Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666
Vol. 5, Issue 1 (2012): 01-06

DATA MINING IN CANADIAN LYNX TIME SERIES

R.Karnaboopathy and D.Venkatesan*
Department of Statistics, Annamalai University
E Mail: *sit_san@hotmail.com

(Received January 04, 2012)

Abstract

This paper sums up the applications of Statistical model such as ARIMA family time series models in Canadian lynx data time series analysis and introduces the method of data mining combined with Statistical knowledge to analysis Canadian lynx data series.

Key Words: Data Mining; Canadian Lynx Data; Autoregressive Model; FRAR Model; Forecasting.

1. Introduction

The Time Series Data Mining (TSDM) framework is a fundamental contribution to the fields of time series analysis and data mining in the recent past. Methods based on the TSDM framework are able to successfully characterize and predict complex, non periodic, irregular, and chaotic time series. The TSDM methods over come limitations namely including stationary and linearity requirements of traditional time series analysis techniques by adapting data mining concepts for analyzing time series.

A time series $\{X_{t\ ,}$ t= 1,2.,,,,,N} is “a sequence of observed data, usually ordered in time” where t is a time index, and N is the number of observations. Researchers study systems as they evolve through time, hoping to discover their underlying principles and develop models useful for predicting or controlling them.

Traditional time series analysis methods such as the Box-Jenkins or Autoregressive Integrated Moving Average (ARIMA) method can be used to model such time series. However, the ARIMA method is limited by the requirement of stationarity of the time series, normality and independence of the residuals. Residuals are the errors between the observed time series and the model generated by the ARIMA method. The residuals must be uncorrelated and normally distributed.

For real-world time series such as stock market prices, the conditions of time series stationarity and residual normality and independence are not met. A severe drawback of the ARIMA approach is its inability to identify complex characteristics. This limitation occurs because of the goal of characterizing all time series observations, the necessity of time series stationarity and the requirement of residual normality and independence.

The TSDM framework innovate data mining concepts for analyzing time series data. This allows the TSDM methods to predict non stationary, non periodic, irregular time series. The TSDM methods are applicable to time series that appear stochastic, but occasionally (though not necessarily periodically) contain distinct, but possibly hidden, patterns that are characteristic of the desired events.

The following is the organization of the paper. Section two provides review of time series models associated with Canadian lynx data. Section three provides identifying the pattern in the Canadian lynx data through FRAR time series model. Section four provides a comparative study and summary of the paper is presented in Section five.

2. Review of Time Series Model

This section reviews some parametric models that have been fitted to the Canadian lynx data, which consists of the annual record of numbers of Canadian lynx trapped in the Mackenzie River district of North-West Canada for the period 1821-1934, both years inclusive, giving therefore a total of 114 observations.

These data originally appeared in a paper by Elton and Nicholson (1942), which gave a detailed discussion of fluctuations of the size of the lynx population in various regions of Canada over a period of some 200 years. It contains a large amount of Statistical information relating to trapping records but this was treated purely descriptively; See Moran (1953), Hannan (1960), Kashyap (1973) and Bulmer (1974).

Moran (1953) was the first to analysis the lynx data. Because the cycle in the raw data $\{X_t\}$ is very asymmetrical with a sharp and large peak and a relatively smooth and small through, he used a common log-transformation of the data and estimated the AR (2) model.

$$X_t = 1.0549 + 1.4101\ X_{t-1} - 0.7734\ X_{t-2} + e_t$$

With $e_t \sim$ WN (0, 0.04591); hereafter WN (0, σ^2) stands for a white noise process with mean 0 and variance σ^2.

Moran noting that one-step-ahead predictors for the data were not particularly good, however, he suggested that the process would be better represented by “some kind of non-linear model” Since, a fitted AR (2) model does not provide a very good match to the sample auto-covariance function, the auto-covariance function of the model damping much more rapidly than that computed from the data.

In the early days of time series analysis, owing to the limitation of computing facilities, most of the model fitted was restricted to those of very low orders. However, with the advantage of high-speed computers, there is no longer any ground for this restriction and our re-examination of the lynx data clearly re-affirms the necessity for models of the type Full Range Autoregressive models in some cases.

