Institut pour la Maîtrise des Risques
Sûreté de Fonctionnement - Management - Cindyniques

Surveillance de la production par un modèle prédictif en vue d'optimiser l'instant des actions curatives sur un système de production

Monitoring of production by a predictive model with a view to optimize the timing of curative actions on a production system

ELEGBEDE Charles
ArianeGroup
Saint-Médard en Jalles
charles.elegbede@ariane.group

COTTREL-BUSSENAULT Marie
ArianeGroup
Le Haillan
maire.cottrel@ariane.group

LAPEYRE Caroline
ArianeGroup
Le Haillan
caroline.lapeyre@ariane.group

PIRON Alexandre
ArianeGroup
Les Mureaux
alexandre.piron@ariane.group

AYADI Ines
ArianeGroup
Ottobrunn
ines.ayadi@ariane.group

Résumé —Dans un atelier industriel, la détection d'une non-conformité entraine automatiquement son enregistrement et un arrêt de production pour une analyse de premier niveau qui décidera des actions de sécurisation à mettre en place [1]. Une des actions de sécurisation consiste à identifier les causes racines avant la poursuite des opérations de production. Cette phase d'analyse de recherche de causes racines peut être plus ou moins longue et potentiellement bloquante pour la production. Il est donc nécessaire d'anticiper la connaissance des instants d'apparition des non-conformités grâce à l'analyse des dérives afin de commencer au plus tôt la recherche des causes racines avant l'apparition des non-conformités [2] et préparer en temps masqué les interventions correctives sur le système de production. Dans un premier temps nous avons formalisé le problème adressé dans ce papier puis présenter le modèle de série chronologique SARIMAX ainsi que les sous modèles dérivés : AR, MA, ARMA, ARIMA, SARIMA et SARIMAX. Un exemple d'application a été présenté à titre exploratoire sur des données aléatoires et n'a pas pu révéler les bénéfices escomptés des séries chronologiques : aucun modèle sous-jacent pertinent n'a pu être identifié, ce qui explique une mauvaise aptitude à la prédiction. Il nous a cependant, permis d'appréhender la méthodologie de mise en œuvre de ces modélisations et de proposer une feuille de route sur poursuivre les études sur des données présentant des dérives ou des cycles de production.

Mots-clefs — *SARIMAX, cartes de contrôle, détection d'anomalies, série chronologique, industrialisation*.

Abstract — In an industrial workshop, the detection of a non-conformity automatically results in its registration and a production stop for a first-level analysis that will decide on the safety actions to be implemented [1]. One of the safety actions is to identify root causes before further production operations. This phase of root-cause analysis can be more or less long and potentially blocking for production. It is therefore necessary to anticipate the knowledge of the moments of appearance of non-conformities through the analysis of the drifts in order to start as soon as possible the search for root causes before the appearance of non-conformities [2] and prepare corrective actions on the production system in masked time. First, we formalised the problem addressed in this paper and then presented the SARIMAX time series model as well as the derived sub models: AR, MA, ARMA, ARIMA, SARIMA and SARIMAX. An application example was presented as an exploratory study on random data and could not reveal the expected benefits of time series: no relevant underlying model could be identified, which explains a poor predictive ability. However, it allowed us to understand the methodology of implementation of these models and to propose a roadmap to continue studies on data presenting drifts or production cycles.

Keywords — *SARIMAX, control chart, anomalies detection, time series, industrialisation*

I. INTRODUCTION

Dans un atelier industriel, la détection d'une non-conformité entraine automatiquement son enregistrement et un arrêt de production pour une analyse de premier niveau qui décidera des actions de sécurisation à mettre en place [1]. Une des actions de sécurisation consiste à identifier les causes racines avant la poursuite des opérations de production. Cette phase d'analyse de

recherche de causes racines peut être plus ou moins longue et potentiellement bloquante pour la production. Il est donc nécessaire d'anticiper la connaissance des instants d'apparition des non-conformités grâce à l'analyse des dérives afin de commencer au plus tôt la recherche des causes racines avant l'apparition des non-conformités [2] et préparer en temps masqué les interventions correctives sur le système de production.

Dans un premier temps nous allons formaliser le problème adressé dans ce papier puis présenter le modèle SARIMAX ainsi que les sous modèles dérivés : AR, MA, ARMA, ARIMA, SARIMA et SARIMAX. Cette étape permettra au lecteur de situer le cadre des séries chronologiques.

Un exemple d'application sera présenté à titre exploratoire pour illustrer le positionnement actuel d'ArianeGroup sur l'intérêt de ces modélisations dans la surveillance de la production. A travers cet exemple on présentera également les différentes étapes de construction du modèle.

II. POSITION DU PROBLEME DE SURVEILLANCE DE LA PRODUCTION

A. Surveillance par carte de contrôle

La carte de contrôle est l'un des outils de surveillance de base utilisé pour la maîtrise statistique des procédés. C'est un outil graphique utilisé dans l'analyse du contrôle-qualité pour visualiser l'évolution d'un processus dans le temps [3]. Le but étant de détecter les dérives par rapport à une population de référence, traduisant des changements de comportement au niveau du procédé de fabrication et de contrôle [3]. La surveillance de la production par une carte de contrôle est jugée efficace par ArianeGroup et est largement déployée au travers de son processus d'industrialisation.

B. Anticipation des nonconformités suite à des dérives dans un processus de frabrication

Comme nous l'avons vu ci-dessus, la carte de contrôle fait un constat de l'état de la production et donne des alertes sur des dérives. Elle ne permet pas de faire des prédictions.

Pour pallier à cet inconvénient, nous proposons dans cette communication d'explorer l'utilisation des séries chronologiques modélisées par un processus SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous variables) [4] [5]. Ceci, pour déterminer de façon anticipée les dérives et surtout les moments où les caractéristiques suivies sont susceptibles de présenter des non-conformités. Les séries modélisées par SARIMAX présentent un intérêt certain dans la mesure où elles permettent de :

- Suivre les tendances centrales de la production, montrant les dérives éventuelles,
- Faire des prédictions d'anomalies sur un nombre donné de spécimens qui seront mesurés,
- Prendre éventuellement en compte les variables exogènes.

