

Veille sur les modèles de series temporelles

Contexte professionnel :

Vous travaillez pour un fournisseur d'électricité. Votre manager vous demande pour la région des Hauts de France d'être capable de prédire à court terme (sur la semaine qui arrive) la consommation électrique. Il vous indique qu'il pourrait être intéressant d'utiliser d'autres variables exogènes comme la température.

Je vais faire des recherches sur différents modèles utilisés pour les times series et tenter de déterminer quel modèle serai le plus pertinent dans ce contexte.

D'après une analyse sur les données sur la consommation pour la région Hauts de France transmis par le fournisseur d'électricité, on travaille sur une série temporelle univariable et on observe une saisonnalité à l'année et à la semaine, une tendance et il serai utile de coupler ces données avec une variable exogène, la température.

Modèle ARIMA

ARIMA est l'acronyme de "AutoRegressive Integrated Moving Average". C'est une méthode populaire pour modéliser et prévoir des séries temporelles. ARIMA combine trois concepts principaux:

- **AutoRégression** (AR): Cela fait référence à la relation entre une observation à un moment donné et un certain nombre d'observations retardées. Par exemple, un AR(1) utiliserait la valeur précédente de la série pour prédire la valeur suivante, tandis qu'un AR(2) utiliserait les deux dernières valeurs.
- **Différenciation Intégrée** (I): La différenciation est l'action de prendre la différence consécutive d'une série temporelle. Cette étape vise à rendre la série stationnaire, c'est-à-dire à avoir une moyenne et une variance constantes dans le temps. La valeur "I" indique le nombre de différenciations nécessaires pour rendre la série stationnaire.
- **Moyenne Mobile** (MA): Ce composant utilise la dépendance entre une observation et une erreur résiduelle à partir d'une moyenne mobile appliquée à des observations lag. Par exemple, un MA(1) utiliserait l'erreur du pas de temps précédent pour prédire la valeur suivante.

Un modèle ARIMA est généralement noté comme ARIMA(p, d, q) où:

- **p** est l'ordre de la partie auto-régressive (AR).
- **d** est le degré de différenciation.
- **q** est l'ordre de la partie moyenne mobile (MA).

La sélection appropriée des paramètres p, d et q est essentielle pour obtenir un bon ajustement du modèle ARIMA à une série temporelle. Il existe plusieurs méthodes pour déterminer ces paramètres, notamment l'analyse des fonctions d'auto-corrélation (ACF) et d'auto-corrélation partielle (PACF), ainsi que l'utilisation d'approches automatiques ou de méthodes de validation croisée.

Une fois que le modèle est ajusté, il peut être utilisé pour faire des prédictions sur de futurs points de données ou pour comprendre la structure sous-jacente de la série temporelle en question.

<https://datascientest.com/arima-series-temporelles>

<https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>

<https://openclassrooms.com/fr/courses/4525371-analysez-et-modelisez-des-series-temporelles/5001226-les-processus-non-stationnaires-arima-et-sarima>
sources de confiance utilisées par les professionnels

https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
source communautaire

Exemple concret d'utilisation du modèle ARIMA :

Supposons qu'une entreprise de vente au détail souhaite prévoir les ventes mensuelles de l'un de ses produits phares pour les 12 prochains mois afin de gérer efficacement son stock.

L'entreprise dispose de données sur les ventes mensuelles de ce produit pour les cinq dernières années. Lorsqu'elle visualise ces données, elle remarque une tendance à la hausse au fil du temps, sans motif saisonnier apparent.

Grâce à ces prévisions, l'entreprise peut planifier son stock pour les prochains mois, évitant ainsi les ruptures de stock ou les surplus d'inventaire.

Par rapport à notre contexte, ARIMA peut gérer

- une série temporelle : univariée
- une saisonnalité : non
- une tendance : oui
- des variables exogènes : non

Modèle SARIMA

SARIMA signifie "Seasonal AutoRegressive Integrated Moving Average". C'est une extension du modèle ARIMA, spécialement conçue pour gérer la saisonnalité dans les séries temporelles.