An extensive account of the Statistical and historical aspects of the modeling of the lynx data is contained in Campbell and Walker (1977), where the data are reproduced.

Akaike’s information criterion (AIC) and the idea of subset model selection were used first by Tong (1977) to identify and estimate several AR models for the lynx data, obtained

$$\begin{aligned} X_t = {} & 1.13\ X_{t-1} - 0.51\ X_{t-2} + 0.23\ X_{t-3} - 0.29\ X_{t-4} + 0.14\ X_{t-5} \\ & - 0.14\ X_{t-6} + 0.08\ X_{t-7} - 0.04\ X_{t-8} + 0.13\ X_{t-9} \\ & + 0.19\ X_{t-10} - 0.31\ X_{t-11} + e_t \end{aligned}$$

Where X_t= (Y_t - 2.9036) for t= 1, 2, 3, …, 114 with e_t ~ WN (0, 0.04)

Later, using the random coefficient auto regressive model, Nicholls-Quinn (NQ), obtained the model

$X_t = (1.4132 + B_1(t))\, X_{t-1} + (-0.7942 + B_2(t))\, X_{t-1} + e_t$

Where $E\left[\left(B_1(t)\ \ B_2(t)\right)'\left(B_1(t)\ \ B_2(t)\right)\right]$ and σ^2 are estimated as $\begin{bmatrix} 0.0701 & -0.0406 \\ -0.0406 & 0.0492 \end{bmatrix}$ and 0.0391 respectively.

3. The Full Range Autoregressive Model

This section provides a brief review of FRAR models that are needed here. We define a family of models by a discrete-time Stochastic process (X_t), $t = 0, \pm 1, \pm 2, \ldots$, called the Full Range Auto Regressive (FRAR) model, by the difference equation

$$X_t = \sum\nolimits_{r=1}^{\infty} a_r X_{t-r} + e_t$$

where $a_r = k\sin(r\theta)\cos(r\phi)/\alpha^r$, $(r = 1, 2, 3, \ldots)$, k, α, θ and ϕ are parameters, e_1, e_2, e_3, … are independent and identically distributed normal random variables with mean zero and variance σ^2. The initial assumptions about the parameters are as follows:

It is assumed that X_t will influence X_{t+n} for all positive *n* and the influence of X_t on X_{t+n} will decrease, at least for large *n*, and become insignificant as *n* becomes very large, because more important for the recent observations and less important for an older observations. Hence a_n must tend to zero as *n* goes to infinity. This is achieved by assuming that $\alpha > 1$. The feasibility of X_t having various magnitudes of influence on X_{t+n}, when *n* is small, is made possible by allowing *k* to take any real value. Because of the periodicity of the circular functions sine and cosine, the domain of θ and ϕ are restricted to the interval $[0, 2\pi)$.

Thus, the initial assumptions are $\alpha > 1$, $k \in R$, and θ, $\phi \in [0, 2\pi)$. i.e., $\Theta = (\alpha, k, \theta, \phi) \in S^*$, where $S^* = \{\alpha, k, \theta, \phi \mid k \in R, \alpha > 1, \theta, \phi \in [0, 2\pi)\}$. Further restrictions on the range of the parameters are placed by examining the identifiability of the model and is finally deduced that the region of identifiability of the model is given by $S = \{\alpha, k, \theta, \phi \mid k \in R, \alpha > 1, \theta \in [0, \pi), \phi \in [0, \pi/2)\}$. For more information on these topics, we recommend Venkatesan and Gallo (2012).

FRAR is fitted for the Canadian lynx data for the first one hundred observations obtained the model

$$X_t = \sum_{r=1}^{100} \frac{(10.12643)\sin(2.5611\,r)\,\mathrm{Cos}\,(0.5259\,r)}{(3.6420)^r}\, x_{t-1} + e_t$$

where θ and ϕ values are in degree and standard deviation of e_t estimated as $\hat{\sigma}$ = 0.0680.

It should be pointed out that the main purpose of time series models is to predict the future. So, the suitability of the new solution to the lynx data analysis should be examined only by the ability of the solution to correctly predict the future. This aspect is studied through the Bayesian approach. Venkatesan and Gallo (2012) have obtained the Bayesian predictive distribution of the FRAR models.

