



24/07/2024

STATISTIQUES MARITIMES & PORTUAIRE

Dr KOUAO
ARSTM

STATISTIQUES MARITIMES & PORTUAIRES

CHAPITRE 0 : INTRODUCTION

I. DEFINITIONS ET HISTOIRE

➤ Définitions

« Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation » (Encyclopedia Universalis)

➤ Histoire

- Recensements en Chine au XXIII^e siècle av. J.-C. ou en Égypte au XVIII^e av. J.-C, système de recueil se poursuivant jusqu'au XVII^e .
- Rôle prévisionnel des statistiques au XVIII^e siècle avec la construction des premières tables de mortalité avec Antoine Deparcieux, l'Essai sur les probabilités de la durée de vie humaine (1746).
- Rôle démographique au XIX^e siècle, le Baron de Reiffenberg présentait en 1842 à l'Académie ses calculs rétrospectifs de population chez des peuples gaulois, d'après des chiffres donnés par Jules César dans sa conquête des gaules.
- Premiers textes connus sur le calcul des hasards (ou des chances) au XVI^e siècle avec Cardan et au XVII^e siècle avec Galilée.
- Début officiel avec Pascal, Fermat et Huyguens au XVII^e siècle.
- Tournant au XVIII^e siècle avec Montmort (combinatoire), Bernoulli (loi des grands nombres) puis De Moivre et Laplace (traitement analytique des probabilités et théorèmes limites).
- Théorie des ensembles et de la mesure par Borel et Lebesgue et calcul des probabilités par Lévy au XX^e siècle
- Axiomatisation de la théorie des probabilités par Kolmogorov (1933).

II. DOMAINES D'APPLICATION ET BUT DE LA STATISTIQUE

➤ Domaine d'application

- Economie, assurance, finance : études quantitatives de marchés, prévisions économétriques, analyse de la consommation des ménages, taxation des primes d'assurances et de franchises, gestion de portefeuille, évaluation d'actifs financiers, ...
- Biologie, médecine : essais thérapeutiques, épidémiologie, dynamique des populations, analyse du génôme, ...
- Sciences de la terre : prévisions météorologiques, exploration pétrolière, ...
- Sciences humaines : enquêtes d'opinion, sondages, étude de population, ...
- Sciences de l'ingénieur : contrôle qualité, sûreté de fonctionnement, évaluation des performances, ...
- Sciences de l'information : traitement des images et des signaux, reconnaissance de forme et parole, machine learning, ...

➤ But de la statistique

Les données sont entachées d'incertitudes et présentent des variations pour plusieurs raisons :

- le déroulement des phénomènes observés n'est pas prévisible à l'avance avec certitude

- toute mesure est entâchée d'erreur
- seuls quelques individus sont observés
- ...

⇒ données issues de phénomènes aléatoires

⇒ Intervention du hasard et des probabilités

Objectifs : maîtriser au mieux cette incertitude pour extraire des informations utiles des données, par l'intermédiaire de l'analyse des variations dans les observations.

III. CLASSES DE METHODES STATISTIQUES ET DEMARCHES

➤ 2 classes de méthodes statistiques

📊 Statistique descriptive : elle a pour but de résumer l'information contenue dans les données de façon synthétique et efficace par :

- Représentations graphiques
- Indicateurs de position, de dispersion et de relation
- Régression linéaire

⇒ permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus poussée. Les probabilités n'ont ici qu'un rôle mineur.

📊 Statistique inférentielle : elle a pour but de faire des prévisions et de prendre des décisions au vu des observations par :

- Estimation paramétrique
- Intervalles de confiance, tests d'hypothèse

⇒ Nécessite de définir des modèles probabilistes du phénomène aléatoire et savoir gérer les risques d'erreurs.

La statistique repose sur l'observation de phénomènes concrets et utilise les probabilités comme outils d'analyse et de généralisation. La théorie des probabilités permet de modéliser efficacement certains phénomènes aléatoires et d'en faire l'étude théorique.

➤ Démarche statistique

1 **Recueil des données** ⇒ construction d'un échantillon

2 **Statistique exploratoire** ⇒ formulation d'hypothèses sur la nature du phénomène aléatoires

3 **Choix d'un modèle probabiliste** ⇒ test d'adéquation

4 **Estimation des paramètres inconnus du modèle** ⇒ construction d'estimateurs

5 **Prévision sur les observations futures** ⇒ associer un degré de confiance

➤ Vocabulaire

- Faire de la statistique suppose que l'on étudie un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées variables.

- Le groupe ou l'ensemble d'objets équivalents est appelé la population.
- Les objets sont appelés des individus.
- En général, la population est trop vaste pour pouvoir être observée exhaustivement. On étudie alors la variable sur une sous partie de la population. On étudie alors un échantillon.

On souhaite étudier un caractère X prenant ses valeurs dans Ω , sur une population P .

Exemple : si l'échantillon est un groupe de ports de l'AGPAOC,

- un individu est un port de l'Afrique de l'ouest et du Centre
- la population peut être l'ensemble des ports de la côte ouest et centre africaine
- les variables étudiées peuvent être le volume de cargaisons manutentionnées, la taille des navires embarqués et débarqués, l'âge des équipements de quai, le nombre d'escales, etc.

En général, on ne peut pas observer ce caractère sur tous les individus d'une grande population, mais seulement sur une sous-population de P de taille n . On notera alors :

- la sous population : $\{i_1, \dots, i_j, \dots, i_n\}$ un ensemble de n individus choisis au hasard dans P .
- l'échantillon de données : $x_1, \dots, x_j, \dots, x_n$ les n valeurs observées du caractère X sur les individus de la sous-population.

Deux problèmes se posent alors :

1. Quelles informations sur le caractère X peut-on tirer de l'échantillon ?
2. Quelle prévision pourrait-on faire sur un individu non observé de P à partir des données observées $x_1, \dots, x_j, \dots, x_n$?

Chaque individu est décrit par un ensemble de variables X . Ces variables peuvent être classées selon leur nature :

- variable qualitative s'exprimant par l'appartenance à une modalité. $\Omega = \{\text{conventionnel, Conteneur}\}$; $\Omega = \{\text{minerais, produits pétroliers, céréales, produits de la pêche, etc.}\}$
- variable quantitative, s'exprimant par des nombres réels, par exemple la taille des navires ou le volume de marchandises.
- On distingue les variables quantitatives discrètes lorsque Ω est une suite finie ou infinie d'éléments de $\mathbb{N}$ (ex : $\Omega = \{1, 2, 3\}$; $\Omega = \mathbb{N}$) des variables quantitatives continues si toutes les valeurs d'un intervalle de $\mathbb{R}$ sont acceptables. (Ex: Nombre de grue sur le terminal)

Concept clé en statistique : la variabilité, qui signifie que des individus en apparence semblables peuvent prendre des valeurs différentes.

Exemple : Un processus industriel de fabrication d'accessoires de manutention ne fournit jamais des caractéristiques parfaitement constantes? encore moins, le tonnage quotidiennement enregistré présentera des quantités de marchandises différentes.

L'analyse statistique a pour but d'étudier cette variabilité

- en tenir compte pour prévoir de façon probabiliste le comportement d'individus non observés,
- chercher à l'expliquer à l'aide de variables extérieures,
- chercher à l'augmenter dans le but de distinguer les individus entre eux.

CHAPITRE I : NOTIONS DE BASE DE STATISTIQUE

I. ANALYSE UNIVARIÉE

1. Exemples de tableaux statistiques et subtilités des statistiques

1.1 Exemples de tableaux statistiques

Tableaux statistiques - variables discrètes ou quantitatives

- n la taille de l'échantillon
- k le nombre de modalités.
- $m_i, i \in [1, k]$ les modalités
- n_i le nombre d'occurrence (l'**effectif**) de m_i dans l'échantillon
- et f_i la **fréquence** correspondante.

on a $\sum n_i = n$; $f_i = n_i/n$; $\sum f_i = 1$.

Exemple 1 du nombre mensuel d'accidents intervenus sur un terminal donné (Variables discrètes quantitatives)

$$x_i = \{2, 5, 6, 4, 5, 4, 2, 1, 6, 5, 1, 2\}, n = 12$$

m_i	1	2	3	4	5	6
n_i	2	3	0	2	3	2

. Vérification empirique qu'un accident est de l'ordre de l'aléa?

. Simulation informatique d'une loi uniforme?

Exemple 2 de la répartition des $n=42$ pays selon la variable X , nombre de ports du pays (Variables discrètes quantitatives)

Modalité m_i (nombre de ports)	Effectif n_i	Fréquence f_i (en pourcentage indicatif)
1	13	31.0
2	13	31.1
3	7	16.2
4	6	13.8
5	2	5.5
6 et plus	1	2.4

Exemple 3 de la répartition des $n = 14381440$ T de marchandises de moins de 20 semaines, selon la durée de stationnement (classes d'âge) dans les entrepôts d'un port donné (Variables continues)

On regroupe les valeurs en k classes d'extrémités $a_0, a_1, \dots, a_k$, et on note pour chaque classe $[a_{i-1}, a_i]$ l'effectif n_i , la fréquence f_i .

modalité classes d'âge	effectif n_i	fréquence f_i en pourcent
[0,4]	2 986 925	20.77
[5,9]	3 629 294	25.24
[10,14]	3 833 120	26.65
[15,19]	3 932 101	27.34

Exemple 4 de la répartition des pays selon le nombre de ports d'un échantillon de pays; et fonction de répartition empirique (fréquences cumulées)

Modalité X_i	Fréquence f_i (en pourcentage)	Fréquence F_i (en pourcentage)
1	31.0	31.0
2	31.1	62.1
3	16.2	78.3
4	13.8	92.1
5	5.5	97.6
6 et plus	2.4	100

Exemple 5 de la répartition des chargeurs (importateurs/exportateurs) par sexe et par catégorie professionnelle (tableau de contingence)

	femmes	hommes	total
agriculteurs exploitant	204 209	437 958	1.3 %
artisans, commerçants	484 443	1 174 609	3.4 %
cadres et professions intel. sup.	1 101 537	2 063 798	6.6 %
professions intermédiaires	2 771 948	2 990 937	11 %
employés	5 973 956	1 835 135	16.2 %
ouvriers	1 426 472	5 635 270	15.8 %
retraités	5 434 200	5 200 243	22.1 %
autres sans activité prof.	7 593 554	3 740 108	23.6 %
total	52 %	48 %	48 068 377

1.2 Subtilités des statistiques

- Variations relatives

Le pouvoir d'achat des importateurs a diminué de 12% en 2017 mais remontera de 12% en 2018. Quel sera leur pouvoir d'achat en 2018 :

A) Plus important qu'en 2017

B) Identique à 2017

C) Plus faible qu'en 2017

D) La réponse D

Une quantité x subissant une diminution de $p\%$ puis une augmentation de $p\%$ s'écrit

$$x \times (1 - p) \times (1 + p) = x \times (1 - p^2) \leq x$$

⇒ Pour $p = 12\%$ on obtient une baisse d'environ 1,5%.

- Variations relatives et absolues

Le déficit de fret d'un port donné, qui avait augmenté de 15% l'an passé, n'a augmenté cette année que de 14%. L'autorité portuaire se félicite de sa gestion exemplaire.

NON: Augmentation du déficit

2. Représentations graphiques

Différents types de représentations graphiques :

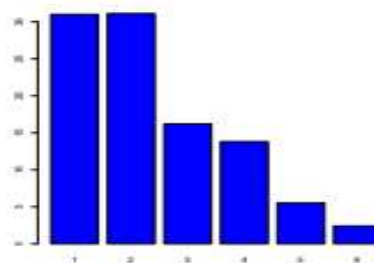
- Diagramme en bâtons et en camembert
- Histogramme des fréquences
- Graphique des fréquences cumulées (= fonction de répartition)
- Boîte à moustache

2.1 Variable discrète

Représentations graphiques

Variables discrètes – Diagrammes en bâtons

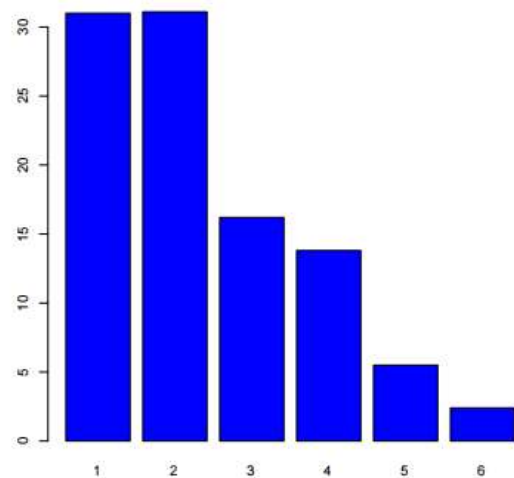
modalité m_i nb pers.	fréquence f_i (en %)
1	31.0
2	31.1
3	16.2
4	13.8
5	5.5
6 et plus	2.4



Diagrammes en bâtons : à chaque modalité correspond un rectangle vertical dont la hauteur est proportionnelle à la fréquence relative de la modalité.

Variables discrètes – Diagrammes en bâtons

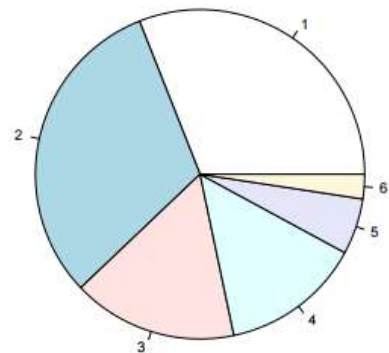
modalité m_i nb pers.	fréquence f_i (en %)
1	31.0
2	31.1
3	16.2
4	13.8
5	5.5
6 et plus	2.4



Diagrammes en bâtons : à chaque modalité correspond un rectangle vertical dont la hauteur est proportionnelle à la fréquence relative de la modalité.

Variables discrètes – Diagrammes sectoriels (ou en camemberts)

modalité m_i nb pers.	fréquence f_i (en %)
1	31.0
2	31.1
3	16.2
4	13.8
5	5.5
6 et plus	2.4



Diagrammes sectoriels (ou en camemberts) : à chaque modalité correspond un secteur de disque dont l'aire est proportionnelle à la fréquence relative de la modalité.

2.2 Variable continue

Histogramme

Variable continue

Quand la variable étudiée est **continue**, les représentations du type **diagramme en bâtons** sont **sans intérêt**, car les données de x sont en général toutes distinctes, donc les effectifs tous égaux à 1.

⇒ La représentation par **histogramme** consiste à **regrouper les observations « proches » en classes** :

On trie le vecteur x (noté alors x^*), et on partitionne l'intervalle $]a_0, a_k]$ ($a_0 < x_1^*, a_k > x_n^*$) en k intervalles $]a_{i-1}, a_i]$ appelés **classes**. La largeur de la classe i est notée $h_i = a_i - a_{i-1}$ (et $h = (a_k - a_0)/k$ si pas fixe).

Définition de l'histogramme

L'histogramme est la figure constituée de rectangles dont les bases sont les classes et dont les aires sont égales aux fréquences de ces classes.

Autrement dit, la hauteur du $i^{\text{ème}}$ rectangle est n_i/nh_i .

Exemple de la durée d'escale en heures de navires d'une journée donnée dans un port X

$x = 91.6, 35.7, 251.3, 24.3, 5.4, 67.3, 170.9, 9.5, 118.4, 57.1$

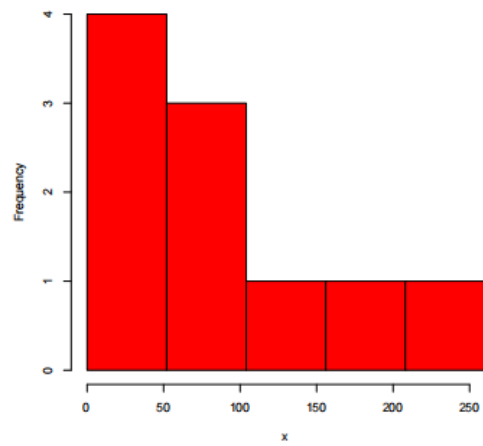
$x^* = 5.4, 9.5, 24.3, 35.7, 57.1, 67.3, 91.6, 118.4, 170.9, 251.3$

- **Choix du nombre de classes** k : $k \approx 1 + \log_2 n$ (règle de Sturges)
- **Choix des bornes** pour a_0 et a_k : $x_1^* \pm 0.025(x_n^* - x_1^*)$
- **Largeur des classes** (fixe) $h = (a_k - a_0)/k$

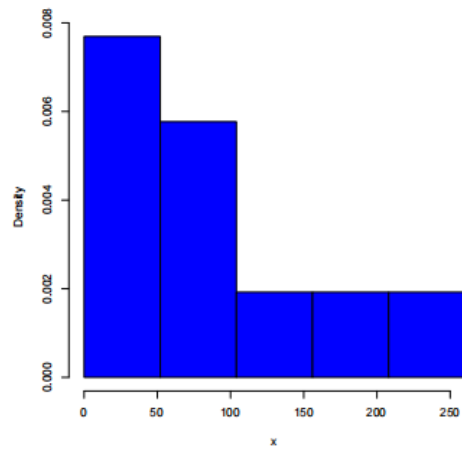
$n = 10, k = 5, a_0 = -0.74 \approx 0$ et $a_k = 257.4 \approx 260, h = 260/5 = 52$.

classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019

classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019

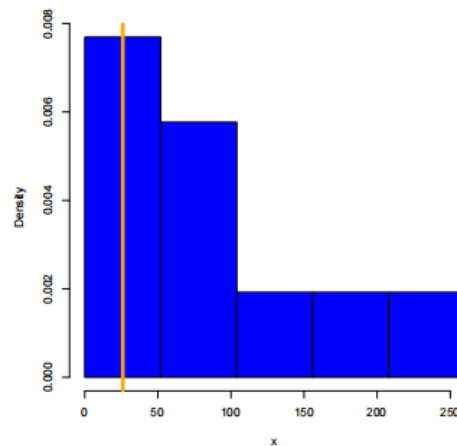


classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019



Mode de l'histogramme

classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019



3. Indicateurs Statistiques

3.1 Indicateurs de localisation ou de tendance centrale

Indicateurs de localisation (ou de tendance centrale)

La moyenne empirique

Définir une valeur autour de laquelle se repartissent les observations

◆ Moyenne empirique

Valeur qu'auraient tous les individus s'ils prenaient la même valeur

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^k n_i m_i$$

Durée moyennes des escales: $X=83,15$ heures

Valeurs extrêmes

• Valeurs extrêmes

Un indicateur de localisation à partir de $x_1^* = \min x_i$ et $x_n^* = \max x_i$ est

$$\frac{x_1^* + x_n^*}{2}$$

Exemple de la durée des escales: On trouve $\bar{X}=128,35$ heures

• Mode

Valeur pour laquelle l'histogramme des fréquences présente un maximum. Modalité la plus représentée dans l'échantillon.

Valeurs aberrantes

Des valeurs exagérément grandes ou petites par rapport aux autres valeurs de l'échantillon peuvent fortement influencer sur **la moyenne qui est sensible aux extrêmes**.

Remarque: Il existe d'autres indicateurs comme la médiane et les quantiles empiriques

3.2 Indicateurs de dispersion (ou de variabilité)

• Variance empirique

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{1}{n} \sum_{i=1}^k n_i (m_i - \bar{x}_n)^2$$

- $\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_n^2$ (moyenne du carré - carré de la moyenne)

- **Écart type empirique** : σ_x (racine de la variance)

- **Coefficient de variation empirique** : $cv_n = \frac{\sigma_x}{\bar{x}_n}$ (sans dimension)

- Dans R $\text{var}(x)$ donne $\sigma_x'^2 = \frac{n}{n-1} \sigma_x^2$ (variance sans biais)

- **Écart type empirique** : σ_x (racine de la variance)
L'écart type sert à mesurer la dispersion d'un ensemble de données.
Plus il est faible, plus les valeurs sont regroupées autour de la moyenne.

3.3 Caractérisation des Indicateurs

L'erreur commise en résumant l'observation x_i par c peut être quantifiée par une distance (ou écart) entre ces deux valeurs $d(x_i, c)$.

Un bon indicateur doit minimiser l'erreur moyenne $e = \frac{1}{n} \sum_{i=1}^n d(x_i, c)$.

- **Écart quadratique** : $e = \frac{1}{n} \sum_{i=1}^n (x_i - c)^2$, minimal quand

$$\frac{\partial e}{\partial c} = 0 \Leftrightarrow -\frac{2}{n} \sum_{i=1}^n (x_i - c) = 0 \Leftrightarrow c = \bar{x}_n$$

- **Écart absolu** : $e = \frac{1}{n} \sum_{i=1}^n |x_i - c|$, minimal quand $c = \tilde{x}_n$
- **Écart sup** : $e = \frac{1}{n} \sup_{i=1}^n |x_i - c|$, minimal quand $c = (x_1^* + x_n^*)/2$

II. ANALYSE BIVARIÉE

1. Corrélation et causalité: Principes

Rappels : indices de localisation, dispersion, relation

Pour un nuage de points (x_i, y_i) , $\forall i \in \{1, \dots, n\}$ on définit :

- Les moyennes empiriques (localisation)

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$$

- Les variances empiriques (dispersion)

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

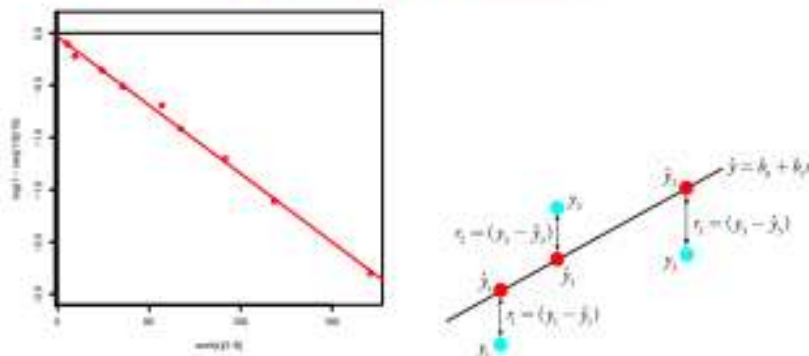
$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}_n^2$$

- La covariance et corrélation empirique entre les x_i et y_i (relation)

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}_n \bar{y}_n, \quad r_{xy} = \frac{c_{xy}}{s_x s_y}$$

Régression linéaire

Le but est de trouver la droite « la plus proche » d'un nuage de points



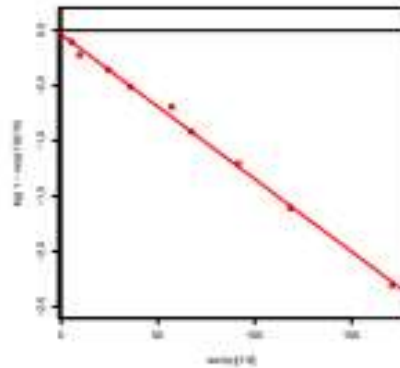
L'idée consiste à estimer y_i par $\hat{y}_i = \beta_1 x_i + \beta_0$ en choisissant β_1 et β_0 qui minimise l'erreur quadratique moyenne :

$$\delta^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0)^2$$

Régression linéaire

La droite $y = \hat{\beta}_1 x + \hat{\beta}_0$ qui minimise l'erreur quadratique moyenne :

$$\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2}, \quad \hat{\beta}_0 = \bar{y}_n - \bar{x}_n \frac{\sigma_{xy}}{\sigma_x^2}$$



La régression linéaire sur le graphe de probabilité fournit :

Crédits : O. Gaudoin $\hat{\beta}_1 = -0.01311$, $\hat{\beta}_0 = -0.03484$

Cum hoc ergo propter hoc

Corrélation

Deux événements (appelons-les X et Y) sont corrélés si l'on observe une dépendance, une relation entre les deux. Par exemple, le Volume de conteneurs débarqués sur un terminal a tendance à augmenter avec les escales: Escale et Trafic sont donc corrélés

Corrélation ou causalité ?

Une erreur de raisonnement courante consiste à dire : « X et Y sont corrélés, donc X cause Y ». On confond alors corrélation et causalité car en réalité, il se pourrait aussi que :

- Y cause X
- X et Y aient une cause commune Z
- X et Y soient accidentellement liés mais n'aient aucun lien de causalité.

CQFR : Ce Qu'il Faut Retenir

CQFR

- Corrélation $\neq$ causalité
- La statistique peut être comme la langue d'Esoppe la meilleure ou la pire des choses. Il convient de se méfier des pièges qu'elle recèle tout en se servant de ses résultats.
- Les représentations graphiques des données statistiques permettent une analyse visuelle de la répartition des données.
- Les indicateurs de localisation, de dispersion et de relation permettent de les quantifier et de résumer l'information.
- Ces deux outils suggèrent une caractérisation de la loi statistique sous-jacente et donc des modèles théoriques plausibles.

2. Liaison entre 2 variables

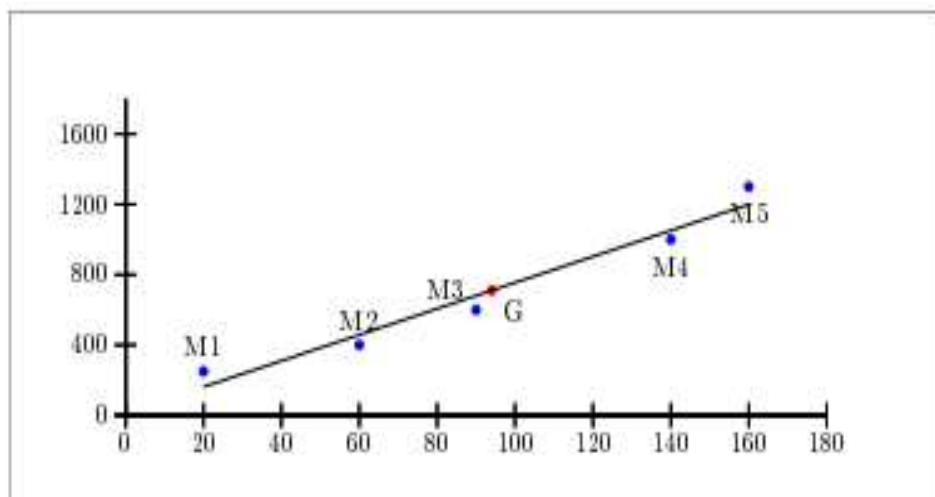
2.1 Liaison entre 2 variables quantitatives

2.1.1 Regression linéaire simple

On souhaite étudier la relation Superficie - Prix de 5 zones de stockage de marchandises dans un espace portuaire donné. La variable quantitative X désigne la surface en m^2 , et la variable quantitative Y , le coût d'usage de ladite zone (location, détention, manutention, ...). Le tableau suivant donne les valeurs de ces 2 variables.

