Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

Comparison of prediction using Matching Pattern and state space models

Heyam A. Hayawi and Najlaa S. Ibrahim

Department of Informatics & Statistic, College of Computer & Mathematical Science, University of Mosul, Mosul, Iraq

Article information

Article history:
Received November 21, 2021
Accepted December 18, 2021
Available online June 1, 2022

Keywords: state space, prediction, matching patterns

Correspondence:
Heyam.A. Hayawi
heyamhayawi@gmail.com

Abstract

Predicting future behaviour is one of the important topics in statistical sciences due to the need for it in different areas of life, and most countries rely on their development programs on advanced scientific foundations and methods in order to reach more effective results. This research deals with a comparison of the accuracy of time series prediction using state space models and the matching patterns method of Singh (2001) algorithm by applying to real data, which are economic observations that were previously addressed by the researchers Box and Jenkins (1976). Where the inputs represent the leading indicator and the outputs represent sales, and the importance of this research is represented in Knowing the most accurate method for predicting time series. The MATLAB program has been used to access the results of the research. The most important results of the research are that the state space model is more accurate in forecasting than the matching patterns in the studied data because it has the lowest values of the test criteria of prediction accuracy results.

DOI: https://doi.org/10.33899/iqjoss.2022.174329

1. Introduction

The philosophy of statistics lies in terms of the application mechanism to try to model different phenomena with models that are as close as possible to the actual reality. These models measure the degree of their strength according to the degree of their affinity with the statistical inferential properties. And that these models are on different forms and types, some of which are probabilistic, and that their formulation depends on pure probabilities (as in time series models).

Forecasting is one of the most important pillars in support of different planning processes. As it is not possible to complete any planning work if it is not based on scientific forecasts based on methodological methods. Therefore, the enterprises resort to choosing the most appropriate methods of forecasting from among these abundant quantities of these methods based on the extent of their needs and capabilities. Time series method is one of the most important methodology of high power which used prediction applications. The Box-Jenkins method is one of the best methodological methods for time series analysis.

The aim of the research is to procedure comparison between prediction with the best model of state space models and prediction in the manner of symmetric patterns of the algorithm Singh) 2001) according to the criteria of the prediction accuracy test)MSE, MAPE, MAE(through the application on real data

2. Local approximation using Pattern Modeling and Recognition System (PMRS)

Local approximation patterns can be used to predict future series time behavior [Singh, 2001]. When using this technique, the time series can be represented as a vector in the following way:

$$Y = (Y_1, Y_2, Y_3, \dots, Y_n) \qquad (1)$$

Therefore, if the current value of the time series when [k=1] it can be represented by the last value in the time series Y_i, which is represented by Y_n, and one of the simple methods of forecasting can be adopted

to diagnose the neighbor closest to Y_n the past values. For example, it can be said that Y_j its prediction Y_(n+1) depends on the value Y_(j+1), and it can be expanding the current value of the time series Y_n to include more than one value, for example: when [k=2] Sc will define the current values as Sc={Y_(n-1),Y_n } and represent the last two values of the time seriesY_i, and changing the direction of the current combination is called Pattern.

Therefore, the current values of the series prediction depend on the past values Sp={Y_(j-1),Y_j } and the value of the next series Y_P^+ that were given before Y_(j+1), provided that we prove that the values {Y_(j-1),Y_j } are the closest neighbor to the values {Y_(n-1),Y_n } and when some prediction measures are used, the states will be referred to as patterns. In theory, the current values can be used for any volume but in practice, current values can be represented only for the optimum size of past values of the same volume which gives a more accurate prediction whether the nearest neighbor is small or large. The prediction procedures can be illustrated according to the fuzzy symmetric pattern method through: [Singh, 2001]

$$Y\hat{}=\emptyset(Sc,Sp,Y\_P\hat{}+,K,C) \qquad (2)$$

As: Yˆrepresent predicting a step forward, Sc represent the current values, Sp represent past values, Y_P^+ represent the series value that results from the past case Sp, K represent pattern size and C represent intended to find an identical matching to the original matching.