La première étape de notre feuille de route sera donc de déterminer si ces outils permettent d'identifier un modèle sous-jacent non identifiable sur la carte de contrôle, et ensuite de critiquer cette modélisation.

Nous utiliserons des outils de machine learning dédiés aux séries chronologiques pour construire le modèle d'apprentissage.

III. QUELQUES NOTIONS SUR LES SERIES CHRONOLOGIQUES

L'objectif de ce chapitre est de donner un aperçu sommaire sur les séries chronologiques afin d'introduire le sujet. Il s'agit d'éléments simples nécessaires pour les lecteurs de cet article.

A. Définition d'une série chronologique

On appelle série chronologique une suite (X_t) d'observations chiffrées d'un même phénomène, ordonnées dans le temps.

Il existe d'autres définitions plus théoriques, en effet une série chronologie bien qu'elle paraisse comme étant un banal recueil de données, est soutenue par la théorie des processus stochastiques encore plus complexe que celle des variables aléatoires. Volontairement dans ce papier qui se veut pratique, nous ne développerons pas les aspects théoriques renvoyant aux ouvrages et thèses nombreux sur la théorie des séries chronologiques [4][5].

Nota 1 : Une série chronologique est aussi appelée série temporelle ou chronique.

Exemple :

- Le nombre de passagers d'une compagnie aérienne enregistrés entre 1946 et 2000,
- Le diamètre mesuré sur des pièces usinées de même définition,
- La quantité d'énergie consommée chaque semaine dans une ville,
- La valeurs journalières boursières observées sur un indice boursier le S&P500 par exemple.

B. Objectif des séries chronologiques

Le principal objectif d'une série chronologique est de faire la meilleure estimée d'un processus à un ou plusieurs instants. Les trois principales estimées se font selon les opérations suivantes :

- Le filtrage : pour la valeur à l'instant courant compte tenu de l'ensemble des valeurs observées, précédemment et à l'instant courant,

- Le lissage : pour l'ensemble des valeurs comprises entre deux instants, compte tenu de l'ensemble des valeurs observées,
- La prédiction : pour une ou plusieurs valeurs futures, compte tenu des valeurs observées précédemment.

L'étude d'une série chronologique permet d'analyser, de décrire et d'expliquer un phénomène au cours du temps et d'en tirer des conséquences pour des prises de décision.

C. *Processus AR(p)*

Ce modèle consiste à prendre en compte les valeurs passées de la série temporelle pour prédire les valeurs actuelles en faisant une auto-régression linéaire sur les p dernières valeurs de la série temporelle. Si le processus est stationnaire, alors le modèle **AR**(p) s'écrit :

$$X_t = C_1 + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \qquad (1)$$

ε_t : erreur à l'instant *t*, un bruit blanc fort en général,

φ_i: le coefficient de régression du modèle,

X_t : valeur du processus à l'instant t,

C_1 : constante du modèle.

p : ordre du processus autoregressif,

D. *Processus MA(q)*

Les modèles à moyenne mobile suggèrent que la série présente des fluctuations autour d'une valeur moyenne. On considère alors que la meilleure estimation est représentée par la moyenne pondérée d'un certain nombre de valeurs antérieures. Ceci revient en fait à considérer que l'estimation est égale à la moyenne vraie, à laquelle on ajoute une somme pondérée des erreurs ayant entaché les valeurs précédentes. Le processus considéré doit être stationnaire et modèle *MA*(q) peut s'écrire :

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \qquad (2)$$

Avec :

ε_t : erreur à l'instant *t*, un bruit blanc fort en général,

θ_i: le coefficient de régression du modèle,

X_t : valeur du processus à l'instant t,

μ : moyenne du processus.

q : ordre du processus de moyenne mobile,

E. *Processus ARMA(p,q)*

Le processus *ARMA*(p,q) (AutoRegressive Moving Average) est un processus construit par combinaison des processus *AR* et *MA*. Le modèle s'écrit alors :

$$X_t = C_2 + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \qquad (3)$$

Les notations sont les mêmes que celles adoptées dans les équations (1) et (2).

C_2 : une constante du modèle.

F. *Processus ARIMA (p,d,q)*

La série ARIMA consiste à introduire une étape de différenciation dans le traitement de la série, ce qui permet d'avoir une série stationnaire qui pourra alors être modélisée par un processus ARMA. Si la série X_t présente une tendance, alors on introduit une série $Y_t = X_t - X_{t-1}$. Cette différentiation permet en général de rendre la série stationnaire. En général une différentiation suffit. Si ce n'est pas le cas, on peut différencier une deuxième fois la série en introduisant une variable $Z_t = Y_t - Y_{t-1}$.

Pour remonter aux valeurs de la série, il faudra alors faire les opérations inverses en intégrant la série au bon ordre, d'où le « I » d'ARIMA.

$$Y_t = C_2 + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \qquad (4)$$

Les notations sont les mêmes que celles adoptées dans les équations (1), (2) et (3) avec

G. *Processus SARIMAX(p,d,q)(P,D,Q)s*

Si lors de l'analyse de la série on identifie une saisonnalité, celle-ci sera introduite dans le modèle avec la périodicité de la saisonnalité ainsi que des variables exogènes, la série sera différenciée, à l'ordre D, avec des décalages correspondants à la saisonnalité en y intégrant également les variables exogènes :

$$Y_t = C_3 + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{P} \vartheta_i Y_{i-s.i} + \sum_{i=1}^{Q} \omega_i \varepsilon_{t-s.i} + \sum_{i=1}^{r} \gamma_i E_i + \varepsilon_t \qquad (5)$$