X (en m^2)	$x_1=20$	$x_2=60$	$x_3=90$	$x_4=140$	$x_5=160$
Y (en K-euros)	$y_1=250$	$y_2=400$	$y_3=600$	$y_4=1000$	$y_5=1300$

On commence par visualiser les variables X et Y en les représentant sous la forme d'un **nuage de point** : dans un repère cartésien, chaque observation (x_i, y_i) est figurée par le point M_i de coordonnées (x_i, y_i) . On cherche une approximation de ce nuage dans un but de simplification ; sa forme donne une information sur le type d'une éventuelle liaison entre les variables X et Y .



Dans l'exemple étudié, on observe un nuage **oblong** (allongé), nous permettant d'envisager **une liaison linéaire** entre la surface d'un appartement et son prix. Plus précisément, il semble raisonnable de considérer que la relation entre la surface x_i d'un appartement et son prix y_i , est à peu près de la forme $y_i = ax_i + b$. Les coefficients (ou paramètres) a et b seront choisis de la sorte que la droite d'équation $y = ax + b$ passe « **le plus près possible de l'ensemble des points du nuage** » ; nous allons maintenant formaliser cette idée.

Considérons une droite D d'équation $y = ax + b$ et soit Δ la droite parallèle à l'axe des ordonnées et passant par le point M_i . Les droites Δ et D se coupent en un point M'_i ; la distance de M_i à M'_i vaut $|y_i - ax_i - b|$. Les coefficients a et b seront choisis de sorte que la quantité :

$$(y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + (y_3 - ax_3 - b)^2 + (y_4 - ax_4 - b)^2 + (y_5 - ax_5 - b)^2,$$

soit minimale.

Plus généralement, soient $x_1, x_2, \dots, x_N$ et $y_1, y_2, \dots, y_N$, les valeurs observées de deux variables quantitatives X et Y , pour un échantillon de N individus. Les coefficients de la **droite des moindres carrés**, c'est-à-dire de la droite qui permet d'ajuster au mieux, au sens du critère des moindres carrés, le nuage de points $M_1 = (x_1, y_1); M_2 = (x_2, y_2); \dots; M_N = (x_N, y_N)$ sont les nombres a et b qui rendent minimale la quantité

$$(y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + \dots + (y_N - ax_N - b)^2.$$

Ils sont donnés par les deux formules :

$$a = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_N - \bar{x})(y_N - \bar{y})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2} \quad (3.1)$$

et

$$b = \bar{y} - a\bar{x}. \quad (3.2)$$

La formule (3.2) signifie que la droite des moindres carrés passe par le **centre de gravité** du nuage de points $M_1 = (x_1, y_1); M_2 = (x_2, y_2); \dots; M_N = (x_N, y_N)$, c'est-à-dire par le point G de coordonnées $(\bar{x}, \bar{y})$ où, $\bar{x}$ et $\bar{y}$ sont les moyennes arithmétiques des variables X et Y .

Une fois qu'on a déterminé a et b , pour tout $i = 1, 2, \dots, N$, on pose :

$$\hat{y}_i = ax_i + b; \quad (3.3)$$

cette quantité $\hat{y}_i$ est appelée la **valeur estimée de Y , par la droite des moindres carrés, lorsque X vaut x_i** . Quand l'ajustement est de bonne qualité, cette valeur estimée $\hat{y}_i$ est assez proche de y_i la valeur réelle de Y lorsque X vaut x_i .

Appliquons maintenant, à l'exemple qui nous intéresse, les formules qu'on vient de donner dans un cadre général.

La moyenne arithmétique $\bar{x}$ des surfaces des 5 appartements vaut $\bar{x} = \frac{470}{5} = 94 m^2$, la moyenne arithmétique $\bar{y}$ de leurs prix vaut $\bar{y} = \frac{3550}{5} = 710$ milliers d'Euros; ainsi, G le centre gravité du nuage des 5 points associé aux variables X et Y , admet pour coordonnées $(94, 710)$.

Le tableau suivant va nous permettre de calculer les valeurs de a et b :

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
-74	-460	34040	5476	211600
-34	-310	10540	1156	96100
-4	-110	440	16	12100
46	290	13340	2116	84100
66	590	38940	4356	348100
		Total = 97300	Total = 13120	Total = 752000

(3.4)

ainsi, grâce aux formules (3.1) et (3.2), on trouve que :

$$a = \frac{97300}{13120} \simeq 7,416 \quad \text{et} \quad b = 710 - 7,416 \times 94 \simeq 12,896, \quad (3.5)$$

donc la droite des moindres carrés admet pour équation :

$$y = 7,416x + 12,896.$$

Calculons enfin, $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_5$, les prix estimés en milliers d'Euros des 5 appartements. Grâce à (3.3) et à (3.5), on trouve que : $\hat{y}_1 = 7,416 \times 20 + 12,896 \simeq 161$; $\hat{y}_2 = 7,416 \times 60 + 12,896 \simeq 458$; $\hat{y}_3 = 7,416 \times 90 + 12,896 \simeq 680$; $\hat{y}_4 = 7,416 \times 140 + 12,896 \simeq 1051$ et $\hat{y}_5 = 7,416 \times 160 + 12,896 \simeq 1199$.

Le tableau suivant permet de comparer les prix réels des appartements à leurs prix estimés au moyen de droite des moindres carrés :

X (en m^2)	$x_1 = 20$	$x_2 = 60$	$x_3 = 90$	$x_4 = 140$	$x_5 = 160$
Valeur réelle de Y (en milliers d'Euros)	$y_1 = 250$	$y_2 = 400$	$y_3 = 600$	$y_4 = 1000$	$y_5 = 1300$
Valeur estimée de Y (en milliers d'Euros)	$\hat{y}_1 = 161$	$\hat{y}_2 = 458$	$\hat{y}_3 = 680$	$\hat{y}_4 = 1051$	$\hat{y}_5 = 1199$

3.1.2 Covariance et coefficient de corrélation

Il est toujours possible de tracer la droite des moindres carrés quelle que soit la forme du nuage de points $M_1 = (x_1, y_1)$; $M_2 = (x_2, y_2)$; ...; $M_N = (x_N, y_N)$. L'approximation de ce nuage par cette droite est-elle pour autant légitime ?

Un premier élément de réponse à cette question est donné par l'examen de $R(X, Y)$ le **coefficient de corrélation linéaire des variables X et Y** (parfois on dit le coefficient de corrélation

linéaire entre les variables X et Y). Pour pouvoir définir ce coefficient, il faut d'abord définir la **covariance de X et Y** (parfois on dit la covariance entre X et Y).

$x_1, x_2, \dots, x_N$ et $y_1, y_2, \dots, y_N$ désignent les valeurs prises par X et Y pour une population de N individus. La **covariance de X et Y**, notée par $\text{cov}(X, Y)$, est définie par :

$$\text{cov}(X, Y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_N - \bar{x})(y_N - \bar{y})}{N}, \quad (3.6)$$

où $\bar{x}$ et $\bar{y}$ désignent les moyennes arithmétiques de X et Y ; notons que

$$\text{cov}(X, X) = \text{Var}(X).$$

La covariance de X et Y peut aussi être calculée au moyen de la formule (parfois désignée par formule de Huygens) :

$$\text{cov}(X, Y) = \left(\frac{x_1 y_1 + x_2 y_2 + \dots + x_N y_N}{N} \right) - \bar{x} \bar{y}; \quad (3.7)$$

en fait la formule (2.2) n'est rien d'autre que la formule (3.7), dans le cas où $X = Y$.

Exemple 3.2. Soient X et Y les variables « Superficie » et « Prix », dont il est question dans l'Exemple 3.1 (l'exemple des appartements). Nous allons calculer $\text{cov}(X, Y)$ au moyen de deux méthodes : la première d'entre elles consiste à utiliser la formule (3.6), et la seconde consiste à à utiliser la formule (3.7).

Présentons d'abord la première méthode. On a déjà vu que (voir le tableau (3.4)) :

$$(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + (x_3 - \bar{x})(y_3 - \bar{y}) + (x_4 - \bar{x})(y_4 - \bar{y}) + (x_5 - \bar{x})(y_5 - \bar{y}) = 97300;$$

ainsi, il résulte de la formule (3.6) que :

$$\text{cov}(X, Y) = \frac{97300}{5} = 19460.$$

Présentons maintenant la seconde méthode. Pour calculer $x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5$, nous utilisons le tableau suivant :

x_i	y_i	x_iy_i
20	250	5000
60	400	24000
90	600	54000
140	1000	140000
160	1300	208000
		total = 431000

qui nous permet de trouver que : $x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5 = 431000$; ainsi, on obtient que :

$$\frac{x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5}{5} = \frac{431000}{5} = 86200. \quad (3.8)$$

D'autre part, dans la Sous-section 3.1.1, on a vu que $\bar{x} = 94$ et $\bar{y} = 710$; on a par conséquent :

$$\bar{x}\bar{y} = 94 \times 710 = 66740. \quad (3.9)$$

Finalement, en utilisant la formule (3.7), ainsi que (3.8) et (3.9), on obtient :

$$\text{cov}(X, Y) = 86200 - 66740 = 19460.$$

Remarque 3.1. (Inégalité de Cauchy-Schwarz) La valeur absolue de la covariance de deux variables quantitatives X et Y , est toujours inférieure ou égale au produit de leurs écarts-types :

$$|\text{cov}(X, Y)| \leq \sigma_X \sigma_Y ;$$

cette inégalité peut aussi s'écrire sous la forme

$$-\sigma_X \sigma_Y \leq \text{cov}(X, Y) \leq \sigma_X \sigma_Y.$$

Ecrivons l'inégalité de Cauchy-Schwarz dans le cas de l'Exemple 3.1 (l'exemple des appartements). Pour cet exemple, on a déjà montré que $\text{cov}(X, Y) = 19460$; il nous reste à calculer les écarts-types σ_X et σ_Y . On a déjà vu que (voir le tableau (3.4)) :

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2 = 13120$$

et

$$(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 + (y_4 - \bar{y})^2 + (y_5 - \bar{y})^2 = 752000 ;$$

on obtient donc, au moyen de la formule (2.1), que $\text{Var}(X) = \frac{13120}{5} = 2624$ et $\text{Var}(Y) = \frac{752000}{5} = 150400$, d'où $\sigma_X = \sqrt{2624} \simeq 51,22$ et $\sigma_Y = \sqrt{150400} \simeq 387,81$. Ainsi, dans le cas de l'Exemple 3.1, l'inégalité de Cauchy-Schwarz s'écrit :

$$19460 = |\text{cov}(X, Y)| \leq \sigma_X \sigma_Y \simeq 51,22 \times 387,81 \simeq 19863,63.$$

Le coefficient de corrélation linéaire des deux variables X et Y , noté $R(X, Y)$, est défini par

$$R(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}. \quad (3.10)$$

Ainsi, dans le cas de l'Exemple 3.1, on a

$$R(X, Y) \simeq \frac{19460}{51,22 \times 387,81} \simeq 0,979.$$

Remarque 3.2. (Propriétés importantes du coefficient de corrélation linéaire)

- (i) Il résulte de l'inégalité de Cauchy-Schwarz que $R(X, Y)$ est toujours compris entre -1 et $+1$.
- (ii) Le coefficient directeur a (la pente) de la droite des moindres carrés vérifie :

$$a = \frac{\text{cov}(X, Y)}{\text{Var}(X)} = \frac{\sigma_Y}{\sigma_X} R(X, Y);$$

par conséquent a et $R(X, Y)$ sont toujours de même signe.

Remarque 3.3. (Interprétation du coefficient de corrélation linéaire)

- (i) Lorsque $R(X, Y)$ est voisin de 0 , il y a absence de corrélation entre les variables X et Y ; l'approximation du nuage de points par la droite des moindres carrés est alors illégitime et il faut rejeter l'ajustement linéaire.
- (ii) Lorsque $R(X, Y)$ est voisin de $+1$, il y a une corrélation directe entre les variables X et Y : cela signifie grosso modo que Y augmente lorsque X augmente, et que X augmente lorsque Y augmente.
- (iii) Lorsque $R(X, Y)$ est voisin de -1 , il y a une corrélation inverse entre les variables X et Y : cela signifie grosso modo que Y augmente lorsque X diminue, et que X diminue lorsque Y augmente.

2.2 Liaison entre 2 variables qualitatives (M2 spécialement)**2.2.1 Tableau de contingence (ou tableau croisé)**

Exemple: Soit 218 645 Établissements industriels et commerciaux, opérateurs de la chaîne de transport international et spécialisés dans le commerce maritime, de plus de 10 salariés qui se répartissent en fonction de leur localisation et de leur taille de la manière suivante:

REGIONS	TABLEAU DES EFFECTIFS OBSERVES					Total
	Classe d'effectif des établissements					
	10-49	50-99	100-499	500-199	+2000	
ILE DE FRANCE	43003	3825	1812	688	101	55340
RHONE ALPES	18055	3453	869	188	18	22289
PROVENCE COTE D'AZUR	12174	1600	284	106	18	14512
NORD-PAS DE CALAIS	10357	2362	487	157	8	13318
PAYS DE LOIRE	8131	1885	312	88	5	10226
BRETAGNE	7841	1609	248	48	5	9749
AQUITAINE	7695	1308	203	67	7	9520
CENTRE	7348	1545	296	85	6	9280
MIDI PYRÉNÉE	6978	1010	179	61	4	8240
LORRAINE	6258	1332	251	88	15	7943
ALSACE	5870	1026	231	82	6	7014
HAUTE-NORMANDIE	5113	1130	209	74	5	6531
PICARDIE	4843	1075	203	68	5	6214
LANGUEDOC ROUSSILLON	5058	766	121	29	4	6006
BOURGOGNE	4772	837	171	83	7	5847
CHAMPAGNE ARDENNE	4088	607	194	56	4	5239
POITOU CHARENTES	4256	732	126	48	2	5164
BASSE-NORMANDIE	3807	793	122	34	5	4758
AUVERGNE	3821	672	87	40	5	4525
FRANCHE COMTE	3152	618	114	28	7	3917
LIMOUSIN	1894	358	93	13	1	2327
CORSE	580	51	4	3	0	618
Total	118954	34025	8274	2100	233	218683

nombre 2362 se trouve sur la ligne Nord-Pas de Calais et sur la colonne 50-199 ; cela signifie que sur les 218645 établissements recensés 2362 se trouvent dans la région NPdC et possèdent chacun un effectif compris entre 50 et 199 salariés.

Le nombre 13318 qui se trouve sur la colonne Total et sur la ligne NPdC signifie que sur les 218645 établissements recensés 13318 se trouvent dans la région NPdC ; ce nombre est donc égal à la somme de tous les autres nombres qui se trouvent sur la ligne NPdC.

Le nombre 34025 qui se trouve sur la ligne Total et sur la colonne 50-199 signifie que sur les 218645 établissements recensés 34025 possèdent un effectif compris entre 50 et 199 salariés ; ce nombre est donc égal à la somme de tous les autres nombres qui se trouvent sur la colonne 50-199.

Le nombre qui se trouvent sur la ligne Total et sur la colonne Total correspond au total des établissements recensés c'est-à-dire 218645 ; ce nombre est donc égal à la somme de tous les autres nombres qui se trouvent sur la ligne Total, il est aussi égal à la somme de tous les autres nombres qui se trouvent sur la colonne Total.

De façon générale, soient Z et T deux variables qualitatives dont les modalités sont respectivement $z_1, \dots, z_i, \dots, z_k$ et $t_1, \dots, t_j, \dots, t_l$. Les valeurs de ces variables ont été observées sur une population de n individus.

La répartition des effectifs suivant les modalités de Z et de T , se présente sous forme d'un tableau à double entrée, appelé tableau de contingence ou encore tableau croisé :

$Z \setminus T$	t_1	$\dots$	t_j	$\dots$	t_l	Total
z_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1l}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	n_{k1}	$\dots$	n_{kj}	$\dots$	n_{kl}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet l}$	n

L'effectif n_{ij} qui se trouve sur la i -ème ligne et la j -ème colonne du tableau de contingence, est le nombre d'individus qui possèdent à la fois la modalité z_i de la variable Z et la modalité t_j de la variable T . Les effectifs n_{ij} , $i = 1, \dots, k$ et $j = 1, \dots, l$ sont appelés **les effectifs croisés observés**.

L'effectif $n_{i\bullet}$ qui se trouve sur la i -ème ligne et la colonne Total est le nombre d'individus qui possèdent la modalité z_i de la variable Z ; on a donc :

$$n_{i\bullet} = n_{i1} + n_{i2} + \dots + n_{il}.$$

L'effectif $n_{\bullet j}$ qui se trouve sur la j -ème colonne et la ligne Total est le nombre d'individus qui possèdent la modalité t_j de la variable T ; on a donc :

$$n_{\bullet j} = n_{1j} + n_{2j} + \dots + n_{kj}.$$

L'effectif n qui se trouve sur la ligne Total et la colonne Total est le nombre d'individus de la population étudiée; on a donc

$$n = n_{1\bullet} + n_{2\bullet} + \dots + n_{k\bullet}$$

et

$$n = n_{\bullet 1} + n_{\bullet 2} + \dots + n_{\bullet l}.$$

La fréquence de la modalité z_i de la variable Z est donnée par :

$$f_{i\bullet} = \frac{n_{i\bullet}}{n}.$$

Ainsi sur les 218645 établissements recensés $f_{1\bullet} = \frac{55349}{218645} \simeq 0,253$ (soit 25,3%) c'est-à-dire plus d'un établissement sur 4 se trouve dans la région parisienne. Trois autres régions concentrent les établissements, Rhône-Alpes ($f_{2\bullet} = \frac{22280}{218645} \simeq 0,102$ soit 10,2%), Provence Côte d'Azur ($f_{3\bullet} = \frac{14512}{218645} \simeq 0,066$ soit 6,6%) et Nord-Pas de Calais ($f_{4\bullet} = \frac{13318}{218645} \simeq 0,061$ soit 6,1%).

La fréquence de la modalité t_j de la variable T est donnée par

$$f_{\bullet j} = \frac{n_{\bullet j}}{n}.$$

Dans notre exemple, il ressort de l'étude des fréquences $f_{\bullet j}$, une répartition asymétrique des entreprises en fonction de leurs effectifs ($f_{1\bullet} = \frac{176004}{218645} \simeq 0,805$ soit 80,5%) ont moins de 50 salariés et seuls ($f_{5\bullet} = \frac{233}{218645} \simeq 0,001$ soit 0,1%) en ont plus de 2000.

La donnée des modalités z_i de la variable Z et des fréquences correspondantes $f_{i\bullet}$ (ou encore des effectifs correspondant $n_{i\bullet}$) est appelée **distribution marginale** de la variable Z .

La donnée des modalités t_j de la variable T et des fréquences correspondantes $f_{\bullet j}$ (ou encore des effectifs correspondant $n_{\bullet j}$) est appelée **distribution marginale** de la variable T .

La fréquence conditionnelle de z_i sachant que $T = t_j$ est donnée par

$$f_{i|j} = \frac{n_{ij}}{n_{\bullet j}}.$$

$f_{i|j}$ se lit « f indice i si j ». On a donc $f_{1|j} + f_{2|j} + \dots + f_{k|j} = \frac{n_{\bullet j}}{n_{\bullet j}} = 1$.

Le tableau suivant est appelé **tableau des profils colonnes**

$Z \setminus T$	t_1	$\dots$	t_j	$\dots$	t_l	Distribution marginale de Z
z_1	$f_{1 1}$	$\dots$	$f_{1 j}$	$\dots$	$f_{1 l}$	$f_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	$f_{i 1}$	$\dots$	$f_{i j}$	$\dots$	$f_{i l}$	$f_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	$f_{k 1}$	$\dots$	$f_{k j}$	$\dots$	$f_{k l}$	$f_{k\bullet}$
Total	1	$\dots$	1	$\dots$	1	1

$f_{i|j}$ se trouve sur la i -ème ligne et la j -ème colonne du tableau. De façon général, ce tableau permet de comparer les profils colonnes (les colonnes) au profil marginal colonne (dernière colonne) et de les comparer entre eux. Dans le cas de notre exemple, au moyen du Tableau 3 (voir un peu plus loin), on peut capter pour chaque classe d'effectif la répartition géographique des entreprises correspondantes. On se rend compte notamment que la concentration dans la région Île de France des grandes entreprises est nettement plus forte que celle des petites.

La fréquence conditionnelle de t_j sachant que $Z = z_i$ est donnée par

$$f_{j|i} = \frac{n_{ij}}{n_{i\bullet}}$$

On a donc

$$f_{1|i} + f_{2|i} + \dots + f_{l|i} = 1.$$

$f_{j|i}$ se lit « f indice j si i ».

Le tableau suivant est appelé **tableau des profils lignes**

$Z \setminus T$	t_1	...	t_j	...	t_l	Total
z_1	$f_{1 1}$	...	$f_{j 1}$	...	$f_{l 1}$	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	$f_{1 i}$	...	$f_{j i}$	...	$f_{l i}$	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	$f_{1 k}$	...	$f_{j k}$	...	$f_{l k}$	1
Distribution marginale de T	$f_{\bullet 1}$	...	$f_{\bullet j}$	...	$f_{\bullet l}$	1

$f_{j|i}$ se trouve sur la i -ème ligne et la j -ème colonne du tableau. De façon générale, ce tableau permet de comparer les profils lignes (les lignes) au profil marginal ligne (dernière ligne) et de les comparer entre eux. Dans le cas de notre exemple, le Tableau 2 (voir un peu plus loin) donne pour chaque région la répartition des entreprises par classe d'effectif. On se rend compte qu'il n'y a guère de différence entre les régions. Dans chaque région, les petites entreprises sont largement majoritaires alors que les grandes sont largement minoritaires.

Tableau 2 (Profils lignes)

REGIONS	PROFILS LIGNES					Fréquence
	Classe d'effectif des établissements					
	10-49	50-199	200-499	500-199	+2000	
ILE DE FRANCE	78,4%	13,8%	3,3%	1,2%	0,2%	100,0%
RHONE ALPES	81,0%	15,8%	2,6%	0,6%	0,1%	100,0%
PROVENCE COTE D'AZ	83,8%	13,3%	2,0%	0,7%	0,1%	100,0%
NORD-PAS DE CALAIS	77,4%	17,7%	3,7%	1,2%	0,1%	100,0%
PAYS DE LOIRE	79,7%	16,3%	3,1%	0,9%	0,0%	100,0%
BRETAGNE	80,4%	16,5%	2,5%	0,3%	0,1%	100,0%
AQUITAINE	83,4%	13,7%	2,1%	0,7%	0,1%	100,0%
CENTRE	79,3%	16,7%	3,1%	0,9%	0,1%	100,0%
MIDI PYRENEE	84,7%	12,4%	2,2%	0,7%	0,0%	100,0%
LORRAINE	78,8%	16,8%	3,2%	1,1%	0,2%	100,0%
ALSACE	80,8%	14,8%	3,3%	1,2%	0,1%	100,0%
HAUTE-NORMANDIE	76,3%	17,3%	3,2%	1,1%	0,1%	100,0%
PICARDIE	77,9%	17,3%	3,3%	1,4%	0,1%	100,0%
LANGUEDOC ROUSSE	84,2%	13,2%	2,0%	0,5%	0,1%	100,0%
BOURGOGNE	80,2%	15,8%	2,9%	1,0%	0,1%	100,0%
CHAMPAGNE ARDENN	78,0%	17,1%	3,7%	1,1%	0,1%	100,0%
POITOU CHARENTES	82,4%	14,2%	2,4%	0,9%	0,0%	100,0%
BASSE-NORMANDIE	80,9%	16,6%	2,8%	0,7%	0,1%	100,0%
AUVERGNE	84,4%	12,6%	1,9%	0,9%	0,1%	100,0%
FRANCHE COMTE	80,5%	15,8%	2,9%	0,7%	0,2%	100,0%
LIÉGEOIS	81,4%	15,3%	2,7%	0,6%	0,0%	100,0%
CORSE	90,8%	8,3%	0,8%	0,8%	0,8%	100,0%
Fréquence	80,5%	15,5%	2,9%	1,0%	0,1%	100,0%

Tableau 3 (Profils colonnes)

REGIONS	PROFILS COLONNES					Fréquence
	Classe d'effectif des établissements					
	10-49	50-199	200-499	500-1999	+2000	
ILE DE FRANCE	25,6%	25,9%	28,3%	31,7%	43,3%	25,3%
RHONE ALPES	10,3%	10,1%	9,1%	8,9%	6,4%	10,2%
PROVENCE COTE D'AZ	6,9%	6,7%	4,5%	5,1%	6,0%	6,8%
NORD-PAS DE CALAIS	5,9%	6,5%	7,8%	7,4%	3,4%	6,1%
PAYS DE LOIRE	4,6%	4,0%	5,0%	4,2%	2,1%	4,7%
BRETAGNE	4,6%	4,7%	3,9%	2,3%	2,1%	4,6%
AQUITAINE	4,5%	3,8%	3,2%	3,2%	3,0%	4,4%
CENTRE	4,2%	4,5%	4,8%	4,0%	2,1%	4,2%
MIDI-PYRENEE	4,0%	3,0%	2,9%	2,9%	1,7%	3,8%
LORRAINE	3,6%	3,9%	4,0%	4,1%	6,4%	3,6%
ALSACE	3,2%	3,0%	3,7%	3,9%	2,8%	3,2%
HAUTE-NORMANDIE	2,9%	3,3%	3,3%	3,5%	2,1%	3,0%
PICARDIE	2,8%	3,2%	3,2%	4,2%	2,1%	2,8%
LANGUEDOC ROUSSIL	2,9%	2,3%	1,9%	1,3%	1,7%	2,7%
BOURGOGNE	2,7%	2,8%	2,7%	2,8%	3,9%	2,7%
CHAMPAGNE ARDENNE	2,3%	2,6%	3,1%	2,7%	1,7%	2,4%
POITOU CHARENTES	2,4%	2,2%	2,0%	2,0%	0,8%	2,4%
BASSE-NORMANDIE	2,2%	2,3%	1,9%	1,6%	2,1%	2,2%
Auvergne	2,2%	1,7%	1,4%	1,9%	2,1%	2,1%
FRANCHE COMTE	1,8%	1,8%	1,8%	1,2%	3,0%	1,8%
LIÉGEOIS	1,1%	1,0%	1,0%	0,8%	0,4%	1,1%
CORSE	0,3%	0,1%	0,1%	0,1%	0,0%	0,3%
Fréquence	100,0%	100,0%	100,0%	100,0%	99,8%	

3.2.2 Test d'une éventuelle liaison (test du χ^2 « chi 2 »)

Il n'y a pas de liaison entre les variables Z et T , lorsque tous les profils colonnes sont identiques au profil marginal colonne. Autrement dit, pour tout $i = 1, \dots, k$ et tout $j = 1, \dots, l$ $f_{i|j}$ la fréquence conditionnelle de z_i sachant $T = t_j$ est égale à $f_{i\bullet}$, la fréquence de z_i . Cette égalité est équivalente à l'égalité

$$\frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n}$$

ou encore à l'égalité

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}.$$

Il n'y a également pas de liaisons entre les variables Z et T , lorsque tous les profils lignes sont identiques au profil marginal ligne. Autrement dit, pour tout $i = 1, \dots, k$ et tout $j = 1, \dots, l$ $f_{j|i}$ la fréquence conditionnelle de t_j sachant $Z = z_i$ est égale à $f_{\bullet j}$, la fréquence de t_j . Cette égalité est équivalente à l'égalité

$$\frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n}$$

ou encore à l'égalité

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n},$$

qu'on a déjà vue plus haut.