Bayesian predictive distribution of the $(r+l)^{th}$ observation, using the first r observations, is obtained. The mean of this distribution is taken to be the $(r+l)^{th}$ predicted value of the Lynx data. Since the direct evaluation of the mean of the one-step ahead predictive distribution involves four dimensional numerical integration, instead of the marginal predictive distribution of X_{N+1}, the conditional predictive distribution of X_{N+1}, given by Venkatesan and Gallo(2012) got by fixing the parameters K, α , θ and ϕ at their estimates, is used and the mean is calculated. The posterior mean of the predictive distribution is computed numerically after fixing the parameters at their respective estimated value. This prediction is done for the cases r = 11, 12, …, 114 and are given in the Table I. Table I contains both the true values and the one-step ahead predicted values for the transformed data

4. Comparative Study

Nicholls and Quinnon (1982) have used the above data to compare the quality of the predicted values obtained by several methods, viz., (1) Moran-1 (2) Tong (3) NQ-1 (4) Moran-2 and (5) NQ-2 as presented above.

Moran-1 refers to the linear predictor obtained from the second order autoregressive model, Tong refers to the linear predictor from autoregressive model of order eleventh, NQ-1 denotes the linear predictor obtained from the second order random coefficient model while Moran-2 and NQ-2 denotes the non-linear predictors for the lynx data. The models and other details can found in the Nicholls and Quinn (1982).

Nicholls and Quinn (1982) have used these methods to predict the last 14 values of the Canadian lynx data and calculated the error sum of squares. To compare the efficiency of prediction of the new FRAR model with those of the others stated above is given in Table II. The error sum of squares for the last 14 predicted values is 0.0637 under the FRAR model whereas they are 0.2531, 0.2541, 0.2561, 0.2070 and 0.1887 respectively under the other methods. So, at least in the above context the superiority of the FRAR model is established beyond doubt.

5. Conclusion

FRAR model provides the best fit for the lynx data and therefore, the FRAR model certainly provides a viable alternative to the existing time series methodology, from the predictive power of the model and from the point of view of pure data analysis, completely avoiding the problem of order determination in the case of Canadian Lynx time series data.

S.No	Y	$\hat{Y}$
1	2.430	-
2	2.506	-
3	2.767	-
4	2.940	-
5	3.169	-
6	3.450	-
7	3.594	-
8	3.774	-
9	3.695	-
10	3.411	-
11	2.718	2.582
12	1.991	1.767
13	2.265	2.181
14	2.446	2.413
15	2.612	2.650
16	3.359	3.482
17	3.429	3.468
18	3.533	3.596
19	3.261	3.182
20	2.612	2.444
21	2.179	1.999
22	1.653	1.461
23	1.832	1.801
24	2.328	2.385
25	2.737	2.839
26	3.014	3.069
27	3.328	3.380
28	3.404	3.405
29	2.981	2.849
30	2.557	2.379
31	2.576	2.500
32	2.352	2.260
33	2.556	2.569
34	2.864	2.895
35	3.214	3.296
36	3.435	3.481
37	3.458	3.449
38	3.326	3.263
39	2.835	2.668
40	2.476	2.325
41	2.373	2.283
42	2.389	2.360
43	2.742	2.726
44	3.210	3.292
45	3.520	3.569
46	3.828	3.856
47	3.628	3.542
48	2.837	2.656
49	2.406	2.252
50	2.675	2.614
51	2.554	2.481
52	2.894	2.973
53	3.202	3.248
54	3.224	3.229
55	3.352	3.344
56	3.154	3.062
57	2.878	2.765
58	2.476	2.023
59	2.303	2.255
60	2.360	2.315
61	2.671	2.672
62	2.867	2.934
63	3.310	3.466
64	3.449	3.479
65	3.646	3.684
66	3.400	3.296
67	2.590	2.399
68	1.863	1.806
69	1.591	1.454
70	1.690	1.677
71	1.771	1.766
72	2.274	2.398
73	2.576	2.642
74	3.111	3.241
75	3.605	3.683
76	3.543	3.499
77	2.769	2.589
78	2.021	1.877
79	2.185	2.105
80	2.588	2.671
81	2.880	2.963
82	3.115	3.143
83	3.540	3.633
84	3.845	3.881
85	3.800	3.713
86	3.579	3.494
87	3.264	3.249
88	2.538	2.306
89	2.582	2.547
90	2.907	2.917
91	3.142	3.204
92	3.433	3.473
93	3.580	3.562
94	3.490	3.408
95	3.475	3.406
96	3.579	3.539
97	2.829	2.663
98	1.909	1.587
99	1.903	1.833
100	2.033	2.069
101	2.360	2.439
102	2.601	2.621
103	3.054	3.108
104	3.386	3.409
105	3.553	3.528
106	3.468	3.454
107	3.187	3.150
108	2.723	2.518
109	2.686	2.646
110	2.821	2.864
111	3.000	3.053
112	3.201	3.231
113	3.424	3.464
114	3.531	3.512

