Predictive models for metrological data of engineering systems

Lukankin Alexander, Slastnikov Sergey
National Research University Higher School of Economics, 101000 Moscow, Russia
E-mail: sslastnikov@hse.ru

Abstract. Paper is devoted to the predictive models for metrological indicators on the real estate engineering infrastructure. The solution is in demand among many enterprises both in terms of security and economic considerations. The key task is to build a mathematical model performing predictions on the real data samples. We study both classical predictive models (ARIMA, SARIMA) and modern machine learning based approaches (RBF, LSTM), and compare them.

Keywords: ARIMA, SARIMA, RBF, LSTM

1. Introduction
Modern tools and technologies allow data collection and real estate object information simulation. Modeling of engineering systems is an important part of building information model (BIM) teconoliges. The extracted information from the accumulated data allows building model owners to identify cases of irrational electricity consumption and save money by eliminating this problem.

Unfortunately, ordinary data collection does not allow to predictive estimate the real property state. Using of modern forecasting tools could solve this problem, as well as expand the range of applicability areas of building information models. For example, considering sets of archived and real-time data describing the building’s power supply system provides possibility to predict short-circuit situations and carry out preventive maintenance in time. This example describes the applicability of complex analytical results in the field of building safety.

This investigation is aimed at identifying the best models for time series forecasting using metrological data from the building's ventilation system as an example.

The paper builds econometric models for time series forecasting (ARIMA and SARIMA) and neural network models (RBF and LSTM). The predictive performance of each model is evaluated by some metrics. The final stage is a comparative analysis of the metrics values for each model.

The article has the following structure. The next section reviews the literature covering the subject area under consideration. The tools, methods, and metrics used in this study are described below. The final part summarizes the results and discusses the prospects for the study development.

2. Short literature review
The classical approach for time series analysis is described in the famous book [1]. This reference offers the basic principles of constructing econometric forecasting models. In particular, the ARIMA and SARIMA models are demonstrated.

The article [2] presents the construction of an ARIMA model for monitoring the occupied beds number in a Singapore hospital. According to this study, the successful application of the econometric model to the problem of time series forecasting is shown.
The practical SARIMA model implementation is described in detail in the article [3]. The researchers note the fact of obtaining successful results of the constructed SARIMA model in the subject area they study.

Fundamental scientific paper on the theory of neural networks is described in a book [4]. The reference explains in detail the classes of direct signal propagation networks and recurrent networks. The book also discusses in detail neural network models based on RBF function.

The study [5] presents the implementation of the LSTM neural network model. The resulting model has successfully proven itself in solving the problem related to time series forecasting.

3. Problem description

In this study, we consider a metrological data set describing the state of the building’s ventilation system. The dataset components are sets of observations of several physical quantities in particular outdoor air temperature, exhaust air temperature, supply air temperature, temperature before the recuperator, temperature after the recuperator, temperature of the return carrier, illumination in the ventilation chamber, pressure in the return pipeline, pressure in the direct pipeline. Quantities are characterized by the according observations time series.

Monitoring and predictive analysis of ventilation system characteristics values aimed to solve key problems such as cost estimation, competent management, and energy consumption planning. Moreover, such approach is extremely useful in the field of building safety since it is able to assess the future of the object under study.

The described problem can be formalized as follows. There are \( N \) data sets \( \{ X_i \in \mathbb{R}^n | i = 1, \ldots, N \} \) and observations \( \{ Y_i \in \mathbb{R}^m | i = 1, \ldots, N \} \). Required to find a form transformation:

\[
F: X_i \rightarrow Y_i, \forall i \in [1, \ldots, N] \tag{1}
\]

Expression (1) can also be represented as

\[
y_i = F(x_i), \forall i \in [1, \ldots, N] \tag{2}
\]

Data accessed in real time using is provided by the API. The uploaded data slice covers the period from 11.12.2019 10:00 to 26.02.2020 14:00. The main statistical characteristics used for the initial data estimation as expectation, standard deviation, minimum value, maximum value presented in table 1.