Dans le cas de notre exemple, les profils colonnes ne sont pas identiques au profil marginal colonne. Cela signifie qu'il existe une liaison entre la variable « Régions » et la variable « Classe d'Effectif des Etablissements ». Pour tout $i = 1, \dots, k$ et tout $j = 1, \dots, l$ on pose

$$n_{ij}^* = \frac{n_{i\bullet} n_{\bullet j}}{n}.$$

Les quantités n_{ij}^* sont appelées **les effectifs (croisés) théoriques** ; il s'agit en fait des effectifs qu'on aurait obtenus s'il n'y avait pas eu de liaison entre les variables Z et T . Par exemple, l'effectif théorique croisé Ile de France, Classe d'effectif 10-49 vaut $n_{11}^* = \frac{55349 \times 176604}{218645} \simeq 44555$ et l'effectif théorique croisé Nord-Pas de Calais, Classe d'effectif 200-499 vaut $n_{41}^* = \frac{13318 \times 6274}{218645} \simeq 382$.

Le tableau suivant est appelé tableau des effectifs théoriques

$Z \setminus T$	t_1	$\dots$	t_j	$\dots$	t_l	Total
z_1	n_{11}^*	$\dots$	n_{1j}^*	$\dots$	n_{1l}^*	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	n_{i1}^*	$\dots$	n_{ij}^*	$\dots$	n_{il}^*	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	n_{k1}^*	$\dots$	n_{kj}^*	$\dots$	n_{kl}^*	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet l}$	n

n_{ij}^* se trouve sur la i -ème ligne et la j -ème colonne du tableau. Plus la différence entre le tableau de contingence (le tableau des effectifs croisés observés) et le tableau des effectifs théoriques est grande, plus grande est la probabilité d'existence d'une liaison significative entre les variables Z et T . Pour formaliser cette idée, il convient d'introduire la quantité suivante appelée distance du χ^2 (« chi 2 »).

$$\begin{aligned}\chi^2 &= \frac{(n_{11} - n_{11}^*)^2}{n_{11}^*} + \frac{(n_{12} - n_{12}^*)^2}{n_{12}^*} + \dots + \frac{(n_{1l} - n_{1l}^*)^2}{n_{1l}^*} \\ &+ \frac{(n_{21} - n_{21}^*)^2}{n_{21}^*} + \frac{(n_{22} - n_{22}^*)^2}{n_{22}^*} + \dots + \frac{(n_{2l} - n_{2l}^*)^2}{n_{2l}^*} \\ &\vdots \\ &+ \frac{(n_{k1} - n_{k1}^*)^2}{n_{k1}^*} + \frac{(n_{k2} - n_{k2}^*)^2}{n_{k2}^*} + \dots + \frac{(n_{kl} - n_{kl}^*)^2}{n_{kl}^*}\end{aligned}$$

La distance du χ^2 mesure l'écart entre le tableau de contingence et le tableau des effectifs théoriques. Plus elle est grande, plus cet écart est important. Lorsqu'il n'y a pas de liaisons entre Z et T , comme on l'a vu précédemment, les effectifs croisés observés sont égaux aux effectifs théoriques (pour tout $i = 1, \dots, k$ et pour tout $j = 1 \dots l$ $n_{ij} = n_{ij}^*$) et cela est équivalent à $\chi^2 = 0$.

Les χ^2 partiels sont les quantités χ_{ij}^2 définies pour tout $i = 1, \dots, k$ et tout $l = 1, \dots, l$ par

$$\chi_{ij}^2 = \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}.$$

χ_{ij}^2 mesure le carré de l'écart entre l'effectif observé n_{ij} et l'effectif théorique n_{ij}^* relativement à l'effectif théorique n_{ij}^* . Par exemple, le χ^2 partiel Île de France, Classe d'effectif 10-49 vaut $\chi_{11}^2 = \frac{(43943 - 44555)^2}{44555} \simeq 8,4$ et le χ^2 partiel Nord-Pas de Calais, Classe d'effectif 200-499 vaut $\chi_{43}^2 = \frac{(487 - 382)^2}{382} \simeq 28,86$.

Lorsque pour un certain i_0 et un certain j_0 l'effectif observé $n_{i_0 j_0}$ est plus grand que l'effectif théorique $n_{i_0 j_0}^*$ ($n_{i_0 j_0} > n_{i_0 j_0}^*$) on dit qu'il y a attraction entre la modalité z_{i_0} de la variable Z et la modalité t_{j_0} de la variable T . Lorsque pour un certain i_1 , et un certain j_1 , l'effectif observé $n_{i_1 j_1}$ est plus petit que l'effectif théorique $n_{i_1 j_1}^*$ ($n_{i_1 j_1} < n_{i_1 j_1}^*$) on dit qu'il y a répulsion entre la modalité z_{i_1} de la variable Z et la modalité t_{j_1} de la variable T .

Dans le cas de notre exemple, il y a répulsion entre la modalité Île de France de la variable Région et la modalité 10-49 de la variable classe d'effectif (car $n_{11} = 43943 < 44555 = n_{11}^*$). En revanche, il y a attraction entre la modalité Nord-Pas de Calais de la classe Région et la modalité 200-499 de la variable classe d'effectif (car $n_{43} = 487 > 382 = n_{43}^*$).

Il résulte de ce qui précède que la distance du χ^2 est égale à la somme de tous les χ^2 partiels

$$\begin{aligned}\chi^2 &= \chi_{11}^2 + \chi_{12}^2 + \dots + \chi_{1l}^2 \\ &+ \chi_{21}^2 + \chi_{22}^2 + \dots + \chi_{2l}^2 \\ &\vdots \\ &+ \chi_{k1}^2 + \chi_{k2}^2 + \dots + \chi_{kl}^2\end{aligned}$$

Tableau 4 (effectifs théoriques)

REGIONS	EFFECTIFS THEORIQUES					Total colonne
	Classe d'effectif des établissements					
	10-49	50-199	200-499	500-199	+2000	
ILE DE FRANCE	44683	8815	1588	534	50	55549
RHON ALPES	17835	3467	839	215	24	22280
PROVENCE COTE D'AZ	11682	2268	416	140	15	14512
NORD-PAS DE CALAIS	10721	2073	382	128	14	13318
PAYS DE LOIRE	8218	1688	293	98	11	10208
BRETAGNE	7846	1517	280	94	10	9749
AQUITAINE	7863	1481	273	92	10	9520
CENTRE	7461	1442	296	89	10	9289
MIDI-PYRENEE	6033	1282	238	79	9	8240
LORRAINE	6393	1236	228	77	8	7942
ALSACE	5646	1092	201	68	7	7014
HAUTE-NORMANDIE	5257	1018	187	63	7	6531
PICARDIE	5002	987	178	60	7	6234
LANGUEDOC ROUSSILLON	4835	935	172	58	6	6006
BOURGOGNE	4787	925	171	57	6	5846
CHAMPAGNE ARDENNE	4217	815	150	51	6	5239
POITOU CHARENTES	4137	804	148	50	6	5164
BASSE-NORMANDIE	3830	760	137	46	5	4789
AUVERGNE	3643	704	130	44	5	4525
FRANCHE COMTE	3183	610	112	38	4	3917
LOUISIN	1873	362	67	22	2	2327
CORSE	497	95	18	6	1	618
TOTAL ligne	176054	34025	6274	2109	233	218645

Tableau 5 (des χ^2 partiels)

REGIONS	TABLEAU DES χ^2					Total c2
	Classe d'effectif des établissements					
	10-49	50-199	200-499	500-199	+2000	
ILE DE FRANCE	8,46	5,20	11,83	33,68	29,93	108,75
RHONE ALPES	0,80	0,06	7,74	3,37	3,22	15,19
PROVENCE COTE D'AZ	20,74	47,73	42,11	7,31	0,02	117,80
NORD-PAS DE CALAIS	15,96	40,43	28,76	6,34	2,70	94,20
PAYS DE LOIRE	0,87	3,71	1,25	0,91	3,17	9,91
BRETAGNE	0,01	5,57	4,07	22,64	2,80	34,07
AQUITAINE	9,83	20,31	18,83	8,71	0,87	58,66
CENTRE	1,72	7,30	1,51	0,22	2,41	13,16
MIDI-PYRENEE	17,94	54,47	13,96	4,30	2,60	93,27
LORRAINE	2,85	7,47	2,34	1,15	6,05	18,87
ALSACE	0,10	4,06	4,39	3,04	0,29	11,89
HAUTE-NORMANDIE	3,98	12,71	2,48	1,02	0,55	21,63
PICARDIE	5,98	12,66	3,42	13,14	0,40	34,08
LANGUEDOC ROUSSILLON	10,31	20,66	15,30	15,47	0,00	62,84
BOURGOGNE	0,66	0,14	0,00	0,12	0,07	0,38
CHAMPAGNE ARDENNE	3,86	8,19	12,88	0,59	0,45	25,88
POITOU CHARENTES	2,36	6,38	3,32	0,07	2,23	14,36
BASSE-NORMANDIE	0,14	3,32	1,55	3,08	0,00	8,09
AUVERGNE	8,75	24,81	14,14	0,32	0,01	48,00
FRANCHE COMTE	0,00	0,12	0,02	3,67	1,91	5,73
LOUISIN	0,23	0,10	0,21	3,97	0,88	5,41
CORSE	7,89	21,22	10,64	1,47	0,88	41,84
TOTAL c2	121,71	366,22	219,44	133,36	61,23	842,80

$Z \setminus T$	t_1	$\dots$	t_j	$\dots$	t_l	Total
z_1	χ_{11}^2	$\dots$	χ_{1j}^2	$\dots$	χ_{1l}^2	$\chi_{1\bullet}^2$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	χ_{i1}^2	$\dots$	χ_{ij}^2	$\dots$	χ_{il}^2	$\chi_{i\bullet}^2$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	χ_{k1}^2	$\dots$	χ_{kj}^2	$\dots$	χ_{kl}^2	$\chi_{k\bullet}^2$
Total	$\chi_{\bullet 1}^2$	$\dots$	$\chi_{\bullet j}^2$	$\dots$	$\chi_{\bullet l}^2$	χ^2

χ_{ij}^2 se trouve sur la i -ème ligne et la j -ème colonne. Pour tout $i = 1, \dots, k$, $\chi_{i\bullet}^2$ désigne la somme des χ^2 partiels se trouvant sur la i -ème ligne du tableau :

$$\chi_{i\bullet}^2 = \chi_{i1}^2 + \chi_{i2}^2 + \dots + \chi_{il}^2.$$

Pour tout $j = 1, \dots, l$, $\chi_{\bullet j}^2$ désigne la somme des χ^2 partiels se trouvant sur la j -ème ligne du tableau :

$$\chi_{\bullet j}^2 = \chi_{1j}^2 + \chi_{2j}^2 + \dots + \chi_{kj}^2.$$

D'après ce qui précède, on a

$$\chi^2 = \chi_{1\bullet}^2 + \dots + \chi_{k\bullet}^2 = \chi_{\bullet 1}^2 + \dots + \chi_{\bullet l}^2.$$

Pour calculer χ^2 , on commence, par calculer, pour chaque ligne du tableau, la somme des nombres s'y trouvant et on reporte les résultats dans la colonne Total. Ensuite, on calcule la somme de nombres se trouvant dans la colonne Total.

On peut également, pour calculer χ^2 , commencer par calculer pour chaque colonne du tableau, la somme des nombres s'y trouvant, reporter le résultat dans la ligne Total puis faire la somme des nombres s'y trouvant dans la ligne Total.

De façon générale, lorsque la valeur de la distance du χ^2 est plus grande qu'un certain seuil (la méthode permettant de déterminer ce seuil dépasse le cadre de ce cours), on accepte l'hypothèse d'existence d'une liaison entre les variables Z et T . Dans le cas de notre exemple, on trouve que $\chi^2 = 842$ et cela nous amène à accepter l'hypothèse de l'existence d'un lien entre la taille (l'effectif) d'un établissement industriel ou commercial et sa localisation géographique.

En examinant plus attentivement le tableau des χ^2 partiels, on s'aperçoit que dans certaines cases les valeurs sont sensiblement plus élevées qu'ailleurs. On est tenté de considérer que ce sont les cases les plus importantes, que ce sont ces situations qu'il faut interpréter. C'est notamment le cas des cases (Midi-Pyrénées, 50-199) ; (PACA ; 50-199) ; (PACA, 200-499) ; (NPdC, 50-199) ; (IdF, 500-1999) ; (IdF, 200-499) ...

Pour pouvoir identifier de façon précise les cases $(\cdot, \cdot)$ les plus importantes du tableau des χ^2 partiels, on est amené à considérer, pour tout $i = 1, \dots, k$ et tout $j = 1, \dots, l$ la quantité

$$CTR_{ij} = \frac{\chi_{ij}^2}{\chi^2} \times 100$$

Cette quantité est appelée **contribution relative de la case** (i, j) à la valeur de χ^2 . Dans le cas de la case (Midi-Pyrénées, 50-199), on trouve que $CTR_{22} = \frac{54.47}{842} \times 100 = 6,47\%$

Le tableau des contributions est le tableau suivant :

$Z \setminus T$	t_1	...	t_j	...	t_l	Total
z_1	CTR_{11}	...	CTR_{1j}	...	CTR_{1l}	$CTR_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	CTR_{i1}	...	CTR_{ij}	...	CTR_{il}	$CTR_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	CTR_{k1}	...	CTR_{kj}	...	CTR_{kl}	$CTR_{k\bullet}$
Total	$CTR_{\bullet 1}$	...	$CTR_{\bullet j}$	...	$CTR_{\bullet l}$	100%

Tableau 5. (tableau des contributions)

(4)

TABLEAU DES CONTRIBUTIONS						Contribut Ligne
RÉGIONS	Classe d'effectif des établissements					
	10-49	50-199	200-499	500-1999	+2000	
ILE DE FRANCE	1,00%	0,82%	3,74%	4,00%	3,55%	12,9%
RHÔNE ALPES	0,10%	0,01%	0,92%	0,40%	0,36%	1,8%
PROVENCE CÔTE D'AZ	2,48%	5,87%	5,00%	0,87%	0,30%	14,8%
NORD-PAS DE CALAIS	1,90%	4,80%	3,42%	0,75%	0,32%	11,2%
PAYS DE LOIRE	0,10%	0,44%	0,19%	0,11%	0,38%	1,2%
BRETAGNE	0,00%	0,69%	0,48%	2,68%	0,33%	4,2%
AQUITAINE	1,14%	2,41%	2,14%	0,80%	0,12%	6,6%
CENTRE	0,20%	0,87%	0,10%	0,03%	0,28%	1,6%
MIDI-PYRÉNÉES	2,13%	6,47%	1,88%	0,51%	0,21%	11,1%
LORRAINE	0,34%	0,68%	0,28%	0,14%	0,60%	2,2%
ALSACE	0,01%	0,48%	0,52%	0,36%	0,03%	1,4%
HAUTE-NORMANDIE	0,47%	1,51%	0,30%	0,23%	0,07%	2,6%
PICARDIE	0,60%	1,43%	0,41%	1,88%	0,06%	4,0%
LANGUEDOC-ROUSSIL	1,23%	2,48%	1,82%	1,84%	0,11%	7,5%
BOURGOGNE	0,01%	0,02%	0,00%	0,01%	0,01%	0,0%
CHAMPAGNE ARDENNE	0,47%	0,97%	1,51%	0,07%	0,06%	3,1%
POITOU CHARENTES	0,28%	0,78%	0,30%	0,01%	0,26%	1,7%
BASSE-NORMANDIE	0,02%	0,36%	0,18%	0,37%	0,00%	1,0%
AUVERGNE	1,04%	2,85%	1,68%	0,04%	0,00%	5,7%
FRANCHE COMTE	0,00%	0,01%	0,03%	0,44%	0,23%	0,7%
LOUISIN	0,03%	0,01%	0,03%	0,47%	0,10%	0,6%
CORSE	0,93%	2,52%	1,26%	0,17%	0,68%	5,0%
Contribution totale	14,5%	38,4%	28,1%	15,8%	7,3%	100,00%

III. ANALYSE DE SÉRIES CHRONOLOGIQUES

1. Description de série chronologique

On distingue en général trois effets constitutifs d'une série chronologique :

- Un effet à long terme, appelée tendance (on ajoute parfois à long terme), composante tendancielle ou trend ;
- Un effet dit saisonnier, qui réapparaît à intervalles réguliers ; cet effet se traduit par une composante de la série appelée composante saisonnière.
- Un effet inexpliqué : cet effet, que l'on suppose en général dû au hasard, se manifeste par des variations accidentelles.

Dans les séries économiques longues, on cite souvent un effet supplémentaire : c'est ce que l'on appelle le cycle de Kondratiev, qui résulte du fait que, suivant la théorie de Kondratiev, à une période de prospérité économique succède mécaniquement une période de dépression.

1.1 Description de la tendance

La description initiale de la tendance repose sur l'interprétation de la représentation graphique de la série.

Définition : on appelle tendance (ou variation à long terme ou trend) de la série x_t la série c_t résultant de la totalité des effets permanents auxquels est soumise la série x_t .

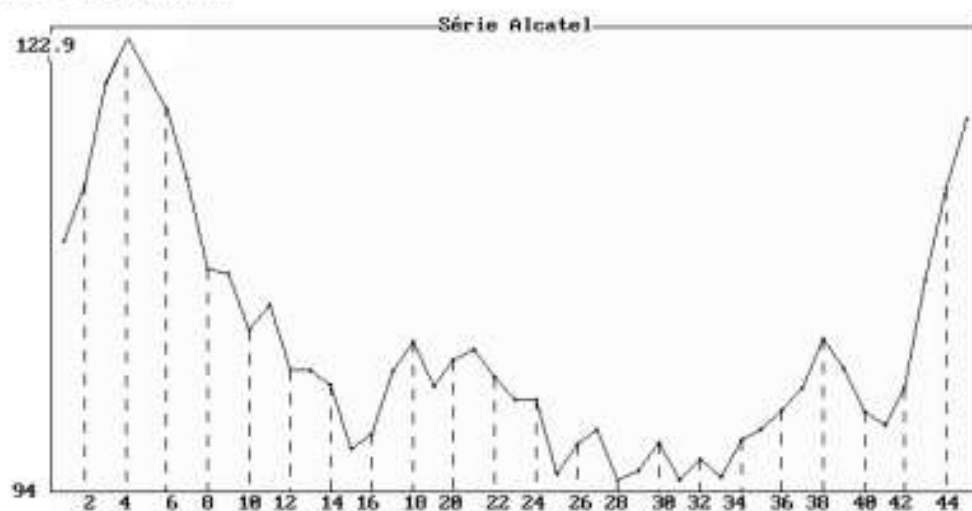
Exemple de l'évolution du taux de fret à différentes périodes à différentes périodes de facturation pour un trafic donné du 4 janvier (N°1) au 5 mars (N°45)

1	109.500	10	103.750	19	100.100	28	94.000	37	99.950
2	113.200	11	105.400	20	101.800	29	94.600	38	103.150
3	119.700	12	101.175	21	102.450	30	96.425	39	101.250
4	122.350	13	101.100	22	100.600	31	94.025	40	98.450
5	122.900	14	100.150	23	99.200	32	95.350	41	97.550
6	118.250	15	96.050	24	99.200	33	94.175	42	100.000
7	113.550	16	96.950	25	94.375	34	96.600	43	107.050
8	107.700	17	101.000	26	96.350	35	97.250	44	112.900
9	107.400	18	103.000	27	97.250	36	98.500	45	117.400

La tendance peut être décomposée en trois phases:

- Le taux de fret baisse du début des observations (7 janvier) jusqu'à l'observation N°15, c'est à dire le 22 janvier
- De l'observation N°16 (25 janvier) à l'observation N°41 (1er mars), le taux de fret diminue légèrement.
- Il augmente régulièrement jusqu'au 2 mars

On ne distingue pas sur la figure 1.8 de variations apparaissant à intervalles réguliers : la série n'est soumise à aucun effet saisonnier visible. La composante accidentelle est visualisée par les petites variations de cours d'un jour à l'autre. Par exemple le cours n°11 est supérieur au cours n°10 : cela ne remet pas en cause la baisse de la tendance compte tenu des cours du 7 au 22 janvier.



Pour faire apparaître plus clairement la tendance, il faut atténuer la composante accidentelle. On utilise pour cela les moyennes mobiles définies de la façon suivante :

Définitions :

- on appelle moyenne mobile centrée de longueur impaire $l_i = 2k + 1$ à l'instant t la valeur moyenne mm_t des observations $x_{t-k}, x_{t-k+1}, \dots, x_t, x_{t+1}, \dots, x_{t+k}$:

$$mm_t = (x_{t-k} + \dots + x_{t-1} + x_t + x_{t+1} + \dots + x_{t+k}) / l_i$$

- on appelle moyenne mobile centrée de longueur paire $l_p = 2k$ à l'instant t la valeur moyenne mm_t des observations $x_{t-k}, x_{t-k+1}, x_{t-k+2}, x_t, x_{t+1}, \dots, x_{t+k}$, la première et la dernière étant pondérées par 0.5 :

$$mm_t = (0.5 x_{t-k} + x_{t-k+1} + \dots + x_{t-1} + x_t + x_{t+1} + \dots + x_{t+k-1} + 0.5 x_{t+k}) / l_p$$

Dans la première formule, le nombre de termes de la somme est égal à $2k + 1$: il s'agit bien d'une moyenne. Dans la seconde, la somme des coefficients est égale à $2k$, puisque le premier et le dernier sont égaux à 0.5 : il s'agit d'une moyenne pondérée. Dans les deux cas, le nombre d'observations prise en compte avant l'instant t est égal au nombre d'observations prises en compte après l'instant t : c'est pour cela que les moyennes sont dites centrées.

La première valeur d'une moyenne mobile de longueur 4 ($= 2 \times 2$) ou 5 ($= 2 \times 2 + 1$) que l'on peut calculer, est à l'instant $t = 3$, puisque la première observation connue est x_1 :

$$mm_3 = (0.5 x_1 + x_2 + x_3 + x_4 + 0.5 x_5) / 4 \quad (l = 4)$$

$$mm_3 = (x_1 + x_2 + x_3 + x_4 + x_5) / 5 \quad (l = 5)$$

De façon générale, ne peut calculer de moyenne mobile en $t = 1, t = 2, \dots, t = k$ puisque les formules ne peuvent être appliquées que si l'on connaît x_{t-k} . De même, si T est le nombre total d'observations, on ne peut calculer $mm_T, \dots, mm_{T-k+1}$ puisqu'il faut connaître x_{t+k} .

L'avantage des moyennes mobiles est d'atténuer la composante accidentelle tout en conservant les tendances linéaires : la série est dite « lissée », et est d'autant plus lissée que la longueur de la moyenne mobile est élevée comme on peut le constater sur la figure 3.8 sur laquelle nous avons représenté les moyennes mobiles de longueur 14.

Exemple: soit le tableau suivant montrant un extrait des moyennes mobiles de longueur 5 et les représentations graphiques sur les figures 2.8 et 3.8 de l'évolution du taux de fret et des moyennes mobiles de longueur 5 et 14

Instant	cours (€)	moyenne mobile	Instant	cours (€)	moyenne mobile
1	109.50000		42	100.00000	103.19000
2	113.20000		43	107.05000	106.98000
3	119.70000	117.53000	44	112.90000	
4	122.35000	119.28000	45	117.40000	

On notera que si dans les cours du taux fret, les moyennes mobiles ne sont pas centrées, on utilise les moyennes des 50 ou 100 dernières observations avant l'instant t pour définir la tendance à l'instant t

On notera que, dans les journaux financiers, les moyennes mobiles ne sont pas centrées : on utilise les moyennes des 50 ou 100 dernières observations avant l'instant t pour définir la tendance à l'instant t .

L'inconvénient des moyennes mobiles de longueur 14 est qu'elles ne sont définies qu'à partir de la 8^e observation et jusqu'à la 38^e. On ne dispose d'aucune information sur la tendance ni au début ni à la fin de la période d'observation. Il faut donc choisir la longueur des

moyennes mobiles suivant le nombre d'observations et l'objectif de l'analyse. Nous verrons que la longueur de ces m.m. dépend aussi de la période des variations saisonnières pour faire apparaître la tendance.

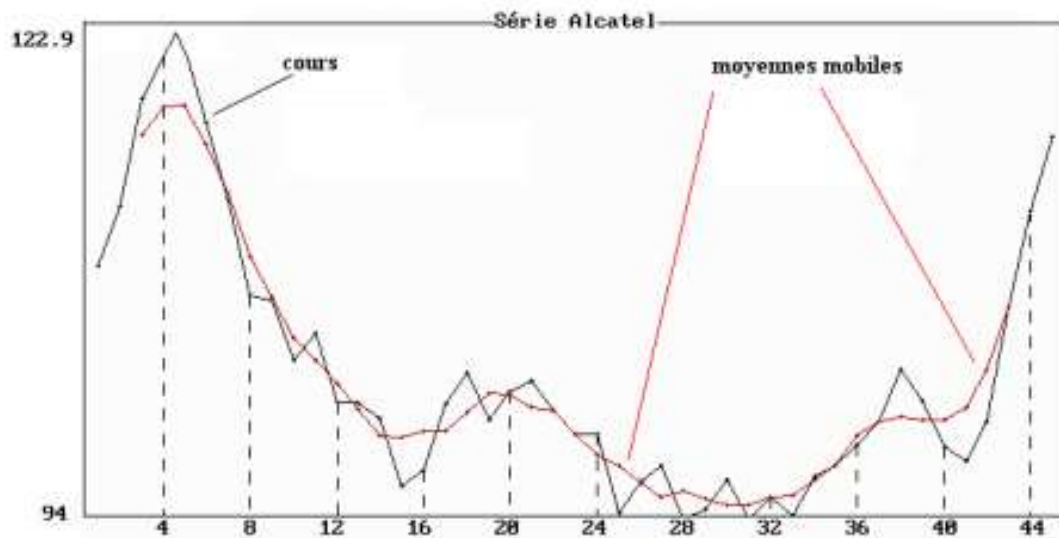


Figure 2.8: Moyennes mobiles de longueur 5



Figure 3.8: Moyennes mobiles de longueur 14

Théorème : si la tendance d'une série chronologique x_t est linéaire et a pour équation $c_t = \beta t + \alpha$, les moyennes mobiles centrées ont pour tendance la même droite et en sont d'autant plus proches que la longueur des moyennes mobiles est élevée.