En plus des notations déjà définies ci-dessus, on adopte les notations suivantes :

- $s.i$: ce produit, représente la longueur du i ème pas de différentiation pour saisonnalité de pas s,
- P : l'ordre de régression de la saisonnalité,
- D : l'ordre de différenciation de la saisonnalité,
- Q : l'ordre de régression des erreurs liées à la moyenne mobile de la saisonnalité.
- ϑ_i: le coefficient de régression du modèle *AR* associé à la saisonnalité.
- ω_i: le coefficient de régression du modèle *MA* associé à la saisonnalité,
- E_i : valeur de la variable exogène à l'instant i,
- γ_i: le coefficient de régression du modèle associés aux variables exogènes,
- θ_i: le coefficient de régression du modèle,
- C_3 : une constante du modèle,

H. *Fonction d'autocorrélation et Fonction d'autocorrélation partielle*

L'identification consiste essentiellement en l'étude de la loi de variation de deux caractéristiques statistiques que sont :

- La fonction d'autocorrélation (*ACF*),
- La fonction d'autocorrélation partielle (*PACF*).

Il est donc primordial de bien définir et comprendre ces deux grandeurs qui constituent la clé de voûte de l'analyse des séries chronologiques.

1) *Fonction d'autocorrélation ACF*

On se donne une chronique $X_1, \dots, X_n$ et on va définir le coefficient de corrélation entre X_{t-k} et X_t par :

$$\rho(k) = \frac{E((X_{t-k}-\mu)(X_t-\mu))}{\sqrt{E((X_{t-k}-\mu)^2)}.\sqrt{E((X_t-\mu)^2)}} \qquad (6)$$

Il s'agit d'un coefficient de corrélation entre deux variables d'un même processus, d'où la notion d'auto « corrélation ». La valeur *k* désigne le décalage entre les deux valeurs du processus considérées.

Définition

La fonction f : $k \rightarrow f(k) = \rho(k)$ est appelée fonction d'autocorrélation.

Propriété de convergence de la fonction d'autocorrélation

Pour une série chronologique de taille *n*, on montre que cette fonction converge vers une loi normale de moyenne 0 et d'écart-type simple à calculer :

$$\hat{\rho}(k) \rightarrow N\left(0, \frac{1}{\sqrt{n^*}}\right) \qquad (7)$$

Avec :

$$n^* = n - d - sD$$

Cette propriété de convergence est capitale dans la détermination des hyper-paramètres du modèle SARIMAX.

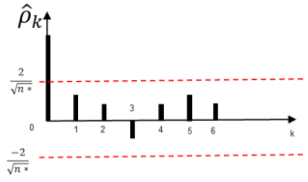

Fig. 1. Graphe ACF de convergence de l'autocrrélation d'un process MA(2)

2) *Fonction d'autocorrélation partielle PACF*

On se donne une chronique $X_1, \dots, X_n$ et on va définir un coefficient de corrélation entre X_{t-k} et X_t qui fasse abstraction de l'influence de $X_{t-k+1}, \dots, X_{t-1}$. Cette notion est une réponse à la question suivante : il arrive que 2 phénomènes soient fortement corrélés, mais que cette corrélation soit due à l'influence d'un facteur extérieur et non pas à un fort lien entre les deux phénomènes. Le coefficient de d'autocorrélation partielle entre X_{t-k} et X_t abstraction faite de l'influence de $X_{t-k+1}, \dots, X_{t-1}$.est le coefficient de d'autocorrélation partielle entre les deux variables X_{t-k} et X_t auxquelles on a retranché leurs meilleures explications φ en termes de X_{t-k} et X_t.

$$r(k)_{|X_{(t-k)+1,\dots,X_{t-1}}} = \rho\left(X_{t-k} - \varphi_{X_{(t-k)+1,\dots,X_{t-1}}}(X_{t-k}), X_t - \varphi_{X_{(t-k)+1,\dots,X_{t-1}}}(X_t)\right) \qquad (8)$$

Pour simplifier l'écriture on la notera simplement *r(k)*. Yule et Walker montrent que *r*(*k*) peut être approximé par la suite :

$$\Psi_{11} =. \rho_{1,} \qquad (9a)$$

$$r(k) = \Psi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \Psi_{k-1,j}.\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \Psi_{k-1,j}.\rho_j} \qquad (9b)$$

$$r(k) = \Psi_{kk} \text{ pour } k>0$$

Définition :

La fonction g : $k \rightarrow g(k) = r(k)$ est appelée fonction d'autocorrélation partielle (*PACF*)

I. Methode de Box-Jenkins

La méthode de Box-Jenkins est la méthode éprouvée pour la construction du modèle. L'analyse des séries temporelles unidimensionnelles par la méthode de Box-Jenkins comporte traditionnellement trois étapes:

- E1 : Identification de la forme de modèle la mieux adaptée à la série étudiée,
    - Estimation de la tendance,
    - Détermination de l'ordre de dérivation (stationnarisation de la série),
    - Détermination de l'hyper paramètre *q* d'une moyenne mobile,
    - Détermination de l'hyper paramètre *p* du modèle *AR*.
- E2 : Estimation des coefficients du modèle,
    - Estimation des coefficients des modèles avec les outils numériques adaptés,
    - Construire le modèle.
- E3 : Validation : le modèle retenu convient-il bien? Sinon pourquoi? et comment l'améliorer.
    - Vérification que les p-value des coefficients du modèle pour conclure qu'ils sont significatifs,
    - Prévision de valeurs futures et confronter aux valeurs de test,
    - Analyse de la qualité du modèle (vérifier que les résidus suivent un bruit blanc fort, analyse des critères associées à la MAPE et éventuellement d'autres critères spécifiques au problème)

Avant de construire le modèle, il est crucial d'analyser les tendances, les schémas saisonniers et l'influence potentielle des co-variables sur les valeurs de la chronique.