To illustrate the matching process for predicting the time series of the future, we assume that the time series is represented as a vector since (n) the total number of observations in the series. In most cases, the series is represented as a function of time, that is , the series can be defined S={S_1,S_2,…,S_(n-1) } as the difference vector representing the transition from The case (n) to the case (n+1). As:

$$S_i = Y_{i-1} - Y_i \quad ,\forall i \ , 1 \le i \le n-1 \qquad (3)$$

The nearest neighbor is defined mathematically, the series is encoded Y as a vector with respect to the change in direction.

For this purpose, Y_i it is encoded as 0 if Y_(i+1)<Y_i as 1 if $Y_{i+1} < Y_i$ as 1 if $Y_{i+1} > Y_i$ and 2 if $Y_{i+1} = Y_i$ as follows:

$$Y_i = \begin{cases} 0 & if\ Y_{i+1} < Y_i \\ 1 & if\ Y_{i+1} > Y_i \\ 2 & if\ Y_{i+1} = Y_i \end{cases} \qquad (4)$$

Therefore, the complete time series is encoded as b_i binary values are either zero or one except in some cases where the series value does not change. We will denote it by (2). The patterns differ according to their size, but most of the time they are ($2 \le k \le 5$) and that the number of possible shapes at the font size (k) is (2k+1), knowing that there are many matching to one pattern. Patterns with small sizes will have simple and distinct shapes while patterns with large sizes will have complex shapes[Singh,2000].

3. Algorithm of Fuzzy Pattern Matching for (Singh, 2001)

This algorithm was suggested by the researcher Singh, and it is present in a lot of research McAtckney & Singh (1998) , Singh(1999) & Singh(2001) and its steps are as follows:

We start by choosing the pattern that has the lowest magnitude, that is k=2 taken from the last two values of series $P' = (b_{n-2}, b_{n-1})$.

We look at the coded time series$(b_1, b_2, \dots, b_{n-3})$ to find the closest matching for $\acute{P}$. Assume the closest matching is $P'' = (b_{j-1}, b_j)$ to group $P' \& P''$ corresponds to (S_{n-2}, S_{n-1}) & (S_{j-1}, S_j) respectively, j represents the location of the matching. Then we use the following equation: -

$$\nabla = \sum_{i=1}^{k} W_i (S_{n-i} - S_{j-i}) \qquad (5)$$

As: j represent the matching site, S represent the view difference value, k represent the size of the pattern and W_i equal to one for all sizes.

Calculating the expectation value Y_{n+1} as it depends on Eq.(5). We take the lowest value of ∇ what is the value of (j) corresponds to less ∇. if it is ($b_j = 1$) then the expectation for the value (Y_{n+1}) is higher and that:

$$Y_{n+1} = Y_n + BS_{j+1} \qquad (6)$$

If it is (b_j=0) then the expectation value is for (Y_(n+1)):

$$Y_{n+1} = Y_n - BS_{j+1} \qquad (7)$$

And if it is ($b_j = 2$) then the expectation value is for (Y_{n+1}):

$$Y_{n+1} = Y_n \qquad (8)$$

As:

$$B = \frac{1}{k}\sum_{i=1}^{k}\frac{S_{n-1}}{S_{j-1}} \qquad (9)$$

As: Y_n represent the last value of the original series. Y_{n+1} represent the value to be predicted. S_{j+1}represent the difference in viewing value (j+1), i.e. $S_{n+1} = Y_{j+2} - Y_{j+1}$. b_j represents the encoded value of the watch (j). n represent sample size. k represent size pattern, and i represents a counter for values k starting from 2 to 5.

When making steps 1 to 3 with this, we have predicted one step forward when (k = 2), and so we repeat the steps until the prediction values are completed when (k = 2).

Resize k and repeat steps 1 to 4 and after applying all the sizes for k we will use error measures to find the best size whose prediction values are close to the original values.

