

Cours en Analyse de données séquentielles

Chapitre 3: Fouille de séries temporelles

Dr D. AKROUR

**2ème année Master
Systèmes d'Information, Optimisation et Décision**

2023-2024

Plan

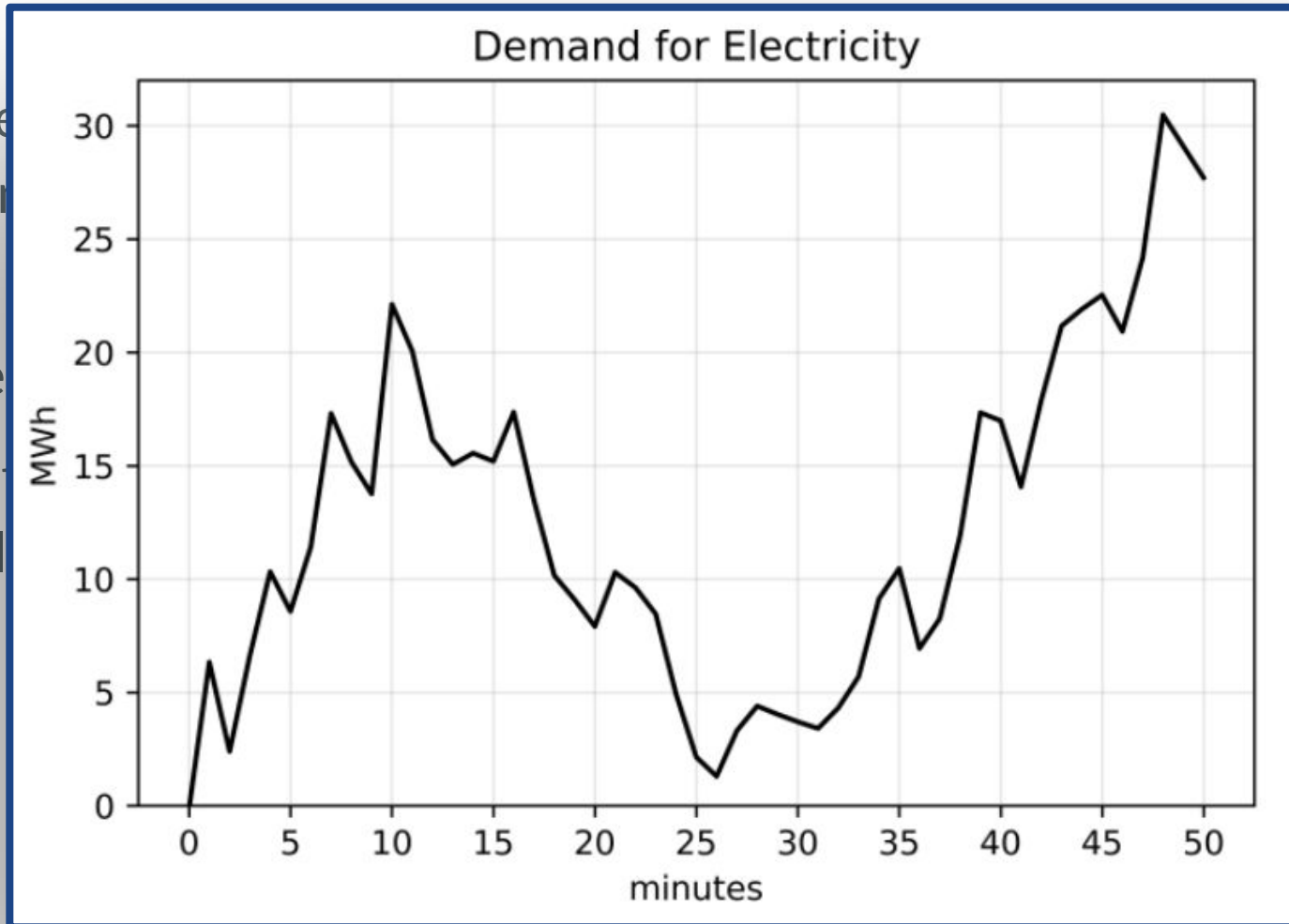
- Définition
- Domaines d'applications
- Indices descriptifs d'une série temporelle
- Composants d'une série temporelle
- Estimation de la tendance
- Mesures de similarité
- Recherche de similarité dans une série temporelle
- Recherche de motifs fréquents dans une série temporelle
- Clustering des séries temporelles
- Requête par contenu des séries temporelles
- Classification des séries temporelles
- Détection d'anomalies dans les séries temporelles

Définition

- Une **série temporelle** est une séquence ordonnée de valeurs d'une variable prises à des intervalles de temps équidistants
- On peut dire qu'elle représente l'évolution d'une quantité spécifique au cours du temps.
- Mathématiquement : elle est définie par les valeurs $y_1, y_2, \dots, y_n$ de la variable y aux pas de temps $t_1, t_2, \dots, t_n$ (($y=f(t)$))

Définition

- Une
- d'un
- On
- spé
- Ma
- de l



rs

... yn

Définition

- Voici quelques exemples de séries temporelles :
 - Température moyenne mensuelle
 - Bénéfice annuel d'une entreprise
 - Prix journalier du carburant
 - Consommation d'électricité d'une maison par heure
 - Ventes de maisons trimestrielles

Définition

Une série temporelle est caractérisé par:

- La **fréquence**: le temps écoulé entre y_{t+1} et y_{t+1} ainsi on a:
 - Les données annuelles
 - Les données semestrielles
 - Les données mensuelles
 - Les données hebdomadaires
 - Les données journalières
 - Les données intra-journalières (heure, minute, seconde)
- La **durée**: La période du temps au cours de laquelle les données ont été collecté

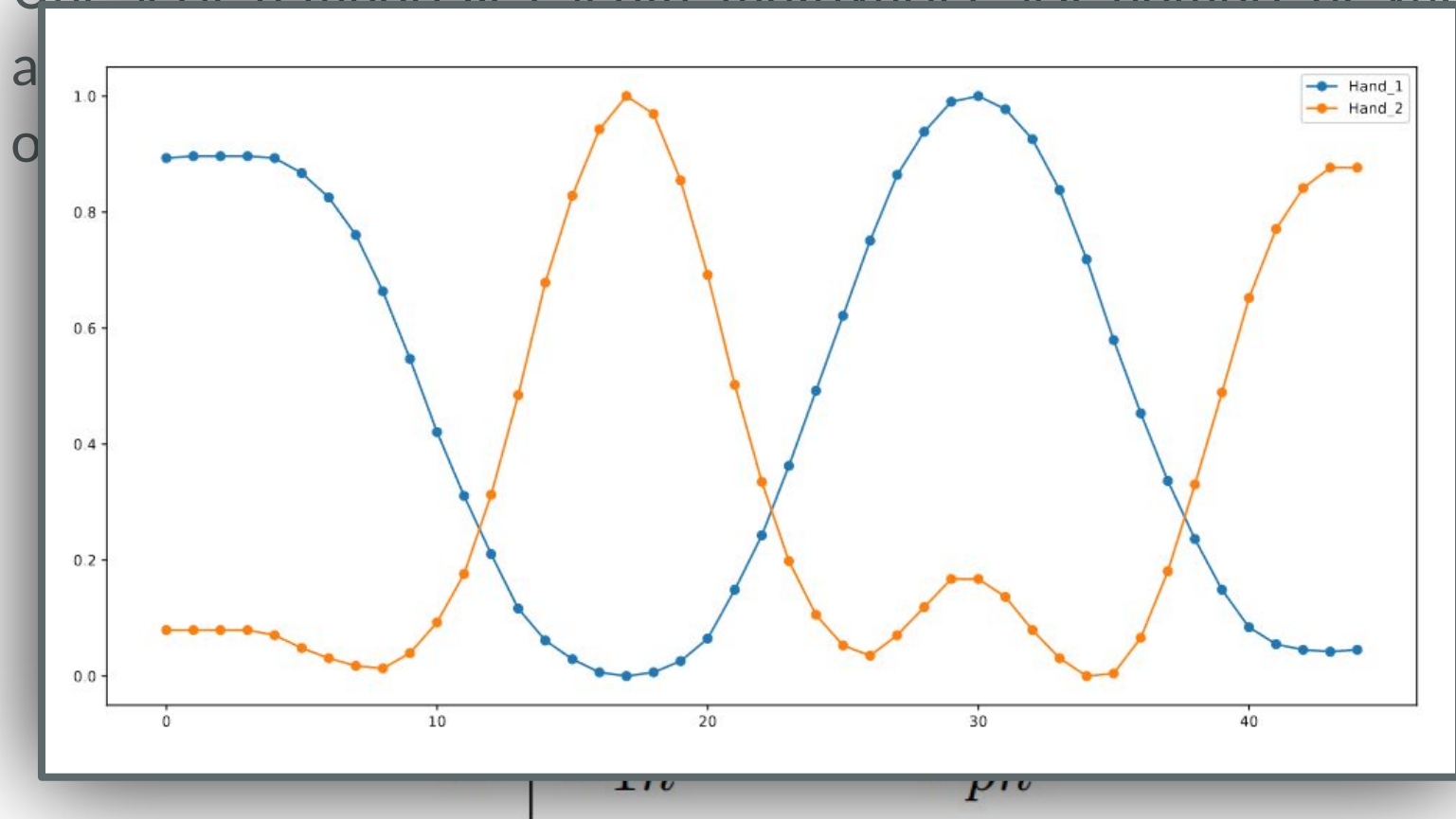
Série temporelle univarié et multivarié

- Une série temporelle est dite **multivariée** si le nombre de valeurs associées à chaque période est supérieur à 1. Dans ce cas les observations sont vectorielles

	X_1	$\dots$	X_p
1	x_{11}	$\dots$	x_{p1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	x_{1n}	$\dots$	x_{pn}

Série temporelle univarié et multivarié

- Une série temporelle est dite multivariée si le nombre de valeurs



Objectifs

Les séries temporelles sont utilisées pour effectuer les opérations suivantes :

- **Comprendre le comportement passé des données observées**
 - Analyser, décrire et expliquer un phénomène au cours du temps et en tirer des conséquences pour des prises de décision (marketing...).
 - Fouille de motifs fréquents
 - Détecter les comportements inhabituels
- **Etablir un modèle de prévision pour prévoir le comportement future**
 - Assister les spécialistes dans leurs décisions
 - Planifier les opérations futures
 - Comparer les performances réelles aux performances attendues et analyser les causes de variations

Domaines d'application

- Finance et économétrie : évolution des indices boursiers, des prix, des données économiques des entreprises, des ventes et achats de biens, des productions agricoles ou industrielles,
- Assurance : analyse des sinistres,
- Médecine/biologie : suivi des évolutions des pathologies, analyse d'électrocardiogrammes,
- Sciences de la terre et de l'espace : indices de marées, variations des phénomènes physiques (météorologie),
- Traitement du signal : signaux de communications, de radars, de sonars, analyse de la parole,
- Traitement des données : mesures successives de position ou de direction d'un objet mobile (trajectographie)

Indices descriptifs d'une série temporelle

Il est bien utile de disposer de quelques mesures numériques qui résument une série temporelle :

- **Indices de tendance centrale**
 - Moyenne
- **Indices de dispersion**
 - Variance
 - Ecart type
- **Indices de dépendance**
 - Auto-covariance
 - Auto-corrélation

Indices de tendance centrale

- La tendance **centrale** d'une série temporelle est donnée par la **moyenne empirique**

$$\bar{y}_n = \frac{1}{n} \sum_{t=1}^n y_t$$

Indice de dispersion

- La **variance** (et sa racine carrare, l'**écart-type** empirique) décrit la **dispersion** des données de la série temporelle autour de sa moyenne

$$Variance = \sigma^2 = \frac{1}{n} \sum_{t=1}^n (y_t - \bar{y})^2$$

Indice de dispersion

- La **variance** (et sa racine carrare, l'**écart-type** empirique) décrit la **dispersion** des données de la série temporelle autour de sa moyenne

$$Ecart\ type = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \bar{y})^2}$$

où:

- y_t sont les valeurs de la série temporelle aux instants t .
- n est le nombre total d'observations.
- $\bar{y}$ est la moyenne des observations.

Indice de dispersion

- une valeur élevée de variance ou d'écart type indique que les données sont dispersées autour de la moyenne,
- tandis qu'une faible valeur signifie que les données sont concentrées près de la moyenne.

Indice de dispersion

- u
- le
- ta
- co

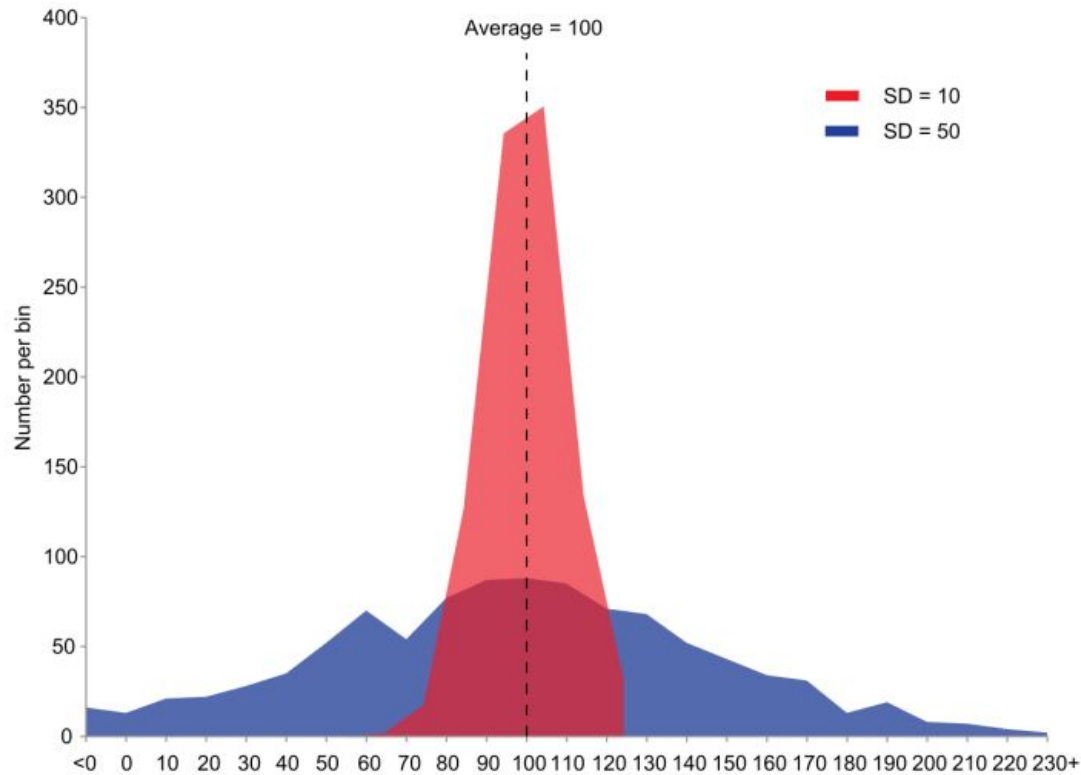


Figure 3.1. Exemple de deux échantillons ayant la même moyenne (100) mais des écarts types différents illustrant l'écart type comme mesure de la dispersion autour de la moyenne².

Indice de dépendance

Covariance et Corrélation

Elles mesurent la relation linéaire entre deux variables x et y

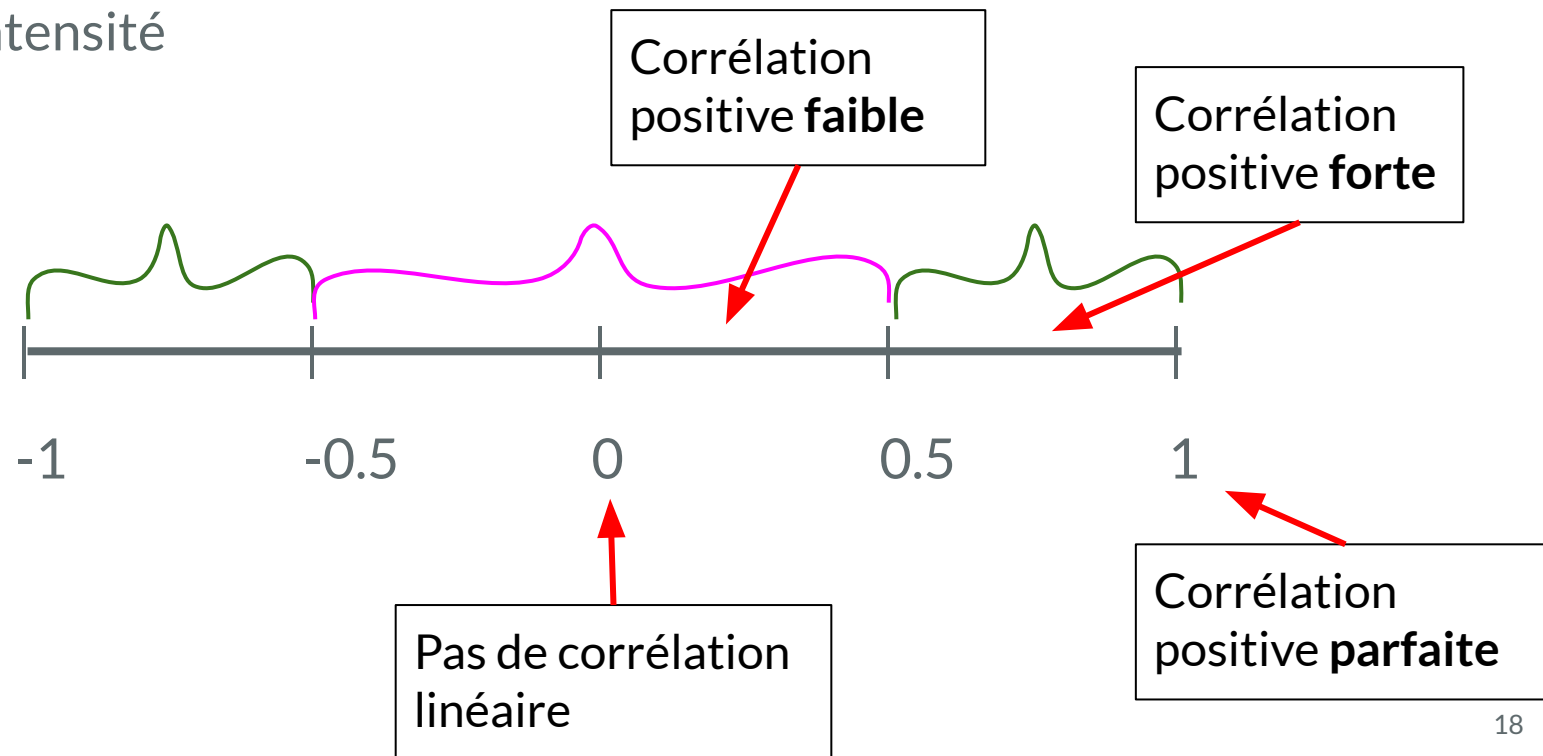
- La **covariance** indique uniquement la relation en terme de direction
- Si la valeur de la covariance est:
 - nulle : il n'y a pas de relation linéaire
 - positive: il y a une dépendance positive entre x et y ((plus x augmente plus y augmente ou plus x diminue y diminue))
 - négative: il y a une dépendance négative entre x et y ((plus x augmente plus y diminue ou plus x diminue y augmente))

Indice de dépendance

Covariance et Corrélation

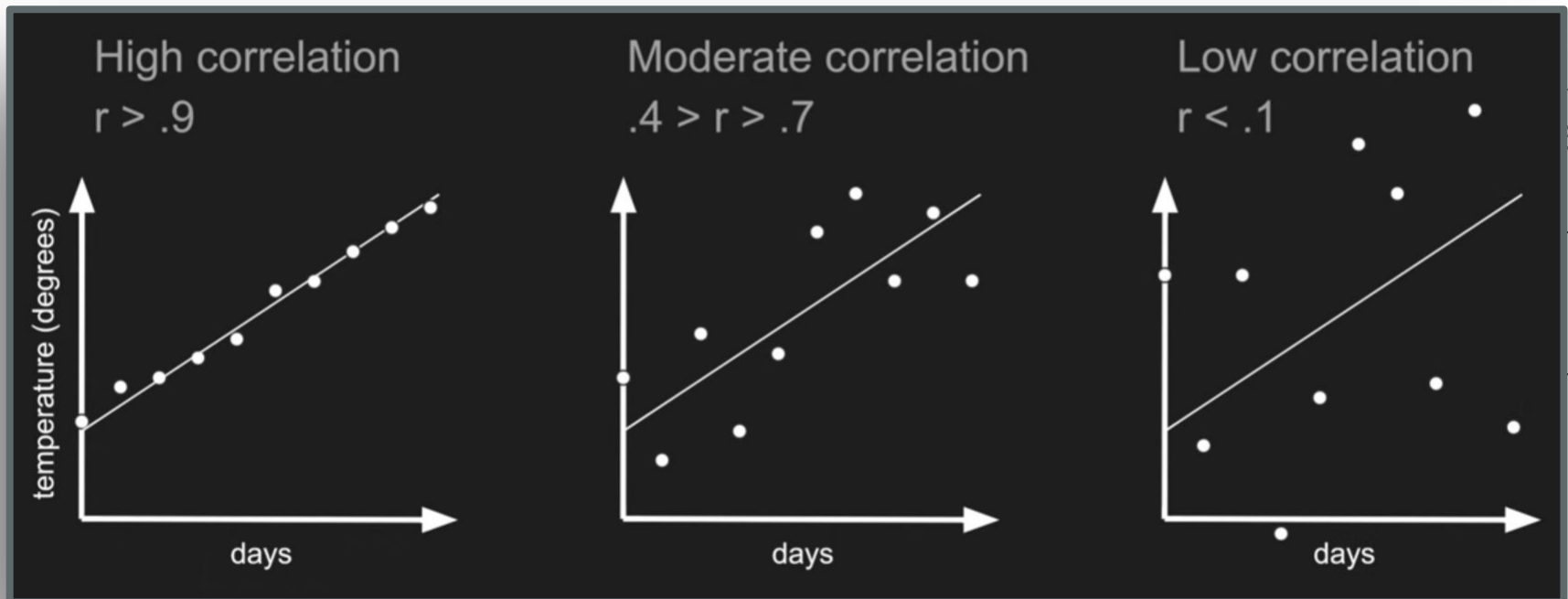
Elles mesurent la relation linéaire entre deux variables x et y

- La **corrélation** indique la relation en terme de direction et en terme d'intensité



Indice de dépendance

Covariance et Corrélacion



Pas de corrélation
linéaire

Corrélation
positive **parfaite**

Indice de dépendance

Autocovariance et Autocorrélation

étudier la relation pour une même variable à différents instants dans le temps quantifient comment une valeur à un instant t est liée à une autre valeur de la même série à l'instant $t + L$, où L représente le décalage temporel (ou Lag).