J. Principes d'identification d'un processus ARIMA

Les principes d'identification des processus *ARIMA* sont les suivants :

- La présomption de processus *ARIMA* qui résulte :
    - De l'observation visuelle de la série indiquant une tendance,
    - Du fait que la corrélation empirique décroit très lentement. En effet, si la corrélation empirique décroit lentement, on peut avoir affaire soit à un processus *ARIMA*, soit à une série avec tendance. L'observation visuelle peut permettre de lever le doute.
- On opère ensuite sur ce processus des opérations de différentiation successives, jusqu'à obtenir un processus stationnaire (sa corrélation empirique doit décroitre assez rapidement) le nombre minimum d'opérations de différenciation nécessaires donne la valeur de *d*, en générale une différentiation suffit et au maximum deux.

Théorème d'indentification des AR(p)

Une loi de variation peut être approximée par un modèle autorégressif, si et seulement si :

- Sa fonction d'autocorrélation décroît (ACF), en valeur absolue, exponentiellement vers zéro,
- Sa fonction d'autocorrélation partielle est identiquement nulle au-delà du décalage *t-p*. La valeur de *p* qui est aussi égale au nombre pics significatifs dans la représentation en histogramme des *PACF*, est l'ordre du processus.

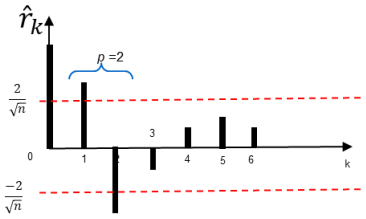

Fig. 2. Graphe ACF de convergence de l'autocrrélation d'un process MA(2)

Nota : Pour un processus AR, la fonction de corrélation partielle est à horizon fini.

Une première méthode consiste à réaliser des tests statistiques. En pratique on considère pour p variant de 1 à P les tests successifs permettant de comparer le modèle d'ordre p au modèle d'ordre $p + 1$.

H_0 : X suit un modèle $AR(p)$ contre H_1 : non H_0 .

Par le théorème de normalité asymptotique des équations de Yule-Walker [7], on a directement, pour un processus $AR(p)$, un théorème central limite pour les *PACF* empiriques.

$$\hat{r}(k) \rightarrow N\left(0, \frac{1}{\sqrt{n}}\right) \qquad (10)$$

Où :

n: est la taille de l'observation.

L'intervalle de confiance à 95% est alors approximativement : $\hat{r}(k) \in \left[\frac{-2}{\sqrt{n}}, \frac{+2}{\sqrt{n}}\right]$

Une première méthode consiste à réaliser des tests statistiques. En pratique on considère pour p variant de 1 à P les tests successifs permettant de comparer le modèle d'ordre p au modèle d'ordre supérieur.

Théorème d'indentification des MA(q)

Un processus peut être approximé par un modèle de moyenne mobile si et seulement si :

- Sa fonction d'autocorrélation partielle (PACF) décroît, en valeur absolue, exponentiellement vers zéro,
- Sa fonction d'autocorrélation est identiquement nulle au-delà du temps *t-q*. La valeur de *q* qui aussi égale au nombre pics significatifs dans la représentation en histogramme des ACF, est appelé ordre du processus. On note alors ce processus MA(p)

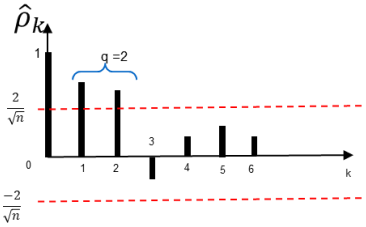

Fig. 3. Graphe ACF de convergence de l'autocrrélation d'un process MA(2)

Nota : Pour un processus *MA*, la fonction d'autocorrélation est à horizon fini.

Théorème d'indentification des ARMA(p,q)

Un processus peut être approximé par un modèle de moyenne mobile si et seulement si :

- Sa fonction d'autocorrélation partielle est identiquement nulle au-delà du décalage *t-p*. La valeur de *p* qui est aussi égale au nombre pics significatifs dans la représentation en histogramme des *PACF*, est l'ordre du processus.
- Sa fonction d'autocorrélation est identiquement nulle au-delà du temps *t-q*. La valeur de *q* qui aussi égale au nombre pics significatifs dans la représentation en histogramme des *ACF*, est appelé ordre du processus. On note alors le processus *MA*(*q*)

K. Estimation des paramètres du modèles

Une fois les hyper-paramètres *p* et *q* déterminés, on va déterminer les coefficients du modèle en utilisant le langage python avec la bibliothèque SARIMAX de statmodels. Pour ce faire, la base de données sera divisée en deux parties en respectant l'ordre chronologique d'occurrence des valeurs :

- Une base d'apprentissage allant de x_1, à x_l,

- Une base de test allant de x_{l+1}, à x_n.

Il est commun de construire la base d'apprentissage en prenant 80% de la base de données complète. Et le reste des données serviront à faire les tests de validation du modèle.

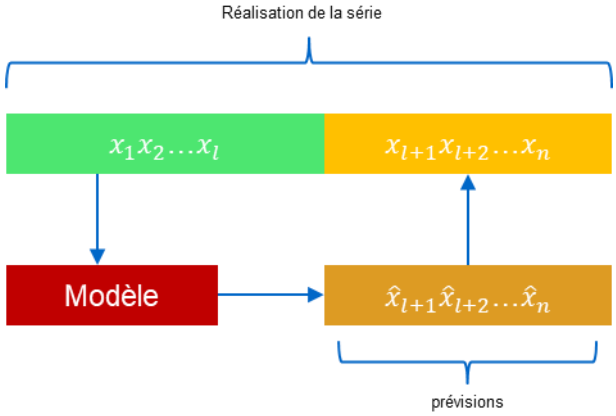

Fig. 4. Construction et validation du modèle

L. *Validation du model*

Pour évaluer la performance en prévision du modèle, il faut pouvoir mesurer l'écart entre les prévisions et la réalité. Comme indiqué ci-dessus, la démarche à adopter consistera à diviser la base de données disponible en deux parties. La première partie est utilisée pour entrainer le modèle et la deuxième partie servira à confronter le modèle à la réalité. Tout en conservant la chronologie des valeurs, la première partie sera composée de 80% des données, et la deuxième des 20% restantes.

