

Web Traffic Prediction/ Forecasting

Analyse séquentielle de données
Cédric Campos Carvalho

Table des matières



01

Données

Lecture, prétraitement

02

Statistiques

Corrélation, distribution

03

Premier modèle

SARIMA, GARCH

04

Wavenet

Principes de base

05

Second modèle

Wavenet

06

Conclusion

Résultats, améliorations

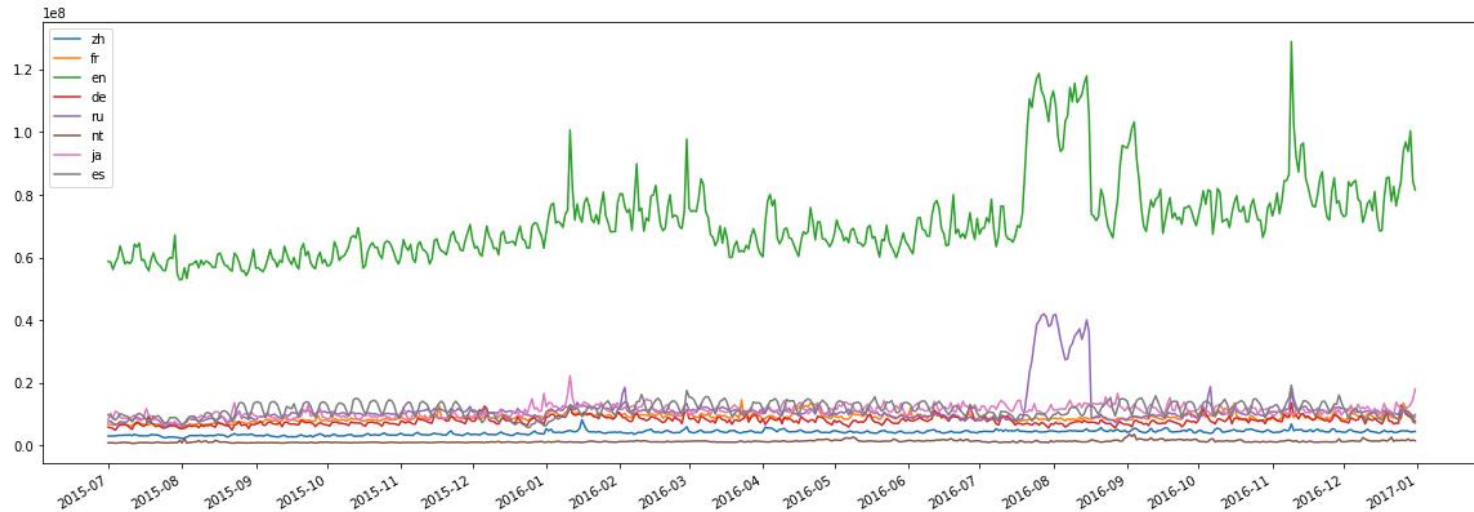
Données (1)

- Compétition Google (25'000\$)
- Prédiction du trafic web sur 140'000 articles Wikipédia
- Solution du gagnant:
 - Transformation $\ln(1 + p)$.
 - Attention based RNN
 - Encoder/Decoder, cuDNN GRU.

Données	
Article	Nom de l'article
Domaine	Mandarin (zh), Français (fr), Anglais (en), Allemand (de), Russe (ru), Northern Territory of Australia (nt), Japonais (ja), Espagnol (es)
Agent	Spider, vrai trafic
Type d'accès	Desktop, all access, mobile
Données temporelles (journalière)	Nombre de visites (Juillet 2015 à Septembre 2017)

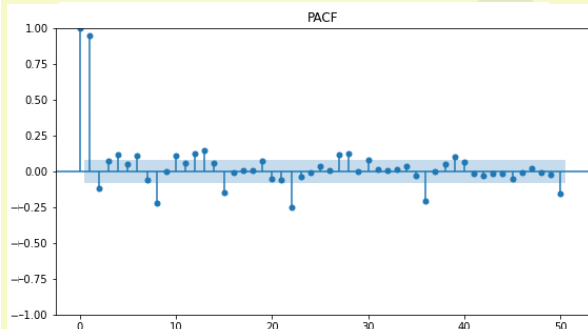
Données (2)

- Données sommées par domaine
- Saut pendant l'été 2016 (Jeux Olympiques ?)
- Série sélectionnée : EN



