Times series forecasting et données covid

Julien JACQUES

Universite Lumière Lyon 2

Diplôme et niveau

Ce cours de séries temporelles est donné :

- Université Lumière Lyon 2
  - M2 Informatique
    - parcours SISE (Stat. et Info. pour la Science des DonnéEs)

Semestre 3 (30 Crédits ECTS)

APPLICATIONS STATISTIQUES
- CM Applications, marketing (TD)
- Fouille de données massives (CM)
- Fouille de données massives (TD)
- TD Applications, marketing (TD)
- Text Mining - Données non structurées (TD)

INFORMATIQUE APPLIQUEE
- CM Programmation statistique sous R (TD)
- Entrepôt de données avancées (CM)
- Entrepôt de données avancées (TD)
- Logiciels spécialisés - Statistique, Data Mining, BI (TD)
- Machine learning sous Python (TD)
- TD Programmation statistique sous R (TD)

METHODES STATISTIQUES
- CM Analyse de variance et plan d'expériences (TD)
- CM Biostatistique, données catégorielles (TD)
- CM Séries temporelles et données séquentielles (TD)
- TD Analyse de variance et plan d'expériences (TD)
- TD Biostatistique, données catégorielles (TD)
- TD Séries temporelles et données séquentielles (TD)

Semestre 4 (30 Crédits ECTS)

DATA SCIENCE
- CM Data Mining - Apprentissage statistique (TD)
- CM Fouille du web et analyse des réseaux sociaux (TD)
- CM Visualisation et analyse des données de sécurité (TD)
- Initiation à la recherche - Big Data (CM)
- Initiation à la recherche - Big Data (TD)
- TD Data Mining - Apprentissage statistique (TD)
- TD Fouille du web et analyse des réseaux sociaux (TD)
- TD Visualisation et analyse des données de sécurité (TD)

PROFESSIONNALISATION
- Anglais (TD)
- Gestion de projet (CM)
- Gestion de projets (TD)
- Séminaires de recherche et ateliers techniques (CM)
- Séminaires de recherche et ateliers techniques (TD)
- Technique de recherche d'emploi (projet professionnel) (TD)

STAGE

Objectifs du cours

L'objectif du cours est d'apprendre à prédire la suite d'une série temporelle :

- à partir de l'observation d'une série

$$(x_t)_{1 \leq t \leq n} = (x_1, \ldots, x_n)$$

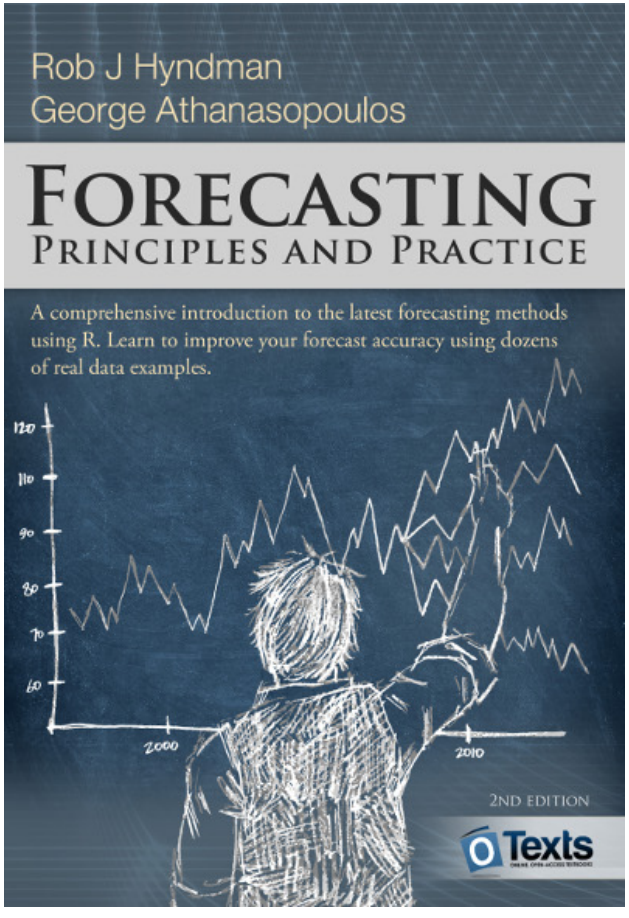
où t est le temps (seconde, jour, année...).

