DE LA RECHERCHE À L'INDUSTRIE

cea

INTRODUCTION AUX PLANS D'ÉCHANTILLONNAGE

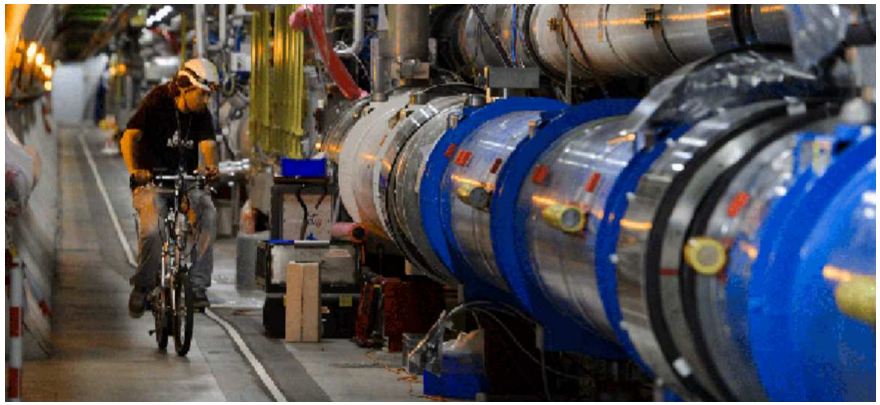
www.cea.fr
www-instn.cea.fr

instn

Introduction

ANALYSER SON ENVIRONNEMENT : 2 APPROCHES

1/ Réglage et contrôle des paramètres – mesure de la réponse



2/ Collecte et enregistrement des paramètres – mesure de la réponse

Yann.kergadallan@cea.fr

CONTRÔLE DES PARAMÈTRES : PLANS D'EXPÉRIENCE

1/ Réglage et contrôle des paramètres – mesure de la réponse

Comportement : « actif »
notion de programme, de procédé, …

Lieu : le laboratoire, l'usine

Méthode : synthèse, production

Outil : projet, **plan d'expérience**

COLLECTE DES PARAMÈTRES : PLANS D'ÉCHANTILLONNAGE

2/ Collecte et enregistrement des paramètres – mesure de la réponse

Comportement « passif »
notion de surveillance, de clustering, ,,,,

Lieu : l'environnement au sens large

Méthode : observation voir analyse destructive,

Outil : prélèvements et **plan d'échantillonnage**

PRINCIPAUX DOMAINES D'UTILISATION

Avec un minimum de données à analyser :

- Identification des écarts dans un milieu :
  Détecter un événement, une structure ou une propriété particulière (rare, inattendue, intéressante ou indésirable)
  Surveillance, contrôle de matière première, pollution, crue, défaillances, …

- Caractérisation d'une population ou du milieu lui-même (combien, sur quelle étendue ?) :
  Sa taille, son étendue, sa densité

- Répartition des individus dans un milieu suivant leurs caractéristiques
  répartition des étoiles, étendue d'une maladie

POURQUOI ÉCHANTILLONNER ?

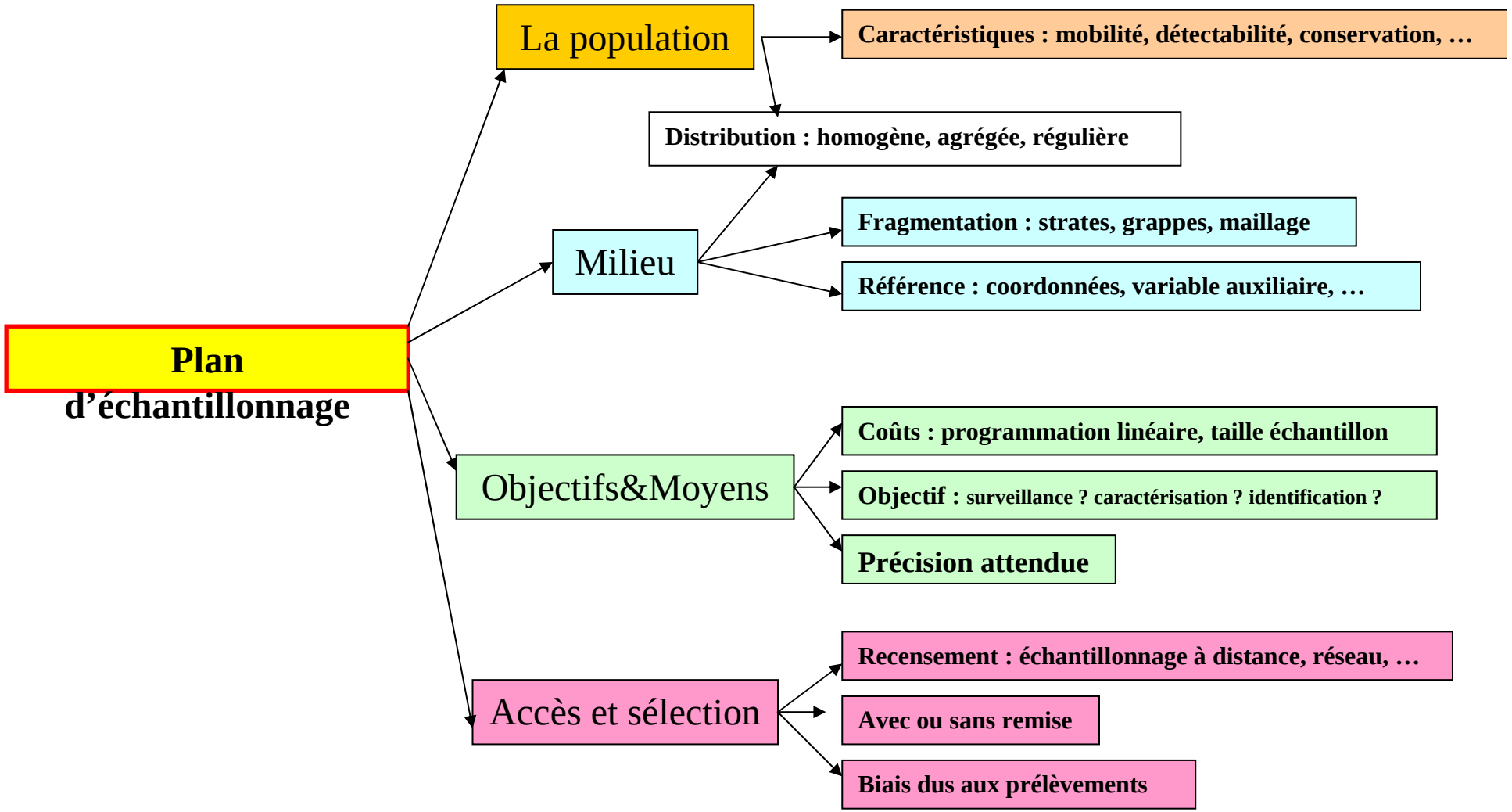
Il est impossible de tout mesurer …. et de tous surveiller
Le plan d'échantillonnage doit permettre à partir d'**un nombre minimal de mesures** de :

1. Déterminer des relations qui peuvent exister entre les différents paramètres et de mettre en évidence les paramètres les plus importants

2. D'inventorier les individus suivant leur propriétés et avec leur effectif. Les ensembles ainsi constitués peuvent être des compartiments qui seront utilisés dans une modélisation

3. De mettre en évidence des structures et de constituer des cartes (voir méthodes de Krigeage)

Ce nombre minimal de mesure est souvent obtenu par des **opérations de regroupement ou de tri conditionnel** sur les unités de la population étudiée


Plan d'échantillonnage
La population
Caractéristiques : mobilité, détectabilité, conservation, …
Distribution : homogène, agrégée, régulière
Milieu
Fragmentation : strates, grappes, maillage
Référence : coordonnées, variable auxiliaire, …
Objectifs&Moyens
Coûts : programmation linéaire, taille échantillon
Objectif : surveillance ? caractérisation ? identification ?
Précision attendue
Accès et sélection
Recensement : échantillonnage à distance, réseau, …
Avec ou sans remise
Biais dus aux prélèvements

NOTIONS PRINCIPALES : TECHNIQUES DE RÉDUCTION

La population (absence/présence) avec son effectif

- Un ensemble de nœud et d'arrêtes pour un graphe
- Un ensemble de parcelles pour une carte
- Un ensemble de durées pour une durée
- Un ensemble de personnes pour une population humaine

Pour faciliter sa compréhension cette **population est regroupée en sous-ensembles appelées strates** suivant leurs propriétés. Plus le nombre de strates est élevé, plus leur effectif est faible

- Regroupement en importance de connexion pour les nœuds d'un graphe
- Regroupement en écosystèmes pour une carte
- Regroupement en saisons pour des longues durées
- Regroupement en habitats pour une population humaine

Mais chacun de ces regroupements peut, à son tour faire l'objet de nouveaux **regroupement successifs donnant plusieurs niveaux ou degré** au plan. Plus le niveau est élevé, plus il y a de subdivision, plus l'effectif ce ces subdivisions est faible.

- Division des connexions suivant leur portée
- Division de la forêt en hêtraies, chênaies puis suivant l'âge des arbres,
- Division en mois puis en jours
- Division en villes puis quartiers puis pavillons

NOTIONS PRINCIPALES : TECHNIQUES DE RÉDUCTION

En fin de regroupement la différences entre les unités d'une classe doivent être suffisamment faible pour suivre **une loi normale**. Par ailleurs, dans le cas des plans multiniveaux, on pourra vérifier l'existence d'une **loi en puissance**.
Cela est primordial **on ne sait bien quantifier que ce qui est purement aléatoire**.
Le problème est de savoir jusqu'où poursuivre cette opération de regroupement.

Une autre technique pour réduire consiste à trier des propriétés ou attributs suivant un **critère de décision**.
Contrairement à la technique précédente où on avait : {regroupement – tirage}, cette technique utilise la séquence : {tirage – regroupement}
Comme pour les plans multiniveaux, cette opération peut être itérée plusieurs fois d'où le nom de ces plans : plans progressifs.

- Si la connexion est supérieure à 2 je répertorie ce nœud et vérifie la connexion des suivant
- Si cet arbre est infecté, il est répertorié et son entourage inspecté
- S'il y a eu un orage, le niveau d'eau est vérifié
- Si ce ménage possède tel magazine, ils sont interrogés sur leur loisirs

Avec cette technique, une **approche bayésienne** semble particulièrement bien adaptée. Le marquage avec recapture appartient à ce type de plan

PRINCIPAUX PLANS D'ÉCHANTILLONNAGE

Plan non structuré dans sa totalité : c'est le plan aléatoire

On pourra cibler sur des unités ou des éléments d'un graphe (arrêtes et/ou nœud)

Plan préalablement structuré par classes (ce qui permet de réduire la totalité à un ensemble de sous groupes)

- Classes de même taille : plan systématique
- Classe d'unités aux propriétés semblables : plan stratifié (une classe = une strate)
- Classe d'unités aux propriétés rangées ou ordonnées : plan multiniveaux (une classe = un niveau). Un niveau pouvant être lui-même constitué de plusieurs strates,

→*Pratiquement un certains nombre de classes sont sélectionnées aléatoirement, puis une autre sélection est effectuée sur les unités des classes qui ont été sélectionnées Chaque étape de sélection aléatoire est un degré*

Les plans progressifs qui sont une sélection itérative jusqu'à réalisation d'un critère

NOTIONS PRINCIPALES : TECHNIQUES DE SÉLECTION

Les deux techniques de base sont le tirage aléatoire ou le recensement

La taille du tirage aléatoire peut être proportionnel à l'effectif des classes (strates ou niveaux). Une optimisation de la taille de l'échantillon est obtenue par **allocation**,

Attention, certains plans dits d'échantillonnage sont en fait des sélections particulières :

- Le plan mélange ou composite est un plan où tous les échantillons prélevés sont mélangés. Suivant l'efficacité du mélange, cela revient à constituer un plan aléatoire empirique.

- Le plan en grappe est en fait un plan subdivisé (en strates ou niveaux) pour lequel toutes les unités des subdivisions aléatoirement sélectionnées sont analysées : autrement dit, un recensement y est effectué.

- Le plan en « boule de neige » qui est l'analogue du plan en grappe mais pour un réseau ou graphe (étude des interactions).

CHRONOLOGIE « TYPE » DE LA RÉALISATION

1. Définir son **cadre** : objectifs, population, milieu, propriétés

2. Identifier les **contraintes** : normes, budgets, effectif, temps, accès, cout), …

3. Rechercher des **informations** déjà existantes (archives, expert, pré-échantillonnage)

4. **Pré**-échantillonnage et regroupement/clustering

5. **Choix** du plan d'échantillonnage

6. Méthode de **collecte**, taille, allocation

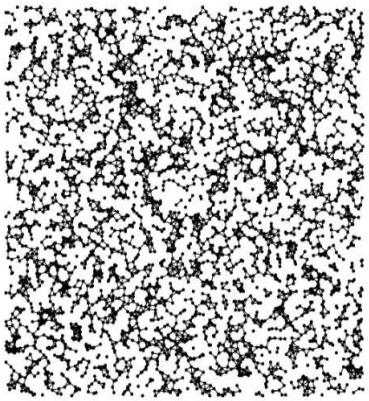
7. **Évaluation** de l'efficacité du plan d'échantillonnage D_{eff}

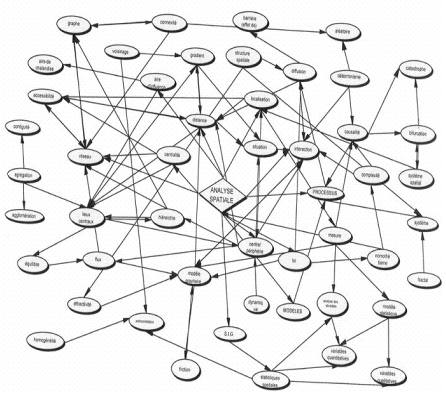
PROPOSITION DE SYNTHÈSE

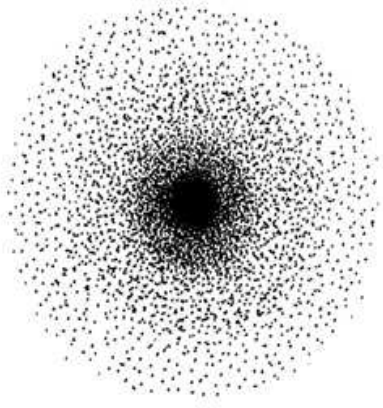
Sélection / Réduction	Aléatoire	Recensement	Ponctuel	Jugement
Aucune	Plan aléatoire			« au pif »
Par division	Plan systématique	Plan en grappe		
Par équivalence	Plan stratifié			
Par ordre	Plan multiniveaux		RSS	
Par tri conditionnel	Plan progressifs	Boulle de neige		
Par jugement	Plan jugement			

Population, statistique et taille de l’échantillon

CHOIX DE LA POPULATION







- Ensemble : collection d'individus
- Système : ensemble d'individus en interactions
- Organisme : système structuré pour la réalisation d'un objectif ou une fonction

EXEMPLES

Un foyer (maison et famille en relation)

Une usine (ensemble d'employés et de machines reliés par la production)

Une communauté (ensemble d'individus reliés par les mêmes valeurs ou mythes)

Une galaxie (ensemble d'étoiles reliées par la gravitation)

Une molécules (ensemble d'atomes en interaction)

Etc …

« machine » qui extrait de l'échantillon
une valeur fixe qui est une estimation de
la valeur vraie inaccessible

Exemple de paramètres à estimer :

- *La moyenne*
- *La population totale T (ou τ)*
- *La variance ou l'écart-type $\sigma^2 = Var(X)=V(X)$*
- *Les proportion p*
- *Le coefficient de variation CV*

On distinguera ainsi :

- La moyenne d'une variable aléatoire, qui est la valeur vraie $\mu°$
- La **moyenne empirique ou expérimentale**, qui est une valeur estimée $\hat{\mu} = \overline{x}$

Tendance des résultats : la moyenne

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Fluctuation ou dispersion des résultats : l'écart type

$$\sigma_x = \sqrt{\sum_{i==1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

Cet écart type caractérise la dispersion <u>intrinsèque</u> de la variable

Par la suite cet écart type sera noté σ_{exp}

ESTIMATION DE LA TAILLE D'UN ÉCHANTILLON

Hypothèse : les variables aléatoires sont indépendantes
Après n données, l'écart type de la moyenne est estimé par : $\sigma_{\bar{x}} = \frac{\sigma_{exp}}{\sqrt{n}}$

Finalement pour une exigence $\Delta°$ donnée, le nombre d'échantillon est :

$$n = \left(\frac{t_{1-\frac{\alpha}{2},n} . \sigma_{exp}}{\Delta° . \bar{x}} \right)^2$$

Règle de Thumb : $n = \frac{16}{(\Delta°)^2}$

Cas de l'utilisation d'un plan d'échantillonnage d'efficacité D_{eff} :

$n_{eff} = \frac{n_{alea}}{D_{eff}^2}$ où n_{eff} est la taille d 'échantillonnage effective

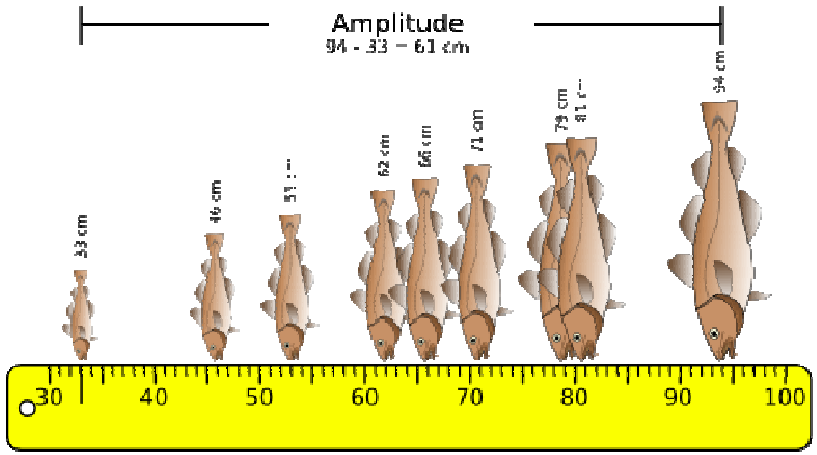
ESTIMATION À PRIORI DE LA VARIANCE

Pour tous les plans d'échantillonnage, il est nécessaire d'évaluer σ_{exp} ou le coefficient de variation CV pour estimer une taille d'échantillon.

→ Faire une bibliographie pour trouver des situations proches
→Faire une étude pilote grossière ou pré-échantillonnage
→approche bayesienne plus adaptée ?

$$\sigma_{exp} \approx \frac{E(X_{max}) - E(X_{min})}{6} = \frac{\text{Etendue des valeurs de X}}{6}$$

$$\sigma_{exp} \approx \frac{R_x}{6}$$



La probabilité de sélectionner une unité u_i pour la mettre dans un échantillon est en général noté : π_i

Population de taille N
Echantillon de taille n → $\pi_i = \pi = n/N = \text{constante}$
Equiprobabilité

Quand n ~ N l'action d'échantillonnage a un effet sur le résultat:
« effet de bord » qu'il faut corriger dans la probabilité de sélection

$$\frac{1}{\pi_i} = \frac{1}{C_n^N} = \frac{n!(N-n)!}{N!}$$

Regroupement et techniques de classification

REGROUPEMENT - CLASSEMENT (*CLUSTERING*)

Deux méthodes de classement :

1/ Celles basées sur **des relations d'équivalence** ou de similitude

- La méthode des Kmoyen ou K-Mean
- La méthode Kohonen
- L'algorithme de Voronoï

2/ Celles basées sur **des relations d'ordre**

- Hiérarchisation

Ces deux méthodes se font par itération de comparaison des individus avec décision suivant un critère

MÉTHODOLOGIE POUR LE CLUSTERING

1. S'assurer que tous ses individus sont quantifiés (variable ou attribut)
2. Se donner une métrique pour pouvoir calculer des distances ou des valeurs
3. Identifier un classificateur à optimiser ou à tester
4. Fixer les conditions d'arrêt (atteinte d'un seuil ou invariance)
5. Établir l'algorithme (programme séquentiel de calcul)
6. Définir les conditions initiales et lancer l'algorithme
7. Définir des indicateurs de performances et de qualité de la classification

MÉTHODE DES K-MOYENS

K-Mean

1. Choisir au hasard K (impair) centres de clusters
2. Calculer la distance entre chaque unité et entre chaque centre de gravité
3. Regrouper chaque unité vers son centre de gravité le plus proche
4. Recalculer les moyennes et les distances
5. Recommencer jusqu'à ce qu'il n'y ai plus aucun changement dans l'attribution d'une unité à une classe

Remarque : Le choix des centres peut produire un rôle important dans le résultat final

EVOLUTION GRAPHIQUE

K-Mean

CLASSIFICATION HIÉRARCHIQUE

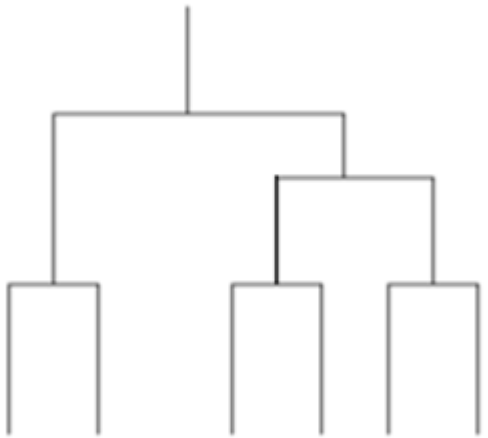
CAH

Approche ascendante :

- *par division*: on démarre avec un seul cluster avec tous les objets et chaque étape éclate les cluster en clusters plus petits

Approche descendante :

- *par agglomération*: au départ chaque objet forme un cluster et chaque étape fusionne des cluster jusqu'au moment où une condition de terminaison est satisfaite



INTÉRÊT DES CARTES SOM

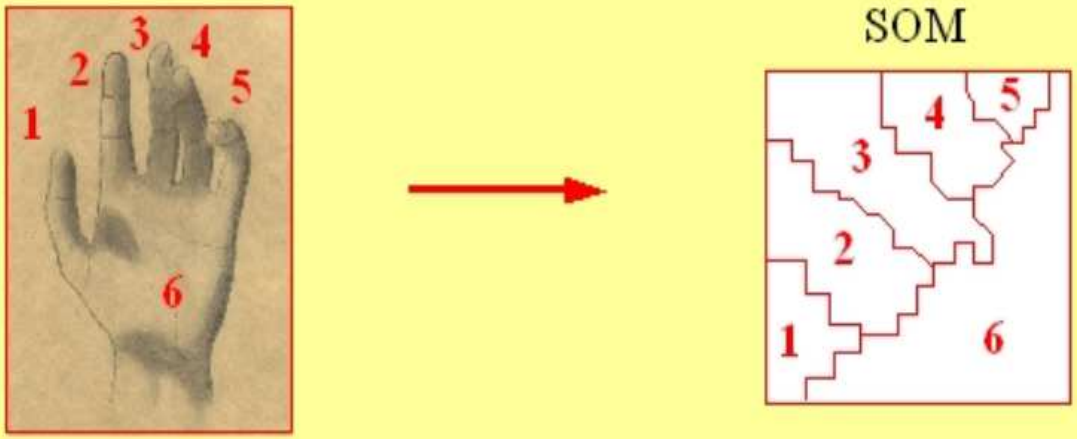
Carte SOM

Principe

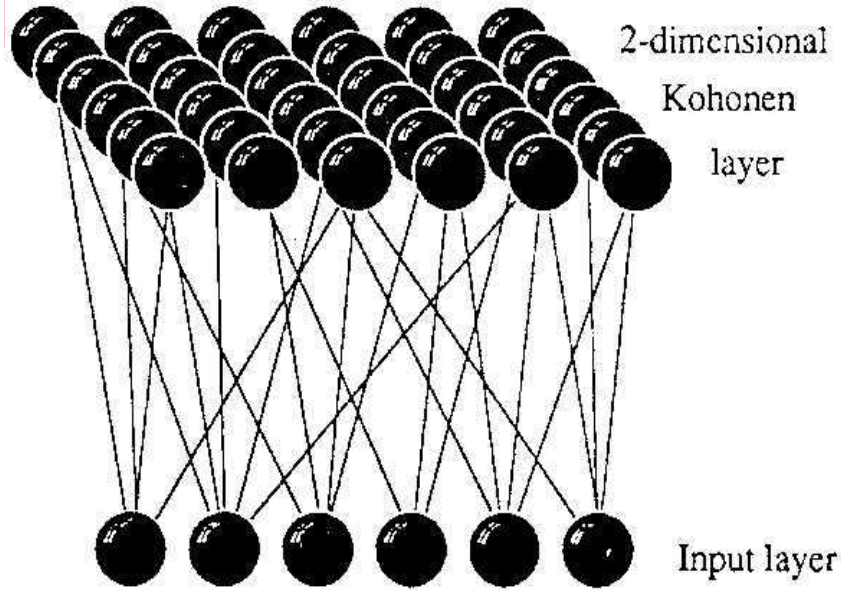
⇒ Classification de données sous formes de cartes dites auto-organisées ou encore topologiques (catégorisation des exemples).

Intérêt

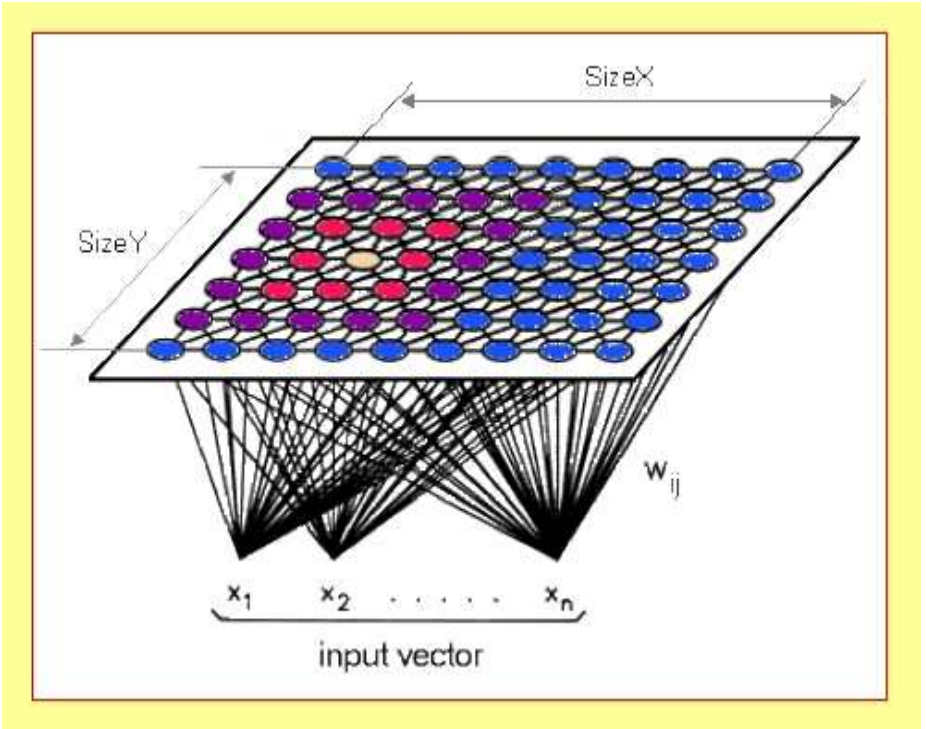
⇒ Préservation de la topologie
(contrairement à d'autres méthodes de classification) :



ASPECT



Carte SOM



REMARQUES

Les unités sont souvent une base de données dont les coordonnées sont les paramètres (pH, t°, …)

Les neurones peuvent avoir plusieurs formes : carrés, hexagones, …

Objectif pour les neurones : chercher à ressembler au mieux aux unités par apprentissage

Pour analyser la ressemblance : utilisation d'une distance (en général euclidienne)

MÉTHODOLOGIE GÉNÉRALE

Carte SOM

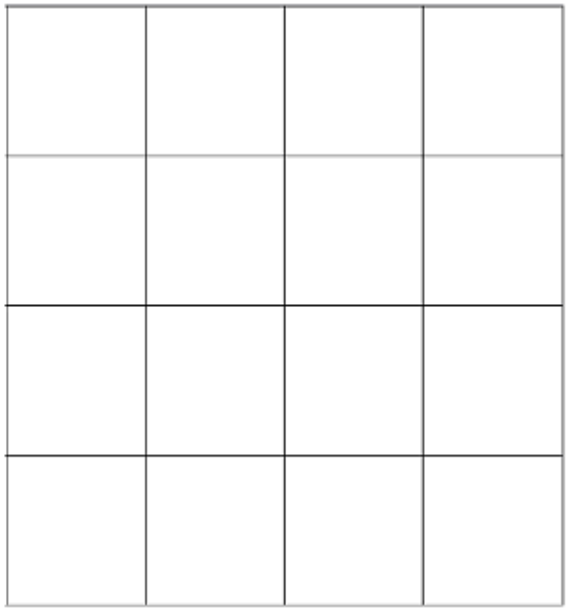
1. Initialisation du réseau en attribuant des poids aléatoire W_{ik} à chaque neurones
2. Sélection d'une unité (vecteur r) dans la base de données
3. Recherche du neurone qui ressemble le plus à cette unité
4. Modification de ce neurone pour qu'il ressemble encore plus à l'unité
5. Diffusion ("contamination") de cette ressemblance sur les neurones voisins
6. Présentation à nouveau d'une unité et même démarche
7. Quand stabilisation : analyse de la couche résultats

ÉTAPE 1 : INITIALISATION DES NEURONES DE LA COUCHE D'ENTRÉE

Carte SOM

- Échantillon:
  n=300 unités, chaque unité possède 2 paramètres : {v1, v2}
- Carte initiale :
  4×4 = 16 neurones repérés par leur coordonnées (i , j)

Initialisation des neurones :
(tirage aléatoires de 16 vecteurs)



Neurone	w1	w2
(0,0)	0,9898	-0,013
(0,1)	-0,1185	0,0554
(0,2)	0,9558	-0,0028
(0,3)	1,0014	0,8654
(1,0)	0,0254	0,0837
(1,1)	-0,0028	0,1384
(1,2)	0,8239	1,0066
(1,3)	-0,1408	-0,0698
(2,0)	1,0268	-0,1004
(2,1)	-0,205	-0,0085
(2,2)	0,8941	0,9218
(2,3)	1,054	-0,0148
(3,0)	1,0924	1,0158
(3,1)	1,0668	0,0893
(3,2)	0,9496	1,0345
(3,3)	1,1077	0,0165

ÉTAPE 2 ET 3 : CALCUL DES DISTANCES

Carte SOM

Parmi les 300 unités on sélectionne aléatoirement celle-ci :

(v_1=0,05015; v_2 =0,9385)

Elle est présentée aux neurones.