| Statistics name                | Mathematical expectation | Standard deviation | Minimum value | Maximum value |
|--------------------------------|--------------------------|--------------------|---------------|---------------|
| Outdoor temperature            | 559.49                   | 99.24              | 429.14        | 1186.32       |
| Exhaust air temperature        | 752.77                   | 11.77              | 733.73        | 780.71        |
| Supply air temperature         | 718.90                   | 8.08               | 703.34        | 740.71        |
| Pre-recuperator temperature    | 732.58                   | 16.83              | 694.42        | 755.23        |
| Post-recuperator temperature   | 704.00                   | 28.31              | 625.98        | 736.55        |
| Reverse carrier temperature    | 800.52                   | 15.45              | 755.42        | 871.22        |
| Illumination                   | 0.0                      | 0.0                | 0.0           | 0.0           |
| Return pipeline pressure       | 1552.17                  | 186.22             | 1219.21       | 2006.18       |
| Forward pipeline Pressure      | 1733.26                  | 184.75             | 1424.48       | 2190.15       |

4. Models and methods

4.1. Econometric models

4.1.1. ARIMA

The following conditions must be met to build the ARIMA model: \( E X_t = E X_{t+\tau}, DX_t = DX_{t+\tau}, \text{Cov}(X_t, X_s) = \text{Cov}(X_{t+\tau}, X_{s+\tau}) \), \forall t, s, \tau. ARIMA model parameters p, q, and d define its unique instance. The model can be represented as follows

\[
\left( I - \sum_{i=1}^{p} \phi_i L^i \right) \left( I - L \right)^d y_t = \alpha + \left( I - \sum_{j=1}^{q} \theta_j L^j \right) \varepsilon_t, \tag{3}
\]
where $y_t$ is the original non-stationary series $y$ value at time $t$, $(1 - L)^d$ is the order $d$ difference operator, $\alpha$, $\phi_l$, $\theta_l$ are model parameters, $\varepsilon_t$ is white noise. The order $d$ difference operator allows you to bring the original series to a stationary by the following rule:

$$L : Ly_t = y_{t-1},$$

where $y_t$ is the original non-stationary series $y$ value at time $t$, $L$ is lag operator. The detection of stationarity of series is performed by using the Dickie Fuller criterion.

$$\Delta y_t = b \cdot y_{t-1} + \varepsilon_t,$$

where $b$ is a linear coefficient which value is checked, $\Delta y_t = y_t - y_{t-1}$.

4.1.2. SARIMA

The SARIMA model is an extension of the ARIMA model with the addition of seasonal components. The model is defined as

$$(1 - \sum_{l=1}^{p} \phi_l L^l)(1 - \sum_{l=1}^{p} \phi_l L^l)(1 - L)^d y_t = \alpha + (1 - \sum_{k=1}^{q} \theta_k L^{2k})(1 - \sum_{l=1}^{q} \theta_l L^l) \varepsilon_t,$$

where $m$ is season period, $p$, $d$, $q$, $P$, $D$, $Q$ are model parameters. This obviously follows the way the model is defined: SARIMA($p,d,q$)$x(P,D,Q)$.

4.2. Neural network models

4.2.1. RBF

The network of radial basis functions (RBF) is mathematical conception subtype of artificial neural networks. This neural network model type can be classified such as direct distribution network of the input signal. RBF network activation function uses radial basis functions. The model weight parameters are adjusted using the error back propagation algorithm. A single-layer network model can be represented as

$$\hat{y} = \varphi(\vec{w}^T \vec{x}),$$

where $\hat{y}$ is the output values vector (observations), $\vec{x}$ is input values vector (the state of the system), $\vec{w}$ is weight coefficient vector, $\varphi$ is activation function.

4.2.2. LSTM

Models based on LSTM architecture can be classified as recurrent networks. In such models an output signal vector is input signal vector. The LSTM main distinguish among other recurrent models is ability to store key values in the memory for different time periods. RNN primary element is cell. LSTM cell can by formalized as equation system (8):

$$\begin{align*}
    f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
    i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
    C_t &= \tanh(W_C x_t + U_C h_{t-1} + b_C) \\
    C_t &= f_t \circ C_{t-1} + i_t \circ \hat{C}_t \\
    o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
    h_t &= o_t \circ \tanh(C_t)
\end{align*}$$

where $x_t$ is input signal vector, $h_t$ output signal vector, $C_t$ is states vector, $W$, $U$ are parameters matrix, $b$ is parameters vector, $f_t$ is forget filter vector, $i_t$ is input filter vector, $o_t$ is output filter vector, $\circ$ is Hadamard product operator.

