放電加工機の時系列解析（その２）

1.1 はじめに

放電加工機の時系列解析に取り組みます。

時系列解析は、以下の４ステップから構成されます。

① ディスクリプション （description)

時系列の特徴を把握します。定常性の確認、成分分解等を行います。

② **モデリング（modeling）**

時系列モデルを構成し、パラメータを推定します。

③ **予測 （prdiction)**

現在までの情報から今後の変動を予測します。

④ 信号抽出

必要な信号や情報を取り出す。（異常検知とか）

今回の報告書は、上記の **モデリング（modeling）**と **予測 （prdiction)** についての取組内容です。モデルについて、代表的な手法として、**自己回帰モデル** と 状態空間モデルがありますが、今回は、**自己回帰モデル**についての報告です。

1.2 どのようなデータを扱ったか

放電加工機は、加工中に障害が発生すると自己診断して 3 段階のレベルに分け、レベルに応じたメッセージを表示します。内容は以下です。

・エラーメッセージ：続行不可能な障害が発生したときに表示し、動作を中断する。

・ハルトメッセージ：再開可能な障害が発生したときに表示し、一時停止する。

・コメントメッセージ：続行可能な障害で発生し、注意を促す。

また、メッセージの記録は、USB により CSV データとして取り出すことが可能です。

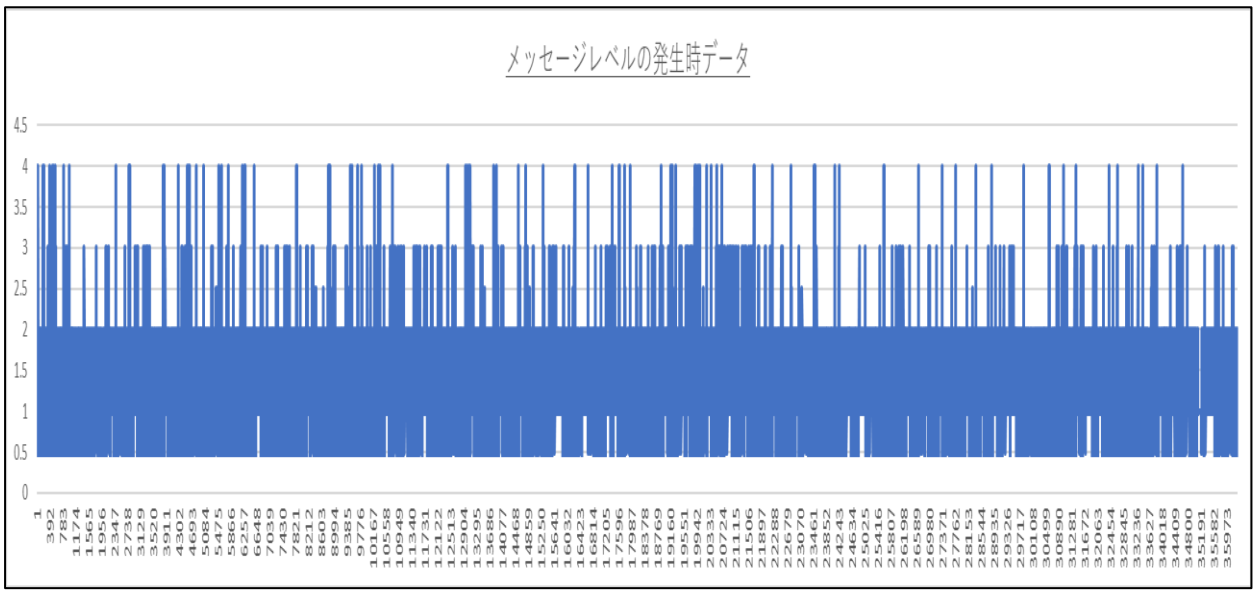
レベルに対応してデータ処理できるように、3 段階のレベルについて障害のレベルが大きいと、点数は大きくなるように点数を設けました。

