Université de Technologie de Compiègne

SY09 - Science des données - P23

US Drought

CHEN Wenlong - FEGHOUL Rayan

supervisé par

Sylvain ROUSSEAU, Haifei ZHANG

1 Introduction

La sécheresse est un phénomène climatique qui peut avoir des conséquences dévastatrices sur l'agriculture, les écosystèmes et les communautés humaines. Les États-Unis ont été touchés par de nombreuses sécheresses au cours des dernières décennies, ce qui a entraîné des pertes économiques et environnementales importantes. L'analyse d'un jeu de données sur la sécheresse aux États-Unis peut fournir des informations précieuses sur la fréquence, la durée et l'intensité de ce phénomène, ainsi que sur ses effets sur différents secteurs de l'économie et de la société. En utilisant des techniques d'analyse statistique et de visualisation de données, il est possible de découvrir des tendances, des corrélations et d'élaborer des modèles prédictifs dans les données, ce qui peut aider à mieux comprendre la sécheresse et à élaborer des stratégies pour y faire face.

Dans cette étude, nous nous sommes intéressés à l'évolution des niveaux de sécheresse aux États-Unis en utilisant le jeu de données "US Drought Monitor" (USDM). Tout d'abord, nous avons procédé à une analyse approfondie du jeu de données. Ensuite, nous avons utilisé différents modèles d'apprentissage pour la prédiction de la sécheresse. Enfin, nous concluons en résumant nos résultats et en énonçant des suggestions d'améliorations possibles.

2 Géneralités sur les séries temporelles

2.1 Définitions

Tout d'abord, nous allons définir les principales notions relatives aux séries temporelles. Une série temporelle est un ensemble d'observations dans lequel chaque observation est associée à un instant unique dans le temps. Les séries temporelles se distinguent par les dépendances qui existent entre les observations successives. Les modèles auto-régressifs ont été conçus dans le but de modéliser ces relations en utilisant les observations passées afin de prédire les valeurs futures. Nous allons maintenant énoncer trois définitions élémentaires :

Définition 1 (Bruit blanc) : Un processus bruit blanc est une séquence de variables aléatoires non-corrélés entre eux, de moyenne nulle et de variance constante.

Definition 2 (Auto-regressif) : Un processus $(X, t \in \mathbb{Z})$ auto-régressif est une séquence de variables aléatoires stationnaires, où chaque valeur dépend de la valeur précédente et d'un terme d'erreur qui est un processus de type bruit blanc. Il est décrit par l'équation suivante :

$$X(t) = \alpha X(t-1) + U(t) \tag{1}$$

où α est le paramètre du processus et U est un processus bruit blanc. Ce processus est stationnaire si et seulement si $|\alpha| < 1$.

Définition 3 (Stationnarité) : Une série temporelle est dite stationnaire lorsque ses propriétés statistiques ne dépendent pas de la valeur absolue de la variable temporelle t. En d'autres termes, ces propriétés ne sont pas affectées par un décalage de la série dans le temps. Toutefois, une série présentant une tendance ou une composante saisonnière n'est généralement pas stationnaire. Par conséquent, on peut conclure qu'il existe une autocorrélation entre les observations dans de telles séries.

Une méthode courante pour rendre les séries temporelles stationnaires consiste à les différencier. Cette opération transforme les séries en de nouvelles séries où chaque observation représente le changement par rapport à l'observation précédente. Supposons une série uni-varié $X(t), t \in [1, n]$. La différenciation d'ordre 1 de X est définie de la manière suivante :

$$\Delta X(t) = X(t) - X(t-1) \tag{2}$$

X est dite intégré d'ordre d si elle est non stationnaire et devient stationnaire juste après d différentiations. Et La différentiation de X d'ordre d est définit comme suit :

$$\Delta^d X(t) = \Delta(\Delta^d X(t)) \tag{3}$$

2.2 Modèles auto-régressif uni-variés

Les modèles univariés sont utilisés pour prédire une seule série temporelle en se basant sur les valeurs passées de cette série. Ces modèles examinent les dernières valeurs de la série afin d'estimer les valeurs futures.

