技術講座

ウェブブラウザで使用できる食肉検査データ簡易統計解析法

足 立 泰 基†

北海道八雲食肉衛生検査所（〒049-3123　二海郡八雲町立岩356）

Simple Statistical Analysis Method for Meat Inspection Data Available with Web browser

Yasumoto ADACHI†

＊*Hokkaido Yakumo Meat Inspection Center, 356 Tateiwa Yakumo-cho, Futami-gun, 049-3123, Japan*

（2020年8月3日受付・2020年12月23日受理）

衛生的な食肉生産のために，食肉検査データを疫学的に解析し，家畜衛生の向上や衛生的な食肉生産に生かす試みが広く行われている［1-4］．北海道の食肉衛生検査所では，食肉検査における廃棄数の経時変化が有意なものであるか否かを生産者が判断できるよう，疫学的な解析方法の一つである時系列分析法による分析結果を還元する試みを行ってきた［5-8］．時系列分析法とは，時間とともに出現するデータの列を統計学的に分析する方法であり，食肉検査データをこの方法で分析できるが，統計学的処理が可能なコンピュータ言語（R, Pythonなど）によるプログラムを動作させるか，SASやSPSSなどの高価な統計計算ソフトウェアを用いる必要があり，官公庁の事務処理等で用いられている表計算ソフト（Excel等）のみでは非常に困難である．しかし，インターネットサーバー上の統合開発環境（クラウドIDEとよばれ，Google Colaboratoryなどが知られている）にインターネット接続できる環境があれば，利用端末に解析用のソフトウェアをインストールしなくてもデータの解析ができる．クラウドIDEの使用目的は，プログラムの開発に必要なソフトウェア等の環境をサーバー上に整備し，ブラウザのみでプログラム開発できるようにすることであるが，同環境外で作成されたプログラムの稼働のみのために用いることもできる．この場合でも，ユーザーのPCにはブラウザのみがインストールされていればよい．ソフトウェアのインストールに制限のあるPCしか利用できない場合には，プログラムを開発するのではなく単に作動させることのみが目的であっても，クラウドIDEを利用するメリットは大きい．そこで，クラウドIDEでも作動する食肉検査データの時系列分析用プログラムを作成し，ブラウザから食肉検査の時系列分析ができることを確認したのでプログラムコード，入力データフォーマットと使用方法を公開する．この簡易統計解析法では，季節自己回帰和分移動平均（SARIMA）モデル［9］を用いている．自己回帰和分移動平均（ARIMA）モデル［9］に季節変動をモデリング要素に加えたSARIMAモデルは，時系列分析の代表的な方法であり，公衆衛生分野において時系列データの解析方法としてよく用いられている［10-20］．SARIMAモデルにはハイパーパラメータとよばれる値がモデルごとに存在し，新たなデータを解析する都度最適なハイパーパラメータの値をソフトウェアが自動的に選択するが，その際ソフトウェアが最適な値を求めるためには，ハイパーパラメータの値の上限を決めておく必要がある．本技術講座では，実際の食肉検査データを用いてハイパーパラメータ値上限の設定を行った検討結果について解説し，その後に，プログラムの使用方法と解析出力の見方について簡単に述べる．

SARIMAモデルのハイパーパラメータ値上限の設定

食肉検査データ：北海道八雲食肉衛生検査所管轄食肉

† 連絡責任者：足立泰基（北海道八雲食肉衛生検査所）
〒049-3123　二海郡八雲町立岩356　☎0137-63-2480　FAX 0137-63-2490
E-mail : adachi.yasumoto@pref.hokkaido.lg.jp

† *Correspondence to : Yasumoto ADACHI (Hokkaido Yakumo Meat Inspection Center)*
356 Tateiwa Yakumo-cho, Futami-gun, 049-3123, Japan
TEL 0137-63-2480　FAX 0137-63-2490　E-mail : adachi.yasumoto@pref.hokkaido.lg.jp

表1 データ解析用csv形式ファイルの書式

日 付	頭 数	肺 廃棄数	心臓 廃棄数	肝臓 廃棄数	大腸 廃棄数	小腸 廃棄数	腎臓 廃棄数	SEP様 肺炎	肺胸 膜炎	肺膿瘍	肝包 膜炎	肝変性	寄生虫性 肝炎	腸腺腫	腹膜炎
2003/4/1	1150	35	10	180	14	15	9	12	18	6	139	25	13	1	6
2003/5/1	1320	25	15	175	17	18	6	10	16	2	150	15	6	5	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

処理場に2003年4月～2020年4月に搬入された豚であって，36カ月間以上搬入しており，搬入開始月以降に12カ月以上の非搬入月がない24生産者由来の3,561,062頭（同期間の総搬入頭数の88.6%）の食肉検査データを用いた．食肉検査データのうち，毎月の搬入頭数，臓器廃棄数（肺，心臓，肝臓，大腸，小腸，腎臓）及び病変数（Swine enzootic pneumonia（SEP）様肺炎，肺胸膜炎，肺膿瘍，肝包膜炎，肝変性，寄生虫性肝炎，腸腺腫，腹膜炎）の14項目の値を表1に示すデータ解析用csv形式ファイルの書式に変換して保存した．