En plus des composants AR, I, et MA du modèle ARIMA, SARIMA ajoute des termes saisonniers:

- **AutoRégression Saisonnière** (AR saisonnier): Cette partie capture la dépendance saisonnière dans les données. Par exemple, si une série montre une saisonnalité annuelle, un terme AR saisonnier pourrait capturer la relation entre une valeur et la valeur à la même période l'année précédente.
- **Différenciation Saisonnière Intégrée** (I saisonnier): Similaire à la différenciation intégrée dans ARIMA, mais appliquée à des intervalles saisonniers. Par exemple, si une série a une saisonnalité annuelle, une différenciation saisonnière pourrait impliquer de prendre la différence entre une valeur et la valeur il y a un an.
- **Moyenne Mobile Saisonnière** (MA saisonnier): Comme la composante MA de l'ARIMA, mais appliquée à la saisonnalité.

Un modèle SARIMA est généralement noté comme SARIMA(p, d, q)(P, D, Q)S, où:

- **p, d, q** sont les ordres de la partie non saisonnière (comme dans ARIMA).
- **P, D, Q** sont les ordres de la partie saisonnière.
- **S** est la période saisonnière (par exemple, 12 pour des données mensuelles avec saisonnalité annuelle).

L'idée est de combiner à la fois les structures de dépendance non saisonnières et saisonnières dans les données de séries temporelles pour obtenir un modèle plus complet. Comme pour l'ARIMA, la sélection des paramètres appropriés pour SARIMA est cruciale et peut être guidée par l'analyse des fonctions d'autocorrélation et d'autocorrélation partielle, ainsi que par d'autres méthodes.

SARIMA est particulièrement utile pour modéliser des séries temporelles qui présentent à la fois des tendances et des motifs saisonniers, comme les ventes mensuelles de certains produits qui pourraient augmenter pendant les vacances chaque année.

<https://datascientest.com/arima-series-temporelles>

<https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>

<https://openclassrooms.com/fr/courses/4525371-analysez-et-modelisez-des-series-temporelles/5001226-les-processus-non-stationnaires-arima-et-sarima>

Exemple concret d'utilisation du modèle SARIMA :

Une station touristique de montagne souhaite prévoir le nombre de visiteurs pour l'année à venir afin de mieux gérer ses ressources, son personnel et ses promotions.

La station dispose de données sur le nombre de visiteurs chaque mois pendant les dix dernières années. Lorsqu'elle visualise ces données, elle remarque non seulement une tendance à la hausse au fil des années (peut-être due à une popularité croissante ou à des campagnes marketing réussies), mais aussi une saisonnalité claire, avec des pics de fréquentation en hiver (pour le ski) et en été (pour la randonnée), et des creux au printemps et à l'automne.

Armée de ces prévisions, la station touristique peut mieux planifier ses besoins en personnel, décider des moments optimaux pour effectuer des réparations ou des améliorations, et concevoir des campagnes promotionnelles pour attirer davantage de visiteurs pendant les mois creux.

En résumé, SARIMA peut gérer une saisonnalité complexe, la tendance et faire des prévisions à court terme.

Par rapport à notre contexte, SARIMA peut gérer

- une série temporelle : univariée
- une saisonnalité : oui
- une tendance : oui
- des variables exogènes : non

Modèle VARIMA

VARIMA est l'acronyme de "Vector AutoRegressive Integrated Moving Average". C'est une extension du modèle ARIMA pour traiter les séries temporelles multivariées, c'est-à-dire lorsque vous avez plusieurs séries temporelles interdépendantes que vous souhaitez modéliser simultanément.

Décomposons les composants de VARIMA :

- **Vector** (Vecteur): Indique que le modèle traite plusieurs séries temporelles en même temps. Chaque série peut dépendre de ses propres valeurs passées ainsi que des valeurs passées des autres séries.

- **AutoRegressive** (AR): Comme dans ARIMA, cela fait référence à la dépendance d'une valeur à un moment donné par rapport à ses valeurs précédentes. Dans le contexte de VARIMA, chaque série temporelle peut également dépendre des valeurs passées des autres séries temporelles.

- **Integrated** (I): Comme dans ARIMA, cela se réfère à la différenciation nécessaire pour rendre une série temporelle stationnaire. Dans le contexte de séries temporelles multivariées, toutes les séries doivent généralement être différenciées le même nombre de fois.

