Development of a statistical forecast model to improve accuracy based on statistical analysis of weather historical data for the Kalmyk region

V M Goryaev¹, G Y Kazakova¹, D B Bembitov¹, E N Dzhakhnaeva² and E V Sangadjieva²

¹Gorodovikov Kalmyk State University, 11, Pushkin Ave., Elista, 358000, Russia
²Gorodovikov Kalmyk State University, 5th microregion, Elista, 358011, Russia

E-mail: goryaeff@mail.ru

Abstract. The article presents the main assumption that the averaging and smoothing models for the time series of meteorological data are local-stationary with slowly varying values. Accordingly, to estimate the current average, a local average was chosen, which was used to forecast for the near future and can be considered as a compromise between the average model and the random walk model (without drift). The same strategy can be used to evaluate and extrapolate a local trend. Average monthly values of temperature, humidity and pressure for the period from January 1928 to December 2018 were used as initial data at 4 synoptic stations. Comparative statistics on dependencies and forecasts are given. The selected method — a moving average — is often referred to as a smoothed version of the original series, since short-term averaging in a given window has the effect of smoothing irregularities in the desired series.

The objective of the study was to adjust the degree of smoothing for the moving average, as an optimal balance between the performance of averaging models and the simplicity of the random walk model or between quality and cost. The task was solved on the basis of the software implementation of the SARIMA model, which required a lengthy adjustment of the initial data and significant manipulations with time series, however, in the end, a successful model was selected.

1. Introduction

Weather forecasts typically use a 12-month weather database and their totality is called the “test reference year” (TRY) or typical meteorological year (TMY). The choice of such typical weather conditions for the given localizations is important in modeling with the weather forecast, thermal characteristics of buildings, structures and other, which ultimately gives the designers the opportunity to observe long periods of the studied data or to select the desired, typical year from the time sequence of TMY. Numerical weather prediction (NWP) data is a form of weather model data. NWP focuses on conducting current weather observations and processing these data using computer models to predict future weather conditions. Knowing the current weather conditions is just as important as numerical computer models processing the data.

Purpose of the Study is to examine the capabilities of the seasonal autoregressive integrated moving average model for weather forecasting and to identify trends in climate change in the region based on the Python 3.6 (Miniconda) code.
2. Materials and methods of research

One of the most popular and frequently used time series models is the autoregressive integrated moving average model (ARIMA) \([1–3]\). The main assumption made to implement this model is that the considered time series is linear and follows a certain known statistical distribution, such as the normal distribution. The ARIMA model has subclasses of other models, such as autoregression models (AR) \([3]\), moving average (MA) \([4]\). For seasonal forecasting of the time series of Box and Jenkins (BJ) \([5]\), they proposed a rather successful version of the ARIMA model and specifically, its seasonal version is SARIMA \([6]\). But a serious limitation of these models is the assumed linear form of related time series, which becomes inadequate in some practical situations; for this, some authors have proposed various non-linear stochastic models \([6]\), but from the point of view of implementation they are not as simple and convenient as ARIMA models. A review of the literature shows that there is no single model that consistently outperforms other models in all situations; therefore, this article attempts to compare approaches in order to find the best method for forecasting weather in the region.

The article uses data on the monthly weather data at the station Elista No. 1 from 1966 to 2018 and Station No. 2 in the period 1927-2018. For the collection of information, the values of time series from open sources are used: the laboratory of the automated information system of Roshydromet and NOAA.

Weather station Elista:
- Synoptic index: 34861, Height above sea level: 133 m
- Geographic latitude: 46.315488, longitude: 44.279401°

Initially, partial indexing and cutting of rows, sampling of time series, separation and re-sampling for different months with different aggregates were performed.

![Figure 1. Temperature chart 1967-2018 in Elista](image)

Creating a SARIMA model consists of four systematic steps (identification, evaluation, diagnostic testing and application or prognosis). The components of the series were studied and installed for stl-seasonal and trend decomposition using the LOESS method ("STL").

3. Results and discussion

For the analysis of rainfall data to build a predictive model, we used the approach of Box - Jenkins. The first-order seasonal difference is the difference between the observation and the corresponding observation of the previous year and is calculated as \(z_t = y_t - y_{t-s}\) (SARIMA \((p, d, q) \times (P, D, Q)_s\)). The daily smoothing of the hourly temperature forecast data for June 2019. The confidence intervals for the smoothed values are shown below.
On the graph, the vertical axis is represented by the following equations:

\[ C_n = \sum_{t=1}^{n-h} \frac{(y(t)-\bar{y})(y(t+n)-\bar{y})}{n}, \quad C_0 = \sum_{t=1}^{n} \frac{(y(t)-\bar{y})^2}{n} \]

The horizontal axis represents the time delay (previous time steps) \( h \)

To determine whether residues are white noise and the type of distribution, it is necessary to construct a normal residual probability graph, with the Akaike information criterion (AIC) and Schwartz Bayes criterion (SBC) being used for the criteria for selecting the model itself.