Pour évaluer la proximité : calcul de la distance euclidienne

$$d_{v,w} = \sqrt{\frac{(v_1 - w_1)^2 + (v_2 - w_2)^2}{2}}$$

Si on désire moins d'influence des valeurs extrêmes prendre la distance de Bray&Curis

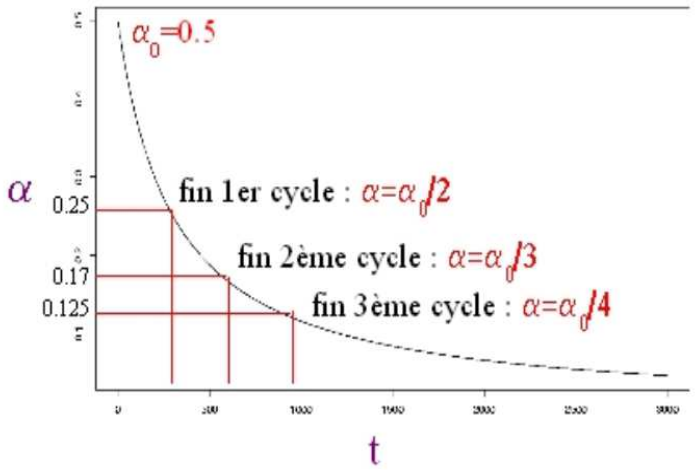
Neurone	D
(0,0)	0,6688
(0,1)	0,1224
(0,2)	0,6441
(0,3)	0,8660
(1,0)	**0,0189**
(1,1)	0,0489
(1,2)	0,8461
(1,3)	0,1778
(2,0)	0,7041
(2,1)	0,1944
(2,2)	0,8360
(2,3)	0,7139
(3,0)	0,9840
(3,1)	0,7189
(3,2)	0,9202
(3,3)	0,7498

Le gagnant
(winner)

ÉTAPE 4 ET 6 : FORMULES POUR LA MODIFICATION

La diffusion sur les neurones voisins se fait suivant l'expression :

Carte SOM



$$\gamma(t) = \alpha(t).\beta(t)$$

$$\alpha = \frac{\alpha_0}{1 + \frac{t}{n}}$$

$$\beta(t) = EXP(-\frac{(r - r_{gagnant})^2}{2.\eta^2})$$

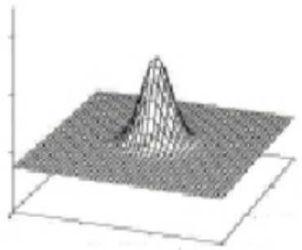
t : le nombre d'unités déjà présentées

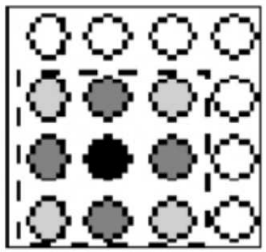
α : coefficient d'amplitude

α_0 : coefficient initial (**en g^{ral} 0,5**)

β : coefficient de diffusion

r : coordonnées des neurones





$$\eta = \frac{\eta_0}{1 + \frac{t}{n}}$$

η : coefficient de voisinage

η_0 : coefficient initial

ÉTAPE 4 ET 6 : PARAMÈTRES POUR LA MODIFICATION

Carte SOM

Valeurs des paramètres :

- t = 1
- Gagnant (1;0)
- η_0= 1,5
- α_0= 0,25
- n=300

$\Rightarrow \alpha \approx 0{,}25$ et $\eta \approx 1{,}5$

Neurone	$r\text{-}r_{gagnant}$	beta	Gamma
(0,0)	1	0,8007	0,2002
(0,1)	1	0,8007	0,1998
(0,2)	2	0,4111	0,1028
(0,3)	3	0,1353	0,0339
(1,0)	0	1	0,2510
(1,1)	1	0,8007	0,2004
(1,2)	2	0,4111	0,1027
(1,3)	3	0,1353	0,0340
(2,0)	1	0,8007	0,2002
(2,1)	1	0,8007	0,2002
(2,2)	2	0,4111	0,1029
(2,3)	3	0,1353	0,0339
(3,0)	2	0,4111	0,1028
(3,1)	2	0,4111	0,1028
(3,2)	2	0,4111	0,1028
(3,3)	3	0,1353	0,0338

MODIFICATION DU NEURONES ET DE SES VOISINS

Pour réattribuer des coordonnées aux neurone voisins

Carte SOM

$$w_i = w_i + (v_i - w_i).\gamma(t)$$

w1	w2	w1-v1	w2-v2	Gamma	w'1	w'2
0,9898	-0,013	0,9397	-0,1069	0,20017027	0,8017	0,0084
-0,119	0,0554	-0,1687	-0,0385	0,19976289	-0,0848	0,0631
0,9558	-0,0028	0,9057	-0,0966	0,10279342	0,8627	0,0072
1,0014	0,8654	0,9512	0,7715	0,03385198	0,9692	0,8393
0,0254	0,0837	-0,0247	-0,0101	0,25101215	0,0316	0,0863
-0,003	0,1384	-0,0529	0,0446	0,20037807	0,0078	0,1295
0,8239	1,0066	0,7738	0,9127	0,10273973	0,7444	0,9128
-0,141	-0,0698	-0,1909	-0,1637	0,03404924	-0,1343	-0,0643
1,0268	-0,1004	0,9766	-0,1943	0,20018431	0,8313	-0,0615
-0,205	-0,0085	-0,2552	-0,1024	0,20023511	-0,1539	0,012
0,8941	0,9218	0,8439	0,8280	0,10285579	0,8073	0,8367
1,054	-0,0148	1,0038	-0,1087	0,03387129	1,02	-0,0112
1,0924	1,0158	1,0423	0,9220	0,10275353	0,9853	0,9211
1,0668	0,0893	1,0166	-0,0046	0,10279363	0,9623	0,0898
0,9496	1,0345	0,8994	0,9406	0,10284634	0,8571	0,9378
1,1077	0,0165	1,0576	-0,0774	0,03375567	1,072	0,0191

Les plans triviaux

- Plan aléatoire (*random sampling*)
- Recensement (*census*)
- Plan de jugement (*judgemental sampling*)

RECENSEMENT, PLAN DE JUGEMENT ET AUTRES …

Ensemble des préjugés possibles (**causes** possibles), Idée que l'on se fait

Prédiction des valeurs paramètres pour l'ensemble des **effets** attendus

Adaptation
Correction de ses préjugés
Maximum de vraisemblance

Ensemble des résultats (effets) obtenus

L'Absolu
Les paramètres vrais

Biais

Calcul des paramètres

Estimateurs

Ensemble des estimations possibles pour les paramètres à déterminer

Echantillonnage & mesure

Population à décrire ou caractériser

PLAN ALÉATOIRE

Correspond plus à la façon dont sont sélectionnées les unités

Les autres façons de sélectionner étant :

- Sélection de l'unité suivante suivant la valeur de l'unité présente (v plan progressif)
- Sélection de l'ensemble des unités en relation avec l'unité présente (v plan boule de neige)

PLAN ALÉATOIRE (RANDOM SAMPLING)

Pour être représentatif, il faut disposer d'une **liste exhaustive** des individus

La liste peut être cause de biais dans sa réalisation

La sélection se fait à l'aide d'un tirage aléatoire (importance de l'algorithme utilisé), chacun des individus ayant la **même chance d'être sélectionné**

La probabilité de sélectionner une unité u_i pour la mettre dans un échantillon est en général noté : π_i

1. Population de taille N
2. Echantillon de taille n $\rightarrow \pi_i = \pi = n/N = \text{constante}$
3. Equiprobabilité

La sélection suppose de disposer d'un générateur d'aléa

•*On ne sait pas caractériser parfaitement le hasard : tout générateur artificiel (algorithme) de nombre aléatoire est nécessairement pseudo-aléatoire*

ECHANTILLONAGE SANS REMISE

Après sélection, les unités peuvent être :

- réintroduites : **échantillonnage avec remise**
- gardées : **échantillonnage sans remise**

Quand n ~ N l'action d'échantillonnage a un effet sur le résultat

Il en résulte un « effet de bord » qu'il faut corriger dans la probabilité de sélection

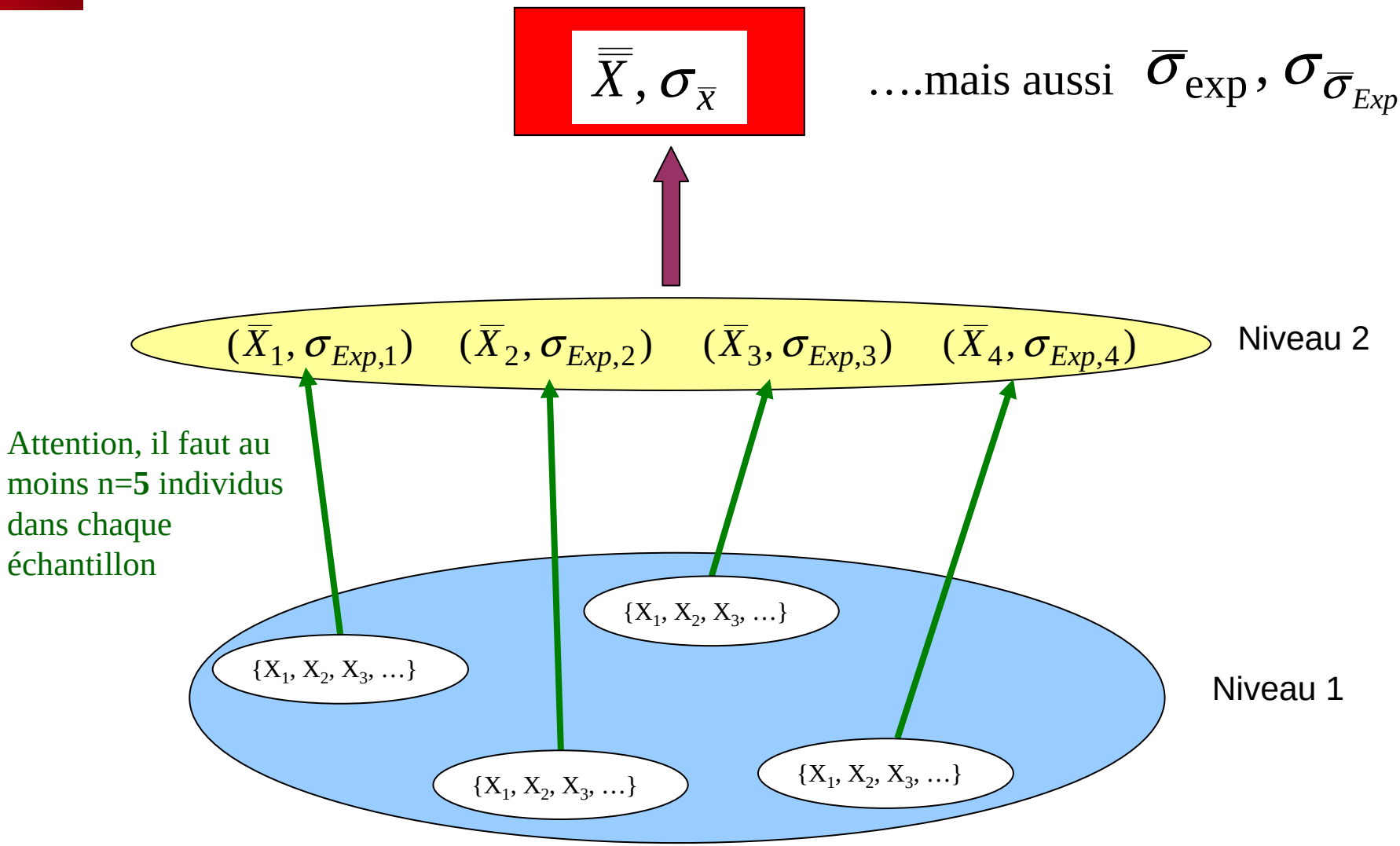
$$\frac{1}{\pi_i} = \frac{1}{C_n^N} = \frac{n!(N-n)!}{N!}$$

Deux types de variance à prendre en compte :

- Celle qui est propre à la population
- Celle qui est reliée au mode d'échantillonnage utilisé

→ Pour évaluer "l'erreur" sur la **moyenne estimée** produite en ne faisant **qu'un seul échantillonnage** on considère une autre statistique : celle de la moyenne estimée qui varie avec le mode et le nombre d'échantillonnage, donc en faisant **plusieurs échantillonnages.**

NOUVEL APPORT D'INFORMATION : PLUSIEURS ÉCHANTILLONAGES



INTUITIVEMENT (INFLUENCE DU NOMBRE D'ÉCHANTILLONNAGE)

Nombre **infini** de mesures :

- Moyenne : valeur constante dite « valeur vraie »
- L'incertitude sur cette moyenne : nulle

Nombre **restreint** de mesures :

- Moyenne : une estimation de la valeur exacte
- Il existe une incertitude sur la moyenne

Quelques mesures :

- Moyenne : estimation **très** grossière de la valeur exacte
- Incertitude sur la moyenne : très importante

VARIANCE DES MOYENNES (INDISCERNABILITÉ DES INDIVIDUS = PLAN ALÉATOIRE)

$$\sigma^2_{\bar{x}}\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) = \frac{1}{n^2}.\sigma^2(x_1 + x_2 + \ldots + x_n) = \frac{n}{n^2}.\sigma^2_{exp} = \frac{\sigma^2_{exp}}{n}$$

Tous les échantillonnage sont indépendants les uns des autres (sinon il faut aussi prendre la covariance)
Les variables aléatoire X_i suivent la même loi

σ_{exp} est le même pour tous les échantillonnage

Finalement :

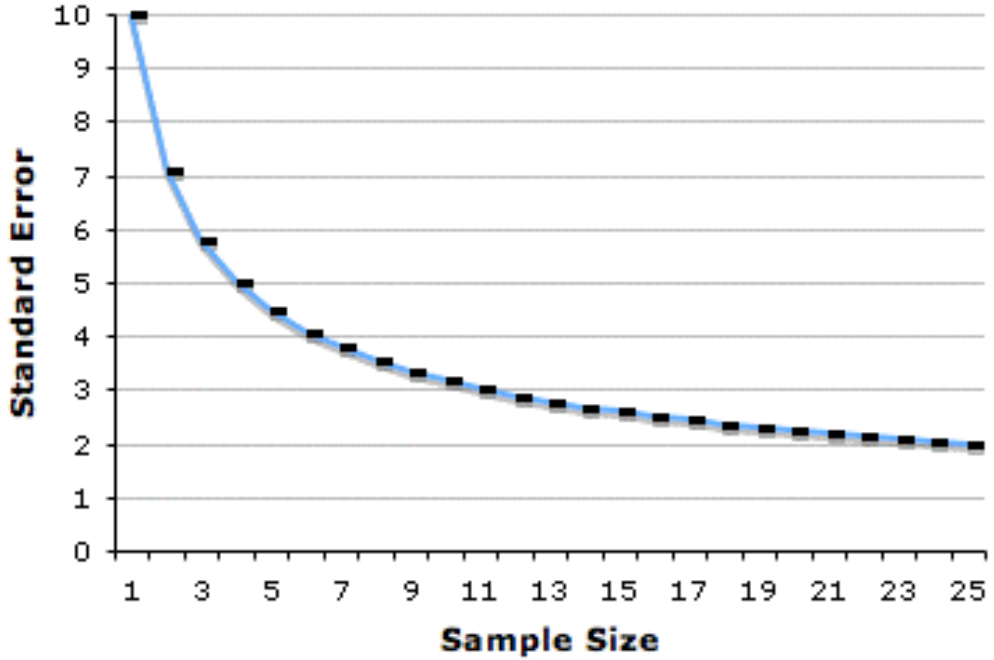
$$\sigma^2_{\bar{x}} = \frac{\sigma^2_{exp}}{n}$$

Cela suppose que σ_{exp} est constant pour tout l'échantillon (variabilité constante)

CONSÉQUENCES : ON RETROUVE LE RÉSULTAT INTUITIF

Cet écart type sur la moyenne prend en compte

1. L'écart type expérimental (qui caractérise les fluctuations intrinsèques au phénomène)
2. Le nombre d'échantillons réalisés



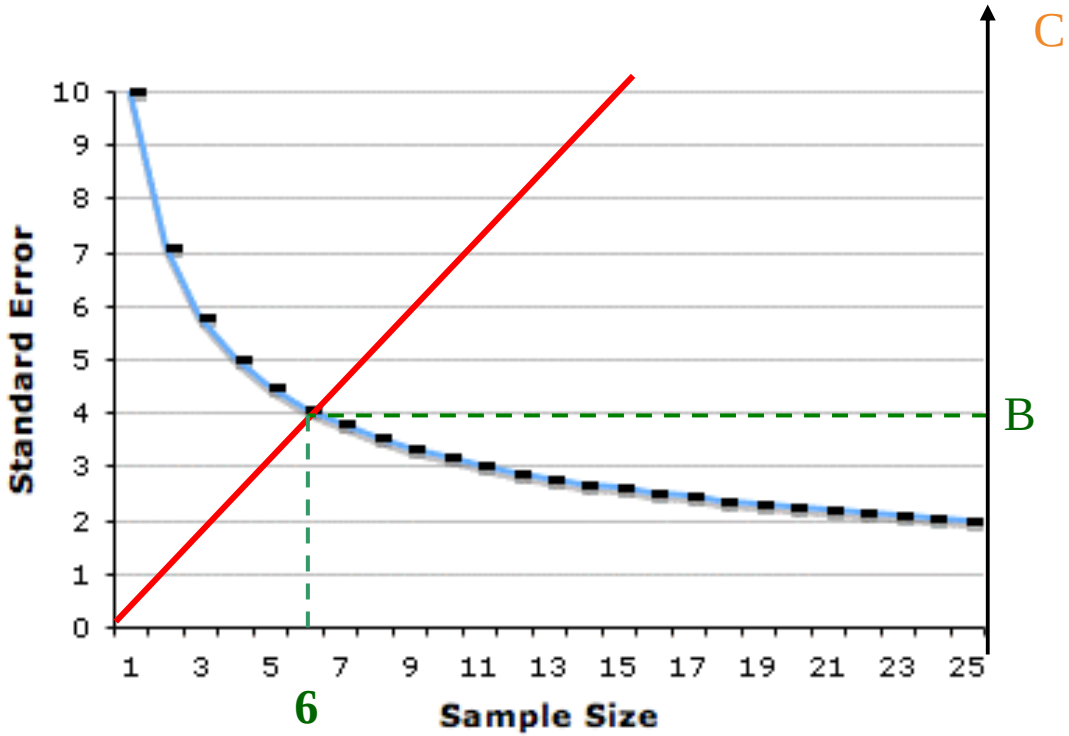
$$\sigma_{moyenne} = \frac{\sigma^{\exp}}{\sqrt{n}}$$

APPLICATION : ESTIMATION DE LA TAILLE DE L'ÉCHANTILLON

Hypothèses :

- Le coût C des observation est linéaire C = a.n
- On dispose d'un budget total de B

Dans l'exemple suivant, à budget fixe, il faut prévoir 6 échantillons (la précision est alors fixée)



ECART-TYPE DE LA MOYENNE - ÉCART TYPE EXPÉRIMENTAL

L’écart type sur la moyenne : **évaluation de la qualité de l’estimation de la valeur** vraie par la moyenne

*Il **dépend** du nombre d'échantillons effectuées*

L’écart type expérimental caractérise la **dispersion intrinsèque des résultats**, il est une conséquences des fluctuations physiques ou autre du phénomène étudié

Cet écart type a lui-même une moyenne et donc une incertitude sur cette moyenne

*Sa valeur vraie est **indépendante** du nombre de mesures effectuées*

EXEMPLE DE GÉNÉRATEURS D'ALÉA

Générateurs artificiels :

- Les tables
  - Suivant le nombre de chiffres servant à repérer les unités
- Les algorithmes
  - Basés sur des lois statistiques

Générateurs physiques :

- La radioactivité
- Les bruits thermiques
- Les bruits électromagnétiques
- Mécanique quantique

La fonction **=ALEA()** :

Elle renvoie un nombre aléatoirement compris entre 0 et 1

Pour des nombres compris entre a et b : **=ALEA()*(b-a)+a**

Attention, ce qui s'apparente à de l'aléa ne l'est pas forcément :

Cas des systèmes chaotiques où derrières l'apparence d'un comportement aléatoire se cache un système déterministe avec un système d'équation non linéaires

Les plans par fragmentation régulière

- Plan systématique (*systematic sampling*)
- Carte de contrôle (*control charts*)

APPLICATION

•Plan très **intuitif** et donc très populaire (impression d'un échantillonnage méthodique et « déterministe »)

•Permet d'avoir une **couverture** uniforme sur un domaine d'étude

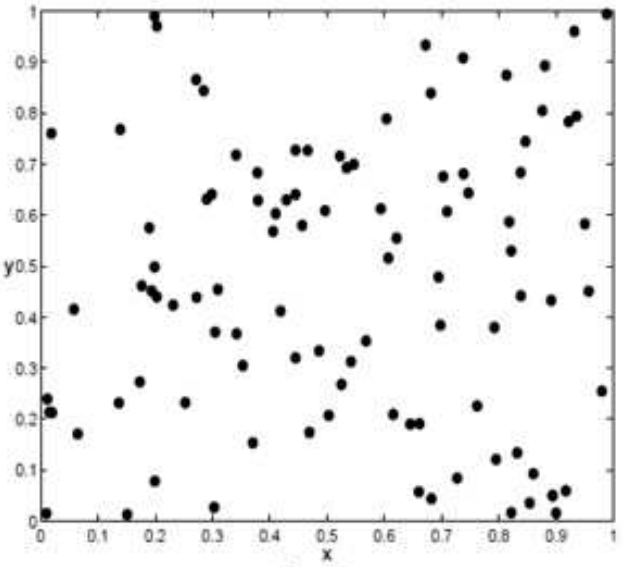
•Facile d'utilisation et souple (adaptation de la taille du maillage à l'hétérogénéité)

Plan adapté pour :

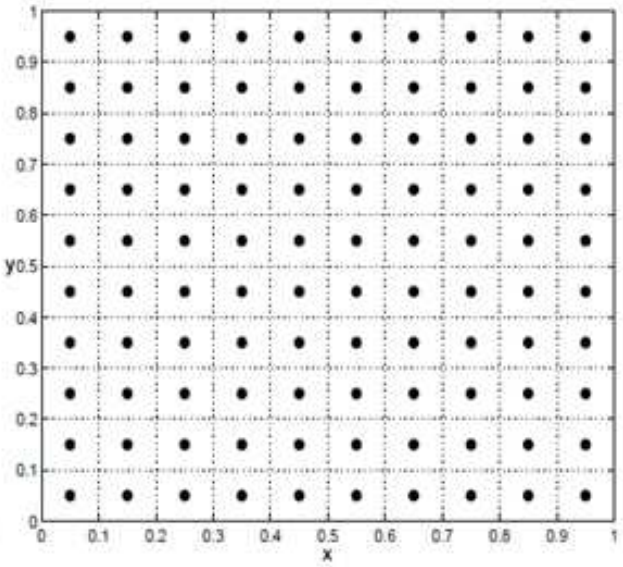
- •La recherche de corrélations
- •Estimer leurs importances
- •Les phases exploratoires (pré-échantillonnage)

Plan souvent comparé avec les plans aléatoires et stratifiés

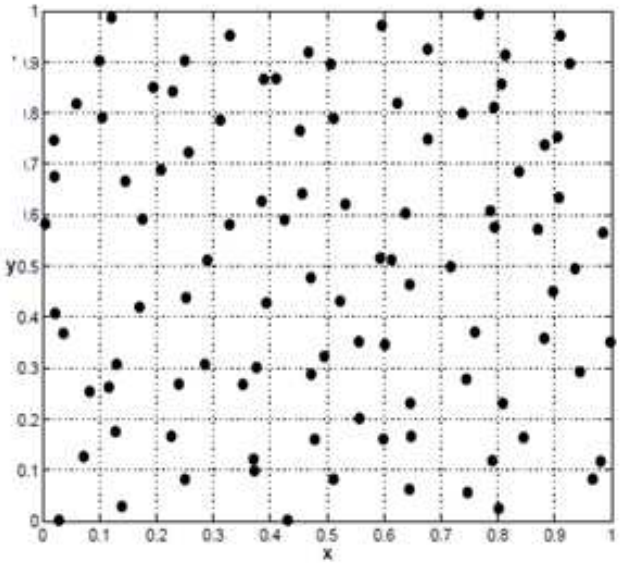
REPRÉSENTATION CARTÉSIENNE



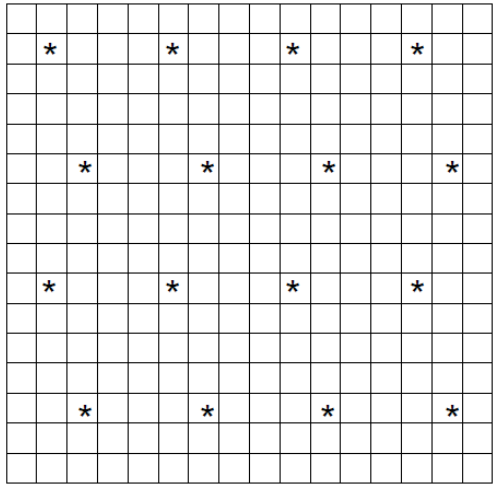
Aléatoire



Systématique avec centrage



Systématique avec de l'aléa



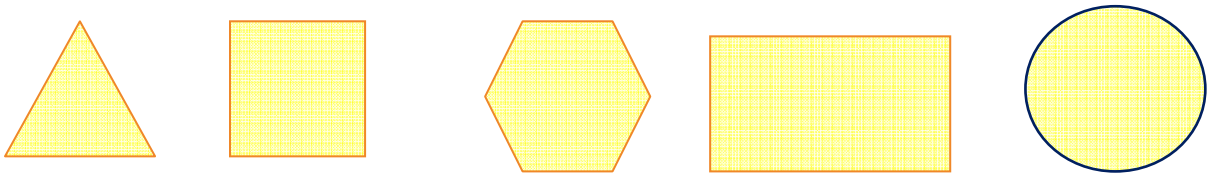
MÉTHODOLOGIE

1. Vérifier l’existence d’un système de coordonnées
2. Les unités sont repérées par ces coordonnées
3. Choisir une périodicité k
4. Découpage du domaine d’étude en k mailles régulières
5. Sélection aléatoire de la première unité
6. Sélection d’une unité dans chacune des mailles (distantes d’un multiple de k)