Le modèle AR (Auto Regressive) considère une série temporelle univariée stationnaire comme une fonction linéaire de ses p valeurs précédentes. Formellement, l'équation du modèle AR(p) est la suivante :

$$Y(t) = \alpha_0 + \alpha_1 Y(t-1) + \ldots + \alpha_p Y(t-p) + U(t) \tag{4}$$

où p est l'ordre du modèle, $\alpha_0, \alpha_1, \ldots, \alpha_p$ sont les paramètres du modèle, $U(t)$ est le terme d'erreur suivant un processus bruit blanc.

Le modèle MA (Moving Average) a la même structure que le modèle AR, mais en considérant les termes d'erreurs au lieu des valeurs précédentes de la série. Le modèle MA(q) peut être exprimé comme suit :

$$Y(t) = \theta_0 + \theta_1 U(t-1) + \ + \theta_p U(t-q) + U(t). \tag{5}$$

Le modèle ARMA (p,q), combine les deux processus AR(p) et MA(p) en considérant à la fois les termes d'erreurs et les valeurs précédentes de la série :

$$Y(t) = y_0 + \sum_{i=1}^{p} \alpha_i Y(t\text{-}i) + \sum_{i=1}^{q} \theta_i U(t\text{-}i) + U(t) \qquad (6)$$

Où y0, $\alpha_0 ... \alpha_p$, et $\theta_0 ... \theta_q$ sont les paramètres du modèle.

Le modèle ARIMA (p,d,q) est plus adapté pour les séries temporelles non stationnaires. Il combine le modèle ARMA(p,q) avec une transformation de différenciation d'ordre d. Cela implique de calculer d fois les différences entre les observations consécutives afin de rendre la série stationnaire, avant d'appliquer le modèle ARMA sur cette série différenciée.

3 Analyse de la série temporelle

3.1 Jeu de données

Le jeu de données USDM utilise un système de cinq catégories pour classer la gravité de la sécheresse dans chaque État. Ces catégories reposent sur divers indicateurs tels que les précipitations, l'humidité du sol et l'écoulement des cours d'eau. Voici les explications pour chacune des quatre catégories :

- D0 (Anormalement sec) : Cette catégorie est utilisée pour indiquer les zones qui connaissent une certaine sécheresse, mais qui ne connaissent pas encore de sécheresse.
- D1 (Sécheresse modérée) : Cette catégorie est utilisée pour indiquer les zones qui connaissent des conditions de sécheresse modérée.
- D2 (Sécheresse sévère) : Indique les zones qui connaissent des conditions de sécheresse sévère.
- D3 (Sécheresse extrême) : Indique les zones qui connaissent des conditions de sécheresse extrême.
- D4 (Sécheresse exceptionnelle) : C'est la catégorie la plus sévère et représente une sécheresse exceptionnelle.

Afin d'analyser les données de manière claire, nous nous concentrons uniquement sur les données de l'État de Californie après l'an 2000. Nous constatons sur la figure 4 que lors de certaines années spécifiques, l'État a connu des périodes de sécheresse importantes, telles que 2001 et 2008, avec plus de 50% de la zone touchée par la sécheresse de catégorie D4.

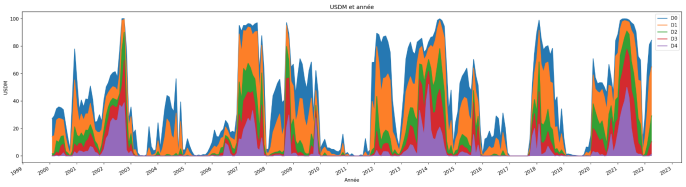

FIGURE 1 – Visualisation données(D0 :bleu, D1 :orange, D2 :vert, D3 :rouge, D4 :violet).

Pour simplifier le problème, nous nous concentrons sur l'analyse d'une série temporelle univariée. L'analyse de séries temporelles multivariées est complexe, c'est pourquoi nous utilisons l'indice DSCI (Drought Severity and Coverage Index) pour représenter le niveau de sécheresse. Voici comment le DSCI est calculé :

$$DSCI = 1 * D0 + 2 * D1 + 3 * D2 + 4 * D3 + 5 * D4 \quad (7)$$

DSCI mesure l'intensité globale de la sécheresse d'une zone. Si une zone a plus de valeurs élevées pour D4, cela indique une sécheresse plus sévère. Par conséquent, nous ajoutons des coefficients devant D0 à D4 pour refléter l'importance de la sécheresse. Plus la sécheresse est sévère, plus le poids ou le coefficient correspondant est élevé.