Seasonal autoregressive integrated moving average with exogenous regressors (SARIMAX) is one of the main assessment classes that can be accessed through statmodels and their result classes.

Information output on the resulting model in Table 1.

| p, q, P, Q = result_table.parameters[0] |

---

**Table 1. Model Results of SARIMA (2,1,3)x (1,1,0)_24**

| Dep. Variable: | Ads № | Synoptic number: | 34861 |
|---------------|-------|------------------|-------|
| Model: SARIMAX (2, 1, 3) x (1, 1, 0, 24) |       |                  |       |
| Date: Sat, 06 Apr 2019 |       |                  |       |

| Q-test Ljung —— box: | 54.106 | Test Jarque-Bera (JB): | 47.908 |
| Prob (F- statistics) (Q): | 0.070 | Prob(JB- statistics): | 0.000 |
| heterogeneity (H): | 0.590 | distortion: | -0.650 |
| Prob(H-statistics): | 0.040 | Excess: | 5.090 |
The result object has many attributes and methods that can be expected from other Statsmodels results, you can make a prediction and a prediction after the evaluation and the end period can be specified as a data type.

From the graph in figure 3 it follows that the model was able to successfully approximate the initial time series, fixing the daily seasonality, the general downward trend and even some anomalies. On model deviations, it can be fixed that the model reacts rather sharply to changes in the structure of the series, but then quickly returns the deviation to normal values. This feature of the model allows you to quickly create anomaly detection systems, even for noisy data without spending on data preparation and model training.

![Figure 3. Moving average and standard deviation](image)

The model parameters from table 1 are significant. Graphs of the residuals in figure 4 show that the distribution of residuals of the proposed model is Gaussian (white noise). This is clearly seen in the figure below. Therefore, the proposed model is justified, as it fits well with the test data. With the exception of some minor bursts, which lies outside the 95% confidence interval, all other points are within the confidence interval.

The predicted figures, as a rule are quite close to the actual points. The model performs its predictive function.

Figure 5 shows the results of the study.

```python
plotSARIMA(ads, best_model, 50)
```

![Figure 4. Average absolute percentage error (4.46%)](image)

As a result, adequate forecasts were obtained. This model was mistaken on average by 4.46%, which is generally not bad. However, the total cost of preparing the data, finding the series in a stationary mode and choosing the parameters may not be so costly.
During the initial selection of the delay parameters, it is necessary to find a balance between the optimal prediction quality and the length of the forecast horizon.

```python
fit1 = sm.tsa.statespace.SARIMAX(Train.Count, order=(2,1,3), seas_order=(0,1,0,24)).fit()
y_hat_avg['SARIMA'] = fit1.predict(start="2019-9-1", end="2019-9-25", dynamic=True)
plt.plot(y_hat_avg['SARIMA'], label="SARIMA")
```

![Figure 5. Forecast SARIMA (2,1,3)x (1,1,0)_{24}](image)

The forecast is shown by the red line and dots compared to the test data set from January 1967 to December 2018 by the blue line. 95% confidence interval overlapped by yellow lines. The predictive SARIMA model is a valuable tool with potential for early warning, detecting weather changes and can provide reliable information for proactive work, because the prediction values from the SARIMA model (2, 1, 3)x(1, 1, 0)_{24}, the most suitable for this model, the forecast values.

4. Conclusion

It was established that the totality of data on average temperatures in the town of Elista from 1966 to 2018 is stationary, which is confirmed by autocorrelation, partial autocorrelation plots and the Dickey Fuller test. To exclude the seasonal component in the time series of average temperatures, it was necessary to carry out one seasonal differentiation. Candidate models were developed as specified in the model building process according to the Jackins approach, and AIC values were obtained for each candidate model. The final model was chosen. SARIMA (2, 1, 3)x(1, 1, 0)_{24}.

By subtracting moving averages from baseline observations, specific seasonal values are obtained, what remains in specific seasonal variations usually represents a stationary horizontal row with two effects that cause specific seasonal variations to deviate from an absolutely straight line: seasonal effects and random error in the original observations.

References

[1] Asamoah-Boaheng M 2014 *International Journal of Statistics and Applications* **4** 292-298
[2] Thibault J C and Senocak I 2009 *Journal of Supercomputing* **9**(2) 693-719
[3] Balyani Y, Niya G F and Bayaat A 2014 *Geographic Space* **12**(38) 127-144
[4] Box G, Jenkins G and Reinsel G 2015 *Time series analysis: Forecasting and Control* (John. Wiley & Sons, Hoboken) pp 27-98
[5] Goryaev V M, Dzhahnaeva E N, Lidzhi-Goryaev V V and Al-Kilani V H 2018 *Modern knowledge intensive technologies* [in Russian – Sovremennye naukoemkie tekhnologii] **12**(1) 56-61
[6] Lazareva V G, Bananova V A and Nguen Van Zung 2017 *Advances in modern natural sciences* [in Russian – Uspehi sovremennogo estestvoznanija] **12** 178-183