LA MAILLE

Elle varie suivant :

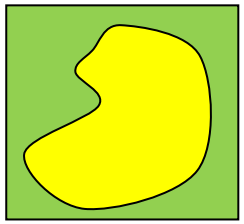
- Leur forme



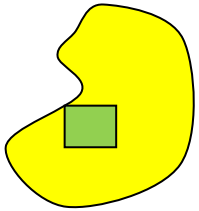
Le triangle équilatéral semble être le plus performant
Les mailles carrées sont les plus utilisées

- Leur orientation
- Leur taille (liée à la périodicité de K)

Trop grand



Trop petit



MISE EN PLACE

Taille :

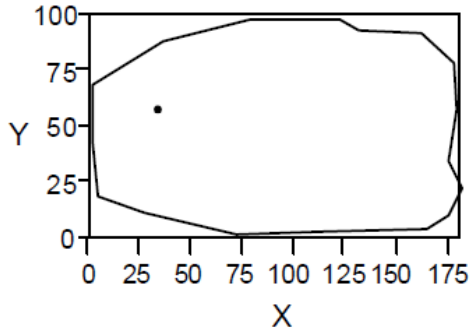
- Maille rectangulaire : $L = \sqrt{\frac{A}{n}}$

  A → Surface du domaine

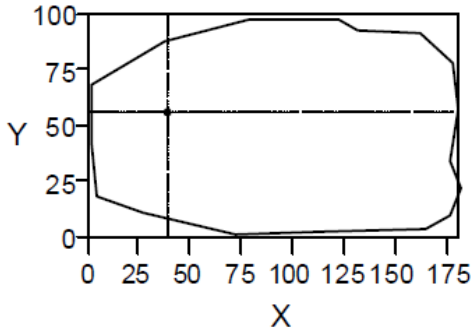
  n → Taille de échantillon

- Maille triangulaire : $L = \sqrt{\frac{A}{Cos(30).n}}$

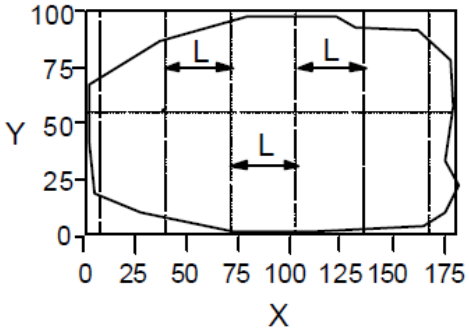
(1) Selection aléatoire du 1er point



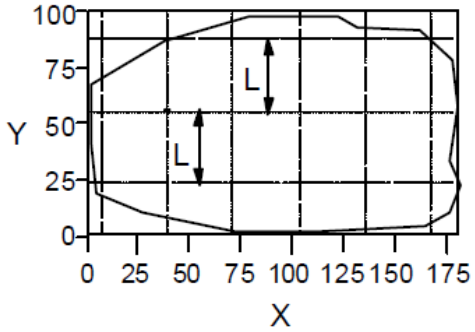
(2) Mettre les axes de coordonnées (passant par ce point)



(3) Mettre les verticales espacées de L



(4) Mettre les horizontales espacées de L



MAILLAGE TRIANGULAIRE, EXEMPLE

Étendue du domaine à échantillonner : (0, 0) à (200, 100) soit
Une surface totale de 140,025

Utilisation d'un générateur d'aléa (nombre compris entre 0 et 1)

Exemple on tire 0,82 et 0,36

Calcul des coordonnées correspondantes(point initial)

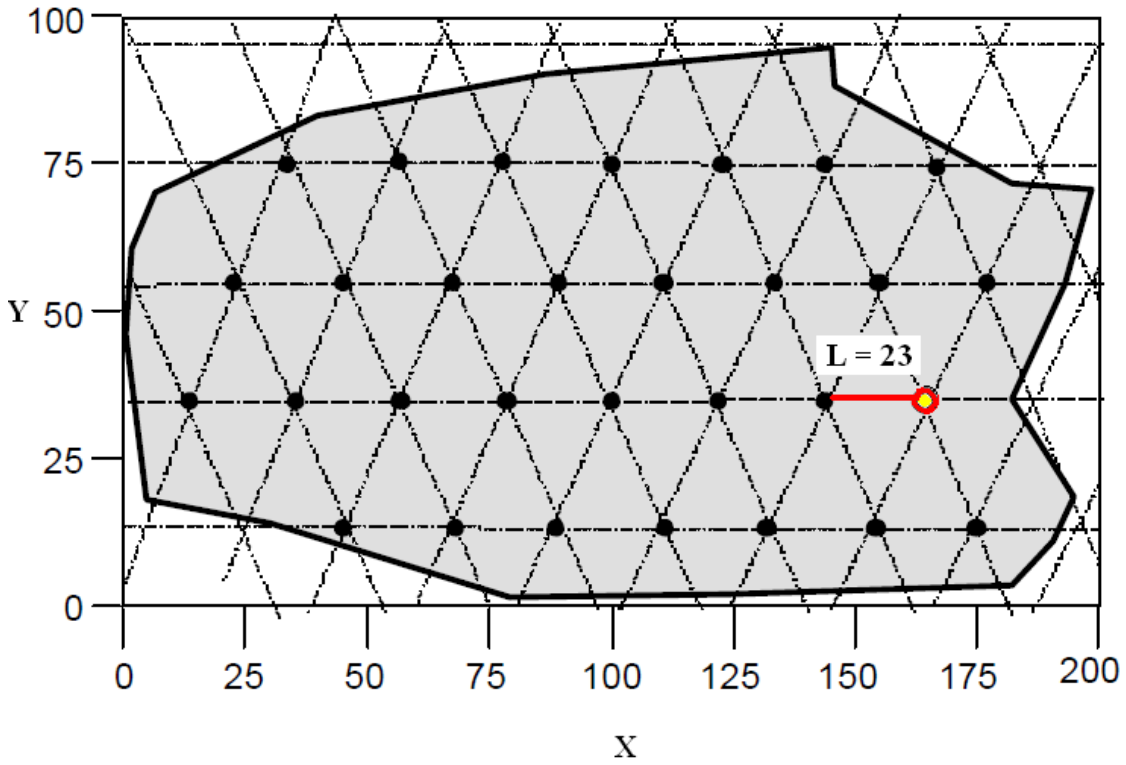
$X = X_{min} + 0{,}82.(X_{max}-X_{min}) = 0 + 0{,}82.(200 - 0) = 164$

$Y = Y_{min} + 0{,}36.(Y_{max}-Y_{min}) = 0 + 0{,}36.(100 - 0) = 36$

Calcul de l'espacement L (choix de triangles équilatéraux et de 30 unités))

$$L = \sqrt{\frac{A}{Cos(30).n}} = \sqrt{\frac{14{,}025}{0{,}866.30}} = 23{,}23 \approx 23$$

MAILLAGE CORRESPONDANT



MESURE DU POIDS D'ORANGE

- Cueillette d'oranges (171 oranges) dans un arbre
- On mesure le poids de chaque orange en gramme

Résultat :

350-341-422-439-476-431-434-420-444-430-296-465-455-444-467-410-253-302-260-510-416-474-525-435-303-246-281-511-476-410-511-434-444-424-356-376-492-459-467-426-509-439-448-400-376-449-515-416-482-398-483-331-431-317-344-363-459-461-446-463-393-352-266-406-276-431-431-538-445-509-278-226-352-468-377-303-413-463-447-404-319-406-476-351-424-415-444-440-398-240-294-430-494-509-294-437-439-361-267-438-287-476-283-426-516-312-287-368-438-410-479-426-441-331-426-492-484-403-513-411-472-328-321-302-477-423-507-475-430-453-337-475-539-432-428-441-456-397-318-458-399-469-451-526-387-543-436-468-470-523-446-443-395-473-346-432-275-476-482-440-406-384-280-245-255-474-481-510-364-401-295

REPÉRAGE DE LA CUEILLETTE

Taille de la population : N = 171

Taille de l'échantillon voulu : n = 20 oranges

Périodicité : k = N/n = 171/20 = 8,55 ≈ 9

Repèrage des ora	Maille 1	Maille2	Maille 3	Maille 4	Maille 5	Maille 6	Maille 7	Maille 8	Maille 9
Unité 1	1	21	41	61	81	101	121	141	161
Unité 2	2	22	42	62	82	102	122	142	162
Unité 3	3	23	43	63	83	103	123	143	163
Unité 4	4	24	44	64	84	104	124	144	164
Unité 5	5	25	45	65	85	105	125	145	165
Unité 6	6	26	46	66	86	106	126	146	166
Unité 7	7	27	47	67	87	107	127	147	167
Unité 8	8	28	48	68	88	108	128	148	168
Unité 9	9	29	49	69	89	109	129	149	169
Unité 10	10	30	50	70	90	110	130	150	170
Unité 11	11	31	51	71	91	111	131	151	171
Unité 12	12	32	52	72	92	112	132	152	
Unité 13	13	33	53	73	93	113	133	153	
Unité 14	14	34	54	74	94	114	134	154	
Unité 15	15	35	55	75	95	115	135	155	
Unité 16	16	36	56	76	96	116	136	156	
Unité 17	17	37	57	77	97	117	137	157	
Unité 18	18	38	58	78	98	118	138	158	
Unité 19	19	39	59	79	99	119	139	159	
Unité 20	20	40	60	80	100	120	140	160	

RÉALISATION D'UN MAILLAGE

Valeurs (poids)	Maille 1	Maille2	Maille 3	Maille 4	Maille 5	Maille 6	Maille 7	Maille 8	Maille 9	Moyenne des échantillons	Ecart type expérimental
Unité 1	350	341	422	439	476	431	434	420	444	417,44	43,98
Unité 2	430	296	465	455	444	467	410	253	302	391,33	83,69
Unité 3	260	510	416	474	525	435	303	246	281	383,33	111,28
Unité 4	511	476	410	511	434	444	424	356	376	438,00	54,43
Unité 5	492	459	467	426	509	439	448	400	376	446,22	42,08
Unité 6	449	515	416	482	398	483	331	431	317	424,67	67,67
Unité 7	344	363	459	461	446	463	393	352	266	394,11	68,73
Unité 8	406	276	431	431	538	445	509	278	226	393,33	109,00
Unité 9	352	468	377	303	413	463	447	404	319	394,00	60,69
Unité 10	406	476	351	424	415	444	440	398	240	399,33	69,06
Unité 11	294	430	494	509	294	437	439	361	267	391,67	90,53
Unité 12	438	287	476	283	426	516	312	287		378,13	95,99
Unité 13	368	438	410	479	426	441	331	426		414,88	46,03
Unité 14	492	484	403	513	411	472	328	321		428,00	74,41
Unité 15	302	477	423	507	475	430	453	337		425,50	71,33
Unité 16	475	539	432	428	441	456	397	318		435,75	63,34
Unité 17	458	399	469	451	526	387	543	436		458,63	54,80
Unité 18	468	470	523	446	443	395	473	346		445,50	53,89
Unité 19	432	275	476	482	440	406	384	280		396,88	80,50
Unité 20	245	255	474	481	510	364	401	295		378,13	105,34

Moyenne des moyennes	**411,74**
σ de la moyenne	**24,8**

PLUS PRÉCIS MAIS AVEC QUELLE PRÉCISION ?

L'échantillon obtenu en sélectionnant l'unité 6 permet l'estimation suivante:

1. Poids moyen : 425 g
2. Variabilité de ce poids : 2.68 = 136 g (32%)

Par contre aucune possibilité d'estimer la précision sur la moyenne !!!

Conséquence : *bien qu'un plan systématique apporte sans doute plus de précision à la valeur moyenne qu'un plan aléatoire, il est impossible de quantifier cette précision*

ÉCHANTILLONNAGE SYSTÉMATIQUE OU STRATIFIÉ ?

Il est impossible d'évaluer la variance de la moyenne avec un seul échantillonnage

Mais en réalisant n échantillonnage (n_{max} ~ N/k), cette variance peut être calculée, en particulier pour un échantillon n (qui correspond à une classe) la variance sur la moyenne s'écrit :

$$\hat{\sigma}^2_{\bar{x}n,\ n\times Syst} = \left(1 - \frac{1}{N}\right)\frac{S^2_{tot}}{n}[1 + (n-1)]$$

Pour un échantillonnage aléatoire dans une des strates d'un échantillonnage stratifié, on aurait eu :

$$\hat{\sigma}^2_{\bar{x}h,aléa} = \left(1 - \frac{n}{N}\right)\frac{S^2_{tot}}{n}$$

Il en résulte que l'échantillonnage systématique est plus précis que l'échantillonnage stratifié lorsque :

$$\hat{\sigma}^2_{\bar{x}n,\ n\times Syst} < \hat{\sigma}^2_{\bar{x},aléa} \quad \text{soit} : \rho < \frac{-1}{n.k-1}$$

Si $\rho < 0$ l'échantillonnage systématique est avantageux
Si $\rho > 0$ l'échantillonnage systématique est DESavantageux

Si on considère que les deux plans apportent une même précision, le coefficient ρ est de la forme :

$$\rho = \frac{1}{1 - n.k} = \frac{1}{1 - N} \text{ qui est sensiblement egalà } 0$$

L'efficacité du plan étant : $D_{eff} = \frac{N-1}{n.(k-1)} [\ 1 + (n-1)\rho\ ]$

SYTEMATIQUE AVEC LE TEMPS : CARTES DE CONTRÔLE

Voir plus loin dans la présentation

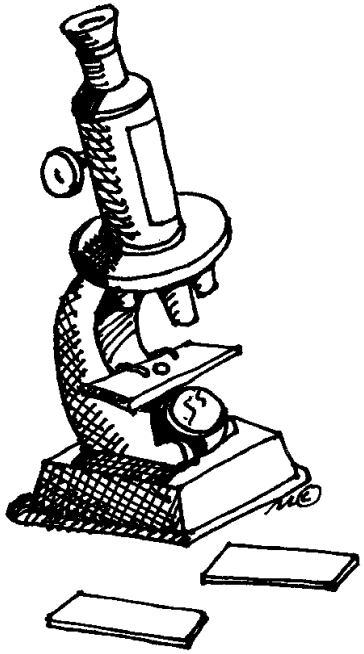
Les plans par regroupement 1

- Classification
- Plan multiniveaux (ou à degrés)
  (multistage sampling)
- Echantrionnage (*Rank Set Sampling*)

ANALOGIE AVEC LE MICROSCOPE

Il s'agit d'adapter la sélection progressivement comme on le ferait en choisissant le grossissement d'un microscope ou d'un télescope

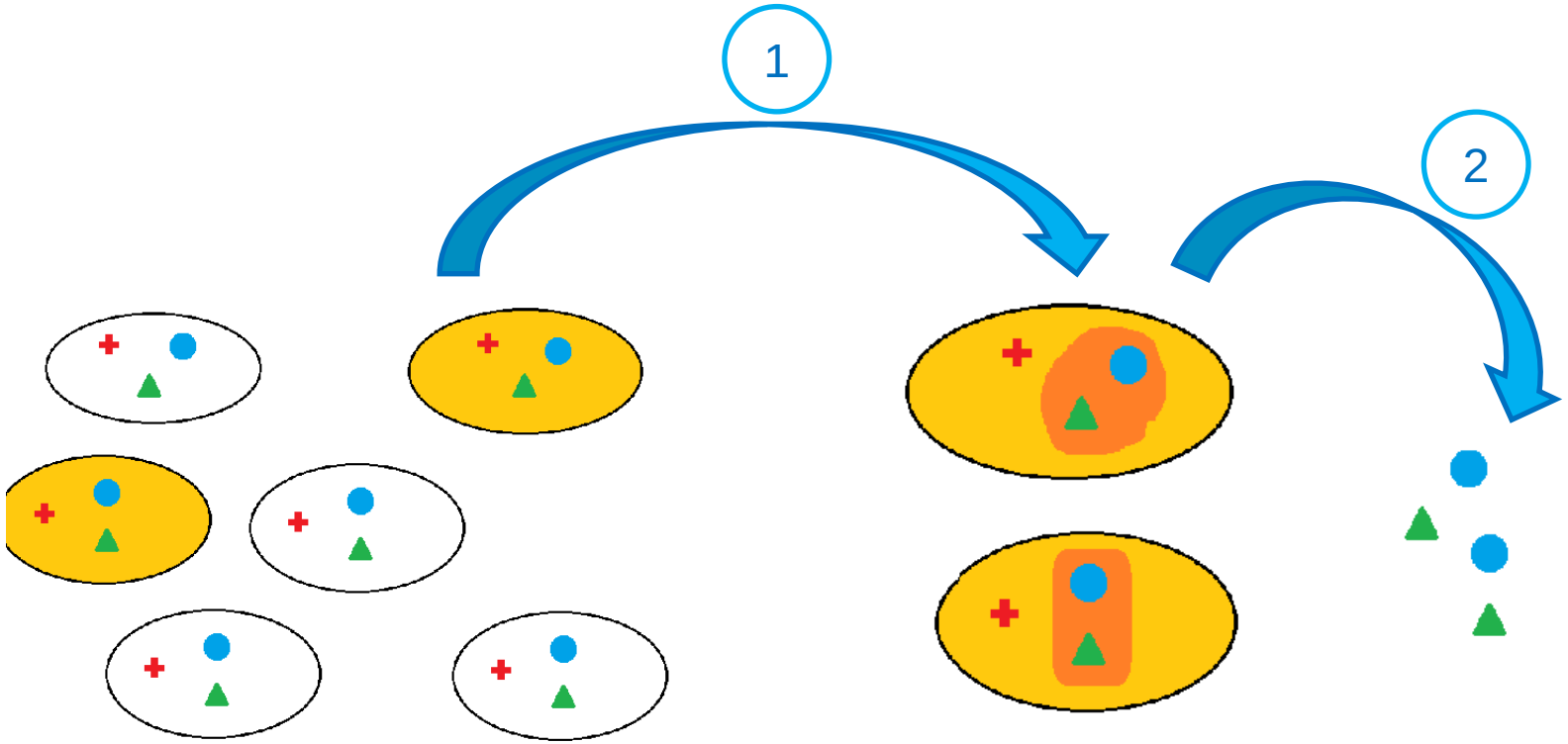
Changer de niveaux c'est un peu changer de monde



Dans un plan à 2 niveaux ou 2 degrés on a :

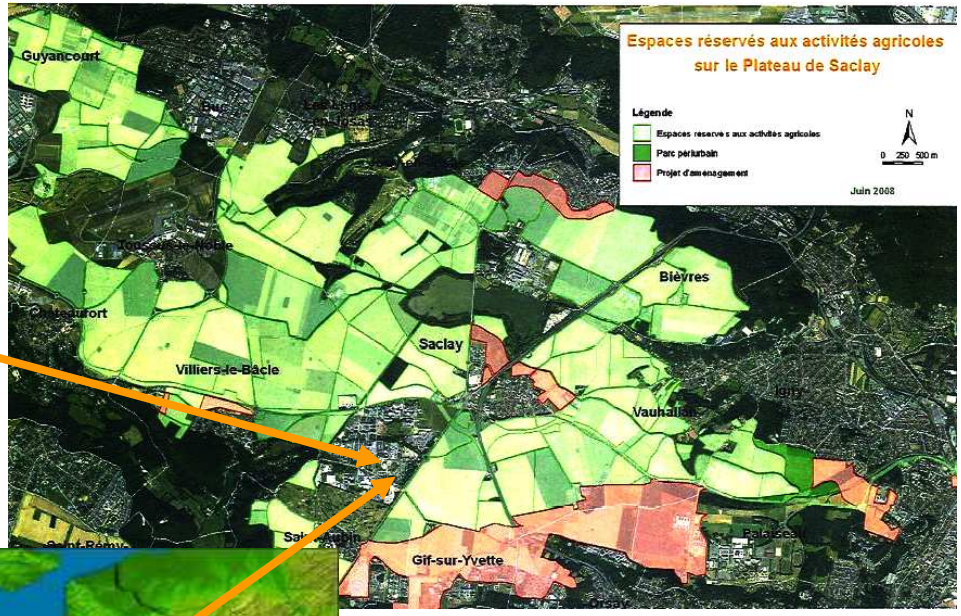
- **Un niveau macroscopique**
- **Un niveau microscopique**

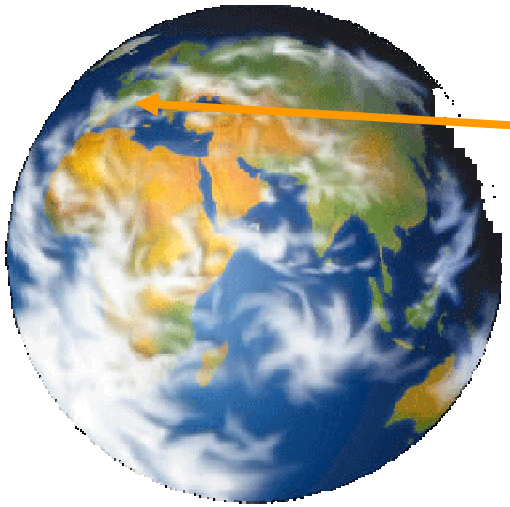
PLAN 2 NIVEAUX OU 2 DEGRÉS



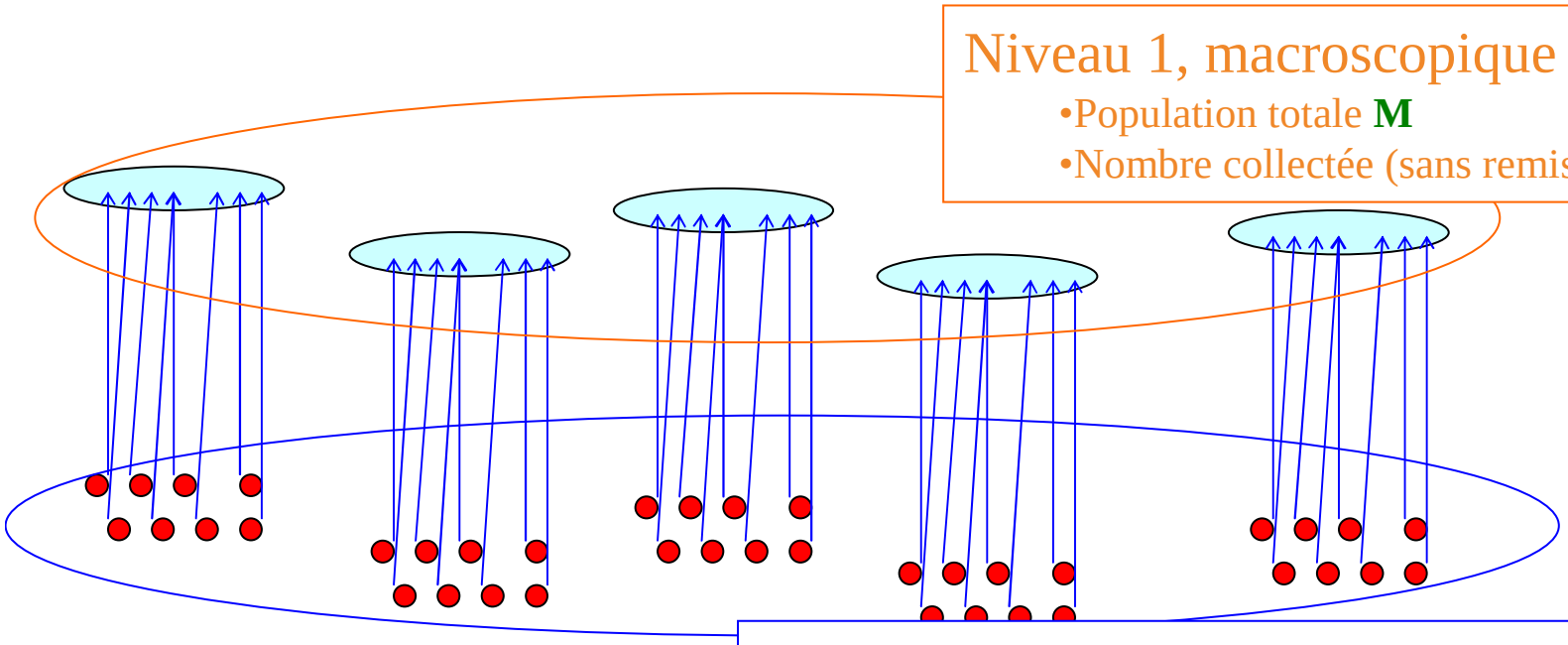
PRINCIPE

Multiniveaux





ANALYSE AVEC 2 NIVEAUX



Niveau 1, macroscopique : les classes

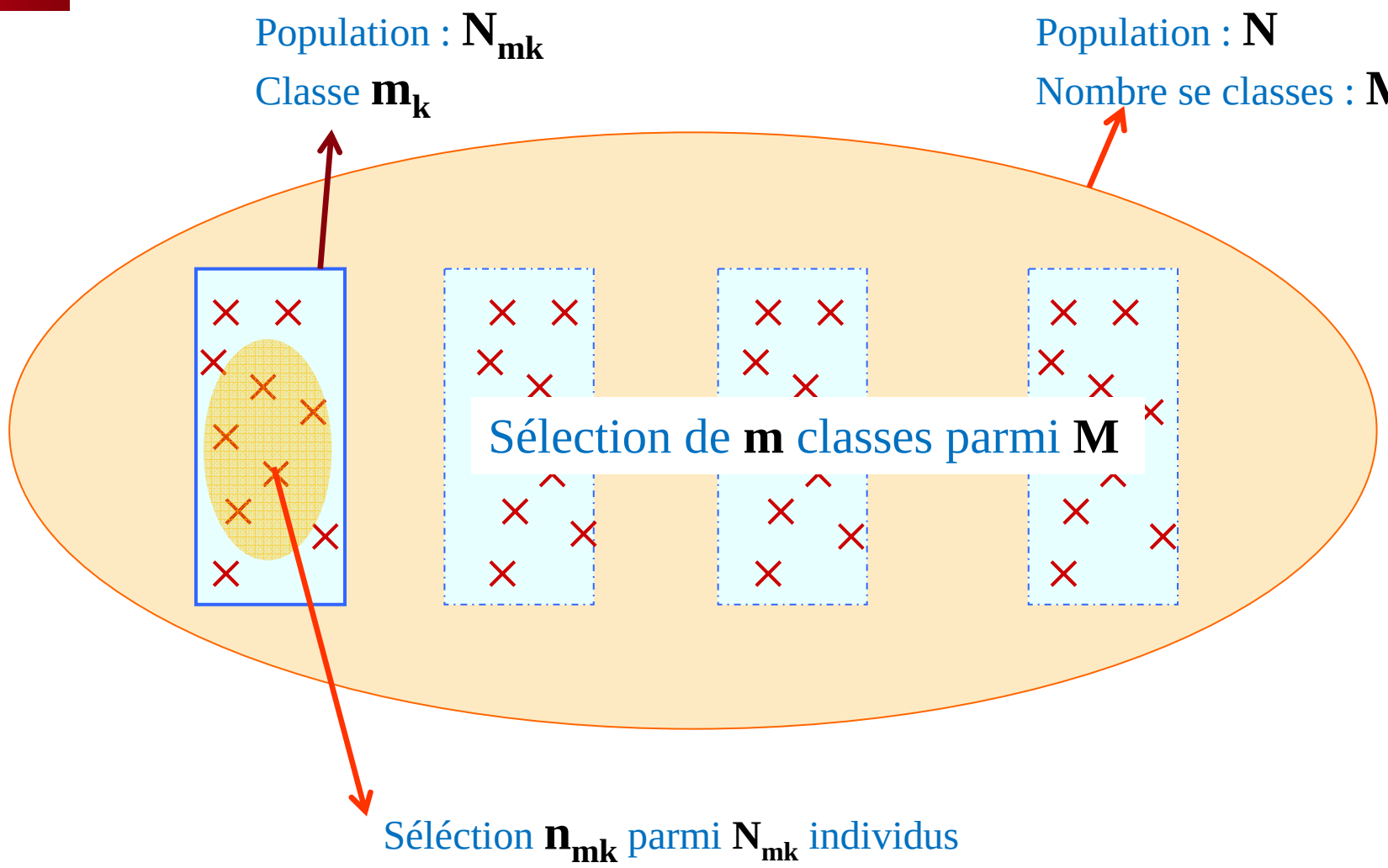
- Population totale **M**
- Nombre collectée (sans remise) : **m**

Niveau 2, microscopique : les individus

- Population totale : **N**, ($\mathbf{N_{mk}}$ pour la classe mk)
- Nombre collectée (sans remise) pour la classe mk : $\mathbf{n_{mk}}$