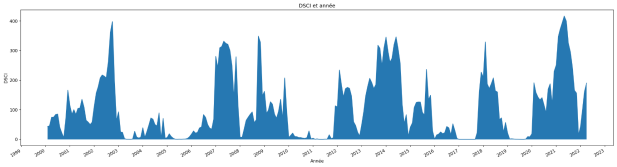

FIGURE 2 – Evolution du DSCI de l'Etat de Californie.

3.2 Analyse de la stationnarité

On va tout d'abord vérifier si notre jeu de données est stationnaire ou non. Pour cela, nous allons étudier l'autocorrélation entre les observations.
Comme le montre la figure 3, il existe une autocorrélation entre les observations, ce qui indique que notre série temporelle n'est pas stationnaire. Par conséquent, nous allons rendre notre série stationnaire en utilisant la méthode de différenciation.

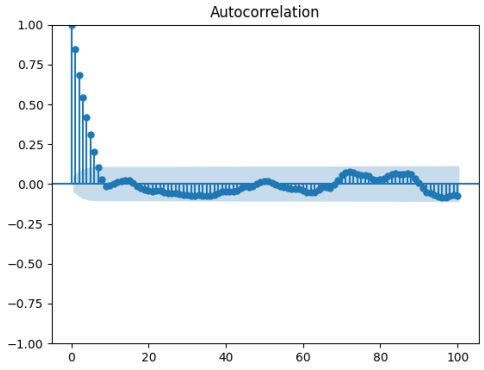

FIGURE 3 – Auto-corrélation.

On peut donc observer sur la figure 4 qu'il n'y a plus d'auto-corrélation. Notre série est devenue stationnaire.

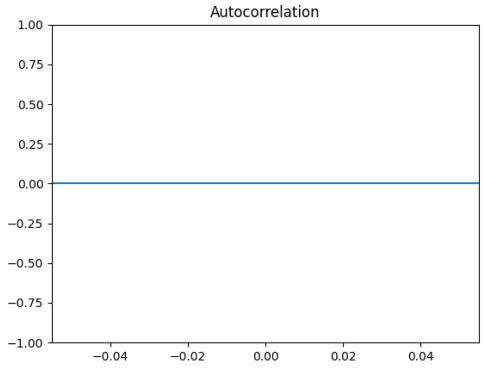

FIGURE 4 – Auto-corrélation après une différenciation.

4 Modèle de prédiction

4.1 Evaluation des modèles

Pour chaque modèle, nous allons l'entraîner en utilisant 80% des données d'apprentissage. Ensuite, nous utiliserons ce modèle pour prédire les 20% restants afin d'évaluer sa performance.

Pour évaluer la qualité de nos modèles, nous allons utiliser l'erreur quadratique moyenne (MSE pour Mean Squared Error en anglais).

$$MSE = (1/n) * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (8)$$

où n est le nombre d'échantillons ou d'observations, y_i représente la valeur réelle de la variable cible pour l'échantillon i. $\hat{y}_i$ représente la valeur prédite de la variable cible pour l'échantillon i.

Elle est donc calculée en prenant la moyenne des carrés des différences entre les valeurs réelles et les valeurs prédites par le modèle. Plus la valeur de MSE est faible, meilleure est la performance du modèle de régression.

4.2 Modèle auto-régressif ARIMA

Nous utilisons le modèle ARIMA (AutoRegressive Integrated Moving Average) en combinaison avec la méthode de décomposition STL (Seasonal and Trend decomposition using Loess) pour améliorer la précision des prévisions des séries temporelles.

Voici comment cette méthode fonctionne :

1. Effectuer la décomposition STL : Utilisez la méthode de décomposition STL (Seasonal and Trend decomposition using Loess) pour séparer la série temporelle en ses composantes saisonnières, tendancielles et résiduelles.
2. Soustraire la composante saisonnière : Une fois que les composantes saisonnières obtenues, il faut les soustraire à la série temporelle initiale. Cela permet de désaisonnaliser les données et de se concentrer uniquement sur la variation non saisonnière.
3. Modélisation du modèle ARIMA : Utilisez les données désaisonnalisées obtenues à l'étape précédente pour ajuster un modèle ARIMA approprié.
4. Ajustement du modèle ARIMA : Appliquez le modèle ARIMA aux données désaisonnalisées et ajustez-le aux observations historiques. Cela permet d'estimer les coefficients du modèle ARIMA.
5. Prévision : Utilisez le modèle ARIMA ajusté pour effectuer des prévisions sur les données désaisonnalisées. Les prévisions fourniront une estimation de la variation non saisonnière future.

