

Série chronologique

R3.25



STID
Aurillac
Statistique &
informatique
décisionnelle
Cybersécurité

Paul-Marie Grollemund

Table des matières

1	Introduction	1
1.1	Exemples de données et de problématiques	2
1.2	Rappels concernant la régression linéaire	3
1.3	Projections et anomalies	4
1.4	Les objectifs	4
2	Prétraitement	7
2.1	Un tableau pour une série chronologique	7
2.2	L'échelle de temps	8
2.3	Normaliser	13
3	Lissage	19
3.1	Moyenne mobile	22
3.2	Lissage exponentiel	23
3.3	Régression locale	24
4	Tendance et périodicité	29
4.1	Tendance linéaire	30
4.2	Périodicité	30
5	Modélisation	33
5.1	Modèle additif	33
5.2	Modèle multiplicatif	34
5.3	Modèle autoregressif	35
6	Anomalies et cassures	45
6.1	Détection d'anomalies	46
6.2	Détection de cassures	46

Introduction

Dans les ressources pédagogiques précédentes, nous avons abordé différents types de données, et le propos de cette ressource pédagogique est de traiter d'un nouveau type de données : les données temporelles. Il s'agit de données correspondant à l'évaluation d'un même phénomène à plusieurs instants successifs.

Définition 1.0.1 (Donnée temporelle) *Une donnée temporelle est une mesure x_t associée à une mesure de temps t . Elle peut aussi se noter x_{t_i} pour un t_i donné.*

Définition 1.0.2 (Série chronologique) *On appelle série chronologique une collection de données temporelles ordonnées de manière chronologique. Pour une série chronologique de n observations, on la notera $(x_{t_1}, x_{t_2}, \dots, x_{t_n})$ ou $(x_{t_i})_{i=1,\dots,n}$.*

Exemple 1.0.3 (Evolution des performances). *Pour illustrer ce que sont des données temporelles, dans cet exemple nous considérons le suivi des performances d'un athlète. Prenons les meilleurs temps annuels de Usain Bolt sur l'épreuve du 200 mètres. Ci-dessous, la table 1.1 et la figure 1.1 donnent des représentation de ces données.*

On note une première chose pour ces données : pour l'année 2014 il n'y a pas de données (il est nécessaire d'en tenir compte si on souhaite étudier l'évolution de ses performances). Autre remarque, la dimension temporelle est dans ce cas mesurée en année, et il est pertinent de ré-exprimer le temps en âge (ce qui est fait dans la troisième colonne du tableau ou pour l'axe des abscisses du graphique). Cette transformation a l'intérêt de rendre comparable ces données, avec celles concernant les performances d'autres athlètes de périodes différentes. Autrement dit, comparer les performances de deux athlètes de périodes différentes, réalisées en 2002 et en 1979 ne serait pas pertinent alors que de comparer ces performances en termes d'âge a plus de sens.

Résultats au 200m	Année	Age
21.73	2001	15
20.58	2002	16
20.13	2003	17
19.93	2004	18
19.99	2005	19
19.88	2006	20
19.75	2007	21
19.30	2008	22
19.19	2009	23
19.56	2010	24
19.40	2011	25
19.32	2012	26
19.66	2013	27
19.55	2015	29
19.78	2016	30

Table 1.1 – Meilleurs temps annuels de Usain Bolt au 200 mètres en fonction de son âge.

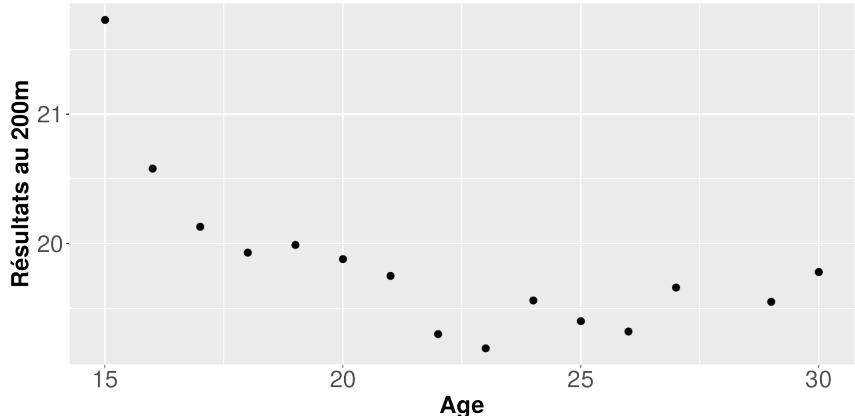


Figure 1.1 – Evolution des meilleurs temps annuels de Usain Bolt au 200 mètres en fonction de son âge.

Remarque 1.0.4 (Différentes appellations). *Pour évoquer l'analyse de ces données on trouve plusieurs appellations : données temporelles, série chronologique ou données longitudinales. Il s'agit du même type de données et des mêmes analyses, et il est donc important de s'en rappeler lorsqu'il s'agit de faire des recherches en ligne pour se documenter.*

Remarque 1.0.5 (Données temporelles quantitatives). *Il est assez commun que la donnée temporelle x_t soit une donnée quantitative, bien qu'il existe des séries chronologiques de données catégorielles. Dans le cadre de ce cours,*

nous ne traitons que du cas de données temporelles quantitatives.

Une particularité de ces données est que certaines approches sont similaires au cas de l'analyse bivariée, à savoir lorsqu'on dispose de deux mesures x_i et y_i pour chaque individu. Pour faire le lien entre les deux notions, on peut considérer par abus de modélisation qu'on dispose de deux variables x_i et t_i , et on peut alors appliquer un modèle de régression linéaire simple aux couples de données (x_i, t_i) pour $i = 1, \dots, n$. C'est d'ailleurs une approche présentée dans le chapitre 4. Cependant s'il s'agit d'un abus de modélisation, c'est que la quantité t_i n'est pas une variable qu'on échantillonne, et autrement dit elle ne correspond pas à une réalisation d'une variable aléatoire. Il y a donc une différence de nature entre un nuage de points (x_i, y_i) (analyse bivariée) et (x_i, t_i) (analyse de données temporelles). Quoi qu'il en soit, la modélisation de la régression linéaire est utile et s'adapte directement au contexte des données temporelles, mais en gardant à l'esprit que les données ne sont pas de même nature.

Dans la suite de ce document, ce premier chapitre est complété par quelques rappels ainsi que l'énoncé des objectifs de cette ressource pédagogique. Le chapitre 2 traite des traitements préalables à effectuer avant d'appliquer des modélisations. Ceci a pour objectif de préparer correctement les données à l'analyse, ce qui peut dépendre de la problématique ou du contexte d'application. Ce chapitre peut se lire indépendamment des autres, et donc il n'est pas nécessaire de le lire en premier. Ensuite, le chapitre 3 décrit des approches de lissage permettant de débruiter la série chronologique. Le chapitre 4 aborde deux notions importantes dans la modélisation des séries chronologiques : la tendance et la périodicité. Ces deux notions permettent de capturer la plupart des informations principales dans une série chronologique. Ce chapitre permet d'introduire différentes modélisations standards d'une série chronologique, dans le chapitre 5. En dernier lieu, le chapitre 6 traite des problématiques de détection d'anomalies et de ruptures (ou cassures) dans une série chronologique. Ces deux notions, assez proches, correspondent à des enjeux importants dans le contexte de la cybersécurité.

Les diapos de cours se trouvent à la fin du chapitre 1 (pour un résumé des choses à savoir avant de commencer à travailler), à la fin du chapitre 5 (faisant une synthèse des quatre chapitres précédents), et à la fin du chapitre 6.

Table des matières de ce chapitre

1.1	Exemples de données et de problématiques	2
1.2	Rappels concernant la régression linéaire	3
1.3	Projections et anomalies	4
1.4	Les objectifs	4

1.1 Exemples de données et de problématiques

Afin de compléter cette introduction, et de montrer que ce type de données se retrouvent dans de nombreux domaines, voici ci-dessous une liste d'exemples de données et de problématiques :

Evolution de la bourse Tout les jours (hors week-end) les différentes places boursières indiquent le niveau de la bourse à l'ouverture, à la fermeture et à tout les instants de la journée. Ces informations sont très suivis et sont facilement accessibles. De plus, l'analyse de ce type de données est un enjeu majeur, soit pour prévoir l'évolution de la bourse, soit pour ré-interpréter les fluctuations boursières en fonction des événements sociétaux et politiques.

Changement climatique Afin d'évaluer l'intensité du changement climatique, de nombreux facettes de l'environnement sont suivies par des équipes de scientifiques. Que cela soit la température, le niveau des rivières ou la quantité d'événements particuliers (avalanche, tornade ou autres), l'analyse de l'évolution de ces phénomènes permet de quantifier le changement climatique. De plus, avec les modélisations de ces données, il est aussi possible de faire des prévisions à moyen et long termes afin d'indiquer ce qui sera problématique à l'avenir, mais aussi d'évaluer l'impact de potentielles actions à envisager.

Processus thérapeutique Lors du développement d'une thérapie, il est nécessaire de suivre les patients sur lesquels ses tests sont réalisés, pour valider la mise en production et la commercialisation de la thérapie. Ce suivi permet entre autres de pouvoir identifier l'effet de la thérapie sur le temps court, mais aussi d'alerter en cas d'effets secondaires ou d'effets négatifs à moyen terme. Par exemple, s'il est question de développer une nouvelle thérapie comportementale auprès de patients atteints de surpoids, il est important d'observer la progression des patients ainsi qu'après. Pendant la thérapie, il est nécessaire d'évaluer la perte de poids progressive, et à la fin d'évaluer l'intensité et la rapidité de cette perte de poids.

Utilisation d'un réseau de transports en commun Les transports en commun, mis en place par une ville et potentiellement en collaboration avec une ou plusieurs entreprises, peuvent être suivis de sorte à mesurer la qualité du réseau mis en place. Cela peut consister à évaluer l'intensité et la fréquence d'utilisation à certains arrêts, afin de suggérer des modifications du réseaux. De manière similaire, il peut être intéressant

de remarquer quels sont les horaires les plus utilisés, de sorte à déployer efficacement les agents et les véhicules sur le réseau.

Election et sondage Afin de pouvoir donner des résultats pertinents concernant les tendances dans les résultats aux élections, il est nécessaire d'évaluer l'évolution de sondages successifs. Cela peut non seulement permettre de valider ou d'invalider la robustesse des résultats dans certains contextes, mais cela peut aussi être mis en regard avec des événements marquants de la campagne.

Qualité et saveur d'un aliment Lors du processus de fabrication d'un produit alimentaire (fromage, viande, vin, ...), il est possible de mesurer des facteurs déterminants dans ce processus et de chercher à mettre cela en relation avec une mesure biochimique ou sensorielle du produit final. Par exemple, il serait possible dans le cas du vin, de mesurer tout les jours la quantité de sucre dans le vin, et de regarder si cette évolution est la même pour des vins de qualité différente ou de composition chimique différente.

La liste ci-dessus n'a pas vocation à présenter l'intégralité des applications possibles de ce type d'analyse, mais ces quelques exemples peuvent nourrir votre culture concernant l'utilité des connaissances à acquérir avec ce cours.

1.2 Rappels concernant la régression linéaire

Au préalable du contenu de ce cours, cette section rappelle les notions à avoir en tête concernant la régression linéaire simple (voir la ressource pédagogique "*Statistique descriptive 2*").

La modélisation Pour une série de données x_1, \dots, x_n et une autre série y_1, \dots, y_n , le modèle de régression linéaire simple est donné par l'équation suivante :

$$y_i = \mu + \beta x_i + \varepsilon_i$$

où μ et β sont les paramètres de ce modèle, à estimer à partir des données. Cette modélisation permet de décrire une liaison linéaire entre le phénomène X et le phénomène Y . Une fois ces paramètres estimés, on obtient le modèle ajusté :

$$y_i = \hat{\mu} + \hat{\beta} x_i + e_i$$

où e_i est un résidu, à savoir l'erreur de prédiction. En particulier, une prédiction \hat{y} de ce modèle, pour une donnée x est $\hat{y} = \hat{\mu} + \hat{\beta}x$. D'un point de vue graphique, cette modélisation revient à tracer une droite, dont l'équation est donnée par l'écriture du modèle ajusté : $y = \hat{\mu} + \hat{\beta}x$, droite qui "passe par" le nuage de points.

Les estimateurs Les estimateurs sont donnés par les formules suivantes :

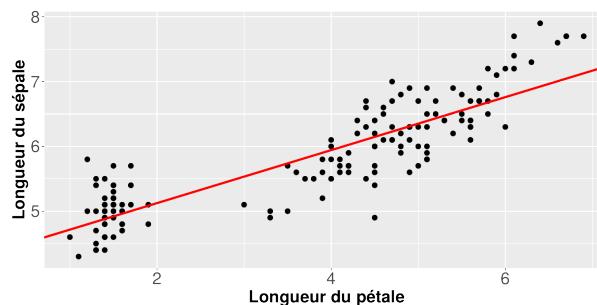
$$\begin{aligned}\hat{\beta} &= \frac{\text{cov}(x, y)}{s_x^2} = \frac{\bar{xy} - \bar{x} \times \bar{y}}{\bar{x}^2 - \bar{x}^2} \\ \hat{\mu} &= \bar{y} - \hat{\beta} \bar{x}\end{aligned}$$

Le critère des moindres carrés Ces estimateurs sont obtenus en minimisant le critère des moindres carrés :

$$C(\mu, \beta) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (\hat{\mu} + \hat{\beta} x_i))^2$$

qui correspond à une quantification de l'erreur de prédiction du modèle. Autrement dit, minimiser ce critère revient à chercher les estimateurs $\hat{\mu}$ et $\hat{\beta}$ qui fournissent des prédictions avec le moins d'erreurs (les plus proches des données observées).

Nuage de points et droite de régression Le graphique suivant correspond au nuage de points croisant les séries de données x_1, \dots, x_n et y_1, \dots, y_n auxquelles on ajoute la droite de régression. Il s'agit d'un résumé



graphique standard de la régression linéaire, permettant 1) de voir une représentation graphique du modèle ajusté, 2) de visualiser rapidement si le modèle est adapté pour décrire la liaison entre les deux variables en question, et 3) de qualifier la liaison (positive ou négative) entre les deux variables.

1.3 Projections et anomalies

Deux des objectifs importants de ce cours, à retrouver dans les chapitres 5 et 6, sont 1) de prédire l'évolution du phénomène observé, et 2) de déterminer si dans les observations à disposition il y a des anomalies.

Le premier revient à se poser la question suivante : à partir d'une série chronologique x_{t_1}, \dots, x_{t_n} , que devrait être la valeur à attendre au temps t_{n+1} , voire au temps t_{n+1}, \dots, t_{n+p} ? Ce type de problématique arrive de manière très récurrente dès lors qu'on dispose de ce genre de données.

Pour le second objectif, il s'agit d'une problématique importante dans certains contextes, et en particulier en cybersécurité. Par exemple, si on suit l'évolution du réseau informatique d'une entreprise, on aura à disposition le nombre de connexions/requêtes effectuées au fur et à mesure du temps. A partir de ces informations, il est possible d'évaluer si le niveau d'activité numérique en rapport avec le réseau de l'entreprise serait anormalement élevé et donc que cela pourrait indiquer qu'une attaque numérique a été effectuée contre l'entreprise.

1.4 Les objectifs

Pour cette ressource pédagogique, les objectifs principaux sont les suivants :

- Comprendre ce qu'est une donnée temporelle et une série chronologique.
- Connaître les prétraitements et les méthodes spécifiques à ces données.
- Savoir mettre en place une analyse sur une série chronologique et faire des prévisions.
- Avoir un regard critique et indiquer si une approche est adaptée ou non dans un contexte donné.
- Savoir synthétiser et communiquer les résultats d'une étude sur des données temporelles.

Pour s'assurer de bien comprendre ce cours et d'en tirer les enseignements essentiels, voici ci-dessous une liste des points à retenir et à garder en tête durant et après ce cours.

- Pour analyser des données temporelles, il est souvent important d'effectuer un prétraitement des données. Cela peut consister à effectuer transformation des données, à effectuer un lissage de la série, ou encore à réaliser une agrégation de données sur des périodes temporelles pertinentes.
- Afin de pouvoir présenter l'évolution principale d'une série chronologique, des méthodes de lissages sont à utiliser. Grâce à ces méthodes, on peut extraire une information, un signal, qu'on découpe du bruit présent dans les données. Autrement dit, les variations chaotiques de la série chronologique, qui ne sont souvent pas intéressantes pour la problématique envisagée, sont mises de côté et on ne conserve que l'information pertinente.
- Plusieurs modélisations sont possibles pour ce type de données. Il convient de les comprendre, de savoir les mettre en place, et de savoir laquelle mettre en place suivant les données et la problématique.
- Un aspect essentiel des ces analyses à savoir mettre en œuvre est de calculer une prévision de l'évolution de la série chronologique. Pour cela, en utilisant la modélisation adaptée au contexte, il faut être capable d'utiliser les estimateurs et les formules du cours afin d'obtenir ces prévisions. Un regard critique concernant ces prévisions est nécessaire et est au centre de ce type d'analyses.
- Dans un contexte d'analyse d'une série chronologique dans le domaine de la cybersécurité, une compétence importante est de pouvoir déetecter des anomalies ou des cassures. Pour cela, des approches spécifiques sont à utiliser, et elles sont soient relatives à l'analyse de données atypiques, soient basées sur des tests d'hypothèses particuliers.

Diapos de cours

Chap. 1 – Introduction

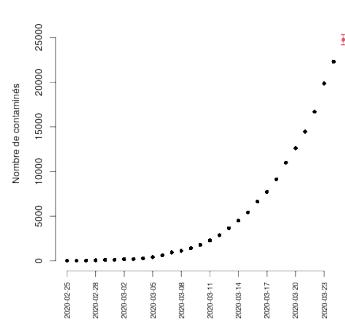
Séries chronologiques

Définition : Série statistique

Une **série statistique** est une liste de valeurs correspondant à des mesures d'un trait/caractère d'un échantillon. L'ordre de cette série n'est généralement pas important.

Définition : Série chronologique (ou série temporelle)

Une **série chronologique** est une liste de valeurs correspondant à des mesures répétées d'un trait/caractère d'un individu statistique en différents temps. L'ordre correspond à l'ordre chronologique des mesures.



Modélisation des données temporelles

Lissage :

Permet d'obtenir une version plus lisse (et continue) que la série chronologique brute.

Tendance et périodicité :

Correspond à des schémas d'évolution globale qu'on peut caler sur la série chronologique afin d'établir une modélisation.

Détection de ruptures :

Une série chronologique peut contenir des phénomènes particuliers qu'il faut être capable de détecter comme par exemple des anomalies dans la série, ou des cassures dans le régime standard de la série.

Module M2102

Objectif du module :

- Analyse des séries chronologiques
- Décomposer le signal (lissage, tendance, périodicité)
- Ajuster des modèles standards (additif et multiplicatif)
- Prédire l'évolution de la série chronologique

Enseignant à contacter : en cas de problèmes ou de questions Mail à paul_marie (dot) grollemund (at) uca (dot) fr
Ou se rendre au bureau des enseignants STID.

Déroulement du module

Séances de cours :

Semaines 36 à 41.

Evaluations : Contrôle continu : 5 à rendre chaque fin de semaine.
Contrôle de TP pendant la semaine 47.

Contrôle de connaissance pendant la semaine 47.

Note finale : moyenne des 3 notes.

SAE associée :

SAE 3.03 - Description et prévision de données temporelles
Des feuilles de TP
Un projet à rendre en semaine 1 (rentrée janvier 2023)

2

CHAPITRE

Prétraitemet

Remarque (Ordre de lecture). *Ce chapitre peut se lire indépendamment des autres chapitres, et peut donc être lu a posteriori. Le cœur du cours étant principalement dans les chapitres 3, 4 et 5, vous pourrez potentiellement souhaiter lire en priorité ces autres chapitres. Cependant, ce chapitre aborde des points essentiels pour réaliser correctement une analyse d'une série chronologique, et il faudra penser à le parcourir correctement.*

Comme dans de nombreuses situations lorsqu'on souhaite analyser de données, il est potentiellement nécessaire d'effectuer un traitement des données, au préalable de l'analyse, d'où l'appellation de "pré"-traitement. Pour ce qui est du prétraitemet des séries chronologiques, il y a des aspects spécifiques à ce type de données, qui sont abordées dans ce chapitre. En effet, étant donné la nature temporelle des données, les prétraitemets doivent tenir compte de cela et ne pas déstructurer les données d'un point de vue temporel.

Pour donner des exemples des particularités, on peut penser à des cas où ce qui est intéressant dans les données n'est pas tellement la valeur de chaque donnée prise isolément, mais plutôt les incrément correspondant au fait de passer d'une donnée à la suivante (dans l'ordre chronologique). Par exemple, l'INSEE recense régulièrement (de manière trimestrielle) le taux de chomage dans la population, en indiquant le niveau de chomage mais surtout la différence avec le taux de chomage du trimestre précédent. Dans ce cas, l'information à analyser est plutôt contenue dans les incrément que dans la série chronologique elle-même.

Ensuite on peut imaginer avoir à disposition des données quotidiennes mais pour lesquelles on ne dispose pas de valeurs pour les week end ou jours fériés. C'est le cas des données relatives aux activités des entreprises, comme les données relatives au marché boursier. Pour ce type de données, il faut être vigilant quant aux calculs de différences qu'on peut faire : bien qu'entre vendredi et lundi il y ait une différence de 3 jours, d'un point de vue de la série chronologique des valeurs de la bourse, le lundi vient juste après le vendredi.

Pour finir, il peut y avoir dans d'autres contextes des indicateurs standards du domaine qu'il faut calculer au préalable. Par exemple, dans le cas du suivi de l'évolution l'épidémie de covid-19, supposons qu'on dispose au quotidien du nombre de tests effectués dans la population, ainsi que le nombre de tests positifs. Pour étudier avec pertinence l'évolution de l'épidémie, il convient de ne pas analyser séparément chacune de ces deux séries chronologiques, mais plutôt de construire au préalable une nouvelle série chronologique : le taux de tests positifs (parmi l'ensemble des test au quotidien). Cela permet de construire une nouvelle série chronologique, qui contient une information pertinente et qui sera celle qu'on étudiera par la suite.

Dans la suite de ce chapitre, il est question de comment structurer les données temporelles en section 2.1, de ce qu'il y a savoir concernant les traitements en terme temporel en section 2.2, des différentes manières de normaliser les données temporelles en section 2.3, et en particulier en construisant une série des différences.

Table des matières de ce chapitre

2.1	Un tableau pour une série chronologique	7
2.2	L'échelle de temps	8
2.3	Normaliser	13

2.1 Un tableau pour une série chronologique

Supposons qu'on dispose d'une série chronologique x_{t_1}, \dots, x_{t_n} qu'on souhaite mettre dans un tableau de données. La bonne structure est de faire une première colonne contenant les temps de mesure (t_1, \dots, t_n) puis une seconde colonne contenant les mesures x_i . Voici ci-dessous avec la figure 2.1 un exemple d'un tableau tel qu'il devrait être dans le cas d'une série chronologique ou d'une collection de séries chronologiques. Pour ce type de données, au contraire il faut éviter les tableaux des formes présentés dans la figure 2.2. A savoir il faut éviter de

	A	B
1	Temps	Mesure
2	16:51:00	7,1
3	16:52:00	7,4
4	16:53:00	7,3
5	16:56:00	7,7
6	16:57:00	8,1
7	16:58:00	7,9

	A	B	C
1	Temps	Mesure 1	Mesure 2
2	16:51:00	7,1	0,2
3	16:52:00	7,4	0,9
4	16:53:00	7,3	NA
5	16:54:00	NA	0,7
6	16:56:00	7,7	NA
7	16:57:00	8,1	1
8	16:58:00	7,9	1,1
9	16:59:00	NA	0,9

Figure 2.1 – Des tableaux bien structurés pour stocker une ou plusieurs séries chronologiques.

mettre en ligne la série chronologique, ou de découper la série chronologique sur plusieurs colonnes, ou encore de ne pas indiquer l'échelle temporelle sur une colonne.