4. State-Space Models

The state space is a special mathematical approach to representing the different dynamic systems based mainly on the notion of the Markovian property. The representation of the state space is a mathematical model to represent a physical system as a set of inputs and outputs by means of a pair of differential differentiation equations of the first order, the first describing the input vector at time (t+1) and symbolizing it $X_{t+1}$with a significance X_t as well as the input U_t and called the state equation, while the second describes the output Y_t significance of both the inputs X_t and the inputsU_t, this equation is called the observation equation. These two equations can be clarified in the case of a single input single output (SISO), as follows:[AlKayat, 2011][Ibrahim & Hayawi, 2021]

$$X_{t+1} = AX_t + BU_t \qquad (10a)$$
$$Y_t = CX_t + DU_t \qquad (10b)$$

As: A represents the independent dynamics of the system with the dimension (n × n). B represents the effect of control verbs in the dimension (n × n). C represents the projection to the observed variables in the dimension (n×1). and D represents real value.

These two equations play an important role in the study of dynamic systems, as the inputs and outputs are expressed through a differential equation when there is a specific system in the intermittent time and there are usually uncontrolled disturbances deal as random variables or disturbances affecting the outputs[Guerbyenne & Hamdi, 2014], [Nelles, 2001]. The case space models give effective basics in time series analysis within a wide range in many fields including engineering and economic matrices and among many researchers among them [Box & Jenkines et al., 2016] and there are several reasons for using case space models because they give a sophisticated set of recursive equations that are used to find Prediction, this method is called the Kalman Filter, which helps to facilitate the calculation process to find the prediction error [Kalman, 1960]. The case space models are also known as the internal model because they are unique from other models by being in parallel with the variables that can be measured, and which cannot be measured, they are included in the model, for details[Box & Jenkines et al., 2016] [Hayawi, 2022].

5. Testing the accuracy of predictive results

Accuracy is often called the word goodness of fit, which refers to how to make the prediction model able to generate data with efficiency, and there are several criteria by which to adjust the accuracy of the model in prediction, including: [Tohme , 2012]

1- **The Mean Square Error**: is defined by the following formula:

$$\text{MSE} = \frac{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)}{n} \qquad (11)$$

Since: Y_t represents the true values of the series, $\hat{Y}_t$ represents the predicted values of the series, and n represents the prediction period.

2- **The Mean Absolute Error**: can be found by the following formula:

$$MAE = \frac{\sum_{t=1}^{n}|Y_t - \hat{Y}_t|}{n} \qquad (12)$$

3- **Mean Absolute Percentage Error:** is calculated as follows:

$$\text{MAPE} = \frac{\sum_{t=1}^{n}|Y_t - \hat{Y}_t/Y_t|}{n} * 100 \qquad (13)$$

6.The practical side

In this research, real data was used, and it is economic observations that were previously addressed by researchers Box and Jenkins (1976), which is Ut, which represents a leading indicator. and the outputs Yt represents sales and includes (150) pairs of inputs and outputs[Box & Jenkines, 1996]. The input and output data can be represented in the following two forms:

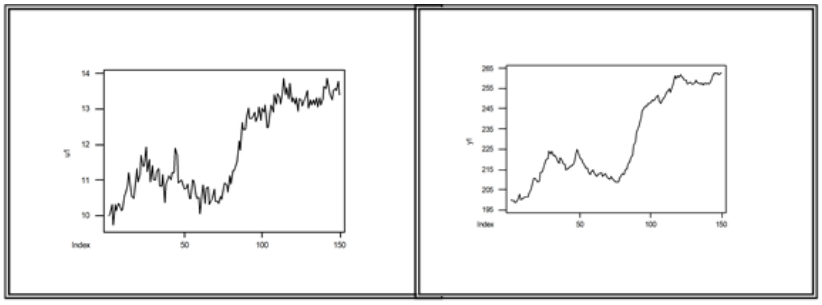

Figure 1. represents the timeline graph of the inputs, which represents the leading indicator

Figure 2. Represents the timeline of the output which represents sales

We notice from Fig. (1) and Fig. (2) that the data are unstable. Box and Jenkins (1976) researchers processed the first difference of the two series. 145 views were taken from the data and the rest was used for prediction.

A case space model was found for the data with different parameters and selecting the best model that has the lowest value for the statistical criteria as shown in table 1.