- **Moving Average** (MA): Cela indique une dépendance entre une observation et une erreur résiduelle à partir d'une moyenne mobile appliquée à des observations précédentes. Comme avec la partie AR, dans le contexte de VARIMA, cette moyenne mobile peut dépendre des erreurs de toutes les séries.

VARIMA est adapté pour analyser et prévoir des systèmes de séries temporelles où les variables sont interdépendantes. Par exemple, supposons que vous ayez des données sur les ventes mensuelles et les dépenses publicitaires de plusieurs produits. Il est probable que les ventes d'un produit soient influencées non seulement par ses propres dépenses publicitaires passées mais aussi par les dépenses publicitaires des autres produits. VARIMA permettrait de capturer ces interdépendances complexes.

Exemple concret d'utilisation du modèle VARIMA :

Supposons qu'une banque centrale souhaite analyser et prévoir simultanément le produit intérieur brut (PIB) d'un pays, le taux d'inflation et le taux de chômage. Ces trois variables sont interdépendantes : par exemple, une politique monétaire expansionniste pourrait stimuler le PIB tout en augmentant l'inflation.

La banque a des données trimestrielles sur le PIB, l'inflation et le chômage pour les vingt dernières années.

Avec ces prévisions en main, la banque centrale est mieux équipée pour prendre des décisions de politique monétaire. Par exemple, si le modèle prévoit une forte croissance du PIB associée à une hausse de l'inflation, la banque pourrait envisager de resserrer sa politique monétaire pour prévenir une surchauffe économique.

Par rapport à notre contexte, VARIMA peut gérer

- une série temporelle : multivariée
- une saisonnalité : oui
- une tendance : oui
- des variables exogènes : non

Modèle VARMAX

VARMAX est une extension du modèle VAR (modèle auto-régressif vectoriel) pour les séries temporelles multivariées. Le modèle VARMAX inclut des composants pour tenir compte des erreurs autorégressives (AR), des moyennes mobiles (MA), et des effets exogènes (X). Les composantes du VARMAX sont :

- **VAR**: Cela signifie que chaque série est modélisée comme une combinaison linéaire de ses propres valeurs retardées et des valeurs retardées des autres séries.
- **MA**: Cela signifie que les erreurs (chocs) dans la série sont modélisées comme une combinaison linéaire des erreurs passées.
- **X**: Se réfère aux variables exogènes. Ce sont des variables qui peuvent influencer les séries temporelles mais qui ne sont pas influencées par elles. Cette option X peut être ajoutée sur différentes extensions du modèle ARMA afin de gérer les variables exogènes.

<https://www.youtube.com/watch?v=WSgpfNdKQWg>

<https://www.youtube.com/watch?v=8EBxw82PKUY&t=12s>

[https://www.statsmodels.org/dev/examples/notebooks/generated/](https://www.statsmodels.org/dev/examples/notebooks/generated/statespace_varmax.html)
[statespace_varmax.html](https://www.statsmodels.org/dev/examples/notebooks/generated/statespace_varmax.html)

Exemple concret d'utilisation du modèle VARMAX :

Imaginons une économie nationale où l'on souhaite prévoir simultanément le taux de croissance du PIB, le taux d'inflation, et le taux de chômage. Ces trois séries temporelles sont probablement interdépendantes. De plus, elles pourraient être influencées par des facteurs exogènes tels que les taux d'intérêt fixés par la banque centrale ou les fluctuations des prix du pétrole sur les marchés mondiaux.

Un économiste pourrait utiliser un modèle VARMAX pour intégrer ces informations. Les composants VAR du modèle captureraient les interdépendances entre le PIB, l'inflation, et le chômage, le composant MA pourrait aider à modéliser les chocs ou les erreurs, et le composant X permettrait d'inclure l'effet des taux d'intérêt et des prix du pétrole comme variables exogènes.

Grâce à ce modèle, l'économiste pourrait mieux comprendre les dynamiques complexes à l'œuvre dans l'économie et faire des prévisions plus précises sur l'avenir de ces variables clés.

Par rapport à notre contexte, VARMAX peut gérer

- une serie temporelle : multivariée
- une saisonnalité : oui
- une tendance : oui
- des variables exogènes : oui

Modèle PROPHET

Prophet est un outil de prévision de séries temporelles développé par Facebook. Il a été conçu pour produire des prévisions fiables pour un large éventail de séries temporelles des affaires, tout en étant flexible pour gérer les spécificités des données (comme les tendances, la saisonnalité, les jours fériés, etc.).