	A	B	C	D	E	F	G
1	Temps	16:51:00	16:52:00	16:53:00	16:56:00	16:57:00	16:58:00
2	Mesure	7,1	7,4	7,3	7,7	8,1	7,9

	A	B
1	Temps	Mesure
2	1	7,1
3	2	7,4
4	3	7,3
5	4	7,7
6	5	8,1
7	6	7,9

	A	B	C	D
1	Temps	Mesure	Temps	Mesure
2	16:51:00	7,1	16:56:00	7,7
3	16:52:00	7,4	16:57:00	8,1
4	16:53:00	7,3	16:58:00	7,9

Figure 2.2 – Les mauvaises pratiques à éviter pour stocker les séries temporelles dans un tableau.

2.2 L'échelle de temps

2.2.1 Le format

Dans cette section sont listées (de manière non-exhaustive) les différentes manières d'écrire correctement le temps d'une série chronologique dans un tableau de données.

Les dates Pour écrire une date de sorte à ce que les calculs fait sur ordinateur se passe correctement, il vaut mieux utiliser le format "ymd" ou autrement dit "yyyy-mm-dd", ce qui veut dire "Year - Month - Day". Par exemple : "2022-12-31". Cela permet entre autres qu'une méthode de tri puisse facilement ordonner plusieurs dates (puisque ces méthodes utilisent généralement l'ordre alphanumérique).

Les heures Si l'information concernant le jours n'est pas pertinente, mais que seul l'heure est importante, voici le format adéquate : "hh:mm:ss" et voici un exemple : "12:15:01".

Format complet Si on souhaite conserver toute l'information concernant la date et l'heure, on peut utiliser le format complet "yyyy-mm-dd hh:mm:ss", ce qui donne : "2022-12-31 12:15:01". Ce format est souvent accompagné d'un suffixe "UTC" (Coordinated Universal Time) ou autre comme "CET" (Central European Time).

Temps depuis une date Une autre possibilité, pour la date comme pour l'heure, est de l'écrire comme le temps écoulé (en jours ou en seconde) depuis un moment précis. Pour une heure de la journée, il s'agit du nombre de secondes écoulées depuis 00:00:00, et par exemple 85 correspond à 00:01:25. Pour une date, il s'agit du nombre de jours écoulés depuis le 1970-01-01, et par exemple 17498 correspond à 2017-11-28. Pour une date complète, il s'agit du nombre de secondes écoulées depuis 1970-01-01 00:00:00, et par exemple 1 511 870 400 correspond à 2017-11-28 12:00:00. Ce format n'est pas forcément facilement interprétable pour une personne mais peut être pratique pour effectuer des opérations sur les dates.

2.2.2 La période de mesure

Lorsqu'on dispose d'une série chronologique, il est possible que l'échelle de temps de mesure de la série puisse ne pas être la plus pertinente en fonction de la problématique. Il convient alors de formater les données de sorte à se ramener à une échelle temps adéquate. En particulier, il est surtout possible de considérer la série sur des périodes de temps plus larges, ainsi que sur des périodes de temps plus fines (même si c'est plus complexe). Il est aussi possible d'effectuer une dilatation du temps concernant les instants de mesures pour se ramener à une série exploitable, mais ceci est très spécifique à certains domaines d'application.

2.2.2.1 Echelle plus large : agrégation

Pour se donner un cas concret, supposons qu'on considère une série chronologique correspondant à la pluviométrie dans une région sur plusieurs années et qu'on souhaite mettre en lumière une potentielle diminution de la quantité de pluie. Si les données enregistrent le niveau de pluviométrie tout les jours, cela peut être une échelle temporelle très fine (trop fine d'ailleurs) pour l'analyse qu'on veut en faire. On peut alors préférer se ramener à un cumul de pluviométrie sur le mois afin de disposer d'une masse d'informations moins importante (ce qui pour des raisons pratiques peut être plus simple à travailler), mais surtout d'avoir des valeurs porteuses d'une information pertinente pour la problématique. En effet, prise isolément, chaque donnée quotidienne de pluviométrie n'informe que très peu concernant un niveau pluviométrique sur plusieurs années. Il se peut très bien qu'il ne pleuve pas ce jour-là, mais qu'il pleuve la veille ou le lendemain. Alors que s'il y a un mois entier sans pluie, cela représente en soi un événement intéressant, quelque soit le niveau de pluie des autres mois.

Le fait de passer d'une échelle à une autre pour une série chronologique, peut donc consister à faire la somme des valeurs au sein de la période souhaitée. Voici ce que cela peut donner en passant du quotidien au mensuel, pour une série chronologique (t_i, x_i) s'étalonnant sur $N = 7$ mois :

$$\begin{array}{c}
 \left. \begin{array}{cc} t_1 & x_1 \\ t_2 & x_2 \\ \vdots & \vdots \\ t_{31} & x_{31} \end{array} \right\} \xrightarrow{\sum x_i} T_1 \quad y_1 \\
 \left. \begin{array}{cc} t_{32} & x_{32} \\ t_{33} & x_{33} \\ \vdots & \vdots \\ t_{61} & x_{61} \end{array} \right\} \xrightarrow{\sum x_i} T_2 \quad y_2 \\
 \vdots \qquad \qquad \vdots \\
 \vdots \qquad \qquad \vdots \\
 \left. \begin{array}{cc} t_{183} & x_{183} \\ t_{184} & x_{184} \\ \vdots & \vdots \\ t_{214} & x_{214} \end{array} \right\} \xrightarrow{\sum x_i} T_7 \quad y_7
 \end{array}$$

Ce traitement est réalisé de sorte à ce que la série y_1, \dots, y_N (où N est le nombre de mois) pour les instants de mesures T_1, \dots, T_N , soit plus facile à analyser et plus pertinente pour répondre à la problématique.

De manière plus générale, suivant le contexte et la problématique, on peut remplacer le cumul des données sur une période ($\sum x_i$) par n'importe quel résumé statistique. Cela peut être la valeur maximale ou minimale, la médiane ou encore le comptage d'un événement (nombre de jours de pluie par exemple). Le choix du résumé correspond au contexte et à la problématique. Par exemple, s'il est question d'étudier le caractère extrême du phénomène, on pourra utiliser le min ou le max, ou encore le compte du nombre de fois qu'une valeur dépasse un seuil donné.

2.2.2.2 Echelle plus fine : interpolation

Pour ce qui est d'obtenir un niveau de détails temporel plus précis que celui des données brutes, il est nécessaire d'estimer ce que devrait être la valeur de la série chronologique, en un temps entre deux instants de mesure. Cela correspond à faire une interpolation de la série.

Définition 2.2.1 (Interpolation) Consiste à faire passer une courbe par un ensemble de points.

Remarque 2.2.2 (Interpolation et nuage de points). Cette méthode ne fonctionne pas avec n'importe quel ensemble de points. En effet, si on dispose d'un ensemble de points pour lequel on a deux points i et j ayant le même temps de mesure ($t_i = t_j$) mais n'ont pas la même valeur ($x_i \neq x_j$), alors aucune fonction ne peut passer par cet ensemble de points.

Remarque 2.2.3 (Interpolation et série chronologique). Lorsqu'on dispose d'une série chronologique, on ne dispose que d'une mesure par temps t_i , et donc le problème posé par la remarque 2.2.2 n'apparaît pas. Cependant, le problème peut arriver lorsqu'on dispose d'un échantillon issu de la mesure de deux variables et qu'on obtient des couples de valeurs (x_i, y_i) (c'est-à-dire un contexte d'application d'un modèle de régression).

Plusieurs types d'interpolation sont possibles, ils sont présentés ci-dessous et la figure 2.3 illustre chacune des trois méthodes introduites dans ce cours.

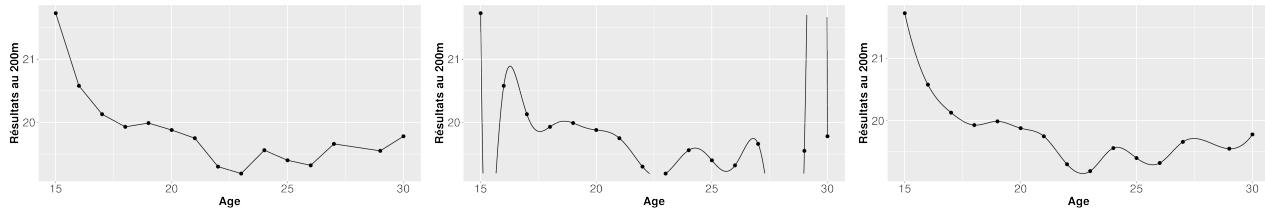


Figure 2.3 – Résultats de l'interpolation linéaire (à gauche), de l'interpolation polynomiale (au milieu) et de l'interpolation par spline cubique (à droite).

Interpolation linéaire. Cette méthode consiste à relier chaque paire de valeurs consécutives de la série chronologique par une ligne droite.

Définition 2.2.4 (Interpolation linéaire) Pour déterminer par interpolation linéaire, la valeur de la série pour un temps t compris entre t_1 et t_2 pour lesquels on dispose de mesures, on utilise la formule suivante :

$$x_t = x_{t_1} + (x_{t_2} - x_{t_1}) \times \frac{t - t_1}{t_2 - t_1}.$$

Un des problèmes de cette méthode est que, suivant le contexte, il n'est pas forcément réaliste que l'évolution soit linéaire. De plus, pour des temps proches d'un temps d'observation t_i (avant ou après) cette méthode induit de potentielles cassures. En effet, on peut constater avec le graphique de gauche de la figure 2.3 des changements de direction abruptes au niveau de certains points. Pourtant, il n'est sûrement pas réaliste que le phénomène en question admette ce type d'évolutions.

Interpolation polynomiale de Lagrange. Elle consiste à déterminer un polynôme de degré minimal qui passe par l'ensemble des points. Ce polynôme est le polynôme de Lagrange.

Définition 2.2.5 (Polynôme de Lagrange) Si on dispose d'une base de données contenant n individus, le polynôme ayant le plus petit degré passant exactement par chacun des points (t_i, x_i) est donné par :

$$P_{n-1}(t) = \sum_{j=0}^{n-1} x_j \prod_{i=1, i \neq j}^n \frac{t - t_i}{t_j - t_i}.$$

Notation 2.2.6 (Opérateur produit). L'opérateur de produit s'écrit \prod et il correspond à multiplier une suite de termes :

$$\prod_{i=1}^n a_i = a_1 \times a_2 \times \cdots \times a_n.$$

Notation 2.2.7 (Condition sous un opérateur). Lorsque en-dessous d'un opérateur de somme \sum ou d'un opérateur de produit \prod on note des informations en plus du classique " $i = 1$ ", cela correspond à définir une condition sur les termes qui doivent être pris en compte dans l'opération.

Exemple 2.2.8 (Condition sous un opérateur). Avec la même condition que celle de la définition 2.2.5, voici à quoi correspond l'opération suivante (somme ou produit) :

$$\sum_{i=1, i \neq j}^n a_i = a_1 + \cdots + a_{j-1} + a_{j+1} + \cdots + a_n \quad \prod_{i=1, i \neq j}^n a_i = a_1 \times \cdots \times a_{j-1} \times a_{j+1} \times \cdots \times a_n.$$

Pour le lire, l'opération en question concerne "toutes les valeurs i de 1 à n , sauf la valeur j ".

A noter que pour n données, le polynôme de Lagrange est de degré $n - 1$. De plus, bien qu'élégante, cette approche donne lieu à un problème qui l'exclut généralement d'une application pratique, et en particulier dès lors qu'on dispose d'un grand nombre de mesures. Le polynôme de Lagrange qu'on obtient sur une longue série de données est assez probablement affecté par le phénomène de Runge qui la rend inexploitable pour estimer correctement l'évolution de la série chronologique entre deux temps de mesure.

Définition 2.2.9 (Phénomène de Runge) Correspond à la situation où le résultat de l'interpolation polynomiale contient d'importantes oscillations pour les valeurs d'abscisses entre deux mesures.

Ce phénomène arrive dès lors qu'il y a beaucoup de points et que le degré du polynôme devient élevé, voir par exemple la figure 2.4.

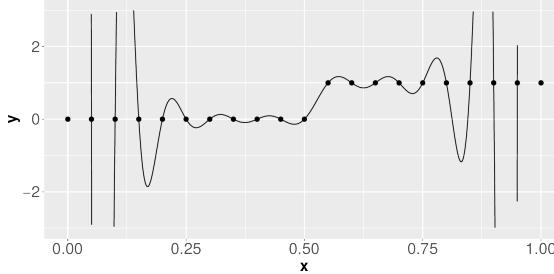


Figure 2.4 – Illustration du phénomène de Runge.

Interpolation par spline. Cette méthode consiste à relier deux mesures consécutives de la série chronologique par des polynômes de degré fixé. Il est possible d'écrire l'expression de la spline d'interpolation mais celle-ci est assez complexe. Pour se fixer les idées :

- Une spline d'ordre 1 correspond à une interpolation linéaire.
- Le graphique de droite de la figure 2.3 correspond à une spline d'interpolation de degré 3.
- Il est assez standard d'utiliser les splines d'ordre 3. Cela donne une évolution plutôt lisse d'une mesure à la suivante, et cela évite d'obtenir des cassures brutales dans l'interpolation.

Problème général avec l'interpolation. En utilisant une des méthodes d'interpolation détaillées ci-dessus, on augmente la série chronologique d'une quantité de données qui ne peuvent pas être considérées comme étant autant fiables que des données observées, ce qui peut constituer un frein majeur à l'utilisation de ce genre d'approche. Autrement dit, dans une analyse statistique, on préfère généralement n'utiliser que les données à disposition, et ne pas faire intervenir des "pseudo"-données qu'on aurait construit de manière artificielle. On ne s'autorise à le faire que si le contexte ou la problématique ne permettent pas de faire sans. De plus, le résultat final de l'analyse peut en partie dépendre de la méthode qu'on utilise pour augmenter les données, mais aussi de la quantité de données qu'on rajoute. Encore une fois, cela constitue une bonne raison pour ne pas utiliser automatiquement et intensivement ce type de méthodes.

Savoir quand interpoler ou quand ne pas interpoler ? Dans quelles situations pourrait-on judicieusement utiliser une méthode d'interpolation : lorsqu'on a de bonnes raisons de penser que l'évolution de la série chronologique entre deux valeurs consécutives n'admet pas trop de variations. Dans ce cas, une interpolation étend la série chronologique à des nouveaux instants t pour lesquels l'estimation n'est potentiellement pas trop incorrecte.

Exemple 2.2.10 (Croissance d'une plante). Supposons qu'on étudie une série chronologique relative à la croissance d'une plante sur 10 jours et que l'on souhaite évaluer en parallèle le cumul de pluie pendant ces 10 jours. Seulement les données à disposition ne sont pas sur la même échelle de temps : le cumul de pluie est donné une fois par jour (mesuré à 20h) et la taille de la plante est mesurée quatre fois par jour (8h, 12h, 16h et 20h), voir la table 2.1. Pour avoir ces deux séries chronologiques sur des échelles de temps comparables, on peut convertir les échelles de temps (Jours d'un côté et Jours-Heures de l'autre) en une échelle de temps commune et effectuer une interpolation de la série des cumuls de pluie.

Dans ce contexte, il est possible de faire une interpolation puisque le cumul de pluie est nécessairement croissant. Donc entre deux temps de mesure, l'évolution de la série ne devrait pas admettre de trop grandes variations qu'une interpolation serait incapable de reconstruire correctement. La table 2.2 présente le résultat obtenu dans ce contexte avec une interpolation linéaire.

En complément, dans quelles situations ne devrait-on pas utiliser une méthode d'interpolation : lorsqu'il paraît

Jours	Cumul de pluie	Jours	Heures	Taille de la plante
1	0.04	1	08:00:00	10.20
2	0.11	1	12:00:00	10.20
3	0.34	1	16:00:00	10.30
4	0.34	1	20:00:00	10.40
5	0.56	2	08:00:00	10.50
6	0.56	2	12:00:00	10.80
7	0.70	2	16:00:00	11.70
8	0.70	2	20:00:00	12.10
9	0.70	3	08:00:00	12.20
10	0.70	:	:	:

Table 2.1 – Séries chronologiques relatives au cumul de pluie et à la croissance de la plante.

Temps	Jours	Pluie	Temps	Jours	Heures	Plante	Temps	Pluie	Plante
0.83	1	0.04	0.33	1	08:00:00	10.20	0.33	0.02	10.20
1.83	2	0.11	0.50	1	12:00:00	10.20	0.50	0.02	10.20
2.83	3	0.34	0.67	1	16:00:00	10.30	0.67	0.03	10.30
3.83	4	0.34	0.83	1	20:00:00	10.40	0.83	0.04	10.40
4.83	5	0.56	1.33	2	08:00:00	10.50	1.33	0.07	10.50
5.83	6	0.56	1.50	2	12:00:00	10.80	1.50	0.09	10.80
6.83	7	0.70	1.67	2	16:00:00	11.70	1.67	0.10	11.70
7.83	8	0.70	1.83	2	20:00:00	12.10	1.83	0.11	12.10
8.83	9	0.70	2.33	3	08:00:00	12.20	2.33	0.23	12.20
9.83	10	0.70	:	:	:	:	:	:	:

Table 2.2 – A gauche les deux tables initiales auxquelles ont été rajoutées une échelle de temps commune (colonne "Temps") et à droite la table contenant les deux séries chronologiques après avoir réalisé une interpolation linéaire sur la série 'Pluie'.

vraisemblable que la série évolue de manière assez chaotique entre deux mesures consécutives (ce qui peut être le cas par exemple lorsque des instants de mesures sont très espacés), ou alors quand il n'y a pas de sens à définir une valeur pour des temps intermédiaires.

Exemple 2.2.11 (CAC40). *Supposons qu'on étudie l'évolution du cours de la bourse, et prenons l'exemple de l'indice du CAC40. La table 2.3 donne un exemple du type de données qu'on peut avoir dans ce contexte. Comme on peut le constater, il y a des jours pour lesquels on ne dispose pas de données : les jours 2020-02-15 et 2020-02-16. Ces deux jours correspondent à un week end, et pour toutes les données relatives à la bourse, il est normal de ne pas avoir d'information pendant ces jours qui ne sont pas des jours travaillés.*

Date	Indice
2020-02-10	6015.67
2020-02-11	6054.76
2020-02-12	6104.73
2020-02-13	6093.14
2020-02-14	6069.35
2020-02-17	6085.95
2020-02-18	6056.82
2020-02-19	6111.24
2020-02-20	6062.30
:	:

Table 2.3 – Indice du CAC40 pour quelques jours du mois de février 2020.

Dans cette situation, il ne serait pas pertinent de faire une interpolation consistant à estimer la valeur pour ces deux jours. En effet, il ne serait pas cohérent d'avoir des valeurs estimées pour ces deux jours, valeurs qui pourraient nous suggérer qu'il y ait des variations de la bourse entre le vendredi et le lundi, alors que cela n'est pas le cas. De plus, si on souhaite faire une interpolation afin de reconstruire la série chronologique entre deux mesures de la semaine (prenons entre mercredi et jeudi), de nouveau cela ne sera pas pertinent dans ce contexte. En

effet, avec la connaissance des marchés boursiers, on peut savoir qu'il y a une évolution très chaotique des indices boursiers, même au sein d'une même journée, et qu'il serait donc assez invraisemblable qu'une interpolation puisse reconstruire assez fidèlement la série chronologique.

2.2.2.3 Dilatation du temps

Contrairement aux modifications précédentes consistant à enlever ou à rajouter des instants de mesure et des mesures, par dilatation du temps on entend ici le fait d'effectuer une transformation des instants de mesure, sans modifier les mesures associées.

Définition 2.2.12 (Dilatation du temps) Pour une série chronologique (t_i, x_i) , avec $i = 1, \dots, n$, une dilatation du temps correspond à une transformation f , qui à un instant de mesure t_i associe un nouvel instant de mesure $f(t_i)$.

Une dilatation temporelle de la série chronologique $(t_i, x_i)_{i=1, \dots, n}$ donne lieu à une nouvelle série chronologique $(f(t_i), x_i)_{i=1, \dots, n}$. Pour que cette dilatation soit valide, il ne faut pas que la transformation change l'ordre des instants de mesure. Autrement dit, il faut que la transformation f soit croissante : $t_i \leq t_j \Rightarrow f(t_i) \leq f(t_j)$.

Exemple 2.2.13 (Croissance d'une plante). On s'intéresse ici à un contexte similaire à celui de l'exemple 2.2.10, pour lequel on étudie la croissance d'une plante en fonction de variations climatiques, et en particulier de l'évolution de la température. Il se trouve que si on plante à la même date (disons le 1^{er} mai) deux plants de maïs dans un même champ, au bout d'un moment il se peut que du fait d'une exposition différente au soleil (ou une différence suivant d'autre facteurs importants), les deux plants de maïs n'en soit pas au même stade phénologique de leurs croissances. Ceci est important à prendre en compte parce que suivant le stade phénologique de la plante à un instant donné, une variation de température n'a pas le même impact sur la plante. Par exemple, si le 1^{er} juillet un des deux plants en est au stade 3 (élongation de la tige principale) alors que l'autre en est au stade 6 (floraison), et s'il y a une forte température ce jour-là, cela n'aura pas le même impact sur les deux plantes. Ainsi, deux séries chronologiques concernant la température sur chacune des deux plantes ne sera pas comparable en l'état, en terme d'impact de la température sur la plante. Pour palier ce problème, il convient d'effectuer une dilatation du temps afin qu'à chaque instant de mesures t_i concernant les deux plantes, la mesure de température corresponde au même stade phénologique pour chacune des deux plantes.

2.3 Normaliser

Lorsqu'on étudie une série chronologique, il apparaît fréquemment qu'il faille normaliser les données pour ensuite les analyser.

Définition 2.3.1 (Normalisation) Consiste à effectuer une transformation aux données (souvent identique à toutes les données) de sorte à formater les données.

Une normalisation peut permettre de rendre plus facilement comparable plusieurs séries de valeurs entre elles, mais aussi à faire en sorte que la structure de l'échantillon normalisé corresponde à la structure d'une loi connue (la loi normale dans la plupart des cas). Autrement dit, une bonne normalisation permet d'annuler les variations temporelles qu'on ne souhaite pas faire émerger dans notre analyse. Si on appelle A le phénomène responsable de ces variations temporelles, au terme de l'analyse on pourra en tirer des interprétations comme : "indépendamment de la contribution du phénomène A , on constate par l'analyse de la série chronologique normalisée que ...". Pour s'assurer de capturer les variations qu'on souhaite annuler, différentes méthodes de normalisation sont possible, et voici les plus communes :

- La standardisation : Cette transformation pour une donnée x permet d'obtenir une donnée z (qui s'appelle z-score) donnée par la formule suivante :

$$z = \frac{x - m}{s}$$

où m est la moyenne de l'échantillon et s est l'écart-type empirique de l'échantillon. Cette transformation n'est souvent pas adaptée pour étudier une série chronologique.

- La normalisation Min-Max : Cette normalisation pour une donnée x permet d'obtenir une valeur x_{norm} qui est nécessairement comprise entre 0 et 1, avec la formule suivante :

$$x_{\text{norm}} = \frac{x - \min_x}{\max_x - \min_x}$$

où \min_x et \max_x sont respectivement les valeurs minimale et maximale de l'échantillon.

Exemple 2.3.2 (Standardisation). Pour les données de la taille de la plante, la moyenne est 13.08 et l'écart-type empirique est 1.53. Voici un exemple des données obtenues après standardisation, avec la table 2.4. Il faut s'attendre à ce que les valeurs obtenues soient globalement entre -2 et 2, ce qui correspondant environ à l'intervalle à 95% de la loi Normale centrée réduite. Les valeurs qui sont proches de -2 correspondent aux valeurs très basses de la série brute, et inversément pour les valeurs proches de 2. Celles qui sont proches de 0 sont celles pour lesquelles les valeurs de la série brute sont proches de la valeur moyenne.

Temps	Jours	Heures	Plante	Plante (standardisé)
0.33	1	08:00:00	10.20	-1.88
0.50	1	12:00:00	10.20	-1.88
0.67	1	16:00:00	10.30	-1.82
0.83	1	20:00:00	10.40	-1.75
1.33	2	08:00:00	10.50	-1.69
1.50	2	12:00:00	10.80	-1.49
1.67	2	16:00:00	11.70	-0.90
1.83	2	20:00:00	12.10	-0.64
2.33	3	08:00:00	12.20	-0.57
:	:	:	:	:

Table 2.4 – Séries chronologiques non-standardisée et standardisée de la taille d'une plante.