L-1 *MAPE*

Le Mean Average Percentage Error est un indicateur couramment utilisé pour l'évaluer la qualité d'un modèle de série chronologique.

$$MAPE = \frac{1}{n-l} \sum_{i=l+1}^{n} 100. \left| \frac{X_i - \hat{X}_i}{X_i} \right| \quad (11)$$

Cet indicateur est une moyenne de pourcentage d'erreur qui permet de mesurer les résidus relatifs de de l'ajustement. Plus ce seuil sera bas, meilleur sera l'ajustement du modèle de la série chronologique. La littérature n'indique pas un seuil de MAPE. Nous fixons arbitrairement un critère sur la MAPE :

Critère 1 : $MAPE \leq 10\%$ (12)

Ce premier critère permet de garantir un modèle qui ne s'écarte pas trop des points ayant permis sa construction.

Critère 2 : Ce deuxième critère en lien avec le besoin de performance donné par l'intervalle de tolérance/spécification de la caractéristique à surveiller. Il exprime le fait que le modèle doit être suffisamment fin au regard de la tolérance spécifiée. Le facteur diviseur *K* fixé arbitrairement à 10 dans cette communication sera à adapter au niveau de finesse visé sur le modèle :

$$MAPE.\mu \leq \frac{IT}{K} \quad (13)$$

IT : intervalle de tolérance,

μ : la moyenne du processus/valeur nominale de la surveillance,

K : facteur définissant le niveau de finesse du modèle.

Critère 3 : En plus du critère MAPE, un autre critère est rajouté pour mesurer la qualité de l'apprentissage. Il repose sur une notion de ratio d'écarts cumulés. L'apprentissage sera d'autant meilleur que ce ratio est proche de zéro.

$$RDEAC = \frac{\sum_{i=l+1}^{n} |X_i - \hat{X}_i|}{\sum_{i=l+1}^{n} |X_i - \mu_{apprentissage}|} \quad (14)$$

$$RDEAC \leq 0.25$$

$\mu_{apprentissage}$: la moyenne des données d'apprentissage.

Les valeurs de ces trois critères sont arbitraires et devront être revues et justifiées en cas de décision de déploiement de ces méthodes. Il est clair que ces trois valeurs de critères dépendent des objectifs d'utilisation du modèle construit, et seront choisis en conséquence.

L-2 Test des coefficients du modèle *ARIMA*

Le deuxième indicateur de qualité du modèle que nous avons retenu est celui consistant à tester les coefficients du modèle en vérifiant que pour chaque coefficient on a la *p*-value < 0,05, auquel cas les coefficients seraient significatifs.

L-3 Analyse de la blancheur du résidu

Il s'agira de vérifier que les résidus du model calculés aux points de test sont distribués selon un bruit blanc fort (gaussienne de moyenne nulle et d'écart-type constant).

Cette même vérification est faite également sur les résidus des points ayant servi à construire le modèle.

IV. Application a un probleme de production

Comme indiqué ci-dessus, l'outil de calcul utilisé est la bibliothèque *statmodels,* elle donne exactement les mêmes résultats que le module sklearn de la bibliothèque scikit learn que nous ne présentons pas dans cette communication.

A. Présentation de l'exemple numérique

Caractéristique mesurée est une grandeur spécifiée à la définition et tolérancée comme suit : [1082-1088]. Il s'agit d'une production unitaire sur une petite série. L'intervalle de tolérance IT = 6 et la valeur moyenne attendue est de μ = 1085.

Nous disposons au totale d'une base de données de 104 valeurs.

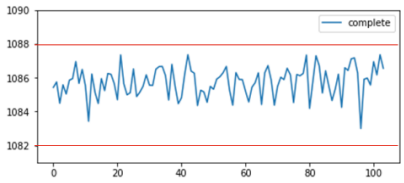

Fig. 5. Données d'entrainement et de test

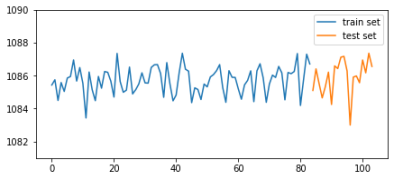

Fig. 6. Données d'entrainement et de test

Comme supra indiqué, le modèle sera construit en utilisant 80% de la population (soit 84 premières valeurs dans l'ordre chronologique) et testé avec les 20% restant soit 20 valeurs.

B. Analyses et construction du modèle ARIMA

Nous allons suivre les trois étapes de la méthode de Box-Jenkins :

Etape 1 : Identification du modèle

La chronique ne présente pas à vue d'œil de tendance si de saisonnalité. Il n'est donc pas nécessaire de la différencier. On cherchera donc à déterminer deux hyper-paramètres pour ce modèle qui sera de la famille *ARMA(p,q)*, car *d* = 0.

Pour confirmer que *d* = 0, on fait un test de stationnarité de *ADF* (augmented Dicker-Fuller test) [8] et on obtient une *p*-value: 1.57 10^{-11} qui nous confirme qu'on peut considérer alors que la série est stationnaire.

Etape 2 : Recherche des paramètres du modèle

La recherche des paramètres *p* et *q* du modèle *ARIMA* (*p*, 0, *q*) s'appuie sur les graphes d'autocorrélation (Fig.8) et d'autocorrélation partielle (Fig.7). L'analyse de ces deux graphes montrent dès les premiers décalages qu'on a des corrélations faibles (dans l'intervalle de confiance à 95%). Dans ce cas, ces graphes ne sont pas d'une grande utilité pour construire le modèle *ARIMA* on doit rechercher pas à pas les couple (*p*,*q*) et évaluer les modèles obtenus en utilisant les outils de validation de modèles que nous avons décrits au §III.L.