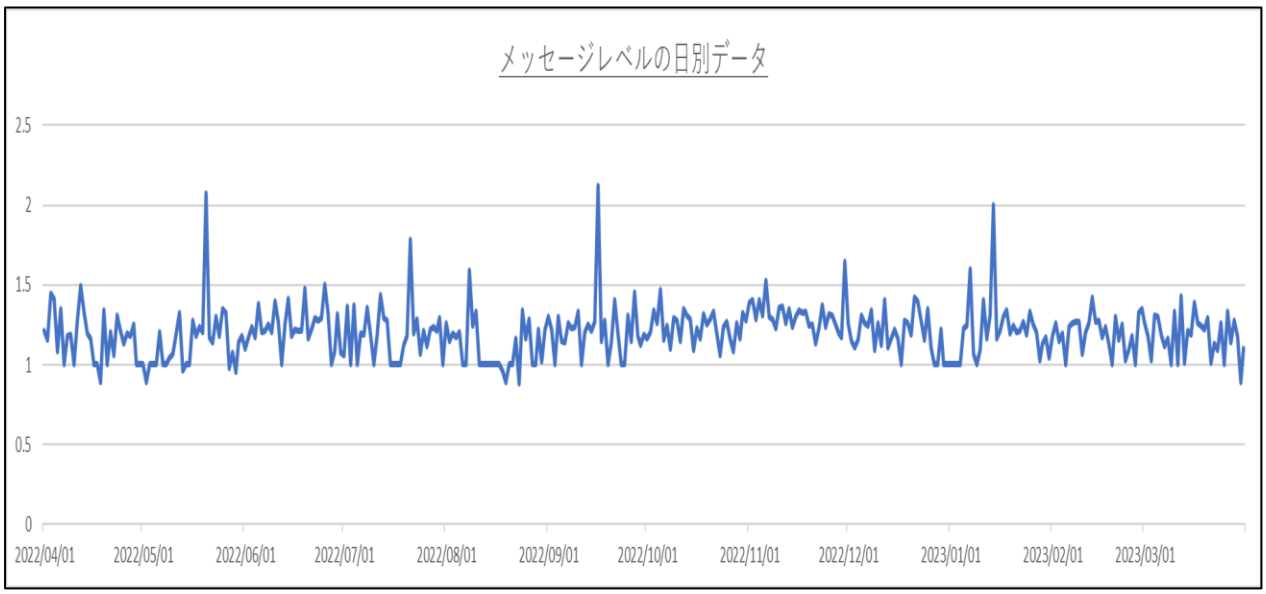
これらは、メッセージが発生した時の時間毎のデータで、点過程データと呼ばれるもので、時系列データではありません。故に、一日分の平均値を取って時系列データとして取り出す

ようしました。本データを以下に示します。2022 年 4 月　から　2023 年 3 月までのデータです。

　上段が発生時毎のデータ（点過程データ）で、1 年で 37000 回発生してます。1 点の発生は、一定時間の間隔ではありません。

　下段は点過程データを、一日毎に平均して、平均したデータです。点の間隔は 1 日毎ですので、時系列データと言えます。





1.3　プログラムについて

Python の　statsmodels.tsa.ar_model

というライブラリー内の以下の関数で計算しています。

自己回帰モデル　→　　Autoreg

予測　→　forecast , predict

95%信頼幅　→　conf_int(alpha=0.05)

最適な次数の選択　→　ar_select_order

詳細は以下で確認ください。

https://www.statsmodels.org/stable/api.html#statsmodels-api

本報告書にはプログラムは記載していません。

～参考文献、WEB～

「R による時系列モデリング入門」　北川源四郎　著　岩波書店

Statsmodels による時系列分析入門

https://qiita.com/innovation1005/items/6c5263d79ccc67263b2c

第 2 回　時系列分析～統計的手法編

https://note.com/yiida/n/n6210246ec5b0

Python で時系列分析の練習（10）予測の信頼区間をグラフに表示

https://momonoki2017.blogspot.com/2018/03/python10.html

2.1　モデリングについて

対象とする事象があるとして、真の事象と、観測できた事象があると考えます。

真の事象は未知ですので、観測できた事象から真の事象を導きます。推定と呼びます。

時系列の事象はある確率分布に従うとします。となると、真の確率分布とモデルから観測されたデータの確率分布の近さを評価し、より近いものを導くのが、推定となります。

数式で説明すると、G(y) を　確率変数 Y　の分布関数　として、

G(y) = $\int_{-\infty}^{y} g(x)dt$　で、　g(x) は密度関数と呼ばれます。

密度関数　として、色々な分布が用いられ、ここでは、正規分布で進めます。

ここで、g(y) を真のモデル（未知です）とし、データから推定された密度関数を、f(y) とします。（与えられたデータからの確率分布です）

モデリングとは、g(y)　になるべく近い f(y)　を求めることです。（＝推定）

さらに、時系列データは、時間間の相関も特徴として捉える必要があり、これを同時分布 f(y1,・・・・・・・・・・yn) といいます。

推定の手法として、最尤法（最小二乗法）というのを用います。

簡単に述べると、二つの確率分布の差異をとる尺度として、カルバック・ライブラリー情報量（K-L 情報量）を用います。K-L 情報量は小さいほど確率分布 f は g に近いとされます。