D'autres méthodes de normalisation sont envisageables et ci-dessous sont présentées certaines d'entre elles à appliquer sur des séries chronologiques. Ce qui diffère principalement entre ces différentes méthodes est : 1) par rapport à quoi on normalise, 2) comment on le fait et 3) de sorte à faire émerger quelle information dans les données. A noter de plus qu'il peut y avoir un côté arbitraire à choisir entre une normalisation ou une autre, et comme cela peut avoir un impact sur le résultat final de l'analyse (estimateur, prédition, conclusion), il y a un choix à faire qu'il faut justifier et ne pas se garder de tester les autres possibilités.

2.3.1 Normaliser par rapport à un indicateur

Une normalisation possible consiste à ramener les valeurs de la série comme une proportion par rapport à un indicateur donné. Pour cela, il suffit de diviser par la valeur M de l'indicateur en question : $z = x/M$. A noter que la valeur de l'indicateur peut ne pas être la même pour toutes les données, voir l'exemple 2.3.4.

Exemple 2.3.3 (Epidémie). Si on souhaite étudier sur un temps cours l'évolution d'une épidémie, une possibilité est d'analyser l'évolution des sous-groupes d'une population : les personnes saines, les personnes infectées et les personnes guéries. On dispose alors de trois séries chronologiques qu'on peut normaliser en divisant chaque valeur des séries chronologiques par la taille de la population. Les données informent alors sur la part de population qui est infectée, ou autre.

Exemple 2.3.4 (Démographie). Dans un autre contexte, pour étudier l'évolution démographique de différentes catégories sociaux-professionnelles, on peut diviser par la taille de la population, comme pour l'exemple 2.3.3. Cependant, si on dispose de données s'étalant sur plusieurs années, il n'est pas cohérent de diviser les données du début de la série et de la fin de la série par la même valeur. On divise alors chaque valeur des séries par la taille de la population au temps relatif à la donnée.

Pour ce type de normalisations, le choix de l'indicateur est crucial, de sorte à faire émerger l'information précise que l'on souhaite, comme illustré par l'exemple suivant :

Exemple 2.3.5 (Taux de chômage). Supposons qu'on étudier le taux de chômage en France et qu'on dispose du nombre exact de chômeurs sur plusieurs années. Si on normalise cette série par la taille de la population, on ne fait pas émerger correctement l'information la plus pertinente. Pour analyser correctement les données, il faut normaliser les valeurs de la série par la taille de la population active, à savoir la population sans compter les personnes qui ne sont pas en âge de travailler (mineurs et retraités).

Exemple 2.3.6 (Score à une l'élection). Lors des élections françaises, les résultats qui sont mis en avant correspondent aux pourcentages de votes de chacun des candidats, par rapport au nombre de suffrages exprimés. Cette manière de normaliser les données ne prend pas en compte le nombre de bulletins nuls ou blancs, ni la quantité d'abstentionnistes. Ainsi, lorsqu'on compare deux élections différentes (voir la figure 2.5), afin de s'assurer de le faire correctement, il est nécessaire de normaliser par rapport à une population pertinente, à savoir la population des personnes inscrites sur les listes électorales, et non pas seulement les bulletins exprimés. Une fois cela fait

(graphiques de droite de la figure 2.5), on peut en tirer des interprétations valides. A l'opposé, si on n'analyse que les graphiques de gauche, il n'est pas possible d'être certain que les variations observées entre 2017 et 2022 ne soient pas seulement dues à une participation différente de la population aux scrutins des deux élections.

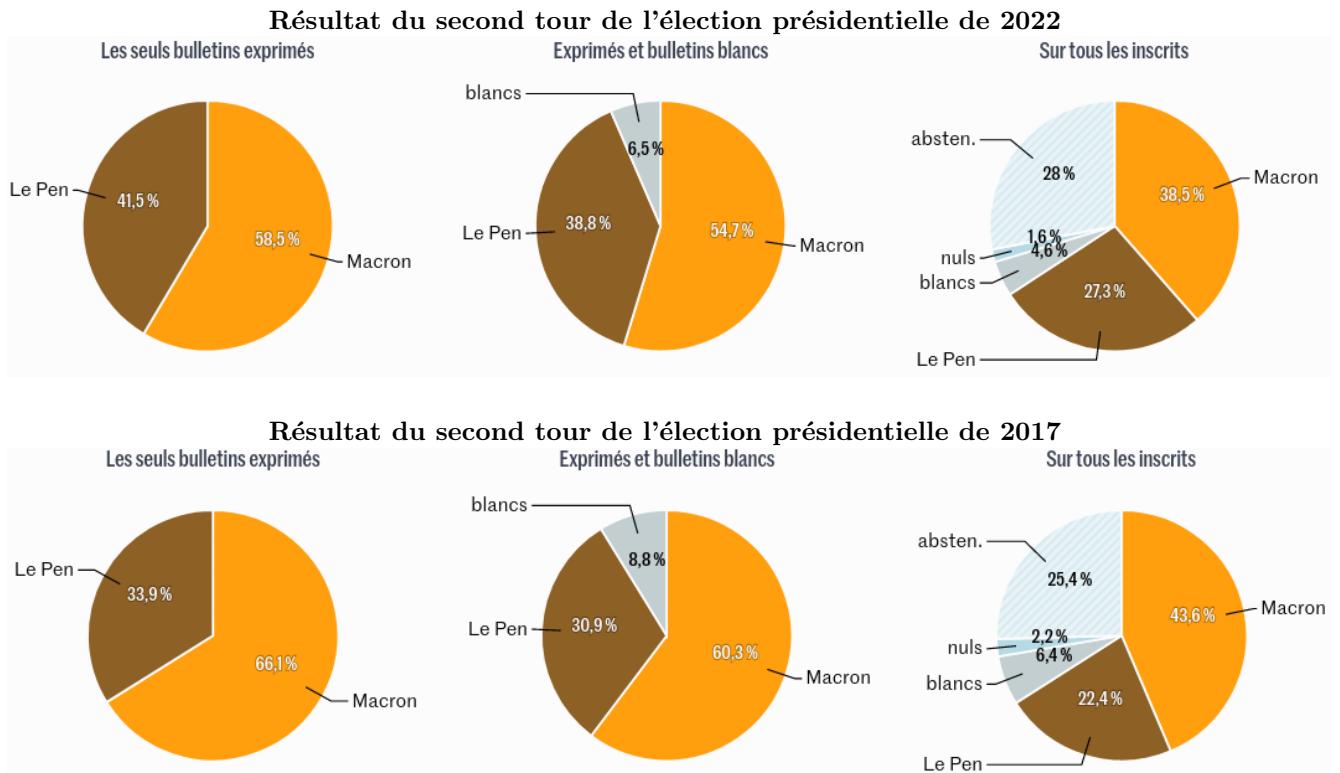


Figure 2.5 – Résultat du second tour des élections présidentielles de 2017 et de 2022, en fonction d'une comptabilisation différentes des scrutins.

Source: *Le Monde* du 25 avril 2022.

2.3.2 Normaliser par rapport à une période

Comme pour les sections précédentes, il peut être nécessaire de normaliser la série chronologique, mais cette fois-ci par rapport à une quantité temporelle, comme une période. Cela permet de ramener une valeur de la série, à une mesure pertinente au niveau d'une période donnée afin de rendre comparables plusieurs mesures, ou bien de faire apparaître un indicateur intuitif du phénomène analysé.

Pour donner des exemples de ce type de normalisation dans différents contextes, on peut penser aux exemples ci-dessous :

Exemple 2.3.7 (Consommation d'alcool). Si on étudie la consommation d'alcool de la population française sur plusieurs années, et que pour cela on dispose de la quantité d'alcool vendu sur l'année en France, on peut normaliser par la taille de la population française pour chacune de ces années (normalisation par rapport à un indicateur, voir section 2.3.1), mais aussi en divisant par le nombre de jours de chacunes des années. En divisant par le nombre de jours, on peut obtenir alors une série chronologique de la consommation moyenne par jour (et par français de part l'autre normalisation), ce qui en fait une série plus interprétable.

Exemple 2.3.8 (Transport public). Une ville souhaite étudier les différents profils d'utilisateurs du réseau de transport en commun mis en place par la commune. A savoir, est-ce qu'il y a beaucoup d'utilisateurs qui prennent les transports pour aller travailler, ou d'autres plutôt lors du week end ? sur quels horaires, et à partir de quels noeuds du réseau ? Pour répondre à ces questions, la commune dispose d'informations concernant le moment où chaque utilisateur valide son ticket à l'entrée du transport en commun. Plus précisément, on dispose d'une série chronologique par utilisateur, sur les 24 créneaux horaires de la journée, ce qui donne par exemple des mesures du type : "l'utilisateur A a validé 32 tickets entre 9h et 10h". Le problème de cette série chronologique est que chaque mesure n'est pas comparable avec celle d'un autre utilisateur parce que cela dépend en partie de depuis quand chacun des deux utilisateurs utilisent le réseaux de transport. Un utilisateur ayant nouvellement déménagé dans la ville aura nécessairement moins de validations, sur chacune des tranches horaires, qu'un autre utilisateur habitant depuis plus longtemps dans la ville. Dans cette situation on peut envisager plusieurs normalisations pour

définir pour chaque utilisateur une intensité d'utilisation du réseau de transport :

- Normaliser par le nombre de jours depuis la première utilisation. Pour chaque utilisateur cela permet d'avoir une utilisation moyenne du réseau par jour, ce qui est plus interprétable, mais ne permet forcément de rendre comparables les mesures de différents utilisateurs.
- Normaliser par le nombre de jours d'utilisation du réseau lors des 3 dernières semaines. Cela permet de rendre comparables les utilisateurs mais ne prend pas en compte les différentes périodes de l'année. Par exemple, un utilisateur qui utilise le réseau pour aller à son lieu de travail, utilisera potentiellement moins le réseau pendant ces vacances, ce qui fera que ces mesures ne seront pas comparables avec un autre utilisateur qui n'est pas en vacances.

Remarque 2.3.9 (Normaliser et longueur de la série). A noter la différence qu'il y a avec ce qui est vu dans les sections 2.2.2.1 et 2.2.2.2 : il ne s'agit pas ici de modifier la longueur de la série chronologique, puisqu'en effectuant une normalisation (transformation sur chaque donnée), on n'enlève pas de mesures et on n'en rajoute pas non plus.

2.3.3 Série des différences

Il apparaît parfois pertinent de ne pas étudier directement la série chronologique, mais plutôt les incrémentations de cette série, à savoir de combien augmente ou diminue la série d'une mesure à la suivante. Il y a principalement deux manières standards de déterminer une série des différences, présentées avec les définitions ci-dessous.

Définition 2.3.10 (Série des différences) A partir d'une série chronologique (t_i, x_i) , avec $i = 0, \dots, n$, la série des différences (t_i, y_i) , avec $i = 1, \dots, n$ est telle que

$$y_i = x_{i+1} - x_i.$$

Définition 2.3.11 (Série des différences relatives) A partir d'une série chronologique (t_i, x_i) , avec $i = 0, \dots, n$, la série des différences relatives (t_i, y_i) , avec $i = 1, \dots, n$ est telle que

$$y_i = \frac{x_{i+1} - x_i}{x_i}.$$

Remarque 2.3.12 (Différence relative et pourcentage). Une différence relative est parfois multipliée par 100 pour être exprimée en pourcentage de la valeur au temps précédent, ce qui peut être plus interprétable.

Remarque 2.3.13 (Différence, un instant en moins). Le fait de passer d'une série chronologique à une série des différences induit qu'un instant de mesure est retiré. En particulier, il s'agit du premier instant, pour lequel on ne dispose pas d'information concernant le temps d'avant, ce qui ne permet pas de calculer une différence.

Ce type de prétraitements est particulièrement fréquent dans le domaine de l'économie avec par exemple la hausse d'un indicateur économique en différence relative. D'autres exemples découlent du suivi tout les trimestres du taux de chômage, ou encore de l'évolution du cours de la bourse.

Une particularité à prendre en compte afin de construire une série des différences pertinente est le fait que, pour une mesure donnée d'une série chronologique, la mesure au temps d'avant n'est pas nécessairement comparable. Par exemple pour évaluer l'ampleur du changement climatique, il n'est pas pertinent d'évaluer la différence entre la température moyenne du mois de juin, avec celle du mois précédent. Il convient plutôt d'évaluer la différence entre le mois de juin et le mois de juin de l'année précédente. Autrement dit, pour calculer correctement une série des différences, il faut calculer une différence avec la précédente période similaire. Cette notion est approfondie dans la section 4.2, qui est consacrée à l'analyse de la périodicité d'une série chronologique.

Présentation synthétique

Pour synthétiser de manière intuitive les différentes notions présentées dans ce chapitre, voici un tableau qui indique les modifications apportées à une série chronologique suivant le prétraitement effectué :

	Temps	Mesure
Agrégation	Diminution	Diminution
Interpolation	Augmentation	Augmentation
Dilatation	Transformation	
Normalisation		Transformation

Mise en pratique des notions du chapitre 2

Exercice 2.3.1 (Agrégation). La série suivante est l'évolution sur plusieurs semaines de l'indice du CAC40. A partir de cette série, calculez la série agrégée hebdomadaire. Vous pourrez calculer plusieurs versions de cette série agrégée en résumant une semaine à l'aide de 1) la valeur moyenne, 2) la valeur minimale et 3) la valeur maximale.

Date	Indice	Date	Indice	Date	Indice	Date	Indice
2020-02-10	6015.67	2020-02-17	6085.95	2020-02-24	5791.87	2020-03-02	5333.52
2020-02-11	6054.76	2020-02-18	6056.82	2020-02-25	5679.68	2020-03-03	5393.17
2020-02-12	6104.73	2020-02-19	6111.24	2020-02-26	5684.55	2020-03-04	5464.89
2020-02-13	6093.14	2020-02-20	6062.30	2020-02-27	5495.60	2020-03-05	5361.10
2020-02-14	6069.35	2020-02-21	6029.72	2020-02-28	5309.90	2020-03-06	5139.11

Exercice 2.3.2 (Implémentation de l'interpolation de Lagrange). Avec R , écrivez une fonction `lagrange_interpolation` qui calcule l'interpolation polynomiale de Lagrange. Cette fonction devra prendre comme argument :

- `x` : le vecteur d'abscisses des points à interpoler.
- `y` : le vecteur d'ordonnées des points à interpoler.
- `xout` : le vecteur d'abscisses pour lesquels on souhaite connaître la valeur de l'interpolation.

Cette fonction renverra un `data.frame` contenant les deux colonnes suivantes :

- `xout` : le vecteur d'abscisses pour lesquels ont été calculées l'interpolation.
- `yout` : le vecteur d'ordonnées de l'interpolation.

En exécutant les lignes suivantes, vous devez obtenir le même résultat :

```
x <- seq(0,1,le=10)
y <- 1:10
xout <- c(2.5,7.5)
interpolation <- lagrange_interpolation(x,y,xout)

interpolation
  xout      yout
1   2.5  23.50000
2   7.5  68.48798
```

Exercice 2.3.3 (Interpolation). Pour les données générées dans l'exercice 2.3.2 (les vecteurs `x` et `y`), déterminez l'interpolation linéaire. Pour cela, vous pourrez calculer la valeur de l'interpolation en $t = 2.9$, $t = 3.2$ et $t = 8.4$, puis le faire numérique sur R .

Exercice 2.3.4 (Normalisation). Pour les données suivantes, calculez sur papier (puis sur R), la moyenne, l'écart-type empirique, la valeur min et max, puis standardisez la série :

Exercice 2.3.4	
t_1	-0.68
t_2	-0.70
t_3	-0.87
t_4	1.96
t_5	3.05
t_6	0.46
t_7	1.73
t_8	2.30
t_9	0.71
t_{10}	2.02
t_{11}	2.14
t_{12}	0.08
t_{13}	1.32
t_{14}	0.71
t_{15}	2.97
t_{16}	1.28
t_{17}	1.34
t_{18}	-0.26
t_{19}	2.85
t_{20}	2.74

Exercice 2.3.5			
Temps	Valeur	Temps	Valeur
1	-0.50	11	-7.60
2	-1.89	12	-7.35
3	-2.93	13	-7.33
4	-4.18	14	-7.52
5	-4.60	15	-8.08
6	-5.31	16	-7.55
7	-6.25	17	-7.26
8	-6.71	18	-6.86
9	-6.98	19	-6.08
10	-7.21	20	-5.24

Exercice 2.3.5 (Série des différences). Faites une représentation graphique de la série ainsi que de la série des différences. Indiquez quelle série vous paraît la plus simple à analyser.

3

CHAPITRE

Lissage

Lorsqu'on étudie une série chronologique, il est souvent pertinent de l'analyser dans ces grandes évolutions, et non pas de se perdre dans ses variations locales. Autrement dit, il s'agit de voir l'évolution de la série, à la hausse ou à la baisse sur de longues périodes, et ce en mettant de côté les variations potentiellement non-pertinentes. Pour faire cela, il est possible de faire de l'agrégation (voir la section 2.2.2.1) sur de longues périodes choisies au préalable, mais pour le faire correctement il faut plutôt utiliser une méthode de lissage.

Une méthode de lissage a justement pour objectif de décomposer l'évolution d'une série chronologique, en une partie concernant l'évolution globale, et une autre partie contenant les variations chaotiques autour de cette évolution globale. En reformulant avec un autre vocabulaire, on dit qu'on opère une décomposition du signal, en faisant apparaître d'un côté le signal (lequel contient de l'information à analyser) et d'un autre côté le bruit (qui ne contient pas d'information et qu'il faut laisser de côté) : données = signal + bruit. Le cours de "Statistique descriptive 2" traite de la même décomposition, pour laquelle on parle plutôt de décomposer les données avec une partie modélisation plus un terme d'erreur. Bien que ces deux approches (régression et lissage) soient proches en terme d'objectif, elles diffèrent dans certains de leurs aspects, et dans le cas spécifique des séries chronologiques, on utilise généralement les méthodes de lissage présentées dans ce chapitre. Une transposition du modèle de régression est introduit dans les chapitre 4 et 5.

Remarque 3.0.1 (Lissage et interpolation). *La différence entre le lissage et l'interpolation (voir la section 2.2.2.2) est qu'une série lissée ne passe pas nécessairement par les valeurs de la série brute, contrairement à la série interpolée. On pourrait initialement souhaité qu'une reconstruction/modélisation d'une série chronologique donne un résultat qui passe par chacune des valeurs de la série. Pour autant, on peut avoir de bonnes raisons de ne pas vouloir imposer cela. En effet, si on se rappelle que les valeurs de la série chronologique sont bruitées, il peut être souhaitable de reconstruire une version cette série qui "passe pas loin" des valeurs bruitées de la série brute, et donc qui puisse avoir une chance de représenter fidèlement le phénomène étudié.*

A noter que ces approches de lissage permettent d'éviter certaines des méprises connues concernant l'étude des séries chronologiques, à savoir par exemple le fait de attribuer une interprétation trop importante aux "pics" observées. L'exemple 3.0.2 ci-dessous illustre comment il serait possible de se tromper, si on se focalise sur les grands écarts observés entre deux pics, en oubliant qu'il faudrait plutôt évaluer l'évolution de la série à l'aide d'une méthode de lissage.

Exemple 3.0.2 (Des pics trompeurs). *On s'intéresse à l'évolution du covid-19 en France et on se focalise sur le mois de février 2021, période à laquelle les déclarations médiatiques et gouvernementales concernant la nécessité d'établir un troisième confinement étaient de plus en plus nombreuses. La figure 3.1 illustre cette évolution des nouveaux cas au quotidien. Pour montrer ce que pourrait être une mauvaise analyse (et dont il faut se méfier), voici ce qu'on pourrait mettre en avant pour indiquer qu'il y a une baisse du nombre de nouveaux cas : "on constate une baisse récente importante de 36% entre le 24 février et le 28 février". Vous pouvez le constater sur le graphique, les chiffres passent effectivement de un peu plus de 30k cas à 20k cas.*

Malheureusement, l'indicateur utilisé ici est trompeur et malhonnête. Il suffit de voir que le mercredi 24 février est une journée avec un pic de contaminations, alors que le 28 février est un dimanche et que très peu de contaminations sont comptabilisées ce jour-là comme n'importe quel dimanche. Mais de plus, le fait de passer de 30k cas à 20k cas sur une semaine, n'a rien de remarquable dès lors qu'on étudie l'évolution sur tout le mois de février. Ce qu'on constate alors, c'est qu'il y a de grandes variations du nombre de nouveaux cas au sein d'une même semaine, et que de comparer les deux valeurs extrêmes lors de la dernière semaine, est une manière trompeuse de présenter les données.

De plus, il est à noter qu'il est tout à fait possible de procéder de la même manière en comparant le 22 février avec le 24 février pour vouloir montrer de manière malhonnête qu'il y a une hausse notable du nombre de cas. Mais pour conclure correctement, au vu de ce graphique, on peut dire qu'il ne semble pas y avoir de tendance à

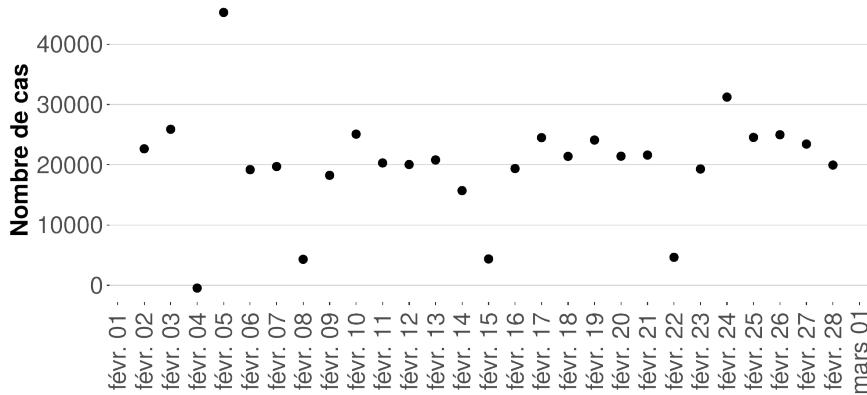


Figure 3.1 – Evolution du nombre de nouveaux de Covid-19 en France pendant le mois de février 2021.

Source: COVID-19 Data Hub, Journal of Open Source Software.

la hausse ou à la baisse, et que les nouveaux cas quotidien durant ce mois-ci sont restés stables. Et si vous pensez que ce type d'erreur n'arrive pas, voici ce que cela donne lorsqu'on ne maîtrise pas la question, on n'est alors pas à l'abri de se ridiculiser : voir sur Youtube [Pujadas qui explique l'épidémiologie à une épidémiologiste](#).

Une autre erreur possible dans ce cas d'étude consiste à se focaliser sur la dernière évolution, sans prendre en compte le passé plus ou moins récent du phénomène étudié. Plus précisément, cela consiste à interpréter l'augmentation (ou la diminution) récente d'un indicateur, en concluant que cela indique une variation notable, ou que cela s'explique par tel ou tel phénomène, alors que si on évaluait l'évolution à plus long terme, on pourrait se rendre compte que cette dernière augmentation/diminution n'a rien de très différent par rapport à ce qui s'est passé jusque là. Dans ce cas, utiliser une méthode de lissage, et mettre en lumière ce que sont les variations chaotiques, permet de relativiser ce que l'on peut croire en ne regardant que la fin de la série chronologique.

Exemple 3.0.3 (Derniers sondages!). Voici le début d'un article publié dans "Entreprendre", le 13 avril 2022, dont le titre est "Emmanuel Macron en hausse" :

Le baromètre OpinionWay publié, ce mercredi, pour CNews donne Emmanuel Macron à 54% d'intentions de vote au deuxième tour de l'élection présidentielle, ...

Le fait d'indiquer dans le titre que le dernier sondage indique une hausse, alors que ce n'est pas un terme repris dans le début de l'article, est ici malhonnête, et ce à plus d'un titre. En effet, non seulement il n'est pas indiqué de combien est cette hausse, ni si cette hausse représente une hausse remarquable par rapport aux sondages effectués lors des semaines précédentes. En ne présentant que cet indicateur, on ne laisse pas la possibilité au lecteur de savoir ce qu'il en est réellement. Si en plus on n'est pas conscient des biais de présentation de ce type-là, on n'aura pas la capacité de remettre en question ce qui est annoncé. Et malheureusement, si on regarde plus dans le détails, avec la figure 3.2 qui est issue du [même sondage](#) publié par OpinionWay que l'auteur cite, on se rend compte qu'il ne s'agit pas d'une hausse importante. De plus, on peut aussi constater que les oscillations des résultats des sondages depuis janvier, vont de 53% à 59% pour le candidat Macron, ce qui est presque recouvert par la marge d'incertitude qui est donnée dans le document : $\pm 2.6\%$. Autrement dit, même s'il y a eu des variations statistiques au fur et à mesure des différents sondages, il n'y a probablement eu aucun changement marquant dans les pourcentages d'intentions de vote, au niveau de la population.