Statistiques (1)

Étude statistique



Distribution

Non normale, alpha
continuous

Outliers

Valeurs extrêmes
(événement)

Saisonalité

Hebdomadaire

Corrélation

Corrélation partielle
forte à $(t-1)$

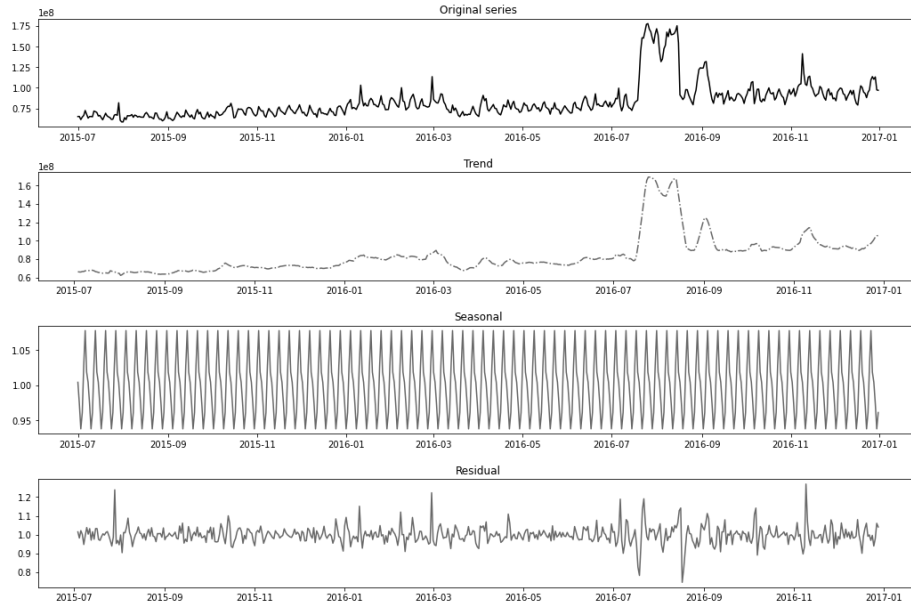
Statistiques (2)

- Non stationnaire

- Augmented Dickey-Fuller : $p_1 = 0.14$
- Kwiatkowski-Phillips-Schmidt-Shin : $p_2 = 0.01$

$$\frac{p_1 + p_2}{2} > \alpha$$

- Décomposition multiplicative



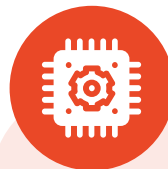
Premier modèle (1)



AR, MA, ARIMA ($p=1, d=1, q=1$)

- Pas de saisonnalité
- « One-step forecast » correcte
- « Dynamic forecast » peu de sens
- Erreur niveau des événements

Vs

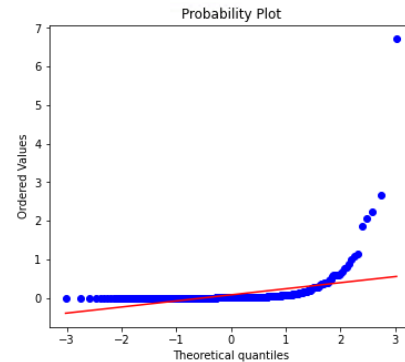
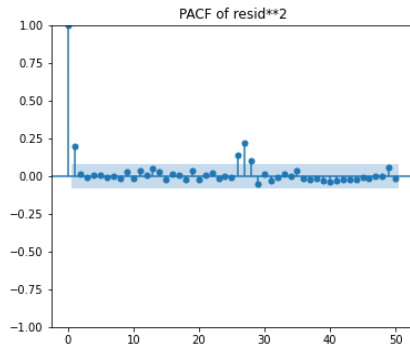
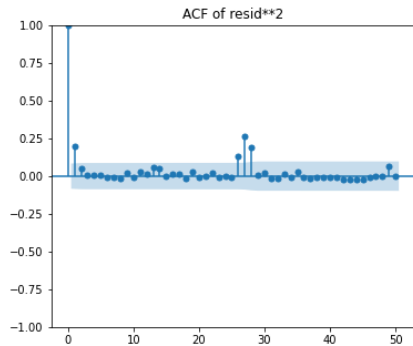
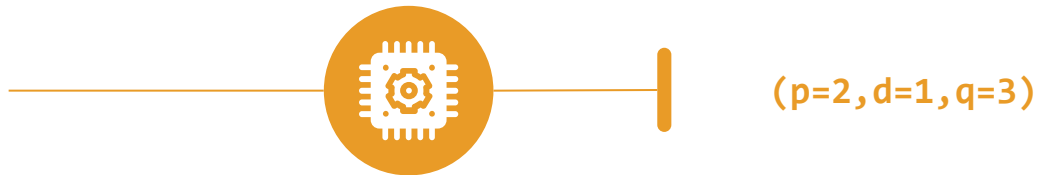


