“Big data” ou “good data” : telle est la question

ou quelques questions à se poser avant d’utiliser des techniques d’apprentissage statistique

Ivan Kojadinovic

Laboratoire de mathématiques et applications de Pau, UMR CNRS 5142
Université de Pau et des Pays de l’Adour, France

Talence, janvier 2020

Plan

Plan

Les données sont souvent des séries temporelles I

Les données sont souvent des séries temporelles II

- La collecte et le stockage de données n'ont jamais été aussi faciles.
- Dans de nombreux secteurs (industrie, environnement, finance, recherche, etc), il est ainsi possible d'accéder à de **très grands volumes de données**.
- Dans de nombreux cas de figure, ces données ont été collectées au cours du temps : il s'agit de **séries temporelles**.

t	$X_{t,1}$	$\dots$	$X_{t,p}$
1	$\dots$		$\dots$
$\vdots$		$\vdots$	
n	$\dots$		$\dots$

- Souvent, les valeurs d'une ligne vont être liées aux valeurs des lignes précédentes : c'est la **dépendance temporelle** ou **sérielle**.

Les données sont souvent des séries temporelles III

- Cela est particulièrement vrai en présence de **variables retard**, par exemple, quand $X_{t,2}$ est en fait définie comme $X_{t-1,1}$.

- Pour exploiter ces données, l'étape suivante consiste souvent à utiliser des techniques d'**apprentissage statistique** supervisé ou non supervisé :
  - clustering (classification en français), règles d'association, arbres de décision, forêts aléatoires, réseaux de neurones, modèles statistiques de discrimination ou régression, etc.

- **Quelques exemples :**
  - **Industrie :** on cherche à prédire la "qualité du verre" en fonction de nombreux paramètres physico-chimiques (capteurs) ;
  - **Finance :** on cherche à modéliser le risque associé à un portefeuille d'actions en fonction du prix de ces actions (bourse) et divers autres indicateurs financiers ;

Les données sont souvent des séries temporelles IV

- **Génie côtier :** on cherche à modéliser l'impact à la côte de tempêtes hivernales en fonction d'**états de mers** afin de dimensionner des ouvrages de protection de la côte ;
- **E-commerce :** systèmes de recommandation, segmentation des clients, etc ;
- **Banque :** modèles de solvabilité des demandeurs de crédits, etc.

- Pour fixer les idées (et en utilisant le langage de l'apprentissage statistique), supposons que l'on souhaite **apprendre** la relation qui existe entre $X_{t,1}$ et les autres variables mesurées.
- Les lignes de la matrice de données sont alors vues comme des **exemples d'apprentissage** liant les valeurs de $X_{t,1}$ aux valeurs de $X_{t,2}, \ldots, X_{t,p}$.

Plan

La stationnarité

- Pour que cet apprentissage ait du sens, il faut que ce que l'on cherche à apprendre soit **"stable / non changeant"** au cours du temps.

