The use of random process models and machine learning to analyze the operation of a taxi order service

Nikita Andriyanov[1,2,*] and *Vladislav* Sonin[3]

[1]Ulyanovsk State Technical University, 432027 Severny Venets 32, Ulyanovsk, Russia
[2]Ulyanovsk Civil Aviation Institute, 432071 Mozhaiskogo 8/8, Ulyanovsk, Russia
[3]Gett Taxi, 432002 Narimanova 1, Ulyanovsk, Russia

Abstract. The possibilities of constructing an effective forecast of the number of taxi service orders based on mathematical models are considered. A comparative analysis of the variances of forecasting errors for various stochastic models and models based on fuzzy logic is carried out. It is shown that the best estimates are provided by the doubly stochastic model, as well as by the fuzzy Sugeno model.

1 Introduction

Currently, there is no complete and universal effective solution to the problem of forecasting the operation of automated technical systems that monitor the operation of a taxi order service. Moreover, the data accumulated by such systems can be represented in the form of time series, and their description can be performed using statistical mathematical models. In addition, given the specifics of the work of the taxi order service [1-5], it is clear that, for example, the distribution of orders will probably be seasonal.

In this paper, we consider forecasts based on random processes generated by autoregressive models, which can be supplemented by models with multiple roots of characteristic equations [6,7], inhomogeneous doubly stochastic models [8,9] and forecasts using fuzzy logic.

2 One dimensional models of data representation and forecasting

Usually tools called stochastic models are used to forecast time series. There are a number of models used to forecast data nowdays:

1) Regression forecasting models [10]. The main advantage of such models is undoubtedly their sufficient knowledge, but there are drawbacks to this approach: e.g. models that have too little complexity may turn out to be inaccurate, and models that have excess complexity can turn out to be retrained.

2) Autoregression forecasting models (ARIMAX, GARCH, ARDLM) [11]. It is worth noting that autoregressive models represent a wide range of time series forecasting models.

* Corresponding author: nikita-and-nov@mail.ru

The advantage of autoregressive models is a well-developed mathematical apparatus, the presence of a full range of processing algorithms for such models, and the ability to quickly forecast. However, at the same time, they also have weaknesses, in particular, they are characterized by spatial homogeneity, the inability to describe processes with a complex internal structure quite accurately without computational costs.

3) Exponential Smoothing Models (ES) [12]. It should be noted right away that these models make it possible to obtain smooth forecasting trends, however, their field of application is narrowed sufficiently in the case of complex processes with many significant changes in values over time.

4) Model on neural networks (ANN). Recent years have been characterized by the rapid development of the application of technologies related to neural networks. The positive results obtained in various fields determine the high popularity of ANN time series forecasting models [13]. Nevertheless, the disadvantage of such models is the complexity of training and the associated computational costs.

5) Classification And Regression Trees Model (CART). Decision trees are one of the methods of automatic data analysis [14]. As the name of the algorithm shows, it solves the problems of classification and regression. Unfortunately, if an attribute was selected once, and it was partitioned into subsets, the algorithm cannot go back and select another attribute that would give the best partition. And therefore, at the construction stage, it cannot be said whether the selected attribute will ultimately yield the optimal partition.

6) Support Vector Model (SVM). This method solves the problems of classification and regression by constructing a nonlinear plane separating the solutions. Due to the nature of the feature space in which the boundaries of the solution are constructed, the support vector method has a high degree of flexibility in solving regression and classification problems of various levels of complexity. There are various types of SVM models, e.g. linear, polynomial, RBF (radial basis functions), and sigmoid [15]. SVM models are also based on neural networks, and are more likely to be used for clustering than for forecasting.

7) Model on fuzzy logic (FL). Fuzzy logic is used to solve a wide range of problems, often not related to time series forecasting. Nevertheless, this model deserves consideration and is effective in solving complex problems. Active research is underway in this area, for example, [16]. However, there are a number of disadvantages of fuzzy systems:

– lack of a standard method for designing fuzzy systems;

– the impossibility of mathematical analysis of fuzzy systems by existing methods;

– the use of a fuzzy approach compared with the probabilistic does not lead to an increase in the accuracy of calculations (!as considered earlier!).

