広島経済大学経済研究論集
第46巻第３号　2024年３月
http://dx.doi.org/10.18996/keizai2024460303

欠損値の処理方法の違いによる予測精度の比較*

得　津　康　義**

1. はじめに

データの分析を行う際に問題となる一つに欠損値がある。日次データの場合では休日や祝日などにより，取引が行われない日のデータは欠損値となる。特に複数の時系列データを多変量時系列データとして分析するときには，ある系列ではデータが存在するが同時点の他の系列にはデータが存在しないことがしばしば発生する。多変量時系列データにおける欠損値のパターンの例として図１に示すようなものがある。

この図が示すように，為替レートはどこかの国でほぼ毎日取引があるが，国内の株価は取引所が休みの時にはデータが存在しない。その他，日本と外国とでは休日が異なるなどの理由で多変量時系列データは揃わないことがある。

共通して存在する時点だけのデータを利用する方法もあるが，低い頻度で観測されるデータでは多くのデータが削除される可能性もある。そこで欠損値のデータを何らかの方法で補完する必要が生じる。

日付	日経平均株価	為替レート	VIX（米国）
2023/5/1	29,123.18	136.73	16.98
2023/5/2	29,157.95	137.61	17.71
2023/5/3	**NA**	134.54	16.94
2023/5/4	**NA**	134.18	16.93
2023/5/5	**NA**	135.19	17.03
2023/5/7	**NA**	135.07	**NA**
2023/5/8	28,949.88	135.04	17.12

図１　多変量時系列の欠損値のイメージ図

補完方法に関する先行研究は数多くあり，辰巳・松葉（2008）では経済データにおける欠損値が生じる原因，補間の必要性，手法について紹介している。高橋・伊藤（2013）では欠損値に対して多重代入法により補完した場合の評価を行っている。田中（2020）では金融データとりわけ財務データに関して，様々な統計ソフトでの補完方法を紹介している。

本稿の目的は時系列データについて補完方法の違いによって予測精度に差が生じるかどうかを検証することである。近年，機械学習の発展により様々な分野でその威力を発揮している。とりわけ深層学習による時系列データの予測は，経済データのみに限らず他の領域においても研究がなされている。また，それらを解説するインターネットサイトや書籍も多くみられ，書籍を代表的なものとして James et al.（2021）が挙げられる。その第10章において深層学習が取り上げられ，第５節の再帰（回帰）ニューラルネットワーク（RNN）に金融データへの応用が書かれている。そこで本稿の予測精度の実験は James et al.（2021）の例題をもとに行った[1)]。本稿の構成は次の通りである。第２節では一般的に言われている欠損値のメカニズムおよび補完方法を述べる。第３節では，本稿で行う実験方法を説明し実験結果を示し，最後に結果と考察，今後の研究課題を述べる。

* 本稿を作成するにあたり，本学名誉教授である前川功一先生から有益なアドバイスを頂いた。ここに記して感謝する。
** 広島経済大学経済学部教授

2.　欠損値の発生メカニズムと処理方法

一般的に欠損値の発生メカニズムは以下の３つに分類される[2)]。

1）MCAR: Missing Complete At Random
2）MAR: Missing at Random
3）MNAR: Missing Not At Random

MCAR はデータの欠損が完全にランダムに起こるケースであり，MAR は欠損が他の要因と関係しているケースであり，MNAR は欠損がそのデータ自体に依存するケースである。

次に欠損値の処理方法[3)] であるが，大きく分けると次の２つがある。

1）リストワイズ法
2）代入法，補完方法
　➢ 単一代入法
　➢ 多重代入法

一つめのリストワイズ法では欠損値を削除して分析するが，欠損値の発生メカニズムが MAR の場合にはサンプルを削除して分析をすると推定結果が偏ることが知られている。二つ目の代入法における多重代入法についても様々の方法が存在するが本稿では扱わず，単一代入法に焦点を当てる。単一代入法の代表的な補完方法としては以下の方法がある。

1）LOCF（Last Observation Carried Forward）法：欠損値の直前の値で補完する方法。
2）平均値代入法：系列の平均値で補完する方法。
3）中央値代入法：系列の中央値で補完する方法。
4）最近傍補完法：最も近い欠測していないデータの値で補完する方法。
5）線形補完法：欠測している期間の前と後を直線（１次関数）で結び欠測値補完する方法。
6）多項式補完法：欠測している期間の前と後を多項式で結び欠測値補完する方法。
7）スプライン補完法：欠測している期間の前と後をスプライン関数で結び欠測値補完する方法。
8）移動平均補完法：移動平均を計算してその値で補完する方法。
9）カルマン平滑化補完法：状態空間モデルなど構築し欠測値補完する方法。