- prédire le futur de cette série

$$x_{n+1}, x_{n+2}, \cdots$$

Référence de ce cours

Hyndman R.J. and Athanasopoulos G. *Forecasting: Principles and Practice*, OTexts, 2013.
https://robjhyndman.com/uwafiles/fpp-notes.pdf



Contenu du cours

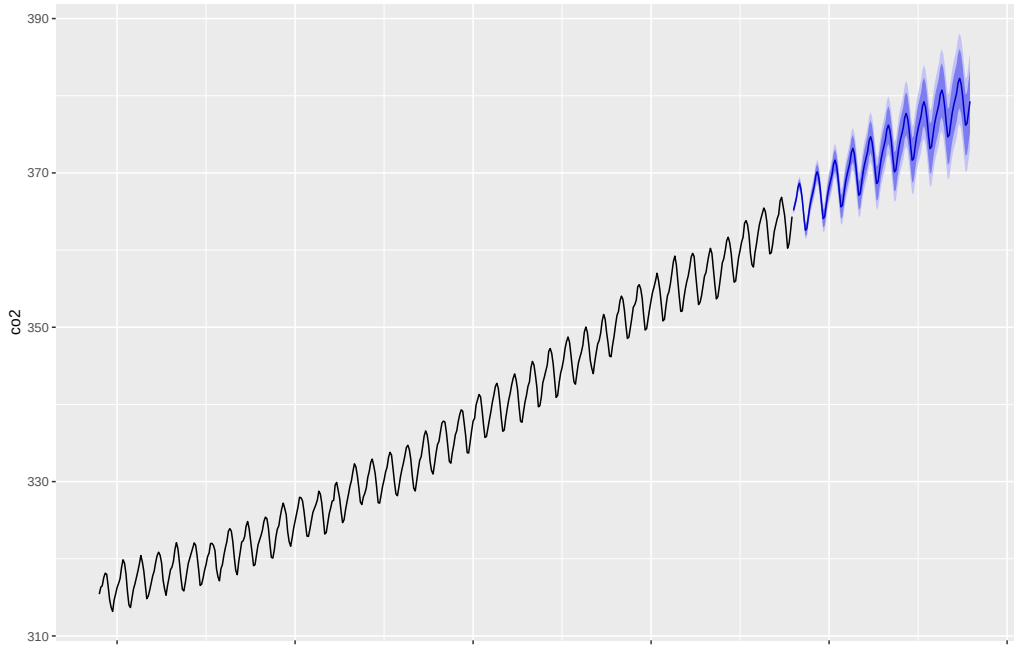
- Méthode de lissage exponentiel
- Modèles SARIMA
- Modèles neuronaux
- Modèles avec covariables
- Modèles pour séries multivariés

Contenu du cours

Pour chaque méthode de prévision

- on travaille avec des séries temporelles classiques en fil rouge
- on compare la qualité de prédiction au fur et à mesure que l'on découvre les méthodes