- **Autocovariance**
 - Indique la relation entre deux valeurs s'il s'agit d'une dépendance positive ou négative.
- **Autocorrélation**
 - En plus de la dépendance selon la direction (autocovariance), l'autocorrélation indique également s'il s'agit d'une dépendance forte ou faible.

Indice de dépendance

Autocovariance et Autocorrélation

étudier la relation pour une même variable à différents instants dans le temps
quantifient comment une valeur à un instant t est liée à une autre valeur de la même série à l'instant $t + L$, où L représente le décalage temporel (ou Lag).

$$\text{Autocovariance}(Y, L) = \frac{1}{n-L} \sum_{t=1}^{n-L} (y_t - \bar{y})(y_{t+L} - \bar{y})$$

$$\text{Autocorrélation}(Y, L) = \frac{\text{Autocovariance}(Y, L)}{\text{Variance}}$$

- y_t et y_{t+L} sont les valeurs de la série temporelle aux instants t et $t + L$,
- $\bar{y}$ est la moyenne de la série temporelle,
- n est le nombre total d'observations.

Indice de dépendance

Autocovariance

- **positive** indique une relation positive : lorsque la valeur d'une observation à l'instant t est élevée, la valeur à l'instant $t + L$ a également tendance à être élevée. Inversement, si elle est faible à t , elle tend également à être faible à $t + L$.
- **négative** relation inverse où les valeurs de la série temporelle fluctuent de manière opposée. Lorsque la valeur d'une observation à l'instant t est élevée, la valeur observée à l'instant $t + L$ tend à être faible, et vice versa
- **proche de zéro** signifie qu'il n'y a pas de relation linéaire évidente entre les valeurs à ces instants, ce qui suggère une indépendance statistique pour ce décalage temporel.

Indice de dépendance

Autocorrélation

- L'autocorrélation se situe toujours dans l'intervalle $[-1, 1]$.
- Plus les valeurs se rapprochent de 1, plus cela indique une forte dépendance positive.
- Inversement, si elles se rapprochent de -1, cela signifie une forte dépendance négative.
- Lorsque les valeurs se rapprochent de 0, la dépendance diminue en intensité jusqu'à devenir inexistante.

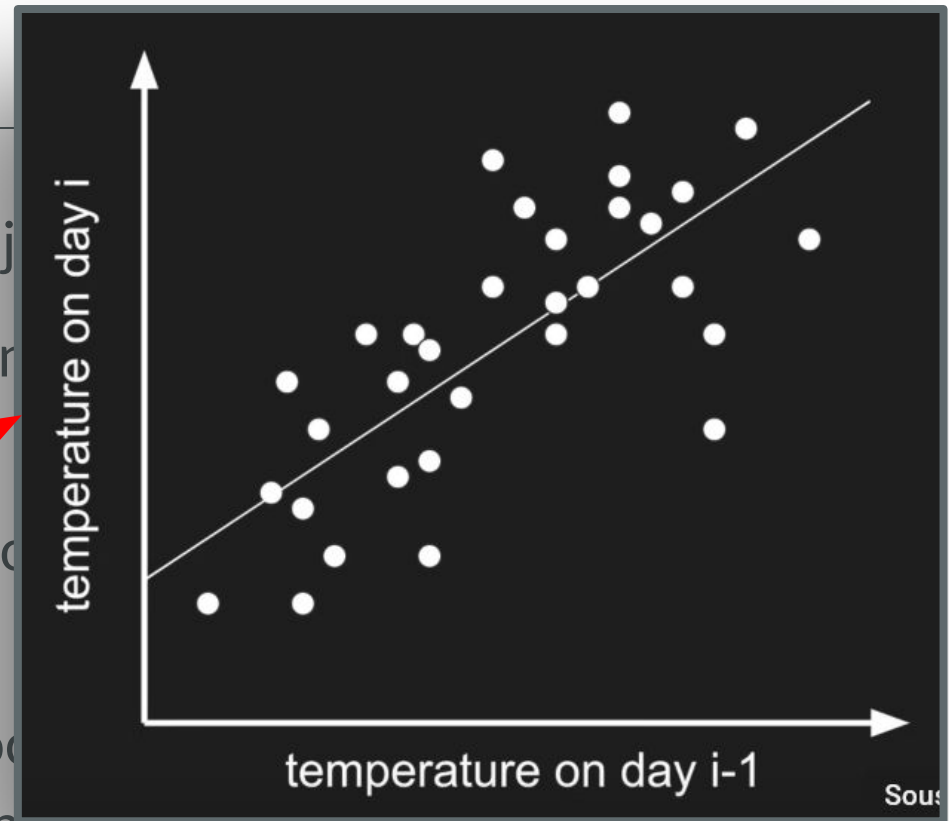
Indice de dépendance

Autocorrélation

- L'autocorrélation se situe toujours entre -1 et 1.

On compare la température d'un jour avec la température du jour d'avant. On remarque une corrélation positive car le changement de températures est faible d'un jour à l'autre.

- Lorsque les valeurs se rapprochent de 0, la corrélation diminue en intensité jusqu'à devenir inexistante.



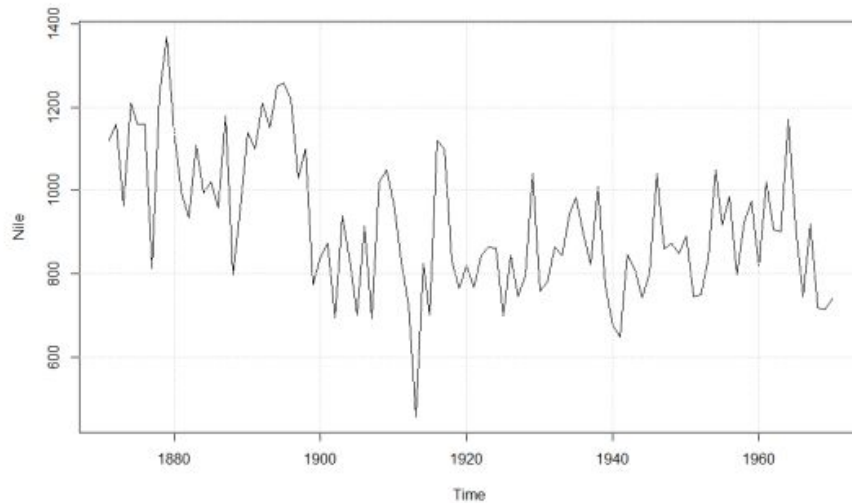
Fonction d'autocorrélation (ACF)

- La fonction d'autocorrélation (ACF) calcule l'autocorrélation d'une série temporelle pour plusieurs valeurs de décalage (ou lags).
- Par exemple, l'ACF au lag 1 mesure la corrélation entre chaque valeur et la valeur qui la suit immédiatement.
- Au lag 2, elle indique la corrélation entre des valeurs séparées par deux pas de temps, et ainsi de suite.

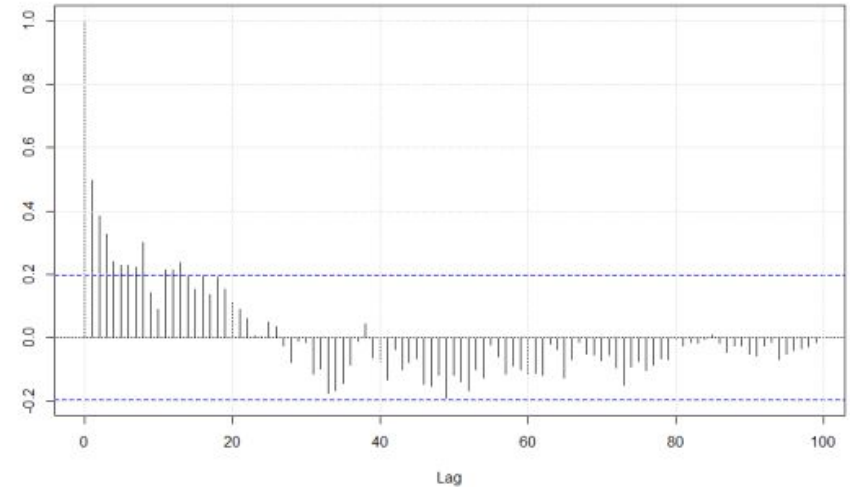
Fonction d'autocorrélation (ACF)

- L'autocorrélation est généralement positive et significative (les barres dépassent les intervalles de confiance) pour de très petites valeurs de L , car les valeurs adjacentes dans la plupart des séries temporelles sont souvent similaires.
- Toutefois, cette corrélation diminue progressivement à mesure que le décalage L augmente, car les observations éloignées dans le temps deviennent de moins en moins liées

Fonction d'autocorrélation (ACF)



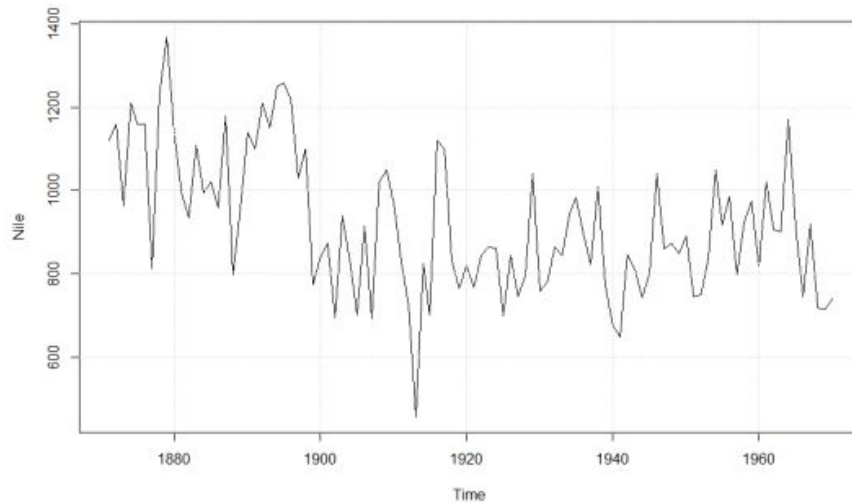
(a) Plot de la série temporelle



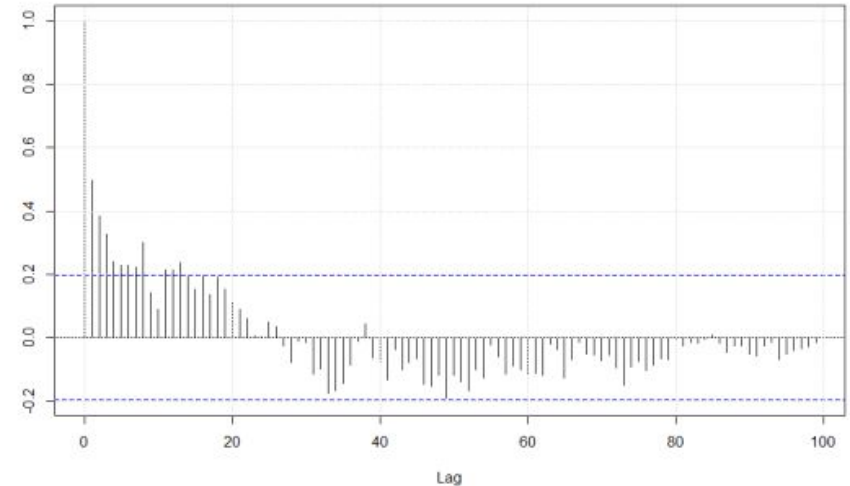
(b) Graphique d'autocorrélation correspondant

Figure 3.6. Débit annuel du Nil à Assouan, de 1871 à 1970.

Fonction d'autocorrélation (ACF)



(a) Plot de la série temporelle



(b) Graphique d'autocorrélation correspondant

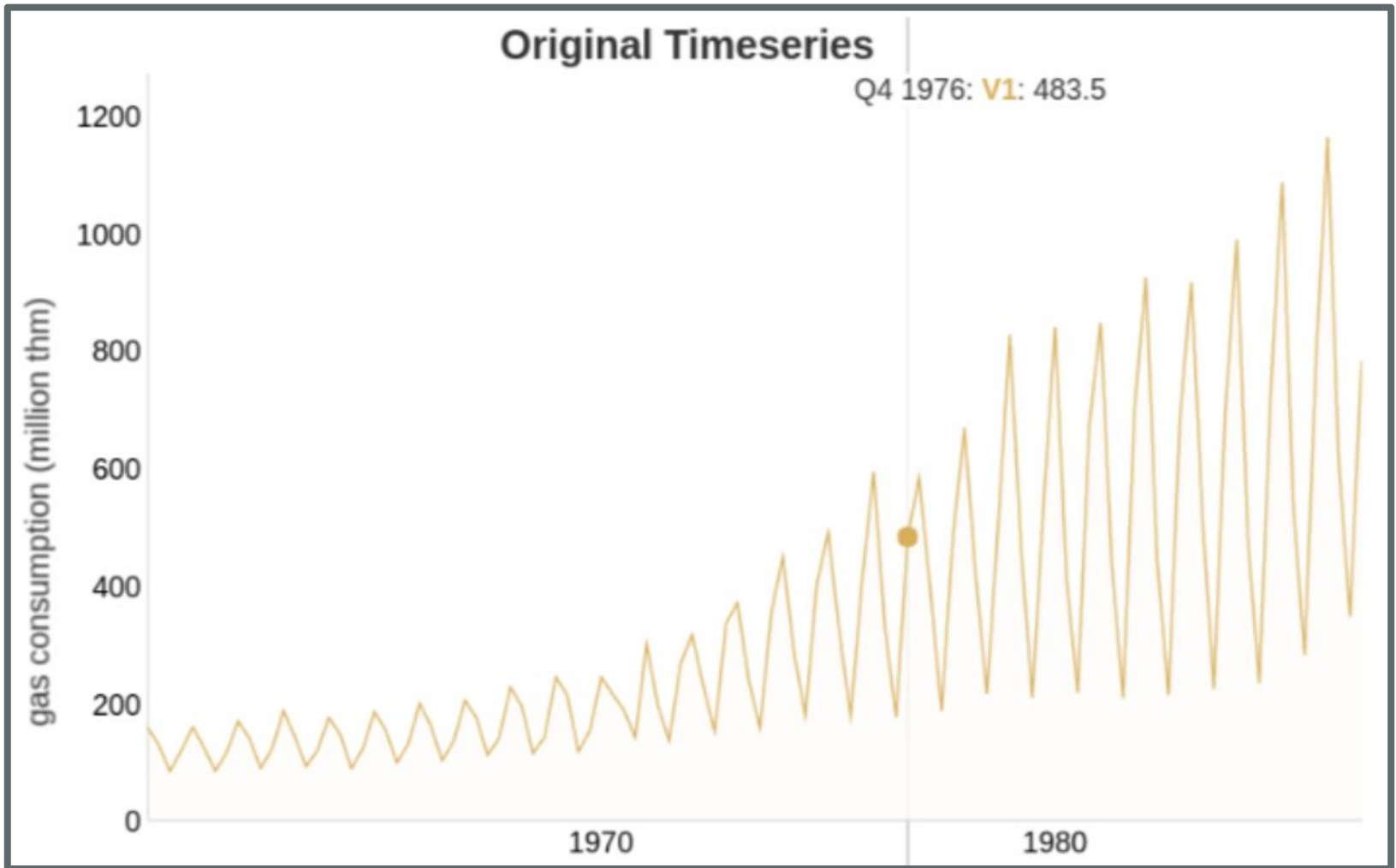
Figure 3.6. Débit annuel du Nil à Assouan, de 1871 à 1970.

Nous remarquons que le débit d'une année est souvent proche de celui de l'année précédente, ce qui entraîne une dépendance positive moyenne pour un lag de une année. Par contre, l'autocorrélation est nulle ou négligeable dans la série.

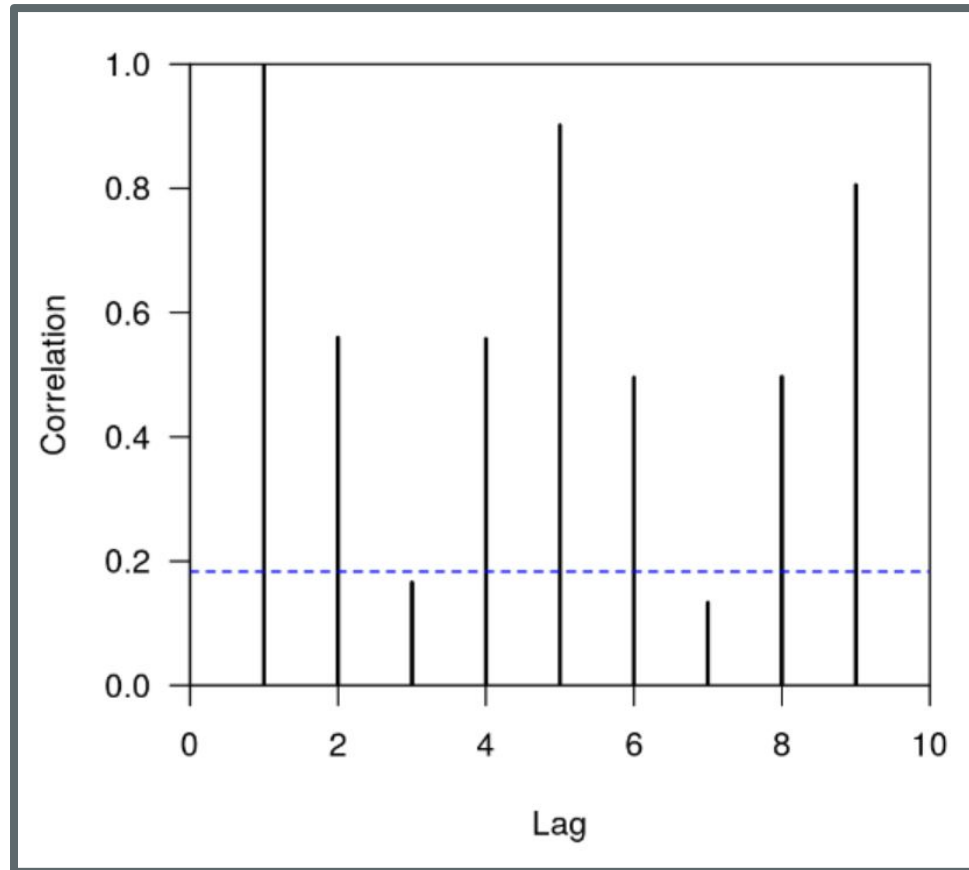
Fonction d'autocorrélation (ACF)

- Utilité des ACT
 - aide à détecter des motifs répétitifs (des cycles saisonniers ou périodiques dans les données...)
 - ainsi, on peut faire des prévisions puisque les valeurs passées à ces décalages influencent fortement la valeur actuelle.