Il est donc important de choisir les paramètres adéquats pour notre modèle. De plus, étant donné que l'on a soustrait à la série temporelle sa saisonnalité, il convient de revoir l'ordre de différenciation. Après vérification, il s'avère qu'une différenciation d'ordre 1 est suffisante pour rendre notre jeu de données stationnaire. Pour choisir les paramètres p et q de notre modèle, on va utiliser les critères d'information d'Akaike (AIC) et de Bayes (BIC). Ce sont des métriques statistiques utilisées pour comparer différents modèles statistiques. Ces critères prennent en compte à la fois la qualité de l'ajustement du modèle aux données et la complexité du modèle. Les critères AIC et BIC sont calculés pour évaluer les ordres des modèles ARMA. Nous obtenons enfin le train AIC (10, 0), le train

BIC (10, 0) comme meilleurs paramètres. Donc p est égale à dix et q est égale à zéro.

Voici ce que donne STL sur notre jeu de données :

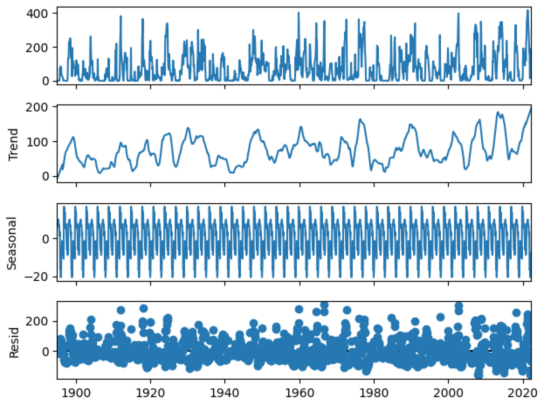

FIGURE 5 – Tendance, saisonnalité, résidus.

Nous avons entraîné le modèle selon les paramètres précedemment définis.

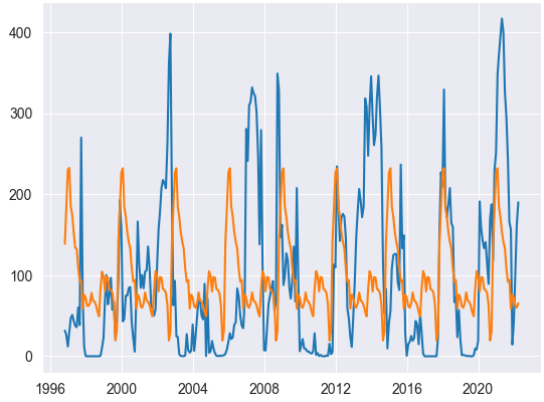

FIGURE 6 – STL-Arima

On peut constater sur la figure 6 que le modèle ARIMA présente des limitations en termes de prévision à long terme avec une MSE de 11980. Cela en raison du caractère excessivement aléatoire du bruit dans nos données d'origine. Cela conduit à l'apparition de valeurs extrêmes dans certaines années, ce qui n'est pas bien pris en compte par le modèle ARIMA.

Pour faire une prédiction à court terme, nous suivons les étapes suivantes : d'abord, nous prédisons uniquement la valeur DSCI du mois suivant (par rapport aux données d'entrainement). Ensuite, nous intégrons la valeur réelle de la prédiction dans l'ensemble d'apprentissage et l'utilisons pour prédire le mois suivant. Dans ce cas, nous avons une MSE de 2691.

FIGURE 7 – STL-Arima

Nous pouvons maintenant étudier les résidus de notre modèle :

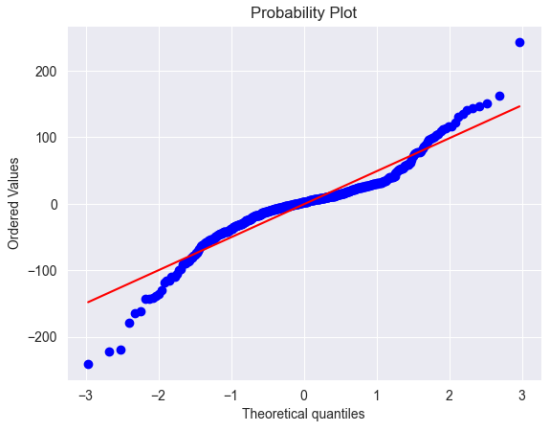

FIGURE 8 – Q-Q plot

On peut constater que le QQ plot des résidus est à peu près sur la ligne diagonale, cela suggère que les résidus suivent une distribution normale. Et par conséquent qu'ils sont donc un bruit blanc.