解析プログラム：以下の（a）～（c）の手順で処理するプログラムをPython3言語で記述した．

（a）上記csv形式ファイルを読み込み，初めて搬入した月を時点t=1として毎月の廃棄率（または有病率）p_t＝廃棄数(病変数)／搬入頭数を計算し，おのおののp_tを以下の（1）式，

$$logit(p_t) = ln\left(\frac{p_t}{1-p_t}\right) \quad \cdots(1)$$

によって対数ロジット変換する手順．

（b）赤池情報量規準（AIC）をもとに，対数ロジット変換されたデータへの当てはまりが最良となるSARIMAモデルを選択する手順．

（c）上記（b）で選択されたモデルを用いて，信頼区間及び最後のデータの翌月以降の対数ロジット予測値を計算し，計算結果を以下の（2）式によってロジスティック変換してパーセント値に戻し，グラフに表示する手順．

$$f(x) = \frac{e^x}{1+e^x} \quad \cdots(2)$$

ここで，xは，ロジスティック変換される数値である．

この動作を行うプログラムをGoogle Colaboratory（https://colab.research.google.com/）で動作させた．SARIMAモデルは，p：自己回帰項の数，q：移動平均項の数，d：差分処理の回数，P：季節自己回帰項の数，Q：季節移動平均の数，D：季節差分処理の回数の合計6個のハイパーパラメータとよばれる数値（0以上の整数）をもち，SARIMA$(p, d, q)(P, D, Q)_{12}$と記載される．AICの値を最小化するハイパーパラメータの組み合わせがデータ解析の都度プログラムにより自動選択されるが，あらかじめ各ハイパーパラメータに最大値を設定し，検討が必要な組み合わせを有限にする必要がある（有限でないとエラーで停止するまで計算が続く）．そこで，統計学的に妥当なモデルを95%以上の確率で選択できる最大値の組み合わせがあるか，またそれらのうち，最短時間で処理できるものがどれかを検討するため，種々の最大値の組み合わせの設定時に解析プログラムが選択したモデルのうち統計学的に妥当なものの割合を調べ（妥当性評価），解析プログラムがそれらを選択するのに要した時間（処理時間）を測定した．なお，本技術講座では，p, q, d, P, Q, Dの最大値をp_{max}，q_{max}，…，D_{max}と表した．

処理時間の測定とモデルの妥当性評価の方法：解析プログラムによってこれらのデータから以下に述べる各条件につき336モデル（24生産者×14廃棄または病変項目）を生成するための処理時間を測定し，SARIMAモデルのハイパーパラメータの最大値と処理時間の関係を調べた．具体的には（a）p_{max}, q_{max}との関係，（b）P_{max}, Q_{max}との関係，（c）d_{max}, D_{max}との関係を検討するために，以下の組み合わせで測定した．

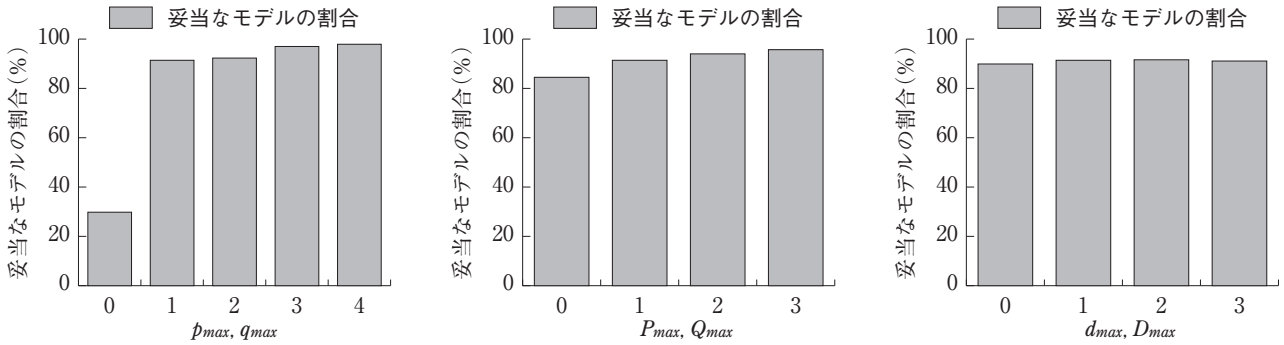
（a）p_{max}とq_{max}が同値でかつ0, 1, 2, 3, 4のいずれかであって，$P_{max}, Q_{max}, d_{max}, D_{max}$が1である場合．

（b）P_{max}とQ_{max}が同値でかつ0, 1, 2, 3のいずれかであって，$p_{max}, q_{max}, d_{max}, D_{max}$が1である場合．

（c）d_{max}とD_{max}が同値でかつ0, 1, 2, 3のいずれかであって，$p_{max}, q_{max}, P_{max}, Q_{max}$が1である場合．

2変数を同値のまま変動させている理由を考察の項で説明するが，最大値が同値であっても選択されるSARIMAモデルの多様性を損なうことはない．処理時間の測定の各条件でプログラムに選択されたモデルの妥当性を確認するため，モデルと観測値の差（残差）の時系列データについてLjung-Box検定［21］を実施した．さらに，SARIMAモデルのハイパーパラメータの最大値を変化させたときの，検定により妥当とされたモデルの割合の変化を調べた．Ljung-Box検定の計算には，python3のstatsmodelsモジュール［22］を用いた．

Ljung-Box検定は，下の（3）式で表されるQ統計量がhを時系列データの時間差としたときに，自由度hのχ^2分布することを利用したものである．



(a) p_{max} と q_{max} が同値でかつ 0, 1, 2, 3, 4 のいずれかであって，P_{max}, Q_{max}, d_{max}, D_{max} が 1 である場合