Deux niveaux $\Rightarrow$ nécessairement deux indices : i et mk

POUR RÉSUMER



ABRÉVIATIONS ÉTUDIÉES

Nombre total d'individus : N
Nombre total de classes : M
Nombre de classe sélectionnées : m
Identification d'une des classes sélectionnée, indice : mk
Population de la classe mk : N_{mk}
Taille de l'échantillon dans la classe mk : n_{mk}
Proportion d'individus sélectionnées dans la classe mk : f_{mk}
Estimation de la variance expérimentale d'une classe mk : $s^2_{dans} = s^2_{mk}$
Estimation de la moyenne d'une classe mk : $\bar{x}_{mk}$
Estimation de la variance de la moyenne d'une classe mk : $s^2_{\bar{x},mk}$
Estimation de la variance expérimentale à partir des classes sélectionnées : s^2_M
Estimation de la moyenne des classes sélectionnées : $\bar{\bar{x}}_m$
Estimation de la variance de la moyenne des classes sélectionnées : $s^2_{\bar{M}}$
Estimation de la moyenne des classes sélectionnées pondérée par le nombre : $\bar{x}_m$
Estimation de la variance à partir des moyennes pondérées : $s^2_{entre} = s^2_m$
Estimation de la moyenne totale : $\bar{\bar{x}}$
Estimation de la variance sur la moyenne totale : $s^2_{\bar{\bar{x}}}$

DESCRIPTION INTERNE D'UNE CLASSE (MICROSCOPIQUE)

Multiniveaux

Moyenne d'une classe m_k: $\bar{x}_{mk} = \frac{1}{n_{mk}} . \sum_{i=1}^{n_{mk}} x_{i,mk}$

Écart type **expérimental** de la classe m_k: $s^2_{mk} = \frac{SSW}{n_{mk} - 1}$

Écart type sur la **moyenne** de la classe m_k : $s^2_{\bar{x},mk} = (1 - f_{mk}) \frac{s^2_{mk}}{n_{mk}}$

Moyenne des écarts types obtenus pour chaque classe (permet d'évaluer l'importance de l'hétérogénéité qu'il y a dans chacun d'eaux) dans le cas où les classes ont toute la même taille n

$$s^2_{dans} = \frac{SSW}{m.(\bar{n} - 1)}$$

Avec $SSW = \sum_{i=1}^{n} (x_{i,mk} - \bar{x}_{mk})^2$

DESCRIPTION EXTERNE, ENTRE LES CLASSES (MACROSCOPIQUE)

Multiniveaux

Moyenne **non pondérée**

Avec : $SSB = \sum_{mk=1}^{m} \left(\overline{x}_{mk} - \overline{\overline{x}}_{m}\right)^2$

Moyenne : $\overline{\overline{x}}_m = \frac{1}{m} \sum_{k=1}^{m} \overline{x}_{mk}$

Écart type **expérimental** : $s^2_{entre} = s^2_M = \frac{SSB}{m-1}$

Écart type sur la **moyenne** : $s^2_{\overline{M}} = \frac{(1 - F_m)}{m.(m-1)} SSB$

*Paramètres sur la population "class" (moyenne **pondérée** par le nombre)*

Moyenne : $\overline{x}_m = \frac{1}{m} \sum_{k=1}^{m} N_{mk} . \overline{x}_{mk}$

Écart type expérimental : $s^2_{entre} = s^2_m = \frac{1}{m-1} . \sum_{mk=1}^{m} \left[\left(N_{mk} . \overline{x}_{mk}\right) - \overline{x}_m\right]^2$

SYNTHÈSE DES DEUX ANALYSES

ECRAT TYPE expérimental

VARIANCE $_{\text{expérimentale}}$ **TOTALE =**
MOYENNE DES VARIANCES (expérimentales des classes) **+ VARIANCE** $_{\text{expérimentales}}$ **DES MOYENNES** (des classes)

- La moyenne des variances c'est : $s^2_{dans} = \frac{SSW}{m.(\overline{n}_m - 1)}$
- La variances des moyennes c'est : $s^2_{entre} = s^2_M = \frac{SSB}{m-1}$

$$s^2_{tot} = \frac{SSB}{m-1} + \frac{SSW}{m.(\overline{n}_m - 1)}$$

ECRAT TYPE sur la moyenne totale

$$s^2_{\overline{\overline{x}}} = \left[\left(\frac{M}{N}\right)^2.(1-F_m).\frac{s^2_{entre}}{m}\right] + \left[\frac{1}{N^2}\frac{M}{m}\sum_{mk=1}^{m} N^2_{mk}.(1-f_{mk}).\frac{s^2_{m_k}}{n_{mk}}\right]$$

SÉLECTION

Multiniveaux

Simplification : classes et unités sélectionnées de même taille, respectivement m et n

Nombre total d'unités sélectionnées

- Pour la taille des classes $N_{mk} = \overline{N} = \frac{N}{M}$

- Pour la taille de la sélection dans les classes $n_{mk} = \overline{n} = \frac{n}{m}$

$$s^2_{\overline{\overline{x}}} = \left[\frac{1}{\overline{N}^2}.(1-F_m).\frac{s^2_{entre}}{m}\right] + \left[\frac{1}{M.\overline{n}}.(1-\overline{f}).s^2_{dans}\right]$$

Par ailleurs, pour donner une forme plus simple à cette équation, on pose :

$$S^2_{entre} = \frac{s^2_{entre}}{\overline{N}^2} \quad \text{et} \quad S^2_{dans} = \frac{s^2_{dans}}{M}$$

$$s^2_{\overline{\overline{x}}} = \left[(1-F_m).\frac{S^2_{entre}}{m}\right] + \left[(1-\overline{f}).\frac{S^2_{dans}}{n}\right]$$

COMPARAISON AVEC UN PLAN PUREMENT ALÉATOIRE

Multiniveaux

Taille de l'échantillon : $n = \bar{n}.m$

Conséquences : $s^2_{\bar{\bar{x}},alea} = \left(1 - \frac{m.\bar{n}}{N}\right).\frac{s^2_{tot}}{m.\bar{n}} = \left(1 - \frac{m.\bar{n}}{M.\bar{N}}\right).\frac{s^2_{tot}}{m.\bar{n}}$

Dans le cas où la population est importante : $s^2_{\bar{\bar{x}},alea} = \frac{s^2_{tot}}{m.\bar{n}}$

En considérant l'équation : $\rho = \frac{S^2_{entre}}{S^2_{total}} = \frac{S^2_{entre}}{S^2_{entre} + S^2_{dans}}$

Ce qui permet d'évaluer l'efficacité du plan (sous réserve des hypothèses approximations faites) $D_{eff} = \frac{s^2_{\bar{\bar{x}}}}{s^2_{aléa}} = 1 + (n-1).\rho$

GÉNÉRALISATION À 3 NIVEAUX

Estimation de la moyenne

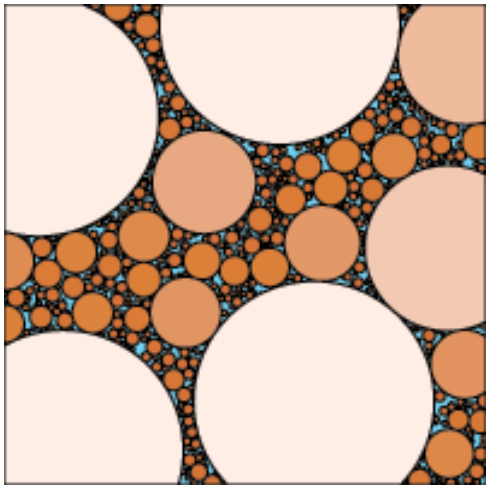
$$x = \frac{1}{n_i.n_j.n_k} \sum_i \sum_j \sum_k x_{ijk}$$

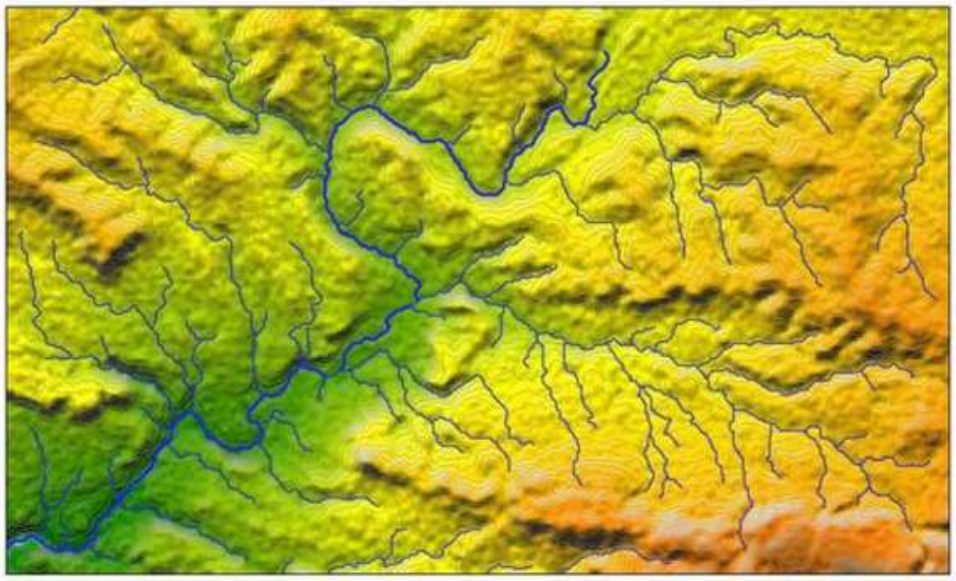
Estimation de la variance

$$\sigma_{\bar{x}} = \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_1.n_2} + \frac{\sigma_3}{n_1.n_2.n_3}$$

Optimisation des couts

$$C_t = C_0 + n_1 C_1 + n_1.n_2 C_2 + n_1 n_2 n_3 C_3$$





LOI EN PUISSANCE DE TAYLOR

$$Ln(s^2) = b.Ln(\mu) + Ln(a)$$

Index d'agrégation
1/b évalue la dispersion des individus

Effectif moyen des individus possédant la variance s^2

Paramètre d'échantillonnage

Statistiquement on prit montrer (Fairfield & Smith 1938) que si :

- S^2 est la fluctuation du nombre d'une espèce sur une surface fixée
- μ est l'effectif moyen d'une espèce pour cette même surface

$$s^2 \propto \mu^{b \approx 0,62}$$

VARIANCE SUR PLUSIEURS NIVEAUX

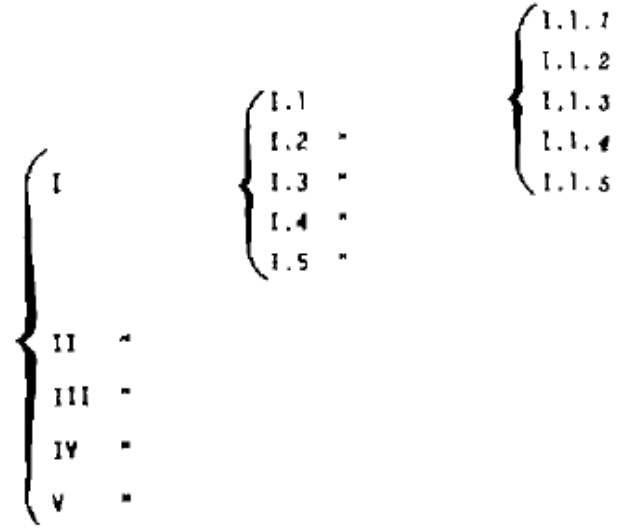
Approche fractale

Plan d'échantillonnage pour évaluer la pollution fécale dans les étangs de Thau

Echelle d'observation croissante

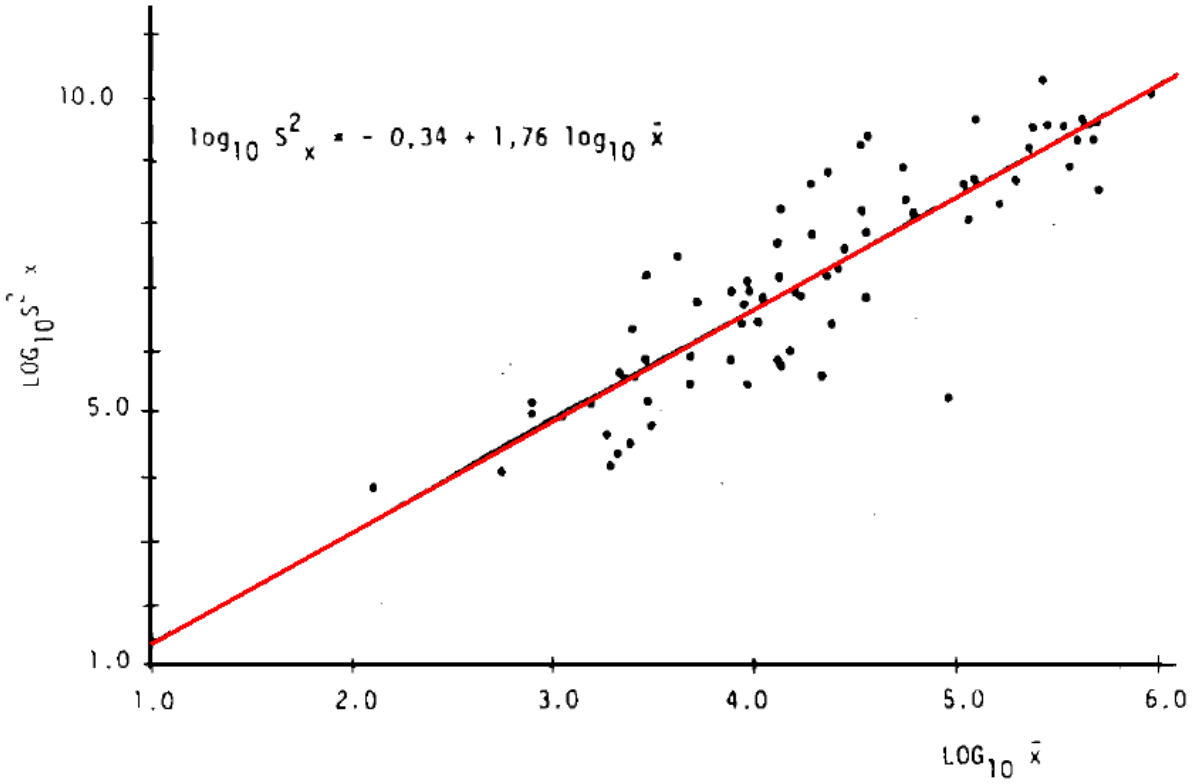
NIVEAU D'OBSERVATION	BAIE	STATIONS	PRELEVEMENTS
SOURCE DE VARIABILITE	*INTER-STATIONS*	*INTER-PRELEVEMENTS*	*INTER-MESURES*
NOMBRE DE REPETITIONS	5 STATIONS	5 PRELEVEMENTS PAR STATION	2 à 5 MESURES PAR PRELEVEMENT

STRUCTURE DU SCHEMA D'ECHANTILLONNAGE



Plan d'échantillonnage pour évaluer la pollution fécale dans les étangs de Thau

75 prélèvements d'eau,
5 mesures par prélèvements



Loi de Taylor : $\log(s^2_{exp}) = 1,76.\log(x_{moy}) - 0,34$

X_{moy} : dénombrement des bactéries hétérotrophes revivifiables

Résultats de l'analyse de la variance hiérarchisée sur les données de 3 campagnes de pré-échantillonnage

	SOURCE DE VARIATION	DDL	SCE	CM	F	PARAMETRES ESTIMES	COMPOSANTES DE LA VARIANCE	POURCENTAGES
1ère CAMPAGNE	STATION (S)	4	95.09	23.77	40.01**	$25\,s^2_s + 5\,s^2_p + s^2_a$	$s^2_s = 0.93$	87.8
	PRELEVEMENT (P)	20	11.88	0.59	47.25**	$5\,s^2_p + s^2_m$	$s^2_p = 0.11$	11.0
	MESURE (M)	100	1.25	0.01		s^2_m	$s^2_m = 0.01$	1.2
2ème CAMPAGNE	STATION (S)	4	53.87	13.47	26.46**	"	$s^2_s = 0.52$	80.0
	PRELEVEMENT (P)	20	10.58	0.53	17.05**	"	$s^2_p = 0.10$	15.4
	MESURE (M)	100	3.10	0.03		"	$s^2_m = 0.03$	4.6
3ème CAMPAGNE	STATION (S)	4	21.75	5.43	25.17**	"	$s^2_s = 0.21$	72.0
	PRELEVEMENT (P)	20	4.32	0.21	4.51**	"	$s^2_p = 0.03$	11.5
	MESURE (M)	100	4.79	0.05		"	$s^2_m = 0.05$	16.5

Avec :

$$s^2_{\bar{x}} = \frac{s^2_s}{n_s} + \frac{s^2_p}{n_s.n_p} + \frac{s^2_m}{n_s.n_p.n_m}$$

ALLOCATION

Adapter le nombre d'unité à prélever en fonction des informations apportées par un plan d'échantillonnage

Allocation **équilibrée** : échantillon de **même taille** pour toutes les classes

Allocation **proportionnelle** : échantillon de taille proportionnelle à la **taille de la classe**

Allocation de **Newman** : échantillon dont la taille est adapté à l'**hétérogénéité de la classe**

INTRODUCTION

Techniques développée par McIntyre en 1952 pour l'agriculture

Objectif : injecter de l'information en cours de sélection des unités pour diminuer la variance et gagner en précision sur l'estimation d'une moyenne

C'est un échantillonnage en 2 niveau, il faut donc choisir un premier plan

Technique intéressante quand le prix de la mesure en laboratoire est élevée

PRINCIPE

Principe :

- Estimer que le domaine étudié est non-homogène
- Disposer d'une variable pour pouvoir se repérer les unités
- Utiliser un premier plan (aléatoire) pour sélectionner les lieux de "prélèvement"
- Injecter de l'information à chaque prélèvement en rangeant suivant un paramètre les unités.
- Recommencer pour avoir une bonne statistique

$\Rightarrow$ On obtient une information sur la structure

$\Rightarrow$ Une distribution suivant les tailles

MÉTHODOLOGIE

1. Sélection aléatoire de m^2 unités du domaine d'étude
2. Regroupement aléatoirement de ces m^2 unités en m ensembles de m unités
3. Faire du rangement dans chacun des m groupes : à chacune des m unités est attribué une place i
4. Sélectionner des unités rangées suivant la règle :
   dans le groupe i, on ne sélectionne que l'unité située à la place i
   (les autres sont laissées)

A la fin de cette opération, on dispose de m unités rangés suivant un place i

Pour avoir une statistique, il est nécessaire d'avoir plusieurs unités pour une place donnée i

$\Rightarrow$ Nécessité de recommencer la procédure r fois

Ce qui donne un nombre total d'unités de $n = m.r$

ESTIMATION DE LA MOYENNE ET DE SA VARIANCE

On distinguera le plan équilibré (même nombre d'individus sélectionnés dans chaque lieu) et non-équilibré

Dans le cas d'un tel plan (ce qui est souvent le cas) la moyenne et sa variance sont estimées par les expressions suivantes

$$\bar{x}_{Rank} = \frac{1}{r.m} \sum_{i}^{m} \sum_{j}^{r} x_{i,j} \qquad Var(\bar{x}_{Rank}) = \frac{1}{r(r-1)} \sum_{j}^{r} \frac{1}{m^2} \sum_{i}^{m} (x_{i,j} - \bar{x}_i)^2$$

$$\bar{x}_i = \frac{1}{r} \sum_{j}^{r} x_{ij}$$

EXEMPLES

La taille des marres d'une région

La reproduction des saumons est liée à la taille des marres

Sélection aléatoire sur la taille des segments de rivières

Les pâturages

La surface d'un terrain contaminé en plomb avec utilisation d'un XRF sur le terrain

Deux paramètres sont nécessaires pour dimensionner son plan :

Le rapport C et la précision relative RP :

$$C = \frac{\text{Cout de la mesure au laboratoire}}{\text{Cout du rangement des unités sur place}}$$

$$RP = \frac{\text{Variance}_{\text{aléatoire}}}{\text{Variance}_{\text{RSS}}}$$

On peut montrer que :

$$1 < RP < \frac{m+1}{2}$$

RSS = Plan aléatoire

Gain obtenu avec un RSS

PARAMÈTRES INITIAUX À DÉTERMINER

Pour bien dimensionner un plan RSS, il faut :

1. Avoir une idée de la loi statistique de la distribution de la caractéristique étudiée

2. Le nombre d'échantillons $\mathbf{n_0}$ voulus (en général dimensionné initialement sur la base d'un plan aléatoire)

3. Une estimation du coefficient de variation **CV** (essais pilote avec 10, mesure, bibliographie, expertise, ….)

4. Se fixer un nombre **m** de classes (évalué au jugement en prenant en compte les contraintes du « terrain »

(pour un plan équilibré)

DÉTERMINATION DE RP OU DE M À PARTIR DU CV

Valeur du RP pour une loi Lognormale (proche de la loi normale pour des CV petits)

Taille m	CV = 0.1	CV = 0.3	CV = 0.5	CV = 0.8
2	1.5	1.4	1.4	1.3
3	1.9	1.8	1.7	1.5
4	2.3	2.2	2.0	1.8
5	2.7	2.6	2.3	2.0
6	3.1	2.9	2.6	2.2
7	3.6	3.3	2.8	2.4
8	3.9	3.6	3.1	2.5
9	4.3	3.9	3.3	2.7
10	4.7	4.3	3.6	2.9

A partir des informations précédente, on détermine le nombre de cycle r :

$$r = \frac{n_0}{m} . \frac{1}{RP}$$

Ce qui permet d'ajuster le nombre d'échantillons : $n = r.m$

EXEMPLE

•A partir de l'hypothèse d'un échantillon aléatoire on détermine qu'il faut 25 échantillons de sol
•Un étude analogue a montré que la CV ≈ 0,4
•On décide de se fixer m = 5

Dans l'hypothèse d'une distribution normale, on détermine à partir de la table que RP ≈ 2,45

On en déduit : $r = \frac{n_0}{m} . \frac{1}{RP} = \frac{25}{5} . \frac{1}{2,45} = 2,04$ arrondi à 3

La taille réelle de l'échantillon est donc : n=3.5 = 15
(alors que pour le même résultat, il aurait fallu 25 échantillon avec un plan aléatoire)

CAS DES PLANS NON-ÉQUILIBRÉS

Utilisation du « t-model » de Kaur (1995)

1. Calcul de la taille de l'échantillon sur la base d'un plan aléatoire, soit n_0 = 64
2. Le nombre de classe est fixée à m=5
3. Par une étude pilote, on a trouvé que CV ≈ 1
4. La table suivante permet de déterminer le nombre t :

CV	0.25	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
t	1	2	3	5	6	7	8	9	10

Le nombre t correspondant à la taille de la plus grande des m classes

5. Détermination du nombre de cycle r pour les m-1=4 plus petites classes restantes.
   Pour CV =1, la table donne une valeur de RP = 1,84

   Cette valeur doit être corrigé à cause du déséquilibre dans la taille des classes

CV	0.1	0.3	0.5	0.8	1.3
Facteur de correction	1.01	1.08	1.2	1.5	1.7

   Soit pour un CV = 1, le facteur est environ de 1,58

   La valeur du RP à prendre est donc : RP = 1,84.1,58 = 2,91
   Valeur qui est meilleur que 1,84 !!

Finalement la nombre de cycle est le suivante :

$$r = \frac{n_0}{m}.\frac{1}{RP} = \frac{64}{5}.\frac{1}{2,95} = 4,4 \text{ arrondi à } 5$$

6. La taille réelle de l'échantillon est donc : n = (t+(m-1)).r = (5+3-1)=35

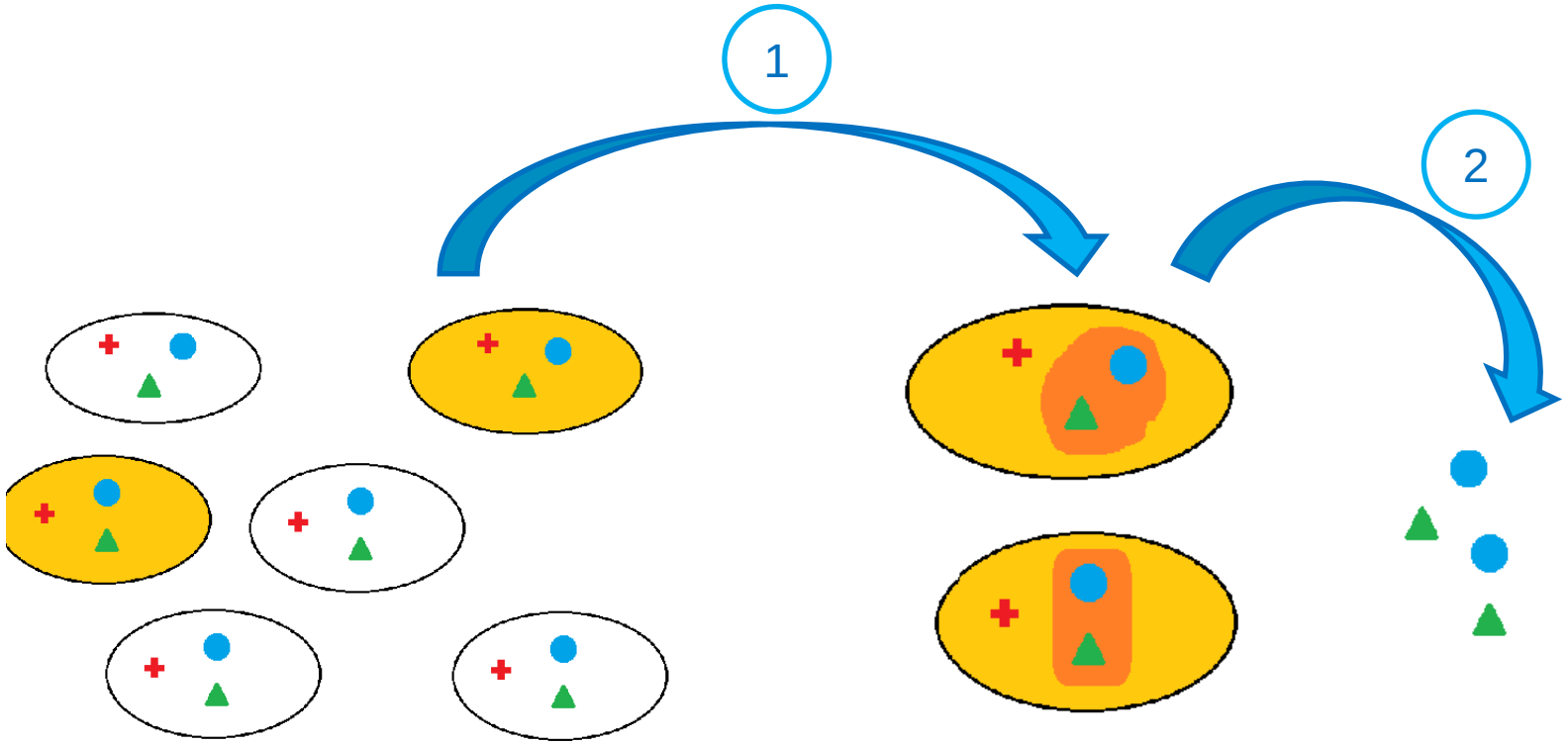
On a donc m+t-1 = 7 lieux à sélectionner aléatoirement et à ranger en m =5 classes

Les plans par regroupement 2

- Plan stratifié (*stratified sampling*)
- Echantillonnage en grappes (cluster sampling – hot spot)