Voici quelques caractéristiques clés de Prophet:

- **Traitement de la saisonnalité**: Prophet est capable de gérer la saisonnalité quotidienne, hebdomadaire et annuelle. Il peut même s'adapter aux changements de saisonnalité au fil du temps.
- **Jours fériés et événements spéciaux**: Vous pouvez fournir à Prophet une liste de jours fériés, et le modèle les prendra en compte. C'est particulièrement utile pour les prévisions dans les domaines où les jours fériés ont un impact notable sur les mesures, comme le commerce de détail.
- **Tendances flexibles**: La structure de Prophet lui permet de s'adapter à la fois aux tendances qui sont linéaires ou non linéaires, et cela peut être ajusté selon le besoin.
- **Absence de données**: Prophet peut gérer les lacunes dans les données et les jours où il n'y a pas de mesures.
- **Points de changement**: Prophet détecte automatiquement les points de changement où les tendances changent. Vous pouvez également spécifier manuellement ces points si vous avez une connaissance préalable.
- **Facilité d'utilisation**: Prophet est conçu pour que les analystes et les développeurs, même sans une expertise approfondie en prévision de séries temporelles, puissent produire des prévisions de haute qualité.

<https://facebook.github.io/prophet/>

<https://ledatascientist.com/facebook-prophet-la-prevision-a-grande-echelle/>

Exemple concret d'utilisation de PROPHET :

On souhaite prévoir les ventes quotidiennes d'une grande entreprise de commerce électronique pour les 6 prochains mois afin d'aider l'entreprise à gérer ses stocks, à planifier ses campagnes marketing, et à déterminer ses besoins en personnel.

On dispose de données de ventes quotidiennes pour les trois dernières années. Ces données montrent des pics lors des soldes du Black Friday, pendant la période des fêtes de fin d'année, et de petites fluctuations hebdomadaires avec un pic le week-end.

Grâce à Prophet, on peut prévoir des ventes pour les 6 prochains mois, avec des estimations de la tendance, de la saisonnalité hebdomadaire, et de l'impact des jours fériés. Avec ces informations, l'entreprise peut planifier ses campagnes, gérer son stock, et anticiper ses besoins en personnel de manière plus efficace.

Par rapport à notre contexte, PROPHET peut gérer

- une série temporelle : univariée
- une saisonnalité : oui
- une tendance : oui
- des variables exogènes : oui

Modèle XGBOOST

Bien que XGBoost soit généralement associé à des problèmes de classification et de régression standard, il peut également être adapté et utilisé pour la prévision de séries temporelles.

Avantages de l'utilisation de XGBoost pour les séries temporelles :

- **Interprétabilité** : XGBoost fournit des scores d'importance des caractéristiques, ce qui peut aider à comprendre quelles caractéristiques (par exemple, quels décalages ou motifs saisonniers) sont les plus informatifs pour la prévision.
- **Flexibilité** : XGBoost peut intégrer de multiples types de données et de caractéristiques, ce qui est utile lorsque l'on travaille avec des séries temporelles complexes ou lorsqu'il existe des données exogènes pertinentes.
- **Performance** : Comme dans d'autres domaines, XGBoost peut offrir de bonnes performances en matière de prévision de séries temporelles, surtout si les données sont bien préparées et que le modèle est bien ajusté.

Exemple concret d'utilisation de XGBOOST :

On a une série temporelle univariée représentant les ventes mensuelles d'un produit. XGBoost, en transformant le problème en un problème de régression supervisée, peut prédire les ventes futures.

Par rapport à notre contexte, XGBOOST peut gérer

- une serie temporelle : univariée
- une saisonnalité : oui (avec des fonctions)
- une tendance : oui (avec des fonctions)
- des variables exogènes : oui

CONCLUSION

Dans notre contexte, nous avons une serie temporelle univariée, différentes saisonnalités, une tendance, et une variable exogène à traiter. C'est pourquoi les modèles PROPHET et SARIMAX (association du modèle SARIMA avec l'option X) semblent être les plus faciles à mettre en œuvre et les plus adapter au contexte car ils sont capables de gérer toutes ces possibilités.