本稿の目的は，上記の補完方法を用いて予測を行う場合，予測精度に差があるかどうかを調べることである。予測には AR モデルに外生変数を入れたモデルについて

$$y_t = c + \alpha_1 y_{t-1} + \cdots + \alpha_k y_{t-k} + \beta_1 x_{t-1} + \cdots + \beta_k x_{t-k} + \gamma_1 z_{t-1} + \cdots + \gamma_k z_{t-k} + \varepsilon_t$$

回帰分析を使用した場合と，深層学習における RNN を用いた場合を比較した。予測精度の評価方法としては，予測値と実績値の平均２乗誤差を用いた。

ここで，RNN について James et al.（2021）を簡単に紹介する[4)]。RNN は深層学習の１つの方法であり日本語では再帰（回帰）ニューラルネットワークと訳されるか，レカレント・ニューラルネットワークと記述される。RNN は言語処理，株価データといた時系列データを扱うために利用されおり，最大の特徴はある時点の隠れ層の出力が次の時点の隠れ層に入力されることである。図２は RNN のイメージ図である。この図は RNN 単体と時間方向に展開した場合の図である。

入力層の $X_l(l=1,\cdots,L)$ は P 個のコンポーネントをもつベクトル（$X_l^T=(X_{l1},X_{l2},\cdots,X_{lp})$）と K 個の隠れ層 $A_l^T=(A_{l1},A_{l2},\cdots,A_{lk})$ がある系列とする。隠れ層 A_{lk} は次の式の通りである。

$$A_{lk} = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_{lj} + \sum_{s=1}^{K} u_{ks} A_{l-1,s}\right)$$

W は入力層における $K\times(p+1)$ 個の共有ウェ

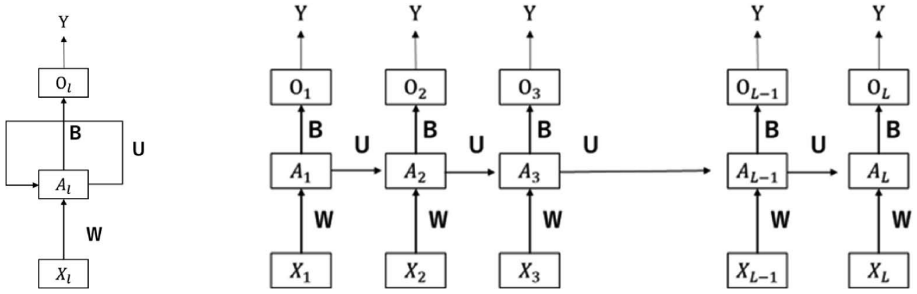

図 2　RNN のイメージ（James et al.（2021）FIGURE10.12）

イト w_{kj} の集合であり，**U** は $K \times K$ 行列であり，各要素は隠れ層から隠れ層へのウェイト u_{ks} である。さらに **B** は出力層のウェイト β_k の $K+1$ ベクトルである。上記の関数 $g(\cdot)$ は活性化関数であり，シグモイド関数や ReLU 関数を想定する。また，系列内の各要素を処理するときに，**W**，**U**，**B** は l の関数ではなく，同じ重み **W**，**U**，**B** が使用される。そして出力 O_l は以下の式によって計算される。

$$O_l = \beta_0 + \sum_{k=1}^{K} \beta_k A_{lk}$$

各パラメーターは以下の式を最小化することにより求める。

$$\sum_{i=1}^{n} \left(y_i - O_{iL}\right)^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{k=1}^{K} \beta_k g \left(\begin{matrix} w_{k0} + \sum_{j=1}^{p} w_{kj} x_i L_j \\ + \sum_{s=1}^{K} u_{ks} a_i, L-1_{,s} \end{matrix} \right) \right) \right)^2$$

3.　実　験　方　法

本稿の実験で利用するデータは2006年 1 月 4 日から2023年10月25日までの日経平均株価指数，10年物国債利回り，米国 VIX の 3 系列である。日本と米国とでは休日・祝日が異なるため欠損値の日付は異なっている。はじめに各系列のオリジナルのデータの欠損値に対して線形補完によって欠損値を埋めたデータを作成する。このように作成された系列を「真の系列」とみなし，この系列に対して，以下の手順に従って人工的に（連続的に生起する）欠損値を発生させる。欠損値が発生する場所と個数について次の 2 つのステップに従って決定する。