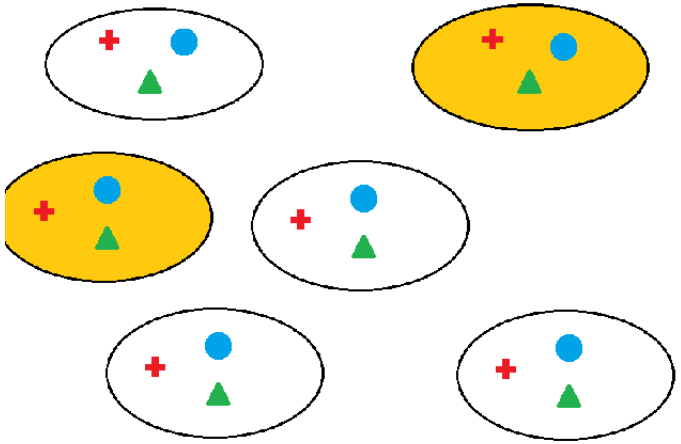
Yann.kergadallan@cea.fr

RAPPELS, PLAN 2 NIVEAUX OU 2 DEGRÉS

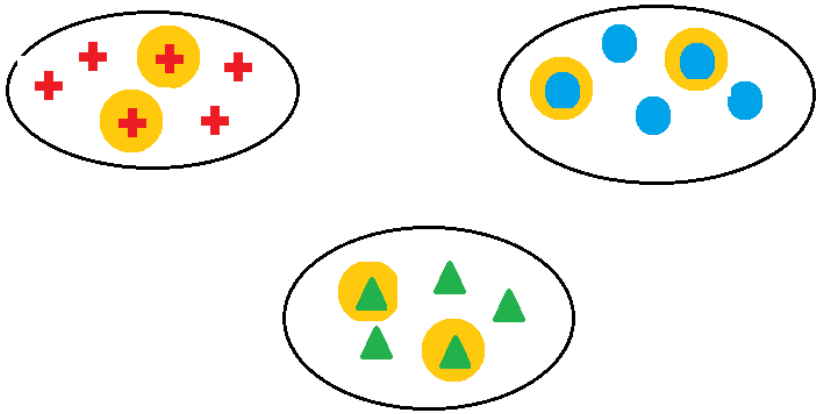


Yann.kergadallan@cea.fr

STRATES ET GRAPPES



- Sélection de 2 unités au niveau **macro** (grappes)
- Grappes hétérogènes (pas de regroupements)
- Exemples : villes, écosystèmes, …



- Sélection de 2 unités au niveau **micro** (strates)
- Strates homogènes (unités semblables rassemblées en sous-groupes)
- Exemples : groupe d'âge, groupe de fruits, groupe de …

Yann.kergadallan@cea.fr

NOMBRE DE CLASSES

Pour les grappes, on sélectionne au niveau macroscopique, mais on observe au niveau des unités

Pour les strates, on sélectionne et observe au niveau microscopique dans différents groupes macroscopiques

Règle de thumb (surtout pour des classes hétérogènes ou grappes), si :

- M : nombre de classes
- N : population totale

$$M \approx \sqrt{\frac{N}{2}}$$

Problème : il faut connaître la population totale N

IDENTIFICATION DU NOMBRE DE CLASSES

Règle de thumb (surtout pour des classes hétérogènes ou grappes), si :

- M : nombre de classes
- N : population totale

$$M \approx \sqrt{\frac{N}{2}}$$

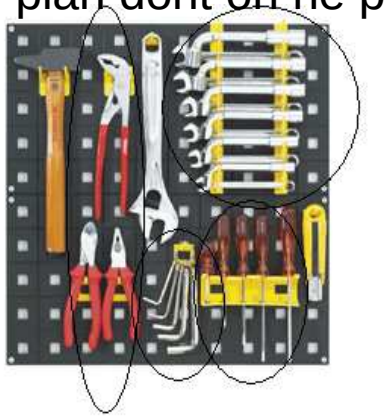
Problème : il faut connaître la population totale N

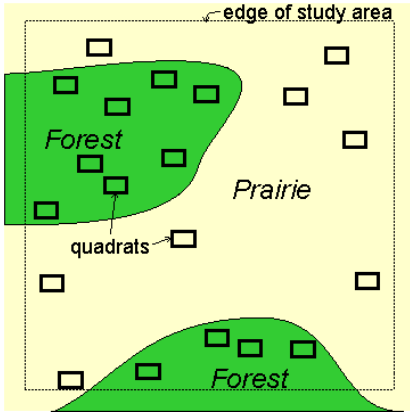
PRINCIPE

Idée : regrouper les unités suivant leurs ressemblance en **sous-ensembles homogènes** appelés strates

Un échantillonnage dans ces strates s'apparente à la réalisation d'un **plan aléatoire**

Rappel : plan aléatoire = plan dont on ne peut plus extraire d'information





Règle de Thumb pour l'estimation de la taille d'une population :

- ***Choix de 6 strates***
- ***Pour chaque strate un échantillon de 5 à 10 unités***

Yann.kergadallan@cea.fr

IDENTIFICATION DES STRATIFICATEURS AVEC UNE ACP

Plan stratifié

Attention : l'ACP n'est pas une méthode de Classification, elle permet Seulement d'identifier les paramètres pertinents pour une analyse

Exemple de classificateurs :

- Précipitation
- Température

Eigen values

Matrix trace = 4,00

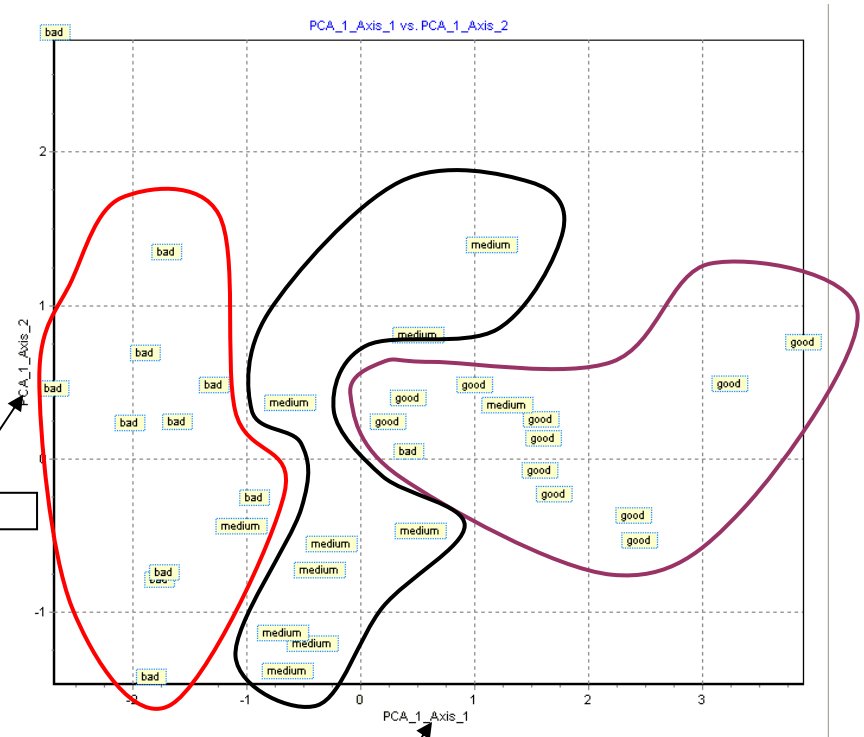
Axis	Eigen value	% explained	Histogram	% cumulated
1	2,791467	69,79%		69,79%
2	0,714470	17,86%		87,65%
3	0,365928	9,15%		96,80%
4	0,128136	3,20%		100,00%
Tot.	4,000000	-	-	-

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2		Axis_3		Axis_4	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
Temperature (°C)	0,9197	85 % (85 %)	0,2580	7 % (91 %)	-0,1255	2 % (93 %)	-0,2681	7 % (100 %)
Sun (h)	0,8577	74 % (74 %)	0,0072	0 % (74 %)	0,5115	26 % (100 %)	0,0524	0 % (100 %)
Heat (days)	0,8964	80 % (80 %)	0,2683	7 % (88 %)	-0,2666	7 % (95 %)	0,2312	5 % (100 %)
Rain (mm)	-0,6376	41 % (41 %)	0,7589	58 % (98 %)	0,1321	2 % (100 %)	0,0089	0 % (100 %)
Var. Expl.	2,7915	70 % (70 %)	0,7145	18 % (88 %)	0,3659	9 % (97 %)	0,1281	3 % (100 %)

Eigen vectors -- Factor Scores

Attribute	Mean	Std-dev	Axis_1	Axis_2	Axis_3	Axis_4
Temperature (°C)	3157,882353	139,092598	0,550458	0,305215	-0,207543	-0,748843



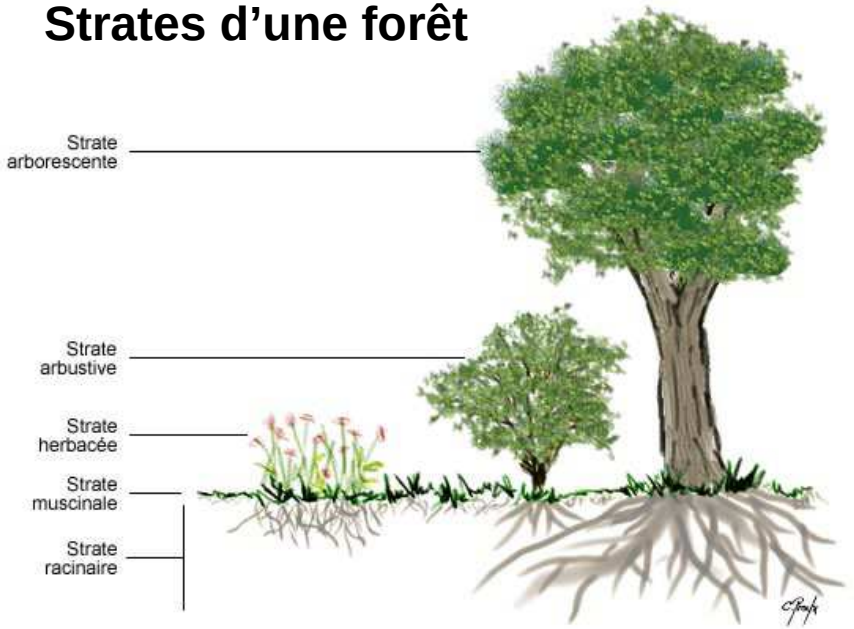
Précipitation

Température
Soleil
Durée d'ensoleillement

EXEMPLES

Plan stratifié

Strates d’une forêt



- La classification périodique des éléments
- Les voitures d’une ville
- Les fruits d’un verger
- Les élèves d’une classe
- Les articles d’un magasin
- Les victimes de la grippe A
- ….

Yann.kergadallan@cea.fr

APPLICATION

Problématique : distribution des fourmis dans les bois

Choix des stratificateurs : (variables qui influencent la présence de fourmis), intuitivement (plan de jugement) :

1. L'altitude
   3 zones : 380-800, 801-1200, >1200m
2. L'exposition au soleil
   2 zones : favorable et défavorable
3. La pente (stabilité de la fourmilière)
   2 zones 1 – 20° et 25 – 45°)
4. Couvert végétal
   2 zones : lisière et pleine forêt

1. Identification des 3×2×2×2 = 24 strates sur une carte
2. Tirage au sort de 10 quadrat de 1 ha dans les 24 strates
   → soit 240 unités d'échantillonnage
3. Plan en transect :
   - unités parallèles espacées de 15 m,
   - recensement des fourmilières et identification des espèces

Yann.kergadallan@cea.fr

EXEMPLES DE CLASSIFICATEURS

Plan stratifié

Classificateurs "intuitifs" :

- La couleur (photo satellite, terrain pollué, …)
- La forme (espèces animales ou végétales, la taille, …)

Classificateurs numériques :

- Les facteurs abiotiques en écologie (températures, exposition à la lumière, …)
- Les facteurs sociologiques (genre, revenu, âge, quartier, …)

- Stratification temporelle : les saisons, les mouvements d'une symphonie,
- Stratification spéciale : les paysages, les écosystèmes, les qualités de sols
- Stratification sociale : les villes, les quartiers

STRATIFICATION : UTILISATION D'UN ALGORITHME (CLUSTERING)

Plan stratifié

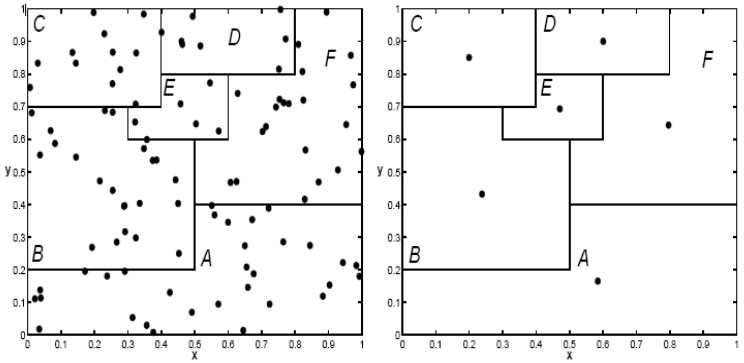
Échantillonnage aléatoire de quelques individus (pré-échantillonnage) puis :

1. ACP : analyse des corrélations entre les variables permet d'identifier une variable auxiliaire (ACP)

2. K-mean : identification des éventuelles classes/strates et du classificateur

Classification K-mean :

- Les unités sont regroupées en strates avec un centre de gravité comme représentant
- Stratificateur = classificateur = distance euclidienne



(a) Stratified random $m = 100$ (b) Stratified centroid $m = 6$

On remarque qu'une des caractéristiques d'un plan stratifié est le poids des strates

UTILISATION DES TESTS STATISTIQUES

Tester deux hypothèses complémentaires :

- H_0 : toute les valeurs des classes sont similaires
- H_1 : il existe au moins un écart significatif entre deux valeurs de classes

Choix du risque α —*Table de la loi de Fisher- Snedecor*→ $F_{seuil} = F_{\alpha}(p-1,N-p)$

Calcul des écarts quadratiques entre (SCE_{inter}) et dans (SCE_{intra}) les strates puis de F_{obs}

$$F_{obs} = \frac{\frac{SCE_{inter}}{p-1}}{\frac{SCE_{intra}}{N-p}} = \frac{N-p}{p-1}\frac{SCE_{inter}}{SCE_{intra}} \propto \frac{\rho}{1-\rho}$$

Plus la valeur de F est élevée (et en particulier si $F_{obs}>F_{seuil}$) , plus la stratification est pertinente (efficace)

ESTIMATION DES PARAMÈTRES STATISTIQUES, DANS UNE STRATE H PARMI M STRATES

Population totale : N

Plan stratifié

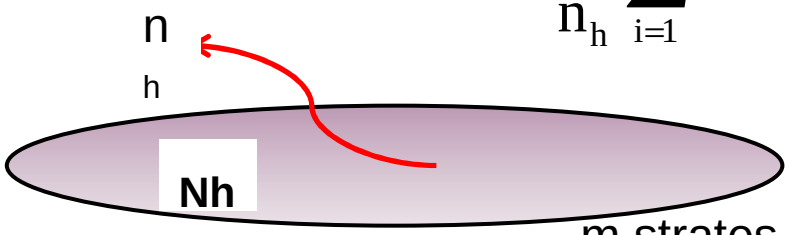
Cette population est divisée en m strates h de taille N_h

Dans chaque strate h, un échantillonnage aléatoire de n_h unités est constitué

1/ Paramètres statistiques au niveau d'une strate :

Moyenne dans une strate

$$\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{ih}$$



m strates de taille N_h

Variance dans une strate

$$\sigma^2_{\bar{x}_h} = (1-f_h).\frac{s_h^2}{n_h}$$

$$s_h^2 = \frac{1}{1-n_h}\sum_{i=1}^{n_h}(x_{ih}-\bar{x}_h)^2$$

Plan stratifié

2/ Paramètres statistiques entre strates :

Pour chaque strate, on défini un poids $W_h = \frac{N_h}{N}$

La moyenne totale est alors : $\overline{\overline{x}} = \sum_{h=1}^{m} w_h . \overline{x}_h$

Et la variance sur cette moyenne : $\sigma^2_{\overline{\overline{x}}} = \sum_{h=1}^{m} w_h^2 . \sigma_h^2$

EXEMPLE D'APPLICATION

Population : 210 communes - Ces communes sont stratifiées en 4 strates :

	Strate 1	Strate 2	Strate 3	Strate 4
Nh	105	63	21	21
Échantillonnage	7	14	23	24
	10	11	36	110
	8	14	2	17
	2	7	9	47
	7	17	24	32
	6	19	Ech3 $n_3=5$	Ech4 $n_4=5$
	2	9		
	6	Ech2 $n_2=7$		
	Ech1 $n_1=8$			

CALCUL DE LA MOYENNE

Strate	1	2	3	4
Effectif	105	63	21	21
Taille de l'échantillon	8	7	5	5
Moyenne	2,775	4,2817	13,4052	37,4767
Ecarts-types de l'échantillon	2,7775	4,2817	13,4052	37,4767
$\sqrt{1-f_h / n_h}$	0,3398	0,3564	0,3904	0,3904
Ecarts-types des estimateurs	0,9438	1,5258	5,2329	14,6294
Poids de la strate	0,5	0,3	0,1	0,1

Calcul de la moyenne totale :

$$\bar{\bar{x}} = \sum_{h=1}^{m} w_h . \bar{x}_h = (0,5.6) + (0,3.13) + (0,1.18,8) + (0,1.46) = 13,38$$

CALCUL DE LA VARIANCE

On calcule

$$\hat{\sigma}^2_{\bar{y}1} = (1-\frac{8}{105})\frac{2{,}7775^2}{8} = 0{,}6886$$

$$\hat{\sigma}^2_{\bar{y}2} = 2{,}3280$$

$$\hat{\sigma}^2_{\bar{y}3} = 27{,}3829$$

$$\hat{\sigma}^2_{\bar{y}4} = 214{,}019$$

Finalement :

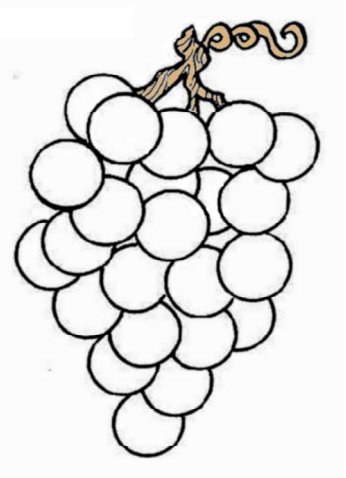
$$\sigma^2_{\bar{\bar{x}}} = \sum_{h=1}^{m} w_h^2.\sigma_h^2 = (0{,}5^2.0{,}6886)+(0{,}3^2.2.3280)+(0{,}1^2.27{,}3829)+(0{,}1^2.214{,}019) = 2{,}7957$$

L'intervalle de confiance est :

$$\bar{\bar{y}} - 2.\hat{\sigma}_{\bar{\bar{y}}} \leq \mu° \leq \bar{\bar{y}} + 2.\hat{\sigma}_{\bar{\bar{y}}} \Rightarrow \mathbf{10{,}2859 \leq \mu° \leq 16{,}9741}$$

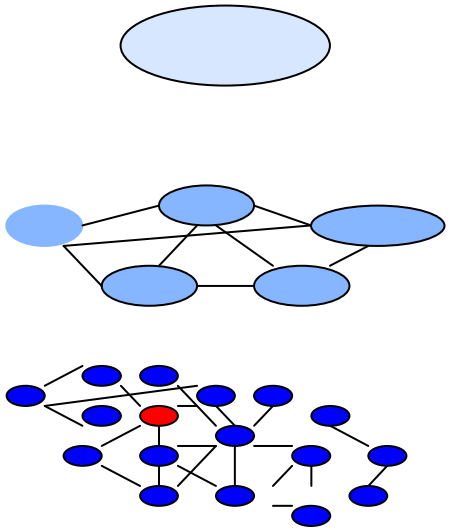
GRAPPES

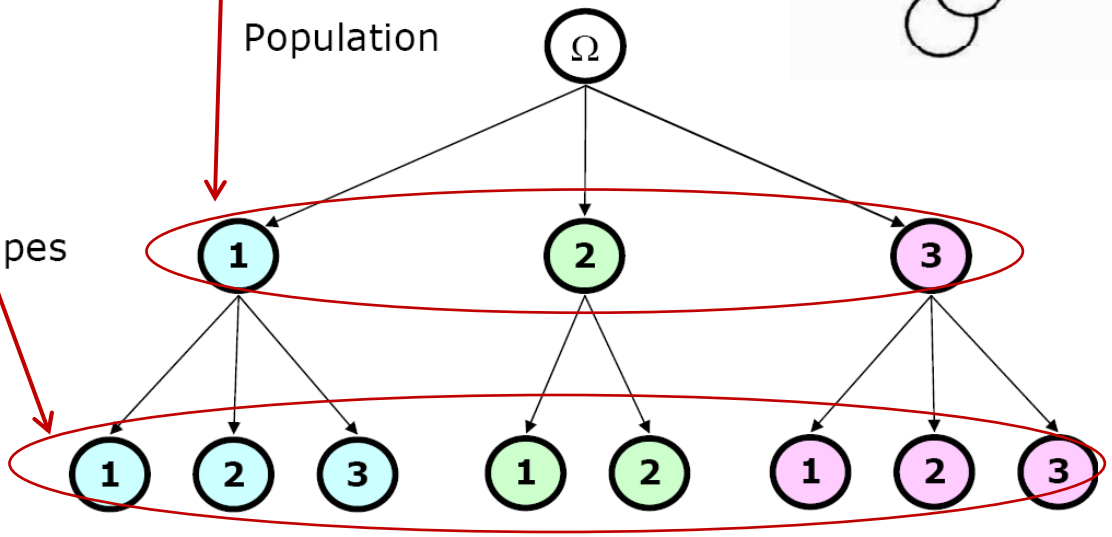
Plan en grappe



Unités primaires

Unités secondaires





EXEMPLES DE GRAPPES

Plan en grappe

Reproduction de la chouette Effraie en Côte d'Or.

- Visite systématique des clochers pour recensement des nichées.
- Une fraction des clochers abritant une nichée a été tirée au hasard.
- Tous les jeunes (grains) des nichées sélectionnées (grappes) ont été pesés.

Etude une caractéristique (ex: attaque par un parasite) de la population de bovins de moins d'un an dans une région.

- Sélection au hasard de cheptels dans cette région
- Identification de la valeur de la caractéristique étudiée sur tous les bovins de moins d'un an

Les plans évolutifs

- Plan par attributs (*attribute sampling*)
- Plan progressif ou adaptatif
  (*sequential, adaptive sampling*)
- Plan « boule de neige » (*snow ball sampling*)

Yann.kergadallan@cea.fr

GÉNÉRALITÉS

Le plan **progressif** (**adaptive** sampling en anglais) est un **plan qui modifie sa stratégie d'échantillonnage** en fonction des résultats observés.

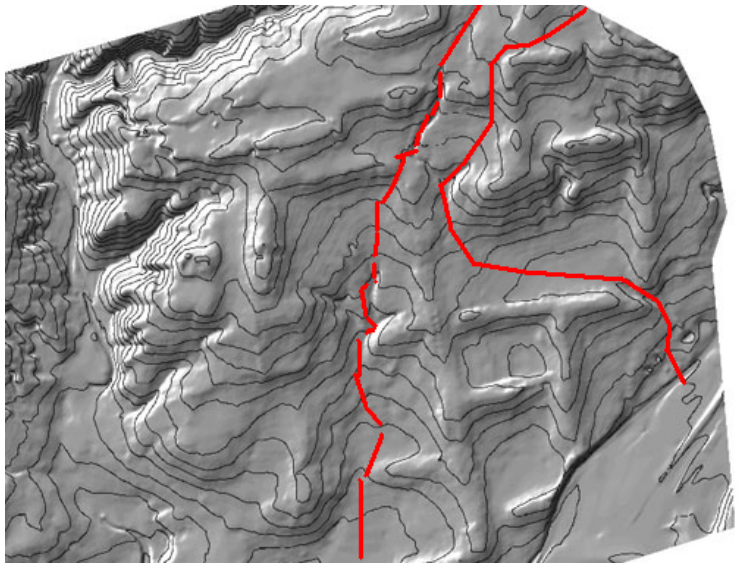
Remarque : l'algorithme qui permet de réaliser un tel plan doit être déterminé avant la réalisation de l'échantillonnage

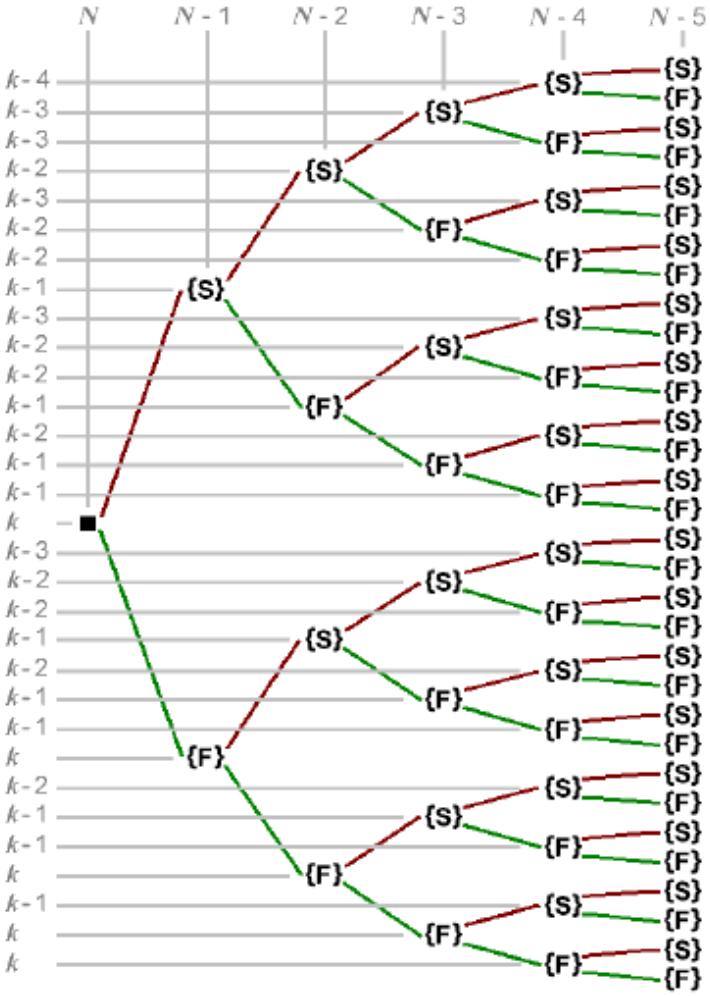
Il ne doit pas y avoir un apport extérieur.

Dans ce genre de plan l'information est apportée progressivement, il y a donc une histoire de l'échantillonnage, un processus d'apprentissage.