Nevertheless, trained networks can produce fairly accurate forecasting results.

8) Models based on doubly stochastic random fields [17,18]. These models are quite good at describing heterogeneous arrays of multidimensional data, but they can often be redundant in the intellectual analysis of one-dimensional information.

Thus, the analysis confirms the interest of researchers in the problems of modeling and forecasting data One of examples of such data is information about orders' distribution in taxi. However, most of the known models are not designed to describe multidimensional correlated data arrays, but are based on some simplifications, including dimensionality.

3 Designing of fuzzy knowledge bases

Currently there are two main types of fuzzy knowledge bases: the Mamdani type and the Sugeno type. The difference between Mamdani's fuzzy inference and Sugeno's is based on the fact that in the first case, the conclusions of the rules are represented by fuzzy terms as well as for input variables, and in the second case, the conclusions are a function of the input variables. The Mamdani knowledge base is described as follows [19]:

$$\text{If } (x_1 = a_{i1}\, u\, x_2 = a_{i2}\, u ... u\, x_n = a_{in}),\ \text{then } y = d_i, \text{with weight } w_i,\ i = \overline{1,N}, \quad (1)$$

where a_{ij} is fuzzy term that is used for a linguistic variable to evaluate factor x_j in the *i*-th rule, $i = \overline{1,N}$, $j = \overline{1,n}$; N is number of knowledge base rules; d_i is consequent of *i*-th rule in the form of a fuzzy term; $w_i \in [0;1]$ is weight of *i*-th rule, which reflects the expert's confidence in its reliability.

The Sugeno knowledge base is described as follows [20]:

$$\text{If } (x_1 = a_{i1}\, u\, x_2 = a_{i2}\, u ... u\, x_n = a_{in}),\ \text{then } d_j = b_{j0} + \sum_{i=\overline{1,n}} b_{ji} x_i, \quad (2)$$

where $b_{j0}, b_{j1}, ..., b_{jm}$ are some real numbers.

For the synthesis of fuzzy knowledge bases, clustering methods are often used, namely, the algorithm of subtractive clustering or fuzzy c-means. The construction of a fuzzy Sugeno knowledge base from experimental data consists of two main stages. The first is the identification of the structure of the knowledge base, which includes the formation of an initial set of rules in IF-THEN statements using subtractive clustering. The second is parameter identification using the ANFIS algorithm (ANFIS - Adaptive Network Based Fuzzy Inference System), which includes setting parameters for membership functions and rule weights. The use of clustering allows you to identify natural groupings of data from a common set, thereby providing identification and a brief presentation of the overall structure of the knowledge base. The advantage of the subtractive clustering algorithm is the lack of the need to determine the initial number of clusters. The algorithm includes the following steps: consideration of data elements as potential candidates for cluster centers; calculating the probability that the data element is the center of the cluster; selection by the center of the first cluster of the point with the greatest potential; potentials are calculated for the following cluster centers (taking into account the deduction of the contribution of the center of the just found cluster); implementation of an iterative procedure for recalculation of potentials and identification of cluster centers until the maximum value of the potential exceeds the initially specified threshold [21]. Despite the fact that the subtractive clustering algorithm is not fuzzy, it is often used for the task of automatically generating fuzzy rules from experimental data.

Thus, the identification of fuzzy knowledge bases using subtractive clustering involves the formation of clusters in the data space and the translation of these clusters into fuzzy rules that describe a specific part of the behavior of the system under study. After projecting the degrees of membership of the clusters on the input space and the corresponding approximation, the membership functions of the terms in the premises of fuzzy rules are determined. The conclusions of the rules are determined by the least squares method.

At the second stage, using the ANFIS algorithm, the parameters of the synthesized knowledge base are configured. The presentation of fuzzy inference in the form of a neuro-fuzzy model allows the use of training algorithms for neural networks. The structure of the ANFIS network is isomorphic to a fuzzy knowledge base and is a five-layer neural network of direct signal propagation. Layers have the following purposes: first layer uses terms of input variables; the second layer is layer for sending fuzzy rules; the third layer is the normalization of the degree of implementation of the rules; the fourth layer is the conclusion of the rules; fifth layer is aggregation of the results obtained according to various rules. A combination of the back propagation algorithm and the least squares method are often used as training procedures for these models.