Step1.　50～3,000の範囲で乱数を100個発生させる。これらの乱数を使って欠損値が始まる場所を決める。

Step2.　1 ～ 5 の範囲で乱数を100個発生させる。これは欠損値の数を決める。今回は真のデータとして日次データを利用しているので最大 5 営業日まで取引がなかったことを表す。

今回の実験で利用する 3 系列に対して，欠損値の発生の仕方は次の 2 通りとする。

1．欠損値の場所も個数もすべての系列で同じ（Type_1）
2．欠損値の場所も個数も各系列で異なる（Type_2）

Type_1では，先の Step1と Step2を 1 回だけ行うのに対して，Type_2では各系列に対して Step1と Step2を行う。図表 1 は Type_1において Step1で2016年 4 月22日が決まり，Step2で欠損値の数が 3 つとなった例を示している。

図表 2 は Type_2のイメージを表している。Type2では各系列に対して Step1と Step2が行われるため，欠損値の発生時期も個数も各系列で異なる。

このような手順で発生させた欠損値を含む

元のデータ

日付	日経平均	利回り	VIX
2016/4/20	16,906.54	−0.135	13.77
2016/4/21	17,363.62	−0.12	15.22
2016/4/22	17,572.49	−0.11	15.7
2016/4/25	17,439.3	−0.075	14.68
2016/4/26	17,353.28	−0.105	15.6
2016/4/27	17,290.49	−0.06	16.05
2016/4/28	16,666.05	−0.085	15.91

欠損データ

日付	日経平均	利回り	VIX
2016/4/20	16,906.54	−0.135	13.77
2016/4/21	17,363.62	−0.12	15.22
2016/4/22	**NA**	**NA**	**NA**
2016/4/25	**NA**	**NA**	**NA**
2016/4/26	**NA**	**NA**	**NA**
2016/4/27	17,290.49	−0.06	16.05
2016/4/28	16,666.05	−0.085	15.91

図表 1　Type_1における欠損値

元のデータ

日付	日経平均	利回り	VIX
2008/5/7	14,102.48	1.65	19.73
2008/5/8	13,943.26	1.625	19.4
2008/5/9	13,655.34	1.55	19.41
2008/5/12	13,743.36	1.585	17.79
2008/5/13	13,953.73	1.58	17.98
2008/5/14	14,118.55	1.7	17.66
2008/5/15	14,251.74	1.675	16.3
2008/5/16	14,219.48	1.69	16.47
2008/5/19	14,269.61	1.66	17.01
2008/5/20	14,160.09	1.63	17.58
2008/5/21	13,926.3	1.605	18.59
2008/5/22	13,978.46	1.665	18.05

欠損データ

日付	日経平均	利回り	VIX
2008/5/7	14,102.48	1.65	19.73
2008/5/8	13,943.26	1.625	**NA**
2008/5/9	13,655.34	1.55	**NA**
2008/5/12	13,743.36	1.585	**NA**
2008/5/13	13,953.73	**NA**	17.98
2008/5/14	14,118.55	**NA**	17.66
2008/5/15	14,251.74	1.675	16.3
2008/5/16	**NA**	1.69	16.47
2008/5/19	**NA**	1.66	17.01
2008/5/20	**NA**	1.63	17.58
2008/5/21	**NA**	1.605	18.59
2008/5/22	13,978.46	1.665	18.05

図表 2　Type_2おける欠損値

データを，上記で述べた補完方法で補完して，「真の」系列とする。一連の補完作業は R ライブラリーの imputeTS を利用した。次に作成された「真の」系列に対する予測精度の比較を行うためにモデルを設定する。設定したモデルは James et al.（2021）の例題にならい，以下の AR（5）に外生変数を入れたモデル

$$\begin{aligned} y_t = {} & c + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} \\ & + \alpha_4 y_{t-4} + \alpha_5 y_{t-5} + \beta_1 x_{t-1} + \beta_2 x_{t-2} \\ & + \beta_3 x_{t-3} + \beta_4 x_{t-4} + \beta_5 x_{t-5} + \gamma_1 z_{t-1} \\ & + \gamma_2 z_{t-2} + \gamma_3 z_{t-3} + \gamma_4 z_{t-4} + \gamma_5 z_{t-5} + \varepsilon_t \end{aligned}$$