SARIMA

- Saisonnalité prise en compte
- « One-step forecast » meilleure
- « Dynamic forecast » bonne
- Erreur niveau des événements

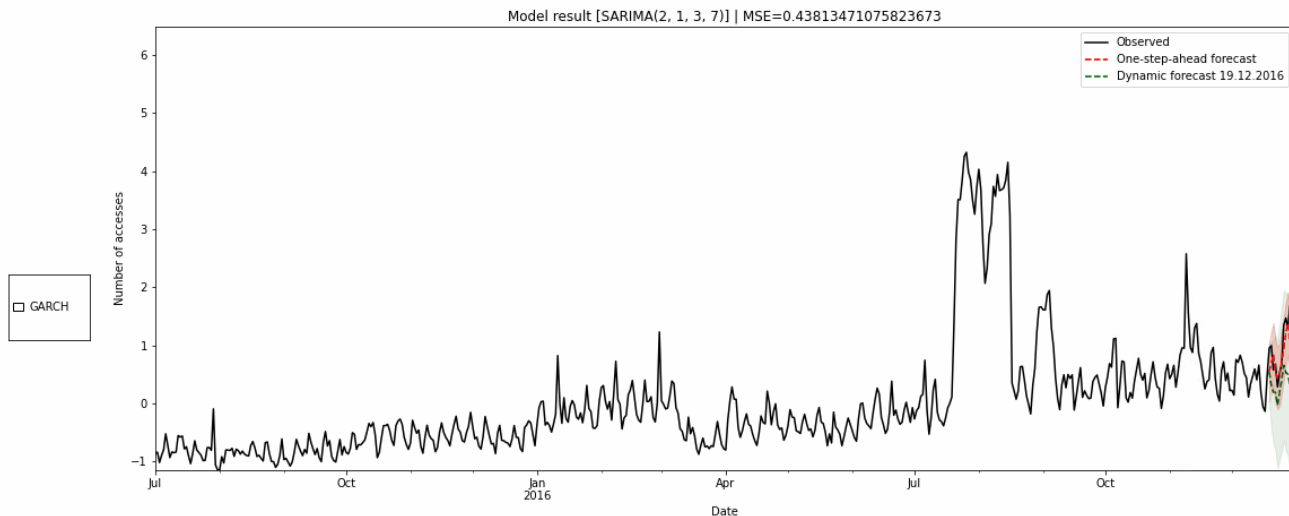
Premier modèle (2)

- *Brute force* avec MSE:
 - $p = [1, 5]$
 - $d = [1, 3]$
 - $q = [1, 5]$
 - $s = 7$
- GARCH(1,1) sur résidu



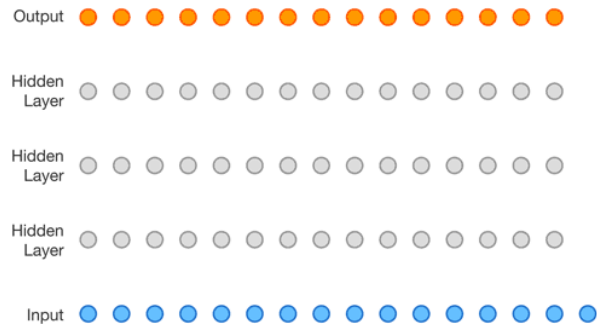
Premier modèle (3)

- Fenêtre interactive (matplotlib + PyQt5)
- Changement du nombre de jours à prédire
- Zone de sûreté
- One-step vs Dynamic