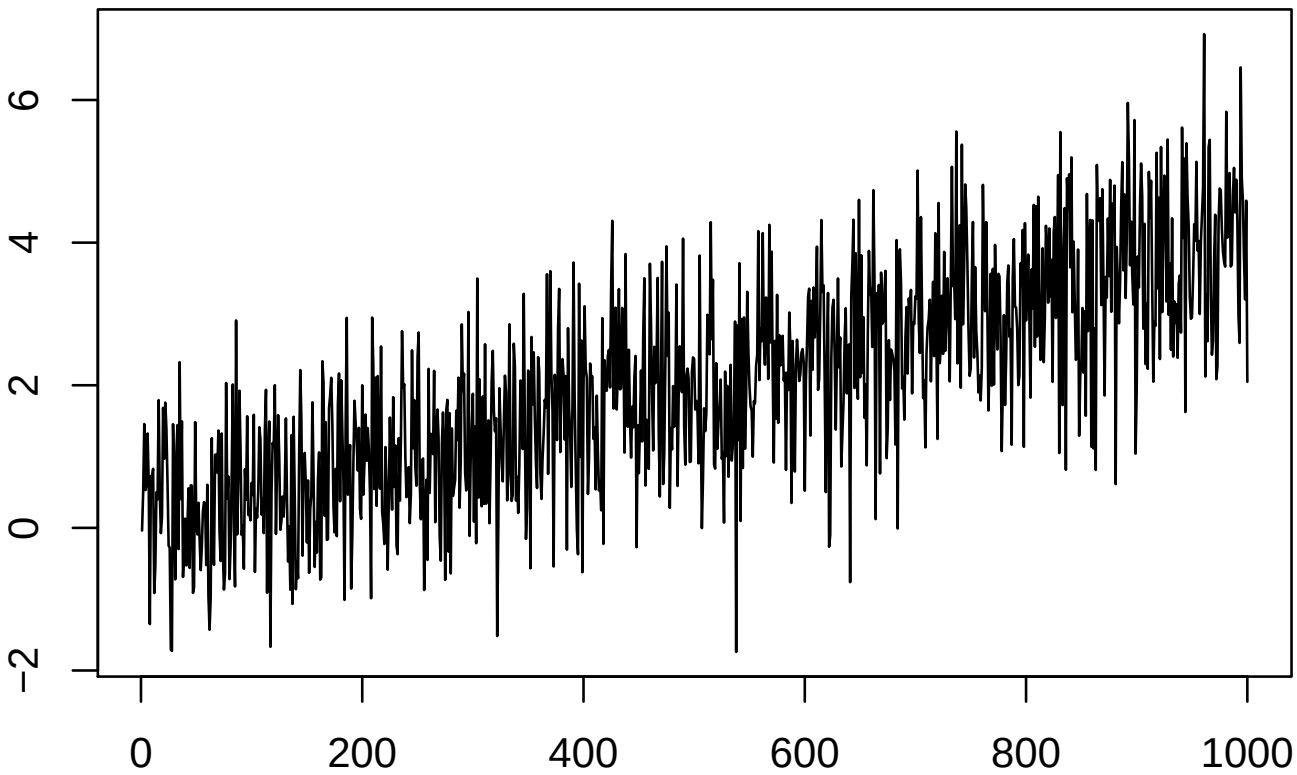
- En d'autres termes, il faut que les **propriétés statistiques** du processus aléatoire sous-jacent soient **"stables" au cours de la période d'observation**.

- En probabilités et en statistique, cette propriété fondamentale est appelée la **stationnarité**.

- Cela revient à supposer que les variables observées sont **"stables"** en moyenne, en dispersion, en corrélation, en dépendance sérielle, etc.
  - Intuitivement, les moyennes, variances, corrélations et autocorrélations entre variables, etc, **ne fluctuent "pas trop"** lorsqu'elles sont calculées sur des **fenêtres glissantes**.