および，AR（5）に曜日ダミーを入れた二つのモデルである。James et al.（2021）の変数はニューヨーク証券取引所の取引高の対数値，ダウ平均の収益率，ボラティリティの対数値である。学習期間を1962年12月３日から1986年12月31日までで標本数は6,051として，1987年１月２日から1,770日間の予測を行っている。本稿では y_t は日経平均株価指数，x_t は10年物国債利回り，z_t は米国 VIX とした。推定期間を2006年１月４日から2017年12月29日（標本数3,128）とし，予測期間を2018年１月１日から2023年10月25日

（標本数1,518）とした。

3.1 実験結果と考察

Type_1，Type_2それぞれの結果は以下の通りで，表中の各値は 1 から RMSE の値を引いたものである。すなわち，数値が 1 に近いほど予測精度が良いことを表している。上からカルマン平滑化，線形，LOCF，移動平均，中央値，平均値，スプライン補完であり，左から AR（ar），AR+ 曜日ダミー（ard），RNN（3 変数：k），RNN（3 変数＋曜日ダミー：n）の順に並んでいる。

各モデルにおける補完方法の差をみると，AR と AR＋曜日ダミーおよび RNN（3 変数＋曜日ダミー）では中央値，平均値での補完が他の補完方法に比べて予測精度が悪くなる。中央値，平均値は系列全体から計算された中央値と平均値なので，欠損値期間における流れを上手く捉えることが出来ずにデータを補完していることが原因と考えられる。RNN（3 変数）については，他のモデルと同様な結果に加えて LOCF での予測精度も他に比べて低くなっている。ダミー変数の効果に関しては，AR と AR＋曜日ダミーは同じ結果を得られた。しかし，RNN（3 変数）と RNN（3 変数＋曜日ダミー）に関しては約 2 倍の精度の違いがみられ，RNN には曜日の情報を入れることにより予測精度が向上する結果となった。全体的な結果に関しては，本稿の例では回帰分析で AR モデルを推定して予測した方が，RNN を利用し予測をするよりもかなり精度が良いことが結果として得られた。この原因については更なる検討が必要である。

表 1　Type_1の結果

Type_1	ar	ard	k	n
カルマン	0.994	0.994	0.417	0.836
線形	0.994	0.994	0.413	0.844
LOCF	0.994	0.994	0.412	0.860
移動平均	0.994	0.994	0.410	0.832
中央値	0.979	0.978	0.358	0.814
平均値	0.970	0.970	0.375	0.848
スプライン	0.994	0.994	0.417	0.828

表 2　Type_2の結果

Type_2	ar	ard	k	n
カルマン	0.994	0.994	0.394	0.856
線形	0.994	0.994	0.394	0.854
LOCF	0.994	0.994	0.347	0.859
移動平均	0.994	0.994	0.390	0.831
中央値	0.970	0.970	0.305	0.722
平均値	0.962	0.962	0.312	0.733
スプライン	0.994	0.994	0.403	0.843

4. お わ り に

本稿では，時系列データに対して欠損値の補完方法の違いで予測精度の違いが生じるかどうかについて実験を行った。実際のデータに対して欠損値に対して線形補完を行い，作られた系列を真の時系列データとした。その系列について人工的に欠損値を発生させ，様々な補完方法でデータ補完をしたうえで，予測精度の比較を行った。時系列データはその順番と時間が重要であるため，全体のデータを使った中央値や平均値での補完は，すべてにおいて予測精度が他の補完方法よりも良くない結果が得られた。欠損値の発生方法の違いで，予測精度が異なるかどうかは，中央値や平均値での補完において違いが現れたが，他の補完方法では特に現れなかった。

今後の課題としては，回帰分析の結果と RNN での結果がかなり異なるため，その原因を探る必要がある。次に今回は自己回帰型モデルを用いたが，今回の結果を踏まえて混合頻度時系列（いわゆる MIDAS）への拡張が考えられる。データに関しては金融データ以外の経済データを利用した分析も試みる必要がある。そ