4.3. Forecast quality metrics

A quality of model forecasting estimates by three metrics. First of this is $R^2$ score which defines as determination coefficient. Value calculation this metric realized by following rule
\[ R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \]  
(9)

The second metric which is used in the research named MSE. Following formula using for calculating the metric value

\[ MSE = \frac{1}{n} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \]  
(10)

The latter metric referred to as Mean Absolute Error. The formula by which it can be computed is presented below

\[ MAE = \frac{1}{n} \sum_{i=1}^{n}|y_i - \hat{y}_i|, \]  
(11)

In formulas (9) – (11) \( y_i \) is i-th test sample value, \( \hat{y}_i \) is model evaluation of the i-th test sample value, \( \bar{y} \) is average sample value.

5. Results

This paper research data describing the ventilation system characteristics. The data collected amount is 1,717 measurements for each of the 9 sensors. Each measurement is recorded once per hour. The time series values collected describes 71 days. This data is used to construct forecasting models. The source time series are regrouped before simulating by neural network models. However, there is no need for that kind of preprocessing raw information if econometric models are constructed. The resulting sample quantity consisted of 758 training pairs and 1 test pair. Each pair contains an input vector of size (1, 2160) and a reference vector of size (1, 2160). The data preprocessing is done in a way that the RBF and LSTM models could make forecasts for 10 days ahead using data from the past 10 days.

5.1. Results of building econometric model

5.1.1. ARIMA

Modeling showed that it is possible to achieve the Dickey-fuller test by difference operator conversion for metrological data. However, model forecasts constructed with the parameter \( d=1 \) or \( d=2 \) does not show a noticeable improvement in the quality of the forecast. Investigation also discover that the difference transformation does not lead the processes to a stationary form. This fact can be explained by the fact that the condition of stationarity in the broad sense is not fulfilled for different difference operator values. Simulations have done with using parameter \( d \) equals 0. In addition, the correlogram study shows the expected data seasonality, which is well simulated using SARIMA models. The constructed ARIMA models and their accuracy estimations are shown in table 2.

| Model number | Model name | Model parameters | Ten-day forecast |
|--------------|------------|------------------|-----------------|
|              |            | \( p \) \( d \) \( q \) | \( R^2 \) MSE MAE |
| 1            | ARIMA 1    | 2 0 2            | 0.84 38129.84   | 127.28 |
| 2            | ARIMA 2    | 1 0 2            | 1.00 79.94     | 7.51  |
| 3            | ARIMA 3    | 2 0 2            | 0.69 65.85     | 5.66  |
| 4            | ARIMA 4    | 2 0 1            | 0.89 310.90    | 17.00 |
| 5            | ARIMA 5    | 1 0 2            | 0.88 783.86    | 26.86 |
| 6            | ARIMA 6    | 1 0 1            | 0.89 282.62    | 15.28 |
| 7            | ARIMA 7    | 1 0 1            | 0.64 37251.12  | 157.91|

5.1.2. SARIMA

The results of ARIMA models research have been used for simulating time series with a SARIMA model. The modeled forecast has studied at a value of the differentiation order above 2. A significant increase in the values of the other parameters of the models was observed with selected \( d \) parameter. This led to a rise the training time of the models. It is worth noting that the accuracy of forecasting such
models does not substantially differ from the accuracy of models with fewer parameters value. The obtained models and their accuracy estimations are presented in table 3.

Table 3. Metric values of SARIMA models

| Model number | Model name  | Model parameters | Ten-day forecast |
|--------------|-------------|------------------|-----------------|
|              |             | p   d   q   P   D   Q   m |   R^2 | MSE    | MAE   |
| 1            | SARIMA 1    | 2   0   1   1   1   2   24 | 0.94 | 39452.37 | 168.21 |
| 2            | SARIMA 2    | 1   1   1   1   1   1   24 | 0.51 | 376.45   | 16.30  |
| 3            | SARIMA 3    | 1   1   1   1   1   1   24 | 0.51 | 83.50    | 6.64   |
| 4            | SARIMA 4    | 2   2   2   2   2   2   24 | 0.56 | 399.43   | 15.86  |
| 5            | SARIMA 5    | 2   2   1   1   1   2   24 | 0.58 | 1162.61  | 25.12  |
| 6            | SARIMA 6    | 2   2   2   2   2   1   24 | 0.65 | 273.06   | 12.18  |
| 7            | SARIMA 8    | 1   2   2   2   1   1   24 | 0.62 | 44373.06 | 166.00 |
| 8            | SARIMA 9    | 1   3   2   2   1   1   24 | 0.62 | 51947.79 | 180.29 |

5.2. Results of building neural network models

5.2.1. RBF

An efficient network architecture has obtained after simulating the values of the available time series with an RBF-type neural networks. A model has built using the resulting architecture. The obtained model has the following structure: the RBF neurons number is 100 units, the size of the layer using the sigmoid activation function equals 2160 units. The calculated values of the neural network model metrics are shown in table 4.