Exemple 3.0.4 (Bourse à la baisse!). Une autre illustration du phénomène en question est issue du suivi quotidien de la bourse. Voici ce qui a pu être entendu à la radio, le chroniqueur principal de l'émission annonce dans les titres des infos :

"..., et pour finir, hier la bourse a fermé à la baisse."

passant ainsi la main à la chroniqueuse dédiée aux questions économiques :

"Oui effectivement, mais cette baisse ne fait que correspondre exactement au rattrapage de la veille."

On apprend donc avec cette précision que cette baisse ne s'inscrit pas dans une tendance puisqu'elle fait suite à une hausse d'amplitude similaire la veille. Mais de plus, on comprend avec le terme de "rattrapage" que la hausse de la veille fait suite à une baisse l'avant-veille. Ce qu'on peut comprendre ici de la situation, c'est que la bourse alterne hausse et baisse comparables sur les 3 jours précédents, et ce qu'il faut en tirer comme conclusion, c'est que le niveau de la bourse semble en réalité osciller quotidiennement autour d'un niveau stable. Or, l'information a été présentée comme une baisse et si on ne prend pas de recul concernant ce qui a été dit, on ne gardera que cette information et on aura tendance à penser qu'actuellement la bourse est à la baisse.

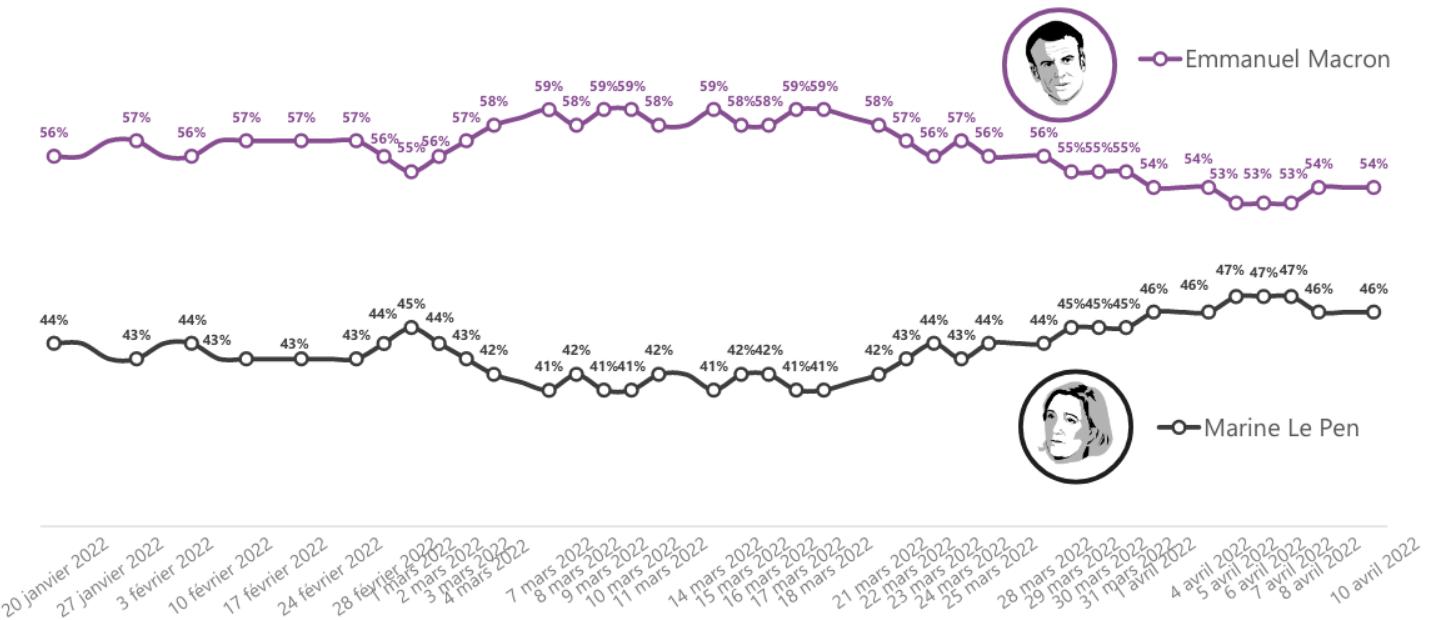


Figure 3.2 – Evolution des sondages des intentions de votes pour le second tour de l'élection présidentielle de 2022.

Source: OpinionWay pour CNEWS et Europe 1 - Sondage soir du vote - 10 avril 2022

Avant d'aller plus loin dans l'explication de ce qu'est une méthode de lissage, il est à noter qu'il y a un problème important avec ce type de méthodes, car il faut fixer la valeur d'un paramètre, et en paramètre est en lien avec le niveau de lissage de la série chronologique. Autrement dit, on doit choisir si on souhaite obtenir une version très lissée ou très peu lissée de la série chronologique. Une version très peu lissée de la série donne une courbe qui "passe pas très loin" des points du nuage de points, voir graphique de gauche de la figure 3.3. A l'opposée, une version très lissée donne une évolution trop grossière de la série chronologique et est potentiellement très éloignée de certains points du nuage de points (graphique de droite). Entre ces deux cas extrêmes, un niveau plutôt raisonnable de lissage pourra donner le graphique du milieu, mais rien ne pourra indiquer si on a raison ou pas en choisissant ce niveau de lissage. Pour le voir autrement, il faut se rendre compte que la méthode de lissage peut être utilisée dans

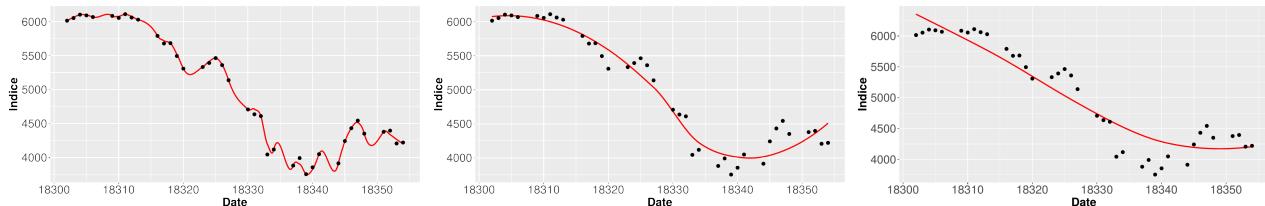


Figure 3.3 – Lissage d'une série chronologique concernant l'indice du CAC40. De gauche à droite, les graphiques donnent le résultat de la procédure de lissage (via régression locale) en allant d'un paramètre de lissage faible à un paramètre de lissage élevé.

l'optique de reconstruire le phénomène sous-jacent derrière la série chronologique. En effet, il y a des contextes pour lesquels, même si on ne mesure le phénomène qu'à des moments précis, on peut s'attendre qu'entre deux instants de mesure, l'intensité du phénomène existe réellement et évolue d'une certaine manière. Par exemple, pour la météo, entre deux mesures de température, il y a du sens à considérer que la valeur de température existe et évolue de manière plutôt douce (si les instants de mesure sont proches dans le temps). On peut alors vouloir estimer/reconstruire l'évolution du phénomène entre deux instants successifs pour avoir accès à une représentation plus fidèle ce qu'est le phénomène : un phénomène continu dans le temps. Donc cela peut donner l'impression qu'une méthode de lissage permette de reconstruire le *vrai modèle* de la production de la donnée. Cependant, ne disposant que de données bruitées, et ne disposant (justement) pas de données entre deux données successives, il n'y a aucune raison de penser qu'une série lissée est meilleure qu'une autre avec un niveau de lissage différent. Il y a donc un problème important : on ne peut pas rigoureusement bien choisir le niveau de lissage de la série chronologique. Pour conclure, étant donné ce problème, on se contentera de lisser des séries chronologiques, de sorte à obtenir une évolution vraisemblable par rapport à ce qu'on pense a priori de la régularité de l'évolution du phénomène en question, mais seulement si cet avis a priori est valide.

Dans la suite de ce chapitre, trois méthodes de lissages sont présentées. La section 3.1 introduit la méthode des moyennes mobiles (ou moyenne glissante), qui est une approche très intuitive. Ensuite, la section 3.2 présente le lissage exponentiel (simple et double) basé sur un calcul par récurrence sur le temps. Enfin, la section 3.3 donne la méthode de régression locale, qui utilise la notion de régression linéaire pour ajuster "très localement" l'évolution de la série chronologique.

Table des matières de ce chapitre

3.1 Moyenne mobile	22
3.2 Lissage exponentiel	23
3.3 Régression locale	24

3.1 Moyenne mobile

Cette méthode consiste faire des calculs simples (moyenne) pour obtenir une version lissée d'une série chronologique. Pour cela, l'idée est de remplacer la valeur de la série en un instant, par une moyenne des valeurs pour des instants proches.

Définition 3.1.1 (Moyenne mobile) *Pour une série chronologique (t_i, x_i) , avec $i = 1, \dots, n$, et une valeur entière k , la série obtenue par moyenne mobile d'ordre k est notée (t_i, m_i) , avec $i = k+1, \dots, n-k$ et se calcule par la formule suivante :*

$$m_i = \frac{1}{2k+1} \sum_{j=-k}^k x_{i+j}.$$

Remarque 3.1.2 (Autre convention pour l'ordre k). *On peut aussi voir une autre convention pour définir l'ordre k , qui correspond alors au nombre total de valeurs prises pour calculer la moyenne. Dans ce cas, si on choisit un ordre $k = 7$, cela signifie qu'on effectue pour chaque donnée x_i une moyenne avec 7 valeurs, à savoir en prenant : x_i , les trois valeurs après $(x_{i+1}, x_{i+2}, x_{i+3})$, et les trois valeurs avant $(x_{i-1}, x_{i-2}, x_{i-3})$. Avec la définition 3.1.1, cela correspondrait à une moyenne mobile d'ordre 3.*

Pour cette méthode, le paramètre qui calibre le niveau de lissage est l'ordre k de la moyenne mobile. Plus k est petit (proche de 1), plus la série lissée donne des résultats proches de la série brute pour les instants de mesure. Au contraire, plus k est grand, plus la série lissée donne des résultats éloignés de la série brute, mais proche de la moyenne globale de la série. La figure 3.4 illustre ce que peut donner cette méthode avec plusieurs niveaux de lissage.

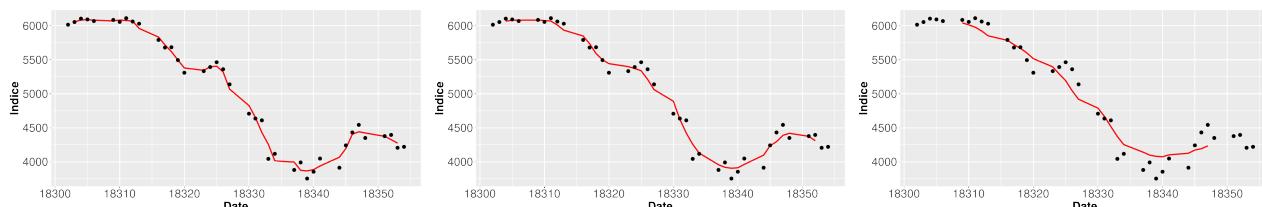


Figure 3.4 – Résultats d'un lissage par moyenne mobile pour différentes valeurs d'ordre k . De gauche à droite : $k = 1$, $k = 2$, et $k = 5$.

Un aspect potentiellement problématique de cette méthode est que cette méthode introduit ce qu'on appelle un effet de bord, à savoir les premières et les dernières valeurs de la série brute disparaissent par l'effet du lissage. Même si ces valeurs sont prises en compte dans les calculs du lissage, on ne dispose pas de valeurs lissées pour ces instants de mesure.

Définition 3.1.3 (Effet de bord) *Un effet de bord est un d'effet secondaire, et potentiellement néfaste, d'un traitement statistique qui impacte les bords du domaine des données. En particulier, pour une série chronologique, cela concerne les premières et les dernières valeurs de la série.*

Concernant les valeurs possibles de k , on préfère généralement prendre des valeurs petites $k = 1, \dots, 10$, mais suivant le contexte et le caractère erratique de la série chronologique, on peut choisir une grande valeur pour k . Quoi qu'il en soit, il faut éviter de prendre une valeur de k qui pourrait se rapprocher de n , afin d'éviter que l'effet de bord ne soit trop important. Par exemple, avec $k = (n-1)/2$ (si n est impair), on passe d'une série chronologique à n valeurs à une série lissée avec une seule valeur, ce qui n'est d'aucun intérêt lorsqu'on souhaite seulement lisser la série. Pour conclure, on pourra s'autoriser à monter la valeur de k jusqu'à environ $n/10$ si les

bords de la série chronologique ne contiennent pas l'information qu'on cherche à analyser (même si cette borne est arbitraire). A noter, que pour un ordre $k = 0$, la moyenne mobile donne les mêmes valeurs que celles de la série brute, ce qui ne correspond donc pas à un lissage.

Remarque 3.1.4 (Généralisations et variantes). *A partir de cette première approche très simple, il est possible de construire un grand nombre de généralisations (modèle ARIMA et autres par exemple) ou de variantes (par exemple moyenne non-symétrique) des moyennes mobiles. Notamment, des généralisations sont possibles afin d'éviter les effets de bord (lissage par noyaux).*

3.2 Lissage exponentiel

Une approche alternative à la moyenne mobile est le lissage exponentiel. Cette approche consiste à calculer successivement les valeurs de la série lissée à partir de la valeur de la série brute, et de celle de la dernière valeur déjà calculée de la série lissée.

Définition 3.2.1 (Lissage exponentiel simple) Pour une série chronologique (t_i, x_i) , avec $i = 1, \dots, n$, et un coefficient γ entre 0 et 1, la série obtenue par lissage exponentiel simple est notée (t_i, z_i) , avec $i = 1, \dots, n$ et se calcule par la formule suivante :

$$z_i = \gamma x_i + (1 - \gamma)z_{i-1}.$$

Ce calcul par récurrence est une moyenne entre deux valeurs, la valeur de la série chronologique et la dernière valeur de la série lissée. Cette moyenne est en particulier une moyenne pondérée, où la pondération est calibrée par le coefficient γ . A savoir, si γ est proche de 0, alors la moyenne calculée sera en grande partie déterminée par la valeur z_{i-1} et si elle est proche de 1, elle sera principalement déterminée par la valeur x_i . On peut comprendre ainsi que le coefficient γ est un paramètre de lissage : il permet de calibrer à quel point la série lissée est proche ou non de la série brute. La figure 3.5 illustre les différents résultats obtenus avec différents niveaux de lissage.

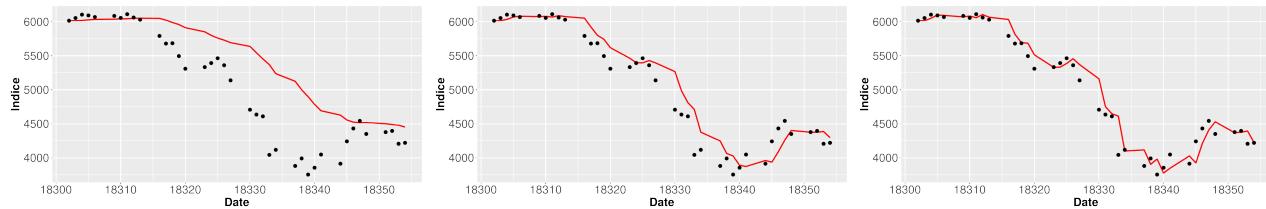


Figure 3.5 – Résultats obtenus par lissage exponentiel simple avec différentes valeurs pour le coefficient γ . De gauche à droite : $\gamma = 0.1$, $\gamma = 0.5$ et $\gamma = 0.9$.

De plus, comme pour toute procédure par récurrence, il est nécessaire de déterminer une valeur initiale. Pour se rendre compte de cela, voici la formule du lissage exponentiel pour la première valeur de la série ($i = 1$) :

$$z_1 = \gamma x_1 + (1 - \gamma)z_0.$$

Pour ce calcul, on a besoin de la valeur z_0 , et comme on ne dispose pas de cette valeur, il faut la calibrer manuellement. Un choix standard par défaut consiste à fixer cette valeur avec la valeur moyenne du début de la série :

$$z_0 = \frac{1}{3}(x_1 + x_2 + x_3).$$

Cette première valeur a un impact important sur les premiers calculs et son impact s'estompe au fur et à mesure de la série lissée. Cependant, si on le fixe très mal (en choisissant une valeur arbitraire qui soit éloignée des valeurs de la série chronologique), cela peut induire un biais important pour une bonne partie de la série.

Un problème qu'on peut rencontrer avec cette méthode de lissage, c'est qu'en présence d'une rupture abrupte de la série, il peut y avoir une période pendant laquelle la série lissée n'arrive pas à capturer cette rupture et donne alors une version lissée ne représentant que très mal la série brute sur cette période, et le problème est accentué si le coefficient γ est faible. La figure 3.6 illustre ce phénomène et fait une comparaison sur cet exemple des trois méthodes présentées dans ce cours pour effectuer le lissage.

Une autre version de cette approche est la méthode de lissage exponentiel double pour laquelle il suffit d'exécuter un lissage exponentiel deux fois, à savoir utiliser un lissage exponentiel sur la série lissée par lissage exponentiel.

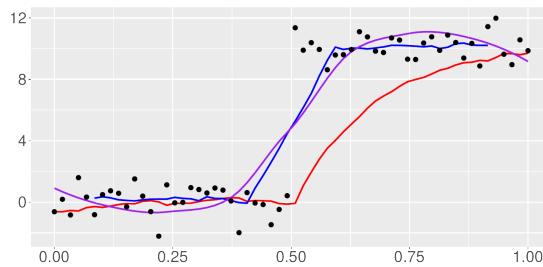


Figure 3.6 – Illustration des biais des différentes méthodes en présence d'une rupture abrupte dans la série chronologique. En rouge : lissage exponentiel simple avec $\gamma = 0.1$. En bleu : moyenne mobile d'ordre 5. En violet : régression local avec un paramètre de lissage $h = 0.7$ (voir la section 3.3).

Définition 3.2.2 (Lissage exponentiel double) Pour une série chronologique (t_i, x_i) , avec $i = 1, \dots, n$, et un coefficient γ entre 0 et 1, la série obtenue par lissage exponentiel double est notée (t_i, w_i) , avec $i = 1, \dots, n$ et se calcule par les formules suivantes :

$$\begin{aligned} z_i &= \gamma x_i + (1 - \gamma)z_{i-1}, \\ w_i &= \gamma z_i + (1 - \gamma)w_{i-1}. \end{aligned}$$

Remarque 3.2.3 (Plusieurs versions). Il existe deux versions différentes de cette approche. Celle présentée ci-dessus est la méthode de Brown, et pour l'autre, celle de Holt-Winters, elle n'est pas présentée dans ce cours.

3.3 Régression locale

Remarque 3.3.1 (Lecture). Cette section contient des aspects techniques qui peuvent être complexes. Il est conseillé de lire ce qui suit en gardant en tête qu'il faut principalement comprendre le principe de cette approche et de savoir l'appliquer avec les commandes de R.

Dans cette section, on note la fonction f correspondant à la version lissée qu'on cherche à déterminer. Cette fonction est une version continue de la série chronologique (qui elle est discrète). Et pour rappel, un méthode de lissage permet de décomposer la série chronologique en une partie lissée et une partie chaotique, ce qui revient à écrire l'équation suivante :

$$x_t = f(t) + \varepsilon(t),$$

ce qui se lit comme suit : en chaque temps t , la valeur x_t de la série chronologique peut se décomposer comme 1) la valeur de la série lissée $f(t)$, plus 2) une variation $\varepsilon(t)$ différente pour chaque temps t . Cette méthode se base sur l'idée que cette fonction f (si elle est deux fois différentiable) peut se voir localement comme une droite. Autrement dit, pour une valeur de h infinitésimale, on peut écrire (formule de Taylor d'ordre 1) :

$$f(t+h) \approx a + b(t+h)$$

où a et b sont des inconnues à déterminer. On peut alors remarquer que cela ressemble fortement à un contexte de régression linéaire. Ainsi, pour le formuler grossièrement, la régression locale consiste à effectuer des régressions linéaires, localement pour chaque valeur de t .

Pour arriver à exprimer la fonction f en un temps t dans ce contexte, il faut réussir à estimer a et b à partir des données de la série chronologique. Et en particulier, on peut noter que ces deux coefficients sont spécifiques au temps t , et qu'ils seront donc potentiellement différents pour deux instants t_1 et t_2 différents. Ainsi, l'objectif est d'estimer les quantités \hat{a} et \hat{b} , pour chaque temps t , et c'est ce pourquoi on pourra aussi les noter $\hat{a}(t)$ et $\hat{b}(t)$.

Une autre observation importante est que pour avoir un caractère "localement en t " pour une régression, il faut que la droite de régression soit en grande partie déterminée par les mesures x_{t_i} pour des instants t_i proches de t . Prenons par exemple la série chronologique suivante :

Jours	Heures	Taille de la plante
1	08:00:00	10.20
1	12:00:00	10.20
1	16:00:00	10.30
1	20:00:00	10.40
2	08:00:00	10.50
2	12:00:00	10.80
2	16:00:00	11.70
2	20:00:00	12.10
3	08:00:00	12.20

et supposons qu'on en soit à vouloir effectuer une régression linéaire locale à un instant $t = 10:00:00$ du deuxième jour. Dans ce cas, on souhaiterait que les temps proches ($t_5 = 08:00:00$ du deuxième jour et $t_6 = 12:00:00$ du deuxième jour) contribuent plus que les temps éloignés (comme par exemple $t_1 = 08:00:00$ du premier jour et $t_9 = 08:00:00$ du troisième jour). Pour faire cela, il faut définir des estimations pour lesquelles les données n'ont pas toutes le même poids, ce qui passe par définir des poids à chacune des données et aussi par l'utilisation d'un critère des moindres carrés pondérés. Dans ce contexte, le poids en question pour une donnée (t_i, x_i) pour une régression linéaire locale en t dépend directement de l'écart entre t_i et t , ainsi que d'un paramètre de lissage h strictement positif, au travers d'un calcul par une fonction de poids.

Définition 3.3.2 (Fonction de poids) *On note W une fonction de poids, qui à un écart $u = \frac{t-t_i}{h}$ associe un poids qui vérifie :*

- Positivité : $W(u) \geq 0$,
- Symétrie : $W(u) = W(-u)$.

Exemple 3.3.3. Une fonction de poids commune dans ce contexte est la fonction bicarré :

$$W(u) = \begin{cases} (1-u^2)^2 & \text{si } -1 \leq u \leq 1 \\ 0 & \text{sinon.} \end{cases}$$

Voici ce que cela donne sur la série chronologique de la taille de la plante présentée plus haut, avec $t = 10:00:00$ du deuxième jour, pour plusieurs valeurs possibles du paramètre de lissage h :

Jours	Heures (t_i)	Taille	$t - t_i$	$h = 10$		$h = 50$		$h = 100$	
				$(t - t_i)/h$	W	$(t - t_i)/h$	W	$(t - t_i)/h$	W
1	08:00:00	10.20	-24.00	-2.40	0.00	-0.48	0.59	-0.24	0.89
1	12:00:00	10.20	-20.00	-2.00	0.00	-0.40	0.71	-0.20	0.92
1	16:00:00	10.30	-16.00	-1.60	0.00	-0.32	0.81	-0.16	0.95
1	20:00:00	10.40	-12.00	-1.20	0.00	-0.24	0.89	-0.12	0.97
2	08:00:00	10.50	-2.00	-0.20	0.92	-0.04	1.00	-0.02	1.00
2	12:00:00	10.80	2.00	0.20	0.92	0.04	1.00	0.02	1.00
2	16:00:00	11.70	6.00	0.60	0.41	0.12	0.97	0.06	0.99
2	20:00:00	12.10	10.00	1.00	0.00	0.20	0.92	0.10	0.98
3	08:00:00	12.20	22.00	2.20	0.00	0.44	0.65	0.22	0.91

A noter que le paramètre de lissage h permet en particulier de normaliser les différences temporelles, qui initialement prennent potentiellement de fortes valeurs (de -24 à 22 dans l'exemple 3.3.3), de sorte à ce que des valeurs obtenues puissent être entre -1 et 1 , puisque la fonction (de poids) bicarré W n'est pas sensible aux valeurs qui ne sont pas dans cet intervalle. Sans cette normalisation, aucune des données aurait eu un poids plus grand que 0 (voir les valeurs de la colonne $t - t_i$).