次に、K-L 情報量を、変換すると、（参考文献を参照ください）

l(θ) = $\sum_{n=1}^{N} logf(\ y_n\ |\ \theta)$　　観測値が独立の場合

l(θ) =　logf(y_n …… y_n | θ)　　観測値が独立でない場合

となり、

l(θ)を対数尤度関数と呼び、最大にするθを選ぶことで、パラメータを推定します。

この最尤法（最小二乗法）ですが、今回のプログラムでは、ライブラリーを導入して、計算しています。

2.2 自己回帰モデルについて

今回は、自己回帰モデル（AutoRegressive Model）にあてはめます。

これは、時点 t におけるモデル出力が時点 t 以前のモデル出力に依存する確率過程であり、

時系列モデル y_1 …… y_N が与えられたとき、以下の式となります。

$$y_n = \sum_{i=1}^{m} ai\, y_{n-1} + v_n$$

ここで、m は 自己回帰の次数、 a_i は 自己回帰係数です。

v_n は、平均 0 , 分散 σ^2 の正規分布に従う白色雑音とします。

与えられたデータに AR モデルをあてはめるため、次数 m を決定し、

（次数mについては、この後で説明します）

自己回帰係数$a_1, \cdots a_m$ と 分散 σ^2 を推定します。

最小二乗法による推定の概略は以下です。

パラメータは θ = $(a_1, \cdots\cdots a_m, \sigma^2)^T$ とします。

対数尤度は、 l(θ) = $\sum_{n=1}^{M} logp(\ y_n\ |y_1,\cdots, y_{n-1})$ となります。

さらに、l(θ) を変換すると、 （参考文献を参照ください）

l(θ) = -((N-M)/2)log2πσ^2 - 1/2σ^2 $\sum_{n=M+1}^{N}(\ y_n - \sum_{i=1}^{m} ai\, y_{n-i}\ )^2$

となり、l(θ) を 最大とする、パラメーターを算出します。

まずは、l(θ) を最大にする分散 σ^2 を求めると、（参考文献を参照ください）

$$\sigma^2 = \frac{1}{N-M}\sum_{n=M+1}^{N}(\ y_n - \sum_{i=1}^{m} ai\, y_{n-i}\ )^2$$ となり

自己回帰係数 $a_1, \cdots a_m$ の対数尤度は

l($a_1, \cdots a_m$) = -((N-M)/2)log2πσ^2 - (N-M)/2 となります。

分散 σ^2 を最小化する 自己回帰係数を求めることになります。

次に、次数の選択ですが、AIC という指標で行います。

統計モデルの良さを評価するための指標であるとされ、「モデルの複雑さと、データとの適合度とのバランスを取る」ために使用されます。

AIC = -2ln L + 2m (L は最大尤度、m はパラメーターの数)

であり、最小となる AIC の値を求めて、m を決めます。

2.3　予測について

AR モデル（自己回帰モデルのことですが、AR の呼称が一般的です）への推定では、次数の設定が必要で、予測の精度に関係します。

AR モデルでの次数とは、以下の式で　m　に　相当するものです。

$$y_n = \sum_{i=1}^{m} ai\, y_{n-1} + v_n$$

何回前まで戻って、推定するかということで、例えば、1 週間単位でサイクルが繰り返す事象であるなら、m=7 にすると有効です。（1 日毎のデータである場合）

次数の設定は、前節で、述べたように AIC という指標で行います。

また、statsmodels のライブラリには、最適な次数を設定する　ar_select_order という関数も用意されてます。この両者の説明をします。

AR モデルへの推定を、実施した時、その推定結果を、表示できます。以下がその例です。

今回のデータで、次数を 7 として、結果(AutoReg Model Result)を表示します。

AutoReg Model Results

Dep. Variable:	point	No. Observations:	334
Model:	AutoReg(7)	Log Likelihood	142.056
Method:	Conditional MLE	S.D. of innovations	0.157
Date:	Tue, 09 May 2023	AIC	-266.111
Time:	11:54:53	BIC	-232.002
Sample:	7	HQIC	-252.501
	334		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.6596	0.128	5.146	0.000	0.408	0.911
point.L1	0.1473	0.055	2.675	0.007	0.039	0.255
point.L2	0.1080	0.056	1.943	0.052	-0.001	0.217
point.L3	0.0123	0.056	0.220	0.826	-0.097	0.122
point.L4	0.0521	0.056	0.933	0.351	-0.057	0.161
point.L5	0.0160	0.056	0.287	0.774	-0.093	0.125
point.L6	0.0441	0.055	0.796	0.426	-0.064	0.153
point.L7	0.0716	0.055	1.303	0.193	-0.036	0.179