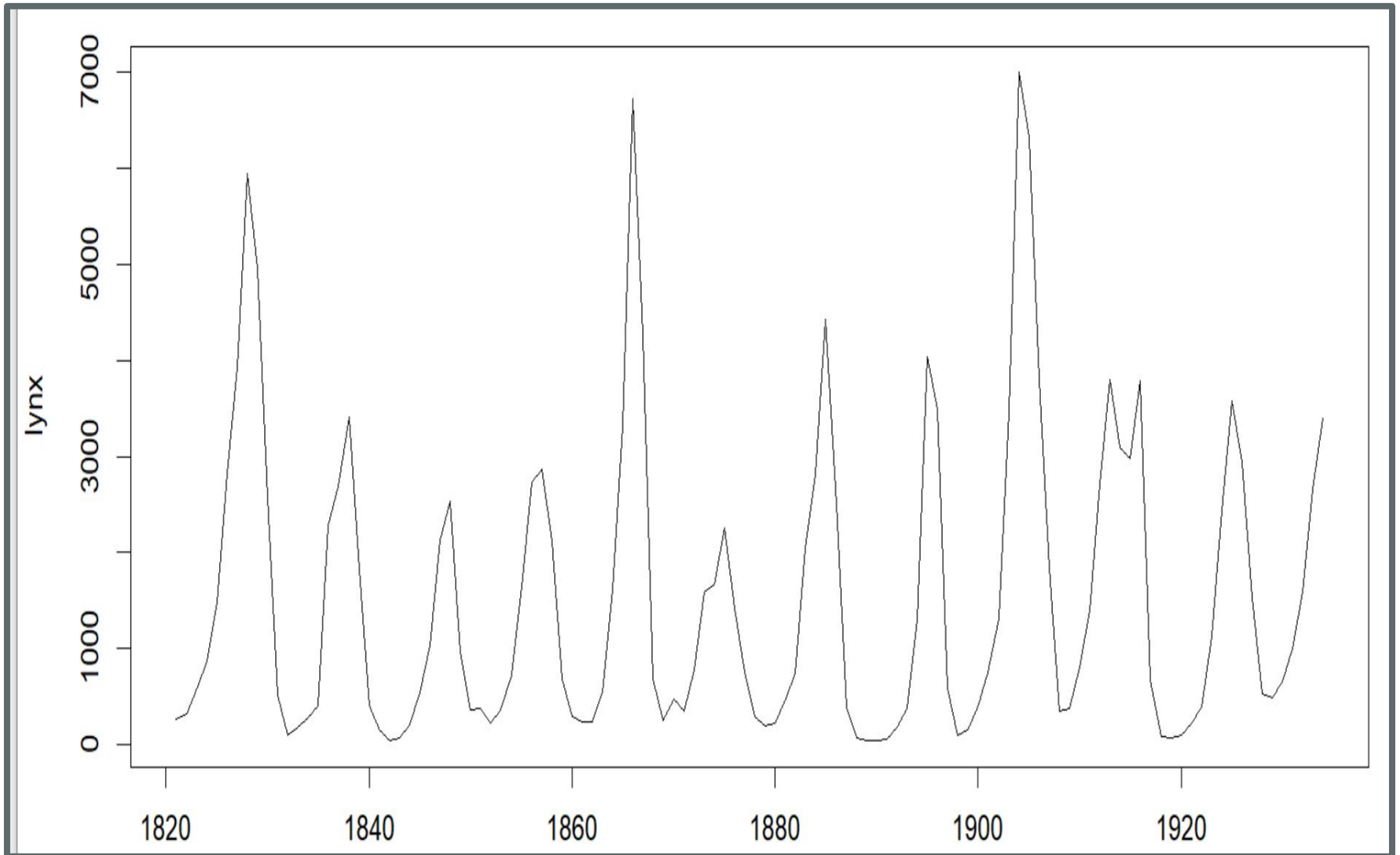
Exemple d'un série temporelle et sa ACF



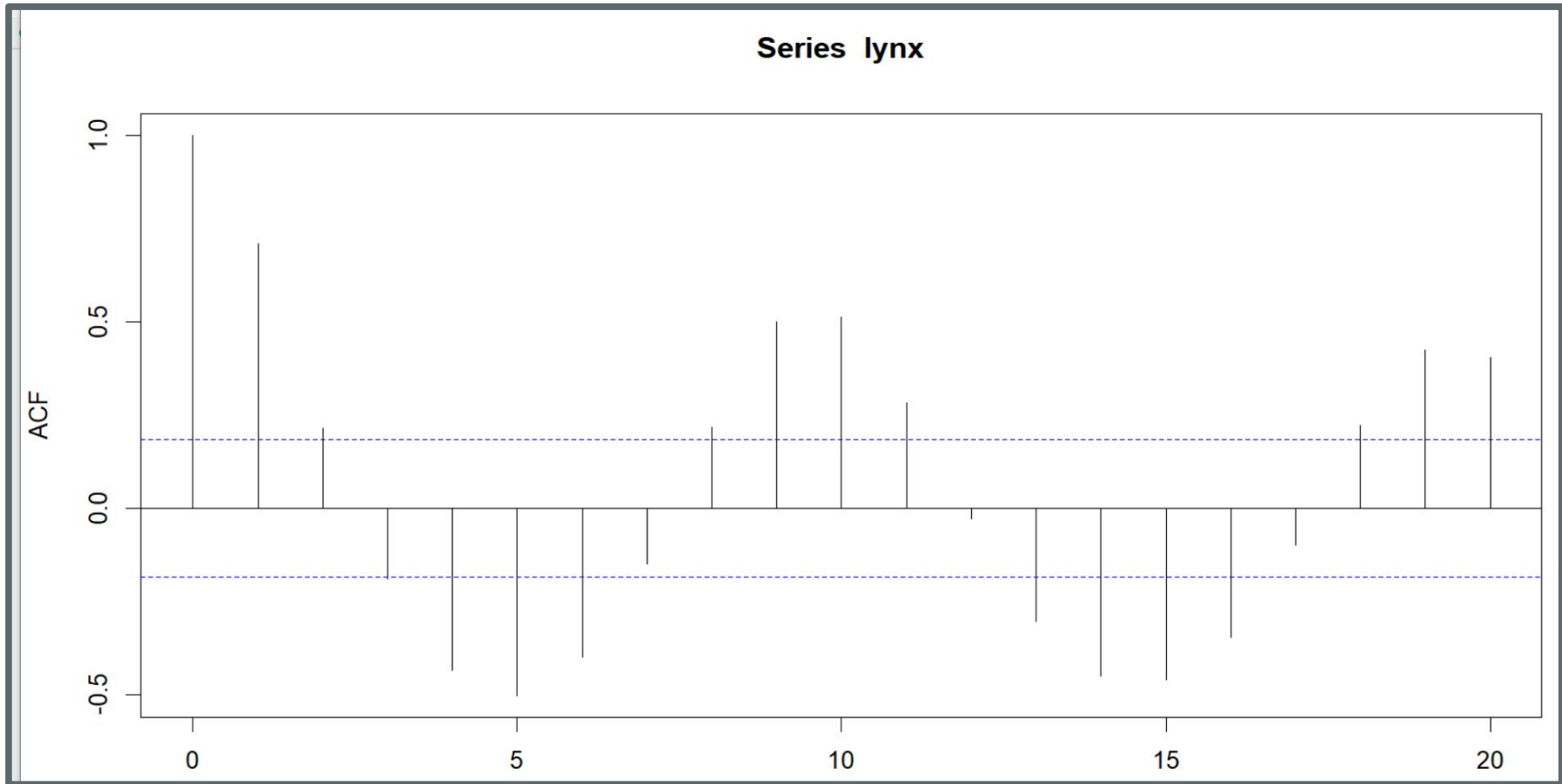
Exemple d'un série temporelle et sa ACF



Exemple d'un série temporelle et sa ACF



Exemple d'un série temporelle et sa ACF



Composants d'une série temporelle

- Une série temporelle peut être décomposée en une série de composantes qui décrivent sa structure:
 - **Tendance**
 - **Saisonnalité** ou **variation saisonnière**
 - **Cycle**
- L'intérêt de ceci est d'une part de mieux comprendre, de mieux décrire l'évolution de la série, et d'autre part de prévoir son évolution

Tendance

- La tendance mesure la variation à long terme
- La tendance est l'orientation générale d'une série d'observations à la hausse ou à la baisse sur une longue période de temps
- Lorsqu'il n'existe pas d'orientation, on dit qu'il n'y a pas de tendance, ce qui ne signifie évidemment pas que toutes les valeurs sont les mêmes...

Saisonnalité

- La saisonnalité se produit lorsqu'une série temporelle est affectée par des facteurs saisonniers tels que la saison de l'année ou le jour de la semaine.
- C'est un comportement répétitif de courte durée qui se produit à intervalles réguliers fixes.
- Exemple
 - les ventes de jouets peuvent montrer une saisonnalité plus élevée pendant les mois de vacances
 - En météorologie, températures plus faibles en hiver qu'en été.
 - En économie, saisonnalité induite par les périodes de vacances, les périodes de fêtes, le climat..

Cycle

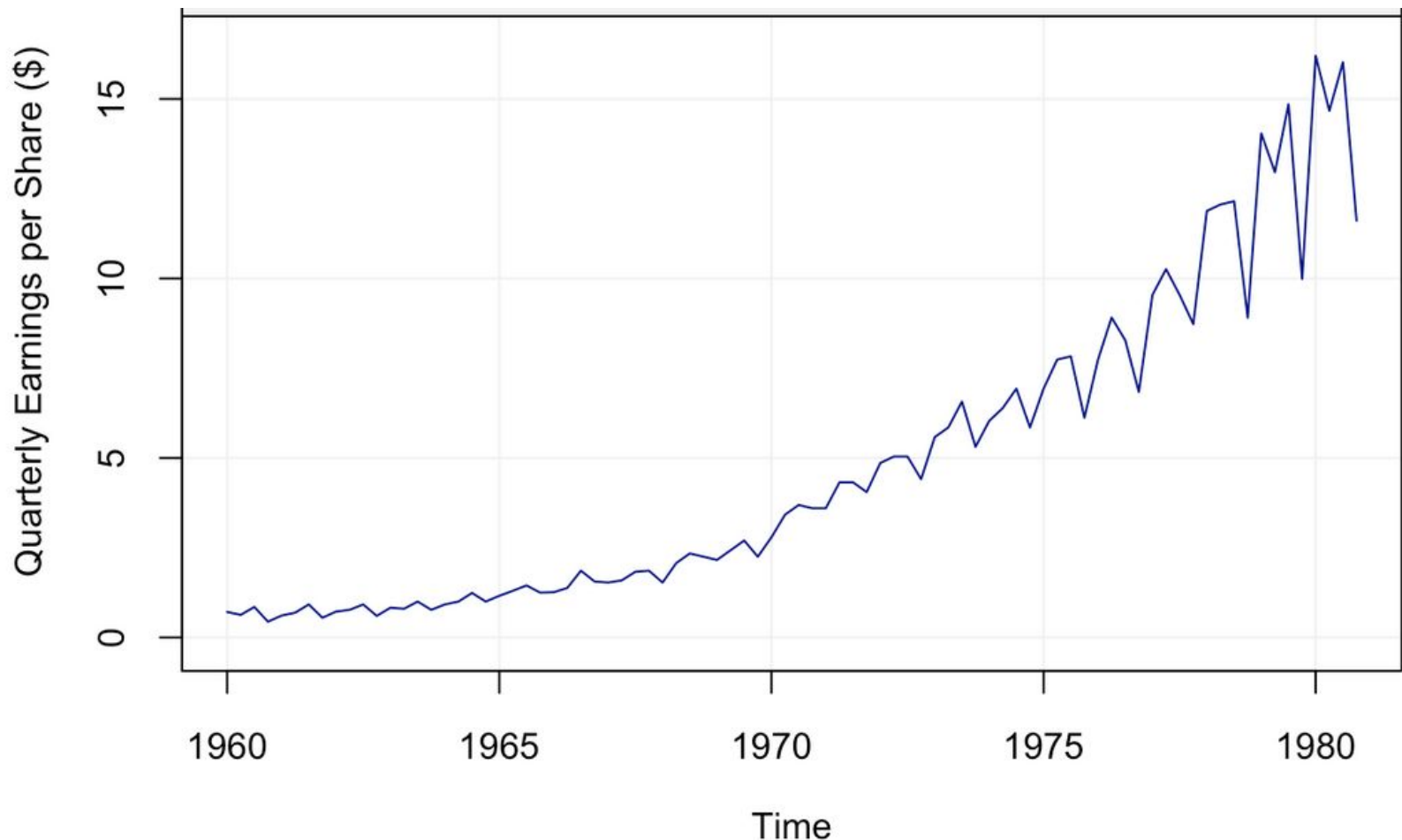
- Le cycle représente des fluctuations qui ne sont pas d'origine saisonnière.
- Contrairement à la saisonnalité, le cycle n'a pas une période fixe. Il peut être associé à des facteurs économiques, des cycles d'affaires, ou d'autres tendances à long terme.

Visualisation

- La visualisation est une étape très importante dans l'analyse des séries temporelles
- La visualisation des séries temporelles nous permet d'observer de nombreuses caractéristiques des données (les motifs, les observations inhabituelles, les changements dans le temps et les relations entre les variables)

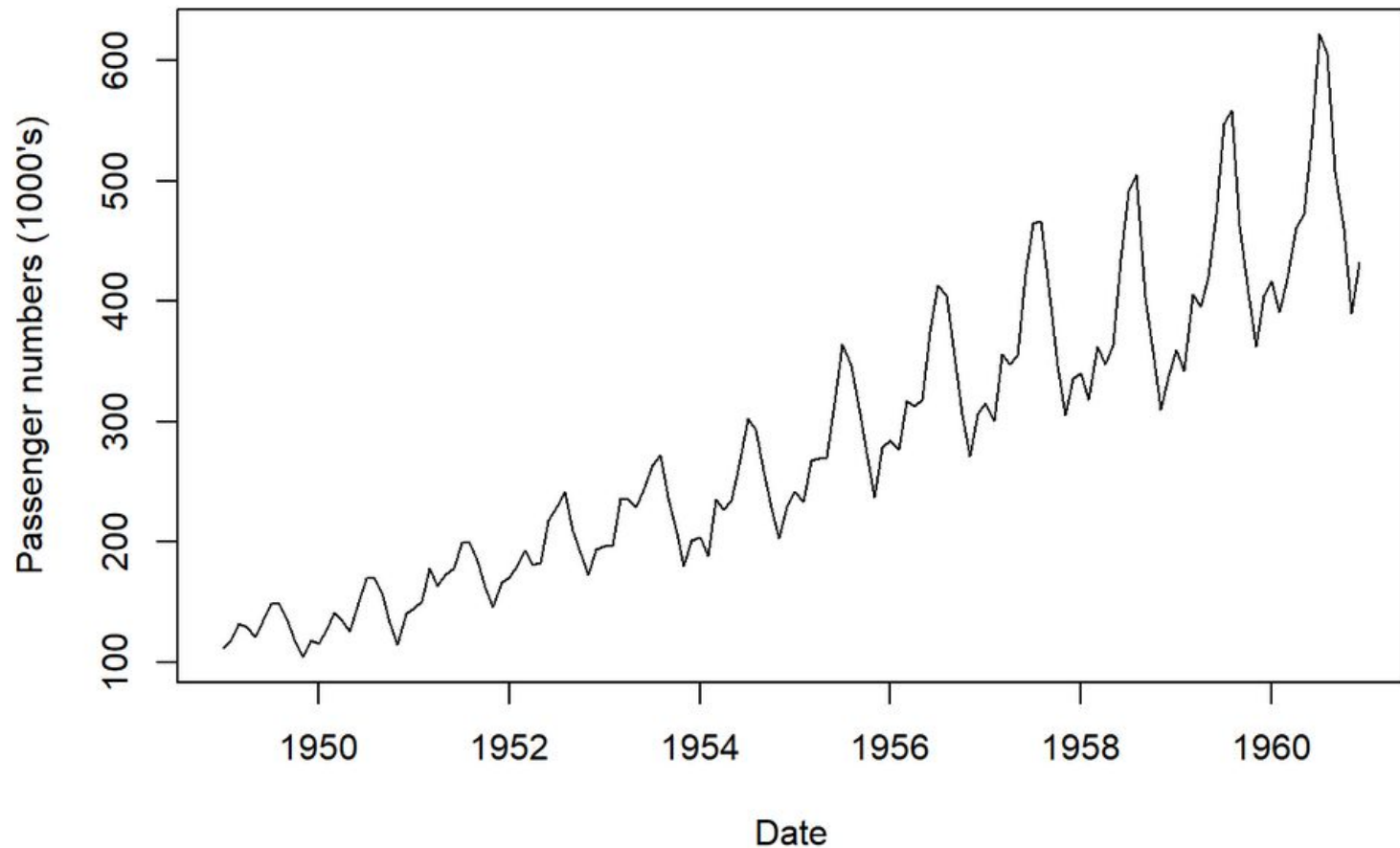
Exemples de séries temporelles

- Gain saisonniers de Johnson & Johnson ((tendance en hausse))



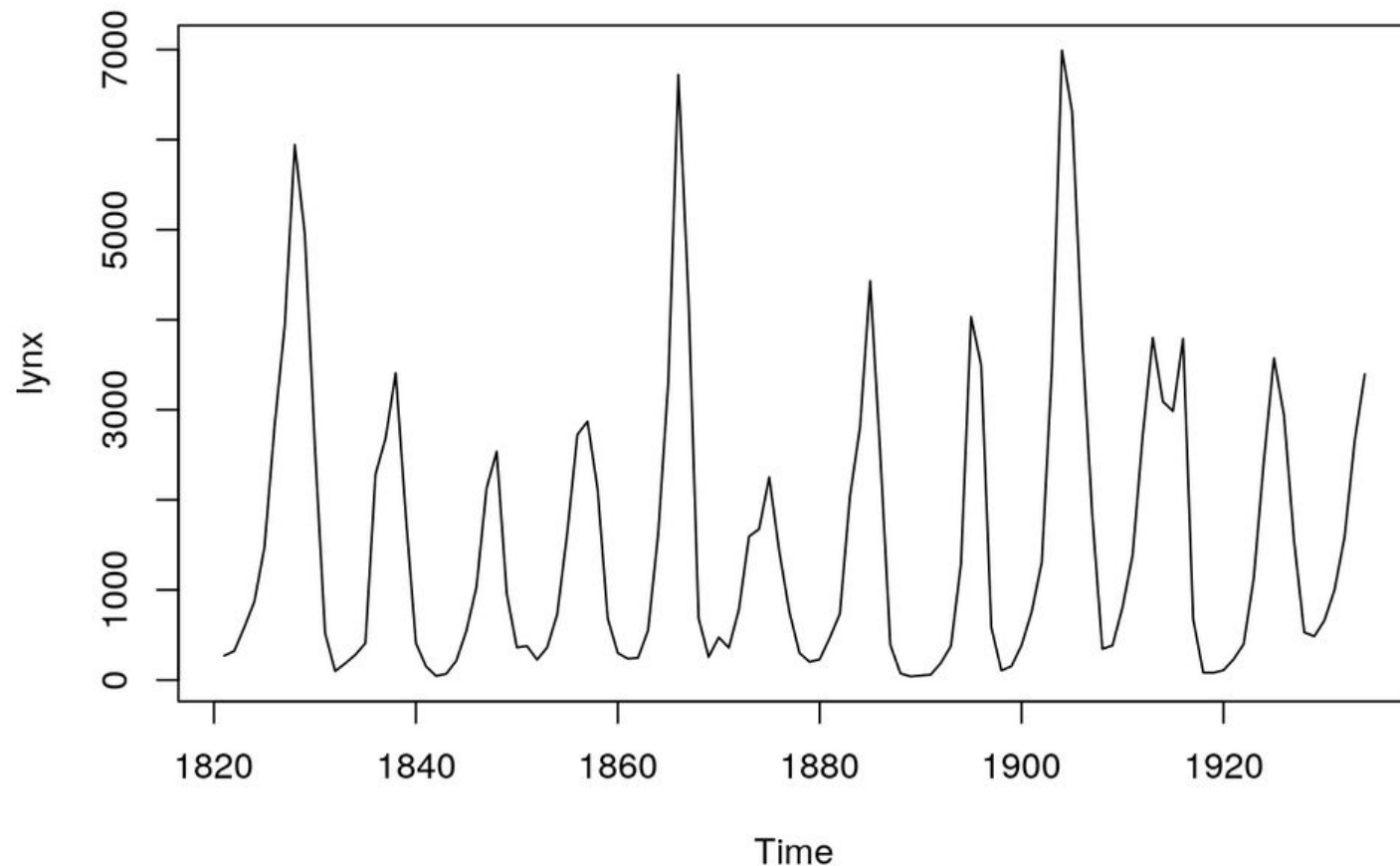
Exemples de séries temporelles

- Nombre de passagers aériens mensuels ((la saisonnalité semble augmenter avec la tendance générale))