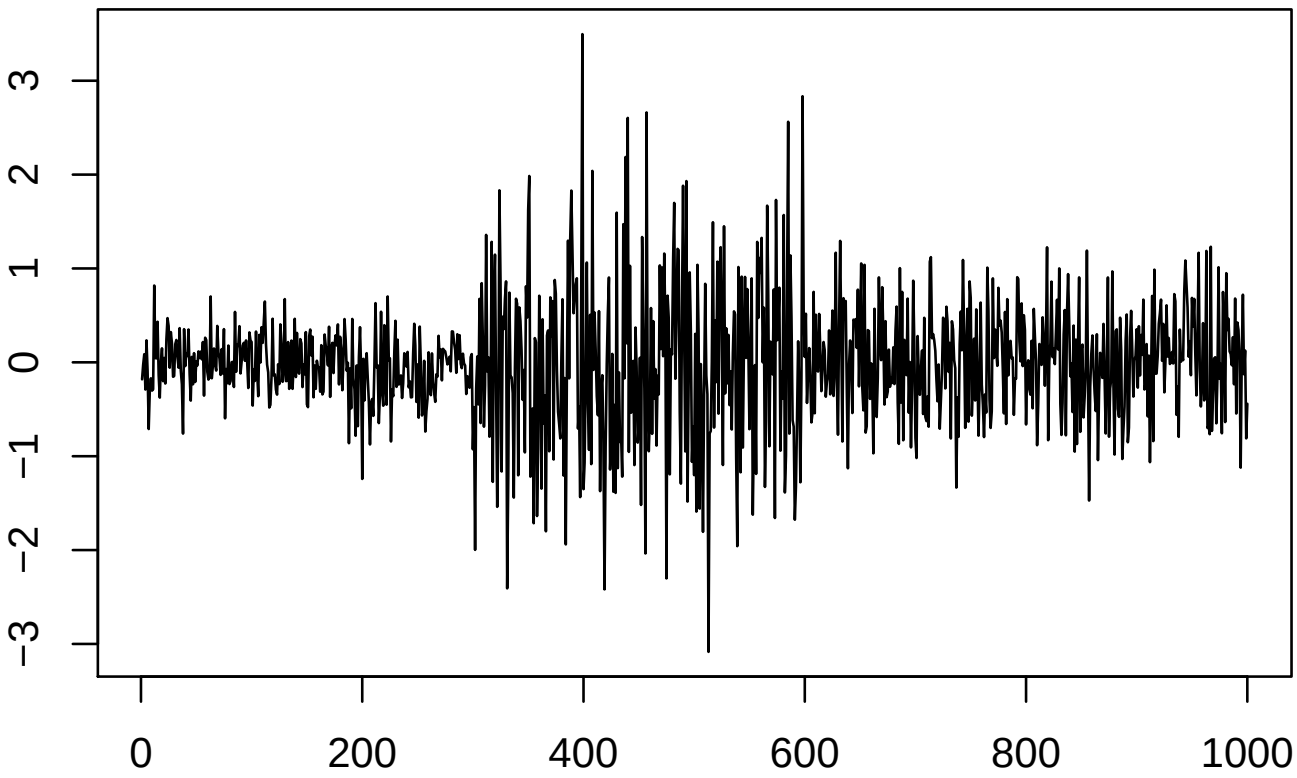
Plan

Exemples de non stationnarité



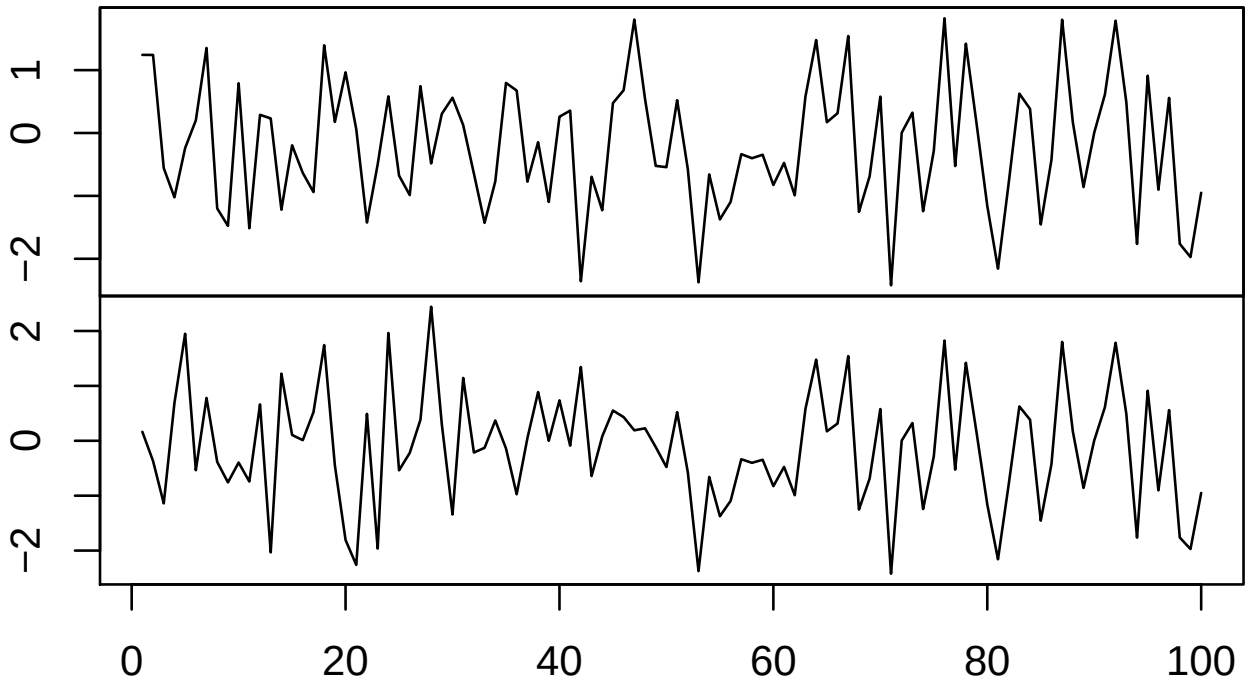
La moyenne change au cours du temps

Exemples de non stationnarité



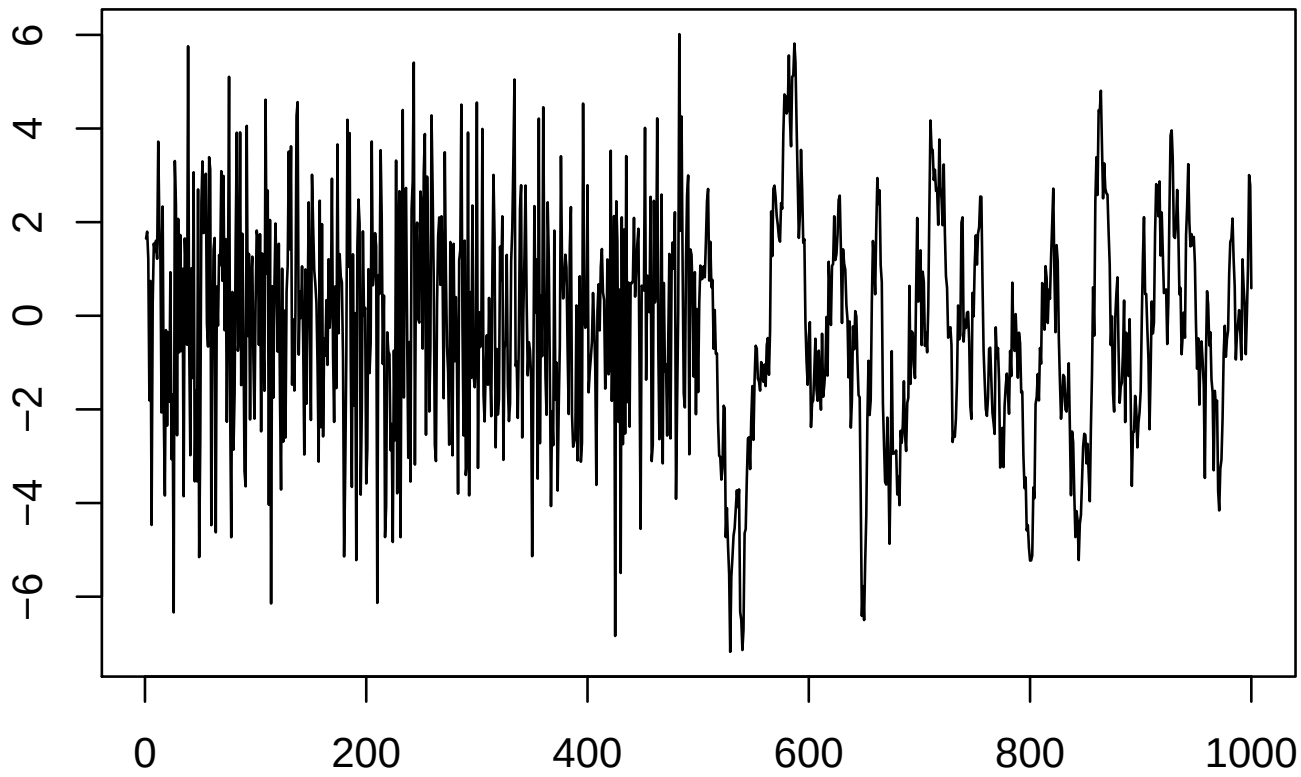
La variance change au cours du temps

Exemples de non stationnarité



La corrélation change au cours du temps

Exemples de non stationnarité



La dépendance sérielle (autocorrélation) change au cours du temps

Plan

Comment tester la stationnarité ? I