(b) P_{max} と Q_{max} が同値でかつ 0, 1, 2, 3 のいずれかであって，p_{max}, q_{max}, d_{max}, D_{max} が 1 である場合

(c) d_{max} と D_{max} が同値でかつ 0, 1, 2, 3 のいずれかであって，p_{max}, q_{max}, P_{max}, Q_{max} が 1 である場合

図 1　ハイパーパラメータの最大値と Ljung-Box 検定による妥当モデルの割合との関係

$$Q(h) = \frac{n(n+2)\sum_{j=1}^{h}\hat{\rho}^{2}(j)}{n-j} \qquad \cdots(3)$$

ここで，n は，観測値の数，$\hat{\rho}^{2}(j)$ は，時間差 j の観測値間における自己相関係数の 2 乗である．h として，statsmodels モジュールの初期設定値の 40 を用いた．

処理時間の測定とモデルの妥当性評価の成績：ハイパーパラメータの最大値と処理時間の関係を表 2 に示す．p_{max}, q_{max} との関係を表 2（a），P_{max}, Q_{max} との関係を表 2(b)，d_{max}, D_{max} との関係を表 2(c) に示した．いずれのハイパーパラメータに関しても最大値を大きくすると，処理に必要な時間が増加する傾向がみられた．d_{max}, D_{max}=2 のとき，1 生産者で計算エラーが生じ，対応する 14 モデルの計算を完了できなかった．また，d_{max}, D_{max}=3 または P_{max}, Q_{max}=3 のとき，4 生産者で計算エラーが生じ，対応する 56 モデルの計算を完了できなかった．なお，P_{max}, Q_{max}=3 でエラーを起こした 4 生産者は，d_{max}, D_{max}=3 でエラーを起こした 4 生産者と同一であり，d_{max}, D_{max}=2 でエラーを起こした生産者は，上記 4 生産者に含まれている．これらの 4 生産者を除いた場合の，処理時間について表 2(b)，2(c) の［　］内に記載した．

なお，これらの 4 生産者の搬入期間は 79，80，90，80 カ月であり，このうち d_{max}, D_{max}=2 でエラーを起こした生産者は 79 カ月搬入の生産者である．これら 4 生産者は 24 生産者中最も搬入月数の少ない 4 生産者であり，24 生産者の平均搬入期間は，171.8 カ月である．これら搬入月数の少ない 4 生産者を除いた 20 生産者の搬入頭数は，3,421,756 頭であり，全搬入頭数の 85.2%である．

ハイパーパラメータの最大値と Ljung-Box 検定による妥当モデル数（割合）との関係を表 2 及び図 1 の棒グラフによって示す．p_{max} と q_{max} を同時に変化させた場合を表 2(a) と図 1(a)，P_{max} と Q_{max} を同時に変化させた場合を表 2(b) と図 1(b)，d_{max} と D_{max} を同時に変化

表 2　ハイパーパラメータの最大値と処理時間・Ljung-Box 検定により妥当とされたモデル数（割合）の関係

(a) p_{max} と q_{max} が同値でかつ 0, 1, 2, 3, 4 のいずれかであって，P_{max}, Q_{max}, d_{max}, D_{max} が 1 である場合

p_{max}, q_{max}	妥当モデル数（割合%）	計算完了モデル数（生産者数）	処理時間（時間）
0	100(29.8)	336(24)	0.3
1	307(91.4)	336(24)	2.1
2	310(92.3)	336(24)	7.4
3	326(97.0)	336(24)	19.3
4	329(97.9)	336(24)	38.2

(b) P_{max} と Q_{max} が同値でかつ 0, 1, 2, 3 のいずれかであって，p_{max}, q_{max}, d_{max}, D_{max} が 1 である場合

P_{max}, Q_{max}	妥当モデル数（割合%）	計算完了モデル数（生産者数）	処理時間（時間）
0	284(84.5)	336(24)	0.1
	[233(83.2)]	[280(20)]	[0.1]
1	307(91.4)	336(24)	2.1
	[255(91.1)]	[280(20)]	[1.9]
2	316(94.0)	336(24)	13.8
	[263(93.9)]	[280(20)]	[13.0]
3	268(95.7)	280(20)	52.4

[　]：P_{max}, Q_{max}=3 で計算完了した 20 生産者のデータのみを処理した場合

(c) d_{max} と D_{max} が同値でかつ 0, 1, 2, 3 のいずれかであって，p_{max}, q_{max}, P_{max}, Q_{max} が 1 である場合

d_{max}, D_{max}	妥当モデル数（割合%）	計算完了モデル数（生産者数）	処理時間（時間）
0	302(89.9)	336(24)	0.4
	[249(88.9)]	[280(20)]	[0.3]
1	307(91.4)	336(24)	2.1
	[255(91.1)]	[280(20)]	[1.9]
2	295(91.6)	322(23)	7.3
	[255(91.1)]	[280(20)]	[6.9]
3	255(91.1)	280(20)	24.8