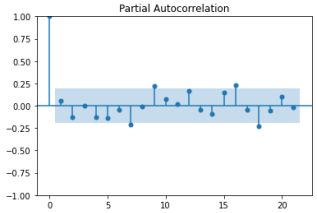

Fig. 7. Autocorrelation partielle

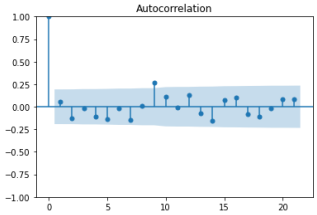

Fig. 8. Autocorrélation

Le choix pas à pas du couple (p, q) se fait en partant de (1, 1) en augmentant de façon judicieuse p et q simultanément ou non pour arriver au couple (p,q) optimal (p et q les plus petits) permettant d'avoir un modèle dont i) les coefficients sont significatifs *ie* leur p-value < 0.05 et ii) si deux modèles sont en compétition on choisira celui dont l'*AIC* (Akaike Information Criteria) est le plus faible. Le graphe superposant la chronique au modèle permet de visualiser (quand c'est possible) la qualité de l'adéquation avant de procéder à la validation par les critères de qualités numériques.

$$AIC = 2k - lnL \qquad (15)$$

Où:

k : nombre de paramètres du modèle,

L : la vraisemblance du modèle.

Nous présentons ici quelques graphes issus de cette étape de recherche pas à pas :

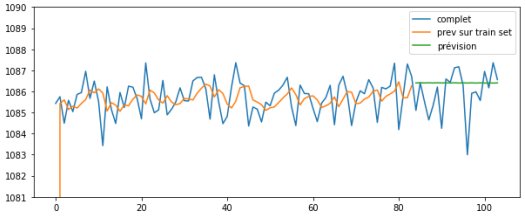

Fig. 9. (p,q) = (2,2) : AIC = 253, les p_value > 0.05

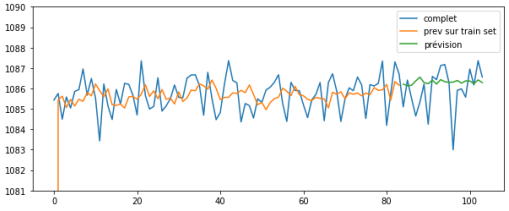

Fig. 10. (p,q) = (10,4) : AIC 263, les p_value > 0.05

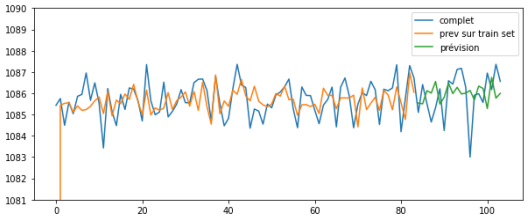

Fig. 11. (p,q) = (18,4) : *AIC* = 249 et p-value nulle

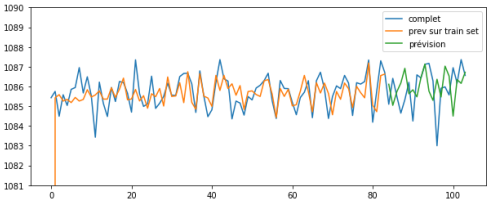

Fig. 12. (p,q) = (20,8) : *AIC* 246, et p-value nulle

Le modèle *ARIMA*(20,0,8) sera retenu pour cette chronique. En effet, dans la zone d'apprentissage il épouse au mieux la chronique et en particulier sur les dernières séries de valeurs. Et plus important, les p_values associées aux coefficients du modèle sont quasi-nulles.

Le modèle recherché est donc de la forme donnée par l'équation (3) ci-dessus qui décrit un processus *ARMA* (20,8).

Les coefficients du modèle obtenus sont donnés dans les tableaux ci-dessous et on a le terme constant qui est nul dans le cas d'espèces :

TABLE I. COEFFICIENTS DU MODÈLE AR

i	φ_i	*i*	φ_i
1	-2,7396	11	2,2508
2	-4,7205	12	3,0937
3	-5,9226	13	3,7383
4	-5,8051	14	3,9313
5	-4,4491	15	3,5907
6	-2,4369	16	2,9409
7	-0,6504	17	2,3129
8	0,4279	18	1,6552
9	0,9653	19	0,9113
10	1,5037	20	0,4022

TABLE II. COEFFICIENTS DU MODÈLE MA

i	θ_i
1	3,0043
2	5,5187
3	7,3771
4	7,7437
5	6,3687
6	4,019
7	1,7452
8	0,3331

TABLE III. VALEUR DE L'ECART-TYPE DU BRUIT BLANC

σ	0,59892

En le confrontant aux données réservées au test, on obtient le graphe ci-dessous dont on peut déjà observer que l'ensemble des valeurs sont dans la limite de surveillance du procédé de fabrication,

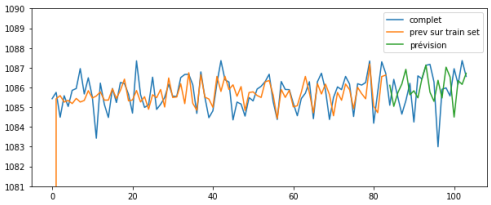

Fig. 13. $(p,q) = (20,8)$: *AIC* 246, et *p*-value nulle

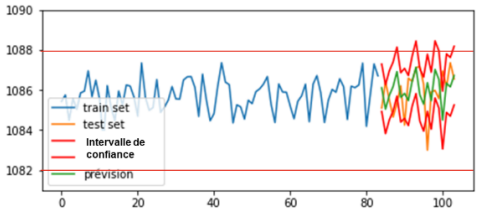

Fig. 14. Intervalle de confiance de la prévision

Etape 3 : Validation du modèle

La validation du modèle s'appuie principalement sur trois critères : *MAPE*, la significativité des coefficients du modèle et la blancheur des résidus.