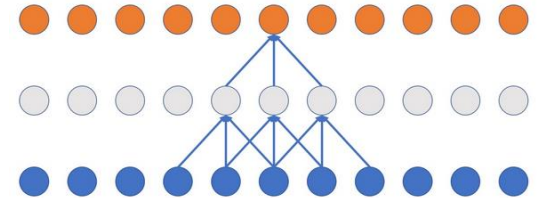


Wavenet (1)

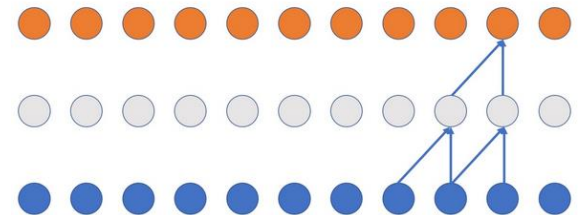
- Développé pour les voix synthétiques par Google Deepmind
- Utilise des réseaux de convolution
- Causal Convolution (keras: padding='causal')
- Dilated Causal Convolution (keras dilation_rate=value)



Standard Convolution

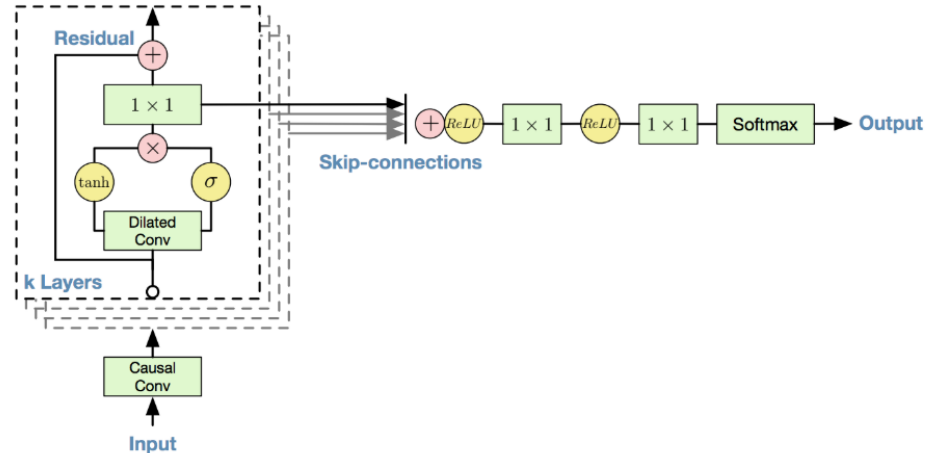


Causal Convolution



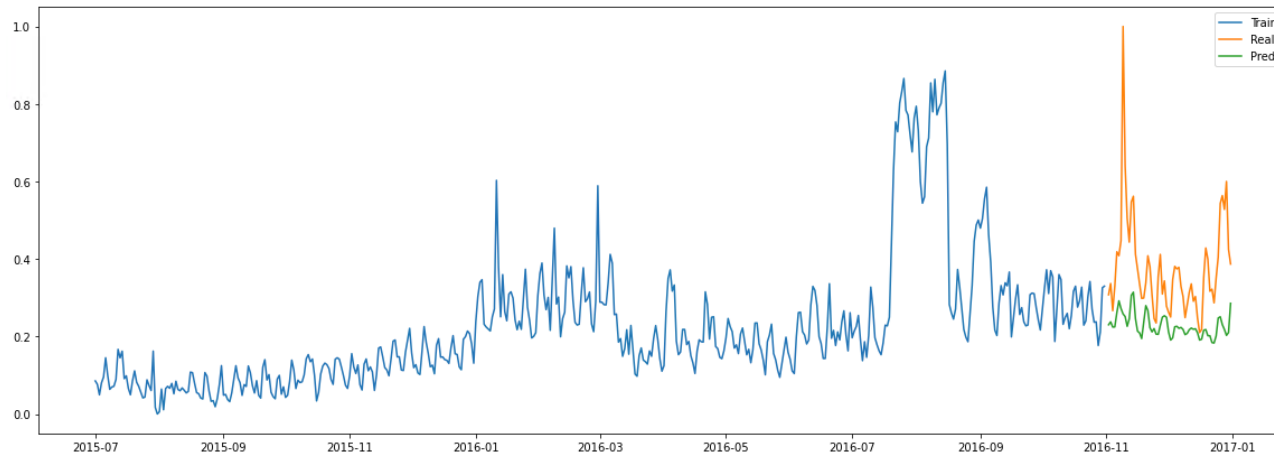
Wavenet (2)