En attribuant ainsi des poids à chacune des données pour l'ajustement de la régression locale en t , on peut calculer le critère des moindres carrés pondérés comme suit.

Définition 3.3.4 (Critère des moindres carrés pondérés) *Le critère des moindres carrés pondérés à minimiser pour trouver les estimateurs $\hat{a}(t)$ et $\hat{b}(t)$ de la régression linéaire locale en t est :*

$$C_t(a, b) = \sum_{i=1}^n W\left(\frac{t - t_i}{h}\right) (x_i - (a + b(t - t_i)))^2.$$

Cette équation revient à évaluer les écarts aux carrés entre chaque mesure x_i et chaque prédition, mais la prise en compte des poids fait que 1) des écarts concernant des données avec un faible poids ne sont finalement pas si importantes et 2) à l'inverse des écarts concernant des données avec un poids élevé sont importantes.

Ce critère des moindres carrés admet une solution simple à calculer à partir des notations suivantes :

Notation 3.3.5 (Matrice des poids W_t). On note W_t la matrice diagonale suivante :

$$W_t = \begin{pmatrix} W\left(\frac{t-t_1}{h}\right) & 0 & \dots & 0 \\ 0 & W\left(\frac{t-t_2}{h}\right) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & W\left(\frac{t-t_n}{h}\right) \end{pmatrix}$$

Notation 3.3.6 (Design X_t et vecteur des mesures). Pour la régression linéaire locale, on note la matrice de design (avec deux colonnes) et le vecteur des mesures de la manière suivante :

$$X_t = \begin{pmatrix} 1 & t - t_1 \\ 1 & t - t_2 \\ \vdots & \vdots \\ 1 & t - t_n \end{pmatrix} \quad \text{et} \quad Y = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Pour chaque t , le minimum du critère $C_t(a, b)$ est donné par (ce qui correspond donc aux estimateurs) :

$$\begin{pmatrix} \hat{a}(t) \\ \hat{b}(t) \end{pmatrix} = (X_t^T W_t X_t)^{-1} X_t^T W_t Y \quad (3.1)$$

Cette équation ressemble beaucoup à l'estimateur des moindres carrés standards (voir cours "R4.EMS.08 Modèle linéaire" au semestre 4 pour le parcours EMS), hormis l'indice t (indiquant que cela dépend de l'instant t qu'on considère), et hormis la matrice de poids W_t qui se positionne au milieu des deux produits matriciels (qui sert à modifier les calculs qui dépendent des données en fonction des poids).

L'estimateur final de la régression linéaire locale, pour un instant t est donné par : $\hat{f}(t) = \hat{a}(t)$ ce qui peut se réécrire de la manière suivante, en utilisant un vecteur $e_1 = (1, 0)^T$:

$$\hat{f}(t) = e_1^T (X_t^T W_t X_t)^{-1} X_t^T W_t Y.$$

Cette formule peut se calculer aisément pour toutes les valeurs de t , et voici avec la figure 3.7 ce que cela donne en pratique sur un exemple. On peut constater sur cet exemple, que même si le modèle et les calculs sont relatifs

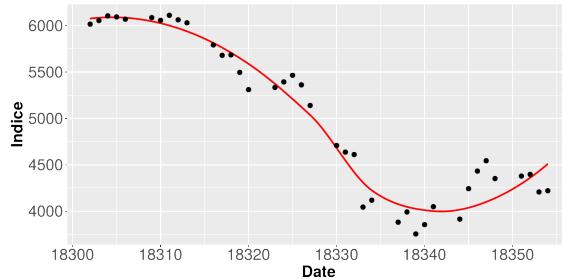


Figure 3.7 – Résultat de la procédure de lissage via régression locale sur les données du CAC40.

à des régressions linéaires, l'ensemble permet d'obtenir une estimation non-linéaire de f .

Mise en pratique des notions du chapitre 3

Exercice 3.3.1 (Moyenne mobile). Pour les données de taille de plante présentées en section 3.3, calculez sur papier puis sur ordinateur, ce que donne le lissage par moyenne mobile d'ordre $k = 2$. Expliquez pour cet exemple pourquoi l'échelle temporelle des données ne rend pas idéale l'utilisation de cette méthode telle quelle.

Exercice 3.3.2 (Lissage exponentiel). Sur ces mêmes données, calculez sur papier puis sur ordinateur, ce que donne le lissage exponentiel simple avec $\gamma = 0.1$ puis avec $\gamma = 0.9$.

Exercice 3.3.3 (Une seule régression locale). Sur ces mêmes données, calculez sur papier puis sur ordinateur, une seule régression locale en $t = 14:30:00$ du deuxième jour et avec une valeur $h = 20$. Calculez $\hat{a}(t)$ et $\hat{b}(t)$, puis tracez le nuage de points avec cette droite de régression (locale).

Exercice 3.3.4 (Moyenne mobile et fonction de poids). Reformuler l'équation d'une moyenne mobile à l'aide d'une fonction de poids que vous expliciterez.

Exercice 3.3.5 (Fonction de poids ?). Montrez si les fonctions suivantes sont bien des fonctions de poids :

- $W(u) = \frac{3}{4}(1 - u^2)\mathbf{1}\{|u| \leq 1\}$
- $W(u) = (1 - |u|)\mathbf{1}\{|u| \leq 1\}$
- $W(u) = \exp\{-u^3\}$

Exercice 3.3.6 (Estimateur local). Minimisez le critère des moindres carrés pondérés pour obtenir les estimateurs donnés par l'équation (3.1).

4

CHAPITRE

Tendance et périodicité

Dans le chapitre précédent, un des objectifs est de mettre en lumière la structure principale de la série chronologique (le signal) en la séparant de la partie chaotique (le bruit). Parmi les signaux possibles, il y en a plusieurs qui peuvent être importants à faire ressortir, voire même à les quantifier. Ces deux signaux sont la tendance et la périodicité, et l'idée principale de ce chapitre est de modéliser la série chronologique au travers de la décomposition suivante : données = tendance + bruit ou données = périodicité + bruit.

Définition 4.0.1 (Tendance) Correspond à l'évolution à long terme de la série chronologique.

Remarque 4.0.2 (Tendance linéaire). Bien que différentes formes de tendances existent, nous nous contenterons ici d'aborder le cas d'une tendance linéaire.

Définition 4.0.3 (Périodicité) Correspond à la répétition d'un schéma récurrent et oscillatoire de la série chronologique.

Remarque 4.0.4 (Saisonalité). Suivant le contexte, on peut parler de saisonnalité plutôt que de périodicité, si la période sur laquelle se répète un schéma correspond à une saison. Un exemple classique est l'évolution de la température qui répète un schéma d'une année sur l'autre : froid en hiver, hausse au printemps, chaleur en été, refroidissement en automne, et ainsi de suite.

Pour le formuler autrement, cela correspond à poser un modèle mathématique sur la manière dont on veut lisser (ou décomposer) la série chronologique. On peut donc considérer dans certains contextes qu'il peut être judicieux de lisser la série avec une structure de lissage prédéfinie. La figure 4.1 donne des exemples de séries chronologiques pour lesquelles on peut penser à décomposer la série avec une tendance linéaire (graphique de gauche) ou avec une périodicité (graphique de droite).

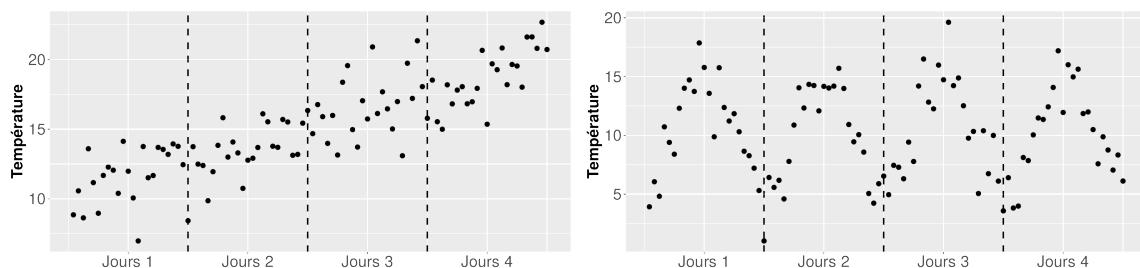


Figure 4.1 – Exemples de séries chronologiques admettant une tendance linéaire (graphique de gauche) ou une périodicité (graphique de droite).

Dans ce chapitre, l'objectif est de reformuler les notions de tendance et de périodicité en terme de modélisation et de pouvoir les estimer à partir de la série chronologique. Ce qui suit se décompose en deux sections, la section 4.1 qui traite du cas de la tendance et la section 4.2 du cas de la périodicité.

Table des matières de ce chapitre

4.1 Tendance linéaire	30
4.2 Périodicité	30

4.1 Tendance linéaire

Mettre en lumière une tendance linéaire dans les données revient à modéliser la série chronologique de la manière suivante :

$$x_i = a + bt_i + \varepsilon_i \quad (4.1)$$

où a et b sont des paramètres inconnus à déterminer et ε_i correspond à l'erreur théorique commise par ce modèle pour prédire la donnée x_i .

Le modèle présenté avec l'équation (4.1) ressemble énormément au modèle de régression linéaire simple. La seule différence fondamentale entre les deux modèles (et déjà énoncée dans le chapitre 1) est que la quantité t_i n'est pas la réalisation d'une variable aléatoire. Autrement dit, on dispose ici de données (t_i, x_i) qui ne sont relatives qu'à une source d'aléatoire (la variable aléatoire X), contrairement au contexte présenté pour des applications du modèle de régression linéaire simple pour lesquelles les données (x_i, y_i) peuvent être considérées comme issues de deux phénomènes aléatoires.

En mettant de côté cette considération abstraite, ce qui reste concernant ces approches (tendance linéaire et régression linéaire simple) est identique. A savoir, on a pour objectif ici d'obtenir une version empirique des quantités a et b qui sont inconnues et théoriques. Pour cela, on utilise les données de la série chronologique et la méthode des moindres carrés pour déterminer des estimations \hat{a} et \hat{b} . Voici ci-dessous leurs expressions respectives :

$$\hat{a} = \bar{x} - \hat{b}\bar{t} \quad \text{et} \quad \hat{b} = \frac{\bar{xt} - \bar{x}\bar{t}}{\bar{t^2} - \bar{t}^2}$$

La figure 4.2 donne un exemple de ce que cela peut donner en pratique. De plus, après avoir calculé les estimations $\hat{a} = 10.0893$ et $\hat{b} = 0.1044$, il est possible 1) d'indiquer que la série chronologique est à la hausse sur la période étudiée à un niveau d'intensité \hat{b} , et 2) de faire des prévisions à court terme concernant cette série. Par exemple, pour un instant $t = 108$ (correspond à midi du jour 5), on peut donner la prévision $\hat{x}_t = 21.36018$. De plus, on peut aussi calculer que pour que la série atteigne la valeur de 30, il faudra attendre le temps $t \approx 191$. Pour obtenir ce résultat, il suffit d'écrire :

$$30 = \hat{a} + \hat{b}t \iff t = \frac{30 - \hat{a}}{\hat{b}} \iff t = \frac{30 - 10.0893}{0.1044} = 190.7887.$$

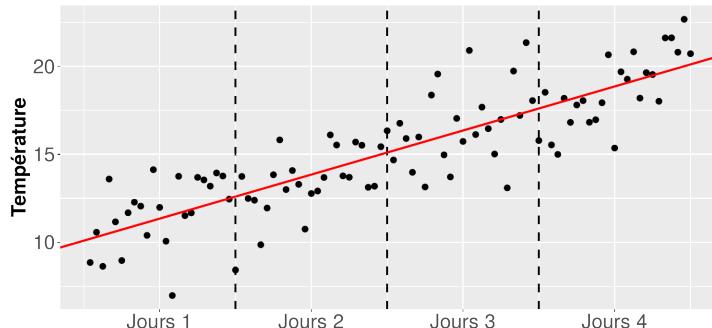


Figure 4.2 – Illustration de l'estimation d'une tendance linéaire sur une série chronologique.

Remarque 4.1.1 (Tendance non-linéaire). *Il peut aussi être souhaitable de déterminer une tendance qui ne soit pas linéaire afin d'appréhender l'évolution globale d'une série qui admet de grosses variations. Pour cela, une pratique courante consiste à utiliser une moyenne mobile sur la série (avec une ordre assez élevé) afin d'en extraire une tendance globale. Cela permet par exemple de modéliser la tendance d'une série qui ne soit pas monotone, et qui par exemple puisse être à la hausse puis à la baisse sur des périodes différentes. Cette approche est notamment utilisée dans le cadre de certaines modélisations présentées dans le chapitre 5.*

4.2 Périodicité

Pour étudier la périodicité d'une série chronologique, il est nécessaire de savoir a priori quelle est la longueur de la période en question. Pour l'exemple de la température, on se doute que la période est de un an, et dans d'autres

contextes, il faudra soit avoir une intuition valide concernant la longueur de cette période, soit de l'estimer par des méthodes annexes.

Définition 4.2.1 (Longueur de la période) La longueur de la période est la durée qu'on note p qui sépare de temps t et $t + p$ pour lesquels on s'attend à ce que les mesures de la série x_t et x_{t+p} soient comparables.

Remarque 4.2.2 (Longueur de la période et variable aléatoire). Pour formuler la définition 4.2.1 autrement, il faut déjà noter que pour deux mesures à des instants différents t_1 et t_2 , on peut considérer que les mesures x_{t_1} et x_{t_2} ne sont pas issues de la même variable aléatoire. Pour l'exemple de la température au fur et à mesure de l'année, si t_1 est un temps relatif à l'hiver et t_2 à l'été, on ne s'attend pas à ce que l'espérance des variables aléatoires X_{t_1} et X_{t_2} soient les mêmes, ainsi ce ne peuvent pas être les mêmes variables aléatoires.

Pour une série chronologique qui admet une périodicité, si on attend une certaine durée, alors on finira par obtenir des valeurs x_{t_1} et x_{t_2} qui sont comparables. Donc la longueur de la période p peut se voir comme la durée qui sépare deux variables aléatoires de lois identiques pour la série chronologique : $X_t = X_{t+p}$.

Si la série étudiée admet une périodicité, on peut décomposer les mesures de la série de la manière suivante :

$$x_i = S_i + \varepsilon_i \quad \text{où } S_i = S_{i+p} \text{ pour tout } i. \quad (4.2)$$

La valeur S_i peut se comprendre comme la norme de la série pour l'instant t_i .

Pour estimer les coefficients S_i , il suffit de moyenniser les mesures qui sont comparables à chacune des étapes de la période. Par exemple, si on dispose d'un cumul de pluie mensuel sur plusieurs années, et qu'on veut calculer le coefficient S_i (pour un i relatif à un mois de janvier) alors il faudra calculer la moyenne de tout les cumuls de pluie pour les mois de janvier. Et ainsi de suite pour tout les autres mois. Les estimations de ces coefficients sont donc données par :

$$\hat{S}_i = \frac{1}{N} \sum_{j=0}^{N-1} x_{i+jp} \quad (4.3)$$

où N est le nombre de périodes dans la série chronologique. Voici avec la figure 4.3 les résultats obtenus sur un exemple pratique.

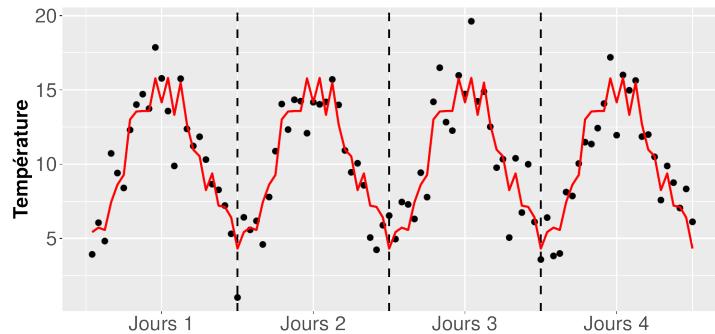


Figure 4.3 – Illustration de l'estimation d'une périodicité sur une série chronologique.

Mise en pratique des notions du chapitre 4

Exercice 4.2.1 (Tendance linéaire). A partir des données brutes (page suivante), calculez l'ajustement d'une tendance linéaire et retrouvez les résultats numériques présentés en section 4.1.

Exercice 4.2.2 (Périodicité). A partir des données brutes (page suivante) de l'exemple présenté en section 4.2, calculez l'estimation de la périodicité.

Exercice 4.2.3 (Tendance linéaire ou non-linéaire). Supposons qu'on ait besoin d'étudier une tendance et que plusieurs formes de tendance puissent être envisagées. Comment choisir parmi l'une d'entre elles ?

Exercice 4.2.4 (Périodicité et changements structurels). Ecrivez ce que vous pensez concernant le calcul d'une périodicité :

- si la périodicité réelle de la série chronologique n'est pas régulière, autrement dit que la durée de la période peut varier d'une période à une autre, ou
- s'il y a une rupture dans la série chronologique, et par exemple que la série qui oscille entre 10 et 15 entre le jour 1 et le jour 3, se retrouve à osciller entre 25 et 30 entre le jour 4 et le jour 7.

Exercice 4.2.5 (Périodicité comme un modèle linéaire). Supposons qu'on dispose d'une série chronologique (t_i, x_i) pour $i = 1, \dots, 10$ admettant une périodicité de longueur 4.

1. Montrez que le modèle (4.2) peut s'écrire de la forme suivante comme un modèle linéaire :

$$X = DS + \varepsilon, \text{ avec } X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{10} \end{pmatrix}, \quad S = \begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{10} \end{pmatrix}$$

et où D est une matrice que vous déterminerez. Conseil : cette matrice doit être remplie de 0 et de 1 et vous devrez déterminer correctement les dimensions de cette matrice de sorte à ce que l'équation ci-dessus soit valide.

2. Pour ce modèle-là, écrivez le critère des moindres carrés.
3. Minimiser le critère des moindres carrés et retrouvez l'expression de l'estimateur (4.3).

Exercice 4.2.1

t	x_t	t	x_t	t	x_t	t	x_t
1	8.85	25	13.74	49	14.68	73	18.52
2	10.57	26	12.49	50	16.76	74	15.53
3	8.63	27	12.39	51	15.90	75	14.99
4	13.59	28	9.86	52	13.98	76	18.18
5	11.16	29	11.94	53	15.98	77	16.81
6	8.96	30	13.84	54	13.14	78	17.80
7	11.67	31	15.82	55	18.37	79	18.05
8	12.28	32	12.99	56	19.56	80	16.82
9	12.05	33	14.08	57	14.97	81	16.96
10	10.39	34	13.29	58	13.71	82	17.93
11	14.12	35	10.75	59	17.04	83	20.66
12	11.98	36	12.77	60	15.73	84	15.35
13	10.06	37	12.91	61	20.90	85	19.69
14	6.97	38	13.68	62	16.12	86	19.27
15	13.75	39	16.10	63	17.68	87	20.83
16	11.51	40	15.53	64	16.46	88	18.19
17	11.67	41	13.77	65	15.01	89	19.64
18	13.69	42	13.69	66	16.98	90	19.53
19	13.54	43	15.69	67	13.09	91	18.01
20	13.19	44	15.51	68	19.73	92	21.62
21	13.94	45	13.12	69	17.21	93	21.62
22	13.76	46	13.19	70	21.35	94	20.80
23	12.45	47	15.43	71	18.05	95	22.67
24	8.42	48	16.34	72	15.78	96	20.72

Exercice 4.2.2

t	x_t	t	x_t	t	x_t	t	x_t
1	3.93	25	6.42	49	4.96	73	6.40
2	6.06	26	5.58	50	7.45	74	3.82
3	4.82	27	6.18	51	7.29	75	3.99
4	10.73	28	4.59	52	6.31	76	8.12
5	9.40	29	7.79	53	9.43	77	7.86
6	8.40	30	10.88	54	7.78	78	10.04
7	12.31	31	14.05	55	14.20	79	11.48
8	14.01	32	12.33	56	16.50	80	11.36
9	14.72	33	14.34	57	12.83	81	12.43
10	13.74	34	14.24	58	12.26	82	14.08
11	17.86	35	12.09	59	15.98	83	17.20
12	15.78	36	14.17	60	14.73	84	11.95
13	13.58	37	14.03	61	19.62	85	16.01
14	9.88	38	14.19	62	14.23	86	14.98
15	15.76	39	15.71	63	14.89	87	15.63
16	12.37	40	13.99	64	12.52	88	11.86
17	11.22	41	10.93	65	9.77	89	11.99
18	11.85	42	9.45	66	10.34	90	10.49
19	10.31	43	10.06	67	5.06	91	7.58
20	8.65	44	8.58	68	10.40	92	9.88
21	8.27	45	5.06	69	6.74	93	8.76
22	7.21	46	4.23	70	9.99	94	7.05
23	5.31	47	5.89	71	6.11	95	8.33
24	1.02	48	6.54	72	3.58	96	6.12

Modélisation

Les chapitres précédents contiennent différentes approches pour effectuer un traitement des séries chronologiques, consistant à la décomposer comme une partie décrite mathématiquement et une partie chaotique. Ce chapitre est basé sur l'idée que cette décomposition peut s'exprimer comme un modèle statistique, et qu'il est donc possible de ré-exprimer les différentes notions précédentes en des modèles.

Le fait de formuler ces notions comme des modèles permet de comprendre qu'il est possible d'effectuer les calculs suivants classiques dans le cadre d'une analyse basée sur une modélisation : calculer des prédictions, calculer des résidus, faire des diagnostiques d'adéquation du modèle aux données, ou encore faire un choix d'un des modèles possibles.

Dans la suite de ce chapitre, la section 5.1 introduit ce qu'il y a à savoir concernant le modèle additif, puis le modèle multiplicatif est présenté en section 5.2. Ensuite, la section 5.3 permet de comprendre ce qu'est un modèle autorégressif. Pour finir, la section 5.3 contient les diapos de cours et une feuille de TD, et celle-ci concerne aussi les chapitres 2, 3 et 4.

Table des matières de ce chapitre

5.1	Modèle additif	33
5.2	Modèle multiplicatif	34
5.3	Modèle autoregressif	35

5.1 Modèle additif

Dans le cadre de l'étude d'une série chronologique, le modèle additif est le modèle qui permet de décomposer la série avec simultanément les deux notions vues dans le chapitre 4 : une tendance et une périodicité. Ce modèle est donné par la formule suivante :

$$x_{t_i} = Z_i + S_i + \varepsilon_i, \text{ où } Z_i = a + bt_i \text{ et } S_i = S_{i+p} \text{ pour tout } i, \quad (5.1)$$

pour une période de durée p . Pour estimer les paramètres de cette modélisation, une possibilité assez simple consiste à procéder en trois étapes :

1. Ajuster une tendance linéaire (voir section 4.2.1) sur la série chronologique : calculer les estimations \hat{a} et \hat{b} .
2. Calculer les prédictions puis les résidus : $e_i = x_{t_i} - \hat{Z}_i$ où les prédictions sont $\hat{Z}_i = \hat{a} + \hat{b}t_i$.
3. Ajuster une périodicité (voir section 4.2.2) sur la série des résidus : calculer les estimations \hat{S}_i .

Les quantités obtenues ($\hat{a}, \hat{b}, \hat{S}_1, \dots, \hat{S}_p$) sont les estimations du modèle additif (5.1). La figure 5.1 montre l'ajustement de ce modèle sur une série chronologique. On peut constater que cela a permis d'obtenir une décomposition faisant apparaître une oscillation périodique de la série, qui évolue à la hausse au rythme de la tendance.

Bien que ce modèle contienne une composante linéaire et une composante périodique, il n'est pas automatique qu'il soit tout le temps nécessaire d'avoir ces deux composantes. Par exemple, il peut y avoir des situations pour lesquelles seule la composante linéaire est utile pour modéliser la série chronologique. Afin de détecter si une composante périodique est nécessaire, voici une procédure à suivre (et qui suit la procédure d'ajustement du modèle additif) :

1. Ajuster une tendance linéaire (voir section 4.2.1) sur la série chronologique : calculer les estimations \hat{a} et \hat{b} .
- 2.a. Calculer les prédictions puis les résidus : $e_i = x_{t_i} - \hat{Z}_i$ où les prédictions sont $\hat{Z}_i = \hat{a} + \hat{b}t_i$.
- 2.b. Tracer le nuage de points des résidus en fonction du temps. Si le nuage obtenu ressemble au graphique de gauche de la figure 5.2, alors il n'est pas nécessaire de rajouter une composante périodique (nuage de points homogène), mais s'il ressemble au graphique de droite, alors il faut réaliser l'étape suivante.