- La première étape est de **discuter avec des experts du domaine** : a-t-on des raisons de croire que le processus aléatoire sous-jacent a changé au cours de la période d'observations ?
  - 150 ans de données hydrologiques ; doit-on s'inquiéter ?
  - 5 ans de données de vente en ligne ; doit-on s'inquiéter ?
  - 1 ans de données de pollution d'une rivière ; doit-on s'inquiéter ?
- En cas d'absence d'experts ou de doute, la stationnarité du processus aléatoire sous-jacent peut être testée à partir des données disponibles à l'aide de **tests statistiques** .
- Le sous-champ disciplinaire de la **statistique** qui s'intéresse à cette question est celui de la **détection de ruptures** (en anglais **change-point analysis**).

Comment tester la stationnarité ? II

- Intuitivement, l'idée est de **segmenter** les séries temporelles et de vérifier que la plupart des caractéristiques statistiques usuelles ne fluctuent "pas trop".

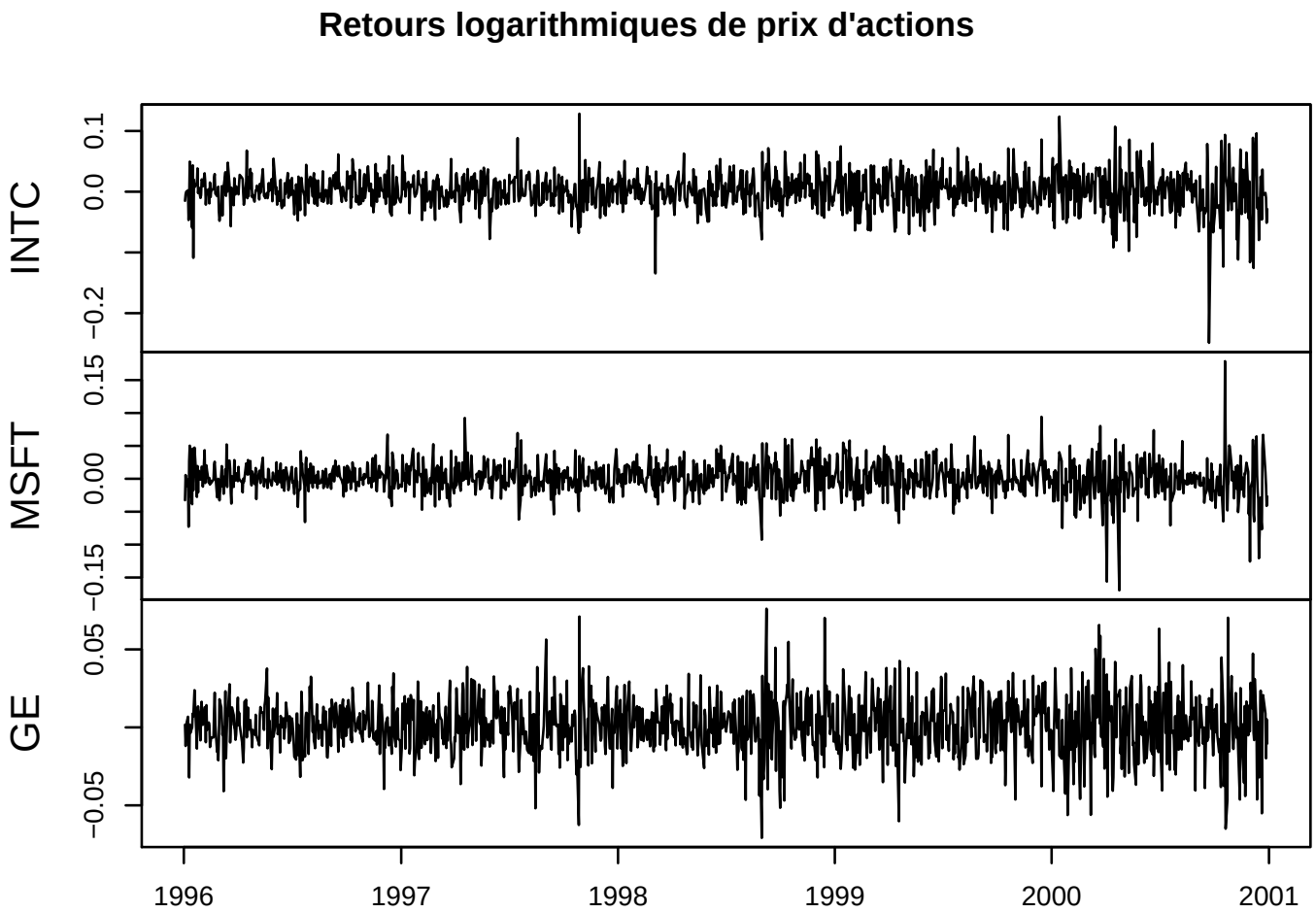
- Comme souvent en statistique, **on cherche des preuves contre la stationnarité** dans les données : stabilité de moyennes ? des variances ? des corrélations ? des autocorrélations ? ...