[　]：d_{max}, D_{max}=3 で計算完了した 20 生産者のデータのみを処理した場合

させた場合を表2(c) と図1(c) に示した．図1(b) と図1(c) には，計算エラーが生じたモデル（生産者・廃棄または病変項目）が含まれていない．すなわち，図1(c) における $d_{max}, D_{max}=2$ の場合，322モデル（336−14モデル，23生産者分）に含まれる妥当なモデルの割合が棒グラフとして描かれ，図1(c) における $d_{max}, D_{max}=3$ または図1(b) における $P_{max}, Q_{max}=3$ については，280モデル（336−56モデル，20生産者分）に含まれる妥当なモデルの割合が棒グラフとして描かれている．$p_{max}, q_{max}, P_{max}$ 及び Q_{max} を増加させることにより妥当とされるモデルの割合が増加する傾向がみられたが，d_{max} 及び D_{max} を2以上に増加させても妥当なモデルの数の増加はみられなかった．また，$d_{max}, D_{max}=3$ でエラーが生じなかった280の生産者・廃棄または病変項目の組み合わせに限定した場合であっても，$d_{max}, D_{max}=1, 2, 3$ のいずれにおいても妥当なモデルの割合は91.1%（255/280生産者・廃棄または病変項目）であり，d_{max}, D_{max} の最大値を大きくしても妥当モデルの増加はみられなかった．検定の結果，表2(a) のとおり P_{max}, Q_{max}, $d_{max}, D_{max}=1$，$p_{max}, q_{max}=3$ 及び4において，95%以上のモデルが妥当であり，$p_{max}, q_{max}=3$ で97.0%（326/336生産者・廃棄または病変項目，処理時間：19.3時間），$p_{max}, q_{max}=4$ で97.9%（329/336生産者・廃棄または病変項目，処理時間：38.2時間）が妥当であった．

考　　察

本検討の目的は，食肉検査データを用いてSARIMAモデルによる解析を一定のインターネット環境さえあれば場所を選ばずに実施できる方法を確立し，本方法によって得たSARIMAモデルの妥当性を確認することである．時系列分析が時間とともに出現するデータの分析方法であることから，経時的な食肉検査データの解析に適すると考えられ，SARIMAモデルはその代表的な方法であることから，北海道の一部の食肉衛生検査所では，2014年よりSARIMAモデルによる解析結果を生産者に還元しており［5］，検査員が行う操作を簡便にするために，北海道独自の様式をもつ食肉検査データベース出力をそのまま処理するプログラムを用いている．独自様式のデータベース出力に特化したこのプログラムを，他自治体で用いるにはプログラムの改造が必要であり，容易ではない．本方法では，表1の書式で作成された食肉検査データのcsvファイルを用いれば，あらゆる食肉衛生検査所の検査データを分析対象とすることができる．

本検討では，36カ月間以上搬入しており，搬入開始月以降に12カ月以上の非搬入月がない24生産者の豚の食肉検査データを用いている．従来の解析ソフト［5］を用いた実際のデータ還元事業では，搬入のない月のデータを線形補完しており，搬入のない月が上記の条件よりも多い場合であっても生産者の希望があれば分析を行っており，本方法の解析プログラムも非搬入月のデータを線形補完する機能を有している．しかし本検討の目的の一つは，解析プログラムが出力したモデルの統計学的妥当性を評価することであるから，補完されたデータが多いのは適当ではなく，上記の制限を設けた．本検討に用いられた24生産者のうち，最も非搬入月が多い2生産者は非搬入月が6カ月あったが，両生産者とも搬入期間は205カ月であり，非搬入月の占める割合は2.9%である．また，搬入期間中で最も非搬入月の占める割合が大きい生産者で，5.0%（非搬入4カ月／搬入期間80カ月）であった．

SARIMAモデルのモデリングは，解析の都度最大値以下のハイパーパラメータをもつすべてのモデルのモデルパラメータを計算し，AICなどの指標が最良のものを最終モデルとして選択するという方法がとられるのが一般的である．この最終モデルの選択は本解析プログラム中の処理で自動的に行っているが，本検討で著者の判断として行ったハイパーパラメータの「最大値の選択」とはまったく別の処理である．各所で本解析プログラムを動作する際にその都度，最大値以下の範囲内で考えられるあらゆるハイパーパラメータの組み合わせについてすべてモデリングを行い，最適の組み合わせを生産者ごと，病変ごとに，解析プログラムが自動的に選択するのである．したがって，同じハイパーパラメータの最大数を条件として選択していたとしても，生産者ごと，病変ごとに異なるモデルが選択され，しかも毎月のデータの追加によって傾向が変化した際には，それに対応して選択されるモデルも変化していく．

本検討では，ハイパーパラメータの最大値と，Google Colaboratoryでの処理に必要な時間との関係を調べた．ハイパーパラメータの数値が大きくなるにつれて，計算すべきモデルパラメータの数も増加し，計算に必要な時間が長くなると考えられる．実際に，表2に示したとおり，ハイパーパラメータの最大値を大きくすると処理に必要な時間は長くなった．モデルの選択は解析の都度行われるため，他所においてこの解析プログラムで検査結果を解析する所要時間もこの結果に準拠する．たとえば，$p_{max}, q_{max}=3$，$d_{max}, D_{max}, P_{max}, Q_{max}=1$ の条件では，表2(a) に示したように，24生産者の336モデルを最終的に選択するのに19.3時間かかっている．このことは他所において，本検討で用いたデータ長（平均171.8カ月）と同程度の長さのデータをこの条件で解析すると，1病変のモデリングに19.3時間／336モデル ≒ 3.5分程度かかることを意味している．一方，p_{max}, $q_{max}, d_{max}, D_{max}=1$，$P_{max}, Q_{max}=3$ を用いた場合，表2(b) に示したように52.4時間を要し，1病変あたりのモデリングにおよそ9.4分を要することになる．両組み合わ

せとも，妥当モデルの割合は95%を超えるが，短い時間で妥当なモデルを発見するには，妥当なモデルを高確率で求めることができる範囲で，できるだけ所要時間が短くなるハイパーパラメータの最大値の組み合わせを用いた方が実用的である．