Ce théorème est démontré dans un complément pédagogique. Il peut être complété par l'étude de tendance de la forme $c_t = \beta_2 t^2 + \beta_1 t + \alpha$ proposée en application pédagogique.

1.2 Description simultanée de variation saisonnière et de la tendance

Une variation saisonnière est caractérisée par le fait qu'elle se produit à intervalles de temps réguliers, d'où d'ailleurs le terme saisonnier.

Définition : on appelle variation saisonnière d'une série chronologique à l'instant t une variation due à un effet momentané se reproduisant régulièrement dans le temps.

Exemple : nous étudions la série chronologique suivante observée trimestriellement pendant 6 ans (tableau 3.8).

	1 ^{er} trimestre	2 ^e trimestre	3 ^e trimestre	4 ^e trimestre
Année 1	89.658	97.593	108.906	114.157
Année 2	96.205	99.399	112.763	119.185
Année 3	99.602	105.192	116.556	121.911
Année 4	103.272	109.644	121.208	126.508
Année 5	105.637	113.428	125.641	131.147
Année 6	111.118	117.215	129.776	133.000

Tableau 3.8 : série chronologique 1
(période $p = 4$)

L'observation de chaque trimestre est soumise à un effet particulier qui revient tous les ans ; il y a donc 4 variations saisonnières correspondant chacune à un trimestre.

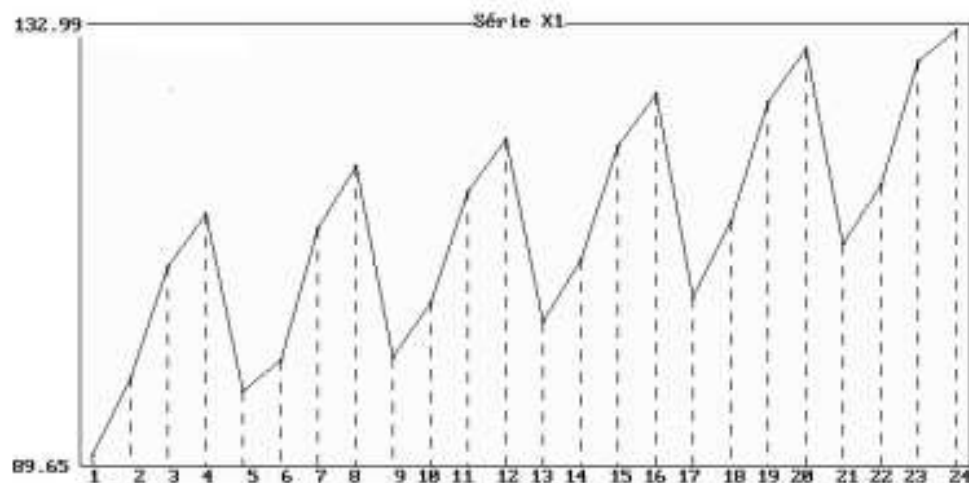


Figure 4.8 : représentation graphique de la série 1
(données observées trimestriellement pendant 6 ans)

Définition : la période notée p des variations saisonnières est la longueur exprimée en unités de temps séparant deux variations saisonnières dues à un même phénomène.

Remarque : nous supposons dans la suite que la série est soumise à des variations saisonnières de même période p . La période est alors le nombre de variations saisonnières. Cette hypothèse n'est pas toujours réalisée au départ : les ventes par tranches horaires d'un hypermarché sont soumises par exemple à une première variation saisonnière due à l'heure et à une seconde due à la journée. Ce cas est traité théoriquement en considérant une période égale au plus petit commun multiple des deux périodes : deux variations saisonnières de périodes 4 et 6 donnent une variation saisonnière de période 12 ($= 3 \times 4 = 2 \times 6$).

Il n'est pas toujours facile de distinguer la tendance lorsque la série chronologique est soumise à des variations saisonnières. La méthode mathématique consiste à calculer les moyennes mobiles en choisissant comme longueur la période des variations saisonnières, de façon à les faire disparaître. Si la moyenne mobile choisie est de longueur différente, les variations saisonnières ne sont pas toujours éliminées (cf. figures 5.8 et 6.8).

Ces moyennes mobiles ont en outre l'avantage d'atténuer les variations accidentelles comme nous l'avons vu précédemment, mais l'inconvénient de n'être définies ni au début ni à la fin de la période observée.

Exemple : on pourra comparer sur les figures 5.8 et 6.8 ci-dessous les moyennes mobiles de longueur 4 (dont les valeurs numériques sont données dans le tableau 4.8) et de longueur 5.

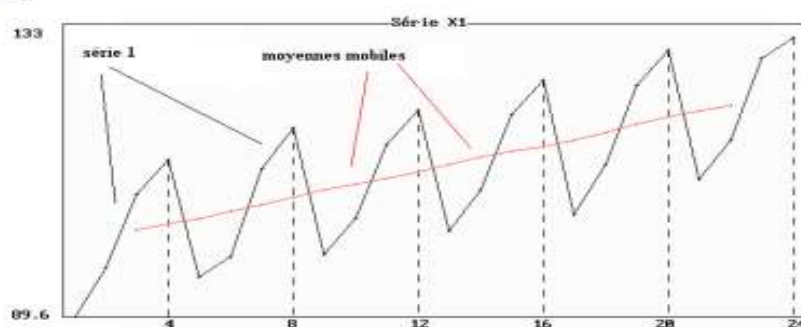


Figure 5.8 : représentation simultanée de la série 1 et des moyennes mobiles de longueur 4

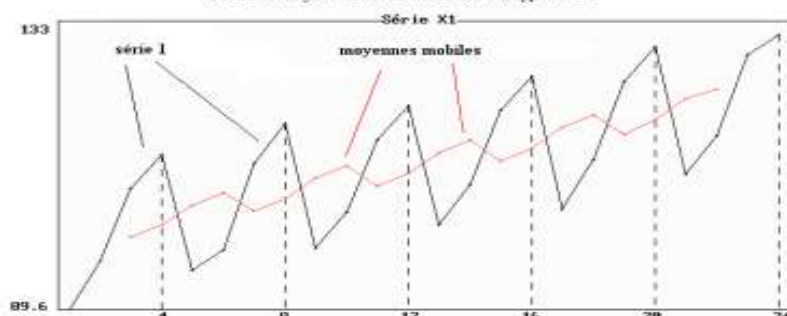


Figure 6.8: ...de longueur 5

Contrairement aux m.m. de longueur 4, les m.m. de longueur 5 n'éliminent pas les variations saisonnières.

Théorème : les moyennes mobiles d'une série soumise à des variations saisonnières de période p ne sont pas soumises à ces variations saisonnières si leur longueur l est égale à la période p , et plus généralement si leur longueur est un multiple de la période.

Ce théorème est démontré dans un complément pédagogique.

Conclusion : les moyennes mobiles d'une série chronologique dont la tendance est linéaire et les variations saisonnières sont de période p font apparaître la tendance et disparaître les variations saisonnières si leur longueur l est égale à la période p .

2. Modélisation et désaisonnalisation

Un modèle de série chronologique est une équation précisant la façon dont les composantes s'articulent les unes par rapport aux autres pour constituer la série chronologique. Il existe de très nombreux modèles, et parmi eux deux modèles classiques simples : le modèle additif et le modèle multiplicatif, auxquels nous nous limiterons.

Dans les deux modèles présentés, la longueur des moyennes mobiles doit être impérativement égale à la période des variations saisonnières.

Nous avons présenté dans le tableau 3.8 les données sous une forme particulière : en lignes, ce sont les années, et en colonnes les trimestres : le terme x_t correspondant à la t^{e} observation est alors noté $x_{i,j}$, i donnant l'année (la ligne) et j le trimestre (la colonne).

La relation entre les indices i et j d'une part et l'instant t d'autre part est la suivante :

$$t = (i-1)p + j$$

Exemple :

	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$i = 1$	$t = 1$	$t = 2$	$t = 3$	$t = 4$
$i = 2$	$t = 5$	$t = 6$	$t = 7$	$t = 8$
$i = 3$	$t = 9$	$t = 10$	$t = 11$	$t = 12$

Exemple pour $n = 3$ et $p = 4$

Nous utiliserons cette notation très souvent dans la suite du texte.

2.1 Modèle additif de série chronologique

La série chronologique x_t se décompose en une tendance notée c_t , des variations saisonnières s_t de période p (égales à $s_1, s_2, s_3, \dots, s_p$) et d'une composante accidentelle e_t .

On peut en déduire les différences $x_{i,j} - mm_{i,j}$. Les moyennes mobiles donnant une première approximation de la tendance $c_{i,j}$, les colonnes du tableau des différences contiennent des approximations des coefficients s_j .

Le modèle le plus simple est le modèle additif, dans lequel la variation saisonnière s'ajoute simplement à la tendance :

pour tout $t = 1, \dots, T$	$x_t = c_t + s_t + e_t$
-----------------------------	-------------------------

Le modèle additif s'exprime donc en général de la façon suivante :

pour tout $i = 1, \dots, n$	pour tout $j = 1, \dots, p$	$x_{i,j} = c_{i,j} + s_j + e_{i,j}$
-----------------------------	-----------------------------	-------------------------------------

Le terme s_j caractérise la variation saisonnière à l'instant j de chaque période i : du trimestre j dans le cas particulier des séries 1 et 2 ($p = 4$), du mois j dans des données mensuelles ($p = 12$) etc.... Les moyennes mobiles seront aussi notées $mm_{i,j}$.

Définition : les termes s_j du modèle additif exprimé sous la forme précédente sont appelés coefficients saisonniers du modèle additif.

On peut calculer la différence entre l'observation et la tendance :

pour tout $i = 1, \dots, n$	pour tout $j = 1, \dots, p$	$x_{i,j} - c_{i,j} = s_j + e_{i,j}$
-----------------------------	-----------------------------	-------------------------------------

Pour un même trimestre, la différence entre l'observation et la tendance est donc à peu près constant et égale à s_j (on suppose que la composante accidentelle est relativement faible).

Nous avons vu précédemment que les moyennes mobiles de longueur l égale à la période des variations saisonnières sont des approximations de la tendance. On peut donc considérer que la différence entre une observation $x_{i,j}$ et la moyenne mobile $mm_{i,j}$ correspondante est à peu près constante pour j fixé :

pour tout $i = 1, \dots, n$	pour tout $j = 1, \dots, p$	$x_{i,j} - mm_{i,j} \approx s_j$
-----------------------------	-----------------------------	----------------------------------

Cette propriété est recherchée sur la représentation graphique de la série x_t pour déterminer si cette série suit un modèle additif ou non. Elle peut être observée sur la figure 5.8 dans laquelle la tendance est caractérisée par les moyennes mobiles de longueur 4 : les différences entre x_3 et mm_3 , entre x_7 et mm_7 , entre x_{11} et c_{11} sont à peu près constantes, de même les différences entre x_4 et mm_4 , x_8 et mm_8 , x_{12} et mm_{12} etc.

Exemple : Les moyennes mobiles et par suite différences $x_{i,j} - mm_{i,j}$ ne sont pas définies aux premier et deuxième trimestres de la première année, ni aux troisième et quatrième trimestres de la dernière (tableaux 4.8 et 5.8).

	1 ^{er} trimestre	2 ^e trimestre	3 ^e trimestre	4 ^e trimestre
Année 1			103.39678	104.44080
Année 2	105.14860	106.25917	107.31233	108.46116
Année 3	109.65950	110.47448	111.27404	112.28937
Année 4	113.42748	114.58360	115.45379	116.22236
Année 5	117.24943	118.38337	119.64839	120.80691
Année 6	121.79719	122.54573		

Tableau 4.8 : moyennes mobiles de longueur 4 de la série 1

	1 ^{er} trimestre	2 ^e trimestre	3 ^e trimestre	4 ^e trimestre
Année 1			5.50932	9.71580
Année 2	-8.94393	-6.86050	5.45047	10.72334
Année 3	-10.05746	-5.28258	5.28226	9.62153
Année 4	-10.15538	-4.93910	5.75471	10.28534
Année 5	-11.61263	-4.95497	5.99271	10.33979
Année 6	-10.67929	-5.33023		

Tableau 5.8 : différences entre les observations et les moyennes mobiles de la série 1

Les différences apparaissant dans une même colonne sont proches les uns des autres et caractérisent le modèle additif.

Les différences $x_{i,j} - mm_{i,j}$ sont donc des approximations des coefficients s_j . Leur moyenne (ou leur médiane), pour chaque colonne j , donne une première estimation s_j' :

$$s_j' = \frac{1}{n} \sum_{i=1}^n (x_{i,j} - mm_{i,j})$$

On obtiendra enfin les estimations définitives s_j en centrant ces termes s_j' :

- on calcule la moyenne des s_j' :

$$m_s' = \frac{1}{p} (s_1' + s_2' + \dots + s_p')$$

- on centre en posant :

$$\text{pour tout } j = 1, \dots, p \quad s_j = s_j' - m_s'$$

Exemple : le tableau 5.8 donne les différences entre les observations et les moyennes mobiles. On en déduit les moyennes suivantes :

$s_1' = -10.2897$	$s_2' = -5.4735$	$s_3' = 5.5979$	$s_4' = 10.1371$
-------------------	------------------	-----------------	------------------

- On calcule la moyenne des s_j' : $m_s' = -0.007039938$
- Les valeurs définitives sont obtenues en posant $s_j = s_j' - m_s'$:

$s_1 = -10.2827$	$s_2 = -5.4664$	$s_3 = 5.6049$	$s_4 = 10.1442$
------------------	-----------------	----------------	-----------------

règle de calcul des estimations des coefficients saisonniers du modèle additif

- on calcule les différences entre les observations et les moyennes mobiles ;
- on calcule la moyenne ou la médiane s_j' des différences de chaque colonne du tableau ;
- on calcule la moyenne m_s' de ces valeurs s_j' ;
- on obtient les estimations s_j en centrant les valeurs s_j' : $s_j = s_j' - m_s'$.

2.2 Modèle multiplicatif de série chronologique

2.2 Modèle multiplicatif de série chronologique.

Le second modèle que nous étudions ici est le modèle multiplicatif suivant :

$$\text{pour tout } t = 1, \dots, T \quad x_t = c_t (1 + s_t) + e_t$$

En présentant les données comme dans le paragraphe précédent, le modèle multiplicatif s'exprime de la façon suivante :

$$\text{pour tout } i = 1, \dots, n \quad \text{pour tout } j = 1, \dots, p \quad x_{i,j} = c_{i,j} (1 + s_j) + e_{i,j}$$

Le terme s_j caractérise la variation saisonnière du trimestre j dans le cas particulier des séries 1 et 2, du mois j dans des données mensuelles etc.

On peut calculer la différence entre l'observation et la tendance :

$$\text{pour tout } i = 1, \dots, n \quad \text{pour tout } j = 1, \dots, p \quad x_{i,j} - c_{i,j} = c_{i,j} s_j + e_{i,j}$$

Considérons le cas particulier $j = 1$ (1^{er} trimestre de l'année i).

$$\text{pour tout } i = 1, \dots, n \quad x_{i,1} - c_{i,1} = c_{i,1} s_1 + e_{i,1}$$

La différence $x_{i,j} - c_{i,j}$ entre l'observation et la tendance est proportionnelle à la tendance $c_{i,j}$: lorsque cette tendance est croissante, la différence augmente, lorsqu'elle est décroissante, il diminue.

Le même raisonnement peut évidemment être tenu pour j fixé quelconque. Les différences permettent ainsi de déterminer si la série chronologique étudiée suit un modèle multiplicatif.

Exemple : on considère la série chronologique ci-dessous :

	1 ^{er} trimestre	2 ^e trimestre	3 ^e trimestre	4 ^e trimestre
Année 1	224.3705	253.2811	201.2421	248.9411
Année 2	274.3802	300.1641	248.9038	298.4386
Année 3	331.9657	371.4032	303.4313	365.9029
Année 4	406.6326	437.9967	361.5774	444.8447
Année 5	488.4166	536.5268	435.5698	549.3614
Année 6	598.0016	659.2896	533.2156	669.2675

Tableau 6.8 : série chronologique 2
(modèle multiplicatif, période $p = 4$)

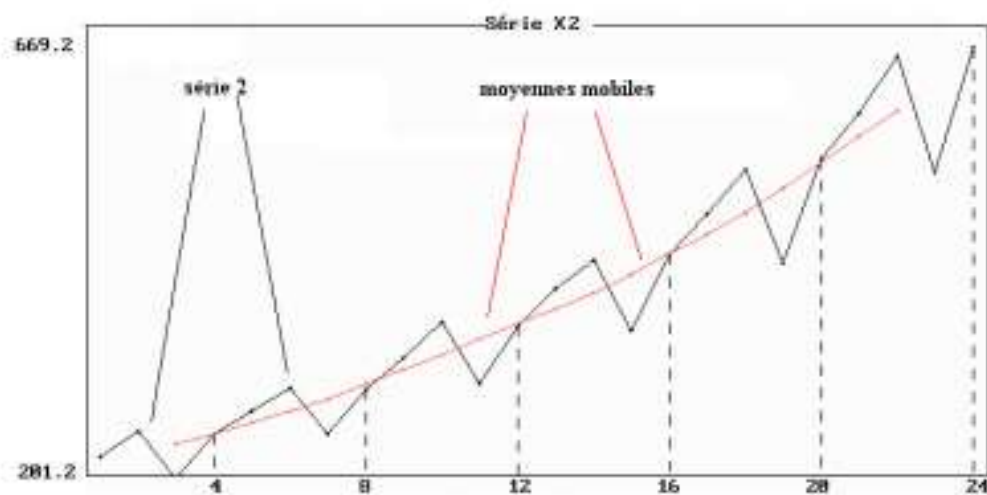


Figure 7.8 : série 2 et moyennes mobiles de longueur 4

Cette série est soumise à des variations saisonnières de période 4 ; la tendance, caractérisée par les moyennes mobiles de longueur 4, est croissante, et la différence entre une observation x_i et la moyenne mobile mm_i a tendance à augmenter pour une même variation saisonnière : 1 Les différences entre x_3 et mm_3 , entre x_7 et mm_7 , entre x_{11} et c_{11} augmentent visiblement, de même que les différences entre x_4 et mm_4 , x_8 et mm_8 , x_{12} et mm_{12} etc. (figure 7.8).

Pour quantifier les variations saisonnières, on considère les rapports $x_{i,j} / c_{i,j}$:

$$\text{pour tout } i = 1, \dots, n \text{ pour tout } j = 1, \dots, p \quad x_{i,j} / c_{i,j} = 1 + s_j + e_{i,j} / c_{i,j}$$

En considérant que les variations accidentelles $e_{i,j}$ sont faibles par rapport à la tendance $c_{i,j}$ et en utilisant l'approximation de la tendance par les moyennes mobiles, on constate donc que les rapports $x_3 / mm_3, x_7 / mm_7, x_{11} / mm_{11}, \dots$ sont à peu près constants et donnent une approximation de $1 + s_3$, de même les rapports $x_4 / mm_4, x_8 / mm_8, x_{12} / mm_{12}$ etc. donnent une approximation de $1 + s_4$:

$$\text{pour tout } j = 1, \dots, p \quad \frac{x_{i,j}}{mm_{i,j}} = 1 + s_j = S_j$$

Exemple : les tableaux 7.8 et 8.8 ci-dessous contiennent les moyennes mobiles de la série et les rapports $x_{i,j} / mm_{i,j}$.

	1 ^{er} trimestre	2 ^e trimestre	3 ^e trimestre	4 ^e trimestre
Année 1			238.210	250.322
Année 2	262.140	274.284	287.670	303.773
Année 3	319.494	334.743	352.509	370.167
Année 4	385.759	402.895	422.986	445.525
Année 5	467.090	489.404	516.167	545.210
Année 6	572.761	599.955		

Tableau 7.8 : moyennes mobiles de longueur 4 de la série 2

	1 ^{er} trimestre	2 ^e trimestre	3 ^e trimestre	4 ^e trimestre
Année 1			0.84481	0.99449
Année 2	1.04670	1.09435	0.86524	0.98244
Année 3	1.03904	1.10952	0.86078	0.98848
Année 4	1.05411	1.08712	0.85482	0.99847
Année 5	1.04566	1.09629	0.84385	1.00761
Année 6	1.04407	1.09890		

Tableau 8.8 : rapports des observations aux moyennes mobiles de la série 2

Les rapports dans chaque colonne du tableau 8.8 sont à peu près constants.

Les rapports $x_{i,j} / mm_{i,j}$ sont donc des approximations des termes $1 + s_j$ que l'on appelle coefficients saisonniers dans le cas du modèle multiplicatif.

Définition : les termes $S_j = 1 + s_j$ du modèle multiplicatif exprimé sous la forme précédente sont appelés coefficients saisonniers du modèle multiplicatif.

On obtient des premières estimations S_j' des coefficients saisonniers en calculant la moyenne (ou la médiane) des rapports figurant dans chaque colonne. Par analogie avec les coefficients saisonniers s_j du modèle additif, dont la moyenne est égale à 0, on cherche des estimations définitives S_j de moyenne 1 :

- on calcule la moyenne :

$$m_{S'} = \frac{1}{p} (S_1' + S_2' + \dots + S_p')$$

- on pose :

$$\text{pour tout } j = 1, \dots, p \quad S_j = S_j' / m_{S'}$$

Les coefficients saisonniers estimés S_j sont ainsi de somme p :

$$\begin{aligned} S_1 + S_2 + \dots + S_p &= \frac{1}{m_{S'}} (S_1' + S_2' + \dots + S_p') \\ &= p \end{aligned}$$

ce qui équivaut à une moyenne des s_j égale à 0 puisque l'on a $S_j = 1 + s_j$.

Exemple :

- le tableau 8.8 donne les rapports des observations aux moyennes mobiles.
- on en déduit les moyennes suivantes :

$S_1' = 1.045913$	$S_2' = 1.097236$	$S_3' = 0.8539006$	$S_4' = 0.9942986$
-------------------	-------------------	--------------------	--------------------

- on calcule la moyennes des S_j' : $m_{S'} = .9978371$
- les valeurs définitives sont obtenues de façon que les S_j' soient de moyenne 1 :

$S_1 = 1.04818$	$S_2 = 1.099614$	$S_3 = 0.8557515$	$S_4 = 0.9964539$
-----------------	------------------	-------------------	-------------------

règle de calcul des estimations des coefficients saisonniers du modèle multiplicatif	
•	on calcule les rapports des observations aux moyennes mobiles ;
•	on calcule la moyenne ou la médiane des rapports S_j' de chaque colonne du tableau ;
•	on calcule la moyenne $m_{S'}$ de ces valeurs ;
•	on obtient les estimations S_j en posant $S_j = S_j' / m_{S'}$.

2.3 Désaisonnalisation

Les coefficients saisonniers permettent d'éliminer d'une observation les effets de la variation saisonnière correspondante. On obtient ainsi les valeurs corrigées des variations saisonnières, ou encore les valeurs désaisonnalisées.

L'avantage de cette désaisonnalisation est de permettre la comparaison de deux observations soumises à des variations saisonnières différentes.

définition : on appelle observation corrigée des variations saisonnières la valeur $x_{i,j}'$ obtenue en éliminant l'effet saisonnier sur la valeur $x_{i,j}$.

modèle additif :	$x_{i,j}' = x_{i,j} - s_j$
modèle multiplicatif :	$x_{i,j}' = x_{i,j} / S_j$

Les valeurs corrigées des variations saisonnières (expression souvent abrégée par c.v.s.) caractérisent à la fois la tendance et la variation accidentelle.

Exemple : on donne ci-dessous les quatre dernières observations de la série 2 (année 6) et les valeurs corrigées des variations saisonnières :

	1 ^{er} trimestre	2 ^e trimestre	3 ^e trimestre	4 ^e trimestre
observations :	598.00160	659.28960	533.21560	669.26750
valeur c.v.s. :	570.51396	599.56452	623.09629	671.64924

L'observation du deuxième trimestre est largement supérieure à celle du troisième, mais c'est l'inverse pour les valeurs c.v.s. : la tendance est restée croissante au troisième trimestre.

Supposons que l'observation du premier trimestre de l'année 7 soit égale à 720.15. Pour savoir si la tendance est restée à la hausse, on calcule la valeur désaisonnalisée :

$$x_{7,1}' = 720.15 / 1.04818 = 687.04771$$

et on la compare à la valeur désaisonnalisée du quatrième trimestre de l'année précédente :

$$x_{6,4}' = 671.649$$

La valeur c.v.s. $x_{7,1}'$ est supérieur à $x_{6,4}'$. La tendance est restée à la hausse si la différence est supérieure à la variation accidentelle. Il faudrait donc comparer cette différence à

l'écart type des variations accidentelles, calculé sur les données antérieures. Il semble que dans la pratique, cette comparaison ne soit guère effectuée.

3. Validation du modèle linéaire et prévision

Pour effectuer la prévision $x_{i,j}^p$ de la série chronologique à l'instant $t = (i-1)p + j$, on remplace dans le modèle théorique les paramètres par leurs estimations :

$$x_{i,j}^p = b[(i-1)p + j] + a + s_j$$

Ces prévisions ponctuelles peuvent être complétées par des prévisions par intervalle de confiance, comme en régression, mais nous ne donnerons pas les formules trop complexes pour être utilisées facilement.

Exemple : Les valeurs estimées par le modèle sont données par la formule :

$$x_{i,j} = 1.011173[(i-1)4 + j] + 100.2237 + s_j$$

avec :

$s_1 = -10.43134$	$s_2 = -5.27912$	$s_3 = 5.77288$	$s_4 = 9.93759$
-------------------	------------------	-----------------	-----------------

Les prévisions concernant l'année 7 sont les suivantes :

$x_{7,1}^p = 115.07169$	$x_{7,2}^p = 121.23509$	$x_{7,3}^p = 133.29826$	$x_{7,4}^p = 138.47414$
-------------------------	-------------------------	-------------------------	-------------------------

Les prévisions que l'on peut effectuer après l'estimation des paramètres ne sont justifiées que dans la mesure où les hypothèses du modèle sont respectées.

