**PROJECT 4: EXECUTIVE REPORT**

1. **PROJECT DESCRIPTION**

This project is to analyze data4 about sedimentary layer thickness and chose the model for this process, and make predictions for the future 13 years sedimentary layer thickness based on chosen model, eventually to predict the temperature change over year. Fig 4-1 shows the plot of this time series 

Figure 4-1 PROJ 4 Time Series Figure 4-2 PROJ 4 Transformed Data Figure 4-3 PROJ 4 Forecast

1. **CHOSEN MODEL**

For this dataset, after log transformation, the final chosen model is **ARIMA(1,1,1),** which means the data has a seasonal pattern, and thickness on the current year depends on thickness of previous year as well as the glacier melt of previous year and this year.

1. **PARAMETER ESTIMATION**

,

1. **MODEL EVALUATION**

The model is chosen and evaluated in many ways including checking stationary, correlogram, periodogram, MSE, Parsimony, residual diagnostic etc. See Appendix. For the finalized model, Mean Squared Error is 0.2473; Akaike Information Criterion is 856.89.

1. **FORECAST**

The future 13 years forecast result can be seen from Figure 4-3 above. We can see the prediction could match with the data pattern well. The thickness value will be around the bottom part. As the prediction time goes further away from the original data, the prediction interval would be larger; the upper interval will be larger than lower interval.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Time | 625 | 626 | 627 | 628 | 629 | 630 | 631 | 632 | 633 | 634 | 635 | 636 | 637 |
| Pred.Lower 95% | 3.88 | 3.94 | 3.93 | 3.89 | 3.85 | 3.81 | 3.76 | 3.72 | 3.68 | 3.64 | 3.60 | 3.57 | 3.53 |
| Prediction | 9.93 | 10.63 | 10.78 | 10.80 | 10.80 | 10.78 | 10.76 | 10.75 | 10.73 | 10.71 | 10.70 | 10.68 | 10.66 |
| Pred.Upper 95% | 25.39 | 28.70 | 29.61 | 30.01 | 30.28 | 30.54 | 30.78 | 31.03 | 31.27 | 31.51 | 31.75 | 31.99 | 32.23 |

Table 4-1 PROJ 4 Forecast Result

**PROJECT 3: TECHNICAL APPENDIX**

1. **Data Plot and Exploration**

Firstly, plot the data, and it has non-stationary pattern. After doing log transformation, the data becomes seasonal. After first differentiation, this is process turns to be stationary. Augmented Dickey–Fuller test also shows that the process is stationary after differentiation. For the differentiated data, plot the sample ACF, pACF, and periodogram separately and see the pattern.

* 1. **Sample Autocorrelation and Partial Autocorrelation**

Obviously, the log transformed process acf drops down slowly indicating non-stationary. From the differenced sample ACF and pACF plot, we can see that the ACF cuts off at lag 1; pACF becomes non-significant at around lag 4 and have some exponential decay. So the models could be some ARIMA model have MA(1), like ARIMA(2,1,1), ARMA(3,1,1) etc.



Figure 4-4 Proj 4 Sample ACF and pACF before and after differentiation of log-data

* 1. **Periodogram**



Figure 4-5 Proj 4 Periodogram before and after differenciation

We can see there is a peak in the end, or trough in the beginning, maybe some indication of 1 or 2 MA or AR elements., and maybe a trough around 0.2, which imply an MA(2), So we based on this plot, we can choose some models like ARIMA(1,1,2), ARIMA(1,1,1), ARIMA(0,1,1).

1. **Choose Candidate Models**

After exploration on the data above, I initially chose the ARIMA(0,1,1), ARMA(1,1,1), ARMA(2,1,1), ARMA(1,1,2), ARMA(2,1,2) to try to fit the data.

1. **Candidates models Comparison**

First, log-transformed data4 is split to training data (90%) and test data (10%). Use the training data to fit various candidate models, and use test data for prediction SSE test.

* 1. **Coefficients significance test**

In the coefficients significance test, coefficients in ARIMA(2,1,2), ARIMA(2,1,1), ARIMA(1,1,2) models are not significant. So these three models will not be considered and tested later.

* 1. **AIC**

AIC values are compared in these candidate models. ARIMA(0,1,1) has larger AIC values compared with other models. See Table4-2

* 1. **MSE**

Calculating the MSE, all models MSE are close compared with each other. See Table4-2.

* 1. **Test data Prediction SSE**

Use the models to make prediction of test data. Calculate the summation of squared error. ARIMA(0,1,1) have bad performance in this criteria. See Table4-2.

* 1. **Residual Diagnostics**

After running the residual diagnostics, all models except ARIMA(0,1,1) have random residuals, while other models residuals show some patterns or fail the randomness test. Below is the residual diagnostic of the model ARIMA(1,1,1).

 

Figure 4-6 Proj 4 Residual Diagnostics

* 1. **Periodogram match**

All models except ARIMA(2,1,2) have good fit of spectrum with original training data periodogram. See ARIMA(1,1,1) fit in Figure 4-7 below.

 

Figure 4-7 Proj 4 Sample and Theoretical Graphs Match Result

* 1. **Sample ACF and pACF match**

All models acf, pacf have good match with original sample acf and pacf. See ARIMA(1,1,1) fit in figure 4-7 above.

* 1. **Parsimony**

Out of parsimony concern, the higher order ARIMA processes are not preferred.

* 1. **Model Comparison Summary**

Summarize all the criterions above. Green cell means that the model performs good in the particular criterion.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **COMPARE Models** | **ARIMA(0,1,1)** | **ARIMA(1,1,1)** | **ARIMA(2,1,1)** | **ARIMA(1,1,2)** | **ARIMA(2,1,2)** |
| Coef Signif Test | Y | Y | N | N | N |
| AIC | 764.47 | 747.78 | 749.82 | 749.82 | 750.93 |
| MSE | 0.2276353 | 0.2208953 | 0.2216939 | 0.221694 | 0.2221279 |
| Test prediction SSE | 39.37559 | 37.79581 | 37.79001 | 37.79174 | 37.73922 |
| Residual Diagnostics | N | Y | Y | Y | Y |
| Periodogram match | Y | Y | Y | Y | N |
| Correlogram match | Y | Y | Y | Y | Y |
| Parsinomy | Y | Y | Y | Y | Y |

Table 4-2 Model Candidates Comparison

1. **Finalize Model and Forecast**
   1. Overall, ARIMA(1,1,1) performs well in all the criterions in Table4-2. Choose as the final model and fit to all the original data and make 13 points forecast, then transform the prediction to original data scale using exp(). Results are as shown in the report.

,

* 1. Calculate the MSE, AIC and residual diagnostics of the final model. All results are as well as training model above.