妥当なモデルを高確率で求めることができる範囲を知るために，ハイパーパラメータの最大値を変化させたときに，Ljung-Box検定によって妥当とされたモデルが全体に占める割合がどのように変化するかを調べた．Brockwellら［23］は，「観測値からモデルによる計算値を差し引いた残差数列に自己相関が認められない場合には，残差数列を独立した乱数とみなすことができ，時系列分析としてさらにできることは，平均値と分散を計算することだけである」と述べており，残差数列の独立性を調べる方法の一つとしてLjung-Box検定をあげている．本方法で用いたPython3のstatsmodelsモジュールでこの検定を行うことができるため，妥当性の評価方法として採用した．

ハイパーパラメータの最大値を大きくすることによって妥当なモデルの割合が増加する場合もあるが，表2に示すとおり，処理の所要時間もいっそう増加することになる．妥当なモデルの割合が95%で十分であるとするならば，ハイパーパラメータの最大値をそれぞれ，$P_{max}, Q_{max}, d_{max}, D_{max}=1$，$p_{max}, q_{max}=3$とする組み合わせを選択すると，上記のとおり1病変のモデリングをおよそ3.5分程度であり，選択されたモデルの97.0%が妥当であることから，この条件が，ハイパーパラメータの最大値の最適な組み合わせであると判断した．

なお，ハイパーパラメータの最大値を変化させるにあたり，最適条件選択の流れをわかりやすくするために，関連する変数の組み合わせである（a）p_{max}, q_{max}，（b）P_{max}, Q_{max}及び（c）d_{max}, D_{max}についてそれら2個の変数の最大値を同時に変化させ，検討を行っている．ハイパーパラメータの最大値は，モデル選択時にプログラムが選択し得る最大値であるから，複数の最大値を連動させ，同時に変化させたとしても解析プログラムによって実際に選択されるハイパーパラメータは異なる値の組み合わせが含まれており，モデル選択の範囲を狭めるものではない．具体的には，$p_{max}, q_{max}=2$と設定した場合において，解析プログラムは$p=2$，$q=1$や，$p=0$，$q=1$などの組み合わせを含む最大値以下のすべての値でモデリングし，AICの値が最良のモデルを自動的に選択する．また，SARIMAモデル及びARIMAモデルの自己相関項と移動平均項は互いに変換可能であり，組み合わせて使用することによって簡潔な式で多様なモデルを記述可能にする働きがある［8］．したがって，自己相関項もしくは移動平均項のいずれか片方の最大値のみを変化させることは，モデル式による表現の多様性を損なうことにつながる可能性があり，好ましいとはいえない．

今回の検討では，P_{max}, Q_{max}を3，d_{max}, D_{max}を2または3とすることにより，プログラムでエラーが発生しているが，P, Q, Dは，それぞれ1増加するごとにモデリングに必要な時系列データが12増加するため，モデリングに必要なデータが不足したことによるエラーと考えられる．エラーが生じなかった生産者・廃棄または病変項目の組み合わせに限定すると，表2(c）に示したとおり，d_{max}, D_{max}の最大値を2以上に大きくしても妥当モデルの割合の増加はみられず，上記の最適な組み合わせにおける$d_{max}, D_{max}=1$は問題のない条件であると考えられた．

データ長とプログラムのエラー発生との関係を調べたところ，最適条件と考えられる$p_{max}, q_{max}=3$，$d_{max}, D_{max}, P_{max}, Q_{max}=1$の場合，69カ月以下のデータでエラーを発生することがわかった．しかし，Lag=40のLjung-Box検定ルーチンを除去することによって，15カ月以上であれば作動するようになるので，短いデータを処理したい場合に対応できるよう，プログラムリスト中に，Ljung-Box検定ルーチンを除去するために削除するコード部分を記載した．

以上より，一定のインターネット接続環境があれば，どこでも実施可能な方法を確立したこと，並びに本方法は適切な数値を設定することにより，得られた時系列モデルの95%以上が残差の自己相関性の点においてモデルとして妥当性を有することが明らかとなった．食肉検査データの生産者への還元事業は国内の複数の自治体において進められてはいるものの，食肉検査データを生かした調査研究は国内ではまだ少ない．本方法は，食肉検査データの経時的変化を統計学的に解析できる実用的な方法であり，さらに検討する価値のあるものと考えられるので，本方法が国内の食肉衛生検査所で広く活用に供されるよう，プログラムコードを公開することに加え，以下にプログラムの使用方法と活用方法について記載する．

プログラムの使用方法と活用方法

著者は，本方法に係る解析プログラム2種（フル仕様と簡易仕様），試用データファイルと使用説明書を食肉検査データ解析法（https://doi.org/10.13140/RG.2.2.20554.90560）としてインターネットサイト上で公開している．詳細は使用説明書に記載したので，そちらをご参照いただくとして，本稿では大まかな流れについて説明する．

Google Colaboratoryを用いるために，Googleアカウントを取得し，解析したいデータをGoogle Driveにアップロードする．Googleアカウントによる紐づけにより，アカウントごとの環境が確保され，他のアカウントにデータが漏れることはないが，生産者名等の個人情