Calcul de la MAPE et des critères de qualité du modèle

La *MAPE* a été calculée sur les 20% des valeurs qui ont servi à tester le modèle comme indiqué ci-dessus.

Critère 1 : *MAPE* = 0,09% < 10 % donc le critère est respecté,

Critère 2 : *RDEAC* = 21,24/16.25 = 1,30 > 0,25, donc le critère n'est pas respecté,

Critère 3 : *MAPE* * μ = 0,09 * 1085 = 0,9765 < IT/10 = 0,6, donc critère non respecté.

Significativité des coefficients du modèle

Toutes les *p*-values de chaque coefficient (cf, TABLE I. TABLE II.) sont quasi nulles. Cela indique donc que les coefficients du modèle sont significatifs. En effet, c'est un des crièteres qui aide à la sélection des paramètres *p* et *q*.

Analyse des résidus de la base d'entrainement

Le test de normalité d'Anderson Darling ne rejette pas l'hypothèse de normalité des résidus (sur les points d'apprentissage). En effet, la *p*-value = 0,14 est au delà du seuil de 0,05. De plus, la droite de Henry montre que l'ensemble des points se situe dans l'intervalle de confiance à 95% de la droite de Henry exceptés quelques uns en débuts d'apprentissage. La moyenne des résidus est *m* = 0,05 et $sigma$ = 0,67, ce qui conduit à un intervalle de confiance sur la moyenne incluant le 0, permettant de conclure qu'on a un bruit blanc fort. L'écart-type des résidus (0,67) reste important par rapport à la tolérance (11,2% IT).

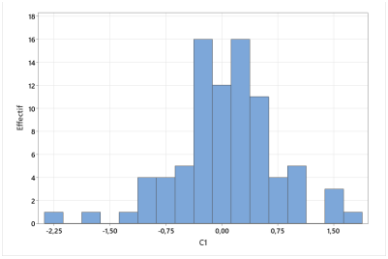

Fig. 15. Histogramme des résidus

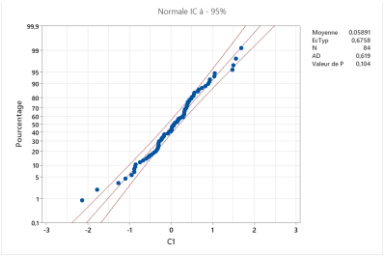

Fig. 16. Droite de Henry

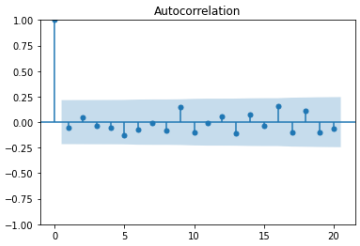

Fig. 17. Autocorrelation des résidus

Analyse des résidus de la base de test

Le test de normalité d'Anderson Darling ne rejette pas l'hypothèse de normalité des résidus (sur les points de test). En effet, la *p*-value = 0,65 est au delà du seuil de 0,05. De plus, la droite de Henry montre que l'ensemble des points se situe dans l'intervalle de confiance à 95% de la droite de Henry. La moyenne des résidus est *m* = -0,10 et son écart-type $sigma$ = 1,36

ce qui conduit à un intervalle de confiance sur la moyenne incluant le 0, permettant de conclure qu'on a un bruit blanc fort. L'écart-type des résidus (1,36) des valeurs de tests est deux fois supérieur à celui des données d'entrainement.

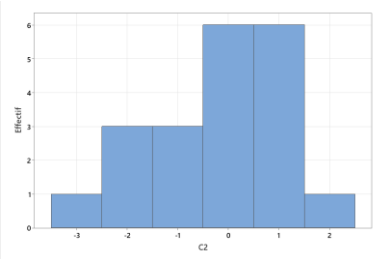

Fig. 18. Histogramme des résidus

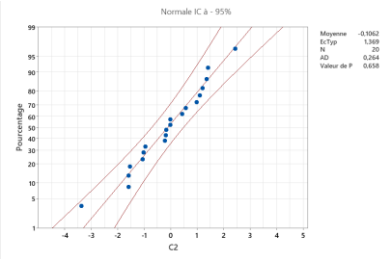

Fig. 19. Droite de Henry

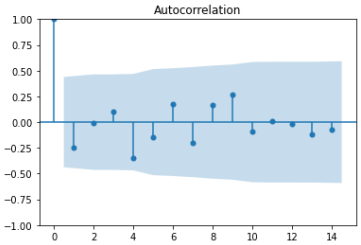

Fig. 20. Autocorrelation des résidus

Observations sur le modèle

Si on observe le graphique 12, on observe que :

- Le modèle colle plutôt bien aux données d'apprentissage,
- Le modèle colle beaucoup moins bien aux données de test : valeurs prévues haute alors que basses, inversement, …
- Le modèle n'a pas prévu le point très bas,
- L'ensemble des points prédits sont dans l'intervalle de tolérance spécifié,
- Les prédictions générées par le modèle restent cohérentes des valeurs de la base de données de test,
- Les dispersions sur la population issue de la prédiction du modèle sur les données de test, sont plus faibles que sur celles de la population issue de la prédiction sur les données d'apprentissage (Fig.13),
- L'intervalle de confiance à 95% associé à la prédiction est en dehors de limites de tolérances.

Analyse du critère 2 montre que les valeurs de la population de test (ne rentrant pas dans la construction du modèle) sont plus éloignées de leurs prévisions résultant de la modélisation que de la valeur moyenne de la population d'apprentissage la modélisation n'est pas pertinente.