- **k Layers** : Une couche Wavenet par dilation rate
 - Extrait l'information sur des périodes différentes
 - Rates : Suivent une courbe exponentielle
- Concept: Extraire des informations via les CNN
- Résultats supérieurs à la seconde position
- **Beaucoup plus rapide à entraîner**
 - Moins de paramètres
 - Moins d'époques pour converger



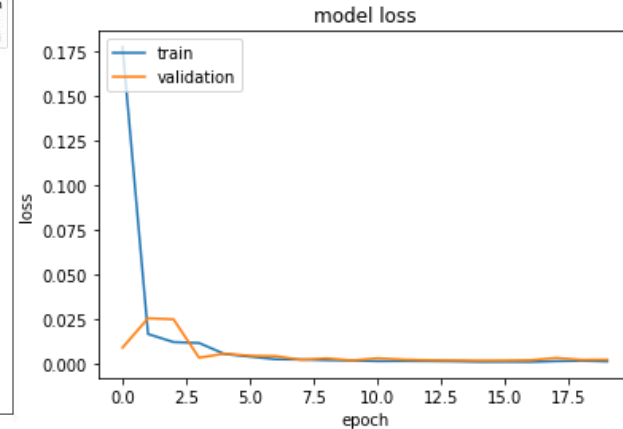
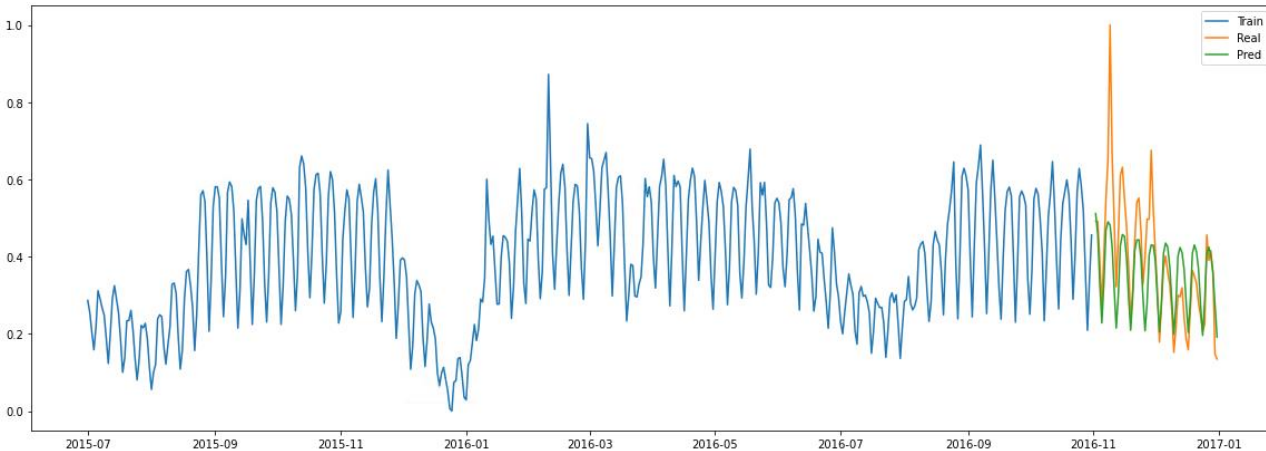
Second modèle (1)

- Données normalisées (MinMax)
- Input des données de 365 jours (≈ 1 année)
- 60 jours pour les données de test (≈ 2 mois)
- Suivi par un simple MLP (3 layers)
- Entraînement avec `validation_split=0.1`