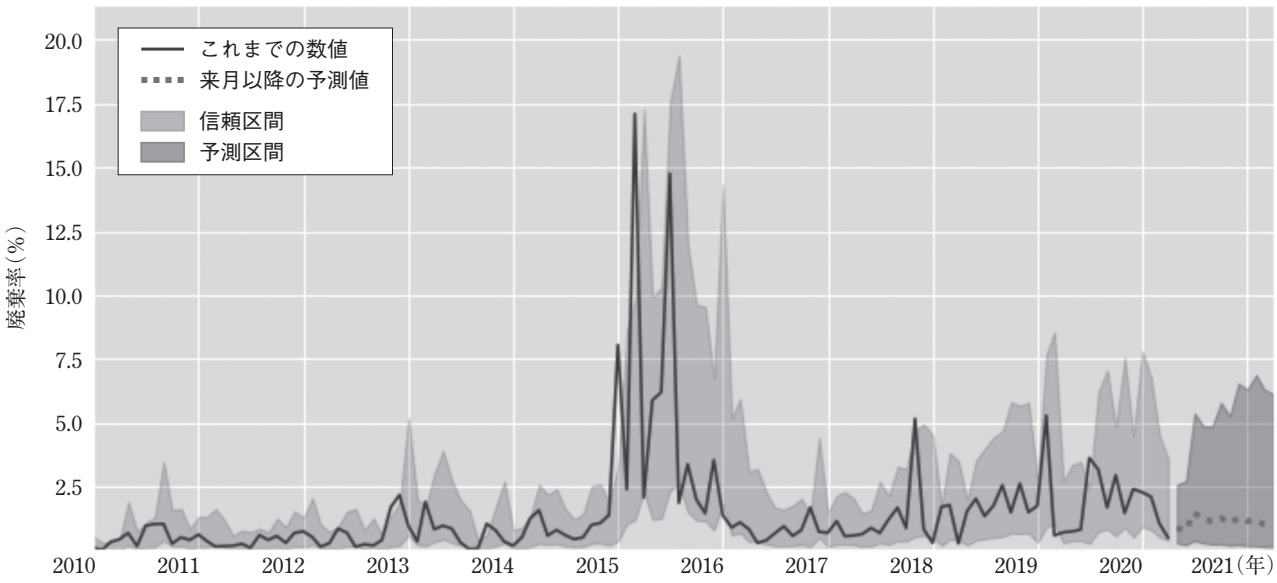


図2　Y農場産肉豚の肺膿瘍による月別廃棄率とSARIMAモデルによる予測廃棄率の推移

報が取り除かれた数値データのみをアップロードすべきである．Google Driveを用いることができない環境の場合は，ローカル環境から直接Google Colaboratoryに逐次アップロードする方法もあるが，その場合には複数ファイルの自動処理ができなくなるために，作業効率が低下する．複数ファイルの自動処理中に，データ処理担当者が他の業務を実行できる点が本プログラムの特徴の一つであり，Google Driveを使用できる環境が望まれる．

データをアップロードしたら，Google Colaboratoryの操作画面よりGoogle Driveの接続を行い，同じ画面にあるプログラムコード入力欄に，本解析プログラムのコードをそのまま入力する．最後に入力欄の左端にある起動アイコンをクリックすることによりプログラムが起動する．

簡易版（simple_version.txt）とフル仕様版（meat_inspection_analysis.txt）の2種類の解析プログラムを公開する．簡易版は最低限の機能を有しており，プログラムの解読，改造や各所における独自仕様版の作成を容易にするための短いプログラムである．Google Driveを用いることができない場合でも，簡易版を利用できる．フル仕様版では，使用マニュアルに記載したGoogle Driveの所定のフォルダ内に存在するすべての検査データを逐次読み込み，自動的にデータ解析を行う．

フル仕様版を起動すると，予測期間，グラフに表示する最初の西暦年と信頼水準の入力をプログラムが求める．これらを入力すると処理が開始される．処理結果は，Google Drive内に自動的に作成されるfigures，results_csv及びresults_paramの3つのフォルダに保存される．

figuresフォルダには時系列分析によって作成された廃棄率のグラフが保存される．results_csvには，figuresフォルダのグラフに示された信頼区間等の数値がcsv形式ファイルで保存されるので，このファイルのデータを用いたExcelやパワーポイント等による図面作成や，他のソフトウェアによるさらなる解析を自由に行うことができる．results_paramフォルダには，SARIMAモデルの各パラメータ値が保存される．

先行研究で示されているSARIMAモデルの解析結果を公衆衛生行政の判断や意思決定に生かす方法の一つとして，モデルが出力する信頼区間を逸脱する実測値が観測された場合に，何らかの異常の兆候とみなすというものがある．そのような例として，信頼区間を逸脱するカンピロバクター感染の届出数の増加から，未知の感染源の検知しようとするもの［11］や，信頼区間を逸脱する救急外来患者数の増加から，炭疽菌等によるバイオテロリズムの早期シグナルとする［19］というものがあげられる．食肉検査データに関する先行研究として，信頼区間からの逸脱から寄生虫性肝炎多発農場における投薬の有効性を確認した例［5］や，非定型抗酸菌症の農場アウトブレイクの検出方法を検討した例［6］がある．本技術講座において用いたデータの解析結果のうち，信頼区間を逸脱した例の一つとして，図2にY農場産肉豚の肺膿瘍による月別廃棄率とSARIMAモデルによる予測廃棄率の推移を示す．この例の場合は，逸脱が過去のものであるため対策を行うことはできないが，直近の値が信頼区間を逸脱している場合にこのことを生産者に情報提供できれば，生産者は迅速に衛生対策の見直しの必要性を検討できる．廃棄率値の変化のみの観察からは，平時にみられる値のばらつきなのかそれを逸脱する大きな変化であるのかを判断するのは難しいが，信頼区間を示すことにより，生産者はこれらを分別しやす