Y - Lynx data (Transformed)

$\hat{Y}$ - One-step-ahead Predicted value

Table I : One-Step-ahead predicted values of the transformed Lynx data

S.No	Year	Lynx data	Moran-I	Tong	NQ-1	Moran-2	NQ-2	FRAR
1	1921	2.3598	2.4448	2.4559	2.4596	2.3835	2.3842	2.4390
2	1922	2.6010	2.7971	2.8088	2.8173	2.6271	2.6323	2.6210
3	1923	3.0538	2.8850	2.8991	2.8989	3.1193	3.0955	3.1080
4	1924	3.3860	3.3285	3.2306	3.3474	3.3883	3.3971	3.4090
5	1925	3.5532	3.4471	3.3879	3.4571	3.4955	3.4999	3.5280
6	1926	3.4676	3.4289	3.3321	3.4296	3.4787	3.4781	3.4540
7	1927	3.1867	3.1859	3.0060	3.1759	3.2683	3.2555	3.1500
8	1928	2.7235	2.8628	2.6875	2.8468	2.6405	2.6587	2.5180
9	1929	2.6857	2.4348	2.4286	2.4153	2.3747	2.3650	2.6460
10	1930	2.8209	2.7296	2.7643	2.7299	2.5977	2.6292	2.8640
11	1931	3.0000	2.9440	2.9838	2.9508	3.1277	3.0927	3.0530
12	1932	3.2014	3.0897	3.2169	3.0966	3.1981	3.1762	3.2310
13	1933	3.4244	3.2331	3.3656	3.2390	3.3065	3.2956	3.4640
14	1934	3.5309	3.3896	3.5035	3.3942	3.443	3.4413	3.5120
Error sum of squares			0.2531	0.2541	0.2561	0.2070	0.1887	0.0637

Table II: One-Step a head predictors of the transformed lynx data with other models

Acknowledgement

We are grateful to the Editor and referees for their many helpful comments that shaped the style and presentation of this paper.

References

1. Bulmer, M.G. (1974). A statistical analysis of the 10-year cycle in Canada, Journal of Animal. Ecology, 43, p. 701-715
2. Cambell, M.J and A.M. Walker (1977). A survey of statistical work on the Mackenzie River series of Annual Canadian lynx Trappings for the years 1821-1934 and a new analysis, Journal of Royal Statistical Society, A, p. 411-431.
3. Elton, C and M. Nicholson (1942). The ten year cycle in numbers of lynx in Canada, Journal of animal Ecology. 11, p. 215-244.
4. Hannan, E.J.(1960). Time Series Analysis, Methuen, London.
5. Kashyap, R.L. (1973). Validation of stochastic difference equation models for emprical time series. Proceeding 1973 Conference on Decision and control, san Diego, California, USA.
6. Moran, P.A.P (1953). The Statistical analysis of the Canadian lynx cycle, Australian Journal *of* Zoology, 1, p. 163-173.
7. Nicholls,D.F and B.G.Quinn (1982). Random Coefficient Autoregressive Models: An introduction, Springer-Verlag, New York.
8. Tong, H.(1977a). Some Comments on the Canadian lynx data, Journal of the Royal Statistical Society, A, 140, p. 432-436.
9. Tong, H.(1977b). Discussion on Stochastic modeling of river flow time series by Lawrance and Kottegoda. Journal of the Royal Statistical Society, A, p. 140.