- En cas d'absence de preuves, on jugera la **stationnarité plausible** et on passera à la phase de modélisation.

- Tester la stationnarité est une **tâche complexe méthodologiquement et computationnellement**.

- Il n'existe pas de test universel dans la mesure où **il existe une infinité de façons différentes d'être non stationnaire**.

Comment tester la stationnarité ? III



Comment tester la stationnarité ? IV

```
> library(npcp)
[...]
> dim(Xrdj)
[1] 1262    4

> head(Xrdj)
        Date          INTC          MSFT           GE
1 1996-01-03 -0.015037877 -0.032559878  0.001692413
2 1996-01-04 -0.004340400  0.005738883 -0.011949427
3 1996-01-05  0.000000000 -0.011510892  0.001718221
4 1996-01-08  0.002165577 -0.001455179  0.006835194
5 1996-01-09 -0.046623697 -0.072881517 -0.009627754
6 1996-01-10 -0.016037134  0.026920520 -0.032119624

> cpDist(Xrdj)
```

Comment tester la stationnarité ? V

```
        Test for change-point detection sensitive
        to changes in the distribution function

data:  Xrdj
cvmmax = 325.12, p-value = 0.002498
User time:  73.5 sec

> cpCopula(Xrdj)

        Test for change-point detection sensitive
        to changes in the copula

data:  Xrdj
cvmmax = 57.28, p-value = 0.01049
User time:  127.8 sec
```