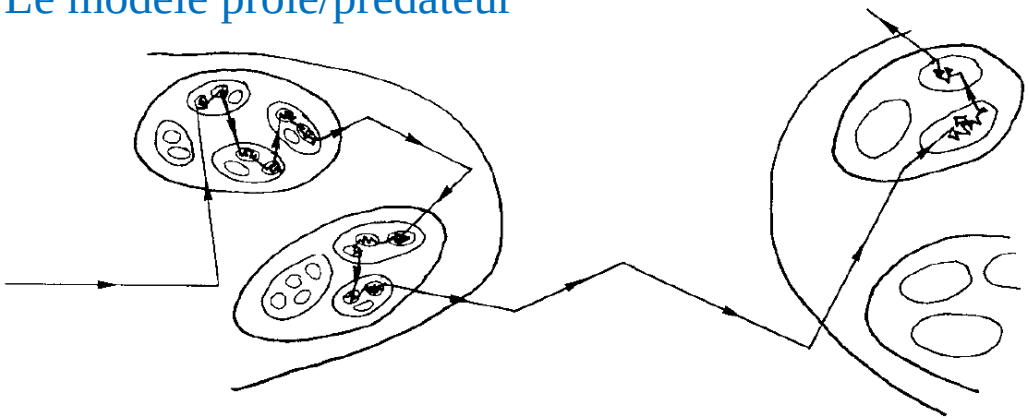
*Une **approche Bayesienne** est souvent nécessaire pour la modélisation*

ASPECT GRAPHIQUE





Le modèle proie/prédateur



CAS DU PRÉ-ÉCHANTILLONNAGE

Un cas plus simple d'échantillonnage séquentiel est le pré-échantillonnage

« pour faire un bon échantillonnage commencez par un mauvais »

- Le plan se fait souvent en deux niveaux
- Le critère est souvent une précision sur la variance expérimentales ce qui permet de dimensionner au mieux le nombre d'échantillons n à prélever

1. Echantillonnage séquentiel de n_1 unités
2. Estimation du nombre d'échantillon n
3. Sélection des $n_2 = n-n_1$ unités nécessaires à l'échantillon

PROBABILITÉ D'INCLUSION

Approche séquentielle

$$\pi_k = 1 - \left[\frac{\binom{N - x_k}{n_1}}{\binom{N}{n_1}} \right]$$

Probabilité d'intersection (d'inclusion) :

$$\pi_{j,k} = 1 - \left[\binom{N - x_j}{n_1} + \binom{N - x_k}{n_1} + \binom{N - x_j - x_k}{n_1} \right] \Big/ \left[\binom{N}{n_1} \right]$$

Où:

k est le nombre de cluster ou réseaux indépendants identifiés

n_1 est la taille de l'échantillon aléatoire initial

N est le nombre total d'unités (dans l'exemple, des mailles)

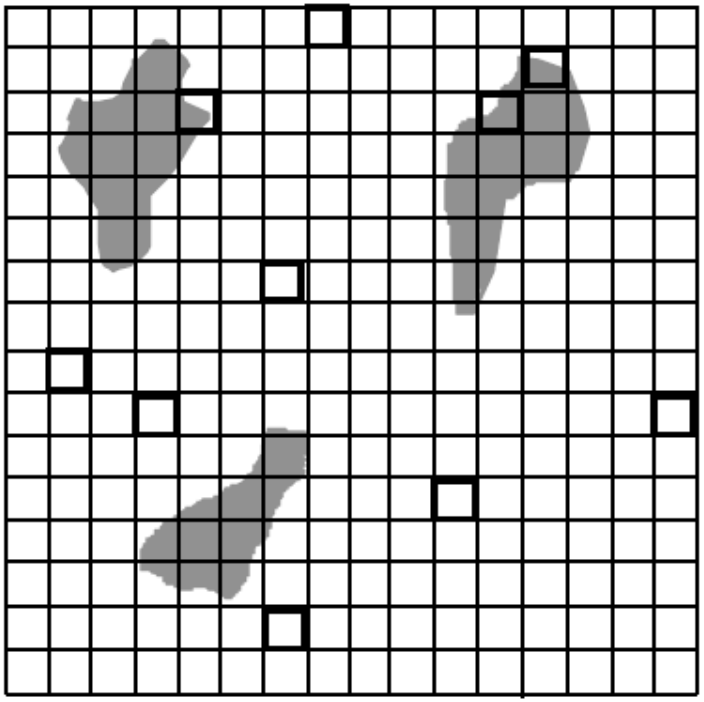
x_j ou x_k sont les valeurs de la caractéristique étudiée pour les unités i et k respectivement

Pour la moyenne : $$\hat{x} = \sum_k^n \frac{1}{N} \cdot \frac{1}{\pi_k} . x^*_k$$

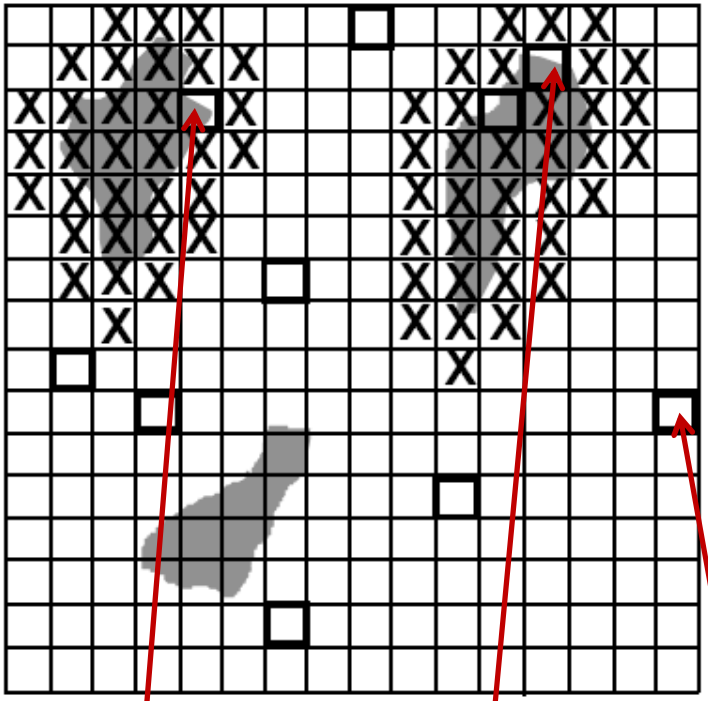
Pour la variance : $$\hat{V}(X) = \frac{1}{N^2} \sum_j \sum_k \left[\frac{1}{\pi_{j,k}} \cdot \frac{\pi_{j,k} - \pi_j \pi_k}{\pi_j \pi_k} \right] . x_j^* x_k^*$$

APPLICATION À UNE SURFACE CONTAMINÉE

Approche séquentielle



Domaines d'intérêt en gris, réalisation d'un échantillonnage aléatoire à partir d'un maillage



Résultats d'un plan progressif lancer sur les grappes identifiées avec le plan précédant

Il y a au total 256 mailles, l'échantillon aléatoire initial est de $n_1 = 10$

Zone 1 (taille 18)

Zone 2 (taille 19)

Zone i (taille 1)

IDENTIFICATION DES GRAPPES

Approche séquentielle

Finalement sur les 10 unités initiales on a :

Nombre total de réseaux identifiés : k=9 (2 grands et 7 petits)

$X_1 = 18$

X_2 (2 unités pour cette valeur) $= 19$

$X_3 = X_4 = X_5 = X_6 = X_7 = X_8 = X_9 = 1$

CALCUL DES PROBABILITÉ D'INCLUSION

Approche séquentielle

$$\pi_1 = 1 - \left[\binom{256-18}{10} \Big/ \binom{256}{10} \right] = 1 - \left[\left(\frac{238!}{10!228!} \right) \Big/ \frac{256!}{10!246!} \right] = 0.5241791$$

$$\pi_2 = 1 - \left[\binom{256-19}{10} \Big/ \binom{256}{10} \right] = 1 - \left[\left(\frac{237!}{10!227!} \right) \Big/ \frac{256!}{10!246!} \right] = 0.5441714$$

$$\pi_k = 1 - \left[\binom{256-1}{10} \Big/ \binom{256}{10} \right] = 1 - \left[\left(\frac{255!}{10!245!} \right) \Big/ \frac{256!}{10!246!} \right] = 0.0390625$$

Ce qui permet d'en déduire la moyenne :

$$\hat{\mu} = \frac{1}{256} \left[\frac{x_1^*}{0.5241791} + \frac{x_2^*}{0.5441714} + \frac{x_3^*}{0.0390625} \right]$$

CALCUL INTERMÉDIAIRE POUR L'ESTIMATION DE LA VARIANCE

Approche séquentielle

$$\pi_{jk} = \pi_{kj} = 1 - \left[\binom{256-1}{10} + \binom{256-1}{10} + \binom{256-2}{10}\right] \Big/ \left[\binom{256}{10}\right] = 0.0013786$$

Pour j = 3, 4, ….9 et k= 3, 4, ….9, j≠k

$$\pi_{1,J} = \pi_{j,1} = 1 - \left[\binom{256-18}{10} + \binom{256-1}{10} + \binom{256-19}{10}\right] \Big/ \left[\binom{256}{10}\right] = 0.0190701$$

Pour j = 3, 4, ….9

$$\pi_{2,j} = \pi_{j,2} = 1 - \left[\binom{256-19}{10} + \binom{256-1}{10} + \binom{256-20}{10}\right] \Big/ \left[\binom{256}{10}\right] = 0.0198292$$

Pour j = 3, 4, ….9

RÉSULTATS SUR L'ESTIMATION DE LA VARIANCE APRÈS DE NOMBREUX CALCULS !

Approche séquentielle

$$\widehat{\text{var}}(\hat{\mu}) = \frac{1}{256^2}\left[\sum_{j=1}^{9}\sum_{k=1}^{9}\frac{y_j^* y_k^*}{\alpha_{jk}}\left(\frac{\alpha_{jk}}{\alpha_j\alpha_k}-1\right)\right]$$

$$= \frac{1}{256^2}\left[\sum_{j=1}^{9}\frac{\left(y_j^*\right)^2}{\alpha_j}\left(\frac{1}{\alpha_j}-1\right)+2\sum_{j=1}^{8}\sum_{k=j+1}^{9}\frac{y_j^* y_k^*}{\alpha_{jk}}\left(\frac{\alpha_{jk}}{\alpha_j\alpha_k}-1\right)\right]$$

$$= \frac{1}{256^2}\left[\frac{\left(y_1^*\right)^2}{\alpha_1}\left(\frac{1}{\alpha_1}-1\right)+\ldots+\frac{\left(y_9^*\right)^2}{\alpha_9}\left(\frac{1}{\alpha_9}-1\right)\right]+\frac{2}{256^2}\left[\frac{y_1^* y_2^*}{\alpha_{12}}\left(\frac{\alpha_{12}}{\alpha_1\alpha_2}-1\right)+\ldots+\frac{y_8^* y_9^*}{\alpha_{89}}\left(\frac{\alpha_{89}}{\alpha_8\alpha_9}-1\right)\right]$$

$$= \frac{1}{256^2}\left[\frac{\left(y_1^*\right)^2}{0.524179}\left(\frac{1}{0.524179}-1\right)+\ldots+\frac{\left(y_9^*\right)^2}{0.0390625}\left(\frac{1}{0.0390625}-1\right)\right]$$

$$+\frac{2}{256^2}\left[\frac{y_1^* y_2^*}{0.2719547}\left(\frac{0.2719547}{(0.524179)(0.5441714)}-1\right)+\ldots+\frac{y_8^* y_9^*}{0.0013786}\left(\frac{0.0013786}{(0.0390625)(0.0390625)}-1\right)\right]$$

⇒Il vaut mieux se faire aider d'un logiciel

Visual Sampling Plan

ON NE PEUT PAS TOUT DEMANDER OU ACCEPTER !!

Approche décisionnelle

Au départ : une relation client/fournisseur formalisée par un contrat

Clarification des exigences sur des unités produites (on ne peut pas tout accepter !) :

- Pour le fournisseur : une qualité maximale techniquement possible (plus d'exigences le contraindrait à modifier complètement son procédé), cette contrainte qui est à l'origine du NQA (ou NDQ).

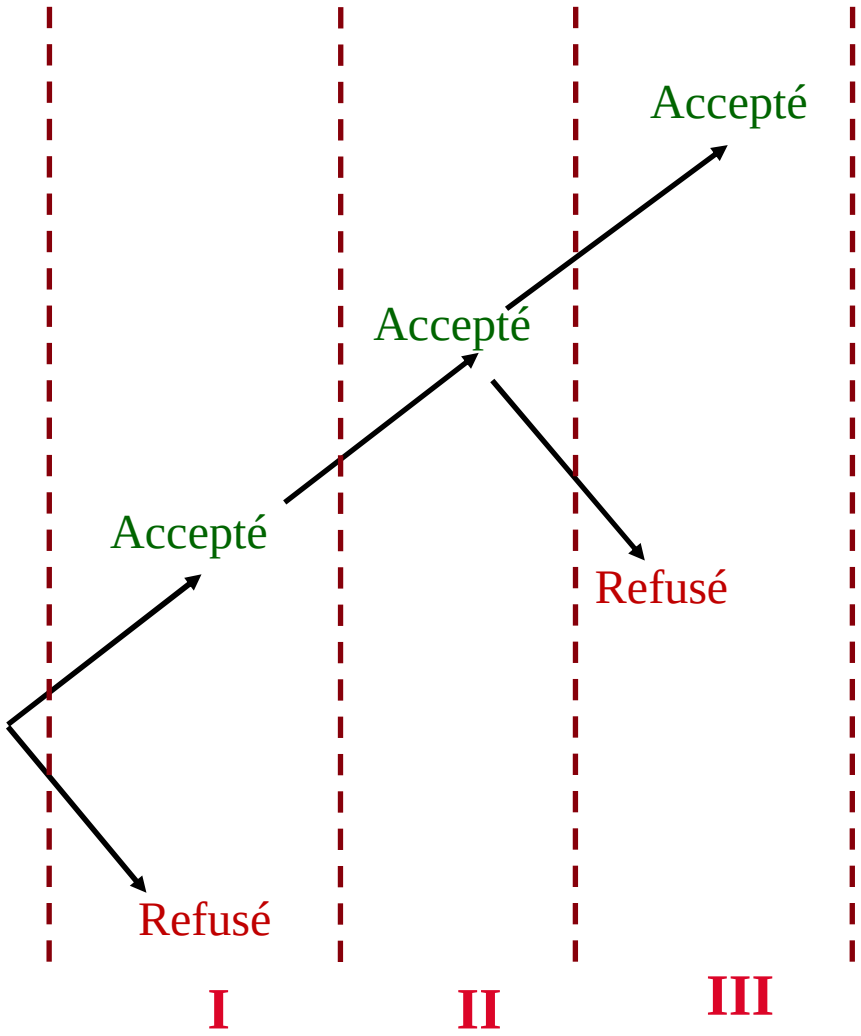
- Pour le client : un minimum de qualité (accepter moins ce serait ne pas pouvoir utiliser ces unités pour un usage ultérieur) , cette contrainte est à l'origine du NQL

Comme,il s'agit d'un échantillonnage et non pas d'un recensement, il existe un risque pour les deux contraintes NQA et NQL:

- Conventionnellement, le risque NQA du fournisseur est fixé à $\alpha = 5\%$
- Conventionnellement, le risque NQL du client est fixé à $\beta = 10\%$

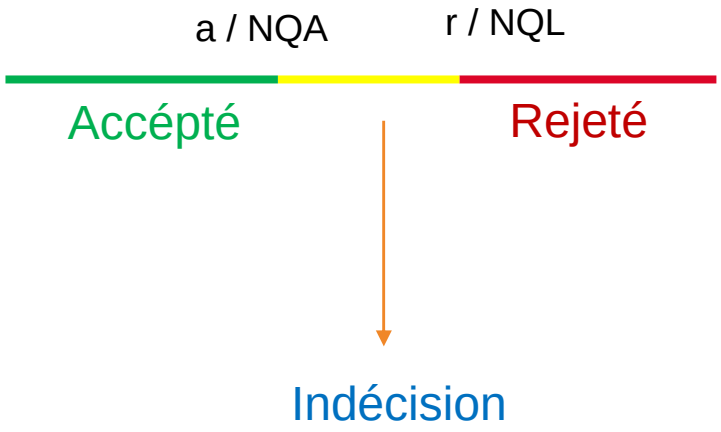
ARBRE DE DÉCISION (EXEMPLE : LE CONTRÔLE STATISTIQUE)

Approche décisionnelle



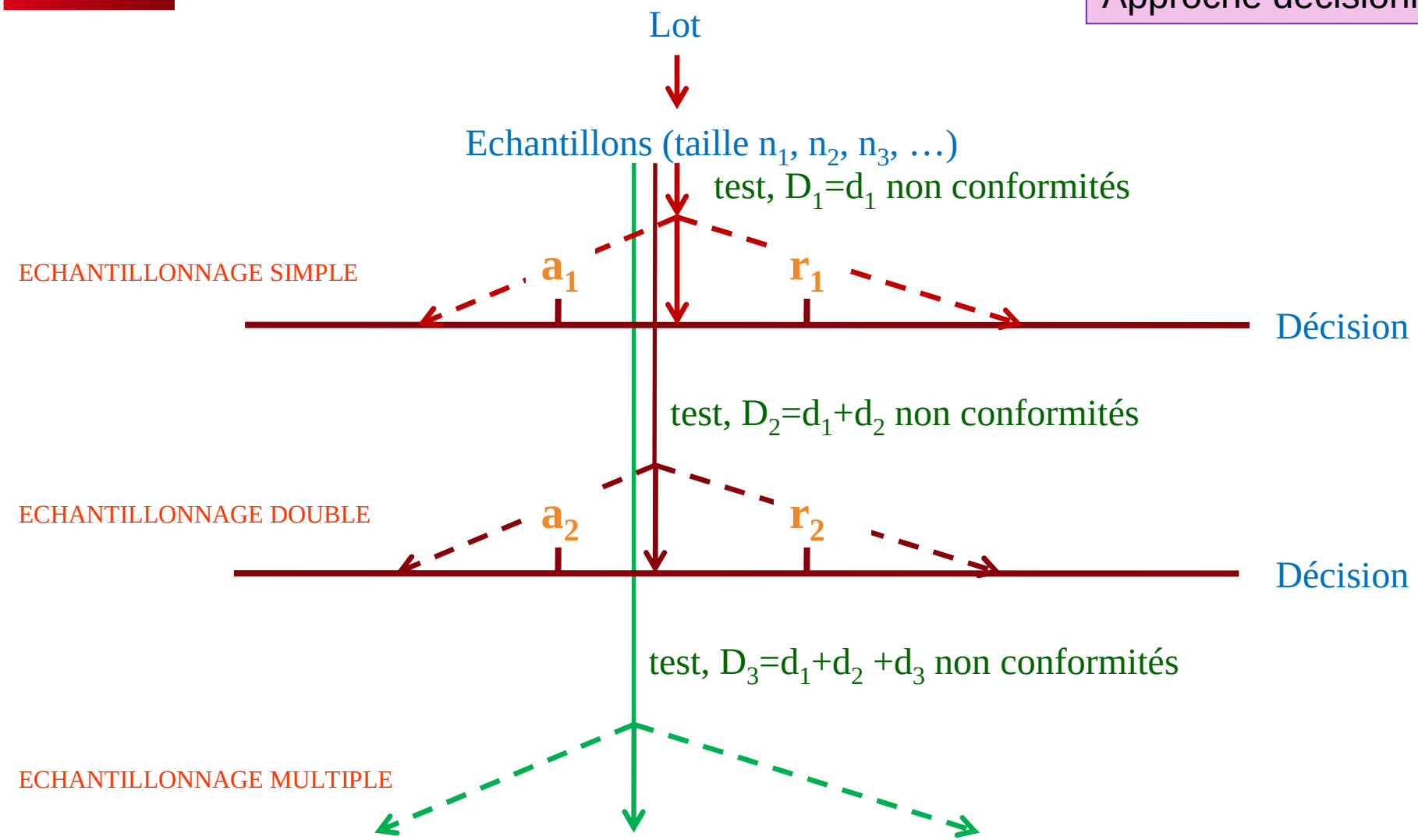
Paramètres importants :

1. La proportion de non-conformité observées dans l'échantillon
2. La valeur d'acceptation (a ou NQA)
3. La valeur de rejet (r ou NQL)



PROCÉDURE : PLAN SIMPLE, DOUBLE ET MULTIPLE

Approche décisionnelle



ÉCHANTILLONNAGE ET PARAMÈTRES

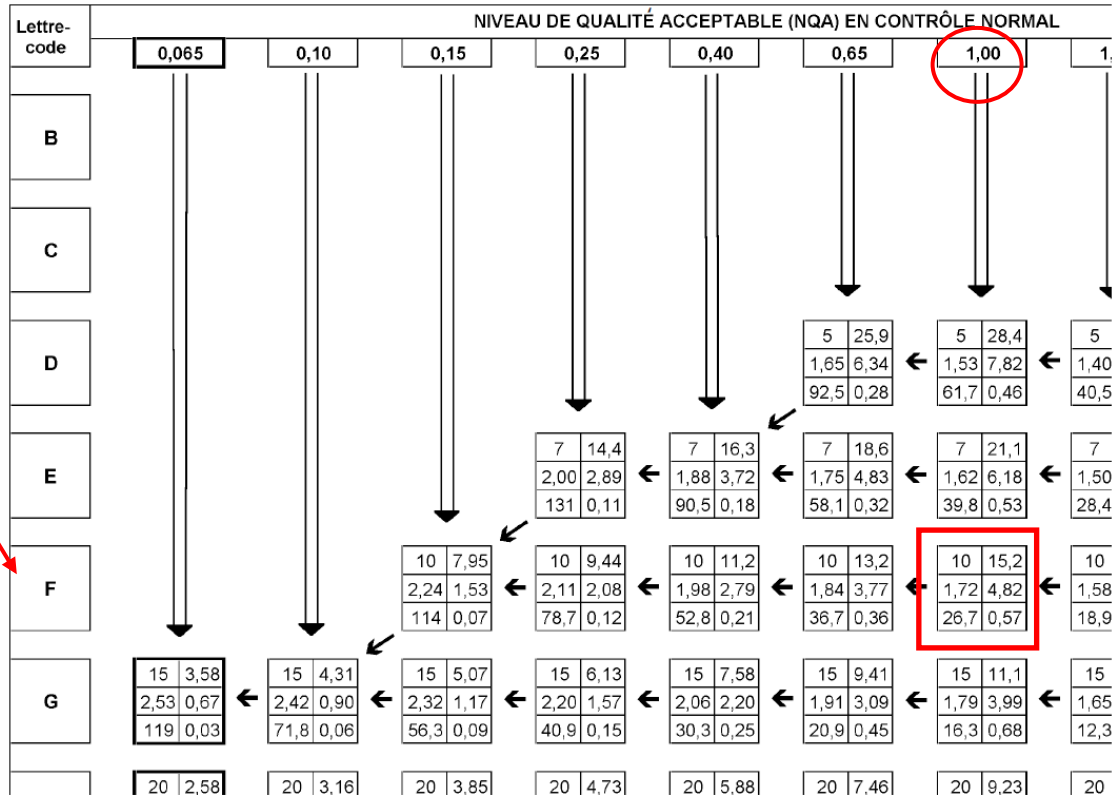
Approche décisionnelle

- Le nombre total d'unités (la population)
- Qualité acceptable NQA (en général de 0,07% à 4 %)

→ Lettre codée
Les paramètres

Effectif N des lots	Lettres-codes
Inférieure à 16	(B)
16 à 25	(C)
26 à 50	D
51 à 90	E
91 à 150	F
151 à 280	G
281 à 400	H
401 à 500	I
501 à 1 200	J
1 201 à 3 200	K
3 201 à 10 000	L
10 001 à 35 000	M
35 001 à 150 000	N
150 001 et au-dessus	P

taille de l'échantillon	n	$P_{10\%}$
constante d'acceptation	K	$P_{50\%}$
rapport de discrimination	$DS=P_{\alpha}/P_{1-\beta}$	$P_{95\%}$



ÉCHANTILLONNAGE ET PARAMÈTRES

Les paramètres sont les suivant :

- La taille de l'échantillon **n**
- Des seuils pour les tests :
    1. une valeur d'**a**cceptation (**a**)
    2. une valeur de **r**ejet (**r**)
- Une valeur du résultats du tests sur les n unités : **d** non conformes
- La proportion d'unités non conformes : **p = d/n**

EXEMPLE1 D'APPLICATION ET NORMES

Approche décisionnelle

Calculer le rapport qui suit et le comparer avec la valeur de K

$$Q_s = \frac{X_{min-accep} - \overline{X}}{s_{exp}}$$

Exemple :

- contrôle de N=100 thermocouples ⇒ Lettre code = **F**
- $X_{max\text{-}accept}$ = 60°C
- NQA = 1%

  {1% et F} ⇒ n=10 - **K = 1,72** - $P_{10\%}$ = 15,2 % - $P_{95\%}$ = 0,57%

- Mesure sur l'échantillon : (53, 57, 49, 58, 54, 50, 56, 55, 50, 59)

  moyenne = 54,9 & s = 3,414 ⇒ **Q_s = 1,494** < K=1,72

  ⇒ Q_s < K le lot est rejeté

Normes internationales :

ISO 2859, Parties 1 - 2 - 3 - 4

Normes françaises :

NF X 06-23, NF X 06-24, NF X 06-25

EXEMPLE 2 D'APPLICATION

Approche décisionnelle

Objectif : optimisation d'un plan de maintenance préventive dans un hôpital

Unités : équipements peu complexes (pousse-seringue, manomètres, …) mais en grand nombre

Paramètres :

- L'importance de l'impact (ici sur le patient)
- L'effectif (c'est-à-dire la taille de la population)
- La multiplicité (nombre de fois échantillonnées)
- Le Niveau de Qualité Acceptable

EFFECTIF ET CRITICITÉ

Approche décisionnelle

On dispose d'un ***ensemble de 225 unités*** :

- Les tensiomètres (100 unités)
- Les monitorings (40 unités)
- Les pousses seringues (85 unités)

En se basant sur ces effectifs, on peut proposer le nombre d'échantillonnage suivant :

- Les tensiomètres → Multiple
- Les monitorings →Simple
- Les pousses seringues →Double

Criticité faible, pas d'impact, Niveau I
Criticité normale, impact sur la qualité du service (pas sur le patient), Niveau II
Criticité élevée, impact direct sur le patient, Niveau III

A partir de ces critères, le classement suivant est obtenu:

- Les tensiomètres →Niveau I
- Les monitorings →Niveau II
- Les pousses seringues →Niveau III

LA LETTRE CODE

Approche décisionnelle

Avec l'effectif et le niveau de criticité on recherche la lettre code :

Effectif du lot	Niveau I	Niveau II	Niveau III
2 à 8	A	A	B
9 à 15	A	B	C
16 à 25	B	C	D
26 à 50	C	D	E
51 à 90	C	E	F
91 à 150	D	F	G
151 à 280	E	G	H
281 à 500	F	H	J
501 à 1200	G	J	K
1201 à 3200	H	K	L
3201 à 10000	J	L	M
10001 à 35000	K	M	N
35001 à 150000	L	N	P
150001 à 500000	M	P	Q
500001 et plus	N	Q	R

- Les tensiomètres →Niveau I, 100 : D
- Les monitorings →Niveau II, 40 : D
- Les pousses seringues →Niveau III, 85 : F

LA TAILLE DES ÉCHANTILLONS ET NAQ

Approche décisionnelle

Taille des échantillons														
Lettre code	A	B	C	D	E	F	G	H	J	K	L	M	N	P
Simple n	2	3	5	8	13	20	32	50	80	125	200	315	500	
Double n_1	-	-	3	5	8	13	20	32	50	80	125	200	315	
n_2	-	-	3	5	8	13	20	32	50	80	125	200	315	
Multiple n_1	-	-	-	2	3	5	8	13	20	32	50	80	125	
n_2	-	-	-	2	3	5	8	13	20	32	50	80	125	
n_3	-	-	-	2	3	5	8	13	20	32	50	80	125	
n_4	-	-	-	2	3	5	8	13	20	32	50	80	125	200
n_5	-	-	-	2	3	5	8	13	20	32	50	80	125	200
n_6	-	-	-	2	3	5	8	13	20	32	50	80	125	200
n_7	-	-	-	2	3	5	8	13	20	32	50	80	125	200

- Les tensiomètres → Multiple (ni =2)
- Les monitorings →Simple (n=8)
- Les pousses seringues →Double (n_1= n_2=13)

Échantillon accepté si le nombre d'unités défectueuses est inférieur à 1

Échantillon rejeté dès qu'il y a deux unités non conforme

Compte tenu
- du risque client (10%) et fournisseur (5%)
- de la lettre code

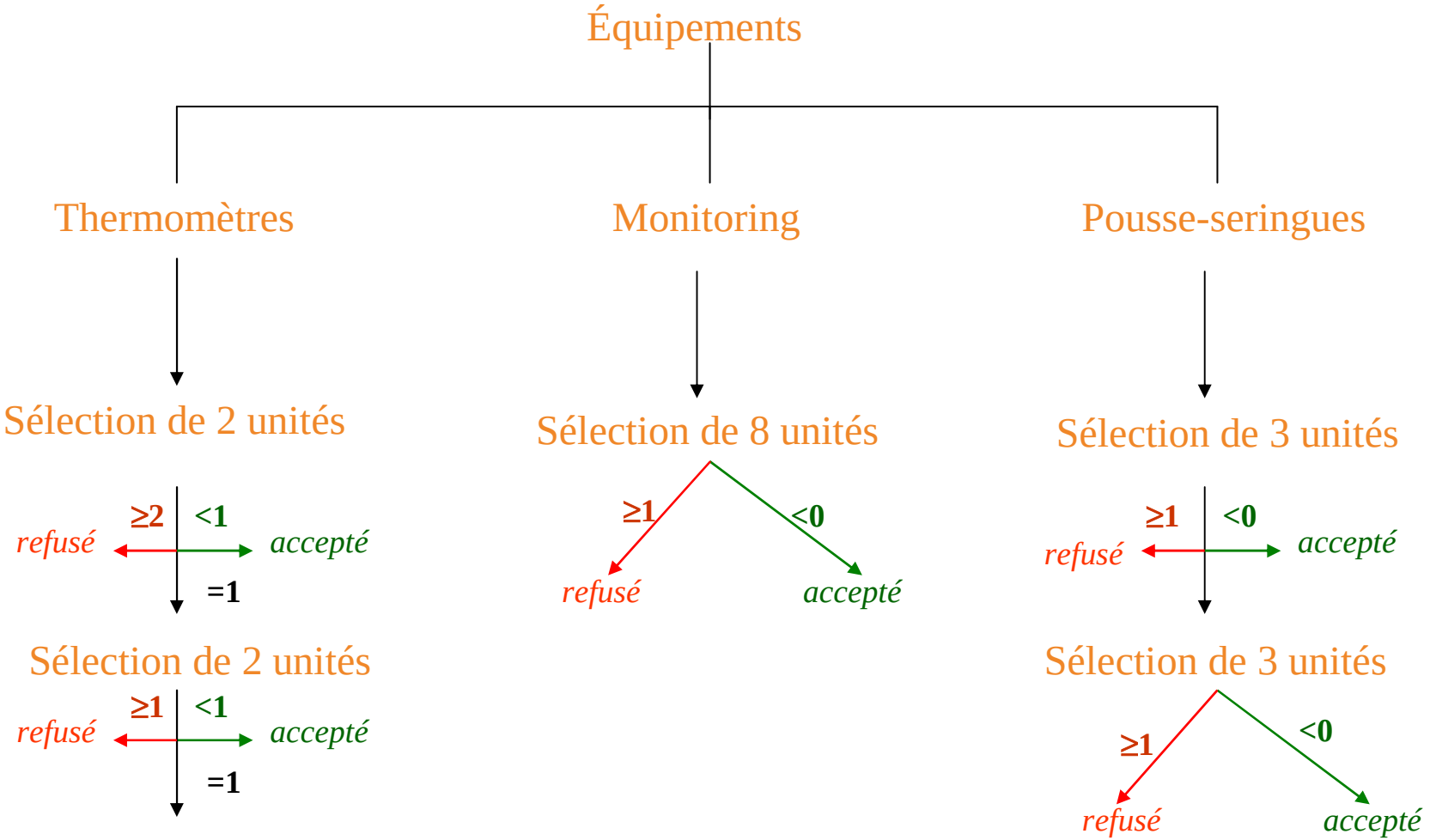
→ Un intervalle de valeurs possibles [P_{95}, P_{10}]