4.3 Modèles d'apprentissage automatique

4.3.1 Prétraitement des données

Nous allons utiliser deux modèles : XGBoost et Random Forest. XGBoost (eXtreme Gradient Boosting) est un modèle basé sur l'approche de boosting, qui consiste à construire un ensemble de modèles faibles (généralement des arbres de décision) de manière séquentielle.
Random Forest, en revanche, est basé sur une approche d'ensemble appelée bagging. Il construit également un ensemble d'arbres de décision, mais chaque arbre est construit de manière indépendante en utilisant des sous-échantillons aléatoires de l'ensemble de données d'entraînement.

Lorsque nous utilisons des méthodes d'apprentissage automatique pour traiter des séries chronologiques, il est essentiel de convertir ces données en un format adapté à l'apprentissage. Pour ce faire, nous appliquons une trans-

formation en utilisant une technique appelée "lagging", où chaque observation de la série temporelle est convertie en une ligne de données comportant des variables explicatives correspondant à des observations antérieures. Par exemple, nous pouvons inclure des valeurs remontant à un an ou à trois ans dans la série temporelle. Cette approche permet aux modèles d'exploiter les informations historiques et d'apprendre à partir de celles-ci, facilitant ainsi les prédictions futures.

Parallèlement, nous procédons également à la conversion de la variable temporelle en une représentation catégorielle en utilisant un encodage "one-hot". Cela signifie que chaque instant de temps est représenté par un vecteur binaire où une seule valeur est activée (1) et toutes les autres sont désactivées (0). Bien que cette approche puisse entraîner une augmentation de la dimensionnalité des données lorsque de nombreuses catégories temporelles sont présentes, l'encodage "one-hot" s'avère efficace pour les modèles basés sur des arbres tels que XGBoost.

De plus, nous utilisons une fenêtre glissante de longueur trois pour capturer les informations de tendance des trois mois précédents. Pour cela, nous ajoutons la moyenne des valeurs de cette fenêtre en tant que variable explicative supplémentaire (colonne "Moyenne"). Voici une représentation visuelle sans le codage one-hot pour plus de clarté :

1mago	2mago	3mago	1yago	2yago	3yago	moyenne
111.5	11.1	0.2	0.0	35.2	0.0	40.933333
132.3	111.5	11.1	0.0	76.7	0.0	84.966667
131.2	132.3	111.5	0.0	39.3	0.0	125.000000
205.7	131.2	132.3	0.0	32.9	0.0	156.400000
231.7	205.7	131.2	0.0	29.6	0.0	189.533333
...	...	...	...	...	...	...
156.3	165.6	242.5	230.0	8.3	57.1	188.133333
14.5	156.3	165.6	251.3	18.9	21.4	112.133333
43.4	14.5	156.3	347.4	191.1	1.2	71.400000
104.0	43.4	14.5	372.3	158.6	1.8	53.966667
161.6	104.0	43.4	394.5	142.8	1.2	103.000000

FIGURE 9 – Transformation des données

En utilisant ces techniques d'apprentissage automatique, nous cherchons à exploiter les relations complexes et non linéaires entre les données passées et futures, ainsi que les informations de tendance et de saisonnalité, afin d'améliorer la précision des prévisions pour les séries chronologiques.

4.3.2 XGBoost

Après avoir entraîné le modèle et ajusté les paramètres pour obtenir le modèle optimal à l'aide de la méthode GridSearchCV de la bibliothèque sklearn, nous obtenons une erreur quadratique moyenne (MSE) de 3246.