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.2330	-0.0000j	1.2330	-0.0000
AR.2	0.8802	-1.1361j	1.4372	-0.1451
AR.3	0.8802	+1.1361j	1.4372	0.1451
AR.4	-0.3660	-1.4640j	1.5091	-0.2890
AR.5	-0.3660	+1.4640j	1.5091	0.2890
AR.6	-1.4389	-0.5819j	1.5521	-0.4388
AR.7	-1.4389	+0.5819j	1.5521	0.4388

赤枠の箇所を補足します。

Conditional MLE　最尤法　(OLS 最小二乗法)

Log Likelihood　対数尤度

想定した回帰モデルから見て、実際に得られた標本がどの程度もっともらしいか。

大きい方が良い。

S.D. of innovations　撹乱項の標準偏差

AIC (Akaike Info. Criterion)　赤池情報量基準

回帰式の当てはまりの良さを示す。小さいほど良い。

coef (Coefficient) 回帰係数

説明変数が被説明変数に与える影響（説明変数が 1 単位変化したときに被説明変数がどれだけ変化するか）を表す係数。

std err (of Coefficients) 係数の標準誤差　二乗誤差　係数の推定値の標準誤差。

小さいほど精度の高い推定。

P>|Z|　Prob. p 値（z 検定に基づく）

説明変数として意味の無い（係数がゼロである）確率。

小さければ意味のある説明変数である（「有意」である）と判断。とりあえず、0.05 以下なら、「有意性が高い」

結果を見ると、次数を増やしても、有効ではないようです。

次に、AIC 値を　mを 1 から 30 まで、繰り返し処理を行い、算出しました。

```
lag = 1 aic : -271.2543825057729
lag = 2 aic : -273.02485400778147
lag = 3 aic : -272.0252146940184
lag = 4 aic : -270.568784450021
lag = 5 aic : -268.3808499154213
lag = 6 aic : -266.242682519484
lag = 7 aic : -266.1113208965717
lag = 8 aic : -262.3572918818377
lag = 9 aic : -259.16596807396826
lag = 10        aic : -261.1890292950411
lag = 11        aic : -257.93863149883845
lag = 12        aic : -257.56278686582255
lag = 13        aic : -254.09252568850098
lag = 14        aic : -250.61352242487828
lag = 15        aic : -247.08620261718346
```

```
lag = 16        aic : -244.6463101755822
lag = 17        aic : -242.74904115884925
lag = 18        aic : -242.84407530749058
lag = 19        aic : -243.91901234544508
lag = 20        aic : -242.9921741409339
lag = 21        aic : -239.29415600532525
lag = 22        aic : -236.12097117981472
lag = 23        aic : -233.63094124353626
lag = 24        aic : -229.96335487956514
lag = 25        aic : -226.89358379736808
lag = 26        aic : -223.37456860837824
lag = 27        aic : -220.33213328217698
lag = 28        aic : -216.84194768649047
lag = 29        aic : -214.60655753408446
lag = 30        aic : -213.3109257677773
```