Table 4. The metric values of the neural network models RBF

| Model number | Sensor name | Ten-day forecast |
|--------------|-------------|-----------------|
|              |             | R^2 | MSE | MAE |
| RBF(100, 2160)| Sensor 1    | 0.68 | 67889.43 | 178.87 |
|              | Sensor 2    | 0.70 | 94.42  | 6.67  |
|              | Sensor 3    | 0.56 | 91.20  | 8.54  |
|              | Sensor 4    | 0.68 | 373.65 | 18.68 |
|              | Sensor 5    | 0.67 | 1090.52 | 31.87 |
|              | Sensor 6    | 0.64 | 467.42 | 20.20 |
|              | Sensor 8    | 0.66 | 58467.24 | 202.09 |
|              | Sensor 9    | 0.67 | 51080.52 | 193.08 |

5.2.2. LSTM

Investigations of the LSTM neural network model have shown that the dual LSTM architecture is less accurate than the single one. However, adding a sigmoid layer to a single model increases the forecast accuracy. The sigmoid layer helped to reduce the LSTM model architecture amount, which accelerated the model learning process, though it has depressed the forecast quality. Further actions have aimed to expanding the scope of the LSTM model. The final model architecture has become (LSTM(2160), Sigmoid(2160)). Table 5 presents the test results.

Table 5. Results of testing models obtained by the LSTM neural network model

| Model number | Sensor name | Ten-day forecast |
|--------------|-------------|-----------------|
|              |             | R^2 | MSE | MAE |
| LSTM(2160), Sigmoid(2160)| Sensor 1    | 0.69 | 73845.35 | 181.67 |
|              | Sensor 2    | 0.65 | 122.99 | 8.95  |
|              | Sensor 3    | 0.55 | 123.48 | 9.84  |
|              | Sensor 4    | 0.73 | 364.43 | 16.58 |
|              | Sensor 5    | 0.80 | 737.67 | 21.80 |
|              | Sensor 6    | 0.70 | 421.48 | 17.27 |
|              | Sensor 8    | 0.65 | 57936.49 | 203.34 |
|              | Sensor 9    | 0.66 | 52035.54 | 194.99 |
5.3. The results of the comparative analysis of models

A comparative analysis of the model testing results has shown that the considered neural network models simulate the available data worse in comparison with econometric models. However, it is quite obvious that further research is needed within the theme under discussion. These considerations are based on the calculated metrics values of all constructed models.

![Figure 1. Model prediction errors values for all sensors represented in a logarithmic scale](image)

6. Conclusions

After reviewing the conducted investigations, it is necessary to remind that 4 models of 2 different classes have been built and analyzed: econometric models (ARIMA, SARIMA) and neural network models (RBF and LSTM). The expected results related to ability of the models to predict accurate forecasts have proved correct. The experiments analysis has made it possible to enumerate recommendations for further research. Future work on the theme development is to collect more data, as well as their following pre-processing. A detailed analysis of the data for outliers and anomalies is required. It is recommended to repeated studies to the models construction considered in this work, after preprocessing the data. The possibility of expanding the models range in the paper is not excluded.

References

[1] G. Jenkins G. E. P. Box. Time series analysis, forecasting and control. Holden-Day, San Francisco, CA, 1970

[2] A. Earnest, M. I. Chen, D. Ng, L. Y. Sin, “Using Autoregressive Integrated Moving Average (ARIMA) Models to Predict and Monitor the Number of Beds Occupied During a SARS Outbreak in a Tertiary Hospital in Singapore,” in BMC Health Service Research, 5(36), 2005

[3] Sharafi, M., Ghaem, H., Tabatabaee, H. R., & Faramarzi, H. (2017). Forecasting the number of zoonotic cutaneous leishmaniasis cases in south of Fars province, Iran using seasonal ARIMA time series method. Asian Pacific Journal of Tropical Medicine, 10(1), 79–86. doi:10.1016/j.apjtm.2016.12.007

[4] Simon Haykin, Simon S. Haykin Prentice Hall, 2009

[5] Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2017). Short-Term Residential Load Forecasting based on LSTM Recurrent Neural Network. IEEE Transactions on Smart Grid, 1–1. doi:10.1109/tsg.2017.2753802