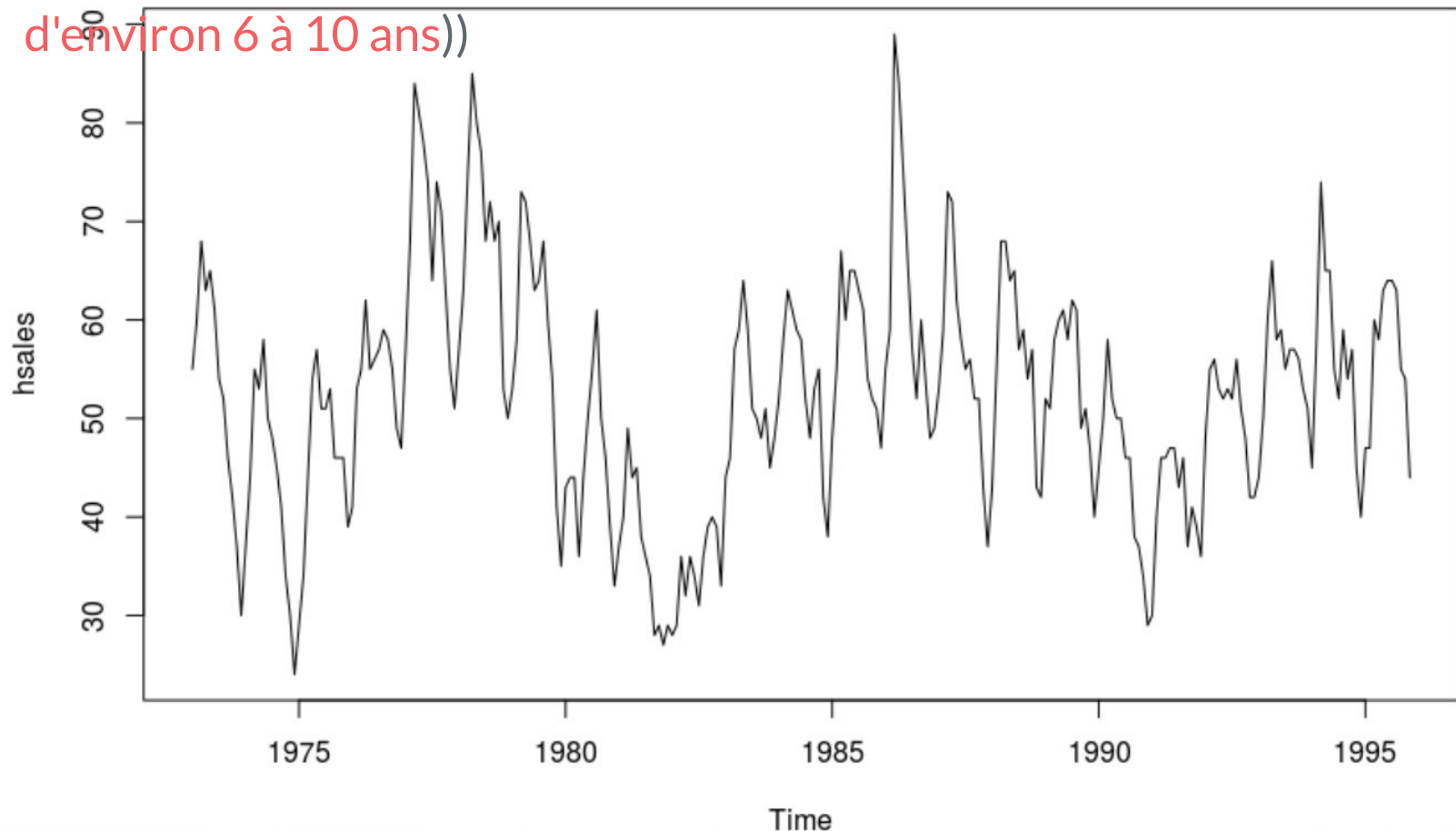
Exemples de séries temporelles

- Nombre annuel du Lynx capturés au Canada((les années de captures abondantes reviennent à intervalle régulier, cycles de 10 ans))



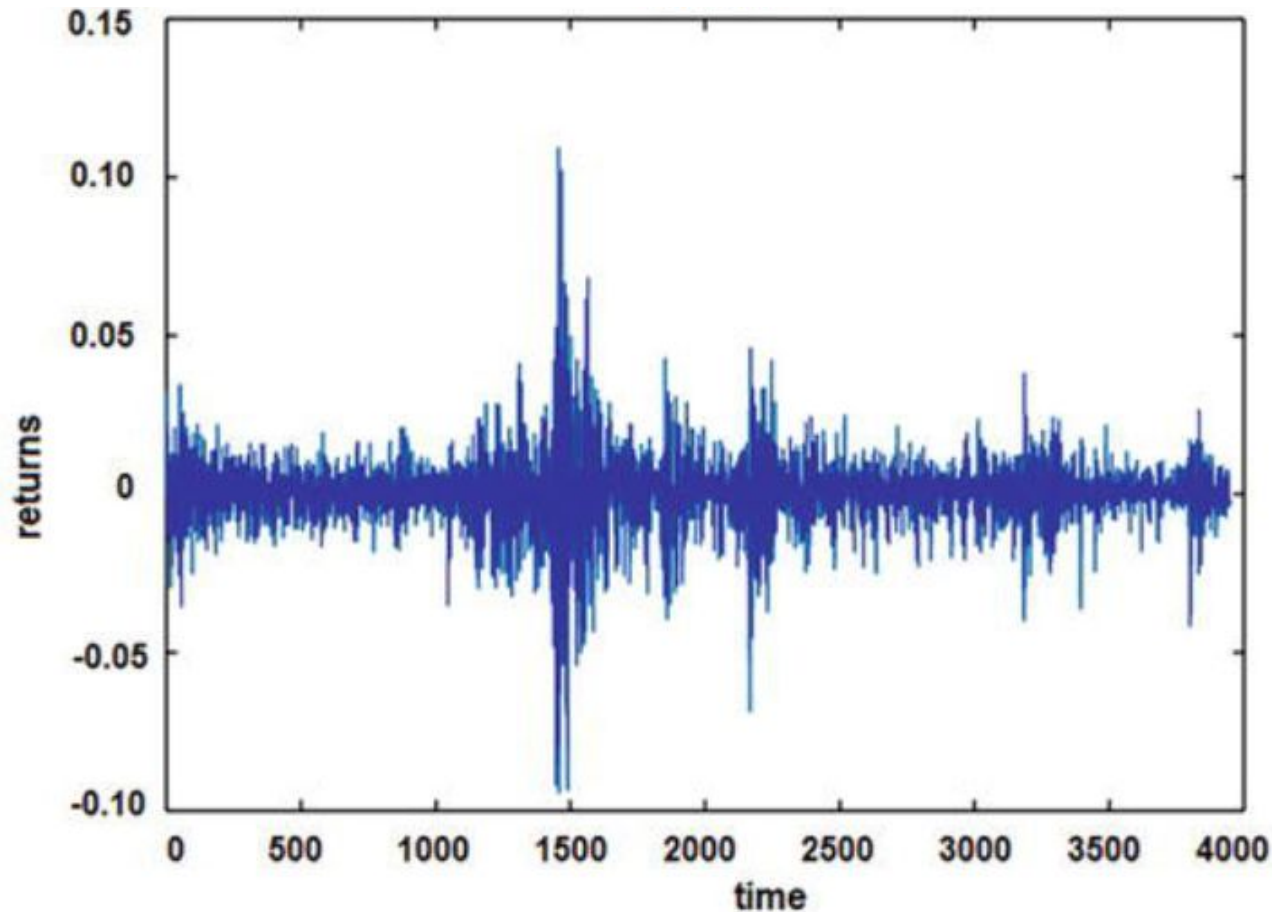
Exemples de séries temporelles

- Vente mensuelles de logements ((forte saisonnalité au cours de chaque année, ainsi qu'un fort comportement cyclique avec une période d'environ 6 à 10 ans))



Exemples de séries temporelles

- Retours de bourse ((Irrégularité ç cause des données imprévisibles))



Estimation de la tendance

- Il est souvent très utile d'estimer la tendance d'une série temporelle par une fonction afin de pouvoir prévenir des valeurs futures de la série.
- Cette estimation est appelée aussi lissage de la série.
- On présente ici deux méthodes:
 - **Méthode des moindres carrés**
 - **Méthodes des moyennes mobiles**

Méthode des moindres carrés

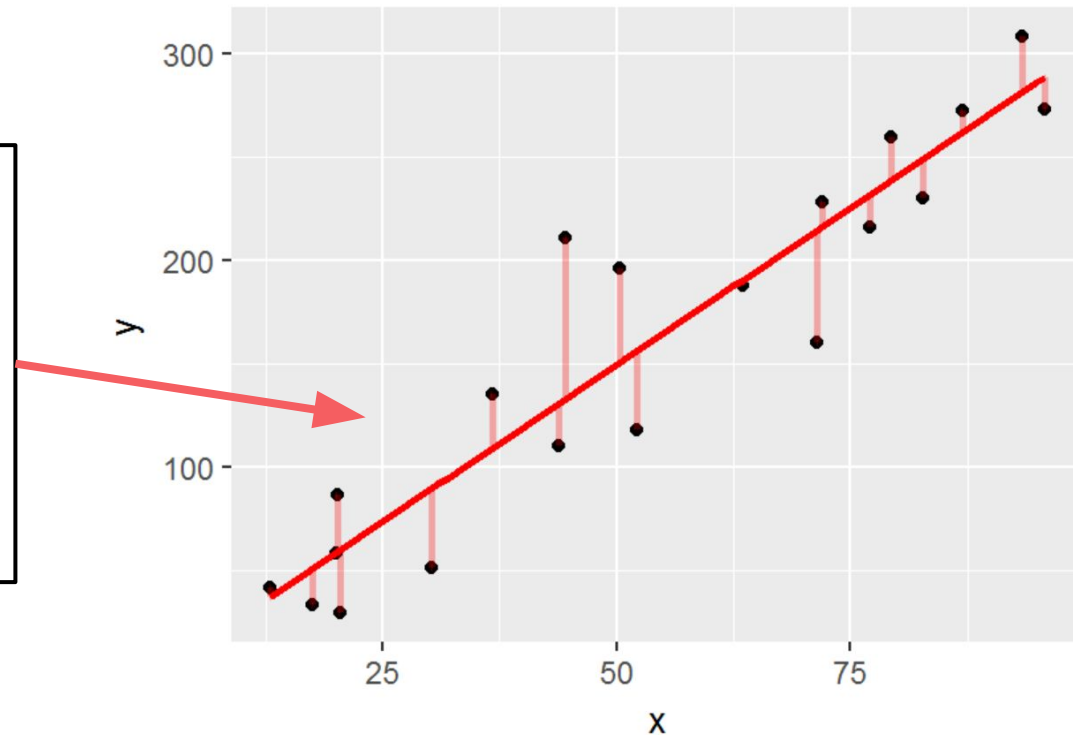
Si la tendance de la série semble linéaire, elle peut être représentée par une droite $y=wt+b$ (*Droite de régression*)

- t et y sont les attributs numériques de la base de données.
- w et b sont les coefficients de régression spécifiant respectivement la **pente** de la droite et l'**ordonnée à l'origine**.

Méthode des moindres carrés

Si la tendance de la série semble linéaire, elle peut être représenté par une droite $y=wt+b$ (*Droite de régression*)

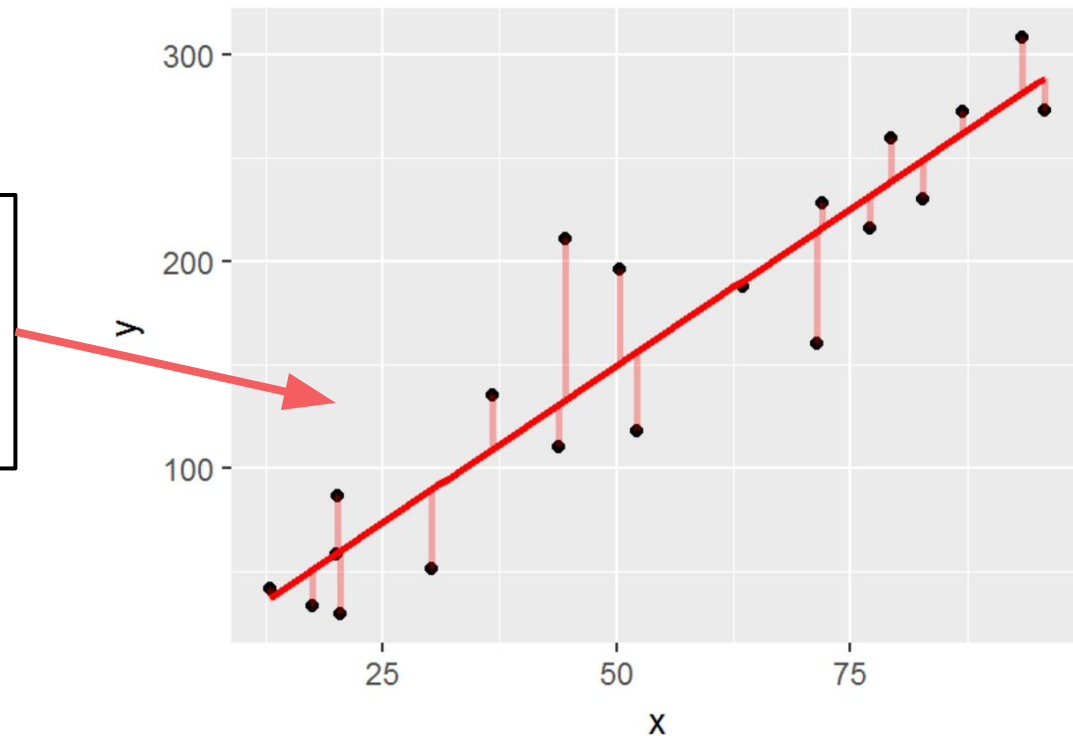
il faut trouver la droite tel que la somme des écarts entre tous les points et la droite soit la plus petite possible



Méthode des moindres carrés

Si la tendance de la série semble linéaire, elle peut être représenté par une droite $y=wt+b$ (*Droite de régression*)

La droite de régression passe toujours par le point moyen de coordonnées $(\bar{x}, \bar{y})$



Méthode des moindres carrés

Si la tendance de la série semble linéaire, elle peut être représenté par une droite $y=wt+b$ (*Droite de régression*)

La méthode des moindres carrés détermine les coefficients de régression afin d'estimer la valeur de la série pour une date t :

$$w = \frac{\text{Covariance}(t, y)}{\text{Variance}(t)}$$

$$b = \bar{y} - w \bar{t}$$

Méthode des moindres carrés

Exemple: Soit la série temporelle suivante, trouver la droite de régression en utilisant la méthode des moindres carrés.

t	100	110	120	130	140	150	160
x	105	95	75	68	53	46	31

$$\text{moy}(t) = 130$$

$$\text{moy}(x) = 67.5714$$

$$\text{cov}(t, x) = (1/7) * ((100-130)(105-67.5714) + (110-130)(95-67.5714) + \dots) = -488.5714$$

$$\text{var}(t) = (1/7) * ((100-130)^2 + (110-130)^2 + \dots) = 400$$

$$a = \text{cov}(t, x) / \text{var}(t) = -1.2214$$

$$b = \text{moy}(x) - a \text{ moy}(t) = (67.5714) - (-1.2214)(130) = 226.3571$$

Méthode des moindres carrés

Exemple: Soit la série temporelle suivante, trouver la droite de régression en utilisant la méthode des moindres carrés.

t	100	110	120	130	140	150	160
x	105	95	75	68	53	46	31

La droite de régression est :

$$x = - 1.2214 t + 226.3571$$

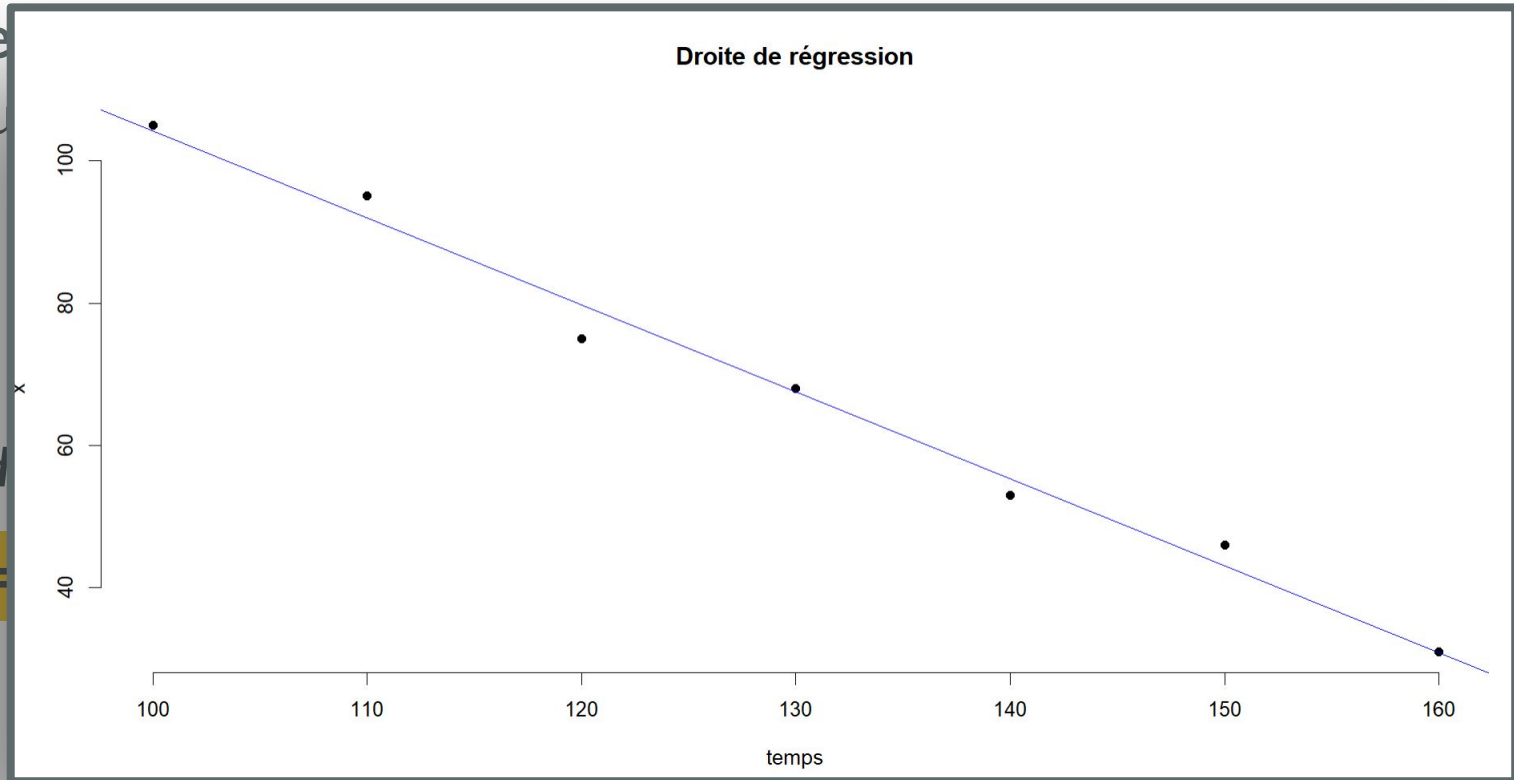
Méthode des moindres carrés

Exe
en u

ssion

La d

$X =$



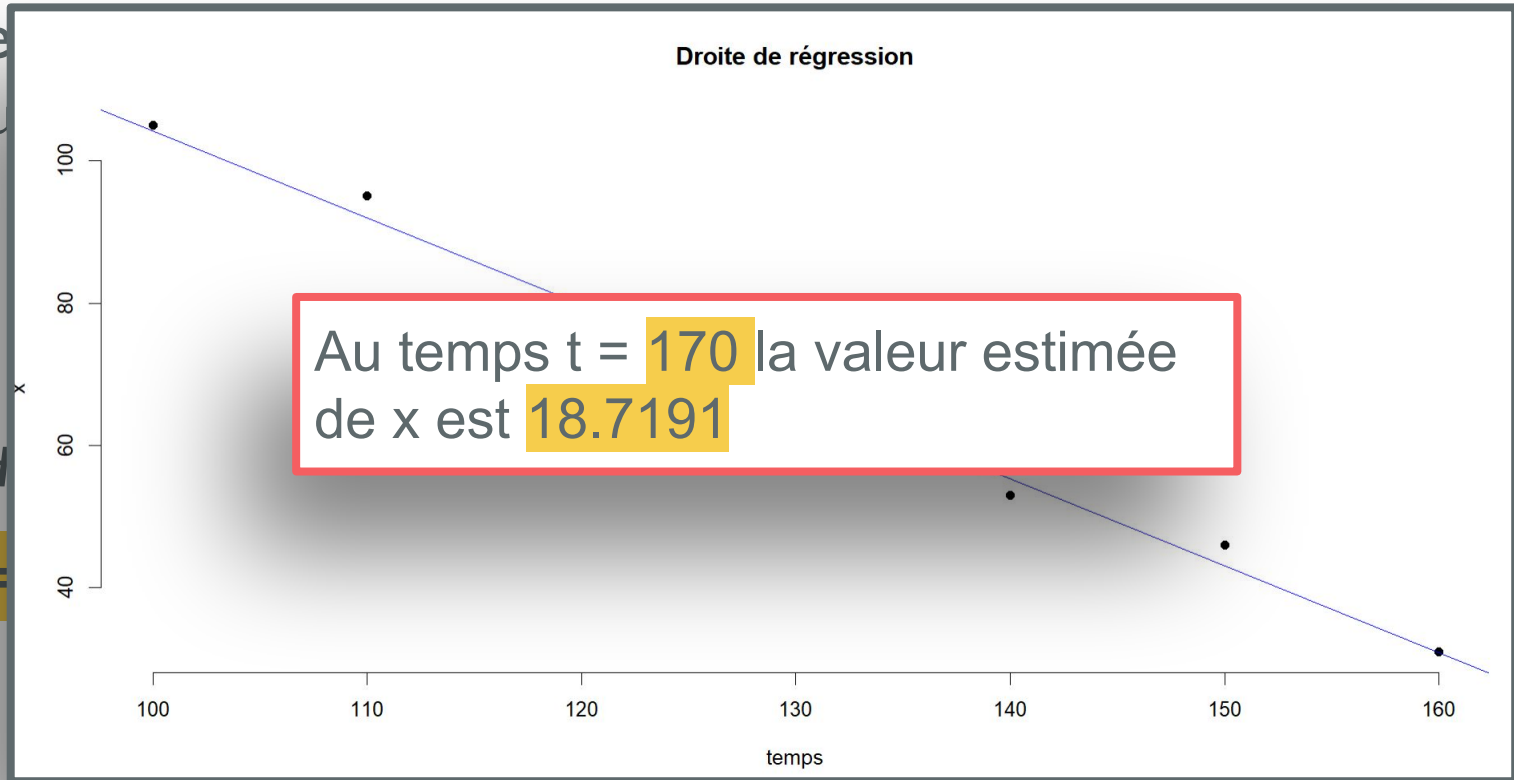
Méthode des moindres carrés

Exe
en u

ssion

La d

X =



Méthode des moyennes mobiles

- La méthode des moyennes mobiles consiste à calculer une moyenne arithmétique sur un nombre limité de données et ensuite l'affecter à une certaine période.
- Le paramètre à sélectionner est le nombre de données dans la moyenne mobile.
- Plus ce nombre est élevé, plus les moyennes éliminent les fluctuations.