Table 1. State space models of various ranks with criteria

Standards / Ranks	AIC	Loss fun.	FPE
1	**0.5310**	**1.3924**	**1.70185**
2	**0.2839**	**0.890379**	**1.33557**
3	**-0.4758**	**0.341028**	**0.633337**
4	**-0.9273**	**0.13246**	**0.16858**
5	**0.6018**	**0.481181**	**2.4059**

We notice from Table (1) above that the best model of the fourth-level case space has the lowest AIC, Loss function and FPE standard and is formulated as follows:

$$A=\begin{bmatrix} -0.571 & 2.7689 & -0.80335 & -1.3188 & 1.1337 \\ -0.67978 & 0.87412 & -0.50666 & 0.64417 & 0.24174 \\ -2.9175 & 4.7416 & -1.9105 & -1.6517 & 2.1577 \\ -0.29977 & -0.3237 & 0.48917 & -0.25331 & -0.82795 \\ 1.4372 & -2.2158 & 0.76012 & 1.571 & -1.1798 \end{bmatrix}$$

$$B=\begin{bmatrix} 102.53 \\ 20.271 \\ 159.53 \\ 3.2534 \\ -93.239 \end{bmatrix}, \quad K=\begin{bmatrix} 25.972 \\ 7.8994 \\ 46.546 \\ -4.1389 \\ -22.47 \end{bmatrix}, \quad X(0)=\begin{bmatrix} 3.2701 \\ 1.2645 \\ 13.872 \\ -7.7074 \\ -1.3809 \end{bmatrix}$$

$$C = [1.504 \quad 3.603 \quad 2.9059 \quad -0.76076 \quad -3.1034]$$

The value of D = 0, and using the MATLAB system, the best case space model was represented by the ARMAX model formula. To ensure the accuracy of the results obtained in terms of the accuracy of the state space model, a random drawing of the remaining series can be observed, as shown in the following figure:

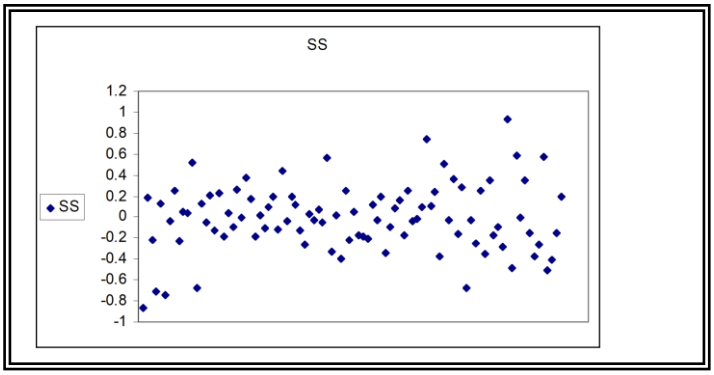

Figure 3. The random drawing the residual series of the

After that, the state space moc

Table 2. The prediction values of the state space model.

series	**Original values** Y_i	**Forecast values** $\hat{Y}_i$
146	263.3	263.7
147	262.8	262.3
148	261.8	262.8
149	262.2	262.6
150	262.7	262.2

After that, the application was made using the matching patterns method on the research data where the patterns vector was found using the Eq.(4) and the differences and the series of patterns were found for the data as shown in Table (3) the following:

Table 3. The original data with finding the series of differences and the series of patterns.

T	Y_i	Patterns Vector P'	vector of differences S	T	Y_i	Patterns Vector P'	vector of differences S
1	200.1	0	-0.6	74	209.7	0	-0.9
2	199.5	0	-0.1	75	208.8	2	0.0
3	199.4	0	-0.5	76	208.8	2	0.0
4	198.9	1	0.1	77	208.8	1	1.8
5	199	1	1.2	78	210.6	1	1.3
6	200.2	0	-1.6	79	211.9	1	0.9
7	198.6	1	1.4	80	212.8	0	-0.3
8	200	1	0.3	81	212.5	1	2.3
9	200.3	1	0.9	82	214.8	1	0.5
10	201.2	1	0.4	83	215.3	1	2.2
11	201.6	0	-0.1	84	217.5	1	1.3
12	201.5	2	0.0	85	218.8	1	1.9
13	201.5	1	2.0	86	220.7	1	1.5
14	203.5	1	1.4	87	222.2	1	4.5
15	204.9	1	2.2	88	226.7	1	1.7
16	207.1	1	3.4	89	228.4	1	4.8
17	210.5	2	0.0	90	233.2	1	2.5
18	210.5	0	-0.7	91	235.7	1	1.4
19	209.8	0	-0.1	92	237.1	1	3.5
20	208.8	1	0.7	93	240.6	1	3.2