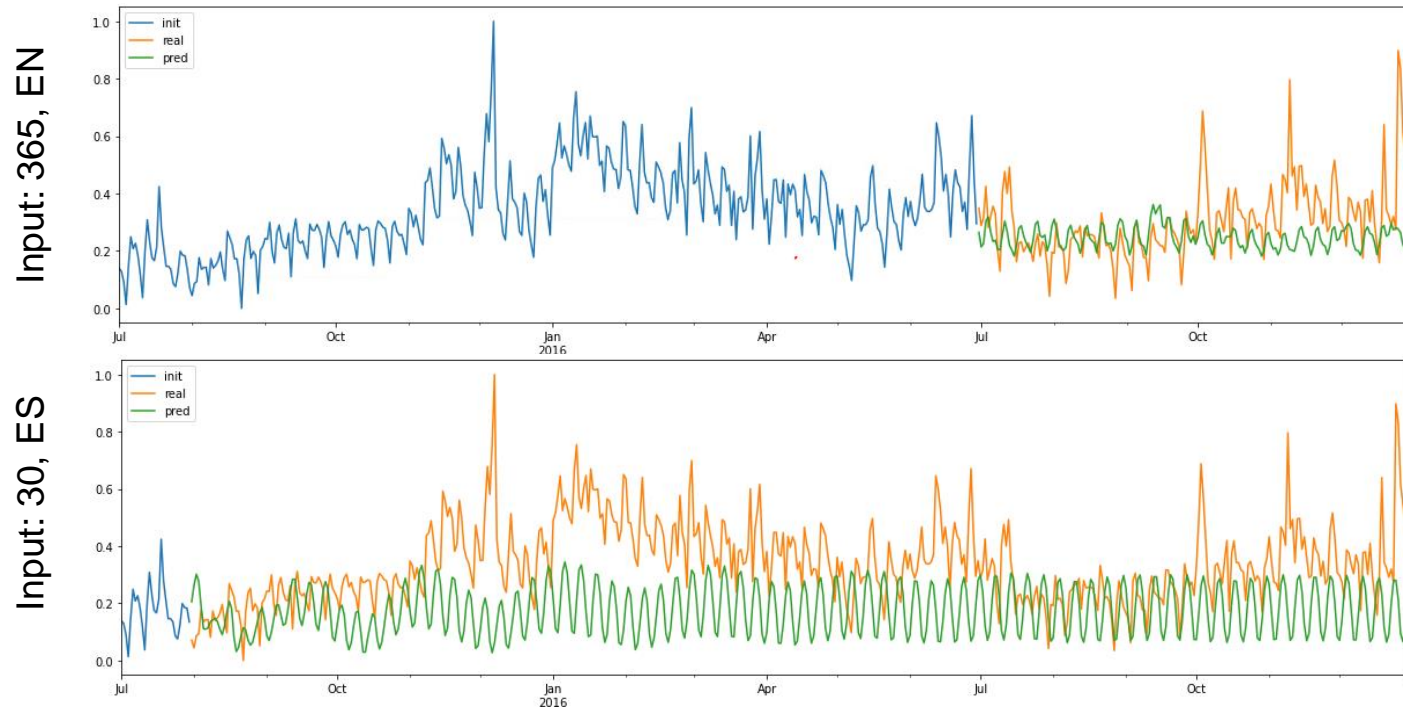
Second modèle (2)

- Données plus régulières : domaine espagnol
- Très rapide à converger (<20 epochs, <1 min 30 sec)
- Dynamic forecasting



Second modèle (3)

- Input trop large, données temporelles non suffisantes.
- Input des données de 30 jours (≈ 1 mois)
- Test sur d'autres domaines (entraînement: EN/ES, test: DE)



Conclusion

SARIMA

Technique

- Brute force sur (p,d,q)
- Saisonnalité hebdomadaire
- GARCH sur résidu

Résultats

- Saisonnalité extraite
- Résultat GARCH moyen
- Plot interactif

Deep Learning

Technique

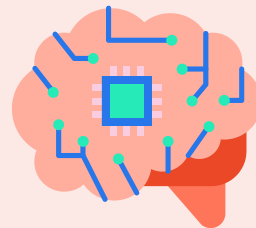
- CNN causal
- Wavenet
- Simple MLP

Résultats

- Entraîné en peu d'époques et rapide
- Saisonnalité extraite avec 1 mois

Améliorations

- Encoder les données en logarithme
- Entraîner sur plusieurs données temporelles



Sources

01

Wavenet

<https://bair.berkeley.edu/blog/2018/08/06/recurrent/>
<https://towardsdatascience.com/web-traffic-forecasting-f6152ca240cb>

02

GARCH

<https://medium.com/analytics-vidhya/arima-garch-forecasting-with-python-7a3f797de3ff>
https://arch.readthedocs.io/en/latest/univariate/univariate_volatility_modeling.html
http://web.vu.lt/mif/a.buteikis/wp-content/uploads/2019/03/02_GARCH.html

03

ARIMA/SARIMA

<https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/>
<https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/>