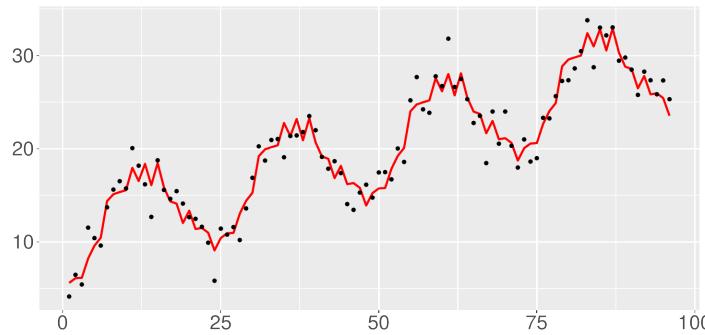


Figure 5.1 – Illustration de l'ajustement du modèle additif sur une série chronologique.

3. Ajuster une périodicité (voir section 4.2.2) sur la série des résidus : calculer les estimations \hat{S}_i .

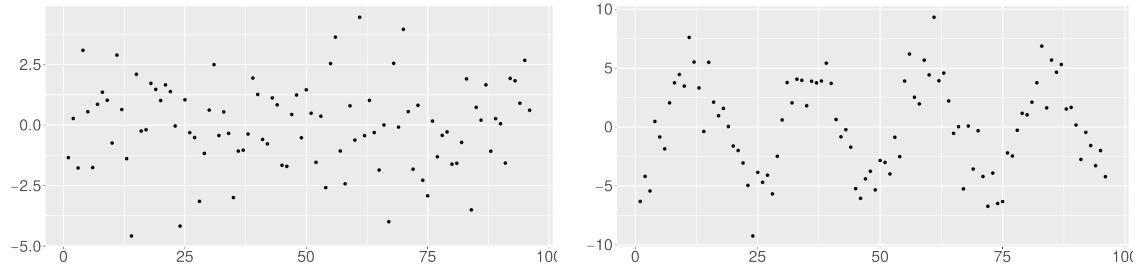


Figure 5.2 – Nuage de points des résidus suite à l'ajustement d'une tendance linéaire sur la série chronologique. Pour le graphique de gauche, le nuage des résidus est réparti de manière homogène sur le graphique, ce qui indique qu'il n'y a pas de structure périodique qu'il faudrait capturer par une modélisation de la périodicité. Pour le graphique de droite c'est le contraire puisqu'on détecte à l'œil nu une structure périodique, et cela indique qu'il faudrait ajouter une partie périodique à la modélisation.

Remarque 5.1.1 (Tendance non-linéaire). Comme indiqué dans la remarque 4.1.1, il est possible de modéliser la tendance de la série avec une moyenne mobile plutôt qu'une approche linéaire. D'un point de vue du modèle additif, cela revient à changer l'expression du terme Z_i en la résultante d'une moyenne mobile sur les x_{t_i} . Les changements dans la procédure d'estimation sont que :

1. dans l'étape 1. il ne faut pas ajuster une tendance linéaire mais il faut appliquer la méthode des moyennes mobiles sur les valeurs x_{t_i} , et
2. dans l'étape 2. pour le calcul des résidus il faut calculer $e_i = x_{t_i} - \hat{Z}_i$ avec \hat{Z}_i les valeurs résultantes du calcul de la moyenne mobile.

5.2 Modèle multiplicatif

Ce modèle est une variante du modèle additif permettant en plus de prendre en compte le fait que l'intensité de la périodicité évolue dans le temps. Avec les mêmes notations que pour le modèle 5.1, le modèle multiplicatif est donné par :

$$x_i = Z_i S_i \varepsilon_i, \text{ où } Z_i \text{ est une tendance, et } S_i = S_{i+p} \text{ pour tout } i. \quad (5.2)$$

Le lien entre les deux modèles est qu'il est possible de passer de l'un à l'autre avec une transformation log :

$$x_{t_i} = Z_i S_i \varepsilon_i \iff \ln(x_{t_i}) = \ln(Z_i) + \ln(S_i) + \ln(\varepsilon_i).$$

Pour estimer les coefficients de cette modélisation, voici la procédure à suivre pour une série dont la longueur de la période est p :

1. Calculer la moyenne mobile d'ordre p :

$$\hat{Z}_i = \sum_{k=-p/2}^{p/2} w_k x_{i+k},$$

où les w_k sont des poids de la moyennisation.

2. Calculer les quantités suivantes qui s'apparentent à des résidus de cette moyenne mobile : $r_i = x_i / \hat{Z}_i$.
3. Ajuster une périodicité \hat{S}_i sur la série des r_i comme indiqué dans la section 4.2.

Dans ce cas, on obtient les résidus de la manière suivante :

$$e_i = \frac{x_i}{\hat{Z}_i \hat{S}_i}.$$

Les résultats qu'on peut obtenir avec ce modèle sont illustrés par la figure 5.3, et ce graphique illustre une situation où le modèle multiplicatif est adapté au données.

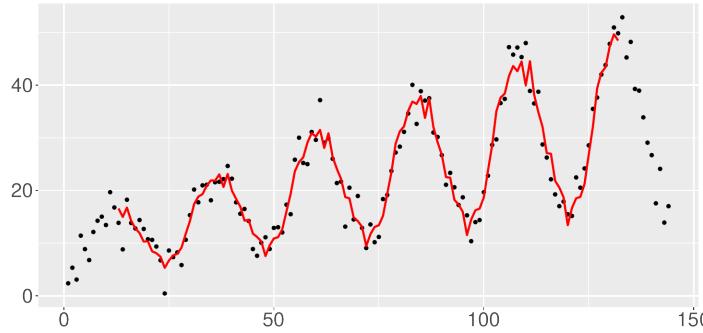


Figure 5.3 – Illustration de l'ajustement du modèle additif sur une série chronologique.

Remarque 5.2.1 (Condition d'utilisation). *Cette modélisation ne s'applique que sur des séries chronologiques n'ayant pas de valeur $x_i = 0$ ou même $x_i \approx 0$.*

5.3 Modèle autoregressif

Après avoir vu dans les sections 5.1 et 5.2 des modélisations de série chronologique pour lesquels la mesure x_t ne s'explique que par le temps t , cette section introduit une classe de modèles qui cherchent à expliquer x_t comme une conséquence des mesures précédentes : $x_{t-1}, x_{t-2}, \dots, x_{t-p}$. Le modèle présenté ici est le plus simple d'entre eux : le modèle autoregressif. Le terme de autorégressif est à comprendre ici comme

”régressif” : faire un modèle de régression, et

”auto-” : sur soi-même.

Autrement dit, il s'agit de faire une régression pour trouver une liaison linéaire entre la mesure x_t et le passé de la même série, mais avant t . On note ce modèle AR(p), où p est le nombre d'instants du passé pris en compte dans le modèle.

Pour commencer simplement, voici l'équation qui décrit le modèle AR(1), où 1 signifie que pour prédire la mesure x_t on inclut dans le modèle une seule valeur concernant le passé, à savoir la valeur à l'instant précédent :

$$x_t = a + bx_{t-1} + \varepsilon_t \quad (5.3)$$

Remarque 5.3.1 (Notation sans les i). *Pour ce type de modèles, on ne note pas x_{t_i} en faisant intervenir l'indice i , parce qu'il s'agit de modèles qu'on utilise préférentiellement sur des séries chronologiques pour lesquelles les temps d'échantillonnage sont réguliers. Autrement dit, la durée séparant t_i et t_{i+1} est la même quelque soit i , voir la définition 5.3.2. On s'autorise alors à ne plus écrire t_i et t_{i+1} puisque cela n'induit pas d'ambiguïté, en préférant écrire t et $t+1$, de sorte à simplifier les notations.*

Définition 5.3.2 (Série chronologique régulière) *Une série chronologique est dite régulière si les temps de mesures t_i sont également espacés (si ce n'est pas le cas elle est dite irrégulière).*

Remarque 5.3.3 (Série irrégulière et le modèle AR). *De plus, il est aussi à noter que si les temps d'échantillonnage n'étaient pas réguliers, alors ce type de modèles ne serait pas pertinent. En effet, ce modèle consiste à considérer que la liaison entre deux instants consécutifs est la même (et est linéaire) quelques soit les deux instants consécutifs considérés. Mais s'il y a des irrégularités dans les temps d'échantillonnage, cela signifie qu'on pourrait avoir que la durée entre t_i et t_{i+1} vaudrait par exemple 2 et que pour un autre instant t_j la durée jusqu'à t_{j+1} est différente de 2, et disons qu'elle soit égale à 10. Alors dans ce cas, il ne serait pas cohérent de considérer que la liaison entre $x_{t_{i+1}}$ et x_{t_i} serait la même que la liaison entre $x_{t_{j+1}}$ et x_{t_j} . Etant donné que la durée entre ces deux*

instants est 5 fois plus long, on peut s'attendre à ce que le phénomène mesuré ait potentiellement évolué avec une plus grande amplitude entre $x_{t_{j+1}}$ et x_{t_j} que entre $x_{t_{i+1}}$ et x_{t_i} . S'il est tout de même souhaité d'ajuster un modèle autorégressif sur une série chronologique irrégulière, il sera nécessaire de faire une interpolation de la série de sorte à obtenir une série chronologique régulière.

Pour estimer les paramètres de ce modèle, on peut se ramener à une expression générique d'un modèle de régression linéaire en posant les notations suivantes :

- on note les données à prédire : $y_i = x_{t_{i+1}}$, et

- qui est à prédire à l'aide de : $x_i = x_{t_i}$,

et cela permet de réécrire le modèle 5.3 comme :

$$y_i = a + bx_i + \varepsilon_i, \text{ pour } i = 1, \dots, n - 1.$$

De cela, on en déduit les estimations des paramètres comme :

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad \text{et} \quad \hat{b} = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2}.$$

Voici ci-dessous avec la figure 5.4 ce que ce modèle donne dans un cas pratique : le graphique de gauche correspond à l'ajustement du nuage de points (x_i, y_i) et le graphique de droite illustre ce que cela donne du point de vue de la série chronologique.

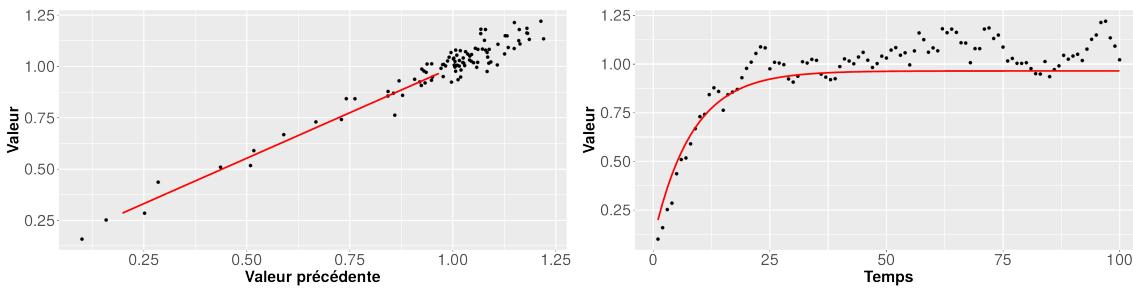


Figure 5.4 – Illustration de l'ajustement du modèle AR(1) sur une série chronologique. A gauche, ajustement du nuage de point des incrémentés (x_i, y_i) , et à droite ce que cela donne sur la série chronologique.

Un modèle plus général consiste à ne pas prendre en compte seulement la précédente valeur, mais de prendre les p valeurs précédentes, il s'agit du modèle AR(p) :

$$x_t = a + b_1x_{t-1} + b_2x_{t-2} + \dots + b_px_{t-p} + \varepsilon_t = a + \sum_{j=1}^p b_jx_{t-j} + \varepsilon_t \quad (5.4)$$

où les paramètres à estimer sont a, b_1, \dots, b_p . Voici ci-dessous avec la figure 5.4 ce que ce modèle donne dans un cas pratique du point de vue de la série chronologique.

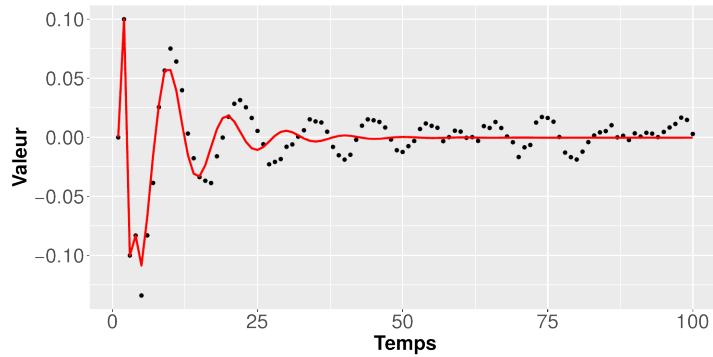


Figure 5.5 – Illustration de l'ajustement du modèle AR(p) sur une série chronologique.

Concernant ce modèle des aspects importants ne sont pas traités dans ce cours, et seront traités dans le cadre de la ressource pédagogique "R4.EMS.08 Modèle linéaire" (parcours EMS) qui aborde tout ce qu'il y a à savoir concernant les modèles linéaires avec plusieurs régresseurs (ce qui est justement le cas du modèle (5.4)). Les aspects en question sont :

1. Comment se ramener à l'expression d'un modèle de régression simple ?
2. Comment calculer les estimations des paramètres ?
3. Comment choisir une valeur de p dans un contexte pratique ?

Pour ce qui est de la question 3., il est important de noter qu'il s'agit d'un type de question qui revient dans différents cas de modélisation : il s'agit de choisir un modèle parmi plusieurs modèles possibles. En effet, lorsqu'on étudie une série chronologique et qu'on souhaite utiliser un modèle autorégressif, on ne dispose a priori d'aucune préférence concernant les modèles AR(1), AR(2), AR(3) et ainsi de suite. Or il convient d'en choisir un, et ce choix peut avoir des conséquences importantes sur les résultats de la modélisation de la série. L'idée principale reste la même que celle vue précédemment dans le cours "R2.21 Statistique descriptive 2" : il faut choisir le modèle dont les prédictions sont les plus proches des données observées.

Mais en attendant le prochain semestre d'en savoir plus, voici une intuition pratique à utiliser pour guider le choix de p lors de l'utilisation d'un modèle autorégressif. Cette approche repose sur l'étude de l'autocorrélation de la série chronologique.

Définition 5.3.4 (Autocovariance empirique) Pour une série chronologique, l'autocovariance empirique K évaluée en une durée h est la covariance empirique entre la série chronologique (en un temps t) avec elle-même (en un temps $t - h$) :

$$K(h) = \text{cov}(x_t, x_{t-h}) = \frac{1}{n-h} \sum_{t=h+1}^n (x_t - \bar{x})(x_{t-h} - \bar{x}),$$

où \bar{x} est la moyenne des valeurs de la série chronologique.

Définition 5.3.5 (Autocorrélation empirique) Pour une série chronologique, l'autocorrélation empirique ρ évaluée en une durée h est la corrélation empirique entre la série chronologique (en un temps t) avec elle-même (en un temps $t - h$) :

$$\rho(h) = \text{cor}(x_t, x_{t-h}) = \frac{K(h)}{\hat{\sigma}^2},$$

s où $\hat{\sigma}^2$ est la variance empirique des valeurs de la série chronologique.

Remarque 5.3.6 (Définition générale). Ces définitions ne sont qu'un cas particulier de ce que sont l'autocovariance et l'autocorrélation (empiriques). Ce cours aborde ces notions sous cet angle-là afin de l'appréhender plus simplement.

Remarque 5.3.7 (Version empirique). Ces définitions correspondent aux versions empiriques (calculables en pratique à partir des données) associées à des versions théoriques, qui ne sont pas spécifiées dans le cadre de ce cours.

Remarque 5.3.8 (Autocorrélation pour $h = 0$). Comme on a nécessairement que $\text{cor}(x_t, x_t) = 1$, alors on a que $\rho(0) = 1$, ce qui permet de remarquer que $K(0) = \hat{\sigma}^2$.

La figure 5.6 illustre ce à quoi ressemble la fonction d'autocorrélation empirique. Sur cet exemple, on constate que pour des valeurs de h proches de 0, plus l'autocorrélation est proche de 1. De plus, elle diminue en oscillant autour de 0 lorsque h augmente. L'interprétation est que pour deux instants proches, la corrélation entre les deux mesures est élevée et la donnée x_t est très liée à la donnée x_{t-h} . A l'opposé, si les instants de mesures sont éloignés l'un de l'autre, la corrélation est proche de 0 ce qui signifie qu'il n'y a pas de lien entre x_t et x_{t-h} . La mesure x_t ne dépend plus de x_{t-h} parce que trop de temps entre les deux s'est écoulé et d'autres phénomènes plus récents ont un impact plus important sur la mesure x_t . En utilisant ce graphique, on peut constater et quantifier que (prises

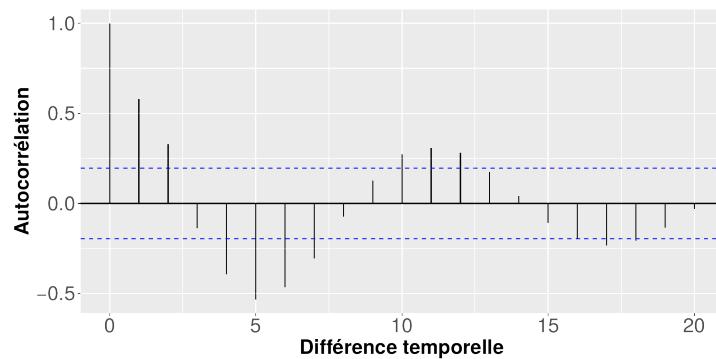


Figure 5.6 – Illustration de la fonction d'autocorrélation empirique sur une série chronologique.

marginalement) chacune des mesures x_{t-h} est plus ou moins liée linéairement à x_t . On peut donc en déduire qu'à partir d'une certaine valeur h_0 , il n'y a plus de liens entre x_{t-h_0} et x_t . Et donc on pourra choisir de modéliser la série chronologique avec un modèle AR(p) avec $p = h_0 - 1$. En procédant ainsi, on s'assure de ne garder que la partie du passé de x_t qui paraît pertinente pour le modèle AR(p), à savoir qui admet une corrélation linéaire importante avec x_t . Pour faire cela, il n'y a pas de seuil optimal en dessous duquel l'autocorrélation serait trop faible et on pourra fixer arbitrairement une valeur entre 0.5 et 0.2.

Diapos de cours et feuille de TD

Chap. 2 – Modélisation de séries chronologiques

Contexte

Contexte :

On dispose d'une suite d'observations numériques d'un même phénomène, mais observé à plusieurs instants.

Par exemple :

- l'évolution quotidienne de l'indice du CAC40 (bourse française)
- l'évolution quotidienne du nombre de cas de Covid
- l'évolution du nombre de connexions à un serveur toutes les heures
- ...

Problématiques :

A partir de ces données, plusieurs questions peuvent émerger :

- Déterminer des tendances d'évolution derrière la cinétique chaotique
- Prédire la valeur du lendemain
- Détecter une anomalies (**Chapitre 3**)
- Détecter un changement abrupte dans la cinétique des données (**Chapitre 3**)

Série chronologique

Définition :

Soient t_1, \dots, t_n une suite d'instants pendant lesquels le phénomène d'intérêt est observé/mesuré. On note y_i la mesure au temps t_i . La série chronologique (ou série temporelle) est l'ensemble de ces données qu'on peut noter de manière synthétique : $y = (y_1, \dots, y_n) = (y_i)_{i=1}^n$. On parlera parfois du couple (t_i, y_i) .

i	t_i	y_i
1	2020-02-10	6015.67
2	2020-02-11	6054.76
3	2020-02-12	6104.73
4	2020-02-13	6093.14
5	2020-02-14	6069.35
6	2020-02-15	
7	2020-02-16	
8	2020-02-17	6085.95
9	2020-02-18	6056.82
10	2020-02-19	6111.24
⋮	⋮	⋮

À la différence d'un couple de mesure (x_i, y_i) comme au chapitre précédent, la mesure d'une série chronologique (t_i, y_i) n'est relative qu'à une seule variable aléatoire. Dans ce cadre, le temps n'est pas une variable aléatoire qu'on échantillonnera/mesure.

Pré-traitements

Série brute :

La série brute est la série chronologique (t_i, y_i) à laquelle on n'effectue aucune modification.

Transformation :

Dans certain contexte, il sera plus pertinent d'étudier la série chronologique (t_i, z_i) où $z_i = f(y_i)$. La transformation la plus souvent utile est $f(y) = \log(y)$ mais d'autres transformations sont possibles. L'objectif de ces transformations est de ramener l'évolution de la série chronologique à une série ayant une forme linéaire.

Série relative :

Il peut aussi être pertinent de ne pas étudier la série brute mais plutôt d'étudier la **série des différences** ou les **taux de croissance**. La série des différences est la série des mesures $z_i = y_i - y_{i-1}$ pour $i > 1$. Les taux de croissance correspondent à la série des mesures $z_i = y_i/y_{i-1} - 1$ pour $i > 1$, à utiliser si la série comporte uniquement des mesures strictement positives.

Moyenne mobile

Série et comportement chaotique :

Une série chronologique a très souvent un comportement chaotique/erratique. Les mesures semblent évoluer de manière imprévisible, voir par exemple l'indice du CAC 40.

L'objectif est de discerner dans une série chaotique :

1. le signal (une partie déterministe qu'on va modéliser)
2. le chaos (le résidus de la modélisation qui restera aléatoire et imprévisible).

Modèle de tendance par moyenne mobile

Pour une série chronologique (t_i, y_i) pour $i = 1, \dots, n$, la moyenne mobile d'ordre impair $N = 2k + 1$ est une série (t_i, m_i) pour $i = k + 1, \dots, n - k$ donnée par :

$$\begin{aligned} m_i &= \frac{1}{m} (y_{i-k} + \dots + y_i + \dots + y_{i+k}) \\ &= \frac{1}{2k+1} \sum_{j=-k}^k y_{i+j}. \end{aligned}$$

Cette méthode permet de "lisser" les données et de voir se dégager une tendance dans l'évolution de la série.

On obtient une **série lissée**.

Lissage exponentiel simple

Définition :

La série lissée obtenue par lissage exponentielle simple est notée $z = (z_1, \dots, z_n)$ et est donnée de manière récursive par : $z_i = \gamma y_i + (1 - \gamma)z_{i-1}$, où $\gamma \in [0, 1]$.

Interprétations :

- La première valeur de la série (z_1) doit être manuellement fixée. Par défaut, on peut prendre $z_1 = \frac{1}{3}(y_1 + y_2 + y_3)$.
- Cette équation donne que z_i est une moyenne (pondérée) entre la valeur y_i de la série brute et z_{i-1} , la série lissée au temps d'avant.
- Le paramètre γ est le poids accordé à la série brute et s'interprète comme le paramètre de lissage :
 - plus γ est proche de 0, plus la série est lisse, et
 - plus γ est proche de 1, plus la série lissée suit la série brute.

Pré-traitements

Série brute :

La série brute est la série chronologique (t_i, y_i) à laquelle on n'effectue aucune modification.

Prédictions :

Dans le cadre d'un modèle additif, on souhaite prévoir l'évolution de la série chronologique, pour des temps t_{n+1}, \dots, t_{n+k} . Pour cela, on va utiliser la formule suivante :

$$\hat{y}_{n+k} = \hat{a} + \hat{b}t_{n+k} + \hat{S}_k.$$

.

.

.

.

.