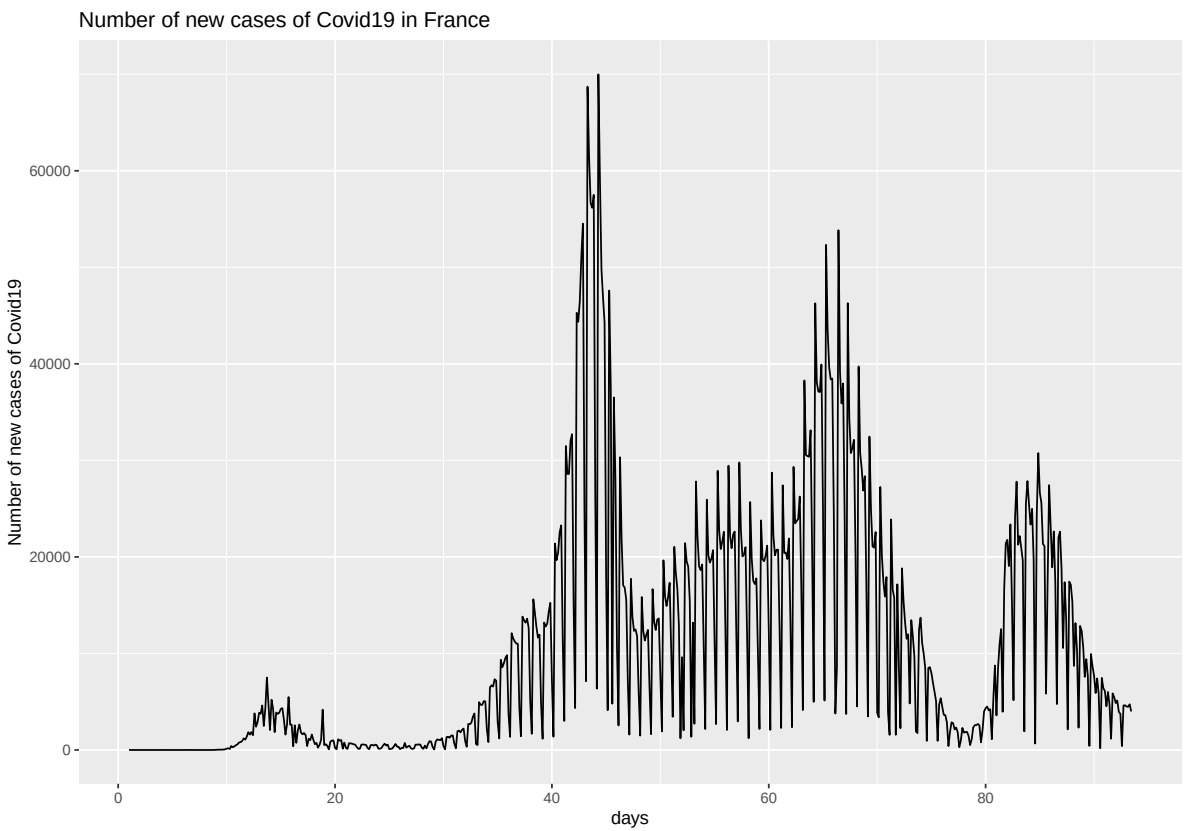
```
library(forecast)
library(ggplot2)
h=hw(co2,seasonal='additive',damped=FALSE,h=120)
autoplot(co2)+autolayer(h)
```



Données covid19

Nombre journalier de nouveaux cas de covid19

```
covid19 = read.csv("2021-10-12-WHO-COVID-19-global-data.csv")
covid19_F=covid19[covid19$Country=="France",]
covid19_F_nc=ts(covid19_F$New_cases,freq=7)
autoplot(covid19_F_nc) +
ggtitle('Number of new cases of Covid19 in France')+ xlab('days'
ylab('Number of new cases of Covid19')
```



Données covid19

Les données covid sont mises à jour chaque jour
https://covid19.who.int/WHO-COVID-19-global-data.csv