Le choix se fait à partir de résultats de tests sur les taux de fiabilité

- Les tensiomètres → [2,64; 40,6], choix de NAQ = 6,5 → Ac=1, Re = 2
- Les monitorings →[0,64; 25], choix de NAQ = 1,5 → Ac=0, Re = 1
- Les pousses seringues →[0,256; 10,9], choix de NAQ = 0,65 → Ac=0, Re = 1

CONSTITUTION ET TEST DES UNITÉS

Approche décisionnelle



POUR ALLER PLUS LOIN

La sollicitation des équipements est différente suivant les services (le taux de fiabilité est donc lui aussi différent)

⇒ Nécessité de stratifier son plan

1. Urgences
2. Réanimation
3. Pédiatrie

Le poids (taille du sous-échantillon) proportionnel au taux de sollicitation

Exemple des monitoring : il faut 8 unités :

- 4 sélectionnés aléatoirement aux urgences
- 3 sélectionnées aléatoirement en réanimation
- 1 sélectionnée aléatoirement en pédiatrie

L'échantillon final est constitué de toutes les unités, on a donc

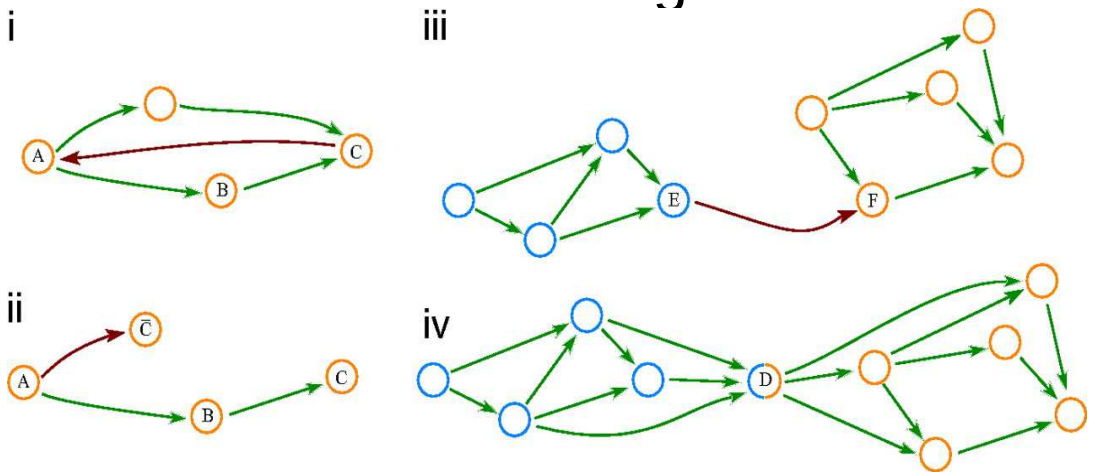
Plan stratifié → Plan aléatoire → Plan composite ou mélange

Echantillonnage en réseau

Yann.kergadallan@cea.fr

NOTATIONS

- V : ensemble des nœuds (verticles, nodes) Ni
- E : ensemble des arrêtes (donc ensemble de couples (i,j) supposées non orientées)
- G = {V, E} graphe réel
- G* = {V*, E*} graphe empirique (construit avec l'échantillonnage)
- Di : degré du nœud i
- N(k) : distribution des degrés en fonction des nœuds
- Cl(G) : coefficient de clustering



LE DEGRÉ : RÉPARTITION DES CONNEXIONS PAR NŒUD

Degré du nœud i :

nombre de connexions d_i de ce nœuds avec d'autres nœuds

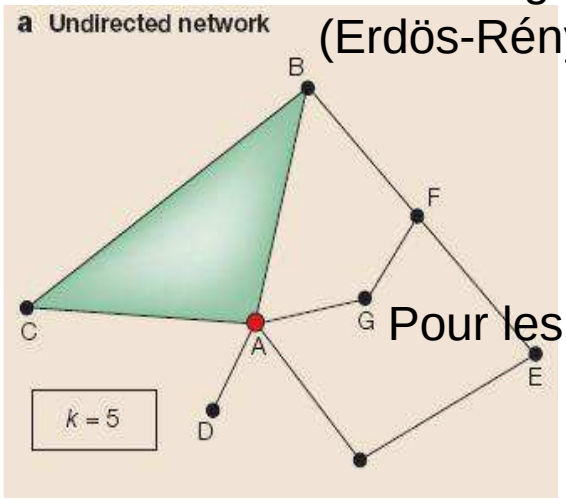
(pour un réseau cristallin c'est le nombre de coordination)

La distribution des degré :

Fréquence des nœuds n(k) ayant un degré fixé k

Pour des graphes aléatoires (Erdös-Rény)

$$n(k) = N.P^k.(1-P)^{N-k}.\binom{N}{k} \approx \frac{(NP)^k e^{-NP}}{k!}$$



Pour les « free-scale »

$$n(k) \approx \frac{N.k^{-\gamma}}{\xi(\gamma)} \text{ avec } \xi(\gamma) = \sum_1^\infty \frac{1}{\gamma^i} \text{ si } \gamma > 1$$

COEFFICIENT DE CLUSTERING : IMPORTANCE DES RELATIONS AVEC LE VOISINAGE

Pour un nœud i :

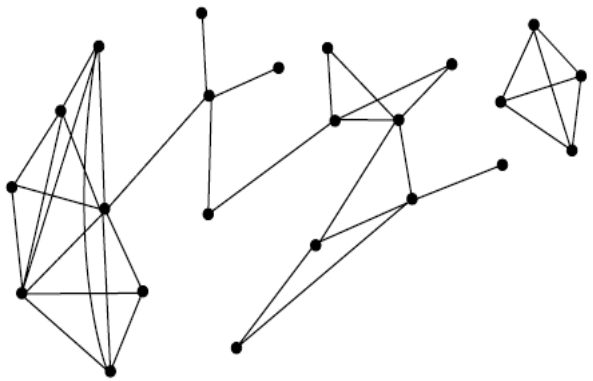
Nombre d'arrêtes possible : $k_i.(k_i-1)/2$

Nombre réellement observé : K_i

Ce qui donne le coefficient (connectance en écologie)

$$C_i = \frac{2.K_i}{k_i(k_i - 1)} \text{ si } k > 2$$

Pour l'ensemble du réseau : $C = \frac{1}{N}\sum_{1}^{N} C_i$



DISTANCE ET DIAMÈTRE

La distance li,j entre deux nœuds (d'un nœud i source à un nœud cible j) est le nombre d'arrêtes rencontrées quand on emprunte le chemin le plus court

La distance moyenne
(qui donne une idée sur la longueur des chemins dans le graphe :

$$\bar{l} = \frac{2}{N.(N-1)} . \sum_{i,j} l_{i,j}$$

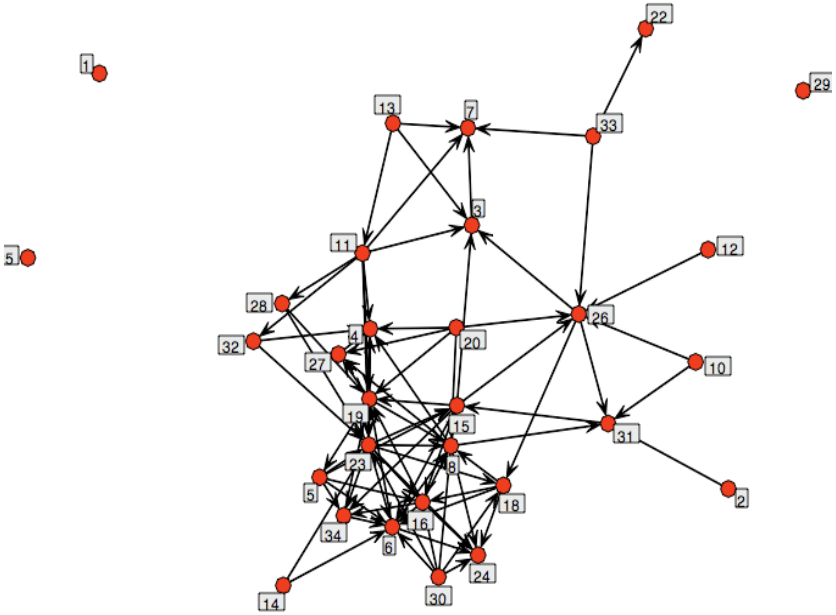
Parmi tous ces chemins, il en existe un plus long : c'est le diamètre Φ du réseau

On peut montrer que : $\Phi \propto \log(N)$

BETWEENESS

C’est une mesure de la fréquence de passage par un nœud donné i quand on circule dans le graphe. Il estime l’importance en tant « carrefour » de ce nœud i

$$C_i^B = \sum \frac{\sigma_i^{s \to t}}{\sigma^{s \to t}}$$



LES PRINCIPAUX RÉSEAUX

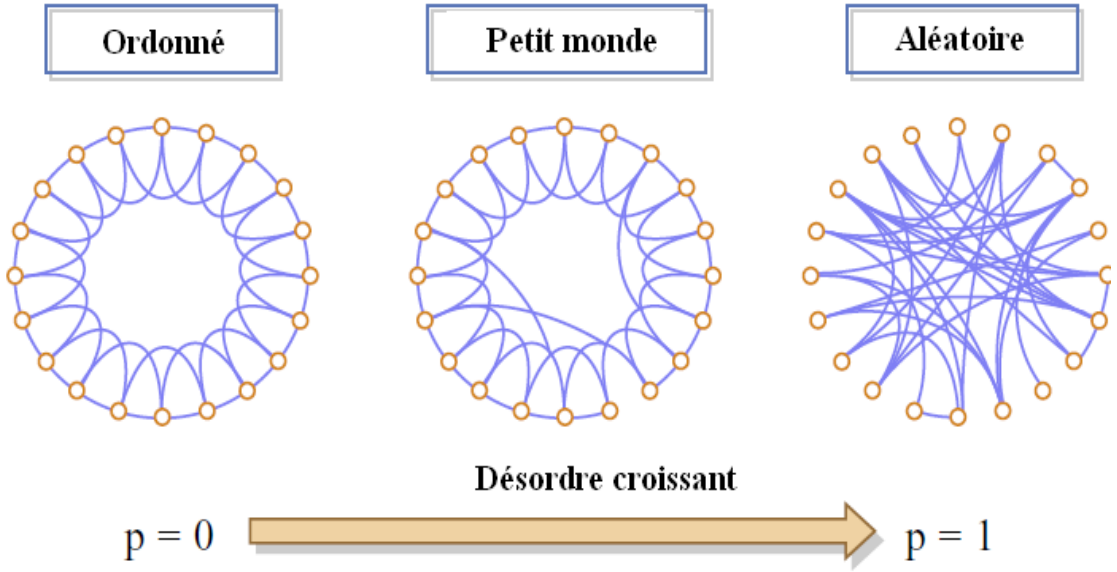
1. Les réseaux réguliers et maillés, lattice network, frozen

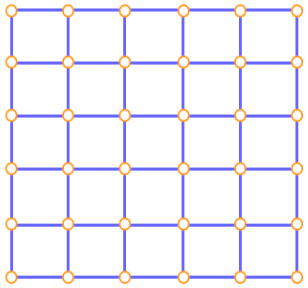
2. Les réseaux aléatoires et désordonnés (les plus intuitif, base de réflexion) random network

3. Les réseaux du petit monde, small world

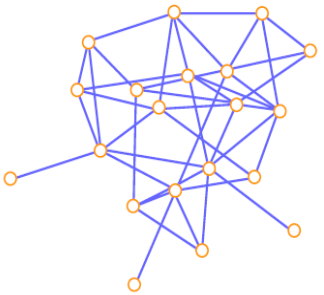
4. Les réseaux avec invariance d'échelle (auto-organisés, difficilement modèlisables), scale-free

LE PETIT MONDE : UN INTERMÉDIAIRE ENTRE L'ORDRE ET LE DÉSORDRE



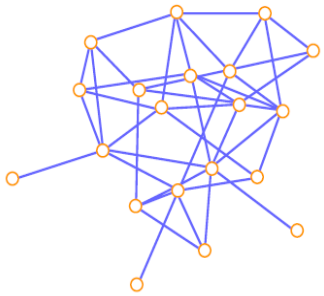


Courts circuits à longue distance



ALÉATOIRE (ERDÖS-RÉNYI)

1 Aléatoire : les principales caractéristiques :



Espérance du nombre d'arrêtes : $E(d) = P.\frac{N.(N-1)}{2}$

Distribution des degrés (loi de Poisson)

$$P(k) = \frac{\bar{k}^k}{k!}.e^{-\bar{k}} \quad \rightarrow \quad e^{k.Ln(\bar{k})}, \text{ si k grand}$$

Longueur moyenne

$$\bar{l} = \frac{Ln(n)}{Ln(\bar{k})}$$

Coefficient de "clustering" $C = \frac{\bar{k}}{n}$

Condition de percolation

$$P_c \approx \frac{Ln(n)}{n}$$

$$\bar{k} \approx Ln(n)$$

INVARIANCE D'ÉCHELLE : LA LOI EN PUISSANCE

2 Invariance d'échelle, principales caractéristiques :

Distribution des degrés $P(k) \approx k^{-\gamma}$

donc décroît plus lentement qu'une exponentielle

Longueur moyenne $\bar{l} \approx \frac{Ln(n)}{Ln[Ln(n)]}$

Coefficient de "clustering", relations empiriques

$C(k) \approx \frac{1}{k}$ *(pour des réseaux hiérarchiques)*

$C(k) \approx n^{3/4}$

ÉCHANTILLONNAGE

Démarche globale

1. Collecter les unités (nœuds ou arrêtes) avec leur information
2. Construire un réseau pour les représenter
3. Identifier les propriétés de ce réseau

PANORAMA

Une seule étape :

- Sélection de nœuds puis d'arrêtes (induced subgraph)
- Sélection d'arrêtes puis des nœuds (incident subgraph)

Plusieurs étapes (Link Tracing)

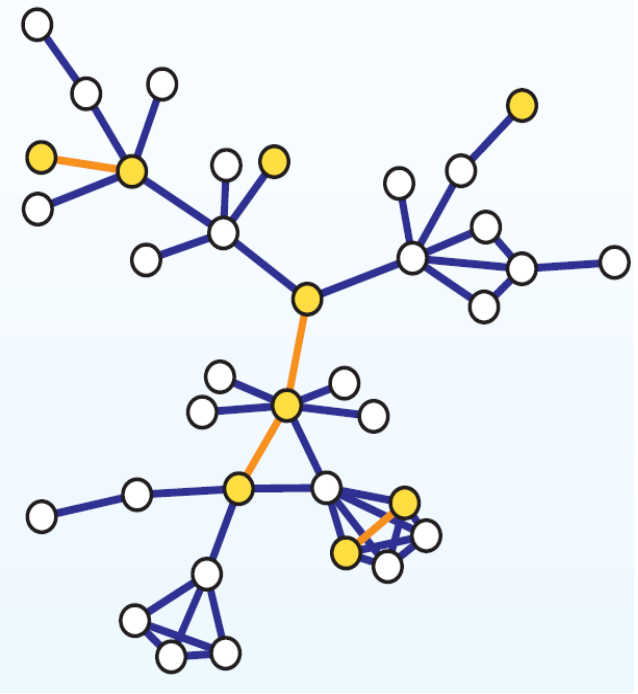
- Sélection exhaustive de TOUS les nœuds en interaction avec un nœud initial (Snowball)
- Sélection partielle dans l'ensemble des nœuds en interaction avec un nœud initial (Targeted)

ÉCHANTILLONNAGE EN SOUS-GRAPHE INDUIT (INDUCED SUBGRAPH)

- Sélection aléatoire de nœuds
- Ne considérer QUE les arrêtes joignant deux nœuds qui ont été sélectionnés

Alternative :

- Sélection aléatoire de nœuds
- Considérer TOUTES les arrêtes joignant les nœud sélectionnés



Probabilité d'inclusion :

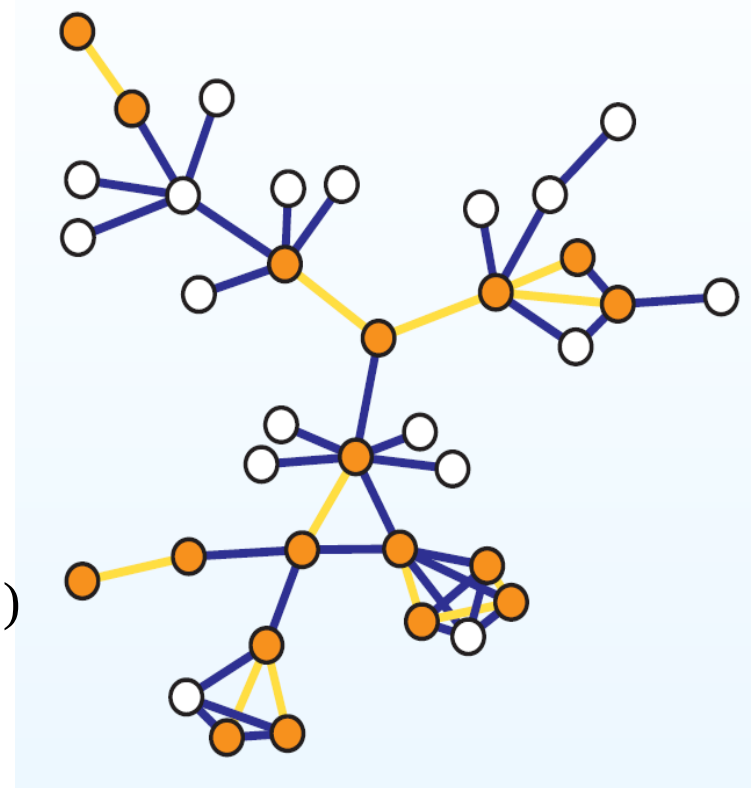
$$\pi_i = \frac{n}{N}$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}$$

Problème : il faut connaître N

ÉCHANTILLONNAGE EN SOUS-GRAPHE INCIDENT (INSCIDENT SUBGRAPH)

- Sélection aléatoire d'arrêtes
- Considérer TOUT les nœuds joignant les arrêtes sélectionnées



Probabilité d'inclusion :

$$\pi_i = \frac{n}{N_e}$$

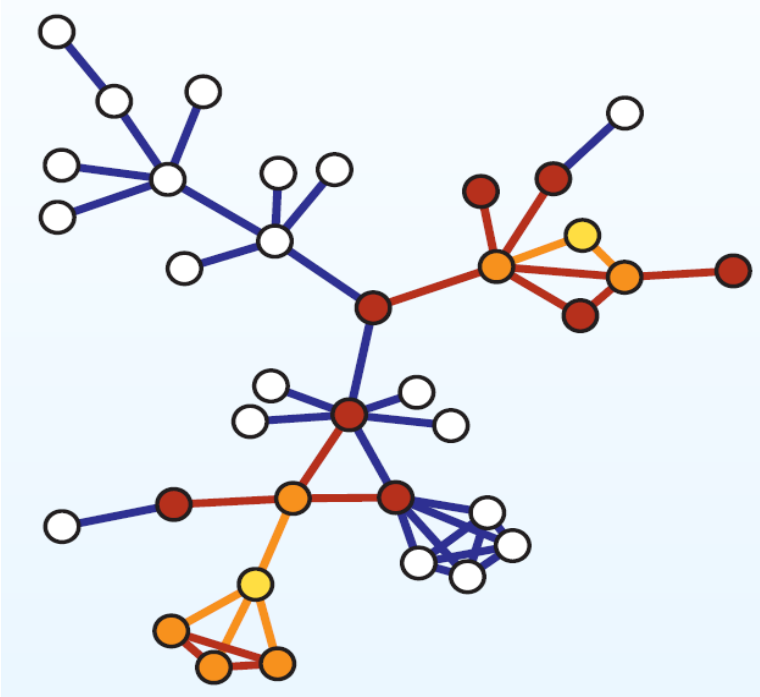
$$\pi_{ij} = 1 - \frac{C_n^{Ne-di}}{C_n^{Ne}}, \quad si\ n \leq N_e - d_i\ (sinon\ \pi_{ij} = 1)$$

ECHANTILLONNAGE EN « BOULE DE NEIGE » (SNOWBALL SAMPLING)

- Sélection aléatoire de n nœuds
- Pour chaque nœuds sélectionné y inclure tous les autres nœuds auquel il est directement connecté

Donc processus itératif (comme un plan progressif)
Chaque étape est appelé vague (wave)

Au-delà de la première vague, les calculs deviennent vite compliqués



LINK TRACING

L'échantillonnage se fait comme pour un échantillonnage en boule de neige, mais à chaque vague seuls une parie des nœuds est incluse à l'échantillonnage ce qui évite une avalanche de données

La probabilité d'inclusion est définie de la façon suivante :

$$\pi_i \approx 1 - (1 - \rho_s - \rho_t).e^{-\rho_s \rho_t b_i}$$

$$\pi_{ij} \approx 1 - e^{-\rho_s \rho_t . b_{ij}}$$

Dall'Asta (2006)

- On considère l'ensemble des chemine entre une nœud source s et un nœud cible t (target)
- Ou b_i est la betweeness du nœud i et b_{ij} celle de l'arrête (i,j)
- $\rho_s = n_s/N$ & $\rho_t = n_t/N$

DE LA RECHERCHE À L'INDUSTRIE

cea

INTRODUCTION AUX PRÉLÈVEMENTS

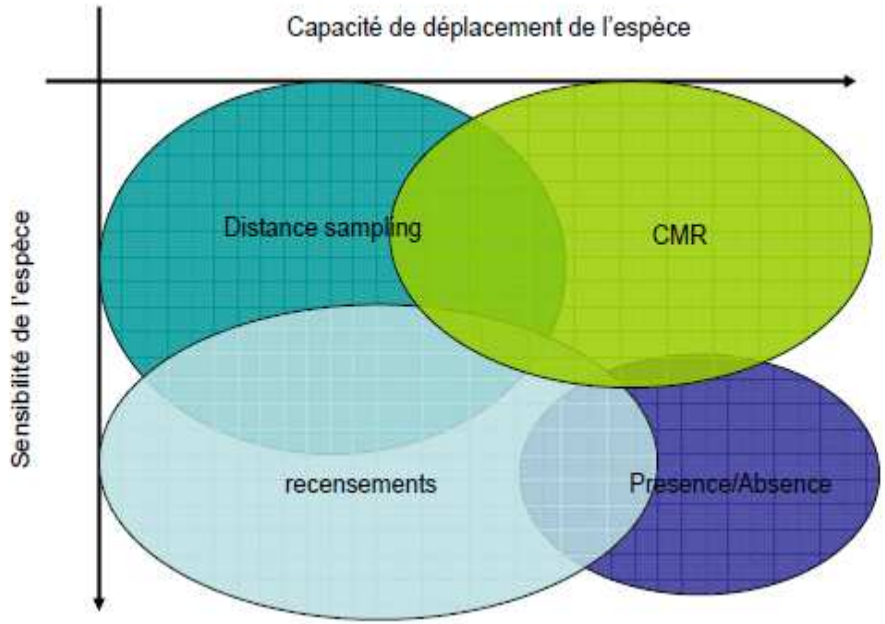
www.cea.fr
www-instn.cea.fr

instn

Les prélèvements sur de longues distances

- Echantillonnage distance (distance sampling)
- Echantillonnage capture marquage remarquage (CMR)

Yann.kergadallan@cea.fr


Capacité de déplacement de l'espèce
Sensibilité de l'espèce
Distance sampling
CMR
recensements
Presence/Absence

Structures et incertitudes

- Le variogramme
- Le plan composite ou mélange
- Méthode de Gy

Yann.kergadallan@cea.fr

PRINCIPE

Avantageux quand le cout de l'analyse est élevé,

1. Pour accéder rapidement à une estimation de la moyenne
   On peut montrer que la moyenne obtenue avec un plan composite est une estimation non biaisée de la moyenne du domaine à échantillonner

2. Pour identifier rapidement une caractéristique ou particulière ou une anomalie avec mesure systématique (ou rejet) des unités qui ont servies à constituer les lots → screening
   Se fixer alors un seuil pour les tests
   Se poser la question d'un retestage des unités initiales

Dans le cas 2 quand on reteste après repérage d'une caractéristique, on se rapproche des plans progressifs utilisés pour le contrôle statistique

Ce plan peut se coupler facilement avec un Rank Set Sampling

Yann.kergadallan@cea.fr

INTRODUCTION

Remarque : il est pas nécessaire dans ce plan d'avoir une sélection statistique des unités

Inconvénient : bonne précision (σ_{xmoy}) mais perte de σ_{exp}

C'est donc le contraire d'un plan systématique

Contraintes :

- Il faut que les échantillons soient physiquement « mélangeables » et il existe une erreur de préparation (mauvaise opération de mélange)
- La mesure doit être précise et les limites de détection faibles
  - Connaître les performances des appareils de mesure et la concentration maximale
  - Connaître le nombre et l'origine des unités à mélanger
  - Connaître le nombre de lots constitutifs

Exemples d'utilisation :

- Initialement : transmission virale par les insectes aux plantes
- Échantillon de sols (contrôle de dépollution)
- Échantillon de sang (cf l'affaire du sang contaminé)
- Les bancs de poissons

UTILISATIONS

Plan composite

Population cible à préciser

- Evaluation de la moyenne d'une population ou d'une strate pour une mesure continue
- Evaluation de la proportion de la population exprimant un caractère spécifique

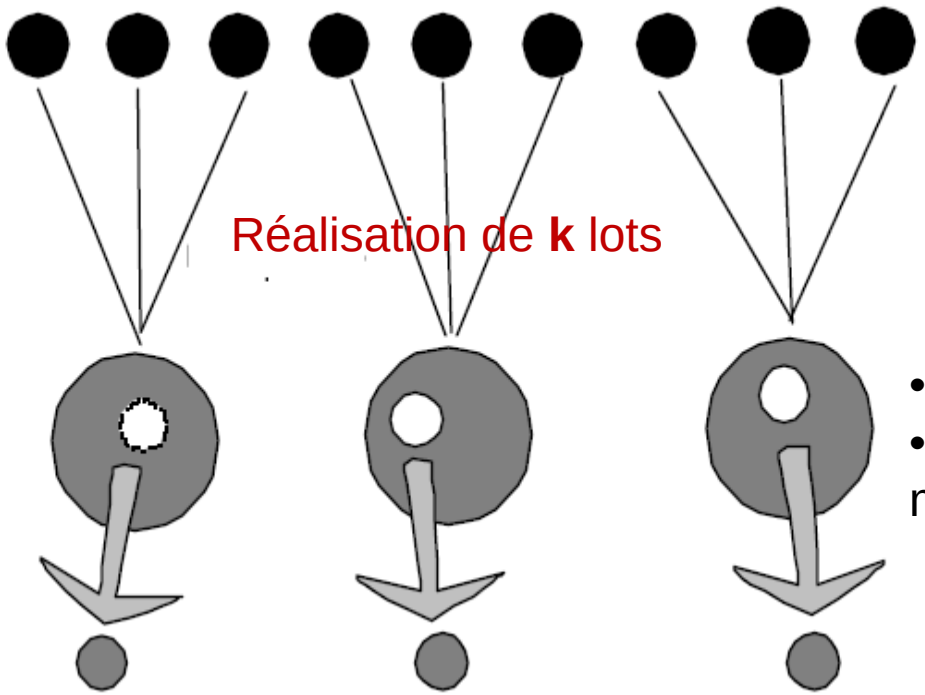
Prise de décision

- Identification des « hot-spots »
- Identification des échantillons ayant la valeur la plus élevée

Nécessite d'avoir suffisamment de matière pour refaire les analyse → elle ne doit pas être coûteuse

PRINCIPE : RÉALISATION D'UN E « MOYENNE » PHYSIQUE PAR MÉLANGE

Plan composite



- Réunion et mélange d'unités
- Chaque lot est constitué du même nombre d'unités

Sélection de **m** nouvelles unités pour analyse

QUAND UTILISER UN PLAN MÉLANGE ?