Méthode des moyennes mobiles

- Optimisation du nombre de données ?
 - **Fluctuations aléatoires** : Si les fluctuations sont considérées comme du bruit, une fenêtre large (un plus grand nombre de données) est idéale pour éliminer les variations non significatives et mieux capturer la tendance générale.
 - **Fluctuations déterministes** : Pour des fluctuations régulières ou cycliques (ex. saisonnalité), une fenêtre plus petite permet de conserver ces motifs sans trop lisser la série.

Méthode des moyennes mobiles

Exemple:

- Soit le tableau ci-dessous indiquant les ventes de dvd suivant différentes périodes.
- On se propose de calculer les moyennes mobiles à 2 périodes, puis à 4 périodes. Pour ce faire :
 - Dans l'hypothèse de deux périodes : On calcule la prévision pour la période (3) en faisant la moyenne des deux premières périodes et ainsi de suite....
 - Pour l'hypothèse de 4 périodes : On calcule la prévision de la période (5) en faisant la moyenne arithmétique des quatre périodes précédentes et ainsi de suite.....

Méthode des moyennes mobiles

Exemple:

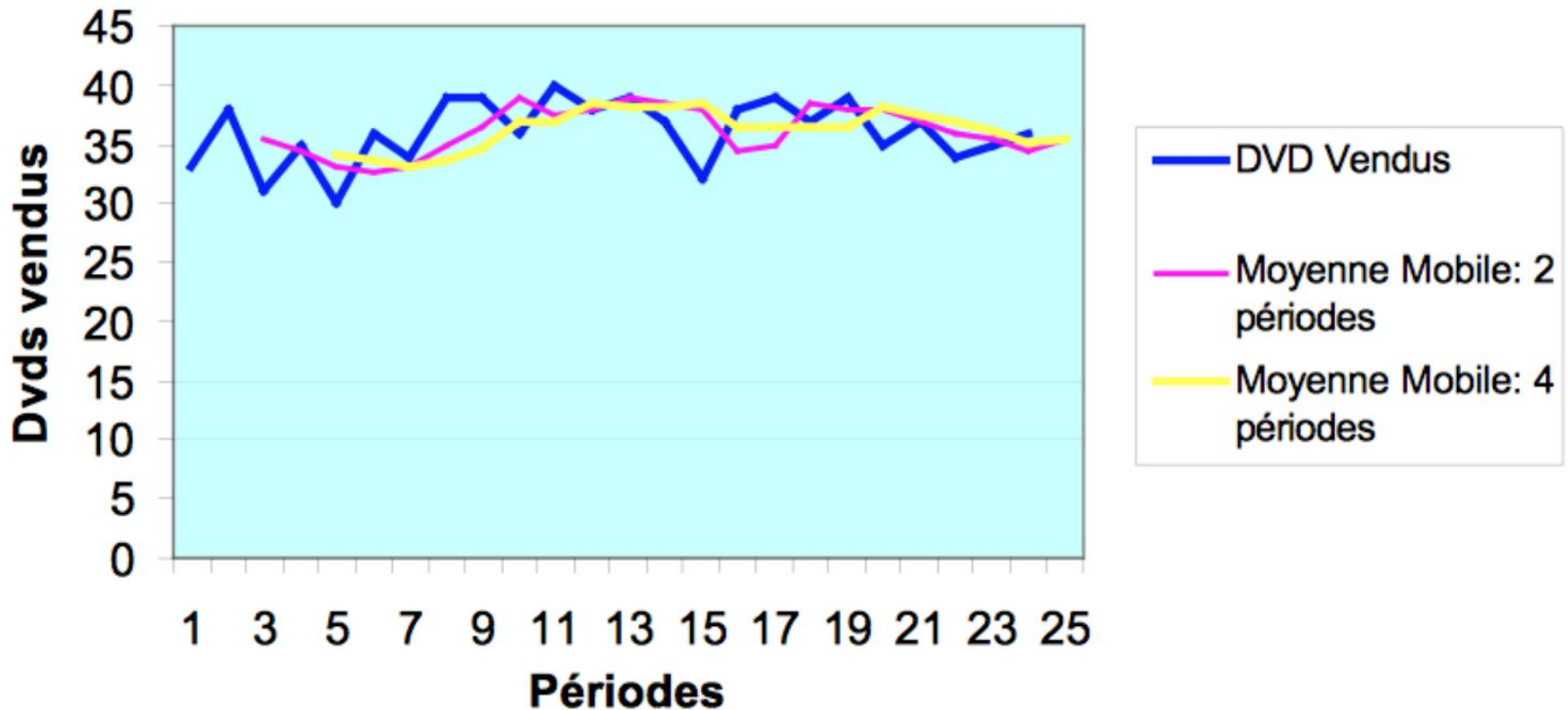
Périodes	DVD Vendus	Moy. Mobile: 2 périodes	Moy. Mobile: 4 périodes
1	33		
2	38		
3	31	35.50	
4	35	34.50	
5	30	33.00	34.25
6	36	32.50	33.50
7	34	33.00	33.00
8	39	35.00	33.75
9	39	36.50	34.75
10	36	39.00	37.00
11	40	37.50	37.00
12	38	38.00	38.50
13	39	39.00	38.25

Périodes	DVD Vendus	Moy. Mobile: 2 périodes	Moy. Mobile: 4 périodes
13	39	39.00	38.25
14	37	38.50	38.25
15	32	38.00	38.50
16	38	34.50	36.50
17	39	35.00	36.50
18	37	38.50	36.50
19	39	38.00	36.50
20	35	38.00	38.25
21	37	37.00	37.50
22	34	36.00	37.00
23	35	35.50	36.25
24	36	34.50	35.25
25		35.50	35.50

Méthode des moyennes mobiles

Exemple:

La prévision avec les moyennes mobiles, nous donne le graphique suivant :



Recherche de similarité dans l'analyse des séries temporelles

- Les séries temporelles sont des données ordonnées dans le temps et cet ordonnancement a une signification que l'on ne peut ignorer.
- Ainsi, on ne peut pas leur appliquer des méthodes de fouille de données classiques qui supposent l'indépendance entre les exemples mais bien des méthodes spécialement adaptées, qui respectent la temporalité de ce type de donnée

Mesures de similarité

- Une mesure de distance entre deux séries temporelles peut être utilisée dans plusieurs tâches de data mining telles que l'apprentissage supervisé et l'apprentissage non supervisé.
- Dans les bases de données traditionnelles, les mesures de similarité sont basées sur un matching exact. Cependant, dans les données des séries temporelles, caractérisées par leur nature numérique et continue, la mesure de similarité est calculée d'une manière approximative.
- Nous présentons ici quelques mesures de similarités entre deux séries temporelles :
 - Mesure de similarité p-normée
 - Dynamic Time Warping
 - Longest Common Subsequence
 - Distance de Hamming

Mesure de similarité p-normée

- Cette mesure de similarité est couramment utilisée car elle a le mérite d'être simple à mettre en œuvre.
- La similarité $\text{Sim}(Q, C)$ entre les séries $Q = q_1, q_2, \dots, q_m$ et $C = c_1, c_2, \dots, c_m$ est égale à:

$$\text{Sim}(Q, C) = \frac{1}{L_p(Q, C)} = \frac{1}{\left(\sum_{i=1}^m (q_i - c_i)^p\right)^{\frac{1}{p}}}$$

Si $p = 1$ alors on utilise la distance de Manhattan : $L_1(Q, C) = \sum_{i=1}^m (q_i - c_i)$

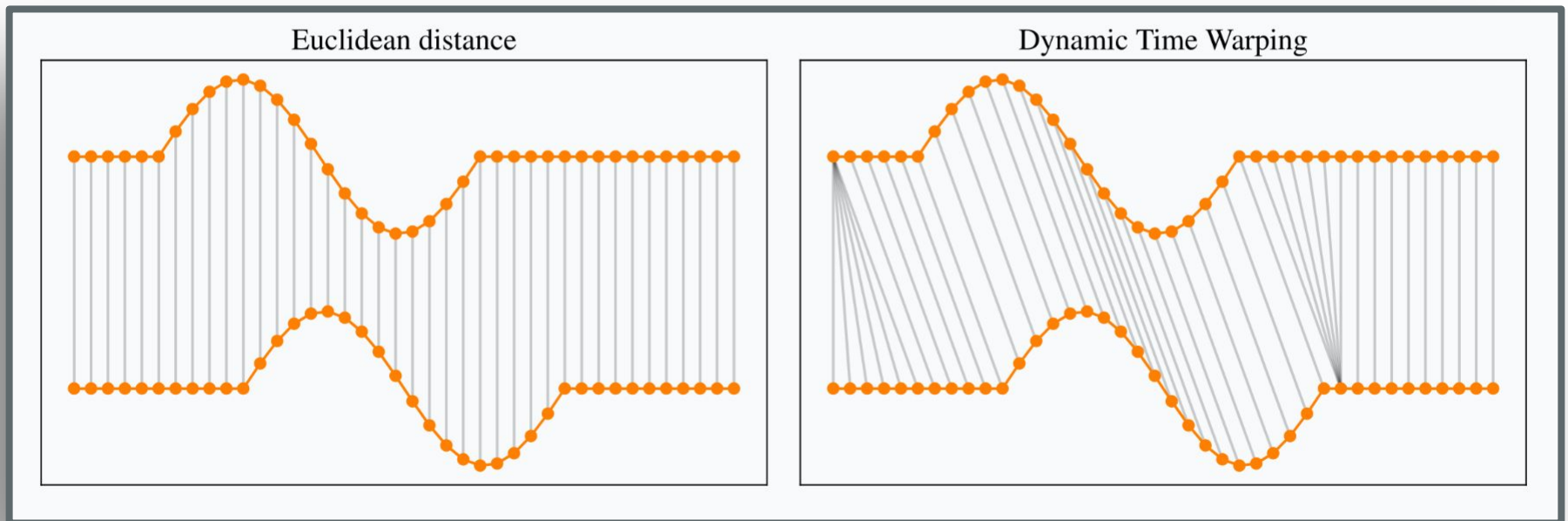
Si $p = 2$ alors on utilise la distance Euclidienne : $L_2(Q, C) = \left(\sum_{i=1}^m (q_i - c_i)^2\right)^{\frac{1}{2}}$

Mesure de similarité p-normée

- **Avantages**
 - Rapide à calculer et fonctionne bien lorsque les séries temporelles ont des formes et des échelles similaires.
- **Limites**
 - Sensible au bruit et aux valeurs aberrantes, et inefficace pour les séries temporelles de longueurs différentes ou celles présentant des déphasages.

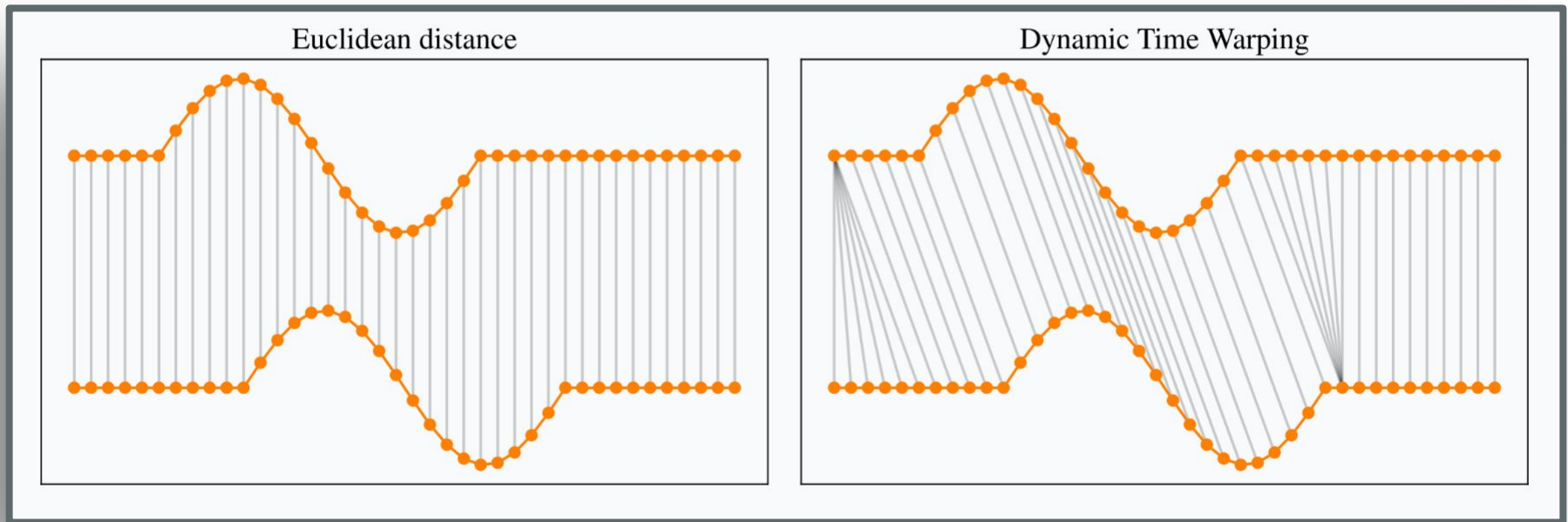
Dynamic Time Warping

La particularité de la méthode Dynamic Time Warping (DTW) ou déformation dynamique temporelle, est de savoir gérer les décalages temporels qui peuvent éventuellement exister entre deux séries. (exemple : Hello VS Heeeelllllloooo)



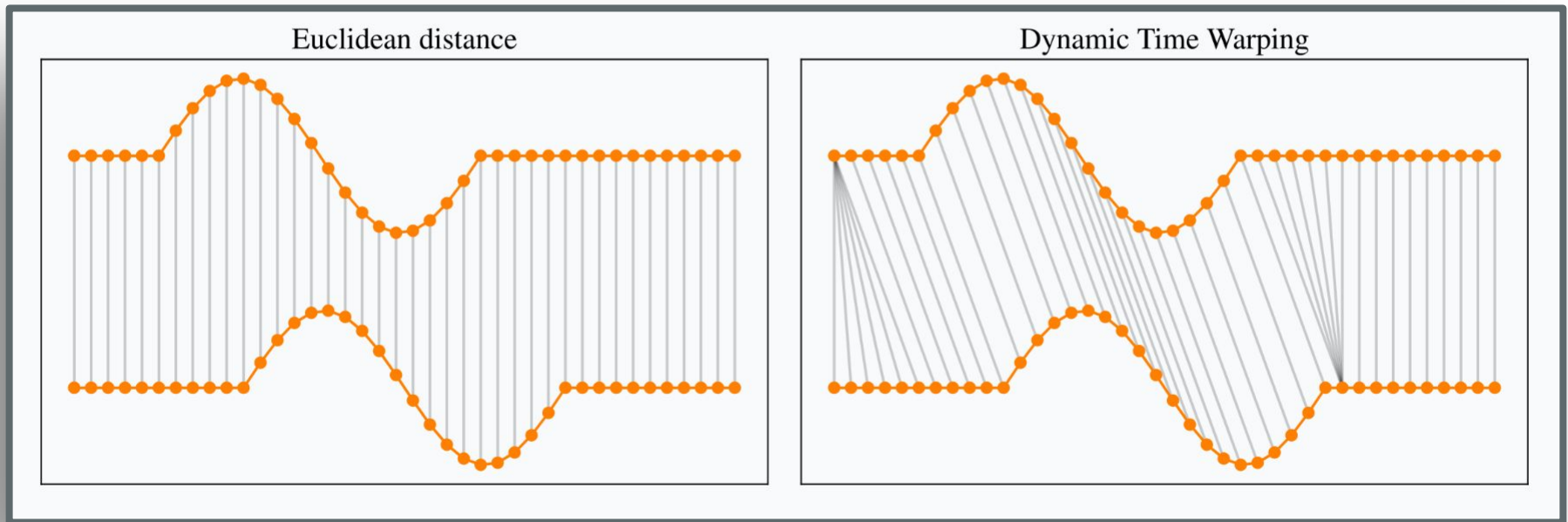
Dynamic Time Warping

Au lieu de comparer chaque point d'une série avec celui de l'autre série qui intervient au même instant t , on permet à la mesure de comparer chaque point d'une série avec un ou plusieurs points de l'autre série, ceux-ci pouvant être décalés dans le temps.