21	209.5	1	3.7	94	243.8	1	1.5
22	213.2	1	0.5	95	245.3	1	0.7
23	213.7	1	1.4	96	246	1	0.3
24	215.1	1	3.6	97	246.3	1	1.4
25	218.7	1	1.1	98	247.7	0	-0.1
26	219.8	1	0.7	99	247.6	1	0.2
27	220.5	1	3.3	100	247.8	1	1.6
28	223.8	0	-1.0	101	249.4	0	-0.4
29	222.8	1	1.0	102	249	1	0.9
30	223.8	0	-2.1	103	249.9	1	0.6
31	221.7	1	0.6	104	250.5	1	1.0
32	222.3	0	-1.5	105	251.5	0	-2.5
33	220.8	0	-1.4	106	249	0	-1.4
34	219.4	1	0.7	107	247.6	1	1.2
35	220.1	1	0.5	108	248.8	1	1.6
36	220.6	0	-1.7	109	250.4	1	0.3
37	218.9	0	-1.1	110	250.7	1	2.3
38	217.8	0	-0.1	111	253	1	0.7
39	217.7	0	-2.7	112	253.7	1	1.3
40	215.0	1	0.3	113	255	0	-38.8
41	215.3	1	0.6	114	216.2	1	39.8
42	215.9	1	0.8	115	256	1	1.4
43	216.7	2	0.0	116	257.4	1	3.0
44	216.7	1	1.0	117	260.4	0	-0.4
45	217.7	1	1.0	118	260	1	1.3
46	218.7	1	4.2	119	261.3	0	-0.9
47	222.9	1	2.0	120	260.4	1	1.2
48	224.9	0	-2.7	121	261.6	0	-0.8
49	222.2	0	-1.5	122	260.8	0	-1.0
50	220.7	0	-0.7	123	259.8	0	-0.8
51	220	0	-1.3	124	259	0	-0.1
52	218.7	0	-1.7	125	258.9	0	-1.5
53	217	0	-1.1	126	257.4	1	0.3
54	215.9	0	-0.1	127	257.7	1	0.2
55	215.8	0	-1.7	128	257.9	0	-0.5
56	214.1	0	-1.8	129	257.4	0	-0.1
57	212.3	1	1.6	130	257.3	0	-9.7
58	213.9	1	0.7	131	247.6	1	11.3
59	214.6	0	-1.0	132	258.9	0	-1.1
60	213.6	0	-1.5	133	257.8	0	-0.1
61	212.1	0	-0.7	134	257.7	0	-0.5
62	211.4	1	1.7	135	257.2	1	0.3
63	213.1	0	-0.2	136	257.5	0	-0.7
64	212.9	1	0.4	137	256.8	1	0.7
65	213.3	0	-1.8	138	257.5	0	-0.5
66	211.5	1	0.8	139	257	1	0.6
67	212.3	1	0.7	140	257.6	0	-0.3
68	213	0	-2.0	141	257.3	1	0.2
69	211	0	-0.3	142	257.5	1	2.1
70	210.7	0	-0.6	143	259.6	1	1.5
71	210.1	1	1.3	144	261.1	1	1.8
72	211.4	0	-1.4	145	262.9	*	*
73	210	0	-0.3				

Using Singh (2001) algorithm which depend on finding the matchings of a pattern (k = 2) and the better matching gives the lowest value of the equation , and so we predict the rest of the series values, and repeats the same steps for the rest of the sizes (k = 3,4,5) so the results of the prediction using fuzzing matching patterns using this algorithm is shown in Table (4) the following:

Table 4. The forecast values of matching patterns

series	Original values Y_i	Forecast values $\hat{Y}_i$
146	263.3	255.66
147	262.8	217.45
148	261.8	227.24
149	262.2	228.52
150	262.7	228.52

By comparing the predictive values of matching patterns with the status space through the criteria used for each of them, it was noted that the best predictive values can be obtained through the status space and according to the criteria for choosing the prediction accuracy as shown in Table (5) the following:

Table 5. The values of the models quality test criteria in forecasting.

models	MSE	MAE	MAPE
matching patterns	2.25001	1.24900	86.4987
status space	0.100519	0.608259	122.657

7. Conclusions

The research concluded some conclusions, including:

It was found through practical application that the prediction of linear dynamic models, which are models of the state space, gives better predictive values than matching patterns.