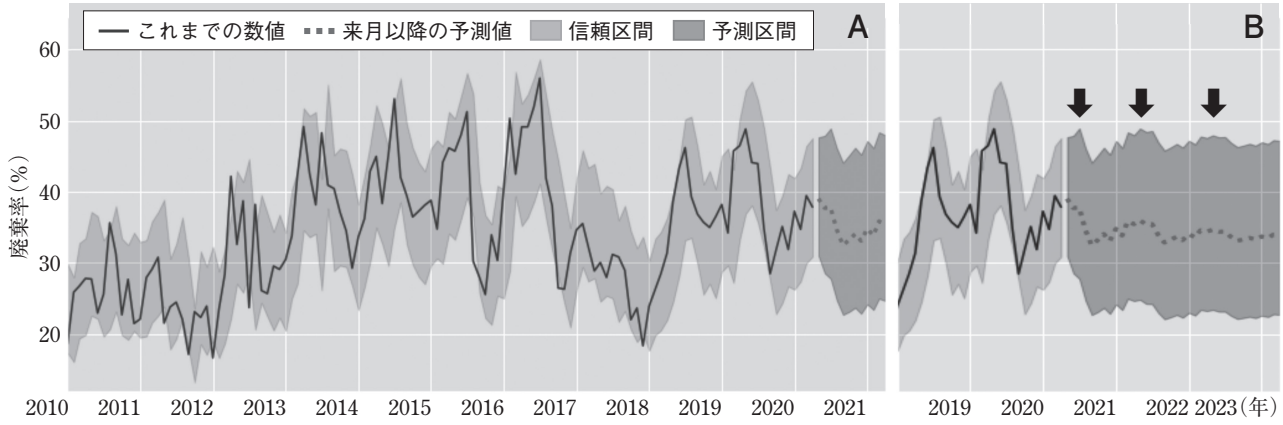


図3　Z農場産肉豚の肺の月別廃棄率とSARIMAモデルによる予測廃棄率の推移
A：予測期間12カ月での出力　　B：予測期間36カ月での出力

くなる．図2の例では，表示期間を2010年以降，予測期間を12カ月，信頼水準として先行研究［5］に示された0.85を入力しているが，廃棄率の表示期間，予測期間及び信頼水準の値を自由に入力することができる．

SARIMAモデルによる解析によって得られる他の有用な結果について，次に述べる．SARIMAモデルは，ARIMAモデルに年周期の変動要素を取り込んだものであるため，パラメータの有意性を確認すれば，有意な年周期性の有無について知ることができる．図3に，Z農場産肉豚の肺の月別廃棄率とSARIMAモデルによる予測廃棄率の推移を示す．台風の進路予測等と同様に遠い未来の予測値の信頼性は高いとはいえないため，予測期間は12カ月程度がよいであろうが（図3A），ここでは廃棄率変動の年周期性が予測値に影響するか確認するため36カ月の出力（図3B）を行ったものを示す．矢印で示した時期に予測値及び上方信頼限界の増加がみられ，年周期の変動が弱まりつつも継続すると予測されている．上で説明したresults_paramフォルダ内には，各モデルのハイパーパラメータ値等が保存されており，このモデルがSARIMA$(2, 0, 0)(1, 0, 1)_{12}$モデルであることが記録されている．SARIMA$(2, 0, 0)(1, 0, 1)_{12}$モデルは以下の(4)式で示される．

$$y_t = m + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \Phi_1 y_{t-12} + \Theta_1 a_{t-12} + a_t,\ a_t \sim N(0, \sigma^2) \quad \cdots(4)$$

ここで，$y_t = logit(p_t)$，mは定数，φ_1，φ_2は1及び2次の自己回帰項の係数，Φ_1は，1次の季節自己回帰項の係数，Θ_1は1次の季節移動平均項の係数，a_tはt時点における誤差(観測値と期待値の差)を表し，$a_t \sim N(0, \sigma^2)$は，誤差が期待値0，分散σ^2で正規分布することを示している．m，φ_1，φ_2，Φ_1，Θ_1，σ^2はモデルパラメータであり，これらの数値がresults_paramフォルダ内の生産者別ファイルに保存されている．表3に，処理後保存されていたZ農場産肉豚の肺の月別廃棄率データから推定されたSARIMAモデルのモデルパラメータ値を示す．表3の（　）内の記号は本技術講座のために表中に付したものであり，実際にはresults_param内のファイルには記録されない．coefは，それぞれのモデルパラメータの期待値であり，std errは標準誤差，zはcoefを標準化した値，$P>|z|$は標準正規分布する変数がzの絶対値よりも大きい値となる確率（いわゆるp値）を示している．ar.S.L12は，季節自己回帰項の係数であり，$P>|z|$が0.003であることから，肺廃棄率変動の年周期性は有意水準を0.05とすると有意であるということがわかる．廃棄率が年周期で変動する（すなわち季節性がある）ということは，疾病に係る飼養条件に季節変動性がある可能性を示唆している．このような情報があると飼養条件を見直すときに，廃棄率を上昇させる可能性のある数多くの要因の中で検討すべき飼養条件を季節の影響を受けるものに絞ることができる．たとえば，夏期に豚房が高温多湿になる，豚房の消毒が特定の季節に偏る，冬期に窓を閉め切るために換気が悪いなどの条件である．