Une hypothèse fondamentale pour la prévision et souvent négligée est que les conditions dans lesquelles la série chronologique évolue sont les mêmes à la date de la prévision que dans le passé. Il faut noter que cette condition n'est pas toujours vérifiée, par l'effet de la prévision elle-même : par exemple, un hypermarché qui prévoit une baisse de son chiffre d'affaires va prendre des mesures de réduction de coût, augmenter sa publicité etc., de façon à augmenter ses ventes : les décisions vont donc à l'encontre de la prévision.

Il est indispensable en outre de contrôler statistiquement le modèle. Pour cela, on étudie les résidus $e_{i,j}$. Ces résidus possèdent les propriétés mathématiques habituelles en régres-

sion : ils sont centrés, indépendants de la variable explicative (ici le temps). Les résidus de chaque trimestre sont en outre de moyenne nulle.

Pour que le modèle soit valide, leur répartition doit être proche de la loi normale. Ils ne doivent présenter pas d'évolution particulière dans le temps, ce que l'on peut vérifier par une représentation graphique ou par des tests (tests de Durbin et Watson, sur le coefficient d'autocorrélation de rang 1).

Rappelons enfin que la prévision $x_{t,j}^p$ est la prévision de la moyenne des observations pour t fixé : l'intervalle de confiance donné par certains logiciels est celui de cette moyenne, et non de la valeur individuelle à l'instant t .

Remarque: La démarche d'approximation peut également être menée par Lissage exponentiel

CONCLUSION

L'analyse des séries chronologiques est un des objectifs fondamentaux de la statistique. Nous insistons sur le fait, que quelle que soit la méthode utilisée, il faut être vigilant sur les prévisions effectuées qui peuvent être dans certains cas totalement aberrantes (cf. Bensaï et Bleuse-Trillon, p. 123, à propos du modèle de Holt et Winters).

Nous conseillons fortement aux praticiens de se limiter aux méthodes qu'ils connaissent lorsqu'ils effectuent leurs prévisions. Le logiciel qu'ils utilisent doit être sûr, contrôlé : il nous est arrivé d'obtenir des résultats différents sur les mêmes données et par la même méthode en employant deux logiciels différents, ou d'aboutir des prévisions manifestement fausses puisque différentes des valeurs données dans les ouvrages de référence.

CHAPITRE II : REGRESSION MULTIPLE - ÉCONOMETRIE (initiation) – M2 spécialement

I. NOTIONS DE BASE

"L'économétrie est un outil d'analyse quantitative, permettant de vérifier l'existence de certaines relations entre des phénomènes économiques et de mesurer concrètement ces relations sur la base d'observations de faits réels" (Eric Dor, "Econométrie"). L'économétrie se base donc sur les mathématiques et les statistiques afin d'identifier des relations entre différentes variables.

1. Modèle de Régression à une variable explicative

Soit un modèle simple à deux variables. Vous souhaitez tester s'il existe une relation linéaire entre les variables y et x . Pour le poser sous la forme d'une équation, votre but est d'estimer les coefficients ? de l'équation ci-dessous (avec un terme d'erreur mais qu'on ne présente pas dans cette intro pour plus de simplicité).

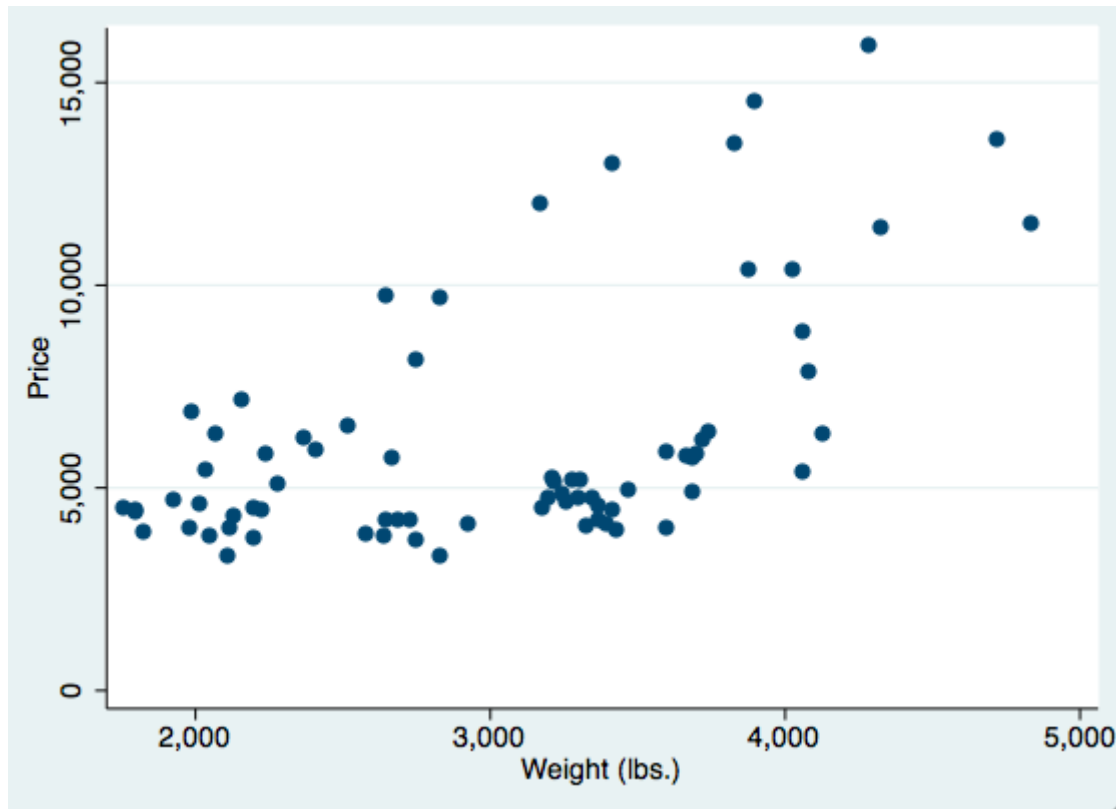
$$y = \beta_0 + \beta_1 x$$

Pour prendre un exemple, supposons que vous ayez envie de voir le lien entre le poids d'un véhicule de manutention portuaire et son prix de vente. Avant de faire votre étude, vous avez donc créé une base de données avec de nombreuses infos sur les voitures, qui pourrait ressembler à celle-ci (si vous avez envie de vous amuser avec cette base, elle est disponible au format Stata ici "auto.dta").

Tableau : auto.dta

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	2.56
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	35	98	3.54
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	46	318	2.47
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206	46	318	2.47
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220	46	225	2.94
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147	33	98	3.15

En réalisant un graphique avec en abscisse le poids d'une voiture (weight) et son prix (price) pour chacune des 74 voitures de notre base de données, l'on obtient, chaque point correspondant à une voiture :



Il semble bien qu'il y ait une relation croissante entre le poids d'une voiture et son prix. Graphiquement on remarque en effet une tendance : plus une voiture est lourde, plus son prix est élevé (en moyenne), mais il faut des chiffres précis afin d'être sûr que cette relation est significative et n'est pas le fruit du hasard".

Il suffit d'un seul calcul dans un logiciel spécialisé (type Stata, Eviews, Scilab, SAS...) pour avoir l'ensemble des réponses aux questions : La "régression linéaire et la méthode des moindres carrés ordinaire". Globalement, le but de cette technique est de tracer une droite sur le graphique précédent, tel que l'écart (au carré) entre les points du nuage et la droite de régression soit le plus faible possible.

L'on peut tâtonner en calculant pour chaque point, l'écart entre les points-nuage et la droite, ou bien utiliser une fonction automatique sur un logiciel économétrique. Dans Stata (logiciel d'analyse de données), la commande "regress price weight" indique que vous souhaitez faire une régression linéaire entre la variable y (price) et la variable x (weight).

. regress price weight Source: CaptainEconomics.fr

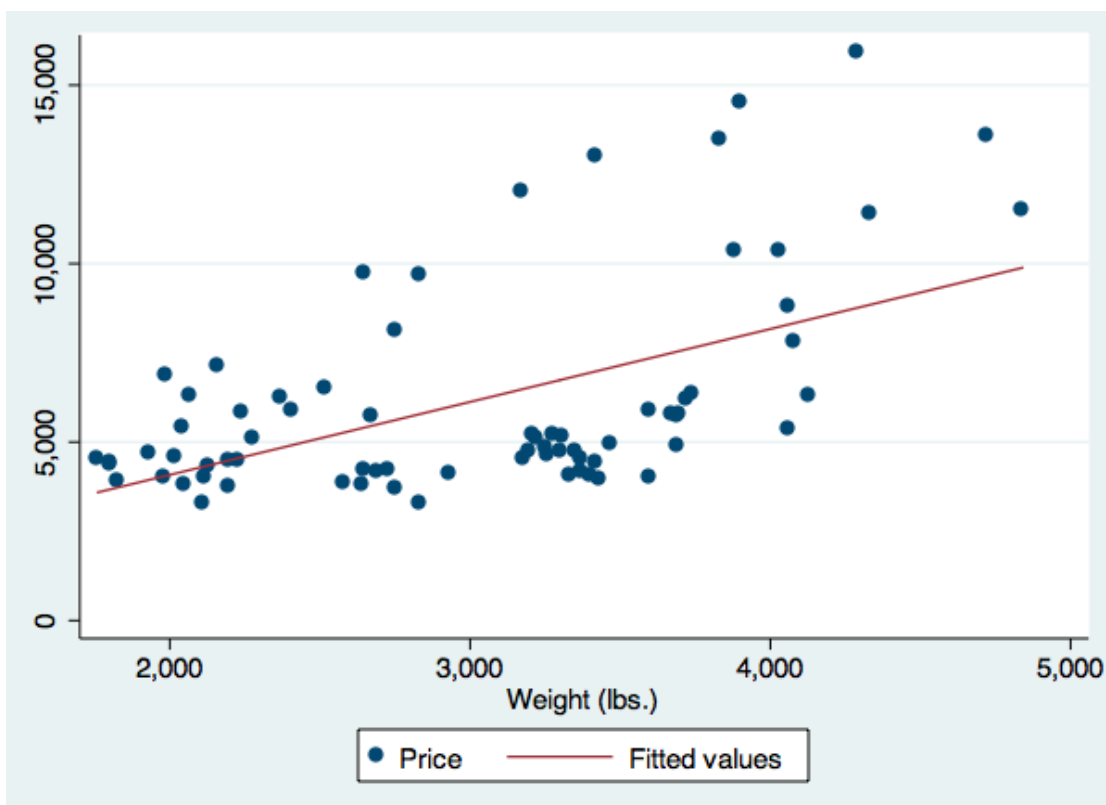
Source	SS	df	MS	Number of obs = 74		
Model	184233937	1	184233937	F(1, 72) = 29.42		
Residual	450831459	72	6261548.04	Prob > F = 0.0000		
				R-squared = 0.2901		
				Adj R-squared = 0.2802		
Total	635065396	73	8699525.97	Root MSE = 2502.3		

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
weight	2.044063	.3768341	5.42	0.000	1.292857	2.795268
_cons	-6.707353	1174.43	-0.01	0.995	-2347.89	2334.475

Le but était donc d'estimer la relation entre le poids d'une voiture et son prix, en déterminant les coefficients β_0 et β_1 de l'équation initiale. Pour rappel, le coefficient β_0 va correspondre à l'ordonnée à l'origine et le coefficient β_1 à la pente de la droite de régression. Sur le tableau ci-dessus, β_0 correspond au coefficient de la constante (_cons) en bleu, soit -6,70 et β_1 correspond au coefficient devant le poids (weight) en rouge, soit 2,05.

L'équation initiale peut donc s'écrire: " **price = -6,70 + 2,05 * weight** "

En moyenne donc, et en se basant sur les données de l'échantillon, une voiture de 2000 livres (lbs) coûtera environ "-6,70 + 2,05 * 2000 = 4000 dollars). Si l'on reprend le graphique précédent, on peut alors tracer la droite de régression linéaire, qui est la droite la plus proche possible des différents points. On voit bien que si l'on se base sur une voiture de 2000 lbs, son prix estimé via la droite rouge est proche de 4000 dollars.



Mais cette relation est-elle significative ?

Pour cela, il convient de regarder le "**t-stat**" du tableau de régression (en vert), et de comparer la valeur de ce "t-stat" (dans notre cas 5,42) à des valeurs déterminées statistiquement. Il est possible de ne pas rentrer dans les calculs, et uniquement de se baser sur cette règle "si le t-stat est supérieur à 1,96 en valeur absolue, alors la variable est significative". Ici c'est le cas, donc, **une voiture qui pèse 1 livre de plus qu'une autre coûtera en moyenne 2,05 dollars de plus, et la relation est significative**".

Cette relation est significative avec un intervalle de confiance de 95%", la valeur de t-stat de 1,96 correspondant à cet intervalle de confiance.

Toutefois, on voit bien qu'il y a de gros écarts entre la droite de régression en rouge et de nombreux points. Ceci s'explique car le modèle est ultra-simplifié (une seule variable explicative). Il est d'ailleurs possible d'estimer le pourcentage de la variation de prix expliqué par la variation de poids, en regardant le R-squared et l'adjusted R-squared (en jaune dans le tableau). Dans notre exemple, on trouve que 29% de la variation de prix peut s'expliquer par une variation de poids (et donc que 71% de la variation de prix provient de variables omises dans notre modèle ou bien de facteurs aléatoires).

Conclusion: Notre exemple peut mettre en avant un premier piège de l'économétrie: la différence entre causalité et de corrélation. Par exemple, si vous rajoutez discrètement des parpaings dans le coffre de votre voiture, le prix devrait selon notre modèle augmenté (ce qui n'a aucun sens). La relation estimée entre le poids d'une voiture et son prix est davantage une relation de corrélation que de causalité. Ce n'est pas le poids à proprement parlé qui entraîne une hausse du prix, c'est le fait que le poids est souvent relié à d'autres variables (taille du moteur, nombre de portes, confort...) qui elles ont réellement une relation de causalité avec le prix d'une voiture.

2. Modèle de Régression à plusieurs variables explicatives

Le modèle tente d'expliquer le prix d'une voiture (notre variable dépendante "Y") en fonction de variables explicatives (X_1 , X_2 , D_1). La variable X_1 correspond comme au poids d'une véhicule de manutention. La variable X_2 représente la consommation en carburant de la voiture et la variable D_1 est une variable indicatrice, prenant la valeur 0 si la voiture est une voiture domestique (donc américaine dans notre exemple) ou 1 si la voiture est une voiture étrangère. Le modèle s'écrit ainsi:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 D_1 + \varepsilon$$

Le dernier terme ε correspond au terme d'erreur, qui représente la déviation entre ce que le modèle prédit et la réalité. Comme précédemment, il s'agira de déterminer (i) les variables significatives, c'est-à-dire voir si les différents coefficients sont différents de 0, (ii) la valeur de la constante 'alpha' et des différents coefficients "beta" qui permettent de

minimiser l'erreur entre notre droite de régression linéaire estimée et les valeurs réelles de Y et enfin (iii) la précision de notre modèle, en utilisant, entre autre, le "R-squared".

Le raisonnement est le même qu'avec seulement une variable, sauf qu'il est difficile de travailler graphiquement avec plusieurs variables. En effet, une régression linéaire à une variable explicative peut s'expliquer dans un graphique en 2D (avec en abscisse la variable X et en ordonnée la variable Y). Lorsque l'on passe à 3 variables explicatives, il faudrait montrer cela dans un environnement en 4 dimensions (une dimension par variable explicative + une dimension pour la variable dépendante).

L'introduction d'une variable indicatrice "D₁" permet de travailler avec des variables qualitatives à la place de variables quantitatives, en codant la variable qualitative sous la forme binaire (par exemple "homme = 0 et femme = 1). C'est une variable qui permet de voir l'impact de la provenance d'une voiture sur son prix. Par exemple, et étant donné que D₁ = 0 si la voiture est américaine et D₁ = 1 si la voiture est étrangère, une valeur positive du coefficient β_3 signifierait simplement que les voitures étrangères coûtent en moyenne plus cher que les voitures américaines.

. regress price weight mpg foreign Source: CaptainEconomics.fr

Source	SS	df	MS	Number of obs =	74
Model	317252881	3	105750960	F(3, 70) =	23.29
Residual	317812515	70	4540178.78	Prob > F =	0.0000
Total	635065396	73	8699525.97	R-squared =	0.4996
				Adj R-squared =	0.4781
				Root MSE =	2130.8

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
weight	3.464706	.630749	5.49	0.000	2.206717 4.722695
mpg	21.8536	74.22114	0.29	0.769	-126.1758 169.883
foreign	3673.06	683.9783	5.37	0.000	2308.909 5037.212
_cons	-5853.696	3376.987	-1.73	0.087	-12588.88 881.4934

Etape 1: Tester la significativité des variables. Pour cela, il suffit de regarder le "t-stat" (t) ou bien la P-value (P>|t|), et comparer ces valeurs à des "valeurs seuils". Pour faire simple, une variable est significative avec un intervalle de confiance de 95% si son t-stat est supérieur à 1,96 en valeur absolue, ou bien si sa P-value est inférieure à 0,05.

Dans notre exemple, on voit que la variable "mpg", qui correspond à la consommation en essence de la voiture n'est pas significative (t-stat trop faible en valeur absolue et P-value trop forte). De plus, l'intervalle de confiance à 95%, allant de -126.17 à 169.99 comprend la valeur 0. Il est donc impossible de rejeter l'hypothèse $\beta_2 = 0$.

Les deux autres variables "weight" et "foreign" sont significatives (t-stat de 5,49 et 5,37 donc supérieur à la valeur seuil de 1,96). De plus, l'intervalle de confiance ne comprend pas la valeur 0. Pour β_1 par exemple, l'intervalle de confiance permet de dire "je suis sûr à 95% que la valeur de β_1 se trouve entre 2,20 et 4,72. Le coefficient (=3.467 pour β_1 par exemple) correspond exactement au milieu de l'intervalle de confiance de la variable.

La variable "consommation de la voiture" n'étant pas significative, l'on relance la régression, mais en supprimant la variable en question. En effet, les résultats de la régression peuvent être modifiés par l'inclusion de variables non significatives, et il est donc préférable d'analyser le résultat d'une régression finale contenant uniquement des variables significatives. Le résultat de notre nouvelle régression se présente comme suit :

. regress price weight foreign Source: CaptainEconomics.fr

Source	SS	df	MS	Number of obs = 74		
Model	316859273	2	158429637	F(2, 71) =	35.35	
Residual	318206123	71	4481776.38	Prob > F =	0.0000	
				R-squared =	0.4989	
				Adj R-squared =	0.4848	
Total	635065396	73	8699525.97	Root MSE =	2117	

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
weight	3.320737	.3958784	8.39	0.000	2.531378	4.110096
foreign	3637.001	668.583	5.44	0.000	2303.885	4970.118
_cons	-4942.844	1345.591	-3.67	0.000	-7625.876	-2259.812

Etape 1 (nouvelle régression) : Les deux variables sont significatives (t-stat > 1,96 en valeur absolue).

Etape 2 : Étude des coefficients. La valeur estimée de β_1 est égale à 3,32 et celle de β_3 à 3637. Comment lire cela ? Cela signifie que "toutes choses égales par ailleurs", une voiture pesant une livre (unité de masse américaine) de plus, coûtera en moyenne 3,32 \$ de plus. Même raisonnement en ce qui concerne l'analyse du coefficient de notre variable indicatrice ; "toutes choses égales par ailleurs", une voiture étrangère coûte en moyenne 3637 dollars de plus qu'une voiture américaine.

Etape 3 : La précision de notre modèle : Pour cela, il est possible de regarder le "R-squared", qui mesure la proportion de la variance de Y (variable dépendante) qui est expliquée par la variation de toutes les variables explicatives. Le R-squared est par construction compris entre 0 et 1 ; plus on se rapproche de 1, plus le modèle est précis. Dans notre exemple, 49% de la variation de Y peut être expliquée par les variations de X_1 et D_1 . Même si cela ne peut pas être globalement rejeté, il manque en effet de nombreuses variables au modèle pour que celui-ci permettent d'estimer avec précision le prix d'une voiture en fonction de ses caractéristiques.

NB : Il n'existe pas de valeur du R-squared à partir de laquelle le modèle peut-être considéré comme bon ou mauvais (cela dépend du modèle). Pour donner un ordre d'idée coutumier dans cette situation, un R-squared proche de 0,8 est signe d'un bon

modèle, tandis que si votre R-square est proche de 0,2 , des variables importantes ont pu être omises).

II. EXEMPLE (FICHIER MS EXCEL)

1. Enoncé

Soit une regression linéaire multiple à réaliser avec MS Excel. Il s'agit d'exprimer le Chiffre d'affaires de la manutention portuaire (Y) en fonction des variables explicatives X1 (coût de l'escale), X2 (Budget Marketing) et X3 (Budget communication), tous en UM.

mois	CA Manutention portuaire	Coût de l'escale	Budget Marketing	Budget Communication
1	73959	8,75	50,04	61,13
2	71544	8,99	50,74	60,19
3	78587	7,5	50,14	59,16
4	80364	7,25	50,27	60,38
5	78771	7,4	51,25	59,71
6	71986	8,5	50,65	59,88
7	74885	8,4	50,87	60,14
8	73345	7,9	50,15	60,08
9	76659	7,25	48,24	59,9
10	71880	8,7	50,19	59,68
11	73598	8,4	51,11	59,83
12	74893	8,1	51,49	59,77
13	69003	8,4	50,1	59,29
14	78542	7,4	49,24	60,4
15	72543	8	50,04	59,89
16	74247	8,3	49,46	60,06
17	76253	8,1	51,62	60,51
18	72582	8,2	49,78	58,93
19	69022	8,99	48,6	60,09
20	76200	7,99	49	61
21	69701	8,5	48	59
22	77005	7,9	54	59,5
23	70987	7,99	48,7	58
24	75643	8,25	50	60,5

2. Rapport de l'utilitaire d'analyse

EXTRAIT RAPPORT DÉTAILLÉ

<i>Statistiques de la régression</i>	
Coefficient de détermination multiple	0,92673872
Coefficient de détermination R^2	0,85884465
Coefficient de détermination R^2	0,83767135
Erreur-type	1274,93553
Observations	24

ANALYSE DE VARIANCE

	<i>Degré de liberté</i>	<i>Somme des carrés</i>	<i>Moyenne des carrés</i>	<i>F</i>	<i>Valeur critique de F</i>
Régression	3	197798833	65932944,3	40,5626221	1,0848E-08
Résidus	20	32509212,1	1625460,61		
Total	23	230308045			

	<i>Coefficients</i>	<i>Erreur-type</i>	<i>Statistique t</i>	<i>Probabilité</i>	<i>Limite inférieure pour seuil de confiance = 95%</i>	<i>Limite supérieure pour seuil de confiance = 95%</i>
Constante	-25096,833	24859,6113	-1,0095425	0,32477316	-76953,073	26759,4076
Coût de l'escale	-5055,2699	526,399554	-9,6034843	6,2195E-09	-6153,3201	-3957,2196
Budget Marketing	648,61214	209,004879	3,10333493	0,00560234	212,635603	1084,58868
Budget Communication	1802,61096	392,848543	4,5885647	0,00017802	983,143256	2622,07866

Résultats

- $CA \text{ Manutention portuaire} = - 5055,27 \times \text{Coût de l'escale} + 648,61 \times \text{Budget Marketing} + 1802,61 \times \text{Budget Communication} - 25\,096,8$
- « -5055,2699 » : En cas d'augmentation de 1UM du coût de l'escale, le CA de manutention diminue de 5055 UM
- « 648,61214 » : A chaque augmentation de 1 UM du budget de l'Action Marketing, le CA de la M.P augmente de 648 UM
- « 1802,61096 » : A chaque UM dépensée pour les publicités communicationnelles, le CA MP, on a en moyenne 1802 UM supplémentaires
- **La publicité communicationnelle est plus rentable que les actions Marketing**

CHAPITRE III : ELABORATION ET ANALYSE DE FLUX MARITIMES ET PORTUAIRES

I. ÉLABORATION ET RÔLE DES STATISTIQUES MARITIMES ET PORTUAIRES

1. Élaboration des statistiques

Désignation d'un service en charge... le service statistique et études économiques

Il a deux divisions à savoir :

- la division statistique
- la division Études Economiques

1.1. La Division Statistique

Cette division comporte deux sections à savoir :

- la section collecte et codification
- la section saisie et production des états statistiques.

1.1.1 Section Collecte et Codification

Les maisons de consignation produisent un document sur lequel l'ensemble de la cargaison du navire, pour une destination, est répertorié. Ce document nommé manifeste est mis à la disposition du port . La section collecte se charge de faire le ramassage de ces documents. En plus de cette collecte externe au port, une autre interne est faite au niveau de la direction de la capitainerie. Celle-ci concerne les navires ayant accostés. L'opération de codification repose sur une nomenclature qui consiste à assigner les codes appropriés aux diverses informations contenues sur le manifeste. Cette codification se fait à partir d'une nomenclature des marchandises, des navires, des pays de provenance. On distingue cinq (5) types de manifestes à savoir les manifestes imports, les manifestes exports, les manifestes transit- import, transit- export et les manifestes en transbordement. Le manifeste renseigne sur certaines informations comme :

- le pays de provenance de destination des marchandises,
- le nom du navire,
- la date d'arrivée du navire,
- le numéro de voyage,
- la liste des articles sur le bateau et leur poids,
- le nom et adresse de l'expéditeur et du réceptionnaire.

1.1.2 -Section Saisie et Production des états statistiques.

A ce niveau sont vérifiés puis saisis les manifestes codifiés et transmis par la section codification. Une édition est ensuite faite pour des traitements ultérieurs. Une compilation mensuelle est faite par rapport aux différents types de trafic, et autres caractéristiques.

1.2 - La Division Etudes Economiques

Cette division est en charge de l'étude des états statistiques. Ceci en vue de faire des commentaires et analyses pour une amélioration du système et la prise d'importantes décisions. Ces données permettent aussi d'avoir une idée générale sur les prestations globales de service au cours d'une période donnée. La mission de cette division est de mettre à la disposition des partenaires du port des informations permettant d'évaluer l'impact et l'importance des activités portuaires sur l'économie nationale et sous régionale.