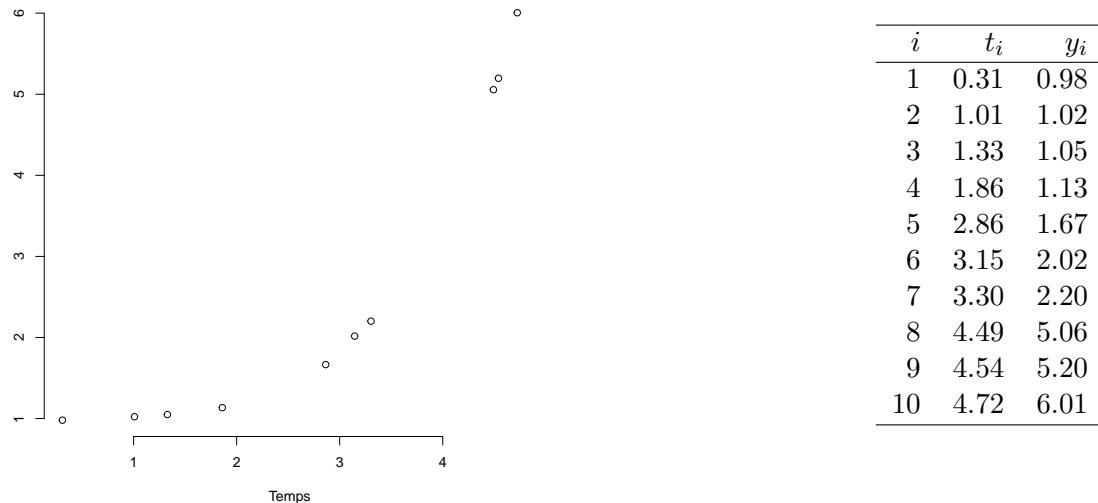
.

Résumé

- Qu'est-ce qu'une série chronologique
Série brute, transformation
- Lissages
Moyenne mobile, lissages exponentiels
- Tendance et saisonnalité
- Modèle additif
Interprétation et méthode d'ajustement
- Modèle multiplicatif : interprétation
- Comment calculer une prévision

TD 1 – Prétraitemenet et lissage

Exercice 1. Soit la série chronologique (t_i, y_i) , pour $i = 1, \dots, 10$, donnée par le tableau suivant et représentée dans le graphique suivant :



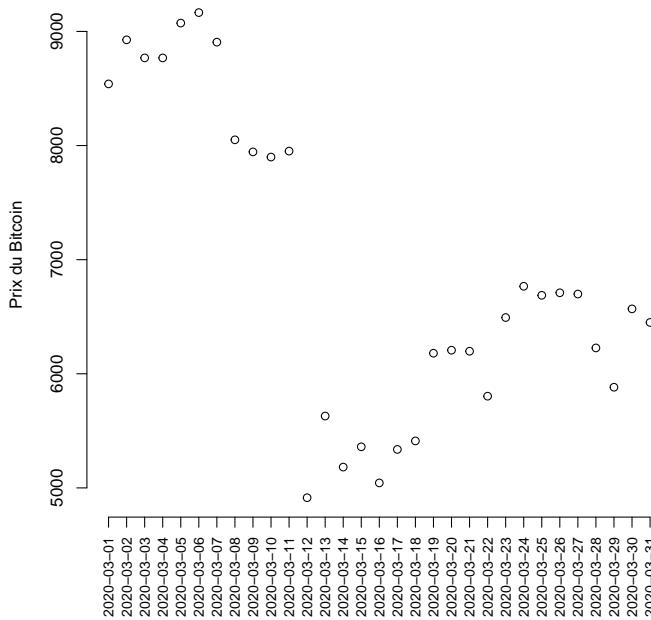
- Calculez les séries transformées $z_{1,i} = \log(y_i)$, $z_{2,i} = \sqrt{y_i}$ et $z_{3,i} = y_i^{\frac{1}{4}}$.
- Pour chacune des séries (t_i, y_i) , $(t_i, z_{1,i})$, $(t_i, z_{2,i})$ et $(t_i, z_{3,i})$, calculez la tendance linéaire de la série.
- Calculez les prédictions de chacune des séries : \hat{y}_i , $\hat{z}_{1,i}$, $\hat{z}_{2,i}$ et $\hat{z}_{3,i}$.
- Calculez les résidus pour chacune de ces séries. Pour rappel, les résidus pour des séries transformées par une transformation f sont : $e_i = y_i - f^{-1}(\hat{z}_i)$. La fonction f^{-1} est la fonction réciproque de f (la réciproque de $\log(x)$ est $\exp(x)$ et la réciproque de x^n est $x^{\frac{1}{n}}$).
- Calculez le MSE pour chacune de ces tendances linéaires.
- En déduire quelle transformation est la plus adaptée pour transformer le nuage de points initial sous une forme linéaire.

Exercice 2. Le prix quotidien du Bitcoin (en dollars) pour le mois de Mars 2020 est donné dans les tableaux suivants :

Jour	Prix	Jour	Prix	Jour	Prix
2020-03-01	8540.26	2020-03-11	7951.17	2020-03-21	6197.62
2020-03-02	8926.25	2020-03-12	4914.09	2020-03-22	5803.67
2020-03-03	8768.18	2020-03-13	5629.58	2020-03-23	6492.94
2020-03-04	8767.89	2020-03-14	5182.32	2020-03-24	6766.78
2020-03-05	9073.27	2020-03-15	5360.06	2020-03-25	6687.70
2020-03-06	9165.15	2020-03-16	5043.68	2020-03-26	6709.95
2020-03-07	8905.95	2020-03-17	5337.09	2020-03-27	6698.73
2020-03-08	8050.19	2020-03-18	5411.62	2020-03-28	6226.55
2020-03-09	7944.45	2020-03-19	6180.27	2020-03-29	5882.01
2020-03-10	7899.23	2020-03-20	6206.61	2020-03-30	6568.81
					2020-03-31 6449.95

- Calculez la série des différences et celle des différences relatives.
- Calculez la moyenne mobile d'ordre 3.
- Calculez la série lissée par lissage exponentiel simple, avec $\gamma = 0.3$.

4. Sur le graphique suivant, tracez la moyenne mobile en couleur et la série lissée d'une couleur différente.



Exercice 3. Dans les tableaux suivants sont données les prix du Bitcoin en dollars du 12 Mars 2020 au 07 Avril 2020. Un autre tableau donne des résumés statistiques des données.

Jour	Prix	Jour	Prix	Jour	Prix
2020-03-12	4914.09	2020-03-21	6197.62	2020-03-30	6568.81
2020-03-13	5629.58	2020-03-22	5803.67	2020-03-31	6449.95
2020-03-14	5182.32	2020-03-23	6492.94	2020-04-01	6671.95
2020-03-15	5360.06	2020-03-24	6766.78	2020-04-02	6833.05
2020-03-16	5043.68	2020-03-25	6687.70	2020-04-03	6740.07
2020-03-17	5337.09	2020-03-26	6709.95	2020-04-04	6872.91
2020-03-18	5411.62	2020-03-27	6698.73	2020-04-05	6778.61
2020-03-19	6180.27	2020-03-28	6226.55	2020-04-06	7297.75
2020-03-20	6206.61	2020-03-29	5882.01	2020-04-07	7360.37

$$\begin{array}{cccc} \bar{t} & \bar{y} & \bar{t^2} & \bar{yt} \\ 14 & 6233.509 & 256.6667 & 91874.87 \end{array}$$

- Calculez la tendance linéaire de cette série chronologique.
- Calculez les prédictions \hat{y}_i à partir de la tendance linéaire.
- Déterminez e_i les résidus de cette tendance.
- A partir des résidus, calculez la périodicité de 7 jours de cette série chronologique.
- Donnez l'équation du modèle additif ajusté avec la tendance et la périodicité (estimées).
- Tracez le nuage de points ainsi que la courbe relative au modèle additif (tendance + périodicité).
- A l'aide du modèle additif ajusté, calculez les prédictions pour les 3 prochains jours.

6

CHAPITRE

Anomalies et cassures

Lorsqu'on étudie une série chronologique, il peut y avoir des caractéristiques marquantes qu'il faut réussir à détecter correctement. Dans le cadre de ce cours, il est question de deux d'entre elles, qui sont d'ailleurs des phénomènes importants pour ce qui est de certaines problématiques autour de la cybersécurité : les anomalies et les cassures. Mais cela correspond aussi à un enjeu dans d'autres domaines d'application, avec par exemple la détection de trouble cardiaque à partir de l'étude d'un électrocardiogramme, la détection de changements dans l'évolution de certains indicateurs économiques, ou encore l'apparition de phénomènes climatiques rares.

Une anomalie correspond à une notion déjà vue au semestre 1 dans la ressource pédagogique "R1.21 Statistique descriptive 1", à savoir il s'agit d'une donnée anormale. Celle-ci correspond à une donnée atypique, mais n'est pas une donnée aberrante. Autrement dit, il ne s'agit pas d'une donnée dont la valeur potentiellement extrême serait issue d'une erreur dans l'acquisition des données. Cette donnée (anomalie) est une donnée qui n'est non seulement pas forcément extrême, mais qui en plus, correspond à une variation très particulière du phénomène observé.

Exemple 6.0.1. *Pour se donner un exemple, considérons une entreprise qui conserve l'historique des connexions extérieures à son service informatique (ce qui donne bien une série chronologique), et si elle constate qu'à un moment donné il y a eu un pic important de connexions (anomalie), cela peut alerter l'entreprise : c'est peut-être une attaque numérique contre l'entreprise.*

Pour cet exemple, on comprendra qu'il s'agit d'une anomalie, si le pic en question correspond à une quantité qui n'est normalement pas observée pour le créneau horaire en question. Autrement dit, c'est une anomalie si elle s'écarte trop de la valeur normale (où cette normalité doit être déterminée/estimée à l'aide d'un modèle).

Le second phénomène est une cassure dans la série chronologique ce qui se manifeste par un changement temporaire ou permanent dans une des caractéristiques de la série chronologique (de la valeur moyenne, de la tendance, ou autre). La figure 6.1 illustre cette notion, en montrant deux types de cassure, l'une en terme de valeur moyenne de la série, et puis la seconde dans la tendance linéaire.

Exemple 6.0.2. *Pour illustrer ce phénomène, on peut prendre l'exemple d'une banque qui fait un suivi des transactions effectuées sur les comptes. Une cassure dans la série pourrait correspondre à une situation où le compte d'un des clients de la banque serait compromis (par le vol de la carte bancaire ou si un attaquant a réussi à s'authentifier numériquement), et que la personne ayant accès au compte utiliserait le compte d'une manière inhabituelle par rapport à l'habitude du client.*

On peut voir avec cet exemple que pour caractériser une cassure dans une série chronologique, il faut pouvoir comparer des caractéristiques importantes sur des périodes différentes, et trouver un instant (la cassure) séparant deux périodes pour lesquelles on constate des différences majeures concernant la caractéristique concernée.

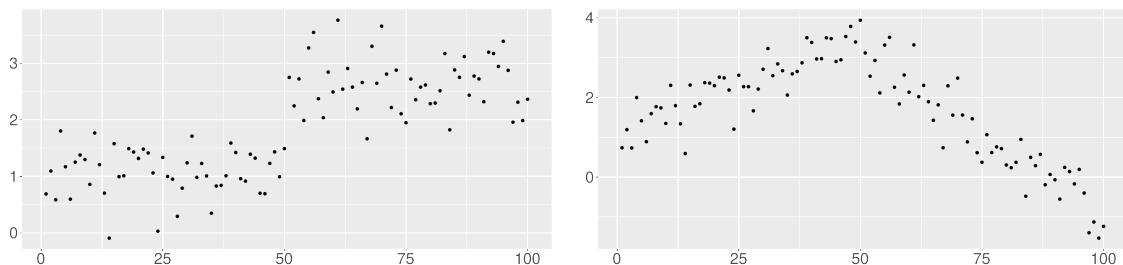


Figure 6.1 – Différentes cassures dans une tendance linéaire.

D'autres méthodes existent et sont principalement basées sur des idées dont les bases pédagogiques sont présentées dans les ressources pédagogiques "R4.21 Méthodes factorielles" (méthode de réduction de dimension) et

"R4.25 Classification automatique" (classification) au semestre 4, "R5.25 Data mining" (classification) au semestre 5 ou encore "R6.EMS.51 Apprentissage statistique pour l'IA" (deep learning) au semestre 6 . Mais quelque soit les méthodes utilisées, elles reposent toutes sur le principe général : 1) chercher à modéliser une normalité de la série chronologique et 2) en déduire un critère évaluant un écartement trop important à la norme établie (ou autrement dit définir un score d'anomalie).

Dans le reste de ce chapitre, il est question de présenter les différentes méthodes pour détecter les anomalies (section 6.1), en montrant comment les méthodes de modélisation et de lissage peuvent être utilisées, et pour détecter une cassure (section 6.2) en formulant cela sous le prisme d'un test d'hypothèses. Pour finir, la section 6.2 contient les diapos de cours et une feuille de TD.

Table des matières de ce chapitre

6.1	Détection d'anomalies	46
6.2	Détection de cassures	46

6.1 Détection d'anomalies

Afin de détecter des anomalies dans une série chronologique, il faut deux ingrédients : un modèle et déterminer quand un écartement est trop important. Pour le premier, les possibilités sont nombreuses, et toutes les approches de modélisation sont envisageables (modèles additif, multiplicatif et autorégressif), mais aussi les méthodes de lissage (moyenne mobile, lissage exponentiel et régression locale). L'idée est de pouvoir obtenir une écriture de la série chronologique comme suit :

$$x_t = f_t + \varepsilon_t$$

où f_t représente la norme établie par le modèle (et peut dépendre de paramètres ou des autres valeurs de la série), et ε_t représente l'erreur du modèle au temps t .

Après avoir ajusté un modèle, il est nécessaire d'obtenir les résidus $e_t = x_t - \hat{f}_t$, où \hat{f}_t sont les prédictions du modèle. A partir de ces résidus, on cherche à déterminer lesquels sont trop grands (en positif ou en négatif). En effet, si un résidu est proche de 0, alors cela indique que la donnée x_t n'est pas très différente de la norme établie en t : à savoir \hat{f}_t . Mais si le résidu prend une grande valeur (positive ou négative), cela signifie que l'écart entre x_t et la prédition du modèle est très importante, ce qui peut indiquer qu'il s'agisse d'une anomalie. Cependant, puisque les données sont considérées comme des variations imprévisibles autour de la véritable mesure du phénomène, il n'est pas impossible d'obtenir de temps en temps des résidus avec de grandes valeurs (sans que la donnée en question soit une anomalie). Cela ne correspond alors pas à une anomalie, mais juste à une donnée ayant une variation individuelle importante. Si on prétendrait que cette donnée est une anomalie, on commetttrait alors une erreur qu'on peut appeler un faux-positif : on l'a considéré à tort comme une anomalie. Afin de tenter d'éviter ce type de faux-positifs, on peut procéder de la manière suivante :

1. On se fixe un taux d'anomalies α , qui généralement correspond à un faible pourcentage, comme 5% ou 1%.
2. On détermine les quantiles empiriques des résidus aux niveaux $\alpha/2$ et $1 - \alpha/2$: $\hat{q}_{\alpha/2}$ et $\hat{q}_{1-\alpha/2}$.
3. On sélectionne les données x_t pour lesquelles les résidus e_t vérifient : soit $e_t < \hat{q}_{\alpha/2}$, soit $e_t > \hat{q}_{1-\alpha/2}$.

Les données ainsi sélectionnées sont celles qu'on considère comme étant des anomalies. En procédant de cette manière, avec par exemple $\alpha = 1\%$, on détermine un intervalle $[\hat{q}_{0.5\%}, \hat{q}_{99.5\%}]$ qui par définition de ce qu'est un quantile, est censé contenir une proportion de 99% des résidus (si l'hypothèse de normalité des résidus semble vérifiée). Et autrement dit, les résidus qui ne sont pas dans cet intervalle (ce qui correspond à l'étape 3 ci-dessus), sont alors dans le "top 1%" des résidus les plus grands.

Pour illustrer cette procédure, la figure 6.2 donne une représentation des résultats de cette méthode (la partie modélisation est une régression locale). Le graphique de gauche montre le nuage de points des résidus, avec les seuils définis par les quantiles empiriques, avec un taux $\alpha = 1\%$, et le graphique de droite montre le résultat sur le nuage de points de la série chronologique. Comme l'illustre le graphique de droite, procéder ainsi peut aussi se voir comme le fait de définir un intervalle de confiance à $1 - \alpha\%$ autour de la courbe du modèle ajusté. C'est souvent plus parlant de présenter le graphique de droite, puisque celui-ci correspond à des valeurs de la série chronologique, ce qui est plus "concret" que la valeur d'un résidu.

6.2 Détection de cassures

Pour ce qui est de la détection de cassure, nous abordons dans cette section le test de Chow. Tel qu'il est présenté ci-dessous, celui-ci est conçu pour trouver une cassure dans la tendance linéaire de la série, mais il est

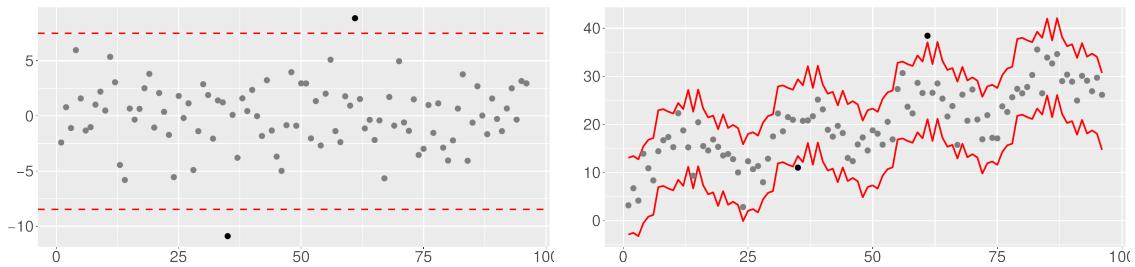


Figure 6.2 – Illustration de la méthode de détection des anomalies par la méthodes des quantiles. Le graphique de gauche montre le nuage des résidus avec les seuils déterminés avec les quantiles empiriques à 0.5% et 99.5%. Le graphique de droite représente la même information mais lorsqu'elle est transposée sur la série chronologique.

possible de reformuler ce test pour tester la présence d'une cassure concernant un autre aspect autour de la modélisation de la série (comme la périodicité).

Le test de Chow est un test statistique basé sur le fait de découper la série chronologique et de modéliser la tendance linéaire sur chacune des sous-parties obtenues. Cela fait intervenir plusieurs modélisations de cette tendance linéaire, qui sont mises alors en compétition avec un test d'hypothèses afin d'évaluer la vraisemblance de la présence d'une cassure. Le premier modèle d'entre eux est le modèle déjà connu de la tendance linéaire, qu'on appellera modèle complet et qu'on notera M_c :

$$M_c : \quad y_i = a + b \times t_i + \varepsilon_i.$$

Les autres modèles qu'on notera M_1 et M_2 modélisent la tendance sur les sous-parties temporelles différentes de la série chronologique :

$$M_1 : \quad y_i = a_1 + b_1 \times t_i + \varepsilon_i \text{ pour } t_i < t_0 \quad \text{et} \quad M_2 : \quad y_i = a_2 + b_2 \times t_i + \varepsilon_i \text{ pour } t_i > t_0.$$

Les expressions de ces deux modèles indiquent qu'on s'intéresse séparément aux deux périodes "avant t_0 " et "après t_0 ", en modélisant pour chacune d'entre elles la tendance linéaire de manière indépendante, de sorte à évaluer si l'y a potentiellement une différence entre "la tendance avant t_0 " et "la tendance après t_0 ".

Il est possible de reformuler cette problématique en un test d'hypothèses, pour lequel les hypothèses en compétition sont :

$$H_0 : a_1 = a_2 \text{ et } b_1 = b_2 \quad \text{vs} \quad H_1 : a_1 \neq a_2 \text{ ou } b_1 \neq b_2.$$

A noter que si l'hypothèse H_0 est vraie, alors on aura $a_1 = a_2 = a$ et $b_1 = b_2 = b$. De plus, puisque l'hypothèse H_1 fait intervenir l'opérateur logique "ou", alors lorsqu'on rejette l'hypothèse H_0 , cela peut être parce que a_1 n'est vraisemblablement pas égal à a_2 , ou parce que b_1 n'est vraisemblablement pas égal à b_2 , ou les deux simultanément, mais sans qu'on ne sache laquelle des configurations corresponde au rejet en pratique de H_0 . Quoi qu'il en soit, le rejet de l'hypothèse H_0 revient à détecter une cassure dans la tendance de la série chronologique en l'instant t_0 , on ne sait juste pas de quel type de cassure il s'agit, mais cela peut la plupart du temps se déterminer graphiquement.

Pour statuer sur l'acceptation ou le rejet de l'hypothèse H_0 , il faudrait connaître les valeurs des coefficients a_1 , a_2 , b_1 et b_2 , or ce sont des paramètres théoriques qu'il n'est pas possible de connaître. Pour avoir une idée de la valeur de ces paramètres, on doit calculer leurs estimations et on pourrait alors vouloir les comparer entre elles pour accepter ou rejeter l'hypothèse H_0 . Cependant, le test en question repose plutôt sur la statistique de test suivante, et celle-ci fait intervenir ces estimations de manière indirecte :

$$F = \frac{(\text{SCR}_c - (\text{SCR}_1 + \text{SCR}_2))/k}{(\text{SCR}_1 + \text{SCR}_2)/n_1 + n_2 - 2k} \stackrel{H_0}{\sim} \mathcal{F}(k, n_1 + n_2 + 2k) \quad (6.1)$$

où SCR_c , SCR_1 et SCR_2 sont les sommes des carrés résiduels pour les modèles M_c , M_1 et M_2 , et où les valeurs n_1 et n_2 correspondent aux nombres de données des deux sous-séries avant et après t_0 . Le paramètre k correspond au nombre de paramètres à estimer par modèle : $k = 2$ pour le cas de la tendance linéaire. De plus, $\mathcal{F}(k, n_1 + n_2 + 2k)$ correspond à la loi de Fisher et les deux paramètres donnés sont les degrés de liberté de cette loi.

Notation 6.2.1 (Sous l'hypothèse H_0). *Le symbole mathématique $\stackrel{H_0}{\sim}$ indique que la quantité en question suit une certaine loi, mais uniquement dans le cas où il s'avère que l'hypothèse H_0 est vraie. L'assertion mathématique suivante $F \stackrel{H_0}{\sim} \mathcal{F}$ se lit "la variable aléatoire F suit une loi de Fisher si l'hypothèse H_0 est vraie", ou autrement formulé : "sous l'hypothèse H_0 , la variable aléatoire F suit une loi de Fisher".*

L'intuition derrière cette statistique de test est basée sur les notions suivantes :

- La somme des carrés résiduels (SCR) correspond à la qualité d'ajustement du modèle aux données (pour vous en convaincre, vous pourrez faire le lien avec le MSE, le coefficient de détermination et le critère des moindre carrés). Et en particulier, la somme SCR_c indique à quel point le modèle M_c s'ajuste correctement sur la série chronologique.
- Comme les modèles M_1 et M_2 concernent chacun des sous-parties complémentaires du domaine temporel de la série chronologique, la somme $SCR_1 + SCR_2$ indique la qualité d'ajustement des deux modèles utilisés de manière conjointe pour modéliser la série chronologique dans son intégralité.
- Le fait d'évaluer la différence $SCR_c - (SCR_1 + SCR_2)$ permet de quantifier un écart de "qualité de modélisation" entre d'un côté le modèle M_c et de l'autre côté les modèles M_1 et M_2 utilisés conjointement. Pour l'interpréter, on peut se rendre compte que si cet écart prend de grandes valeurs (positives) alors cela indiquera que SCR_c est beaucoup plus grand que $SCR_1 + SCR_2$, autrement dit que le modèle M_c s'ajuste beaucoup moins bien aux données.

Remarque 6.2.2 (L'écart est positif.). *De plus, il faut noter que cet écart ne peut pas prendre de valeurs négatives, puisque la modélisation " $M_1 + M_2$ " ne peut pas produire des erreurs de prédictions plus importantes que le modèle M_c puisque cette modélisation correspond à une modélisation plus complexe que M_c (le détails de cette notion n'est pas au programme du cours, mais la remarque 6.2.3 donne quelques détails supplémentaires).*

Pour l'autre cas limite possible, lorsque l'écart $SCR_c - (SCR_1 + SCR_2)$ prend des valeurs proches de 0, cela indique que les deux approches de modélisation ont une qualité similaire d'ajustement aux données. Et dans ce cas-là, puisque la modélisation " $M_1 + M_2$ " correspond à une approche *plus complexe* que le modèle M_c , on pourra préférer utiliser le modèle M_c plutôt que le modèle " $M_1 + M_2$ ". Pour le formuler autrement, à qualité d'ajustement équivalente, on préfère l'approche la plus simple et donc le modèle le moins complexe : il s'agit d'une illustration du principe du Rasoir d'Ockham (au programme du cours "R4.EMS.08 Modèle linéaire", parcours EMS, au semestre 4).