表3　Z農場産肉豚の肺の月別廃棄率データから推定されたSARIMAモデルのモデルパラメータ値

	coef	std err	z	$P>\|z\|$
intercept（m）	−0.081	0.036	−2.268	0.023
ar.L1（φ_1）	0.6186	0.06	10.277	0
ar.L2（φ_2）	0.146	0.075	1.947	0.052
ar.S.L12（Φ_1）	0.5026	0.167	3.003	0.003
ma.S.L12（Θ_1）	−0.2182	0.19	−1.147	0.252
sigma2（σ^2）	0.0604	0.005	12.582	0

（　）内は，(4) 式における記号

以上，プログラムの使用方法と活用方法についてのあらましを説明した．この解析プログラムによる食肉検査データの時系列分析法が今後幅広く活用されることを願い，また，さらに有用な解析方法を求め，今後も検討を続けたい．

引 用 文 献

［1］ Goodall E, McLoughlin E, Menzies F, Mcllroy S : Time series analysis of the prevalence of Ascaris suum infections in pigs using abattoir condemnation data, Anim Sci, 53, 367–372 (1991)

［2］ Sanchez-Vazquez MJ, Nielen M, Guun GJ, Lewis FI : Using seasonal-trend decomposition based on loess (STL) to explore temporal patterns of pneumonic lesions in finishing pigs slaughtered in England, 2005–2011, Prev Vet Med, 104, 65–73 (2012)

［3］ Neumann E, Hall W, Stevenson M, Morris R, Ling Min Than J : Descriptive and temporal analysis of post-mortem lesions recorded in slaughtered pigs in New Zealand from 2000 to 2010, New Zeal Vet J, 62, 110–116 (2014)

［4］ Vial F, Reist M : Evaluation of Swiss slaughterhouse data for integration in a syndromic surveillance system, BMC Vet Res, 10, 33 (2014), (online), (https://bmcvetres.biomedcentral.com/articles/10.1186/1746-6148-10-33), (accessed 2020-08-04)

［5］ 足立泰基，蒔田浩平：季節自己回帰和分移動平均モデルによると畜検査データの時系列分析法，日獣会誌，68，189–197（2015）

［6］ Adachi Y, Makita K : Real time detection of farm-level swine mycobacteriosis outbreak using time series modeling of the number of condemned intestines in abattoirs, J Vet Med Sci, 77, 1129–1136 (2015)

［7］ Adachi Y, Makita K : Time series analysis based on two-part models for excessive zero count data to detect farm-level outbreaks of swine echinococcosis during meat inspections, Prev Vet Med, 148, 49–57 (2017)

［8］ 足立泰基：と畜検査結果の時系列分析法―疫学を利用した食肉衛生のための新たなツール―，獣医疫学雑誌，22，76–82（2018）

［9］ Box GEP, Jenkins GM : Time series analysis: Forecasting and control, rev ed, 47–412, San Francisco, Holden-Day (1976)

［10］ Ali M, Kim DR, Yunus M, Emch M : Time series analysis of cholera in Matlab, Bangladesh, during 1988–2001, J Health Popul Nutr, 31, 11–19 (2013)

［11］ Allard R : Use of time-series analysis in infectious disease surveillance, Bull World Health Organ, 76, 327–333 (1998)

［12］ Bhatnagar S, Lal V, Gupta SD, Gupta OP : Forecasting incidence of dengue in Rajasthan, using time series analyses, Indian Journal of Public Health Research & Development, 56, 281–285 (2012)

［13］ Earnest A, Chen MI, Ng D, Sin LY : Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore, BMC Health Serv Res, 5, 36 (2005), (online), (https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-5-36), (accessed 2020-08-04)

［14］ Lara-Ramírez EE, Rodríguez-Pérez MA, Pérez-Rodríguez MA, Adeleke MA, Orozco-Algarra ME, Arrendondo-Jiménez JI, Guo X : Time series analysis of onchocerciasis data from Mexico: a trend towards elimination, PLoS Neglect Trop D, 7, e2033 (2013), (online), (https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0002033), (accessed 2020-08-04)

［15］ Liu Q, Liu X, Jiang B, Yang W : Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model, BMC Infect Dis, 11, 218 (2011), (online), (https://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-11-218), (accessed 2020-08-04)

［16］ Sato RC : Disease management with ARIMA model in time series, Einstein (São Paulo), 11, 128–131 (2013)

［17］ Soebiyanto RP, Adimi F, Kiang RK : Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters, PLoS One, 5, e9450 (2010), (online), (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009450), (accessed 2020-08-04)

［18］ Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White NJ, Kaewkungwal J : Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan, Malaria J, 9, 251 (2010), (online), (https://malariajournal.biomedcentral.com/articles/10.1186/1475-2875-9-251), (accessed 2020-08-04)

［19］ Reis BY, Mandl KD : Time series modeling for syndromic surveillance, BMC med Inform Decis Mak 3, 2 (2003), (online), (https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-3-2), (accessed 2020-10-09)

［20］ Zeger SL, Irizarry R, Peng RD : On time series analysis of public health and biomedical data, Annu Rev Public Heal, 27, 57–79 (2006)

［21］ Ljung GM, Box GEP : On a measure of lack of fit in time series models, Biometrika, 65, 297–303 (1978)

［22］ Seabold S, Perktold J : Statsmodels: econometric and statistical modeling with python, Proc of the 9th python in science conf, 92–96 (2010), (online), (https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf), (accessed 2020-08-04)

［23］ Brockwell P, Davis R : Introduction of time series and forecasting, 2nd ed, 35–40, Springer, New York (2002)