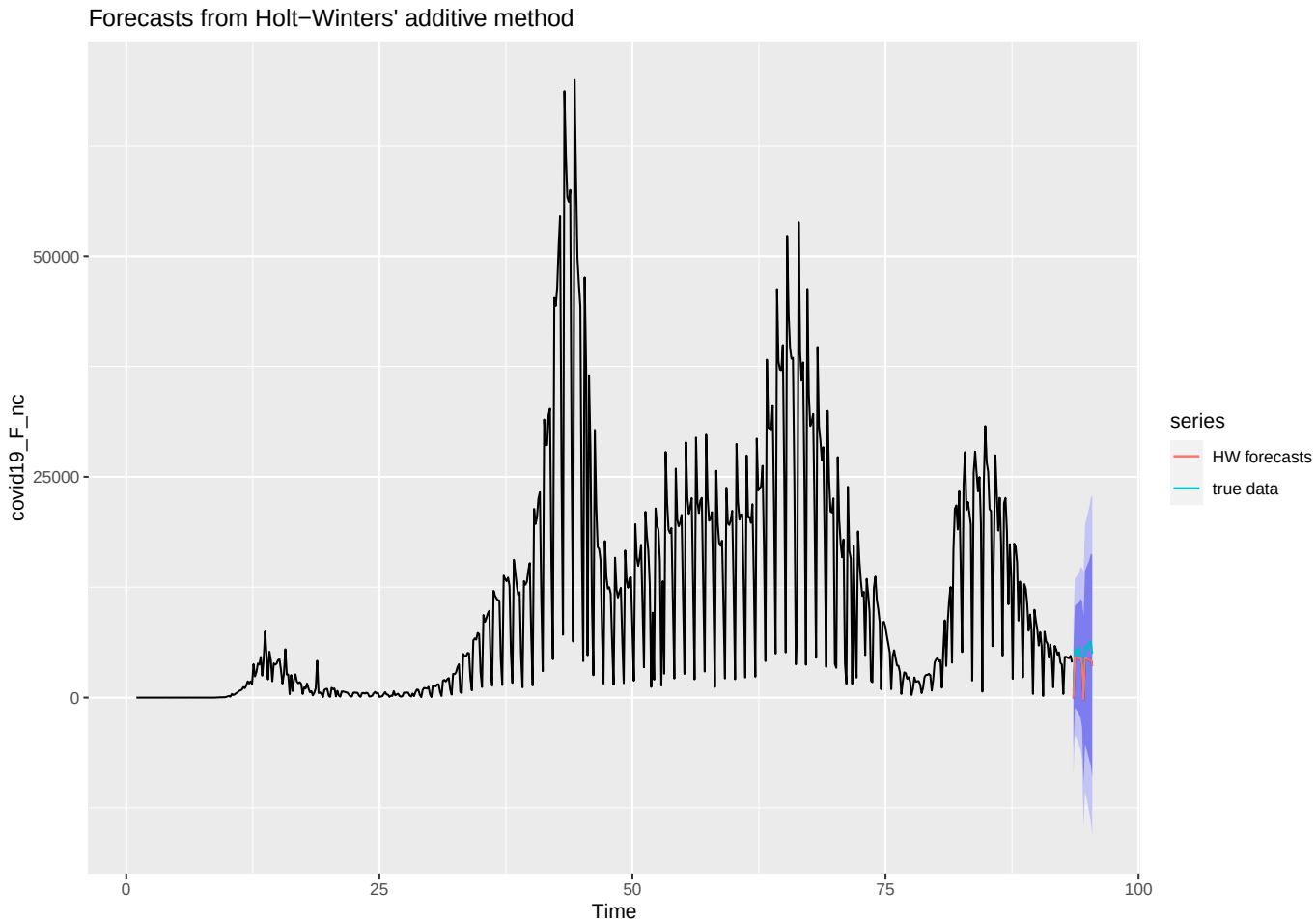
Exercice proposé :

- récupérer les données à jour pendant une séance de TD
- pour la séance de la semaine suivante, prévoir le nombre de cas pour les 7 prochains jours :
  - ils ont la semaine pour y travailler,
  - tester les différents modèles vus jusqu'ici,
  - choisir le meilleur modèle de façon adequat,
  - proposer la meilleure prédiction possible
- en début de séance suivante, on récupère les données à jour et chacun évalue la qualité de sa prédiction

Et on renouvelle cela à chaque séance.