2. Collecte et sources de données

Cibles: . *compagnies maritimes. consignataires. manutentionnaires. opérateurs de terminaux*
. *administration portuaires, Chargeurs*

Outils: *Recherches documentaires, Observations (participatives), Questionnaires, Guide d'entretien*

3. Traitement des statistiques

I. Les types de traitements statistiques en fonction des questions-problèmes (QP)

Dans une recherche, on se pose toujours une ou plusieurs questions qu'il s'agira de résoudre. Le chercheur émet alors une solution qu'il croit possible (hypothèse). Une hypothèse (H) peut être considérée comme une réponse possible à la question-problème. D'autre part, toute question implique au moins une variable qu'il s'agit d'identifier ainsi que ses modalités et le type d'échelle correspondant.

Il existe plusieurs types de questions-problèmes qui se différencient plus ou moins nettement mais qui impliquent des traitements statistiques différents:

A. Recherche de description

Ce type de questions invite le chercheur de décrire les caractéristiques d'une ou plusieurs variables sur un échantillon ou une population. Il s'agira de collecter, présenter, résumer les informations jugées nécessaires et adéquates.

Exemples:

QP1: Quel est le poids moyen de mon échantillon ?

Hypothèse: Le poids moyen se situera en dessous d: 100 kg.

QP2: Quelle est la répartition de mon échantillon au niveau de la nationalité ?

Hypothèse: Il y aurait davantage de Tunisiens, puis des Maghrébins et enfin quelques autres nationalités.

QP3: Quelle est la répartition observée des sujets en fonction de leur poids par rapport au sexe ?

Hypothèse: Il y aura plus d'hommes qui pèseront plus lourds.

B. Recherche d'estimation

Il s'avère souvent impossible d'examiner une population entière. On recourt alors à un échantillon que l'on espère représentatif de cette population.

Le problème de savoir si les caractéristiques d'un échantillon observé correspondent bien aux caractéristiques de la population de référence est un problème d'estimation.

C. Recherche de comparaison

Lorsque dans un échantillon, nous avons des groupes par rapport à une variable, il est légitime de se demander si ces groupes sont similaires ou non. Autrement dit, les résultats obtenus dépendent-ils de la formation des groupes ?

D. Recherche de liaison

Si plusieurs variables différentes sont observées sur un même échantillon, nous cherchons à étudier la relation (la liaison) entre ces variables. Si deux variables s'avèrent reliées, ce sont des variables interdépendantes, c.à.d, l'une peut expliquer l'autre.

II. Les procédures avant tout traitement statistique

Les procédures expérimentales avant tout traitement statistique consiste à :

1. poser une question de recherche
2. émettre une hypothèse
3. élaborer une procédure expérimentale pour tester une hypothèse

La suite de la procédure consiste à :

1. constituer un ou de plusieurs échantillons
2. choisir des procédures statistiques
 - définir l'hypothèse nulle
 - choisir le seuil de signification
 - choisir le test

III. Les mesures à prendre avant tout traitement statistique

Pour choisir un test statistique, il est nécessaire de définir:

1. les types de variables (qualitatives / quantitatives)
2. le nombre de variables
3. la taille de l'échantillon
4. la loi de répartition (normale ou non)
5. les mesures répétées ou non (nombre de facteurs)

IV. Le recueil des données

Les données sont rassemblées dans un tableau à double entrée comprenant autant de lignes que d'unités d'observation indépendantes (les sujets d'étude); autant de colonnes que de "variables" ou de mesures caractérisant chacune des unités d'observation.

Le tableau ainsi constitué se nomme un tableau de codage des données

- Les lignes du tableau représentent les unités d'observation.

Les observations sont réalisées sur un nombre limité d'unités ou d'individus dans le but de tirer des conclusions applicables à une population entière. Ce nombre limité d'unités d'observation forme l'échantillon.

- Les colonnes du tableau de codage représentent les variables.

Une variable, c'est une quantité ou qualité susceptible de fluctuations ou encore une grandeur à laquelle on peut attribuer plusieurs valeurs différentes. Au minimum, une variable a toujours deux modalités ou deux fluctuations. Ainsi, la variable sexe possède deux modalités: fille et garçon.

Exemple : La variable « degré de concentration » peut être estimée sur une échelle de valeur à cinq degrés; elle a alors cinq modalités (très faible, faible, moyen, fort, très fort).

Les variables placées en colonne dans le tableau de codage seront accompagnées d'une légende de codage reprenant l'ensemble de l'information nécessaire pour pouvoir reconnaître la signification précise de n'importe quel code dans n'importe quel endroit du tableau.

Exemple : La légende mentionne le numéro de la variable, son énoncé explicite et un code pour chaque modalité. V_{10} : sexe, 1= homme, 2 = femme

V. Le choix du test statistique

❖ Pour les liens entre variables qualitatives, le choix du test dépend de la taille de l'échantillon :

- pour un échantillon de taille normale (effectifs calculés > 5) : test de χ^2
- pour un échantillon de taille réduite (effectifs calculés > 3) : test de χ^2 corrigé (correction de Yates)
- pour un échantillon de taille très réduite (effectifs calculés < 3) : le « test exact »

❖ **Pour les liens entre une variable qualitative et une variable quantitative le choix du test dépend :**

- **s'il s'agit de comparaison de deux moyennes :**
 - comparaison de deux moyennes observées
 - comparaison d'une moyenne observée à une moyenne théorique
- **si les effectifs sont suffisamment grands ($n > 30$) où la loi de répartition est normale (faire éventuellement un test de normalité) : c'est le test t de Student qui devrait être retenu.**

options du test :

- comparaison d'une moyenne observée et d'une moyenne théorique (one population) ou de deux moyennes observées (two populations)
- mesures appariées (paired) ou non appariées (unpaired)
- comparaison unilatérale (one-tailed) ou bilatérale (two-tailed)

- **Les effectifs sont faibles et la répartition n'est pas normale (faire éventuellement un test de normalité) : ce sont les tests non paramétriques qui devront être retenus.**

Les options :

- **séries non appariées :**

test W de Wilcoxon

test U de Mann et Whitney

test C1 de Fisher-Yates-Terry

- **séries appariées :**

test T de Wilcoxon

- **s'il s'agit de comparaison de plusieurs moyennes (> 2)**

si la loi de répartition de probabilité est normale pour la variable mesurée (faire éventuellement un test de normalité) : c'est l'analyse de la variance (ANOVA) qui devrait être retenue.

options :

ANOVA à plusieurs facteurs

tests « post-hoc » :

Méthode de Bonferroni (test t) : recommandé pour un usage général, si les comparaisons ne sont pas trop nombreuses.

Méthode de Tukey (test t) : à utiliser lorsque toutes les comparaisons paire par paire sont intéressantes.

Méthode de Dunnett : à utiliser lorsque l'on compare le groupe contrôle avec les autres groupes, mais pas les autres groupes en eux.

Méthode de Sheffé (test F) : à utiliser pour les comparaisons compliquées

- si la loi de répartition de probabilité n'est pas normale pour la variable mesurée (faire éventuellement un test de normalité) : **ce sont les tests non paramétriques qui devront être retenus.**

Les tests de normalité proposées :

Martinez-Iglewicz

(Kolmogonov-Smirnov)

Agostino Omnibus

Agostino-Skewness (si $n > 8$)

Agostino-Kurtosis (si $n > 20$)

❖ Pour les liens entre deux variables quantitatives

- une des deux distributions liées au moins est normale avec une variance constante : c'est le test de corrélation (test de Bravais-Pearson) ou le test de régression qui détermine si la pente est statistiquement significative de 0.
- si aucune des variables liées n'est normale et de variance constante (petits échantillons) : c'est le test non paramétrique de corrélation des rangs (test de Spearman)

4. Rôle des statistiques maritimes et portuaires

. bilan d'activités. analyse [comparative] des performances. prévisions d'investissement et dimensionnement

II. CONFIGURATION DES ETATS STATISTIQUES MARITIMES ET PORTUAIRES

1. PRESENTATION DES DONNEES [description synthetiques des informations chiffrées + analyse + interpretation]

QUELLES DONNEES COLLECTER?

. MOUVEMENT DES NAVIRES [voir bulletin statistique des chargeurs/évolution mensuelle ou périodique des escales de navires aux ports de commerce / pêche/autres..../ et comparaison d'une année à une autre. escales par type de navires]

. FLUX DES MARCHANDISES [voir bulletin statistique des chargeurs/évolution mensuelle ou périodique du TRAFIC DE MARCHANDISES aux ports de commerce / ENTRANT ET SORTANT/ TRAFIC CONTENEURISE/ IMPORT EXPORT TRANSBORDEMENT/ et comparaison d'une année à une autre. escales par type de navires]

. TAUX DE CONTENEURISATION [CALCUL DEDUIT DU RAPPORT TRAFIC CONTENEURISE ET TRAFIC GLOBAL]

. TYPE DE TRANSIT [TRAFIC DIRECT OU DOMESTIQUE] . concerne le trafic de marchandises manipulées en direction ou en provenance du territoire national/TRANSBORDEMENT. concerne le trafic de marchandises déchargées aux ports et rechargées en direction d'autres ports/TRANSIT. concerne le trafic de marchandises manipulées en direction ou en provenance des pays de l'arrière-pays

. TYPE DE PRODUITS [IMPORTATION PAR TYPE DE PRODUITS . PRODUITS CIMENTIERS. PRODUITS ALIMENTAIRES. ENGRAIS. EQUIPEMENTS MINIERES. DIVERS. EXPORTATIONS PAR TYPE DE PRODUITS BOIS ET DERIVES. CAFE CACAO . CAOUTCHOUC . PALME ET DERIVE. NOIX DE CAJOU . COTON

COMMENT LES PRESENTER?

Dans un tableau statistique à une seule entrée ou à double entrées (le plus souvent)

2. EXEMPLES DE CONFIGURATION [PORTS DE LONGONI/MAYOTTE . MONTREAL. ABJ ET SAN PEDRO]

2.1 Le port du Longoni

Implanté sur la baie de Longoni, le port de Longoni est un port décentralisé relevant de la compétence du Conseil Général de Mayotte. La Chambre de Commerce et d'Industrie de Mayotte en est le concessionnaire depuis sa mise en service en 1992 et assure de ce fait plusieurs missions :

- la gestion des terres-pleins et hangars,
- la gestion des entrées et sorties des marchandises au port,
- la livraison des marchandises aux clients et
- assure des prestations de services aux navires.

La CCI de Mayotte gère également une zone industrielle, la Vallée 2, où sont implantées des entreprises du milieu portuaire, ainsi qu'une zone de restauration, une zone de dépotage et un parking client. Depuis presque 21 ans après sa mise en service, les évolutions qu'il a connues jalonnent les principales étapes du développement de Mayotte en général.

Le port de Longoni en 2011, c'est aussi :

- plus de 200 navires marchands qui escalent sur le port de Longoni chaque année,
- une activité logistique qui génère plus de 500 emplois directs,
- 11 919 483-€ de chiffre d'affaires de manutention (CCI/SMART),
- 6 947 486-€ de droits de port,
- 2 652 360-€ de RSM (redevance sur marchandises),
- 18 205 EVP en import,
- 16 204 EVP en transbo.

Une activité globalement en croissance de 14% entre 2009 et 2011.



FIGURE 4.1 – Le port de Longoni en 2011



FIGURE 4.2 – Le plan d'exploitation du port de Longoni

4.1.2 Statistiques du port de Longoni

On se donne les statistiques suivantes relatives au port de Longoni :

TRAFFIC			2009	2010	2011	2012
Entrées en kt*	Total Vrares liquides (hors ravitaillement)		97	96	86	-12%
	Total Vrares solides		49	47	37	-24%
	Marchandises diverses	Conteneurs	184	184	337	83%
		Roullet hors conteneurs	0	0	0	
		M/ses diverses hors conteneur et roullet	44	45	45	2%
		Total M/ses diverses	228	229	504	121%
Total entrées		374	373	627	68%	
Entrées en kt*	Total Vrares liquides (hors ravitaillement)		0	0		
	Total Vrares solides		0	0	0	
	Marchandises diverses	Conteneurs	37	41	206	456%
		Roullet hors conteneurs	0	0	0	
		M/ses diverses hors conteneur et roullet	1	0	2	194%
		Total M/ses diverses	38	42	209	450%
	Total sorties		38	42	209	450%
Transbordés	Marchandises	Conteneurisés	235	310	344	46%
Total Entrées + Sorties			647	724	1180	82%
* source : Conseil Général de Mayotte / Direction du Port						
Conteneurs**	nb EVP en entrée		16129	17154	17509	9%
	nb EVP en sortie		16275	18367	16921	4%
	nb EVP en transbordement		12721	16795	16841	32%
	Total nb EVP		45125	52316	51271	14%
** source : Chambre de Commerce et d'Industrie de Mayotte / Direction de la Concession Portuaire						
Passagers***	Croisière Services côtiers ou inter-îles		6188 40750	3284 57527	0 2252	-94%
*** source : Conseil Général de Mayotte / Direction du Port						
Pêche	Tonnage débarqué		0	0	0	

TABLE 4.1 – Statistiques Port de Longoni

- Le port de Longoni a enregistré une augmentation entre 2009 à 2011, avec une variation positive de 82% sur le tonnage du trafic global de marchandises. Comme le montre le tableau 4.1, le tonnage des importations a évolué de 68%, passant de 374Kt à 627Kt en 3 ans, et les exportations de 38Kt à 209Kt soit une progression de 450%.
- En terme d'EVP entre 2009 et 2011, le trafic connaît une évolution de 14%, on constate une légère diminution des exportations entre 2010 et 2011 (18367 à 16921), mais avec un pourcentage positif, et un ralentissement des conteneurs transbordés sur la même période. Les importations restent stables.

4.1.3 Les indicateurs d'exploitation des terminaux conteneurs en 2011

Terminaux conteneurs Quai 1 + Quai 2	
Nombre d'escales navires porte conteneurs	195
Nombre d'heures passées au port	6344
Nombre d'heures passées à Quai	5560
Nombre d'heures opérations commerciales	3823
Tonnes totales de marchandises conteneurisées opérées	538 119
Nombre total d'EVP	51 271
Nombre total d'EVP déchargés	17509
Nombre total d'EVP chargés	16 921
Tonnage manutentionné / heures d'opérations commerciales	141
Tonnage manutentionné / heures passées à quai	97
Tonnage manutentionné / heures passées au port	85
Nombre moyen d'EVP manipulés / heures	13

Pour une escale navire en moyenne

Temps passé (heures) au port	33
Temps passé (heures) à quai	29
Temps passé (heures) à quai pour opérations commerciales	20
Nombre d'équipes	2
Nombre de grues opérationnelles	2
Nombre de shifts	4
Nombre moyen de conteneurs manipulés	263
Tonnes moyennes de marchandises conteneurisées	2760

Indicateurs de productivité (EVP/h/grue)	7
Indicateur de production (tonnes/h/équipes)	10

TABLE 4.2 – Les indicateurs d'exploitation des terminaux conteneurs en 2011

Pour l'année 2011, on compte 211 navires qui sont passés par le port de port de Longoni : 195 navires commerciaux et 16 non commerciaux. On enregistre aussi pour cette année une légère baisse du tonnage manipulé entre 2010 et 2011 (545214,5 à 538119), donc une baisse des conteneurs estimés en EVP. Ceci s'explique notamment par les événements sociaux qui ont secoué l'île de Mayotte en fin d'année 2011.

4.1.4 Les marchandises conteneurisées en 2011

	Entrées	Sorties	Trafic 2011	Trafic 2010	Variation 2010/2011
Import	13179	0	13179	13307	-0,96%
Export	0	785	785	748	4,93%
Transho	8098	8743	16841	18141	-7,16%
Vides	4330	16136	20465	20816	-1,68%
TOTAL	25607	25664	51271	53012	-3,28%

TABLE 4.3 – Les marchandises conteneurisées en 2011

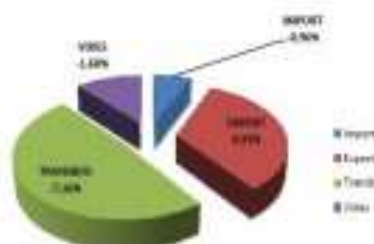


FIGURE 4.3 – Trafic en EVP

4.1.5 Les visites de contrôle (DAF, DSV, DOUANE)

Type de visite	DAF/Phyto	DSV	Douane
1er semestre	264	198	96
2e semestre	270	59	14
TOTAL	526	257	110

TABLE 4.4 – Tableau Visites de contrôle des marchandises

Des visites de contrôle sont effectuées régulièrement sur les conteneurs frigos, ainsi que sur les conteneurs dry pour vérifier toutes les marchandises débarquées, vérifier le poids réel des TC, le respect des normes, voire si tout les produits sont déclarés... Sur l'année 2011, il y a eu 1015 conteneurs visités :

- Les visites phytosanitaires effectuées par la Direction de l'Agriculture et de la Forêt représentent plus de 526 conteneurs dans l'année.
- Les services vétérinaires ont visité plus de 257 conteneurs.
- Pour les visites douane, on dénombre 110 conteneurs.

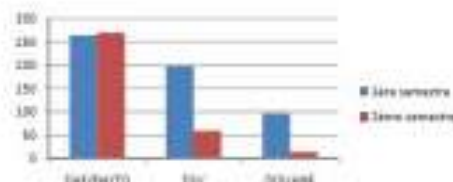


FIGURE 4.4 – Diagramme Visites de contrôle des marchandises

4.1.6 Les engins de manutention de la concession

[illegible]

FIGURE 4.5 – Tableau d'utilisation du parc des engins de maintenance de la concession

L'année 2011 est particulière en terme d'entretien des engins car elle a vu la recrudescence de la maintenance curative. Ceci est essentiellement dû aux pannes à répétition des engins et à leur immobilisation durant plusieurs jours. Dans le même temps, le coût de la maintenance préventive a baissé de 46%. Cette situation est due à la réduction des machines à uniquement 2 stackers, ce qui ne permet pas le respect d'un planning de maintenance. La consommation en gasoil a légèrement augmenté par rapport à 2010, passant de 41582 litres à 46523 litres, soit une hausse de près de 12%. Ceci s'explique par le fait qu'en cas de panne d'une machine, celle qui est disponible travaille beaucoup plus dans le soucis de satisfaire la clientèle. Une augmentation de gasoil a donc été enregistrée.

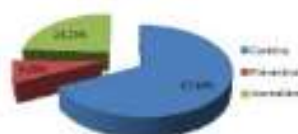


FIGURE 4.6 – Répartition par nature de maintenance des engins

4.1.7 La gestion des matières dangereuses au port

Le nombre de conteneurs de matières dangereuses reçus au port est en constante augmentation, cela a nécessité la mise en place d'une procédure de gestion des matières dangereuses sur le port en collaboration avec la Capitainerie du port et la Préfecture. L'année 2011 a été très prometteuse sur le nombre de TC de matières dangereuses réceptionnés sur le parc, néanmoins, on a enregistré une forte baisse en fin d'année. La grande difficulté sur la gestion des conteneurs de matières dangereuses relève des infractions constatées sur le parc.

Mois	Janv.	Fév.	Mars	Avril	Mai	Juin	Juill.	Août	Sept.	Oct.	Nov.	Déc.	TOTAL
Quantité	135	52	59	85	89	62	44	48	96	52	72	37	831
Tonnage	2329	746	838	1433	1428	947	638	725	1401	754	1076	549	12863

TABLE 4.5 – Tableau Matières dangereuses année 2011



FIGURE 4.7 – Courbe Matières dangereuses année 2011

2.2 Trafic de passagers de ports métropolitains

Avec 28,5 millions de passagers, le trafic des ports maritimes métropolitains est en légère hausse par rapport à 2010 (+0,6%).

	Passagers autre que croisière		Total passagers		Évolution 2011/2010
	2010	2011	2010	2011	
Dunkerque	2533	2601	2533	2601	2,7%
Calais	10236	10063	10236	10067	-1,7%
Boulogne	296	0	296	0	-100%
Dieppe	254	259	254	259	1,7%
Le Havre	270	345	526	715	36,1%
Caen	1022	952	1022	954	-6,7%
Cherbourg	559	625	623	700	12,5%
St-Malo	930	1124	948	1157	22,0%
Roscoff	534	544	534	545	2,0%
TOTAL Manche - Mer du Nord	16635	16513	16973	16998	0,2%
Sete	201	170	212	212	0,0%
Marseille	1383	1531	2485	2887	16,2%
Toulon	1283	1219	1568	1637	4,4%
Nice	916	812	1811	1662	-8,2%
Bastia	2524	2282	2532	2291	-9,5%
L'Île-Rousse	406	366	406	366	-9,8%
Ajaccio	1117	1041	1868	1941	3,9%
Bonifacio	240	241	254	261	2,9%
Calvi	179	154	196	217	10,9%
Total Méditerranée	8250	7817	11333	11475	1,3%
Total Métropole (17 ports)	24885	24330	28306	28474	0,6%

TABLE 4.6 – Tableau Trafics de passagers des ports métropolitains



FIGURE 4.8 – Évolution 2011/2010 des trafics de passagers dans les ports métropolitains français par façade (en milliers de passagers)

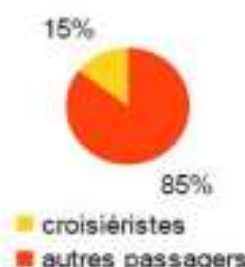


FIGURE 4.9 – Répartition par type de voyageurs dans les ports métropolitains français en 2011

- En Manche-Mer du Nord, le trafic est stable par rapport à 2010.
 - À Calais, premier port de la façade pour les trafics de voyageurs, le trafic recule de -1,7% en 2011, en raison notamment de l'interruption des liaisons SeaFrance à partir de la mi-novembre.
 - Le Grand port maritime du Havre enregistre une forte progression de trafic (+36,1%), grâce notamment à la hausse des fréquentations vers la Grande-Bretagne.
 - À Dunkerque (+2,7% en 2011), la croissance perdure, soutenue par le dynamisme de la ligne Dunkerque-Douvres.
 - À contrario, le port de Boulogne Sur Mer n'a plus de trafic passagers depuis l'arrêt de la liaison Boulogne-Douvres et le départ de LD lines fin 2010.
 - Les autres ports de la façade affichent des résultats globalement positifs (+12,5% et +22% respectivement à Cherbourg et Saint-Malo).
 - Le port de Caen voit sa fréquentation diminuer de -6,7%.
- En Méditerranée, le trafic poursuit sa progression et s'établit à 11,5 millions de passagers en 2011 (+1,3%). Les ports de la façade méditerranéenne ont connu des réussites diverses.
 - Sous l'effet de la forte croissance des liaisons avec la Corse, le Grand port maritime de Marseille voit ses trafics de passagers augmenter de +16,2%. En revanche, les lignes régulières vers le Maghreb affichent une fréquentation en baisse (-12%).
 - Les trafics à Toulon (+4,4%), Ajaccio (+3,9%), Bonifacio (+2,9%) et Calvi (+10,9%) restent bien orientés.
 - À contrario, les ports de Bastia, Nice et L'Île-Rousse enregistrent des baisses de trafic.

On se donne à la page suivante les données et graphes superposés des trafics totaux de passagers lors des années 2009, 2010 et 2011, sur la façade Manche - Atlantique, exprimés en milliers de passagers (chiffres provisoires au 24 janvier 2012).

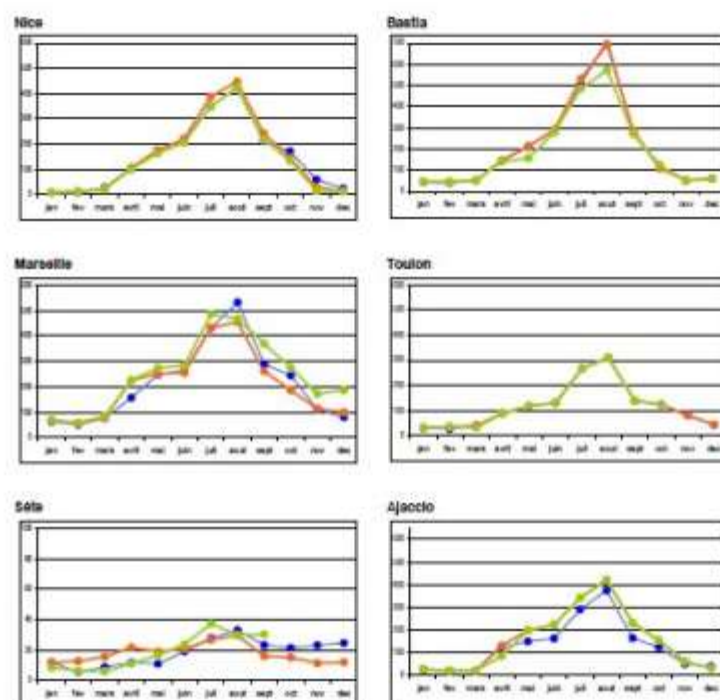


FIGURE 4.13 – Trafic total de passagers en 2009 (- bleu), 2010 (- rouge), 2011 (- vert). Façade Méditerranéenne. Exprimé en milliers de passagers (Chiffres provisoires au 24 janvier 2012)

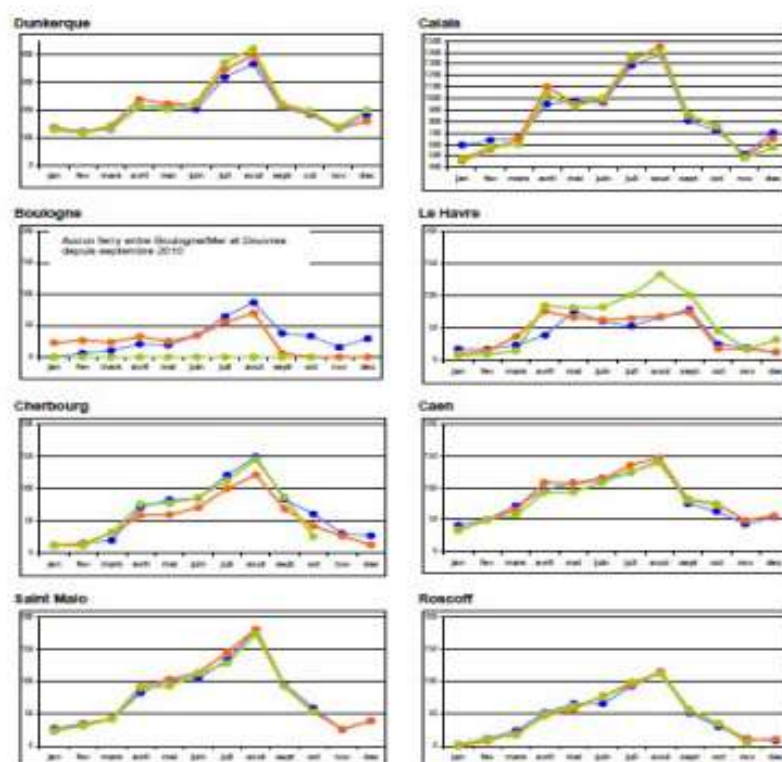


FIGURE 4.14 – Trafic total de passagers en 2009 (- bleu), 2010 (- rouge), 2011 (- vert). Façade Manche - Atlantique. Exprimé en milliers de passagers (Chiffres provisoires au 24 janvier 2012)

2.3. Le port de Montreal

4.3.1 Présentation du Port



Situé sur le Saint-Laurent, une des plus grandes voies navigables du monde, le Port de Montréal est la liaison intermodale la plus directe entre le cœur industriel de l'Amérique du Nord et les marchés de l'Europe du Nord et de la Méditerranée.