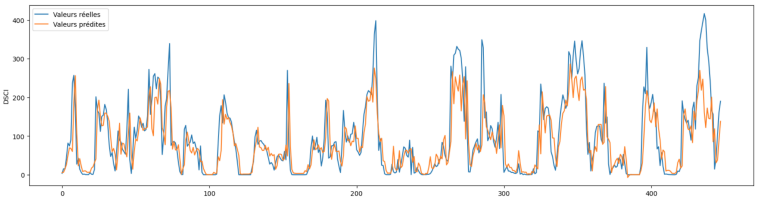

FIGURE 10 – Prédiction sur l'ensemble à l'aide de XgBoost (valeurs prédites en orange)

Nous souhaitons également déterminer les variables qui contribuent le plus au modèle, comme illustré dans la figure ci-dessous :

FIGURE 11 – Importance des variables d'entraînement

Il est clair que la variable principale pour prédire est la valeur DSCI d'il y a un mois ("1mago"), tandis que la variable "Moyenne" contribue le moins.

4.3.3 Random Forest

Après avoir entraîné le modèle et ajusté les paramètres pour obtenir le modèle optimal à l'aide de la méthode GridSearchCV de la bibliothèque sklearn, nous obtenons une erreur quadratique moyenne (MSE) de 2940. Nous pouvons observer la prédiction sur la figure 12.

De plus, nous avons déterminer les variables qui contribuent le plus au modèle. La variable principale pour prédire est aussi le DSCI d'il y a un mois.

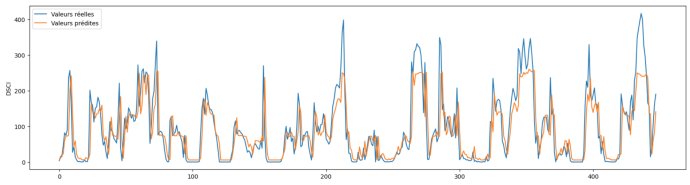

FIGURE 12 – Prédiction à l'aide de Random Forest (valeurs prédites en orange)

4.4 Comparaison des modèles

Voici un tableau des MSE en fonction du modèle choisi :

Modèle	MSE
STL-Arima	2691
XGBoost	3246
Random Forest	2940

Dans notre comparaison entre ARIMA, XGBoost et Random Forest, le modèle ARIMA s'est avéré être le meilleur en termes d'erreur quadratique moyenne (MSE). Cependant, il convient de noter que le modèle ARIMA présente une limitation importante lorsqu'il s'agit de faire des prédictions à long terme.

XGBoost est sensible aux changements et aux irrégularités des données, tandis que Random Forest est plus robuste et moins sensible aux variations. Cependant, les deux modèles ont tendance à sous-estimer les valeurs extrêmes, ce qui limite leur capacité à atteindre les points les plus élevés de la distribution des données. Il est essentiel de prendre en compte ces caractéristiques lors du choix du modèle en fonction des données et des objectifs spécifiques.

5 Conclusion et travaux futures

Dans cette étude portant sur la prédiction du niveau de sécheresse en Californie, aux États-Unis, nous avons utilisé les modèles ARIMA (en combinaison avec STL), XGBoost et Random Forest. Après évaluation, nous avons constaté que le modèle ARIMA a donné les meilleurs résultats.

Cependant, nous ne pouvons faire des prédictions qu'en nous basant sur les valeurs historiques de DSCI. Si nous sommes en mesure de fournir davantage d'informations telles que les précipitations locales, la température et le climat, nous pourrions améliorer la précision des prévisions. L'inclusion de ces variables supplémentaires permettrait d'enrichir le modèle et de mieux capturer les facteurs environnementaux qui influencent le niveau de sécheresse.

À l'avenir, notre objectif est d'explorer l'utilisation de méthodes d'apprentissage en profondeur, telles que les réseaux de neurones récurrents (RNN) et les réseaux LSTM (Long Short-Term Memory) améliorés, afin d'améliorer la prédiction des séries chronologiques. Ces approches permettent au modèle d'apprendre des motifs complexes et de mieux saisir les dépendances à long terme présentes dans les données temporelles, ce qui peut conduire à des prédictions plus précises et fiables.

Références

[1] Youssef HMAMOUCHE. *Prediction of Large Time Series.* https://hal.science/tel-02448325/document

[2] https://people.duke.edu/~rnau/411arim3.html

[3] https://www.kaggle.com/c/favorita-grocery-sales-forecasting/discussion/47582

[4] https://people.duke.edu/~rnau/411arim3.html

[5] https://zhuanlan.zhihu.com/p/67832773