の他，深層学習に関するものとして，RNN ではダミー変数が情報として有効に利用できることが分かったため，他の有効な情報を探る必要もある。計量経済学と深層学習との関連性については，石井貴春（2020）で述べられているように共通する部分も多く，RNN への情報として時系列分析で一般的に使われている AIC を取り入れた場合に予測精度がどのようになるかを明らかにすることは重要と思われる。松浦・六井（2021）においても時系列分析でしばしば利用されているグレンジャーの因果検定を利用した変数選択の提案し予測精度の向上につながった結果を得ているため，従来の時系列分析の手法を RNN への適応は一つの方向性を見出すことができる。また RNN 以外にも LSTM（Long Short-Term Memory）など他の手法も存在するので，手法を変えた場合の違いを明らかにすることなども考えられる。

注

1） 本稿の実験は R4.3.2と torch0.10.0および impute TS3.3 を用いて行っている。また，James et al.（2021）のサンプルプログラム解説付きで次の URL にアップされている。https://hastie.su.domains/ISLR2/Labs/Rmarkdown_Notebooks/Ch10-deeplearning-lab-torch.html

2） 詳しくは Rubin（1976），福島（2015）

3） 高橋・伊藤（2013）は様々な補間方法とその問題点について述べている

4） RNN の解説は柳井・中鹿・稲葉（2022），巣籠（2018）など

参 考 文 献

［1］G. James, D. Witten, T. Hastie, R. Tibshirani（2021）**An Introduction to Statistical Learning with Applications in R**, Springer

［2］D. B. Rubin（1976）“Inference and Missing Data,” *Biometrika*, Vol. 63, 581–592

［3］石井貴春（2020）「機械学習・深層学習・計量経済学に関する先行研究の整理」『ビジネス・ブレークスルー大学レビュー』第6巻2号 p. 121–141

［4］児島利治・大橋慶介（2020）「深層学習による流量欠損値の補完方法の検討」『河川技術論文集』第26巻 p. 137–142

［5］巣鴨悠輔（2018）『詳細ディープラーニング』マイナビ出版

［6］高橋将宜・伊藤孝之（2013）「経済調査における売上高の欠損値補定方法について～多重代入法による精度の評価～」『統計研究彙報』第70号 p. 19–86

［7］辰巳憲一・松葉育雄（2008）「時系列データにおける補間方法の分析と考察」『学習院大学経済経営研究所年報』第22巻 p. 35–43

［8］田中謙一郎（2017）「金融データの欠損値補完」『西南学院大学商学論集』第64巻第3号 p. 35–54

［9］チーム・カルポ（2019）『ディープラーニングの理論と実装』秀和システム

［10］松浦匠吾・六井　淳（2021）「Recurrent Neural Network に基づく複数時系列関係を考慮した時系列予測」『FIT2021講演論文集』第2巻 p. 27–32

［11］柳井啓司・中鹿　亘・稲葉通将（2022）『深層学習』オーム社

補足　R のサンプルプログラムの一部解説

　ここでは，データの補完に利用したプログラムの一部を解説する。

```
#必要なライブラリの読み込み
library(dplyr)
library(tseries)
library(imputeTS)

#欠損値のある元のデータの読み込み
mv_data <- read.table("F:/dlst/outdata/mv_1.csv",header=T,sep=",")

#mthにはいずれかの欠損値の補完方法を入力
# locf:前の値で補完, mn:平均代入法で補完, md:中央値で補完, ln:線形補完
# spl:スプライン補完, ma:移動平均, kl:カルマン平滑法
mth <- "locf"

#補完するデータの作成
imp_data <- as.data.frame(matrix(0,nrow(mv_data),ncol(mv_data)))
colnames(imp_data) <- colnames(mv_data)
imp_data[,1] <- mv_data[,1]
imp_data[,2] <- mv_data[,2]

k <- 3    #何列目から処理を始めるか

for (i in k:ncol(mv_data)) {
  switch(mth,
  locf = { imp_data[,i] <- na_locf(mv_data[,i], option = 'locf')},
  mn = { imp_data[,i] <- na_mean(mv_data[,i], option = 'mean') },
  md = { imp_data[,i] <- na_mean(mv_data[,i], option = 'median') },
  ln = { imp_data[,i]  <- na_interpolation(mv_data[,i], option = 'linear') },
  spl = {imp_data[,i]  <- na_interpolation(mv_data[,i], option = 'spline') },
  ma = { imp_data[,i]  <- na_ma(mv_data[,i], k = 3) },
  kl = { imp_data[,i]  <- na_kalman(mv_data[,i], model ='auto.arima') }
  )
}
```