Dynamic Time Warping

Cela permet de créer un meilleur alignement entre les deux séries temporelles et donc calculer la distance de manière efficace lorsque les deux séries ne sont pas exactement les mêmes mais très similaires



Dynamic Time Warping

- **Avantages**
 - Gère efficacement les longueurs variables et les décalages temporels, ce qui le rend résistant aux distorsions temporelles.
- **Limites**
 - Nécessite beaucoup de calculs, en particulier pour les séquences longues, et peut être sensible au bruit dans les données

Dynamic Time Warping

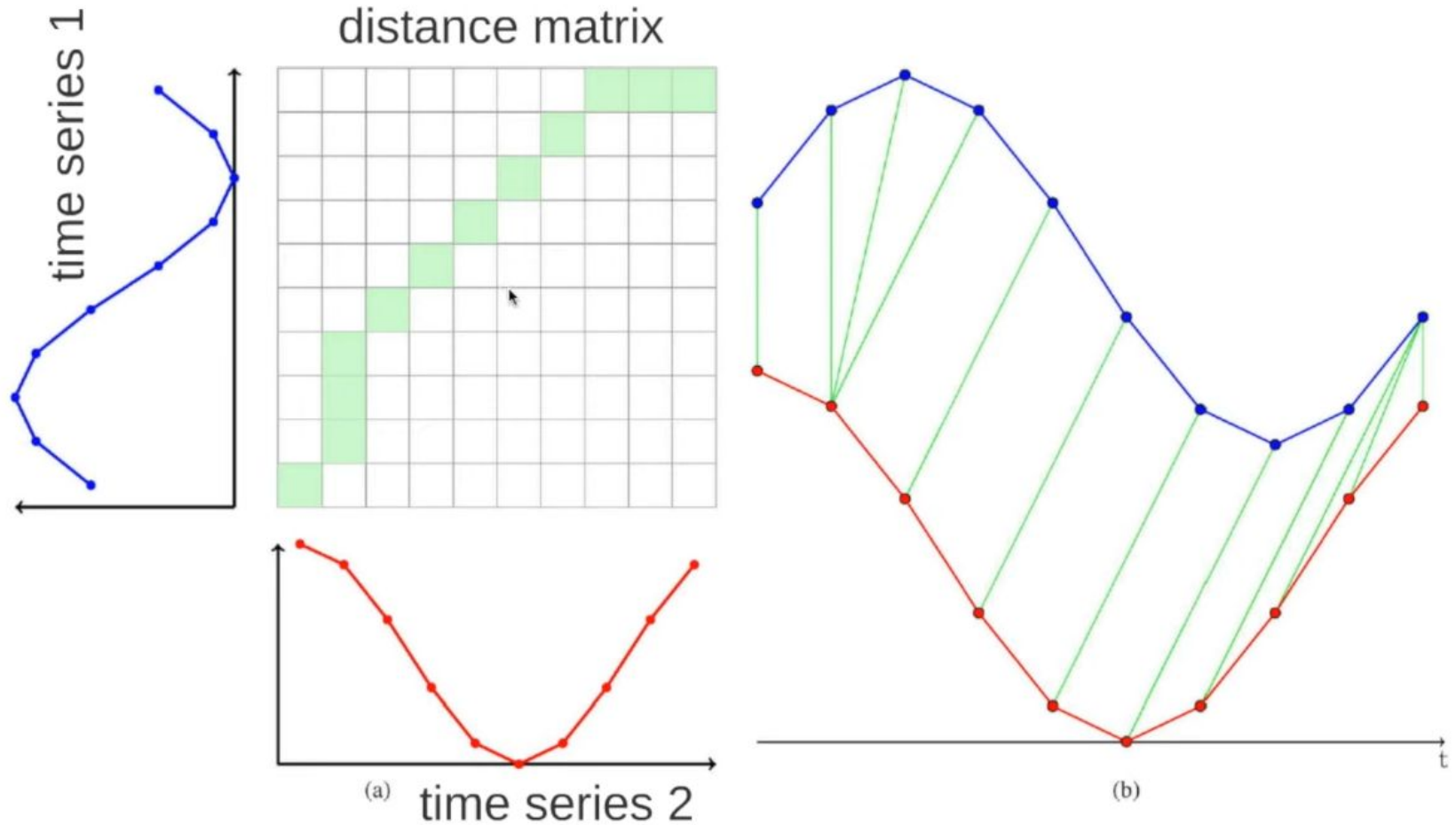
- Le DTW possède une définition récursive qui calcule la similarité entre les séries $\mathbf{Q} = \mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m$ et $\mathbf{C} = \mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n$ de la manière suivante :
- Soit $\mathbf{D}(i, j)$ la distance entre les sous-séquences $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_i$ et $\mathbf{C} = \mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_j$ (avec $i \leq m$ et $j \leq n$)

$$D(i, j) = \begin{cases} |q_1 - c_1| & \text{Si } i = j = 1 \\ |q_i - c_j| + \min\{D(i-1, j), D(i-1, j-1), D(i, j-1)\} & \text{sinon} \end{cases}$$

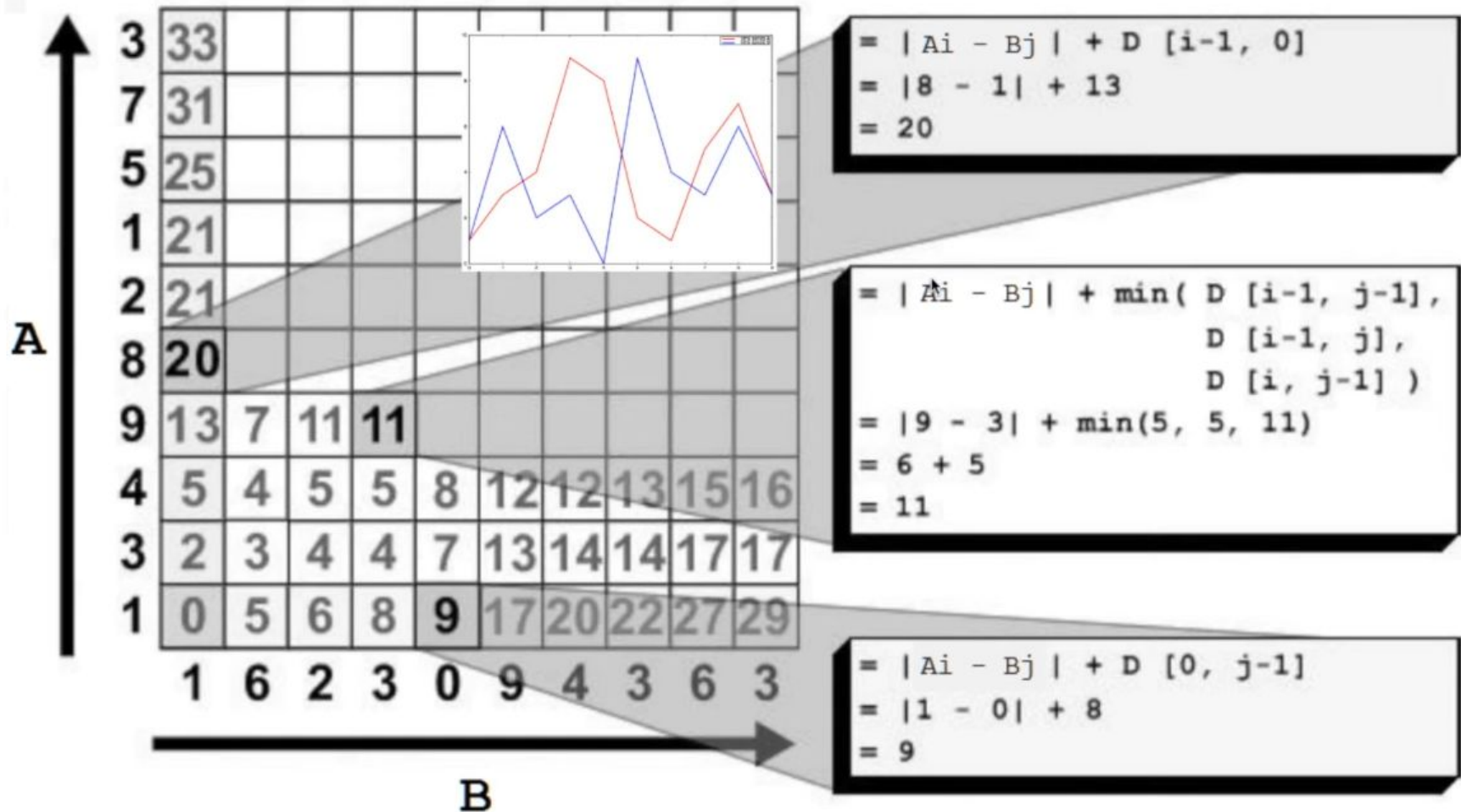
- Soit $\text{Sim}(\mathbf{Q}; \mathbf{C})$ la mesure de similarité DTW entre les séries $\mathbf{Q}$ et $\mathbf{C}$:

$$\text{Sim}(\mathbf{Q}, \mathbf{C}) = \frac{1}{D(m, n)}$$

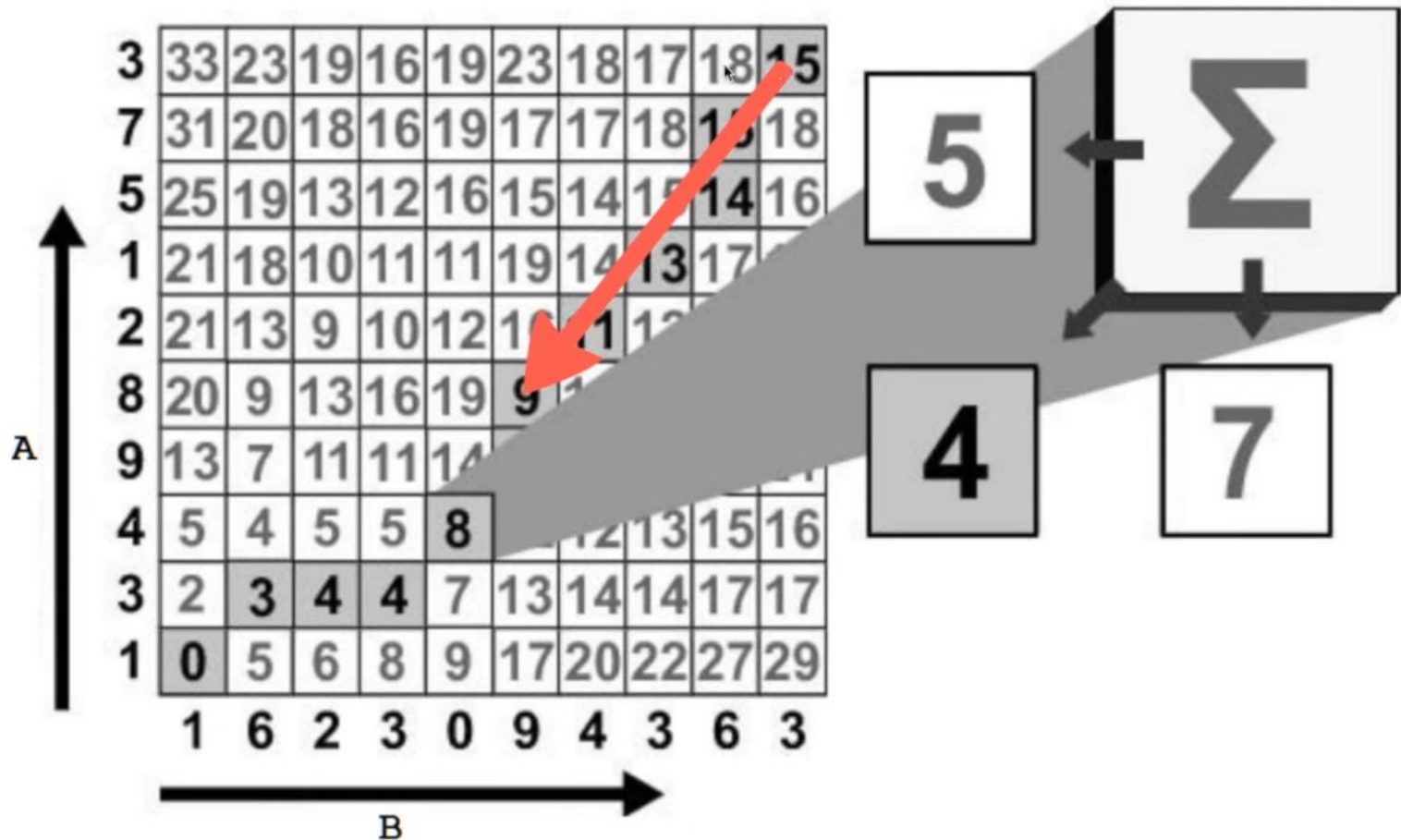
Dynamic Time Warping



Dynamic Time Warping



Dynamic Time Warping



Longest Common Subsequence

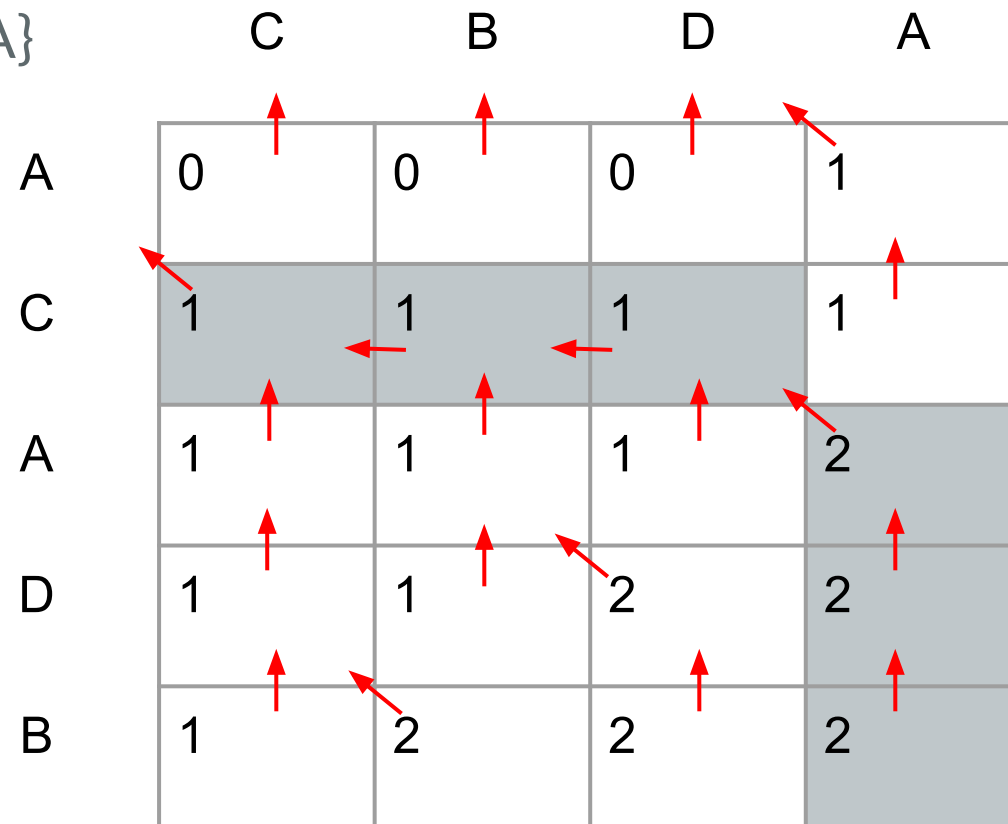
- L'idée de la méthode Longest Common Subsequence (LCSS) ou Plus longue sous séquence commune est de comparer uniquement les portions les plus similaires de chacune des séries.
- Plus les sous-séquences communes sont nombreuses, plus on considérera les deux séries comme similaires.

$$D(i, j) = \begin{cases} 1 + D(i - 1, j - 1) & \text{si } |q_i - c_j| < \epsilon \\ \max\{D(i - 1, j), D(i, j - 1)\} & \text{sinon} \end{cases}$$

Longest Common Subsequence

Exemple: série catégorielle

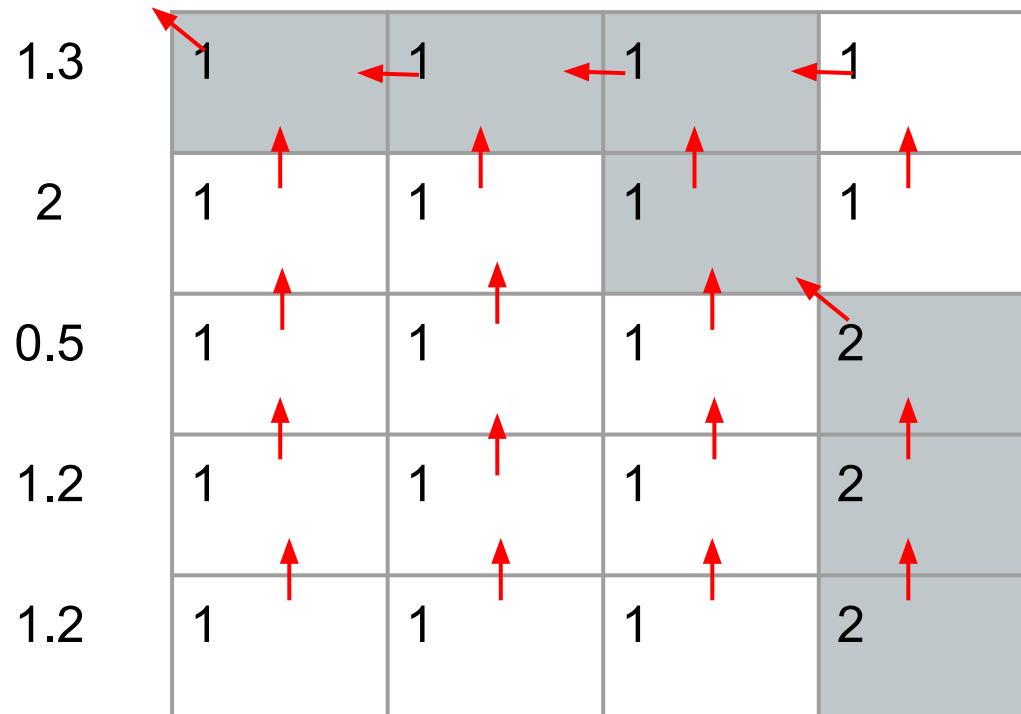
LLCS = 2 LCS = {CA}



Longest Common Subsequence

Exemple: série temporelle avec $\varepsilon = 0.2$

LLCS=2 LCS={1.5, 0.5} 1.5 5 3.2 0.5



Distance de Hamming

- La distance de Hamming entre deux séries Q et C de longueurs égales à n correspond au nombre de différences entre les deux séries

$$D(Q, C) = \sum_{i=1}^n L(q_i, c_i)$$

Avec

$$L(q_i, c_i) \begin{cases} 1 & \text{si } |q_i - c_j| \geq \epsilon \\ 0 & \text{sinon} \end{cases}$$

Recherche des motifs fréquents dans les séries temporelles

- Dans les séries temporelles, un **motif fréquent** est une **sous-série** qui apparaît un nombre significatif de fois.

Recherche des motifs fréquents dans les séries temporelles

- Dans les séries temporelles, un **motif fréquent** est une **sous-série** qui apparaît un nombre significatif de fois.
- Pour rechercher de tels motifs, on se base sur un seuil de **support minimum** (fréquence) et un seuil de **similarité minimale** entre les sous-séries pour les considérer correspondantes (**matching**)

Recherche des motifs fréquents dans les séries temporelles

- Dans les séries temporelles, un **motif fréquent** est une **sous-série** qui apparaît un nombre significatif de fois.
- Pour rechercher de tels motifs, on se base sur un seuil de **support minimum** (fréquence) et un seuil de **similarité minimale** entre les sous-séries pour les considérer correspondantes (**matching**)
- Pranav et al. ont proposé en 2002 un algorithme permettant d'extraire les motifs fréquents dans une série temporelle qui se base sur les définitions suivante :
 - **Matching**
 - **Matching trivial**

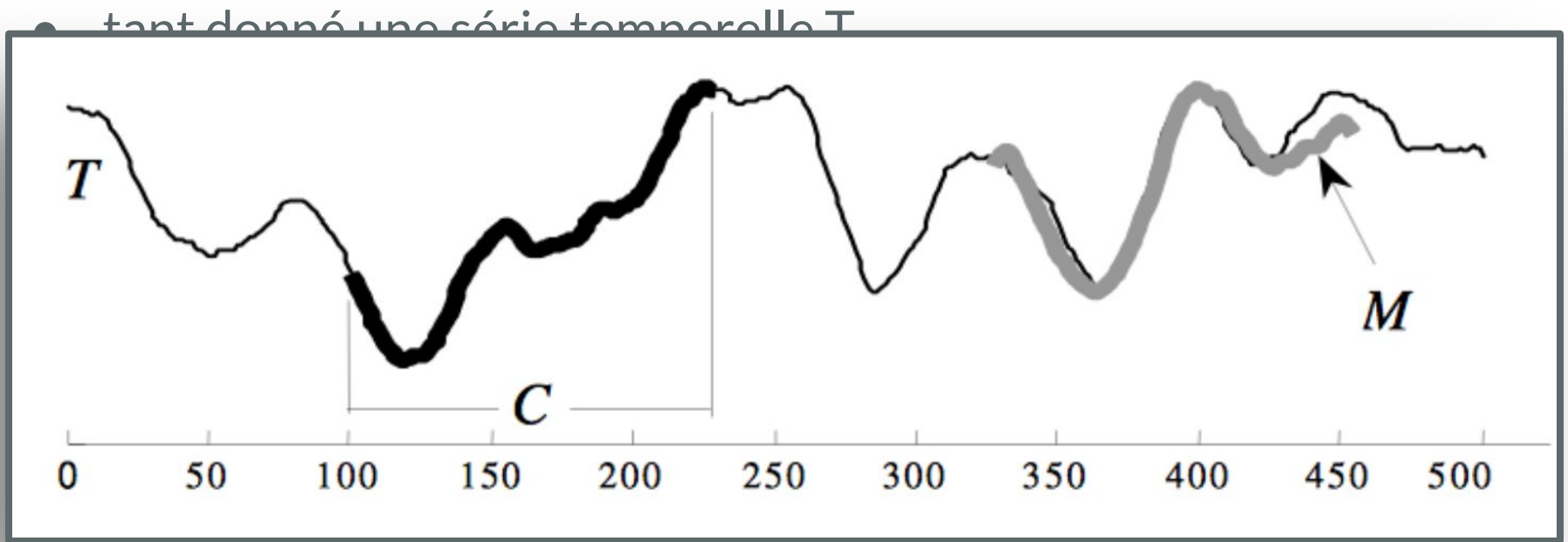
Recherche des motifs fréquents dans les séries temporelles