The mean square error, forecast error, and test criteria for the accuracy of predictive results of linear dynamic models represented by state space models were less than in matching patterns, Which indicates the superiority of the dynamic models over the symmetric patterns and predict the future values of the case study.

The same study can be used for sophisticated matching patterns algorithms and compare their predictive values with dynamic models to see which gives better results in prediction

8. Reference

1. AlKayat, Bassel Thanoon, (2011),” Markovian modeling with practical applications” Dar Iben Alhaytham For printing and publishing,Unversity of Mosul.
2. Box,G.E.P., & Jenkines, G.M. (1976),” Time Series Analysis Forecasting and Control”, Holden Day Inc. 500 Sansome Street San Francisco California U.S.A.
3. Box, G., Jenkins, G., Reinsel ,G. and Ljung G., (2016)," Time Series Analysis Forecasting and control", John wiley & Sons , Inc . Hoboken, New Jersey.
4. Guerbyenne, H., & Hamdi, F., (2014),” Bootstrapping Periodic State-Space”, Communications in Statistics-Simulation and Computation.
5. Hayawi, H.A.A. (2022),” Using wavelet in identification state space models”, Int. J. Nonlinear Anal. (1), 2573-2578.
   Ibrahim, N.J., & Hayawi, H.A.A. (2021),” Employment the State Space and Kalman Filter Using ARMA models”, IJASEIT, 11(1), 145-149.
6. Kalman, R.E., (1960),” A New Approach to Linear Filtering and Prediction Problem Trans ASME series”, Journal of Basic Engineering (82) 35-45.
7. Nelles, O., (2001),” Nonlinear System Identification from Classical Approach to Neural Network and Fuzzy Models”, Springer Verlag Belin Heidelberg Germany.
8. Singh,S., (2001),” Multiple Forecasting using Local Approximation”, Journal Pattern Recognition,. (34) 443-455.
9. Singh ,S., (2000),” Pattern Modelling in Time-series Forecasting”, Cybernetics And Systems – An International Journal ,. (31) 1-25.
10. Tohme ,S.K., (2012),” Using Analysis of Time Series to Forecast numbers of The Patients with Malignant Tumors in Anbar Provinc”, Anbar University Journal of Economic and Administrative Sciences, Vol.4, No.8 **.**

مقارنة التنبؤ باستخدام نموذج المطابقة ونماذج فضاء الحالة

هيام عبد المجيد حياوي و نجلاء سعد ابراهيم

قسم الاحصاء والمعلوماتية ،كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق.

الخلاصة

يعتبر التنبؤ بالسلوك المستقبلي من الموضوعات المهمة في العلوم الإحصائية بسبب الحاجة إليه في مجالات الحياة المختلفة ، وتعتمد معظم الدول على برامجها التنموية على أسس وطرق علمية متطورة للوصول إلى نتائج أكثر فاعلية. يتناول هذا البحث مقارنة دقة تنبؤ السلاسل الزمنية باستخدام نماذج فضاء الحالة وطريقة أنماط المطابقة لخوارزمية Singh (2001) بالتطبيق على بيانات حقيقية ، وهي ملاحظات اقتصادية سبق أن تناولها الباحثان Box and Jenkins (1976). حيث تمثل المدخلات المؤشر الرائد والمخرجات تمثل المبيعات ، وتتمثل أهمية هذا البحث في معرفة الطريقة الأكثر دقة للتنبؤ بالسلسلة الزمنية. تم استخدام برنامج MATLAB للوصول إلى نتائج البحث. أهم نتائج البحث أن النموذج الفضائي للدولة أكثر دقة في التنبؤ من الأنماط المطابقة في البيانات المدروسة لأنه يحتوي على أدنى قيم لمعايير اختبار نتائج دقة التنبؤ.

الكلمات الدالة: فضاء الحالة، التنبؤ، الانماط المتماثلة