À 1600 kilomètres à l'intérieur des terres, le Port de Montréal donne accès à un bassin de plus de 100 millions de consommateurs canadiens et américains.

De tous les ports de la côte Est Nord-Américaine, c'est le Port de Montréal qui offre l'accès le plus rapide, le plus direct et le plus économique aux principaux marchés du centre du Canada, du Nord-Est et du Midwest américains. Il suffit d'un seul arrêt à Montréal pour exploiter ce gigantesque potentiel commercial.

De plus, le Port de Montréal est relié, par les transporteurs maritimes qui le desservent, à plus de 100 pays sur les 5 continents. Il offre des liens directs avec les grands ports Nord-Européens et Méditerranéens parmi lesquels :

- Anvers, en Belgique,
- Liverpool, au Royaume-Uni,
- Rotterdam, aux Pays-Bas,
- Hambourg et Bremerhaven, en Allemagne,
- Le Havre et Marseille/Fos, en France,
- Algésiras, Barcelone et Valence, en Espagne,
- Gênes, Livourne et Gioia Tauro, en Italie,
- Lisbonne, au Portugal.

Le Port de Montréal possède des installations modernes pour manutentionner tous les types de fret. Il gère notamment plus de 1,3 million de conteneurs EVP (équivalents 20 pieds), représentant plus de 12 millions de tonnes de marchandises conteneurisées chaque année.

Son trafic de conteneurs à l'exportation et à l'importation est assuré par des services réguliers à haute fréquence de transporteurs maritimes de classe mondiale : CMA CGM, Hanjin Shipping, Hapag-Lloyd, Maersk, MSC et OOCL.

Le Port de Montréal manutentionne également des marchandises non conteneurisées, du vrac liquide, du vrac solide et des céréales.

4.3.2 Le Port en bref et en chiffres

Avec sa situation géographique privilégiée qui lui permet de rejoindre plus de 100 millions de consommateurs en moins de 36 heures de train, il est le premier port de conteneurs en importance de l'est du Canada et un port d'envergure internationale relié à plus de 80 pays partout au monde.

- Les trois grands secteurs de manutention s'y retrouvent : marchandises conteneurisées, vrac liquide et vrac solide (incluant les céréales). Le port est également actif dans le secteur des croisières internationales et fluviales.
- Près de 30 millions de tonnes de marchandises y sont manutentionnées annuellement.

- Ses installations portuaires s'étalent sur 26 km sur l'île de Montréal et sur 4 km à Contrecoeur, à 40 km en aval de Montréal, sur la rive sud du Saint-Laurent.
- Il est le seul port à conteneurs de la Porte continentale du Canada par où transitent plus de 80% des échanges commerciaux du Canada.
- Il fait partie du club sélect des ports du monde qui reçoivent annuellement plus d'un million de conteneurs équivalent vingt pieds (EVP).
- Le réseau ferroviaire du port est directement relié aux réseaux du CN (Canadien National) et du CP (Canadien Pacifique).
- En 2011, le Royaume-Uni et l'Europe continentale ont fourni 46,8% du trafic des conteneurs, la Méditerranée, 18,9%, l'Asie, 12,9%, le Moyen-Orient, 8,7%, l'Amérique latine, 5,9%, l'Afrique/Océanie, 4% et le Canada, 2,8%.
- 2200 bateaux en moyenne chaque année,
- 5000 mouvements quotidiens de camions (entrées et sorties),
- 80 convois de trains chaque semaine,
- 100 km de voies ferrées le long des quais et sur les terminaux.
- L'année 2011 s'est soldée par une année record avec 28,5 millions de tonnes métriques de marchandises manutentionnées, soit une hausse de 10,1% par rapport à 2010.
- Le secteur des croisières a accueilli 40 navires en 2011 pour un total de 38031 passagers. Le nombre de passagers attendus pour l'année 2012 s'élève à 55500, soit une augmentation de 46%.
- 1,5 milliard de dollars en valeur ajoutée à l'économie canadienne (Étude Secor, 2008) soit 1,0 milliard de dollars au Québec et 0,5 milliard de dollars pour le reste du Canada,
- 241 emplois à l'Administration portuaire de Montréal (2012)
- 18280 emplois soutenus par l'activité maritime et portuaire à Montréal : 5 400 emplois directs (emplois liés aux entreprises directement impliquées dans le système portuaire et maritime de Montréal) et 12 880 emplois indirects (emplois liés aux fournisseurs de la demande de biens et services engendrée par les entreprises directement impliquées dans le système portuaire et maritime de Montréal).

4.3.3 Quelques statistiques du port de Montréal

Trafic cumulé de conteneurs					
TONNES MÉTRIQUES ET EVP					
2013-2012 Janvier-août					
TONNES	2013	2012	VARIATION		
			TONNES	%	
Entrées	3 778 614	3 891 913	(113 399)	-2,9 %	
Sorties	4 145 608	4 026 613	118 995	2,9 %	
Total	7 924 022	7 918 526	5 496	0,1 %	
CONTENEURS EVP	2013	2012	VARIATION		
			EVP	%	
Entrées	443 791	452 812	(9 021)	-2,0 %	
Sorties	456 613	456 316	297	0,1 %	
Total	900 404	909 128	(8 724)	-1,0 %	

EVP : nombre de conteneurs en unités équivalentes à 20 pieds

FIGURE 4.15 – Trafic cumulé de conteneurs

Trafic cumulé				
Tonnes métriques				
2012-2013 - Janvier à août				
Ensemble du trafic	2012	2013	Variation	
			Tonnes	%
Types de marchandises				
MARCHANDISES DIVERSES				
Conteneurisées	7 924 022	7 910 526	8 496	.1 %
Non conteneurisées	102 354	85 494	17 000	20.0 %
Total	8 026 376	8 004 020	22 666	.3 %
MARCHANDISES EN VRAIE				
Solide	3 886 922	3 881 389	(224 447)	-6.8 %
Liquide	6 000 904	6 221 905	430 909	7.1 %
Total	10 044 062	10 107 544	257 018	1.0 %
Trafic à l'entrée				
Types de marchandises	2012	2013	Variation	
			Tonnes	%
MARCHANDISES DIVERSES				
Conteneurisées	3 772 514	3 691 913	(113 200)	-3.0 %
Non conteneurisées	60 423	51 627	13 926	23.0 %
Total	3 832 937	3 743 540	(89 653)	-2.3 %
MARCHANDISES EN VRAIE				
Solide	2 847 273	2 795 311	48 962	1.7 %
Liquide	4 904 123	4 964 704	(60 581)	-1.2 %
Total	11 686 280	11 708 666	(111 142)	-.9 %
Trafic à la sortie				
Types de marchandises	2012	2013	Variation	
			Tonnes	%
MARCHANDISES DIVERSES				
Conteneurisées	4 148 828	4 028 913	118 986	3.0 %
Non conteneurisées	37 071	33 567	3 204	8.6 %
Total	4 185 899	4 060 480	122 989	3.0 %
MARCHANDISES EN VRAIE				
Solide	509 649	1 052 066	(273 606)	-53.2 %
Liquide	1 796 741	1 267 251	499 490	28.7 %
Total	6 745 969	6 400 739	345 130	5.4 %

FIGURE 4.16 – Tableau Trafic cumulé

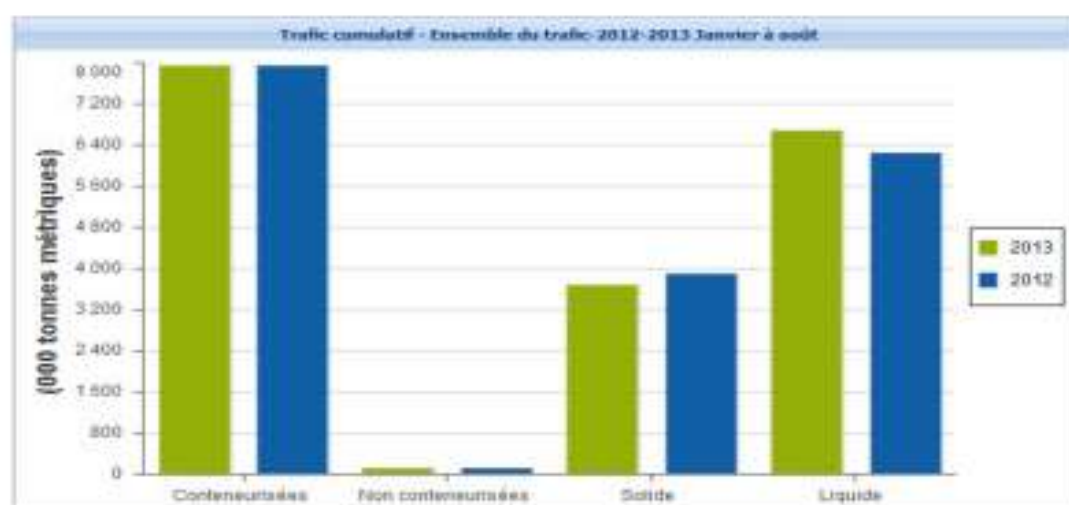


FIGURE 4.17 – Diagramme Trafic cumulé

Trafic cumulé de conteneurs TONNES MÉTRIQUES ET EVP					
2013-2012 : Janvier-août					
TONNES	2013	2012	VARIATION		
			TONNES	%	
Entrées	3 776 514	3 891 913	(115 399)	-2,9 %	
Sorties	4 145 404	4 326 613	(181 209)	-4,0 %	
Total	7 924 022	8 218 526	(294 504)	-3,6 %	
CONTENEURS EVP	2013	2012	VARIATION		
			EVP	%	
Entrées	443 791	452 812	(9 021)	-2,0 %	
Sorties	456 613	466 346	(9 733)	-2,1 %	
Total	900 404	919 158	(8 754)	-1,0 %	

EVP : Nombre de conteneurs en unités équivalentes 20' pieds

FIGURE 4.18 – Trafic cumulé de conteneurs

Trafics trimestriels Tonnes métriques						
Trimestre	Vrac		Marchandises diverses			Total général
	Liquide	Solide	Conteneurisées	Non cont.	Total	
1 ^{er} trimestre mars 2013	1 788 616	1 174 263	2 514 517	37 942	2 552 459	5 515 338
2 ^e trimestre juin 2013	3 372 909	1 639 963	3 025 528	42 834	3 068 362	7 814 233
3 ^e trimestre septembre 2013						
4 ^e trimestre décembre 2013						
Cumulatif	4 902 524	2 805 224	5 541 045	80 776	5 621 821	13 729 568
Même période 2012	4 342 024	2 716 901	5 373 776	68 384	5 442 160	13 100 984
Variation à ce jour	12,9 %	3,3 %	3,2 %	18,1 %	3,3 %	4,8 %

FIGURE 4.19 – Trafics trimestriels



FIGURE 4.20 – Trafic cumulé de conteneurs

4.4 Exercices

Exercice 36 - Port de Longoni

- À l'aide de la table 4.1, retrouver - dans la mesure du possible et en justifiant - les résultats suivants :
 - Le port de Longoni a enregistré une augmentation entre 2009 à 2011, avec une variation positive de 82% sur le tonnage du trafic global de marchandises. Le tonnage des importations a évolué de 68%, passant de 374Kt à 627Kt en 3 ans, et les exportations de 38Kt à 209Kt soit une progression de 450%.
 - En terme d'EVP entre 2009 et 2011, le trafic connaît une évolution de 14%, on constate une légère diminution des exportations entre 2010 et 2011 (18367 à 16921), mais avec un pourcentage positif, et un ralentissement des conteneurs transbordés sur la même période. Les importations restent stables.
- Concernant la baisse des conteneurs estimés en EVP (page 69), calculer le pourcentage de variation, le coefficient multiplicateur et l'indice base 100 associés. Si ce pourcentage se maintient, à combien pourra-t-on estimer le tonnage manipulé en 2014 ?
- Si vous aviez à commenter la table 4.3 et la figure 4.3, quelles seraient vos observations ? La figure 4.3 peut-elle être remplacée par d'autres graphiques ?
- La figure 4.5 met en avant un problème concret relatif aux engins de manutention. Quelle solution pouvez-vous apporter au problème ?
- Le tableau 4.5 décrit l'évolution des tonnages de matières dangereuses en 2011. À l'aide de la méthode des moindres carrés, déterminer une estimation de ce tonnage en juin 2012.

Exercice 37 - Trafics de passagers des ports métropolitains

On s'intéresse aux figures (tableaux) 4.10 et 4.11 et en particulier au trafic "Bouloonnais" des passagers exprimé en milliers de passagers en 2009 et 2010.

- Calculer la série des moyennes mobiles sur quatre périodes.
- Calculer les coefficients de la droite de régression.
- Représenter sur un même graphique la série brute (t, x_t) , la série des moyennes mobiles sur quatre périodes et la droite de régression. Que vous enseigne ce graphique ?
- Le modèle est-il additif ou multiplicatif ?
- Calculer les coefficients saisonniers (1 saison = 1 trimestre) de la série puis les valeurs de la série désaisonnalisée.

Exercice 38 - Le port de Montréal

On considère la figure 4.19. Supposons que l'on dispose des valeurs relatives au 3e trimestre 2013 à savoir :

Trimestre	Vrac		Marchandises diverses			Total général
	Liquide	Solide	Conteneurisées	Non conteneurisées	Total	
3e trimestre septembre 2013	2126768	1430657	3100045	45657	3145702	6703127

- Faites apparaître dans la figure 4.19 les modifications consécutives à l'apport de ces valeurs.
- Représenter graphiquement le trafic cumulé de conteneurs (comme dans la figure 4.20) relatif à septembre 2013.

2.4 PORT DABIDAN ET DE SAN PEDRO [TPE OU PRESENTATION INTERFACE DES DONNEES STATISTIQUES]

III. ETUDE DE CAS DE SERIES CHRONOLOGIQUES DE DONNEES MARITIMES ET PORTUAIRES

1. méthode des moindres carrés

Soit un terminal à conteneurs enregistrant mensuellement le nombre d'escales donné dans le tableau ci-dessous:

Le tableau de calcul pour trouver l'équation de la tendance (droite des moindres carrés) est le suivant:

x_i	y_i	x_i^2	$x_i y_i$	T
1	135	1	135	135,8
2	143	4	286	140,3
3	140	9	420	144,8
4	154	16	616	149,3
5	152	25	760	153,8
Σx_i	Σy_i	Σx_i^2	$\Sigma x_i y_i$	
15	724	55	2217	

$$\begin{cases} \bar{x} = 3 & (17 \text{ mars}) \\ \bar{y} = 145 & F \end{cases}$$

$$a = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad \text{avec } n = 5$$

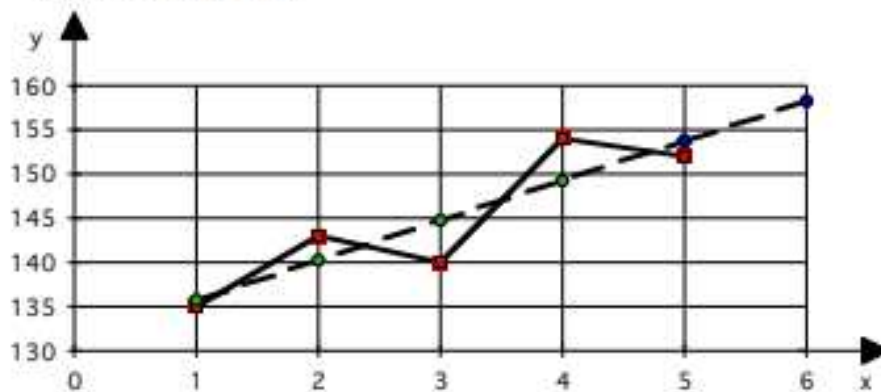
$$a = \frac{2217 - 5 \times 3 \times 145}{55 - 5 \times 3^2} = \frac{45}{10} = 4,5$$

Le nombre d'escale a tendance à augmenter de 4,5 navires par mois. L'équation de la droite de tendance est:

$$T = ax + b \quad \text{avec} \quad b = \bar{y} - a\bar{x}$$

$$T = 4,5x + 131,3$$

On trace le graphe:

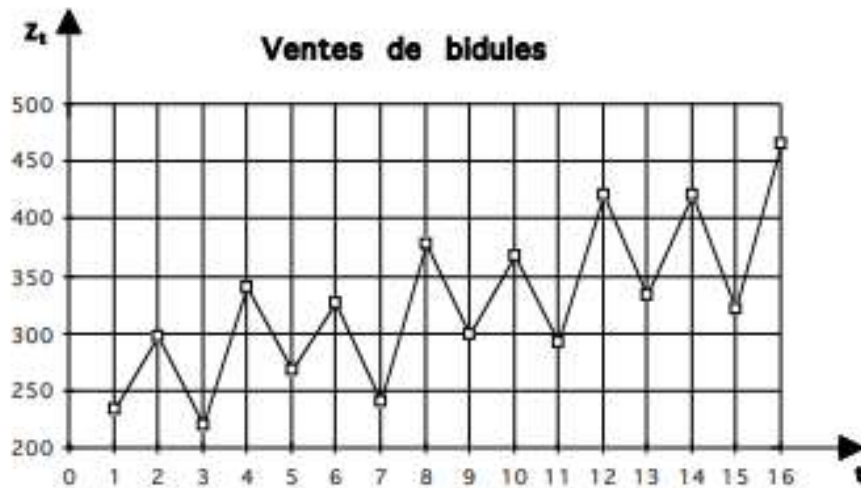


La prévision pour la date 6 s'obtient en remplaçant x par la valeur 6 dans l'équation de la tendance:

$$T(6) = 158 \text{ navires}$$

2. Méthode des coefficients saisonniers

Soit l'évolution du trafic trimestriel de conteneurs Bidules (vendus et livrés) pour positionnement sur une plateforme logistique portuaire



La série manifeste une tendance régulière à la hausse avec des variations saisonnières assez marquées. Une méthode de décomposition semble donc bien adaptée pour son étude.

La tendance peut être déterminée par la méthode des moindres carrés. Il s'agit alors de la droite qui passe le plus près possible de l'ensemble des points. Plus précisément, si on appelle S la somme des carrés des écarts verticaux entre une droite quelconque et les points, la tendance est la droite pour laquelle la somme S est minimale.

L'effet saisonnier peut être évalué à l'aide de coefficients qui mesurent les écarts entre les ventes et la tendance.

Il reste enfin l'effet aléatoire. Dans cette approche, il s'agit d'un effet résiduel: ce sont les irrégularités qui restent inexpliquées une fois analysées la tendance et les saisons.

Les aléas peuvent être rendus plus visibles en effaçant les variations saisonnières. C'est l'intérêt de la courbe C.V.S. (Corrigée des Variations Saisonnières).

Les calculs principaux sont regroupés dans un tableau :

Années	dates t	série z _t	produits t.z _t	carrés t ²	Tendance T=at+b	rapports z _t / T	coef saison	CVS z _t /coef
1	1	235	235	1	249	0,95	0,92	257
	2	298	596	4	259	1,15	1,10	271
	3	221	663	9	270	0,82	0,81	273
	4	340	1360	16	280	1,21	1,17	290
2	5	268	1340	25	290	0,92	0,92	293
	6	327	1962	36	301	1,09	1,10	297
	7	242	1694	49	311	0,78	0,81	299
	8	378	3024	64	322	1,17	1,17	322
3	9	300	2700	81	332	0,90	0,92	328
	10	368	3680	100	343	1,07	1,10	334
	11	292	3212	121	353	0,83	0,81	361
	12	421	5052	144	364	1,16	1,17	359
4	13	334	4342	169	374	0,89	0,92	365
	14	421	5894	196	384	1,10	1,10	382
	15	322	4830	225	395	0,82	0,81	398
	16	465	7440	256	405	1,15	1,17	396
136		5232	48024	1496	5232	16,00		

$$\text{Point moyen} \quad \begin{cases} \bar{t} = \frac{\sum t}{n} = \frac{136}{16} = 8,5 \\ \bar{z} = \frac{\sum z_t}{n} = \frac{5232}{16} = 327 \end{cases}$$

$$a = \frac{\sum tz_t - n\bar{t}\bar{z}}{\sum t^2 - n\bar{t}^2} = \frac{48024 - 16 \times 8,5 \times 327}{1496 - 16 \times 8,5^2} = \frac{3552}{340} = 10,447$$

Les ventes ont *tendance* à augmenter de 10 ou 11 milliers de bidules par trimestre. L'équation de la tendance s'écrit :

$$T = at + b \quad \text{avec} \quad b = \bar{z} - a\bar{t}$$

$$T = 10,447t + 238,20$$

Le coefficient 238,5 (ordonnée à l'origine) correspondrait à des ventes fictives de 238 milliers de bidules au trimestre 0.

On détermine ensuite les coefficients saisonniers en calculant les moyennes des rapports entre les ventes et la tendance pour une même saison.

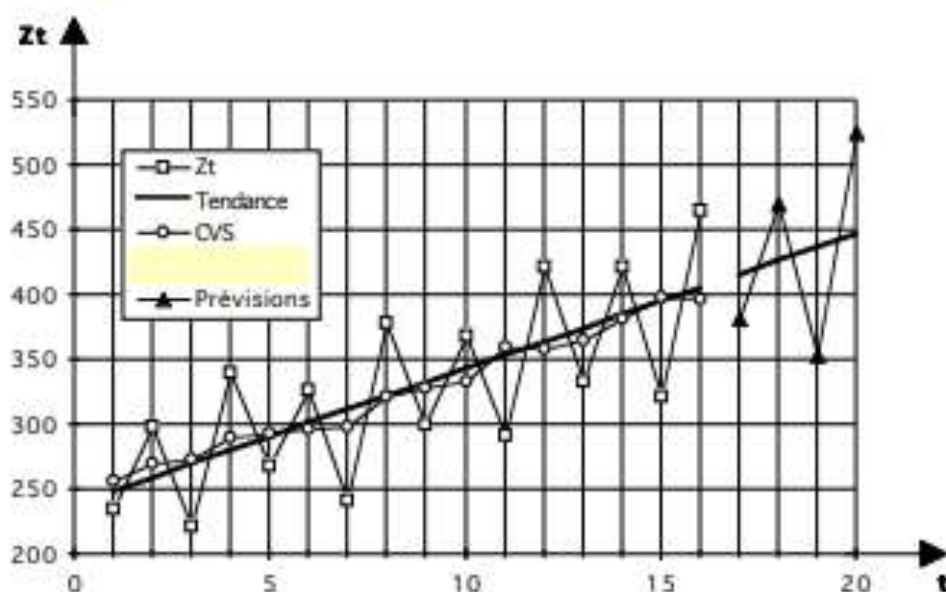
	années				coefficients saisonniers	Par rapport à la tendance
	1	2	3	4		
1 ^{er} trimestre	0,95	0,92	0,90	0,89	0,92	8% au dessous
2 ^e trimestre	1,15	1,09	1,07	1,10	1,10	10% au dessus
3 ^e trimestre	0,82	0,78	0,83	0,82	0,81	19% au dessous
4 ^e trimestre	1,21	1,17	1,16	1,15	1,17	17% au dessus

Les prévisions pour l'année 5 s'obtiennent en prolongeant la tendance et en appliquant chaque trimestre le coefficient saisonnier adéquat:

année	dates	Tendance	coef	Prévisions
5	17	416	0,92	381
	18	426	1,10	470
	19	437	0,81	354
	20	447	1,17	525

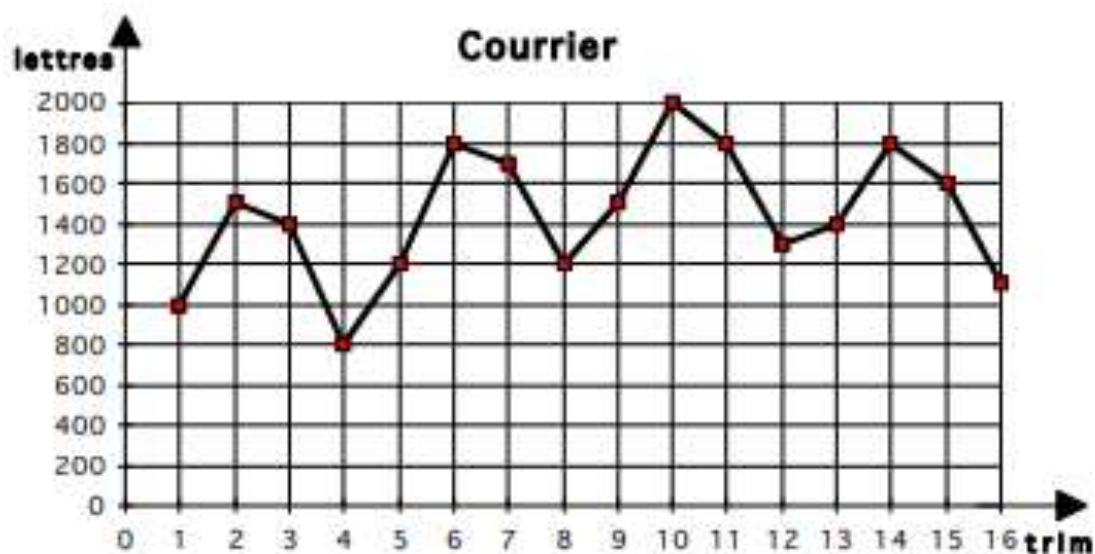
La série C.V.S. se calcule en divisant les ventes de bidules par le coefficient saisonnier du trimestre correspondant. On obtient ainsi le niveau de ventes fictif qui aurait été atteint en l'absence d'effet saisonnier. Cette série fait bien ressortir les aléas qui étaient masqués par l'effet saisonnier.