Plan composite

Critères	Plan mélange adapté si :
Le cout des analyses	Les couts des analyses sont élevés comparés à ceux des prélèvements
Incertitude analytique	L'incertitude de mesure est faible comparée à celle basée sur l'écart type expérimental
Estimation de la moyenne d'une population	L'information relative aux unités est négligeable L'information relative aux relations entre variables (corrélations …) est négligeable
Estimation de la proportion de la population présentant une caractéristique	Possibilité de détecter la caractéristique dans le mélange si l'une au moins des unités la possède La probabilité de ragée à tort les unités est faible La caractéristique est rare
Pour trier les échantillons présentant une caractéristique de ceux qui n'en présentent pas (screening)	Possibilité de détecter la caractéristique dans le mélange si l'une au moins des unités la possède La probabilité de ragée à tort les unités est faible La caractéristique est rare Possibilité de refaire une mesure sur l'une des unités initialement mélangées
Pour identifier l'échantillon qui présente la plus grande valeur	L'incertitude de mesure est faible Possibilité de refaire une mesure sur l'une des unités initialement mélangées
Pour identifier l'intervalle des valeurs	Les grandeurs considérées sont beaucoup plus élevées que la limite de détection
Empêchements physiques	Le mélange ne risque pas d'altérer la représentativité du prélèvement Les unités peuvent être mélangées sans problème

DÉTERMINATION DU NOMBRE K DE LOTS À RÉALISER

Plan composite

Prix unitaire de l'analyse : Cm
Prix unitaire du prélèvement : Cs
$\Rightarrow P_1 = Cm/Cs$

Evaluer autant que possible le rapport : $P_2 = \frac{\sigma_{\text{mesure+prélèvement}}}{\sigma_{\text{exp.dumilieu}}}$

Si les hypothèses suivantes sont respectées :

- La sélection est aléatoire
- La réalisation du mélange n'entraîne pas l'apparition d'erreurs notables

Le tableau qui suit permet de déterminer un nombre k approprié

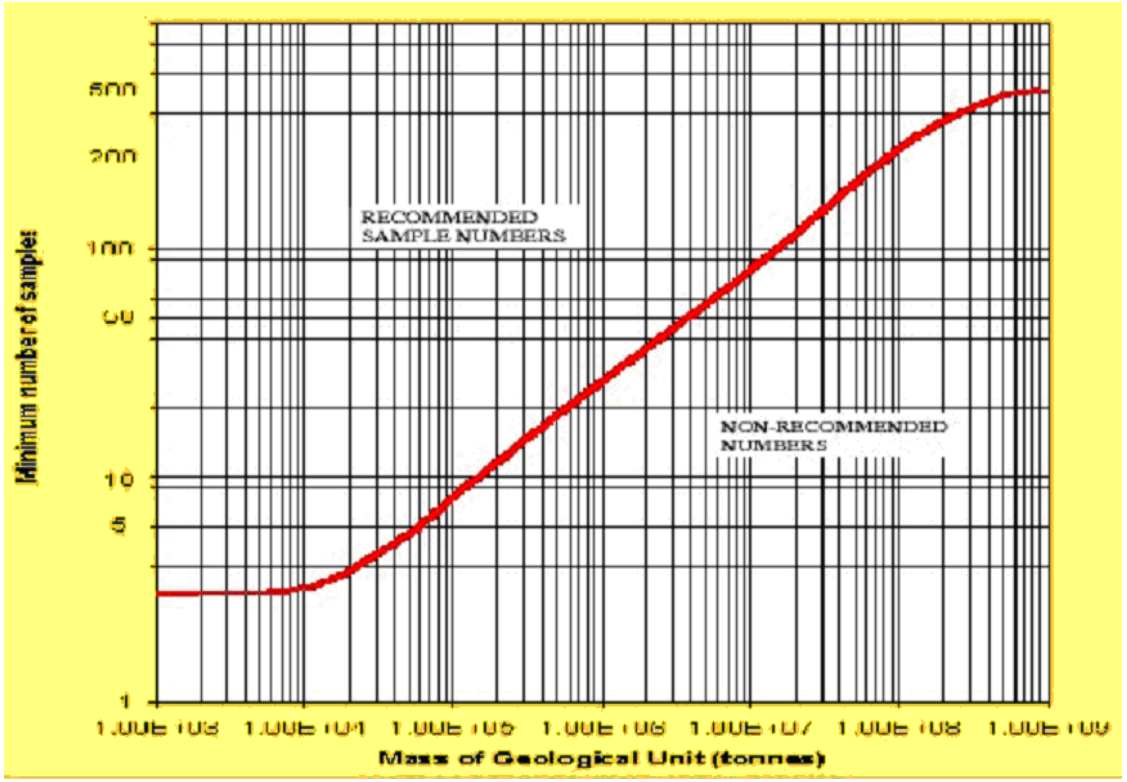
Déterminer la taille de la prise d'essai suivant

- La nature physique des unités
- La valeur attendue pour une caractéristique donnée

(Si le k du tableau est trop grand, prendre la plus grande possible)

EXEMPLE DE DÉTERMINATION DE LA QUANTITÉ DE PRISE D'ESSAI

Plan composite



DÉTERMINATION DU NOMBRE K DE LOTS À RÉALISER

Plan composite

Choix de la valeur de k (nombre d'unités à prévoir dans chaque groupe ou lot)

Rapport du cout C_M/C_S	Rapport P2									
	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**	**1.00**
2	14	7	5	4	3	2	2	2	2	1
3	17	9	6	4	3	3	2	2	2	2
4	20	10	7	5	4	3	3	3	2	2
5	22	11	7	6	4	4	3	3	2	2
8	28	14	9	7	6	5	4	4	3	3
10	32	16	11	8	6	5	5	4	4	3
15	39	19	13	10	8	6	6	5	4	4
20	45	22	15	11	9	7	6	6	5	4
50	71	35	24	18	14	12	10	9	8	7

DÉTERMINATION DU NOMBRE D'UNITÉS SECONDAIRES À PRÉLEVER

Plan composite

2. Le budget permet de déterminer le nombre de mesures prévues : n

   Comme k lots sont prévus, on en déduit le nombre m d'unités secondaires à prélever dans chaque lot

$$m = \frac{n}{k}$$

3. Vérifier que m est suffisamment grand pour fournir une bonne précision à la moyenne
   La précision est obtenue à partir de :

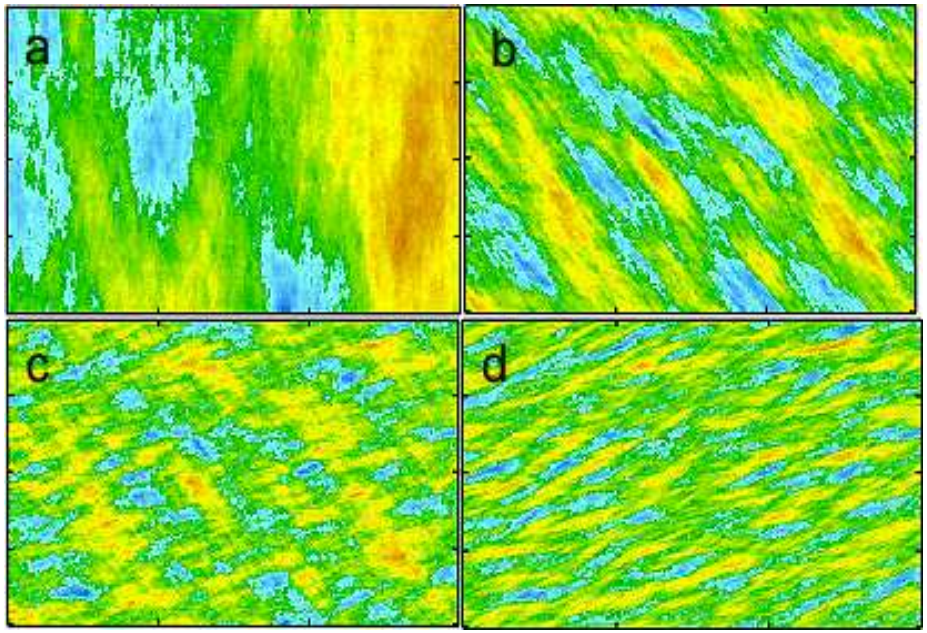
$$\mathrm{Var}(\overline{X}) = \mathrm{Var}(\overline{X}_{\substack{\text{sans mélange} \\ \text{avec chq unites}}}) + \frac{k-1}{m.k}.\mathrm{Var}(\text{mesure})$$

4. Réaliser le mélange en faisant des apports successifs de 4 à 5 fois pour constituer les lots

NÉCESSITÉ D'UNE STRUCTURE DE LA VARIABILITÉ

Variogramme

Même moyenne et variance, mais avec des structures différentes :



•Étude de la variabilité avec l'éloignement, c'est-à-dire quand on quitte un voisinage

•Les hypothèses de stationnarité : cette variabilité ne dépend que de l'éloignement et pas de l'endroit où se trouve ce voisinage

VARIABLES RÉGIONALISÉES

Variable régionalisée : variable associée à une position

•On considère un domaine à échantillonner

En général c'est un plan systématique qui est réalisé

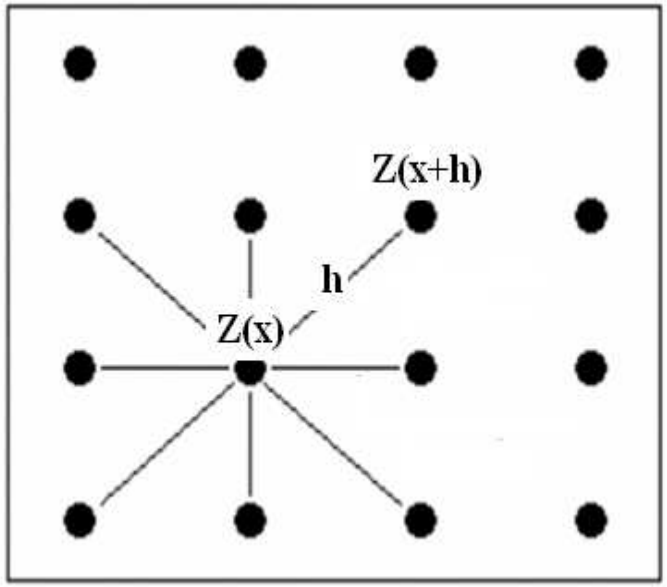
•Chaque individu est repéré par ses coordonnées (x, y)

•La valeur prise par l'individu en (x, y) est une variable aléatoire z(x,y)

On parlera de variables régionalisées

On a donc un champ de variables aléatoires

DISTANCE



Le variogramme est cette fonction $\gamma(h)$ qui donne la variabilité avec l'éloignement

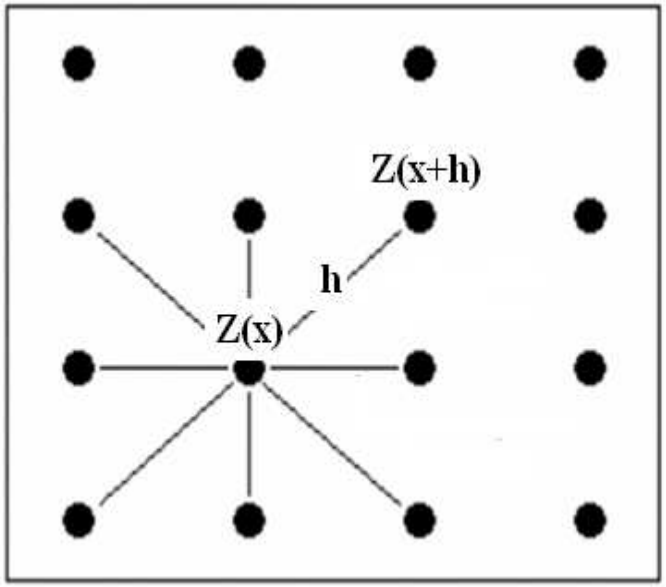
Intuitivement : $\gamma(h)$ est faible pour h faible, puis devient importante et peut se stabiliser

Conditions de stationnarité :

1. E[z(x+h)-z(x)] = 0, l'espérance est invariante par translation,
   L'espérance est quasi la m^me pour deux points peu éloignés
2. Var[z(x+h)-z(x)] = 2.$\gamma(h)$

Seul l'éloignement h intervient, la position absolue (x, y) n'intervient pas dans l'analyse des variogrammes

CALCUL DE γ(H)



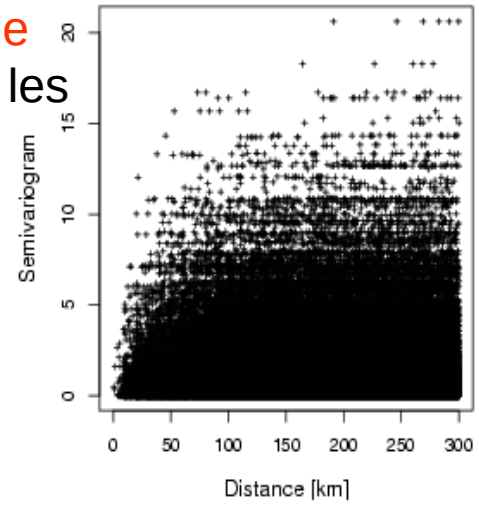
$$\gamma(h) = \frac{1}{N(h)} \cdot \sum_{1}^{N(h)} [z(x_i + h) - z(x_i)]^2$$

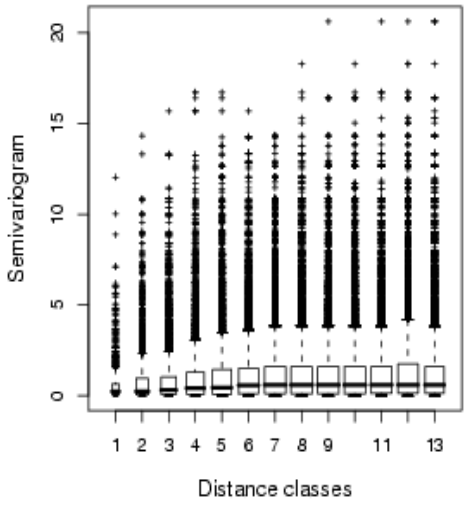
Attention N(h) est le nombre de pairs de points qui ont la même distance h
Ce n'est pas le nombre de voisins !

Ici N(h) vaut à peu prés 8

CONSTRUCTION D'UN VARIOGRAMME

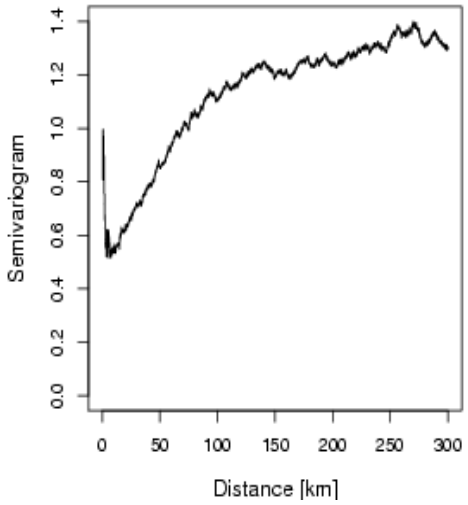
Nuée variagraphique pour détecter les valeurs aberrantes

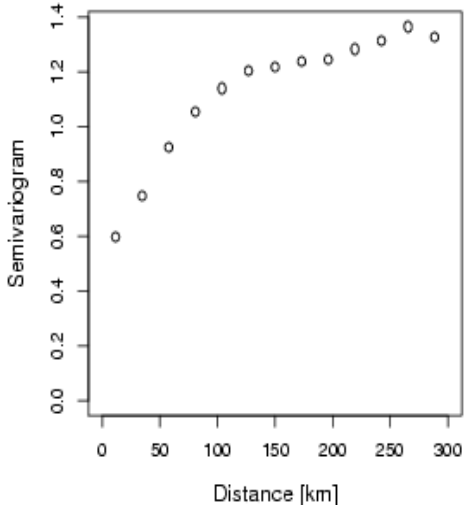




Découpage avec un pas de 1 et une tolérance (résolution) de 1

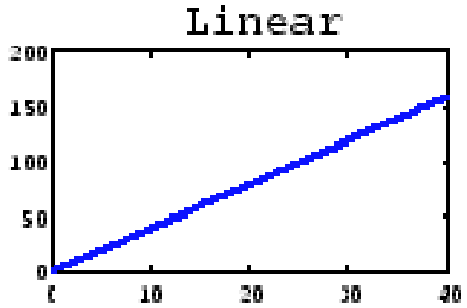
Moyenne de la variabilité = variogramme expérimental

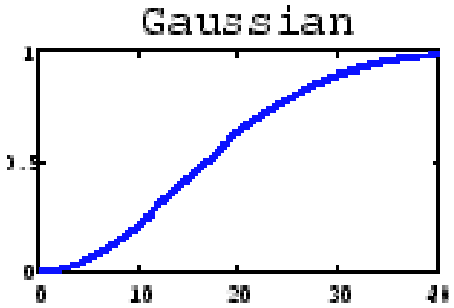


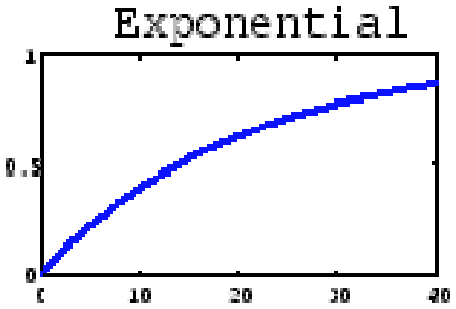


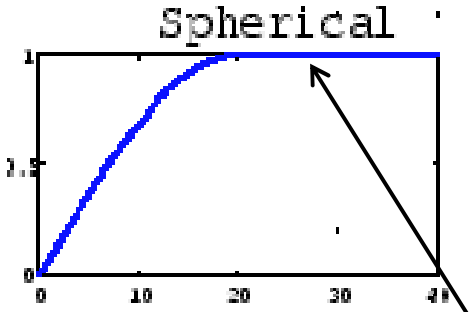
EXEMPLE DE VARIOGRAMME AVEC LEUR MODÈLISATION

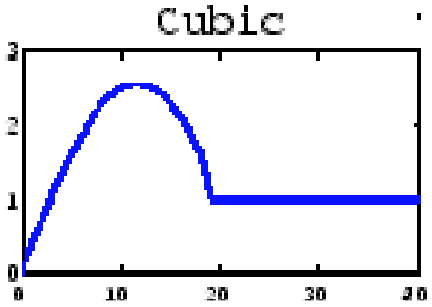
Variogramme







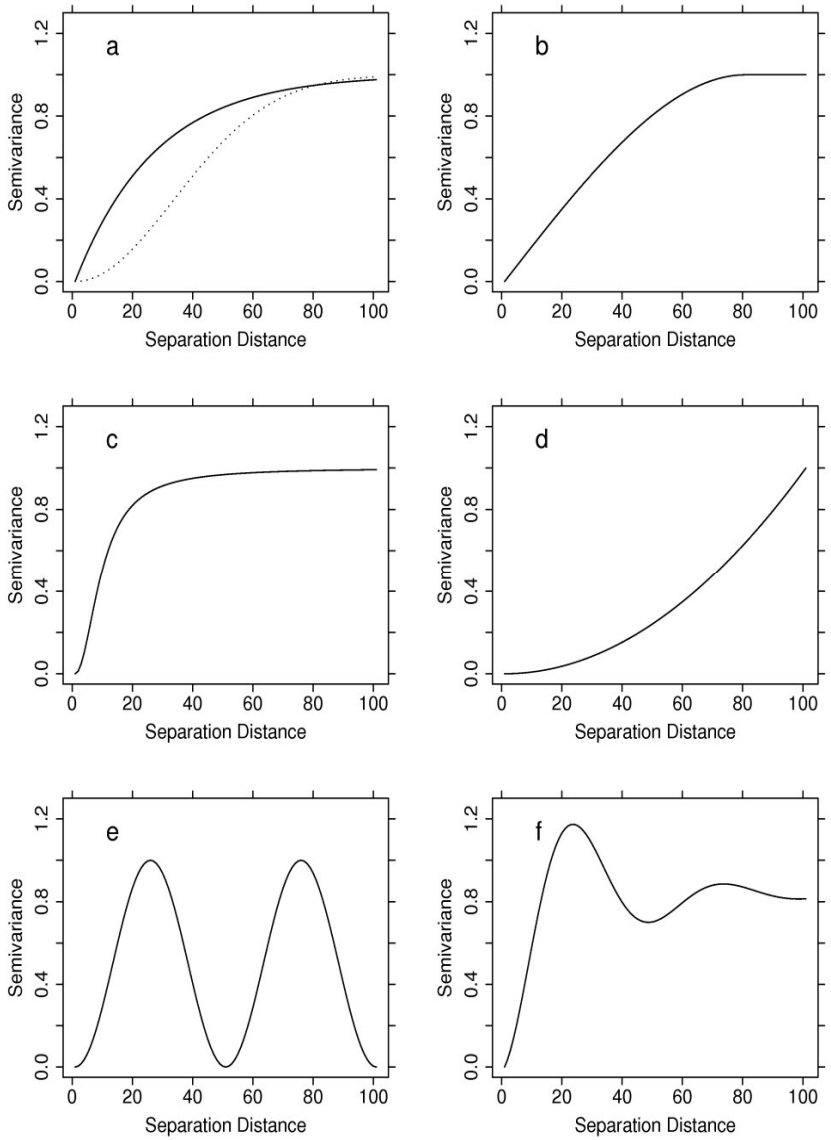




Palier ; fin du domaine d'influence

KRIGEAGE

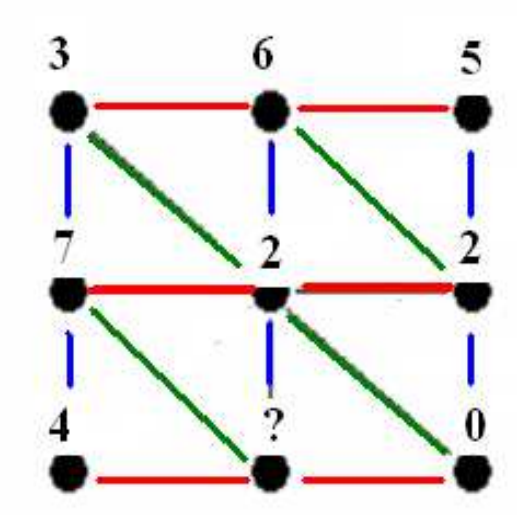
Variogramme



a) Modèle exponentiel (équivallent au modèle Gaussien)
b) Modèle sphèrique
c) Modèle quadratique
d) Modèle puissance
e) Modèle cosinus

L'identification du modèle permet ensuite de faire des extrapolations aux autres points puis une cartographie

EXEMPLE DE CALCULS



$$\gamma_0(1) = \frac{1}{2.4}.[(3-6)^2+(6-5)^2+(7-2)^2+(2-2)^2]$$

$$\gamma_{90°}(1) = \frac{1}{2.5}.[(3-7)^2+(7-4)^2+(6-2)^2+(5-2)^2+(2-0)^2]$$

$$\gamma_{45°}(1) = \frac{1}{2.3}.[(6-2)^2+(3-2)^2+(2-0)^2]$$

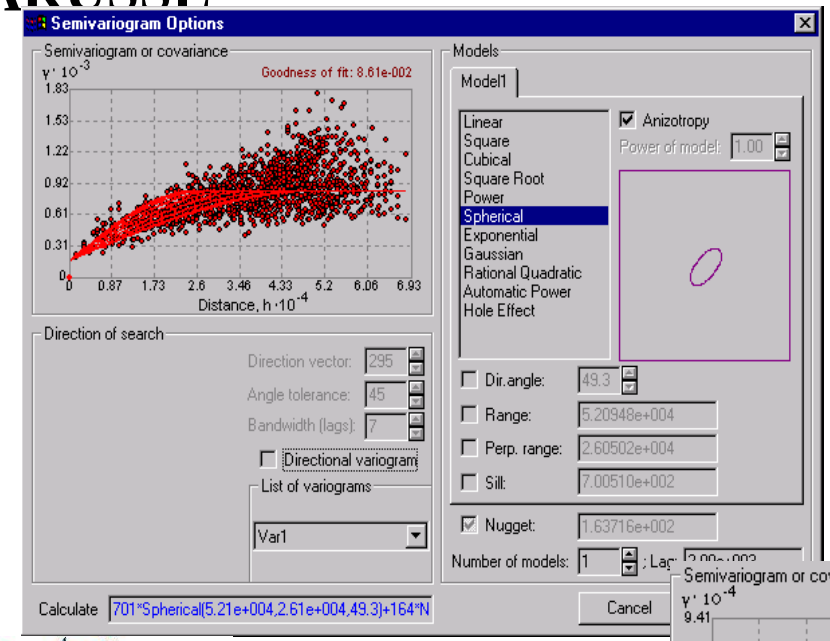
$$\gamma_{0°}(2) = \frac{1}{2.3}.[(3-5)^2+(7-2)^2+(4-0)^{22}]$$

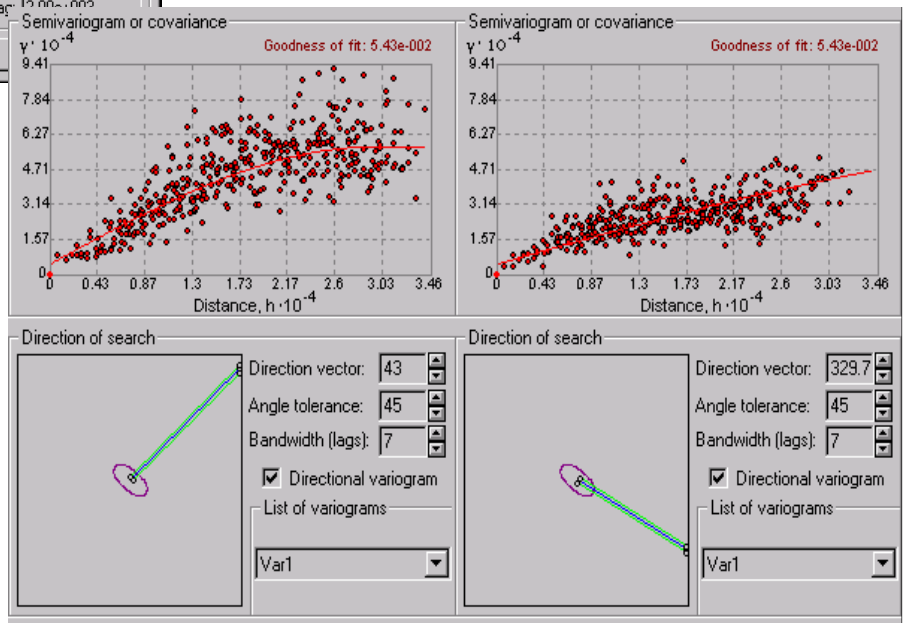
$$\gamma_{90°}(2) = \frac{1}{2.2}.[(3-4)^2+(5-0)^2]$$

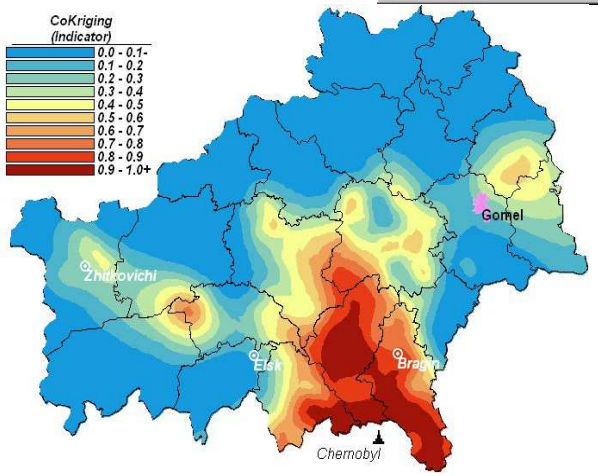
$$\gamma_{45°}(2) = \frac{1}{2.1}.[(3-0)^2]$$

VARIOGRAMME DU ^{137}CS DANS DEUX RÉGIONS BELARUSSE

Variogramme







Contamination au Pu