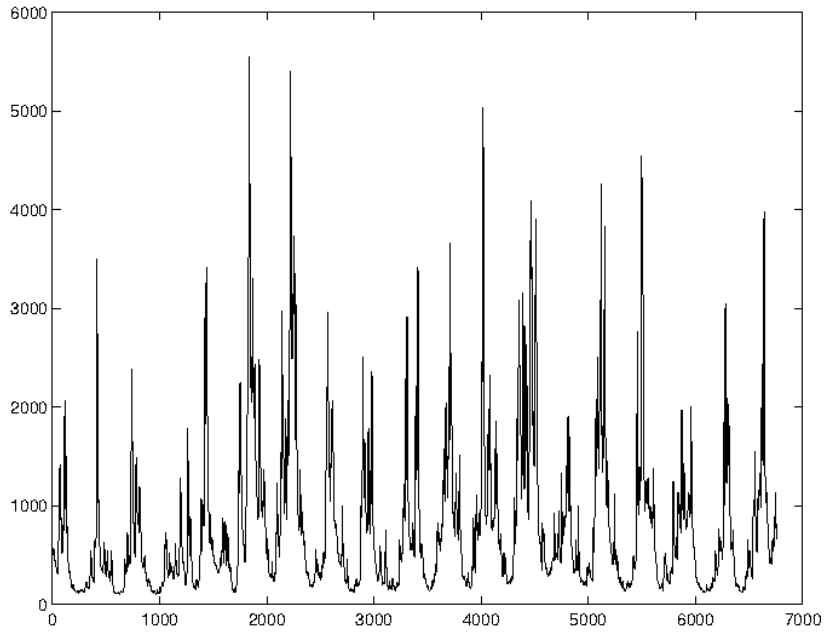


FIGURE 2.23 – histogramme corrigé des débits à pas non constant

Débit rapide... !

Pour terminer, pourquoi ne reviendrions-nous pas sur notre idée de départ lors des relevevés de diamètres d'arbres dans la forêt du Gâvre, celle si vite abandonnée de relier par une courbe les données obtenues ? Nous avons 6570 débits et on pourrait croire la chose absurde mais regardez le graphique obtenu, ci-dessous, en considérant que la variable "débit" est **continue**.



Observez le nombre de pics : 18, c'est-à-dire autant que d'hivers... ! Une information essentielle a été perdue lors des études précédentes, celle du lien chronologique entre ces données, en particulier le rôle des saisons !

2.3 Autres modes de représentation

L'histogramme, les secteurs (2D ou 3D), les diagrammes linéaires, sont loin d'être le seul mode de représentation possible.
Encore une fois, adaptez-vous à vos données. Multipliez les modes de représentation et retenez celle qui apporte le plus d'informations.

Exemple : Lors d'une étude sur l'impact du diamètre du tronc des arbres sur la composition végétale de leur voisinage, une parcelle d'une surface de 100 m^2 a été aléatoirement choisie au sein d'une forêt de chênes. Un repère orthonormé a été déterminé, d'unité le mètre, permettant d'obtenir les coordonnées de chacun des arbres de la parcelle. Le diamètre de chaque tronc, à un mètre du sol, a ensuite été pris.

Les données recueillies ont été les suivantes :

x	-5	-5	-5	-4	-4	-4	-3	-1	-2	0	-1	0	3	3	4	4	5	5
y	1	0	-5	1	0	-5	-4	-5	1	-0,5	-2	3	-2	-2,5	-2	1	1	2
diamètre	25	31	16	13	13	19	11	32	16	9	26	29	20	27	24	14	12	16

L'interprétation de ces données, sans plus d'information, n'a rien d'aisé. Observez maintenant la représentation graphique suivante :

Une fonction matlab a alors été écrite qui associait à chaque arbre le nombre de voisins dans un disque d'influence de rayon égale à t m (ci-dessous $t = 4$). A chaque diamètre, pouvait dès lors être associé deux entiers, le premier égale à l'effectif des arbres dans son disque d'influence, le second égale à leur diamètre moyen.

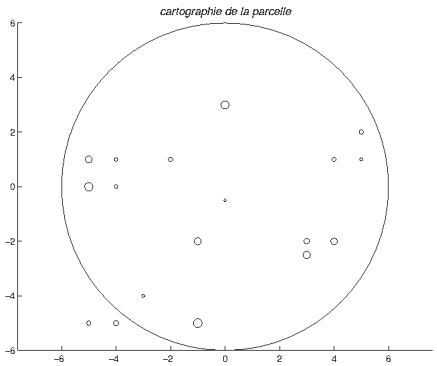


FIGURE 2.24 – cartographie de la zone considérée

Le choix du mode de représentation a été le suivant :

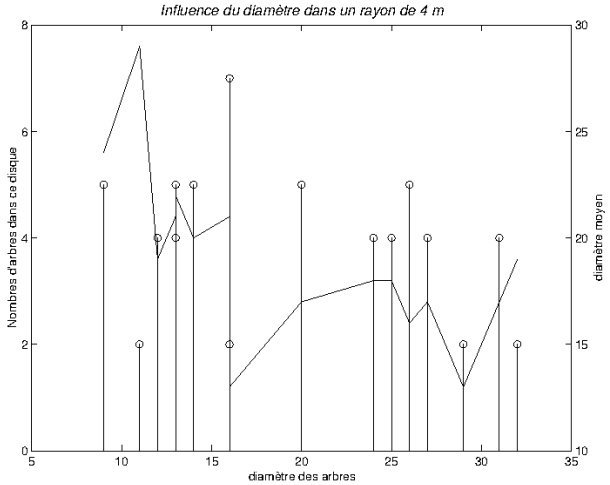


FIGURE 2.25 – Diagramme en bâtons donnant le nombre d'arbres dans un rayon de 4 m en fonction du diamètre de l'arbre situé en son centre et courbe polygonale donnant le diamètre moyen des arbres dans ce disque d'influence

Une interprétation des données devient possible. Il semble en particulier que plus le diamètre de l'arbre est important, moins sa zone d'influence comporte d'individus tandis que le diamètre moyen dans la zone considérée est plus faible.
Cette hypothèse reste cependant à vérifier et à évaluer au regard du nombre d'individus de l'échantillon.