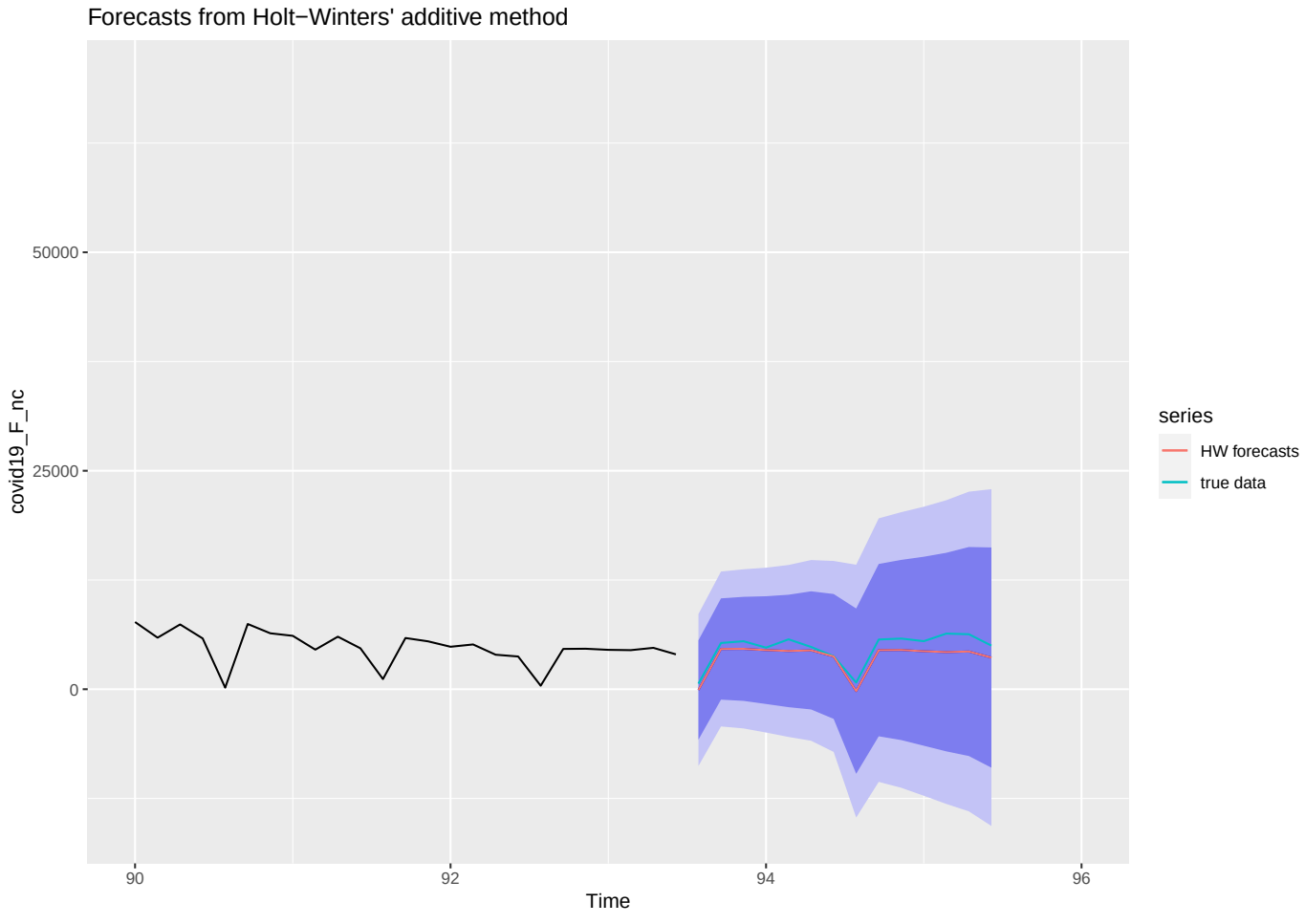
Résultat de l'exercice

```
covid19_2 = read.csv("2021-10-25-WHO-COVID-19-global-data.csv")
covid19_F_2=covid19_2[covid19_2$Country=="France",]
covid19_F_2_nc=ts(covid19_F_2$New_cases,freq=7)
prev=forecast(hw(covid19_F_nc),h=14)
autoplot(prev) + autolayer(tail(covid19_F_2_nc,14), series="true data")+
  autolayer(prev$mean, series="HW forecasts")
```



Résultat de l'exercice

```
autoplot(prev,xlim=c(90,96)) + autolayer(tail(covid19_F_2_nc,14), series="true
  autolayer(prev$mean, series="HW forecasts")
```



Résultat de l'exercice

- la prédiction du nombre de nouveaux cas de covid19 à court terme (une à deux semaines) est aisée
- la prédiction à long terme est plus complexe (effet *vague*) et demanderait à intégrer de nombreuses covariables extérieures (politique sanitaire, données de vaccination, . . .)

Retour des étudiants

- tout de suite plus intéressant que la concentration en co2 du volcan Mauna Loa
- se rendent compte que les méthodes (même simples) qu'on leur enseigne sont :
  - efficaces (*à court terme*)
  - directement applicables à *la* préoccupation du moment
- compréhension des enjeux de la discipline/problématique (intérêts de la prédiction court/long terme)

Pour aller plus loin

Sous la forme de projets étudiants (non mis en place) :

- récupérer des facteurs externes pouvant influencer le nombre de contaminations
- les intégrer dans leur modélisation
- essayer de prévoir ces fameuses vagues
- utiliser les données à l'échelle mondiale
- ...

Matériel pédagogique

Les slides du cours sont disponibles ici :
http://eric.univ-lyon2.fr/~jjacques/enseignement.html

A vous de jouer . . .