Finalement, on peut tracer les différentes courbes sur un même graphique:



3. méthode des moyennes mobiles

L'évolution trimestrielle d'escale sur un port donné se présente comme suit:



Pour une série saisonnière, il faut lisser sur une année entière avec un nombre impair de points:

$$M_t = \frac{0,5z_{t-2} + z_{t-1} + z_t + z_{t+1} + 0,5z_{t+2}}{4} \quad 3 \leq t \leq 14$$

Les calculs principaux sont regroupés dans un tableau:

trimestres	lettres	moy mob	rapports	coef	CVS
1	1000	xxx	xxx	0,91	1099
2	1500	xxx	xxx	1,23	1221
3	1400	1200	1,17	1,13	1239
4	800	1263	0,63	0,74	1087
5	1200	1338	0,90	0,91	1318
6	1800	1425	1,26	1,23	1466
7	1700	1513	1,12	1,13	1504
8	1200	1575	0,76	0,74	1630
9	1500	1613	0,93	0,91	1648
10	2000	1638	1,22	1,23	1628
11	1800	1638	1,10	1,13	1593
12	1300	1600	0,81	0,74	1766
13	1400	1550	0,90	0,91	1538
14	1800	1500	1,20	1,23	1466
15	1600	xxx	xxx	1,13	1416
16	1100	xxx	xxx	0,74	1495

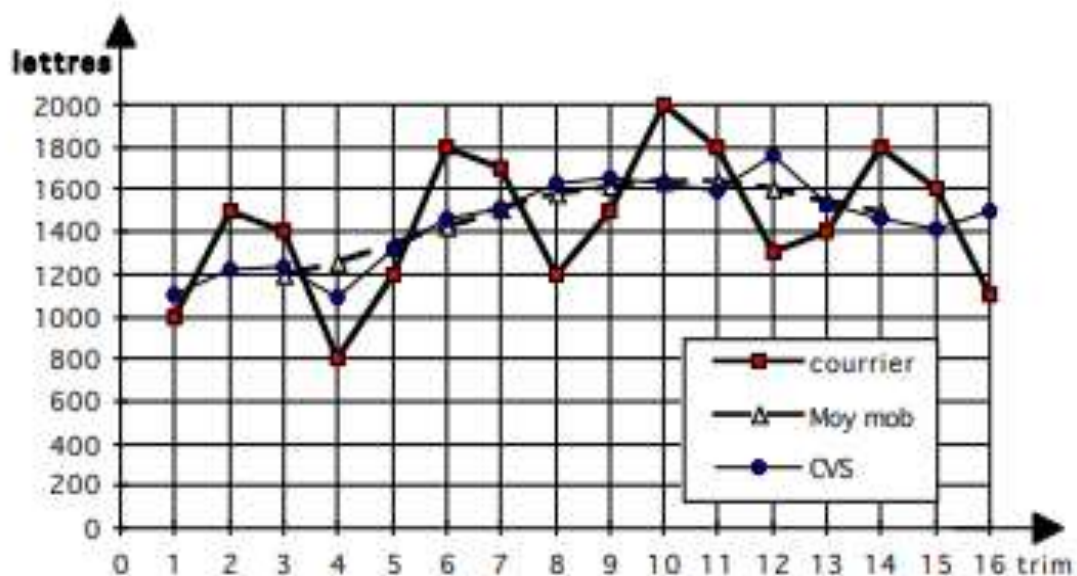
Les coefficients saisonniers sont les moyennes de rapports entre le nombre d'escale enregistré et la tendance pour chaque saison

	années				coefficients saisonnier
	1	2	3	4	
1 ^{er} trimestre	xxx	0,90	0,93	0,90	0,91
2 ^e trimestre	xxx	1,26	1,22	1,20	1,23
3 ^e trimestre	1,17	1,12	1,10	xxx	1,13
4 ^e trimestre	0,63	0,76	0,81	xxx	0,74

Les coefficients saisonniers s'interprètent par rapport à la tendance:

- au 1^{er} trimestre: 9% de moins en moyenne,
- au 2^e trimestre: 23% de plus,
- au 3^e trimestre: 13% de plus.
- au 4^e trimestre: 26% de moins.

Enfin, la courbe CVS (Corrigée des Variations Saisonnières) s'obtient en divisant le volume trimestriel d'escale par le coefficient saisonnier.



Deux points retiennent l'attention en s'écartant sensiblement de la tendance: ils correspondent au courrier reçu aux trimestres 4 et 12. Un spécialiste pourrait sans doute rattacher ces perturbations à des événements marquants dans la carrière du chanteur...

Deux points retiennent l'attention en s'écartant sensiblement de la tendance: Ils correspondent aux escales enregistrés aux trimestres 4 et 12. Un spécialiste pourrait sans doute rattacher ces perturbations à des événements marquants dans le management portuaire notamment la politique des prix des ports et du taux de fret armateurs...

CHAPITRE III : INDICATEURS DE PERFORMANCE PORTUAIRE

I. TABLEAU DE BORD DE PERFORMANCE PORTUAIRE

Graphique 2. Composants du tableau de bord de la performance portuaire



L'accent est mis sur quatre dimensions stratégiques (finance, opérations, ressources humaines et marché), afin d'appréhender les dynamiques de la performance portuaire moderne. Le graphique 5 décrit à la fois le déroulement de l'atelier et les dimensions stratégiques d'une autorité portuaire. Le rendement des opérateurs de fret fournit des données supplémentaires pour expliquer les niveaux de performance, en particulier pour les opérations et les ressources humaines.

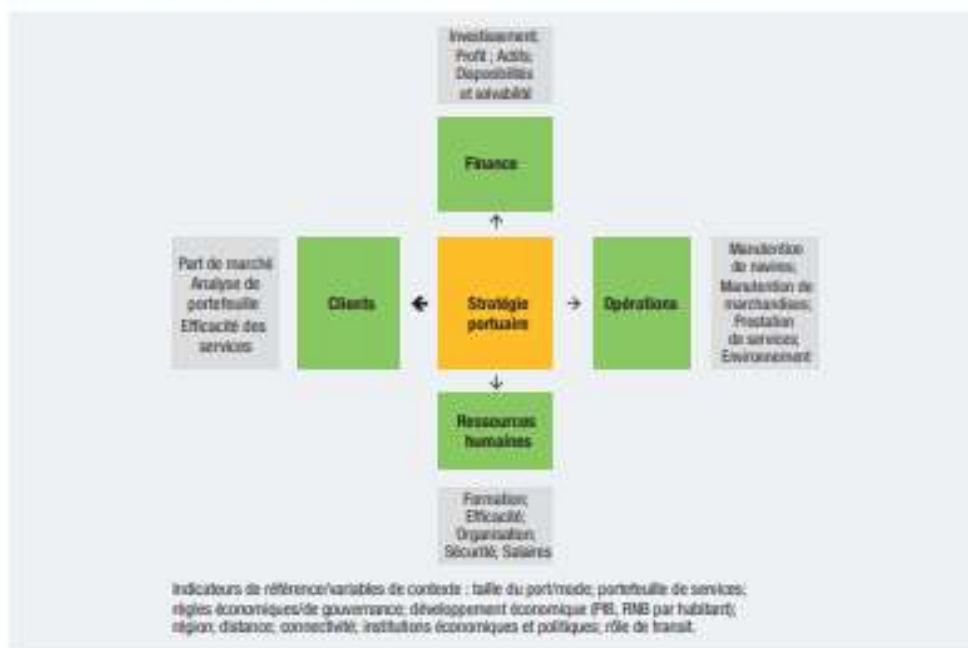
Le tableau de bord est analogue au tableau de bord équilibré de Kaplan et Norton, outil employé pour faire le lien entre stratégie et performance²⁴. L'intérêt de tels outils de gestion, qui sont internes à chaque organisation portuaire, n'est cependant pas à remettre en question. Le tableau de bord est un instrument heuristique précieux permettant de déterminer quels instruments sont importants et pourquoi. L'objectif est de se pencher sur les quatre dimensions stratégiques, et d'examiner les mesures génériques qui seront ou devraient être développées par toute autorité portuaire. Celles-ci seront nécessairement employées pour comparer les ports au niveau national et international.

Cela peut également constituer un instrument utile afin de décomposer la nature complexe du système portuaire, et de mener le projet en phases ou blocs distincts. Une description du tableau de bord est présentée ci-dessous ; elle englobe les dimensions relatives à la finance, aux opérations, aux ressources humaines et à la dimension client. Celles-ci sont examinées plus en détail dans le résumé des résultats du projet. Il est prévu d'inclure des mesures indiquant la durabilité environnementale, dans le cadre de la dimension opérationnelle. La section qui suit aborde donc la question de la durabilité économique.

2.4.1 Finance

L'objectif ultime est de mesurer la performance financière d'un port dans tous les services, bien qu'il s'agisse d'un objectif irréaliste dans l'ensemble, étant donné que de nombreux prestataires de services dans les ports fonctionnent selon des modèles d'entreprise à la fois public et privé. Toutefois, dans ce rapport, l'attention peut être dirigée sur les performances financières de l'entité communément désignée comme autorité portuaire. Pour ce faire, il

Graphique 5. Diagramme relatif au tableau de bord de la performance portuaire



Note: Le tableau de bord repose sur des mesures qui peuvent être comparées entre ports similaires.

est nécessaire de savoir précisément quels flux de revenus et coûts d'activité sont inclus dans le rapport financier. Les données requises seront construites par mode de chargement, et décomposées entre droits de port (navire et marchandises), frais de service (pilote et remorqueurs), accorage (manutention), activités à terre, telles que celles des propriétaires, et autres activités. Les notes explicatives sur la collecte de données, qui seront mises à la disposition des points de contact, incluront un ensemble de règles relatives à la distribution de la dimension finance, qui ont été adoptées lors du PPS 2.

Les coûts peuvent être plus difficiles à analyser ; cependant, il est possible de dissocier les coûts liés à la fonction de propriétaire/régulateur des coûts de fonctionnement. Les données requises dans le sondage ont été axées sur les flux à hauts revenus, exprimés en proportion des ventes totales, sur la marge d'exploitation excluant les dépenses ponctuelles et les coûts de financement des calculs, et sur la proportion des coûts de main-d'œuvre dans le revenu total associé à l'activité portuaire.

Il est prévu que des ateliers soient développés à l'avenir, afin de clarifier tout problème de définition

qui pourrait survenir. Les indicateurs de performance seront également élargis, pour prendre en compte la qualité des actifs et des investissements, la génération de liquidités, ainsi que des comparaisons relatives au financement. Les flux de revenus par mode constituent un autre niveau d'analyse, qui est nécessaire pour obtenir un prix moyen et dériver des comparateurs.

2.4.2 Opérations

Il s'agit du domaine le plus large et le plus utilisé pour évaluer la performance portuaire. Concernant les données primaires relatives au mouvement de navires, d'individus, de marchandises et d'informations, les mesures employées incluent les volumes déplacés au cours du temps, la surface, et la capacité technique. Celles-ci mettent l'accent sur le mode de chargement (gamme de produits) et les dimensions spatiales du lieu de manutention. Généralement, les indicateurs comparatifs renvoient aux taux de manutention des navires et des marchandises, ce qui produit des indicateurs de performance variant selon le temps et l'espace. Cet aspect rend particulièrement difficile la création de définitions communes.

Figure 6. Matrice relative à la performance sur le marché

Performance sur le marché		Part de marché relative	
		FAIBLE	ELEVÉE
Taux de croissance	ELEVÉ	Grand potentiel	Porto champions
	FAIBLE	Performance limitée	Leaders naturels

Dans les phases à venir, le tableau de bord de la performance portuaire inclura également d'autres dimensions opérationnelles, telles que l'énergie, l'environnement et la sécurité. Le développement du tableau de bord inclura également l'utilisation des technologies exploitant Internet (Internet des objets) comme moteur d'efficacité.

2.4.3 Ressources humaines

Ce groupe d'indicateurs est particulièrement intéressant dans le cadre des réseaux du Programme de Gestion Portuaire, étant donné que l'objectif principal du programme est de développer les ressources humaines afin d'améliorer la prestation de services portuaires, de faciliter le commerce, et ainsi d'accroître le bien-être économique. La collecte de données initiale a mis l'accent sur le nombre d'employés par catégories générales, sur la nature des arrangements institutionnels relatifs au travail, par exemple permanent ou temporaire, sur la parité hommes-femmes, et sur l'incidence du temps perdu en raison de grèves ou d'arrêts liés au travail. Le coût relatif de différents types de main-d'œuvre par rapport au coût total et/ou au revenu constitue une mesure comparative précieuse. Ces types de main-d'œuvre sont issus de l'autorité portuaire, de la manutention (privée ou publique) et des ressources temporaires communes. Grâce à cette approche, des indicateurs d'efficacité peuvent être calculés sur une base comparative solide.

Les participants du PPS 2 ont établi une matrice des catégories d'employés, qui sera incluse dans la prochaine phase de collecte de données (Annexe I). Les travaux futurs permettront d'élargir cette série

d'indicateurs, afin d'inclure un ensemble plus large de catégories d'employés et d'identifier leur relation par rapport aux modes de chargement. Les indicateurs de formation et les dimensions de santé et de sécurité présentent un intérêt particulier pour les responsables des ports et les parties intéressées, dans le cadre d'analyses comparatives.

2.4.4 Clients

Il s'agit d'un domaine qui n'est pas couvert dans la première phase du projet, étant donné que la collecte de données est plus coûteuse et problématique. Néanmoins, il peut être décomposé entre satisfaction des clients, part de marché par mode de chargement, efficacité des services, notamment avec le modèle SEAPORT²⁴, et analyse de portefeuille basée sur la valeur ajoutée pour différents types de trafic, tels que le mode.

L'analyse de portefeuille vise à identifier la position sur le marché de chaque port, dans la matrice présentée dans le graphique 6, qui indique les changements au cours du temps de la part de marché du port par rapport à la croissance du marché²⁵. Le fait d'adapter cette approche permettra d'obtenir une base pour mesurer la performance en termes de marché portuaire et de position concurrentielle relative de chaque produit portuaire. La matrice repose sur une période unique, et ne prend pas en compte les changements au cours du temps.

²⁴ Hiney, 2014.

²⁵ R Cruz, J Ferreira and S Azevedo, 2012. A dynamic strategic portfolio analysis: Positioning Iberian seaports, South African Journal of Business Management, 43(1):33-43; également tiré d'une matrice développée par le Boston Consulting Group.

2.4.5 Contexte social et économique

Lors de la première phase, des données fixes ou semi-fixes relatives aux droits de propriété, à la gouvernance économique, aux modèles fonctionnels, et aux dimensions du marché, ont été requises. Comme il a été mentionné précédemment, il est indispensable de différencier les ports en fonction des conditions locales, afin de pouvoir mettre en contexte tout rapport d'évaluation comparatif.

Le contexte local ne repose pas uniquement sur des variables internes. Il existe de nombreuses variables externes permettant de mettre en contexte une analyse comparative. Il s'agit généralement de données accessibles au public, telles que les conditions macroéconomiques au cours du temps, la connectivité et la proximité par rapport aux marchés, et l'ensemble des règles relatives à l'économie politique en lien avec le commerce maritime local. Cette dernière catégorie inclut le contrôle des capitaux, les procédures douanières, le droit du travail, les régimes d'investissement et la facilité de faire des affaires, ainsi que l'ouverture relative du système politique.

Le fait d'inclure de telles variables contextuelles permet de représenter la performance par rapport à des indices d'ouverture économique et politique ; ou toute combinaison des variables contextuelles. Sur de tels graphiques, il est possible de voir la performance des ports les uns par rapport aux autres, tout en tenant compte des conditions échappant au contrôle direct de gestion.

II. INDICATEURS DE RÉFÉRENCE POUR LES PORTS

Table 2. Tableau de bord complet

Entité portuaire uniquement	Indicateurs (23)	Valeur N (p x t)	Moyenne	Min	Max
Finances	Marge opérationnelle (EBITDA)	44	38%	-75%	83%
	Droits de port associés aux navires/Revenu	42	18%	1%	32%
	Droits de port associés aux marchandises/Revenu	41	38%	10%	63%
	Loyers/Revenu	39	10%	1%	57%
	Main-d'œuvre/Revenu	38	23%	7%	63%
	Charges etc/Revenu	30	6%	0.1%	23%
Ressources humaines	Tonnes/Employé	34	38 435	4 202	204 447
	Revenu/Employé	38	\$179 971	\$138	\$1 039 739
	EBITDA/Employé	33	\$93 556	-\$16 696	\$555 835
	Coût de main-d'œuvre/Employé	24	\$23 863	\$4 489	\$93 589
	Coût de formation/ Salaires	33	0.95%	0.03%	4.60%
Opérations liées aux navires	Temps d'attente moyen (heures)	62	17	0	89
	LHT moyenne par navire (m)	55	136	44	289
	Tirant d'eau moyen par navire (m)	55	8	2	22
	JB moyenne par navire	66	14 260	552	43 216
Opérations liées aux marchandises	Nombre de tonnes par arrivée en moyenne - Toutes marchandises	41	4 739	201	20 510
	Tonnes par heure de travail VS	28	116	20	350
	Boîtes par heure - LoLo	46	18	8	35
	Durée d'immobilisation EVP (jours)	29	7	3	18
	Tonnes par heure VL	16	42	17	63
	Tonnes par hectare - Toutes marchandises	41	173 986	75 772	425 800
	Tonnes par mètre de quai - Toutes marchandises	41	3 920	890	7 439
Données IOM	Qualité de l'infrastructure portuaire	Année 2014	3.8	2.7	5.2
	Indice de connectivité des transports maritimes par lignes régulières		21.9	11.8	33.6
	Charges des procédures douanières		3.5	1.8	4.4
	Taux de participation des femmes		21%	6%	44%

III. ÉTUDE DE CAS

Tableau 3: Etude de cas

Entité portuaire uniquement	Indicateurs (23)	Valeur M (p x 0)				Moyenne	Min	Max	Part X				
									2010	2011	2012	2013	2014
Finance	Marge opérationnelle (EBITDA)	44	38%	-75%	83%				49%	58%	60%	59%	
	Droits de port associés aux navires/Revenu	42	18%	1%	32%				16%	16%	14%	15%	15%
	Droits de port associés aux marchandises/Revenu	41	38%	10%	63%				40%	40%	42%	42%	59%
	Loyers/Revenu	39	10%	1%	57%				1%	1%	1%	1%	7%
	Main-d'œuvre/Revenu	38	23%	7%	63%				19%	18%	18%		10%
	Charges ébs/Revenu	30	6%	0.1%	23%				0.1%	0.4%	0.4%	0.3%	1.0%
Ressources humaines	Tonnes/Employé	34	38 435	4 202	204 447				20 174	21 683	21 809	21 873	7 074
	Revenu/Employé	38	\$179 971	\$138	\$1 039 739				\$101 599	\$113 418	\$128 492	\$138 730	\$17 983
	EBITDA/Employé	33	\$93 556	-\$16 096	\$555 835				\$30 265	\$65 587	\$76 985	\$81 464	
	Coût de main-d'œuvre/Employé	24	\$23 863	\$4 489	\$93 589				\$19 198	\$20 982	\$23 580		
	Coût de formation/ Salaires	33	0.95%	0.03%	4.60%				0.2%	0.2%	0.2%	1.0%	
Opérations liées aux navires	Temps d'attente moyen (heures)	62	17	0	89				0	0	0	0	4
	Temps moyen passé à quai	55	136	44	289								67
	LHT moyenne par navire (m)	55	8	2	22					3	3	3	3
	Tirant d'eau moyen par navire (m)	66	14 260	552	43 216				2 212	2 066	2 555	2 710	2 219
Opérations liées aux marchandises	Nombre de tonnes par arrivée en moyenne - Toutes marchandises	41	4 739	201	20 510				335	382	400	412	
	Tonnes par heure de travail VS	28	116	20	350				24	24	24	24	20
	Balles par heure - LdLo	46	18	8	35				20	20	20	20	25
	Durée d'immobilisation EYP (jours)	29	7	3	18				6	6	6	10	11
	Tonnes par heure VL	16	42	17	63				40	40	40	40	
	Tonnes par hectare - Toutes marchandises	41	173 986	75 772	425 800				221 914	233 865	239 895	240 604	
	Tonnes par mètre de quai - Toutes marchandises	41	3 920	890	7 439				6 264	6 601	6 771	6 791	
Données IDM	Qualité de l'infrastructure portuaire	Année 2014				3.8	2.7	5.2	3.0	3.3	3.4	3.5	
	Indice de connectivité des transports maritimes par lignes régulières					21.9	11.8	33.6	18.6	17.2	18.1	20.3	
	Charges des procédures douanières					3.5	1.8	4.4	3.0	3.2	3.2	3.5	
	Taux de participation des femmes					21%	6%	44%				11%	

3.6 Tableau de bord de la performance portuaire : Etude de cas

Ce cas, tiré de la base de données, offre un aperçu des problèmes liés à la collecte de données, mais également des résultats disponibles pour les responsables des ports (tableau 3). Il a été difficile de rassembler certaines mesures relatives aux navires pour ce port. Celles-ci ne sont pas cruciales pour son exploitation et ne sont donc pas collectées actuellement. Les deux indicateurs, la longueur totale et le tirant d'eau du navire, seront modifiés lors du prochain cycle de collecte de données, pour rendre compte de cette difficulté, commune parmi les participants du PPS 2. Concernant les mesures financières, il existe deux lacunes, qui traduisent des problèmes d'échéances et de confidentialité. Par exemple, pour certains ports, un accord formel sera nécessaire afin de diffuser de telles informations, qui ne sont normalement pas visibles dans les comptes publiés. En termes d'indicateurs de référence, il est utile d'analyser le tableau de bord afin d'appréhender la performance du port au cours de la période.

Les scores financiers semblent positifs, en comparaison avec les autres ports du réseau. La marge d'exploitation est nettement supérieure à la moyenne du réseau. Cela suggère que l'exploitation est profitable, et qu'elle permet d'auto-financer le développement de l'infrastructure. La palette de droits de port diffère des valeurs moyennes ; néanmoins, lorsque ceux-ci sont combinés, la moyenne obtenue reste dans le même intervalle. Le revenu associé aux charges immobilières est bas, ce qui indique qu'il s'agit d'une entité portuaire fondée sur un modèle de service complet. Les coûts de main-d'œuvre sont compétitifs

pour un port moderne ; en effet, ils représentent 19% du revenu total, et diminuent au cours de la période.

Concernant les données sur la main-d'œuvre, l'une des difficultés consiste à tirer des conclusions dépassant la simple comparaison des moyennes du réseau. Le port affiche des chiffres égaux ou inférieurs à la moyenne pour toutes les catégories. Cela peut être le signe d'une main-d'œuvre nombreuse, recevant un salaire inférieur à la moyenne. Au cours du prochain cycle de collecte de données, des informations précieuses pourront être obtenues à partir des moyennes par catégorie professionnelle, et des comparaisons par rapport au salaire moyen des pays.

Il manque certaines données relatives aux navires, pour ce port. Il peut toutefois être conclu que la taille moyenne des bateaux et le tirant d'eau sont faibles ; de plus, le temps d'attente ne semble pas constituer un problème. Cela suggère que les petits navires ordinaires ne rencontrent pas de difficultés liées à la disponibilité des postes d'amarrage. Cette situation se traduit, au niveau du rendement des marchandises, par une moyenne faible par navire. L'utilisation des quais semble être satisfaisante, étant donné le fort taux de rotation des marchandises en termes d'hectares et de mètres par quai. Ceci est cohérent avec le temps d'attente nul pour les navires.

Concernant la productivité de la manutention, le tableau de bord du port se situe dans la partie basse de l'intervalle, pour la manutention de vrac secs. Cela peut être dû aux marchandises ou aux équipements de manutention utilisés pour des chargements de petite taille. Le port est marqué par des taux de manutention et des temps de séjour des marchandises qui le placent en bonne position par rapport à l'ensemble du réseau. Concernant les chargements en vrac liquide, le taux de manutention est comparable à la moyenne du réseau.

Les ports étudiés jusqu'à présent sont similaires en termes de type, par exemple pour ce qui est des droits de propriété et de la taille. Toutefois, à mesure que le réseau s'étend, il sera possible de différencier les valeurs moyennes du réseau en fonction de la taille et du modèle fonctionnel. Cette étude de cas suggère que la taille du port est telle que l'investissement dans des équipements de manutention à plus haute rotation pourrait ne pas être justifié sur le plan commercial.

3.7 Mesures externes

Les indicateurs du tableau de bord permettent de comprendre le contexte économique, ainsi que la perception que les utilisateurs ont des ports. Les mesures sont élaborées sur la base d'unités nationales, afin de dégager la performance d'un port de premier plan, possiblement situé dans une capitale, plutôt que celle de tous les ports du réseau.

Le tableau de bord comprend trois valeurs :

- La qualité de l'infrastructure portuaire est un index élaboré par le Forum Economique Mondial, et dont l'échelle s'étend de 1 à 7. Plus le score est élevé, plus le port est développé et efficace, d'après les standards internationaux. Les ports déclarants se situent au milieu de l'intervalle pour cet index ;
- L'indice de connectivité des transports maritimes par lignes régulières est calculé par la CNUCED. Il s'agit d'un indice de valeurs basé sur le nombre de connexions proposées par un port. L'année de référence est 2004, année durant laquelle un score de 100 a été attribué au port avec le niveau le plus haut de connectivité. Chaque année, l'indice est calculé par rapport à cette valeur de référence. Les valeurs sont relativement faibles parmi les ports déclarants. Elles reflètent la taille relative des ports, ainsi que leur position par rapport aux réseaux maritimes globaux. Il est donc probable que les ports soient des ports de transbordement ou des hubs régionaux par nature. Cet indice constitue un point de référence intéressant, car il souligne les différences entre les principaux hubs internationaux, mais également l'évolution des ports dans un contexte de développement économique ;
- Les charges des procédures douanières sont un autre indicateur du Forum Economique Mondial, qui oscille entre 1 et 7. L'indice met en évidence les différents niveaux d'efficacité, estimés par les experts impliqués dans l'étude. Encore une fois, les valeurs se situent approximativement au milieu de l'intervalle. Il est intéressant de noter que les deux mesures développées par le Forum Economique Mondial sont fortement corrélées parmi tous les pays appartenant aux réseaux du Programme de Gestion Portuaire, y compris ceux ne participant pas

actuellement au projet de tableau de bord de la performance portuaire. Il est hautement probable que les participants à l'enquête partagent une vision commune de l'efficacité des ports, basée sur leur expérience avec les douanes, et vice versa.

Les participants du PPS 2 ont adopté une proposition visant à analyser la relation entre la durée de séjour des marchandises et la durée de dédouanement déclarée. Ce ratio fournira une mesure relative de la charge de temps contrôlable par le port, affectant les coûts de transit et la congestion dans des ports modernes.

Les outils de données en ligne de la CNUCED, tels que UNCTADStat, sont également des ressources précieuses pour les responsables des ports cherchant des variables de contexte pour des études comparatives ou d'autres activités de recherche¹⁸.