AIC は、小さい程良いです。次数　1　が最適となりました。

（マイナス値ですので、絶対値は大きい方が小さいです）

　最後に、最適な次数を選択できる　ar_select_order　を用いて算出しました。

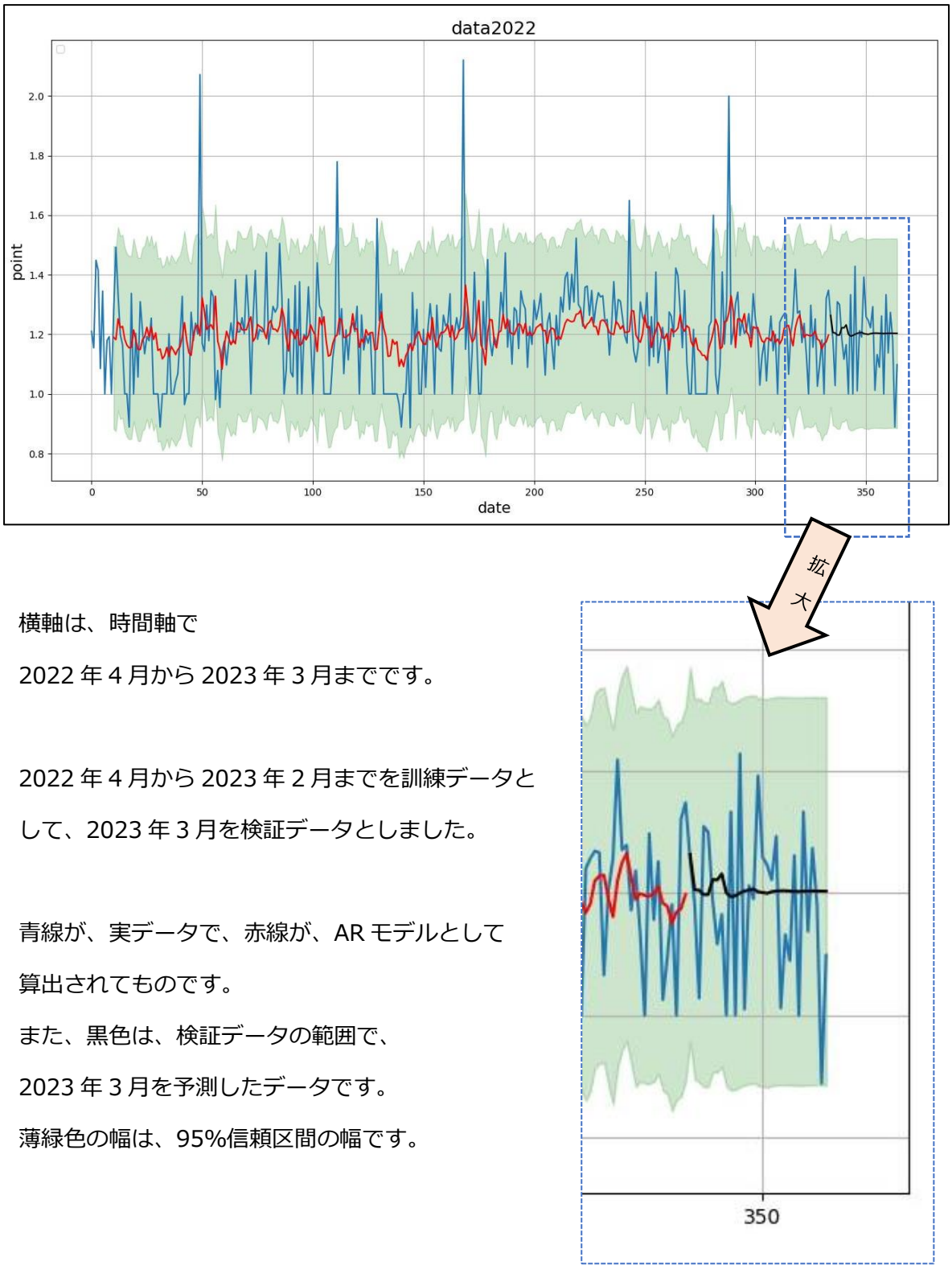
最適な次数は、[1, 2, 7, 10]　となりました。

　予測グラフに、95%信頼区間の幅を追加して表示するようにします。

95%信頼区間とは、「真値は固定されており、仮に 100 回試験をした場合、100 回中 5 回くらいは真値を含まないことがある。」と定義されますが、時系列解析においては、

95%信頼区間を正常範囲とし、これを越えたら異常と判断すると理解するようです。

ar_select_order で求めたグラフを　95%信頼区間を加えて、示します。



横軸は、時間軸で

2022 年 4 月から 2023 年 3 月までです。

2022 年 4 月から 2023 年 2 月までを訓練データとして、2023 年 3 月を検証データとしました。

青線が、実データで、赤線が、AR モデルとして算出されてものです。

また、黒色は、検証データの範囲で、

2023 年 3 月を予測したデータです。

薄緑色の幅は、95%信頼区間の幅です。

3. まとめと今後の取組

AR モデルの当てはめという点からは、うまくできませんでした。
予測もできてません。ただ、95%信頼区間の幅には入っています。
2 月までのデータでは、95%信頼区間の幅から外れたポイントが異常値と判断できる
と考えます。

モデルをより精度良くする対応として、時系列データの前処理、推定の条件の
見直しがあります。ただ、これは別途の対応とします。

今後の取組としては、状態空間モデルで、当てはめを行い。今回との比較をしてみます。