Analyse de la parcimonie du modèle : le modèle identifié est *ARMA(p,q)* avec $p = 20$, $q = 8$ et $\sigma = 0{,}59892$, soit 29 paramètres établis sur une population d'apprentissage de 84 valeurs. Ce nombre très élevé de paramètres a été nécessaire pour coller au mieux à la population ayant servi à l'apprentissage, mais il n'est pas étonnant qu'en contrepartie, il présente une mauvaise aptitude à la prévision. Des modèles plus parcimonieux comme indiqué sur les figures 9 et 10, montrent des qualités d'apprentissage et de prévision mauvaises aussi bien pour les données d'apprentissage que sur les données de tests.

Nota : certains logiciels de statistique donnent une valeur limite de 5 à *p* et *q*.

V. CONCLUSION

L'exemple étudié à ce jour, données *a priori* aléatoires, n'a pas pu révéler les bénéfices escomptés des séries chronologiques : aucun modèle sous-jacent pertinent n'a pu être identifié, ce qui explique une mauvaise aptitude à la prédiction.
Il nous a, cependant, permis d'appréhender la méthodologie de mise en œuvre de ces modélisations.

La construction des modèles de série chronologique n'est pas toujours aisée. Dans le cas où les fonctions d'autocorrélation et d'autocorrélation partielles convergent dès les premiers décalages vers 0, la recherche des hyper-paramètres des modèles se fait par approche successive avec contrôle du critère *AIC*, puis des *p*-value des coefficients des modèles définis par les hyper-paramètres.

Une fois les hyper paramètres identifiés, les coefficients du modèle sont facilement calculés. Ce travail est bien évidement à déléguer aux bibliothèques statistiques de python comme statmodels ou sklearn qui donnent exactement les mêmes résultats.

Les étapes de validation du modèle sont automatisables et sont relativement faciles à implémenter.

En terme de perspective, nous souhaitons poursuivre nos investigations sur ces modèles :

1. Evaluer les apports de ces modélisations sur des données de productions présentant,
   - Des données de type cycle (ex : cycle de cuisson),
   - Des données présentant des dérives.
2. Si des bénéfices sont dégagés :
   - Consolider la méthodologie d'établissement d'un modèle,
   - Consolider les critères de validation mathématique d'un modèle : critères et valeurs limites, valeurs maximales de *p* et *q*,

- Définir dans quels cas utiliser cet outil, le positionner dans le processus de surveillance des procédés,
- Evaluer les risques d'utilisation d'une prévision (ex : dérive) dans le cas où aucune justification « métier » de compréhension du modèle n'a pu être apportée.

REFERENCES

1. EN 91000 : Systèmes de Management de la Qualité pour les Organismes de l'Aéronautique, l'Espace et la Défense,
2. EN 9103 : Systèmes de management de la qualité — Management de la variation des caractéristiques clefs,
3. Maurice Pillet : Appliquer la maîtrise statistique des processus, Ed d'Organisation, 2005- ISBN : 2-7081-3349-7,
4. Sylvestre Tatsa : Modélisation et prévision de la consommation horaire d'électricité au Québec: Comparaison de méthodes de séries temporelles, PhD dissertation, Université Laval (Quebec), 2014,
5. Seif-Eddine Benkabou : Détection d'anomalies dans les séries temporelles : application aux masses de données sur les pneumatiques, Thèse de doctorat, de l'Université Claude Bernard Lyon, 2018,
6. Abdul Aziz Guepe : Méthode de Box-Jenking Analyse des série chronologique, Projet de fin d'étude, Département Génie Civil, Ecole Polytechnique de Thies, 1987,
7. J. Abadie et D. Travers. Une approche simplifiée de la méthode de Box-Jenkins pour l'analyse de la prévision des séries temporelles unidimensionelles . RAIRO Recherches Opérationnelle (II), (vol. n°1, février 1981, p.51-71,
8. Cem Ertur. Méthodologies de test de la racine unitaire. [Rapport de recherche] Laboratoire d'analyse et de techniques économiques(LATEC). 1998, 36 p., Table, ref. bib. : 54 ref. hal-01527262.

Annexe : données analysées

N°	Mesures	N°	Mesures	N°	Mesures	N°	Mesures	N°	Mesures
1	1085,43	22	1087,349	43	1087,36	64	1085,71	85	1085,1
2	1085,75	23	1085,64	44	1086,393	65	1086,29	86	1086,41
3	1084,493	24	1084,99	45	1086,27	66	1084,42	87	1085,51
4	1085,58	25	1085,12	46	1084,36	67	1086,29	88	1084,65
5	1085,04	26	1086,52	47	1085,26	68	1086,72	89	1085,33
6	1085,856	27	1084,89	48	1085,16	69	1085,87	90	1086,21
7	1085,946	28	1085,15	49	1084,55	70	1084,38	91	1084,25
8	1086,954	29	1085,495	50	1085,49	71	1085,49	92	1086,59
9	1085,67	30	1086,17	51	1085,32	72	1086,03	93	1086,43
10	1086,49	31	1085,56	52	1085,92	73	1085,89	94	1087,12
11	1085,53	32	1085,545	53	1086,06	74	1086,56	95	1087,17
12	1083,43	33	1086,5	54	1086,3	75	1086,17	96	1086,28
13	1086,22	34	1086,659	55	1086,67	76	1084,53	97	1083
14	1085,14	35	1086,67	56	1085,27	77	1086,19	98	1085,91
15	1084,48	36	1086,13	57	1084,38	78	1086,12	99	1085,98
16	1085,95	37	1084,686	58	1086,3	79	1086,25	100	1085,57
17	1085,24	38	1086,79	59	1085,9	80	1087,34	101	1086,95
18	1086,25	39	1085,51	60	1085,89	81	1084,19	102	1086,17
19	1086,2	40	1084,47	61	1085,2	82	1085,76	103	1087,36
20	1085,66	41	1084,82	62	1084,57	83	1087,3	104	1086,56
21	1084,7	42	1086,205	63	1085,44	84	1086,71		