Matching

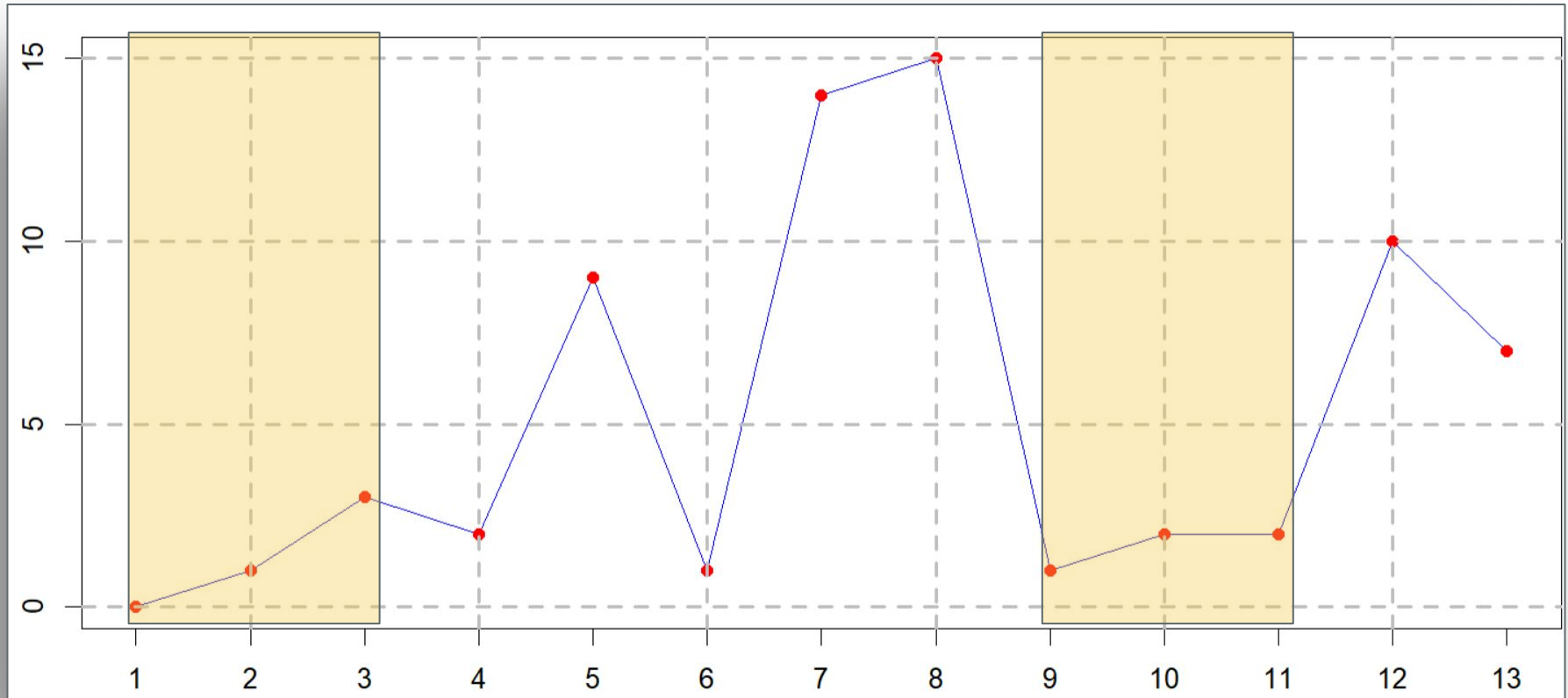
- Etant donné une série temporelle T
- et deux sous-séries de T : C commençant à la position p et M à la position q avec un nombre réel R .
- On dit que C et M représentent un matching (relativement à R) si $D(C, M) \leq R$.
- Dans l'exemple de la figure suivante, la sous-série C matche à la sous série M

Recherche des motifs fréquents dans les séries temporelles

Matching

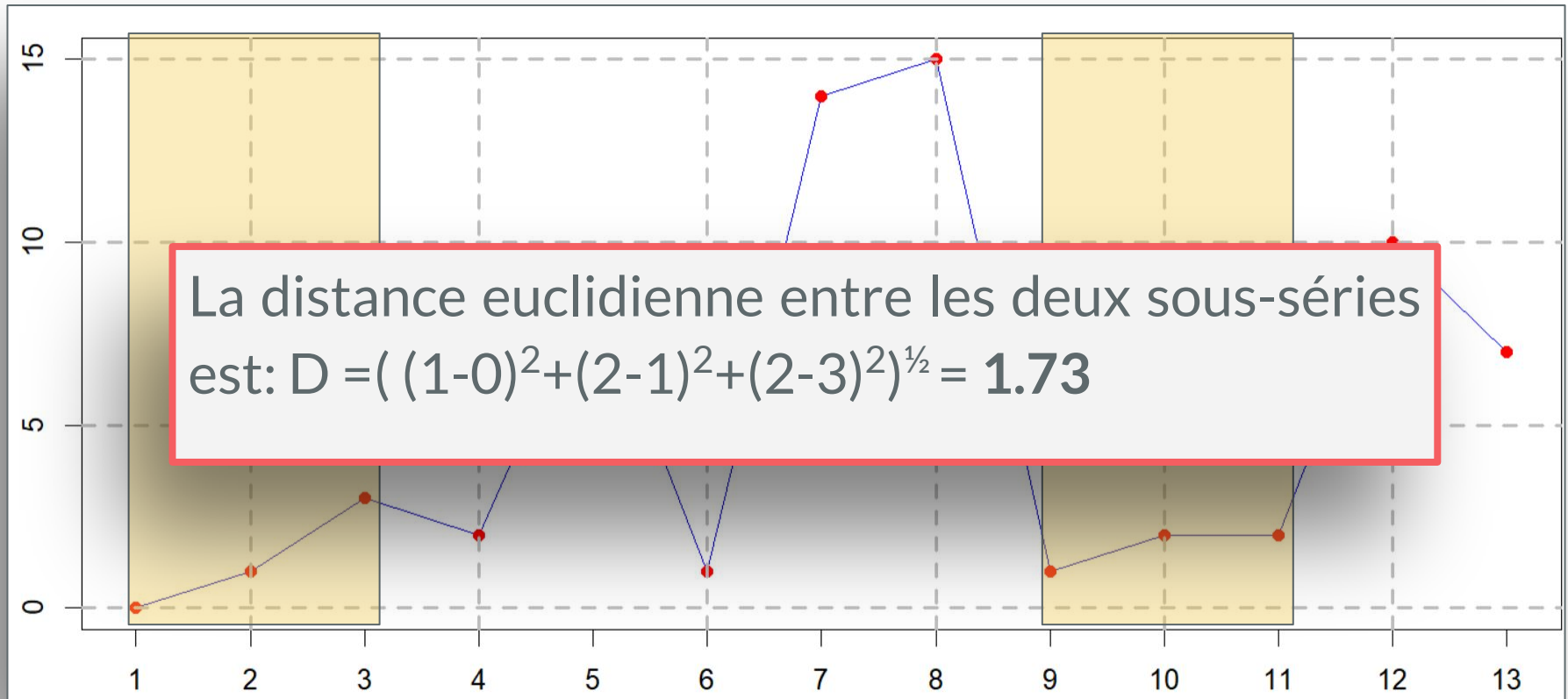


Recherche des motifs fréquents dans les séries temporelles



Recherche des motifs fréquents dans les séries temporelles

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---



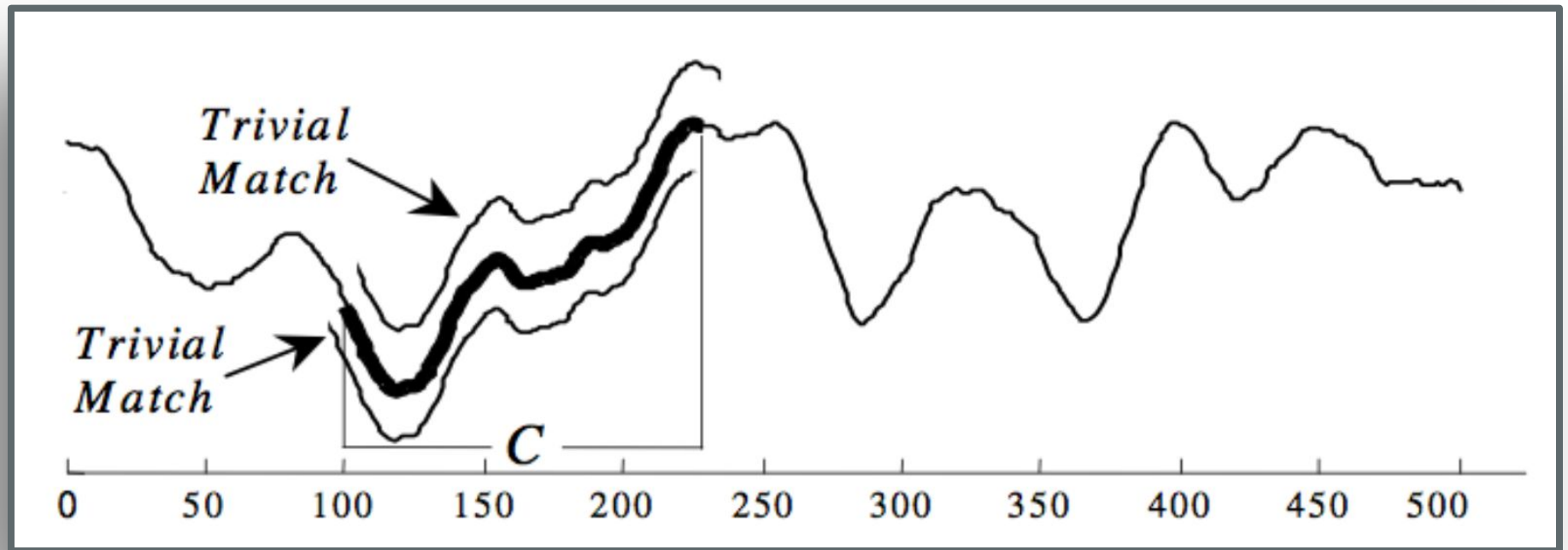
Recherche des motifs fréquents dans les séries temporelles

Matching trivial

- En appliquant la définition précédente, on trouve des paires de sous-séries dont le matching est intuitif tel que la série C commençant à p et la série M qui commence à $p + 1$.
- Ce matching est dit trivial.

Recherche des motifs fréquents dans les séries temporelles

Matching trivial

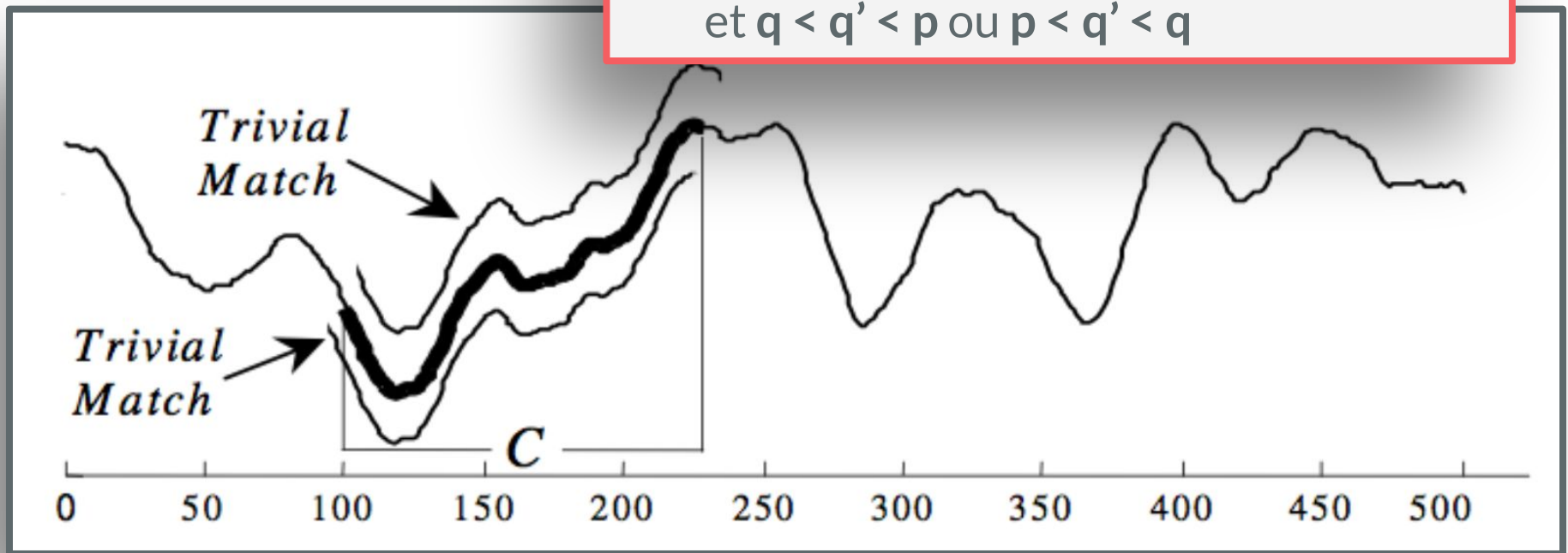


Recherche des motifs temporelles

Matching trivial

Une sous-série C commençant à p et une sous-série M commençant à q représentent un matching trivial si ...

- $p = q$ ou
- qu'il n'existe aucune sous-série M' commençant à q' tel que $D(C, M') > R$ et $q < q' < p$ ou $p < q' < q$



Recherche des motifs fréquents dans les séries temporelles

- Le motif C_1 le plus significatif de longueur n (la sous-série de longueur n la plus fréquente) est la sous-série ayant le nombre le plus élevé de matching non trivial.
- Le motif C_k de longueur n (la sous-série de longueur n la plus fréquente) est la sous-série ayant le le K^{ieme} plus élevé nombre de matching non trivial avec $D(C_k, C_i) > 2R$ pour tout $1 \leq i < k$.

En utilisant ces définitions, on donne l'algorithme suivant permettant de donner le meilleur motif (C_1) de longueur n relativement au seuil de distance R dans une série temporelle T :

Recherche temporelle

- Le motif de longueur n le plus fréquent
- Le motif de longueur n la plus fréquente

En utilisant
de donner
de distance

Algorithme Meilleur_Motifs_ $C_1(T, n, R)$

```
NbreMatchMeilleurMotif  $\leftarrow$  0 ;  
IndiceMeilleurMotif  $\leftarrow$  NULL ;  
pour  $i = 1$  jusqu'à  $Long(T) - n + 1$  faire  
    NbreMotifs  $\leftarrow$  0 ;  
    ListeIndices  $\leftarrow$  Nil ;  
    pour  $j = i$  jusqu'à  $Long(T) - n + 1$  faire  
        si Matching_Non_Trivial( $T_{[i:i+n-1]}, T_{[j:j+n-1]}, R$ ) alors  
            NbreMotifs  $\leftarrow$  NbreMotifs + 1 ;  
            ListeIndices.AjouterIndice( $j$ ) ;  
        fin si  
    fin pour  
    si NbreMotifs > NbreMeilleurMotifs alors  
        NbreMatchMeilleurMotif  $\leftarrow$  NbreMotifs ;  
        IndiceMeilleurMotif  $\leftarrow$   $i$  ;  
        ListeMotifsMatching  $\leftarrow$  ListeIndices ;  
    fin si  
fin pour
```

Fin.

séries

us-série de
e nombre le
ur n la plus
e nombre de
< k .

permettant
ent au seuil

Recherche des motifs fréquents dans les séries temporelles

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	7.4											
---	-----	--	--	--	--	--	--	--	--	--	--	--

Recherche des motifs fréquents dans les séries temporelles

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	7.4	6.9										
---	-----	-----	--	--	--	--	--	--	--	--	--	--

Recherche des motifs fréquents dans les séries temporelles

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	7.4	6.9	14.7									
---	-----	-----	------	--	--	--	--	--	--	--	--	--

Recherche des motifs fréquents dans les séries temporelles

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	7.4	6.9	14.7	19.3	17.7	19.9	15	8.2	8.9			
---	-----	-----	------	------	------	------	----	-----	-----	--	--	--

Recherche des motifs fréquents dans les séries temporelles

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	7.4	6.9	14.7	19.3	17.7	19.9	15	8.2	8.9			
---	-----	-----	------	------	------	------	----	-----	-----	--	--	--

Si $R=8.5$ le nombre de motifs fréquents de cette sous séquence est 3 !

Recherche des motifs fréquents dans les séries temporelles

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

--	--	--	--	--	--	--	--	--	--	--	--	--

Recherche des motifs fréquents dans les séries temporelles

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

--	--	--	--	--	--	--	--	--	--	--	--	--

Recherche des motifs fréquents dans les séries temporelles

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

--	--	--	--	--	--	--	--	--	--	--	--	--

Recherche des motifs fréquents dans les séries temporelles

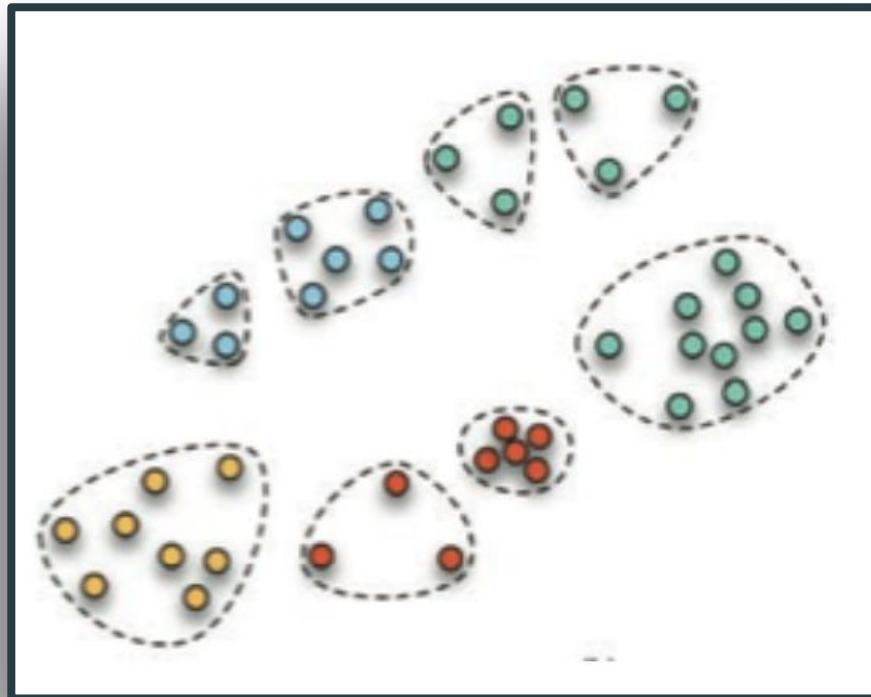
0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

0	1	3	2	9	1	14	15	1	2	2	10	7
---	---	---	---	---	---	----	----	---	---	---	----	---

--	--	--	--	--	--	--	--	--	--	--	--	--

Clustering des séries temporelles

- Le clustering ou la segmentation est le processus de recherche de groupes (appelés clusters) avec des propriétés similaires dans un ensemble de données.



Clustering des séries temporelles

- Le clustering ou la segmentation est le processus de recherche de groupes (appelés clusters) avec des propriétés similaires dans un ensemble de données.
- Le but est de trouver les clusters les plus homogènes possibles et les plus différents les uns des autres.

Clustering des séries temporelles

- Le clustering ou la segmentation est le processus de recherche de groupes (appelés clusters) avec des propriétés similaires dans un ensemble de données.
- Le but est de trouver les clusters les plus homogènes possibles et les plus différents les uns des autres.
- Plus formellement, les clusters doivent maximiser la variance inter-clusters et minimiser la variance intra-clusters.