- Au regard des deux précédents points, on comprend que suivant la valeur de cet écart, soit cela correspond à une configuration de la série chronologique pour laquelle le modèle M_c est préférable (hypothèse H_0 ; lorsque l'écart est proche de 0), soit la modélisation " $M_1 + M_2$ " est préférable (hypothèse H_1 ; lorsque l'écart prend une grande valeur positive). Cependant, pour distinguer entre ces deux configurations, il faut définir une valeur seuil qui permette de marquer une délimitation entre "les valeurs proches de 0" et "les grandes valeurs positives". Pour faire cela, le test de Chow consiste à normaliser d'une certaine manière l'écart $SCR_c - (SCR_1 + SCR_2)$ de sorte à ce que cela donne une quantité aléatoire F dont on connaît la loi sous H_0 (voir l'équation de la statistique (6.2.2)). La figure 6.3 illustre cela, en montrant comment est obtenu ce seuil de sorte à contrôler l'erreur consistant à rejeter H_0 à tort.

Donc si en pratique, on calcule la statistique F à partir de la série chronologique est qu'on obtient une valeur inférieure au quantile empirique en question (zone verte) alors on aura une valeur qui aura une forte densité pour cette loi de Fisher, autrement dit il est assez probable d'observer cette valeur pour une variable aléatoire suivant cette loi. Et comme on sait que F suit cette loi si l'hypothèse H_0 est vraie, on pourra dire que la valeur obtenue de cette statistique F n'est pas incohérente avec l'hypothèse H_0 .

Mais au contraire, si on obtient une valeur de la statistique F qui est plus grande que le quantile empirique en question (zone rouge) alors on aura une valeur qui n'est que très peu probable d'être observée pour la loi de Fisher. On aura alors deux possibilités : soit on a observé un événement rare de la loi de Fisher (ce qui correspond au risque de première espèce), soit c'est que la variable aléatoire F ne suit pas une loi de Fisher. Et comme F est censé suivre une loi de Fisher si H_0 est vraie, alors peut-être que l'hypothèse H_0 n'est pas vraie. Autrement dit, si on observe une valeur de F plus grande que le seuil en question, cela correspond à un indice qui remettrait en cause la validité de l'hypothèse H_0 .

Remarque 6.2.3 (Diminution de SCR). *Pour préciser ce qui est décrit dans la remarque 6.2.2, prenons un exemple théorique de deux modèles M_p et M_q ayant un certain nombre de paramètres en commun. Ecrivons ces modèles de la manière suivante :*

$$M_p : y_i = f(x_i, \theta_1, \dots, \theta_p) + \varepsilon_i \quad \text{et} \quad M_q : y_i = f(x_i, \theta_1, \dots, \theta_p) + g(x_i, \theta_q) + \varepsilon_i$$

où f et g sont des fonctions de lien. On remarque que les paramètres du modèle M_p sont $\theta_1, \dots, \theta_p$ et que le modèle M_q dispose des mêmes paramètres plus le paramètre supplémentaire θ_q . Pour la suite, supposons pour simplifier (mais sans perte de généralité) que g soit une fonction linéaire en x_i : $g(x_i, \theta_q) = x_i \theta_q$. Dans ce contexte, les prédictions des modèles sont $f(x_i, \hat{\theta}_1, \dots, \hat{\theta}_p)$ pour le modèle M_p et $f(x_i, \hat{\theta}_1, \dots, \hat{\theta}_p) + x_i \hat{\theta}_q$ pour le modèle M_q . Maintenant, quitte à supposer qu'on fixe $\hat{\theta}_q = 0$, on peut voir que le modèle M_q donnera les mêmes prédictions que le modèle M_p . Autrement dit, le modèle M_p est un cas particulier du modèle M_q . Mais si on ne fixe pas ce paramètre $\hat{\theta}_q$ à 0, alors on peut arriver à trouver une estimation $\hat{\theta}_q$ telle que cela fasse globalement améliorer les prédictions (par rapport à celles du modèle M_p), et donc faire diminuer la somme des carrés résiduels (SCR). Pour cela, il suffit d'ajuster le modèle suivant :

$$z_i = x_i \theta_q + \varepsilon_i$$

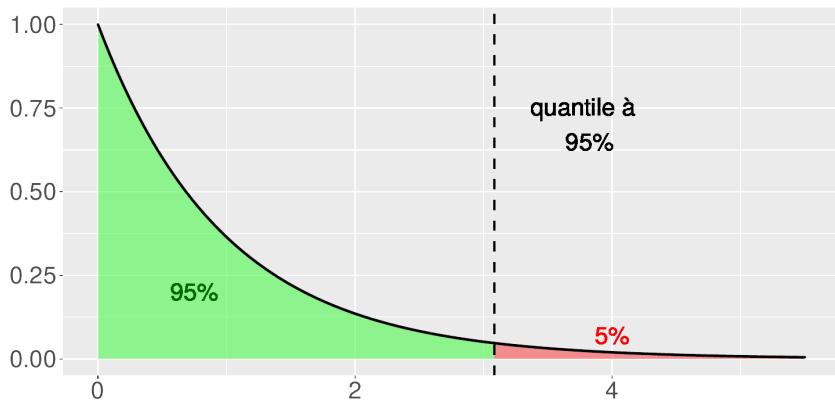


Figure 6.3 – Illustration du quantile de la loi de Fisher. La courbe représente la fonction de densité de la loi de Fisher et le trait vertical matérialise la valeur d'un quantile théorique de la loi de Fisher permettant de délimiter à partir de quelle valeur une valeur de la statistique F du test de Chow viendra remettre en question l'hypothèse nulle.

où $z_i = y_i - f(x_i, \hat{\theta}_1, \dots, \hat{\theta}_p)$ est le résidu du modèle M_p . Finalement, pour le formuler autrement, le modèle M_q correspond à une extension du modèle M_p dans le sens où il dispose des mêmes paramètres et d'un paramètre supplémentaire, et que ce paramètre supplémentaire est une variable sur laquelle on peut jouer pour améliorer les prédictions du modèle M_p . Le fait d'améliorer les prédictions signifie qu'elles sont plus proches des véritables valeurs, ce qui amène à ce que les résidus soient plus petits et donc que SCR diminue.

Puisque qu'on connaît la loi de la statistique F sous l'hypothèse H_0 , on peut en déduire la région de rejet R de la manière suivante :

$$R = [f_{\nu_1, \nu_2}^{1-\alpha}, +\infty] \quad (6.2)$$

où $f_{\nu_1, \nu_2}^{1-\alpha}$ est le quantile d'une loi de Fisher de degré de liberté ν_1 et ν_2 au niveau de probabilité $1 - \alpha$, et où ici $\nu_1 = k$, et $\nu_2 = n_1 + n_2 + 2k$. En pratique, il faudra vérifier si la valeur de la statistique F appartient ou non à R et en particulier si $F \in R$, alors on rejettéra l'hypothèse H_0 au profit de l'hypothèse H_1 au risque de première espèce α . Cependant, comme toute procédure de test statistique, une alternative à la région de rejet est la *p-value* et celle-ci se calcule de la même manière que pour tout les tests d'hypothèses.

Exemple 6.2.4. Pour illustrer ce que cela donne sur un exemple concret, supposons disposer des quatre séries chronologiques représentées dans la figure 6.4. Chacune d'entre elles correspond à une configuration différente par

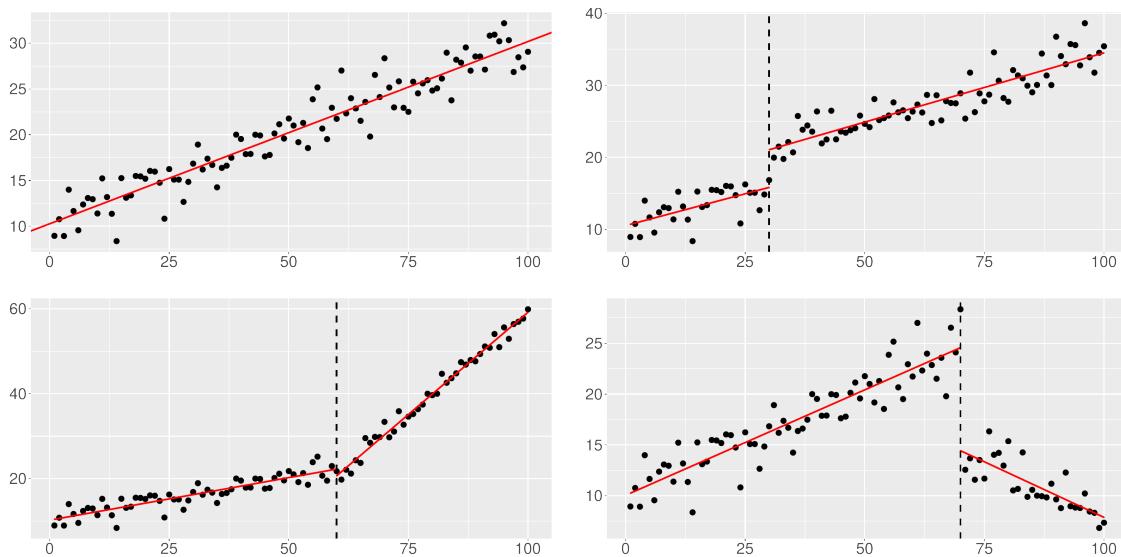


Figure 6.4 – Différentes configurations de cassure : séries chronologiques et leurs modélisations.

rapport aux hypothèses H_0 et H_1 . Les résultats pour chacun de ces cas sont :

1. La *p-value* vaut 0.8575, donc on ne rejette pas l'hypothèse H_0 au profit de H_1 .

2. La p -value vaut 2.35×10^{-10} , donc on rejette l'hypothèse H_0 au profit de H_1 .
3. La p -value vaut 4.92×10^{-51} , donc on rejette l'hypothèse H_0 au profit de H_1 .
4. La p -value vaut 8.33×10^{-45} , donc on rejette l'hypothèse H_0 au profit de H_1 .

De plus, la figure 6.4 illustre l'ajustement du meilleur modèle (M_c ou " $M_1 + M_2$ ") dans chacune des quatre configurations. On constate pour chacun des cas que :

1. Une tendance linéaire paraît adaptée pour modéliser la série chronologique dans son intégralité.
2. Il semble y avoir une cassure dans la valeur de l'intercept (le paramètre a) en $t_0 = 30$ et donc il est normal qu'on trouve que $a_1 \neq a_2$.
3. Il semble y avoir une cassure dans la valeur de la pente de la tendance linéaire (le paramètre b) en $t_0 = 60$ et donc il est normal qu'on trouve que $b_1 \neq b_2$.
4. Il semble y avoir une cassure dans les valeurs de l'intercept et de la pente en $t_0 = 70$.

Diapos de cours et feuille de TD

Chap. 3 – Anomalies et cassures

Introduction

Contexte :

On dispose d'une série chronologique pour laquelle on suspecte qu'il y ait des anomalies qu'on souhaite détecter mathématiquement.

Les anomalies peuvent être de plusieurs sortes :

- la présence d'une donnée atypique, ou
- la présence d'une "cassure" dans l'évolution de la série chronologique.

Méthode de détection de donnée atypique :

Etant donné les caractéristiques d'une série chronologique (chaotique, présence d'une tendance et/ou d'une périodicité), une donnée atypique sera détectée comme une donnée anormale au regard d'un modèle.

La procédure est donc :

1. ajuster un modèle sur la série chronologique,
2. calculer les écarts entre les prédictions du modèles et la série, et
3. déterminer les données pour lesquels l'écart est anormalement grand.

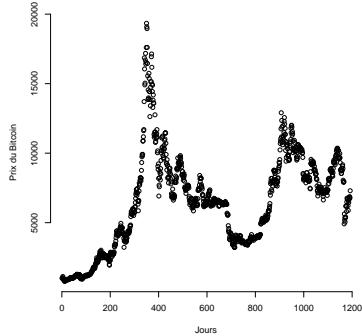
Méthode de détection d'une cassure :

Une cassure à un instant t de la série chronologique peut se déterminer de la manière suivante.

- On sépare le nuage de points en deux nuages de points : celui avant t et celui après t .
- On ajuste un modèle pour chacun des deux nuages.
- Si les modèles sont similaires, il n'y a pas de cassures. Mais, si les modèles sont très différents, on en conclut la présence d'une cassure à l'instant t .

Problématique : Le prix du Bitcoin

Les données : le prix du Bitcoin en dollars américains.



Quels sont les jours pour lesquels le cours du Bitcoin est anormalement bas ou anormalement haut au regard du niveau hebdomadaire du prix ?

Détection d'anomalies

Détection par moyenne mobile :

Pour déterminer s'il y a des anomalies (des données anormales) il faut d'abord déterminer ce qu'est la "normalité" pour ces données. Pour cela, on va utiliser une modélisation par moyenne mobile.

On note m_i la moyenne mobile (d'ordre k) et on peut décomposer la série comme :

$$y_i = m_i + e_i,$$

où e_i sont les résidus : la part chaotique en dehors de la "normalité" établie par la valeur m_i de la moyenne mobile.

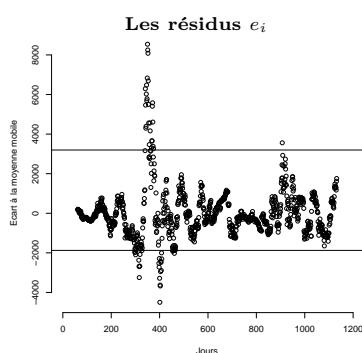
Détection d'anomalies

Détection par moyenne mobile :

Les résidus ne sont pas égaux à 0 et ont une certaine variabilité. Autrement dit, il est normal que la valeur de la série (y_i) ne soit pas exactement égale

à la moyenne mobile m_i (du fait de l'évolution chaotique de la série) et il est normal que les résidus soient autour de 0.

Cependant, il est anormal de voir des valeurs de résidus trop grandes (positives ou négatives). Ces résidus trop grands sont relatifs à des mesures y_i anormalement éloignées de la valeur du modèle : m_i . On peut déterminer un intervalle qui recouvre 95% des résidus les plus proches de 0.



Détection d'anomalies 2

Détection par lissage exponentiel :

On peut procéder de manière similaire avec d'autres manières de modéliser la série chronologique.

Une autre modélisation possible repose sur le lissage exponentiel simple (ou double). On note ℓ_i la série lissée pour le temps t_i et la modélisation est :

$$y_i = \ell_i + e_i$$

En procédant comme pour la modélisation par moyenne mobile on obtient :

Régression locale

Une modélisation plus avancée de la série chronologique permet de déterminer des anomalies en suivant la même procédure.

Régression locale :

(usuellement appelée *loess*) Cette méthode consiste à exécuter des régressions linéaires sur des régions successives du nuage de points. Les régions en question n'ont pas besoin d'être disjointes.

Le résultat final consiste à faire une moyennisation de toutes les régressions locales.

Au final on obtient une modélisation : $y_i = f(x_i) + \varepsilon_i$ où f est une courbe plus complexe qu'une forme linéaire.

•
•
•
•
•
•

Détection de cassures

Cassure :

Une cassure dans une série chronologique est un instant t pour lequel les séries avant et après cet instant t semblent différentes.

Cette cassure peut se "voir à l'œil nu" mais étant donné le caractère chaotique de la série chronologique, il est souvent nécessaire de déterminer mathématiquement s'il y a bien une cassure.

Exemple :

Pour la série représentée par le nuage de points suivant, il y a une cassure qui n'est pas évidente à voir à l'œil nu : en $t = 0$.

Détection de cassures

On cherche à déterminer s'il y a une cassure en $t = 0$.

Un seul modèle :

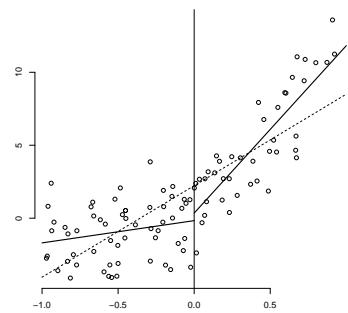
Si on ajuste un modèle de régression linéaire pour ce nuage de points, on trouve le modèle :

$$y_i = 2.227 + 6.272x_i + e_i.$$

Deux modèles :

Si on ajuste un modèle de régression linéaire pour les nuages de points avant $t = 0$ et après $t = 0$, on trouve les modèles :

$$y_i = -0.168 + 1.525x_i + e_i \quad \text{et} \quad y_i = 0.348 + 11.476x_i + e_i.$$



Test de Chow

Contexte :

On dispose d'une série temporelle (t_i, y_i) pour $i = 1, \dots, n$ pour laquelle on présume que l'instant t_0 est une cassure. On note M_c , M_1 et M_2 les modèles suivants :

$$M_c : y_i = a + b \times t_i + \varepsilon_i$$

$$M_1 : y_i = a_1 + b_1 \times t_i + \varepsilon_i \text{ pour } t_i < t_0$$

$$M_2 : y_i = a_2 + b_2 \times t_i + \varepsilon_i \text{ pour } t_i > t_0$$

La question est de savoir si les paramètres a_1 et b_1 du modèle M_1 sont égaux ou non aux paramètres a_2 et b_2 du modèle M_2 . En particulier, s'ils sont égaux, ils sont alors égaux aux paramètres a et b du modèle M_c .

Test de Chow :

• Les hypothèses en compétition sont :

$$H_0 : a_1 = a_2 \text{ et } b_1 = b_2 \quad \text{vs} \quad H_1 : a_1 \neq a_2 \text{ ou } b_1 \neq b_2$$

• La statistique à calculer est :

$$F = \frac{\frac{(SCR_c - (SCR_1 + SCR_2))}{k}}{\frac{(SCR_1 + SCR_2)}{n_1 + n_2 - 2k}}$$

où SCR_c , SCR_1 et SCR_2 sont les sommes des carrés résiduels pour les modèles M_c , M_1 et M_2 , les valeurs n_1 et n_2 correspondent aux nombres de données des deux sous-séries. Le paramètre k correspond au nombre de paramètres à estimer par modèle : $k = 2$ pour le cas de la régression linéaire.

- La région de rejet associée à un risque α de première espèce est de la forme :

$$R = [f_{\nu_1, \nu_2}^{1-\alpha}, +\infty[$$

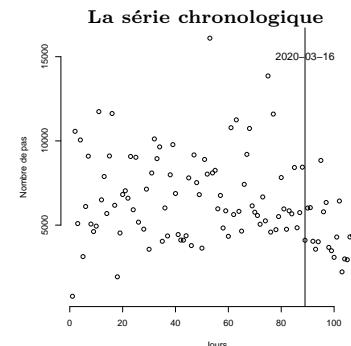
où $f_{\nu_1, \nu_2}^{1-\alpha}$ est le quantile de la loi de Fisher dont les degrés de liberté sont ν_1 et ν_2 , au niveau de probabilité $1 - \alpha$, où $\nu_1 = k$ et $\nu_2 = n_1 + n_2 + 2k$.

Nombre de pas et confinement

Exemple :

Un capteur mesure le nombre de pas effectués par une personne chaque jour. La période de mesure commence en Décembre 2019 et finit en Avril 2020. La question est de savoir si le nombre de pas est le même avant et pendant la période de confinement ayant commencé le 16 Mars 2020.

On note (t_i, y_i) pour $i = 1, \dots, 109$ la série chronologique, où t_i est une journée et y_i le nombre de pas. La date $t_{90} = 2020/03/16$ est la date du confinement et on présume qu'il y a une cassure dans la série chronologique en t_0 .



Nombre de pas et confinement

Application du test de Chow :

1. Calculer les prédictions pour chacun des trois modèles.
2. Déterminer les résidus des modèles.
3. Calculer les sommes des carrés des résidus.
4. Déterminer la région de rejet R .
5. Calculer la statistique F .
6. Accepter ou rejeter l'hypothèse H_0 suivant si F appartient à R ou non.

	t_1	...	t_{89}	t_{90}	...	t_{109}	
y_i	10570	...	6015	6037	...	5457	
\hat{y}_i	7271	...	5914	5898	...	5605	
e_i	3298	...	100	138	...	-148	
e_i^2	10877757	...	10153	19097	...	22106	SCR _c 70914486
M_1	$\hat{y}_{1,i}$	6736	...	6890			SCR ₁ 59996761
	$e_{1,i}^2$	14696573	...	765944			
M_2	$\hat{y}_{2,i}$				6734	...	SCR ₂ 68668498
	$e_{2,i}^2$				486703	...	1718319

Nombre de pas et confinement

Application du test de Chow :

Les sommes des carrés résiduels sont :
 $SCR_c = 709144865$, $SCR_1 = 599967610$ et $SCR_2 = 68668498$.

De plus, $N_1 = 89$, $N_2 = 20$, $k = 2$, et le quantile est $f_{2;105}^{95\%} = 3.08$.
 Donc la région de rejet est : $R = [3.08, +\infty[$.

On obtient finalement $F = 3.180$.

Conclusion :

Comme F appartient à la région de rejet R , alors on rejette l'hypothèse H_0 et on en déduit que la date du 16 Mars 2020 est bien un point de cassure de cette série chronologique.

Résumé

-
- Qu'est-ce qu'une anomalie ? Qu'est-ce qu'une cassure ?
 - Déterminer un modèle pour la série chronologique
Moyenne mobile, lissage exponentielle, Régression locale
 - Régression linéaire et cassure
 - Test de Chow
Interprétation, mise en pratique

TD 2 – Détection d'anomalies et de cassures

Exercice 1. Les données de cet exercice sont celles étudiées dans l'exercice 3 de la feuille de TD 3.

1. Calculez la moyenne mobile d'ordre 3 de cette série.
2. Calculez les résidus de cette moyenne mobile.
3. Déterminez les quantiles des résidus à 5% et à 95%.
4. En déduire quels sont les résidus anormaux et quelles sont les mesures anormales de la série.
5. Sur un graphique, tracez la série chronologique, la moyenne mobile et mettez en lumière les anomalies.

Exercice 2. Refaire les mêmes questions de l'exercice 1 avec les mêmes données mais en utilisant un lissage exponentiel simple (avec $\gamma = 0.3$) à la place de la moyenne mobile d'ordre 3.

Exercice 3. Les données de cet exercice sont celles étudiées dans l'exercice 2 de la feuille de TD 3. Le but de cet exercice est de montrer s'il y a une cassure dans la série chronologique pour le 12 mars 2020.

1. Ajustez un modèle de régression linéaire M_C pour l'ensemble de la série chronologique. Pour cela, vous pourrez utiliser les résumé statistiques suivants :

\bar{t}	\bar{y}	$\bar{t^2}$	\bar{yt}
16	6894.904	336	102656.4

2. Calculez les prédictions et les résidus du modèle M_c .
3. Ajustez un modèle de régression linéaire M_1 pour les 11 premières mesures de la série chronologique. Pour cela, vous pourrez utiliser les résumé statistiques suivants :

\bar{t}_1	\bar{y}_1	\bar{t}_1^2	$\bar{y}_1 t_1$
6	8544.726	46	50256.77

4. Calculez les prédictions et les résidus du modèle M_1 pour les 11 premières mesures de la série chronologique.
5. Ajustez un modèle de régression linéaire M_2 pour la série chronologique privée des 11 premières mesures. Pour cela, vous pourrez utiliser les résumé statistiques suivants :

\bar{t}_2	\bar{y}_2	\bar{t}_2^2	$\bar{y}_2 t_2$
21.5	5987.501	495.5	131476.3

6. Calculez les prédictions et les résidus du modèle M_2 pour la série chronologique privée des 11 premières mesures.
7. Calculez la statistique F du test de Chow.
8. Déterminez la région de rejet R .
9. Déduisez-en la décision issue du test de Chow dans ce contexte.

Exercice 4. Les données de cet exercice sont celles étudiées dans l'exercice 2 de la feuille de TD 3. Le but de cet exercice est de montrer que la cassure du 12 mars 2020 est une anomalie d'un point de vue de la série des différences. Pour cela, il sera nécessaire de reprendre la série des différences calculée à la question 1 de l'exercice 2 de la feuille de TD 3.

1. Déterminez la tendance linéaire de la série des différences.
2. Calculez les résidus de cette tendance linéaire.
3. Calculez les quantiles des résidus à 5% et 95%.
4. Déterminez quelles sont les différences qui sont des anomalies.
5. Déduisez-en que la cassure de la série chronologique brute est une anomalie de la série des différences.