Plan

Que faire en cas de non stationnarité ? I

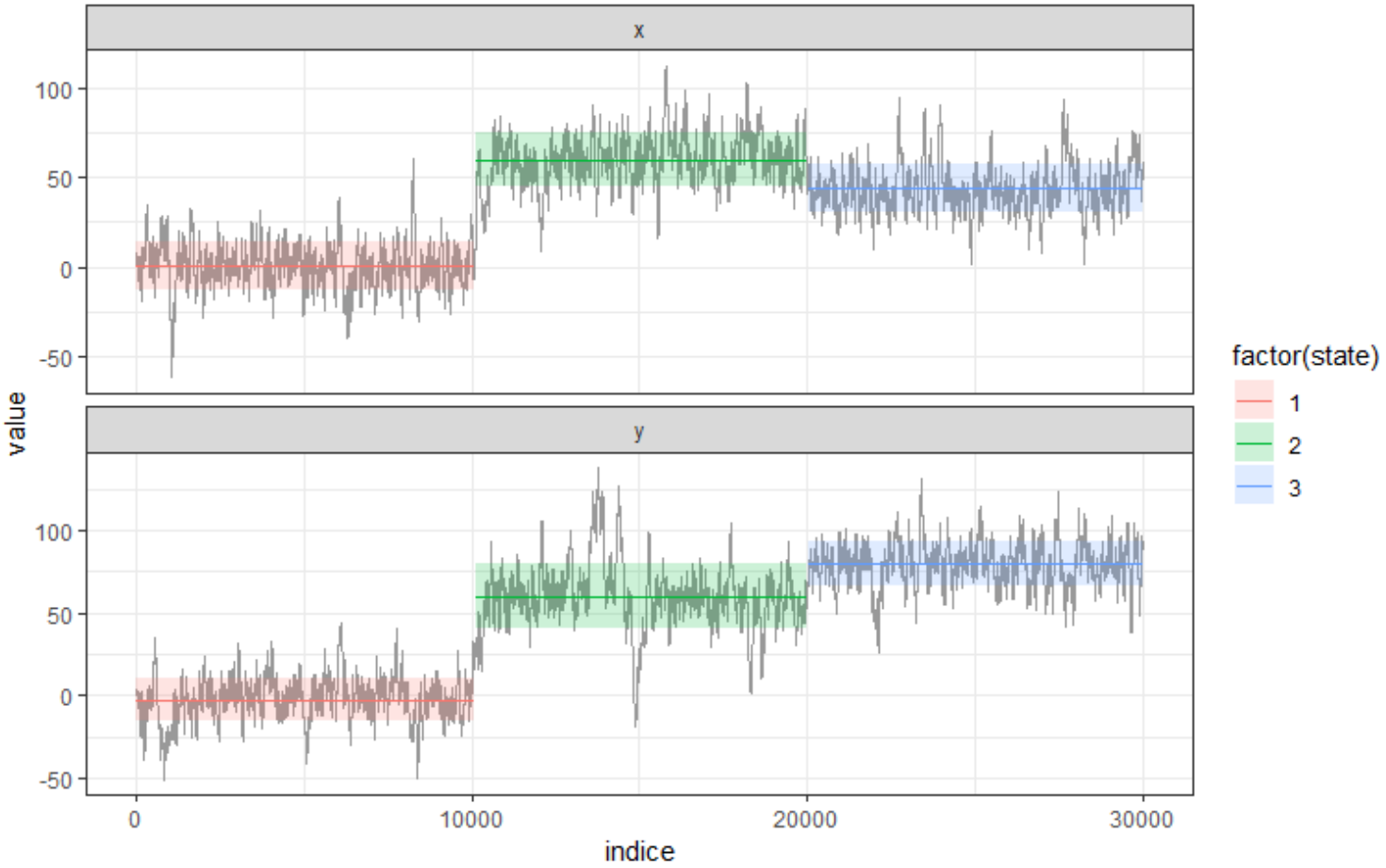
- Dans le cas d'une **tendance**, la modéliser séparément et modéliser alors l'écart à la tendance.
- **Segmenter les séries temporelles** afin de se ramener à des sous-séries qui peuvent être considérées comme stationnaires d'un certain point de vue.
- Pour cela, il faut détecter des **points de rupture** (**change points** en anglais).
- Ces points correspondent à des dates auxquelles certaines caractéristiques des séries temporelles peuvent être considérées comme ayant changées.

Que faire en cas de non stationnarité ? II

- L'identification des points de rupture est une **tâche très ardue méthodologiquement et computationnellement**.

- Les algorithmes travaillent sous de fortes hypothèses sur les propriétés des séries et s'intéressent essentiellement à des **ruptures dans la moyenne**.

Que faire en cas de non stationnarité ? III



Plan

Conclusion I

- Ce n'est pas la taille qui compte ! **Avoir beaucoup de données recueillies dans un contexte de changement ne sert à rien**.

- Il vaut mieux privilégier des **séries temporelles courtes** (relativement au domaine d'application) qui peuvent être considérées comme **stationnaires** plutôt que des séries temporelles longues pour lesquelles le processus aléatoire sous-jacent à modéliser a changé au cours de la période d'observation.

- La **stationnarité peut être testée** mais c'est une tâche ardue méthodologiquement et computationnellement qui mobilise beaucoup de chercheurs en statistique.

Conclusion II

- En cas de non stationnarité, les **séries temporelles peuvent être segmentées**. Il s'agit encore une fois d'une tâche complexe méthodologiquement et computationnellement qui mobilise de nombreux chercheurs.

- **Packages R pour la détection de ruptures et la segmentation :** changepoint, strucchange, ecp, bcp, mcp, npcp, segmentr, segmented, cpm, EnvCpt, wbs, wbsts, segclust2d, ...