Clustering des séries temporelles

- Le clustering ou la segmentation est le processus de recherche de groupes (appelés clusters) avec des propriétés similaires dans un ensemble de données.
- Le but est de trouver les clusters les plus homogènes possibles et les plus différents les uns des autres.
- Plus formellement, les clusters doivent maximiser la variance inter-clusters et minimiser la variance intra-clusters.
- On peut effectuer deux types de clustering sur les séries temporelles
 - Clustering des séries entières
 - Clustering des sous-séries

Clustering des séries temporelles

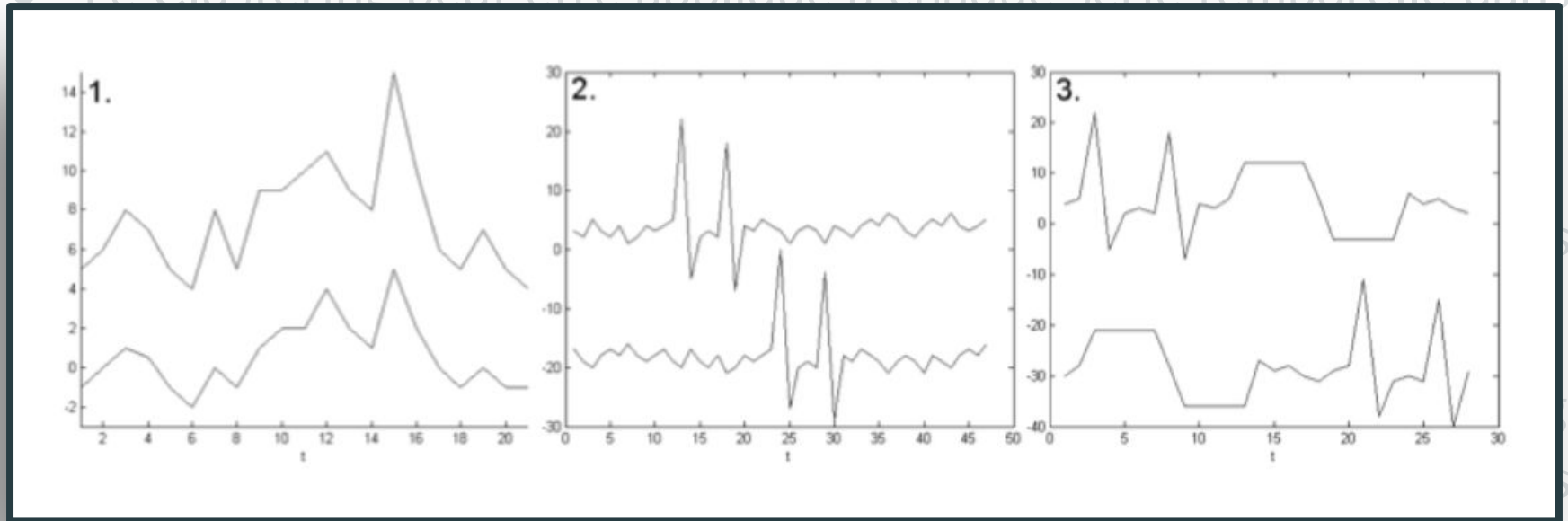
Clustering des séries entières

- Le Clustering peut être appliqué à chaque série temporelle dans un ensemble.
- L'objectif est de regrouper des séries entières dans des groupes tels que les séries temporelles dans chaque cluster soient les plus similaires possibles.
- Un exemple de l'application d'un tel clustering est le clustering des entreprises dans une bourse en vue de détecter les entreprises dont les variations des valeurs des actions se ressemblent

Clustering des séries temporelles

Clustering des séries entières

- Le Clustering peut être appliqué à chaque série temporelle dans



entreprises dont les variations des valeurs des actions se ressemblent

Clustering des séries temporelles

Clustering des séries entières

- Le clustering sur les séries temporelles peut être effectué par plusieurs approches à chacune ses ingrédients.
- Par exemple, l'utilisation d'un algorithme de clustering de type k-Means sur un jeu de séries temporelles amène à se poser les questions:
 - du choix d'une mesure de distance entre deux séries temporelles
 - et celle du choix d'une méthode effectuant l'agrégation de plusieurs séries temporelles afin d'en estimer le centre (i.e. calculer les k moyennes).

K-means adapté aux séries temporelles

Algorithme

- La méthode des k-Means adaptée aux séries temporelles consiste à classer le jeu de séries en k clusters disjoints (Lin et al. 2003).
- Son algorithme est le suivant :

1. Choisir la valeur de k.
2. Initialiser les centres des k clusters (aléatoirement si nécessaire).
3. Affecter chaque série temporelle au cluster dont le centre lui est le plus proche.
4. Ré-estimer les centres des k clusters, en supposant que toutes les affectations des séries sont correctes.
5. Si aucune des séries n'a changé de cluster, alors fin de l'algorithme. Sinon, retour à l'étape 3.

K-means adapté aux séries temporelles

Mesure de similarité

- L'algorithme des k-Means, avec des données classiques (i.e. non temporelles), ne pose pas de problème d'implémentation.
- La ré-estimation du centre de chaque cluster se fait habituellement en calculant la moyenne Euclidienne de tous les objets affectés à ce cluster (étape 4 de l'algorithme).
- Par contre, dans le cadre d'un apprentissage basé sur la classification des séries en fonction de leurs "formes", cette agrégation Euclidienne entraîne une perte d'informations importante

K-means adapté aux séries temporelles

Mesure de similarité

- Nous entendons par forme d'une série les différentes variations relatives qu'elle effectue au cours du temps.
- Cette notion est difficile à définir et assez subjective, mais elle peut être néanmoins en dégagée des deux caractéristiques suivantes :
 - Les valeurs absolues (les amplitudes) des points d'une série ne sont pas intéressantes. C'est la variation qui caractérise plus les séries tel que dans les exemples de la figure précédente.
 - Les variations des valeurs de deux séries peuvent intervenir à différents instants tout en préservant la similarité des séries tel que dans le deuxième exemple de la figure précédente

K-means adapté aux séries temporelles

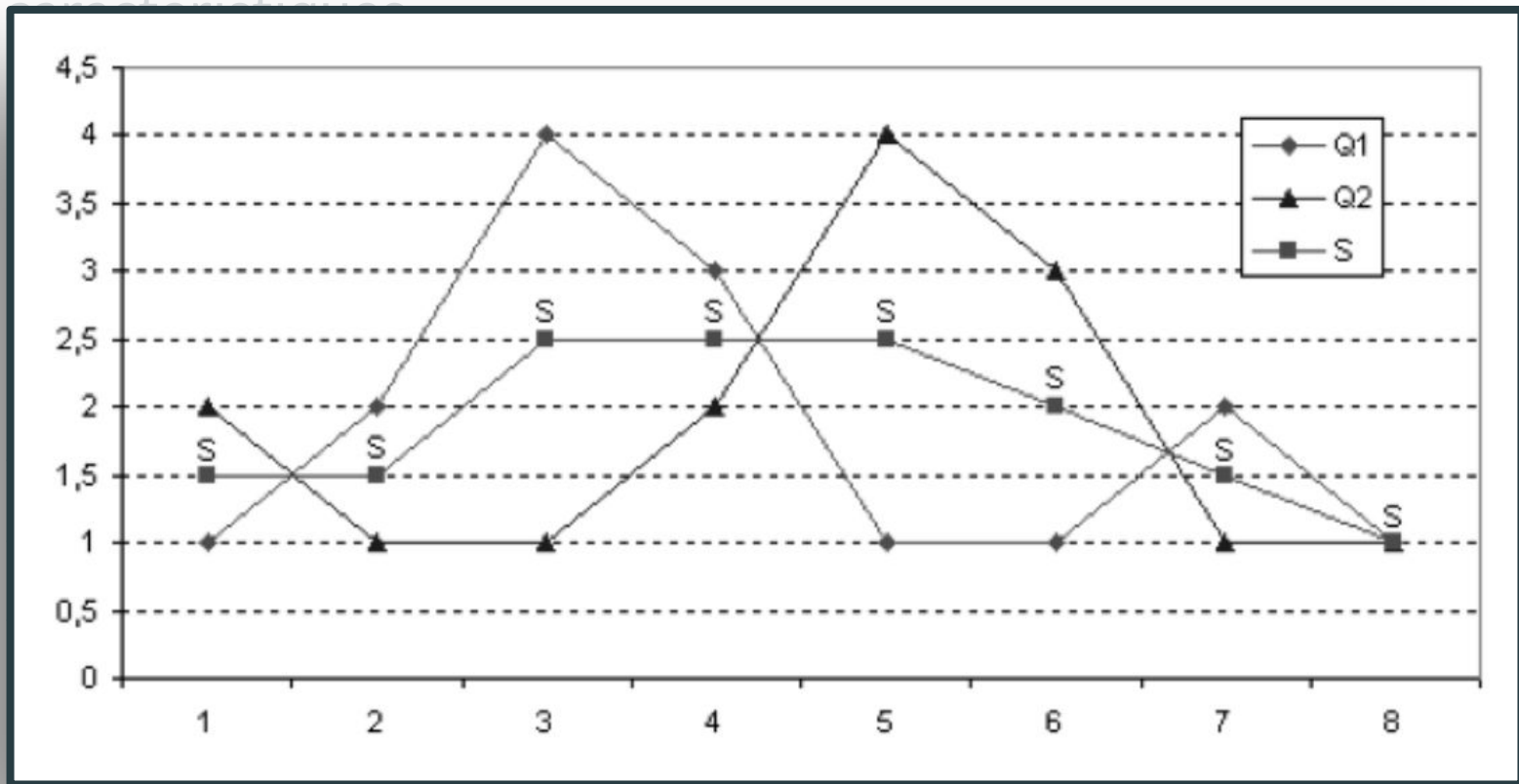
Mesure de similarité

- L'agrégation Euclidienne ne permet pas le respect de ces caractéristiques.
- Prenons par exemple les deux séries suivantes :
 - $Q_1 = \{1, 2, 4, 3, 1, 1, 2, 1\}$
 - $Q_2 = \{2, 1, 1, 2, 4, 3, 1, 1\}$
- Si nous essayons d'agréger les séries Q_1 et Q_2 en une troisième série à l'aide de la moyenne Euclidienne, nous obtenons la série suivante :
 - $S = \{1.5, 1.5, 2.5, 2.5, 2.5, 2, 1.5, 1\}$
- La figure suivante présente le résultat graphique de cette agrégation.

K-means adapté aux séries temporelles

Mesure de similarité

- L'agrégation Euclidienne ne permet pas le respect de ces

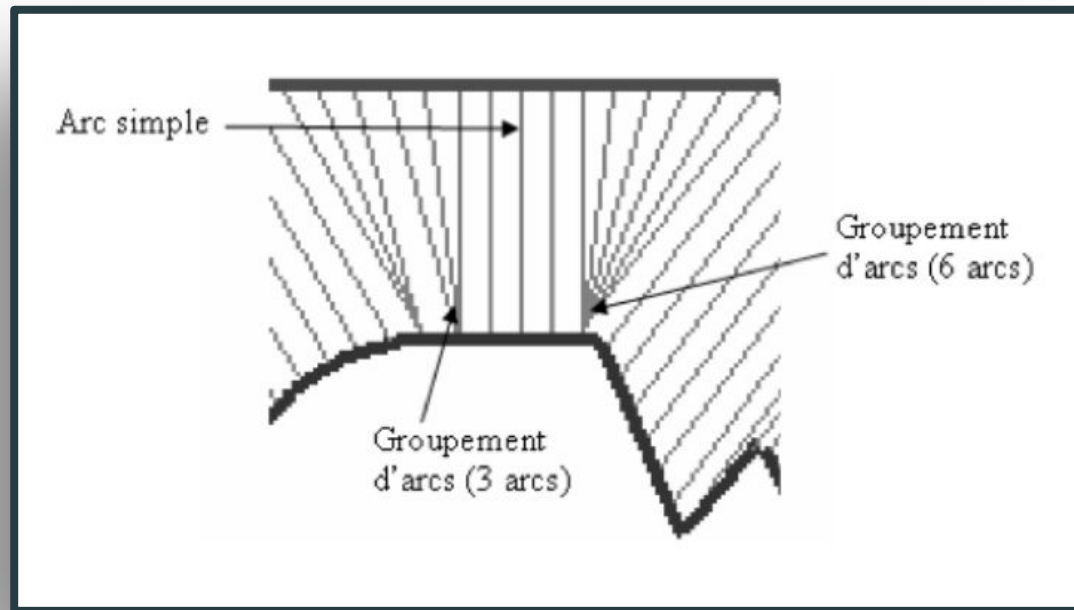


- La figure suivante présente le résultat graphique de cette agrégation.

K-means adapté aux séries temporelles

Mesure de similarité

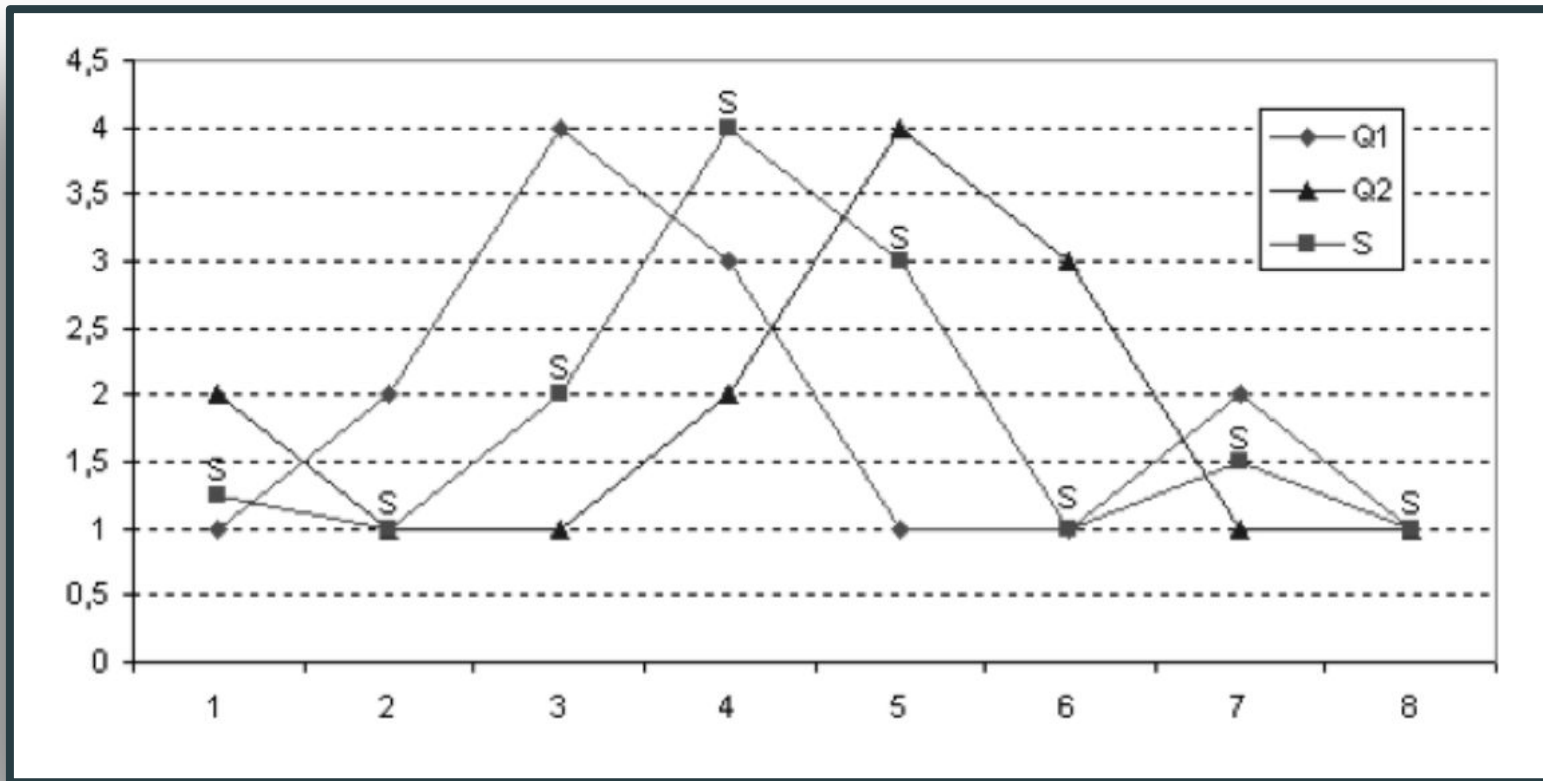
- Lin et al. (2003) proposent d'utiliser la distance DWT entre les deux séries pour générer S. Chaque point de S est généré par la moyenne des arcs groupés par les paires de points des deux séries.



K-means adapté aux séries temporelles

Mesure de similarité

- La figure suivante présente l'agrégation calculée par cette méthode :



Clustering des séries temporelles

Clustering des sous-séries

- Dans cette approche, les clusters sont créés en extrayant des sous-séries à partir d'une ou plusieurs séries plus longues.
- On peut détecter par cette approche par exemple, les clusters de périodes de consommation d'électricité pour un consommateur donnée.
- Dans ce type de clustering, les séries sont découpées en fenêtres non chevauchées dont la taille est choisie par exploration de la structure périodique de la série, par exemple par moyennes mobiles.

Requête par contenu des séries temporelles

- La requête par contenu est le champ de recherche le plus actif dans l'analyse des séries temporelles.
- Il est basé sur la recherche d'un ensemble de séries les plus similaires à une série donnée dans une base de séries temporelles.
- On peut le définir formellement comme suit:
 - étant donné une série requête $Q = (q_1, \dots, q_n)$ et une mesure de similarité D
 - trouver la liste ordonnée $L = \{ T_1, \dots, T_k \}$ des séries temporelles dans une base de séries DB ...
 - tel que $\forall T_k, T_j \in L, k > j \Leftrightarrow D(Q, T_k) > D(Q, T_j)$
 - La liste L est souvent appelée K plus proches voisins de Q

Classification des séries temporelles

- L'objectif est de trouver un modèle de décision permettant de donner la classe d'une série donnée en se basant sur un ensemble de séries temporelles déjà classées.
- Ici, nous suivons la même approche de classification supervisée vue en FDA.
- Deux approches peuvent être utilisées:
 - Approche basée caractéristiques
 - Approche basée distance

Classification des séries temporelles

Approche basée caractéristiques

- Pour chaque série calculer un vecteur de caractéristiques
- La base de séries temporelles est remplacée par une base de vecteurs de caractéristiques.
- Après le calcul de ces caractéristiques pour chaque série, les méthodes d'apprentissage tel que les réseaux bayesiens, les arbres de décision et les SVMs peuvent être utilisées.

Classification des séries temporelles

Approche basée caractéristiques

- J. Rodriguez et L. Kuncheva ont proposé en 2007 une classification basée sur les caractéristiques suivantes :
 - La moyenne
 - L'écart type
 - Covariance avec l'axe du temps
 - Minimum et maximum
 - Amplitude : Maximum - Minimum
 - Fréquence : Nombre de valeurs par régions (intervalles de valeurs)
 - DTW : Distances DTW avec un ensemble de séries de références
 - DTWC : DTW avec les séries de covariance avec les séries de référence

Classification des séries temporelles

Approche basée distance

- Les distances présentées précédemment peuvent être utilisées pour une classification de type KPPV.
- La classe d'une nouvelle série est prise comme la classe majoritaire de ses K plus proches séries de la base.

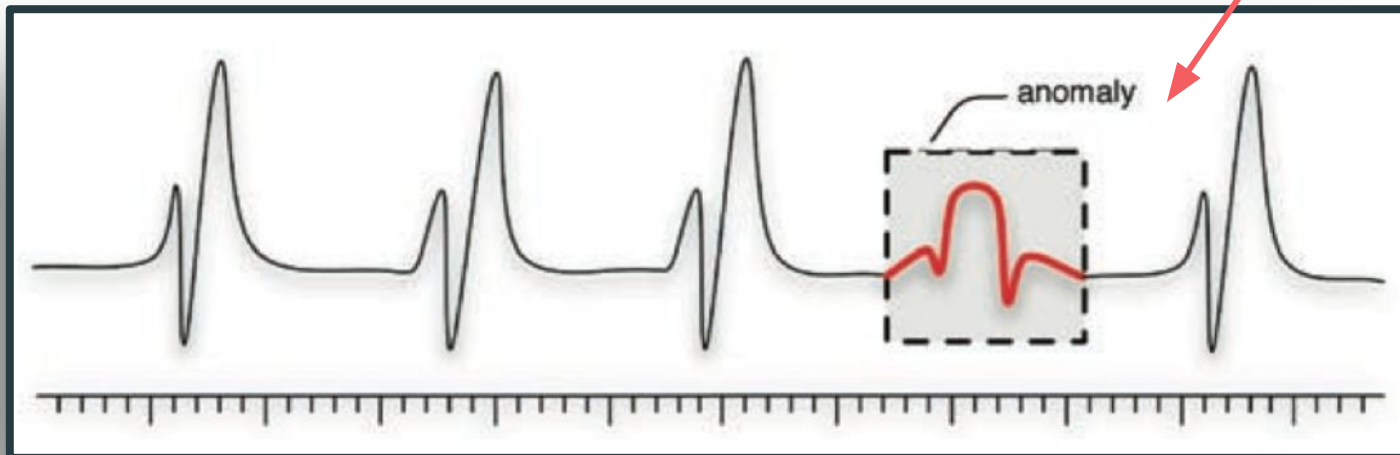
Détection d'anomalies

- La détection d'anomalies visent à trouver les sous-séquences anormales dans une série temporelle.
- La figure suivante présente un exemple d'une série avec un certain comportement périodique et une perturbation.
- L'objectif est de trouver les sous-séquences qui sortent du comportement habituel de la série, ce dernières sont considérées comme étant des anomalies

Détection d'anomalies

- La détection d'anomalies visent à trouver les sous-séquences anormales dans une série temporelle.

Comportement périodique et une perturbation.



Détection d'anomalies

- L'objectif est de trouver les sous-séquences qui sortent du comportement habituel de la série.
- Ces sous-séquences sont considérées comme étant des anomalies.
- L'approche usuelle utilisée, pour rechercher les anomalies dans une série, est de créer un modèle de comportement normal de la série puis caractériser les sous-séries qui sont trop éloignées du modèle comme des anomalies.

Détection d'anomalies

- Plusieurs approches peuvent être utilisées :
 - Effectuer un **clustering des sous-séries** d'une série donnée puis considérer les clusters de faible nombre d'éléments comme des sous-séries anormales.
 - Rechercher les **motifs rares** dans une série et les considérer des anomalies
 - Utiliser un modèle de décision **mono-classe** qui reconnaît les sous-séries d'une série donnée et rejette les autres.