To solve the problem of designing a fuzzy knowledge base Mamdani, an algorithm of fuzzy c-means is used. The clustering problem in this case can be formulated using the characteristic function. The characteristic function takes a value in the interval [0,1], reflecting the degree of belonging of a particular element to a particular cluster. The description of the partition matrix of fuzzy clusters using the characteristic function is presented in the form $F = [\mu_{ki}]$, где $\mu_{ki} \in [0,1]$, $k = \overline{1,M}$, $i = \overline{1,c}$. Wherein the *k*-th row characterizes the degree of belonging of the element $X_k = (x_{k1}, x_{k2}, ..., x_{kn})$ to clusters $A_1, A_2, ..., A_c$. Matrix *F* must satisfy the following requirements [19]:

$$\begin{aligned} &\sum_{i=\overline{1,c}} \mu_{ki} = 1, \ k = \overline{1,M}, \\ &0 < \sum_{k=\overline{1,M}} \mu_{ki} < M, \ i = \overline{1,c}. \end{aligned} \tag{3}$$

The idea of the fuzzy c-means algorithm is to iteratively recalculate the matrices *F* and the centers of the clusters. The scatter criterion is used as the objective function, which must be minimized [20]:

$$\sum_{i=\overline{1,c}} \sum_{k=\overline{1,M}} (\mu_{ki})^m \|V_i - X_k\|^2 \to \min, \tag{4}$$

where $V_i = \dfrac{\sum_{k=\overline{1,M}} (\mu_{ki})^m X_k}{\sum_{k=\overline{1,M}} (\mu_{ki})^m}$ are fuzzy cluster centers; $m \in (1, \infty)$ is exponential weight.

In this case, the Euclidean distance is taken as the metric of anomaly. The membership functions of the terms of the output variables in the conclusions of the rules of the knowledge base are found as well as for the input variables.

4 Taxi service data representation models

Let us forecast the sequence shown in Figure 1 (this is processed information on the distribution of orders, preserving the properties of a real sequence).

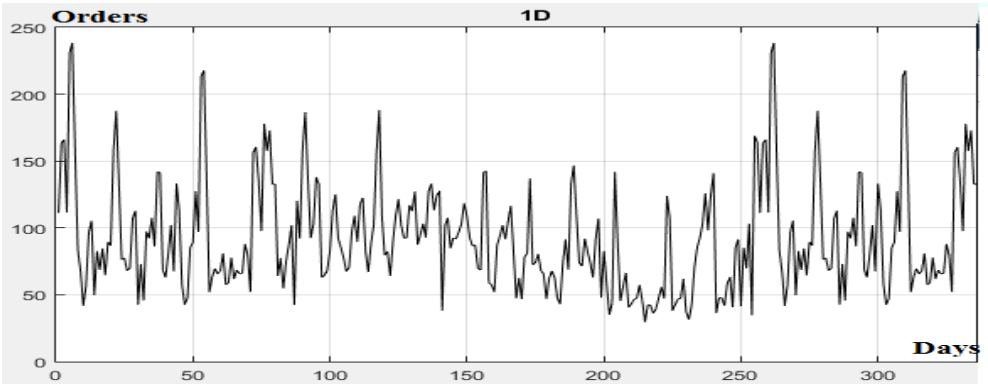

Fig. 1 Distribution of orders by day of the year.

Table 1 presents possible stochastic models for forecasting. It is worth noting that in the doubly stochastic model there is a change in correlation parameters.

Table 1. Data Representation Stochastic Models Used

Model	Mathematical description
One dimensional autoregressive model	$O_i = \rho O_{i-1} + \xi_i, i = 1...N$
One dimensional doubly stochastic model	$O_i = \rho_i O_{i-1} + \xi_i, i = 1...N,\ \rho_i = \tilde{\rho}_i + m_\rho,\ \tilde{\rho}_i = r\tilde{\rho}_{i-1} + \sqrt{\sigma_\rho^2(1-r^2)}\varsigma_i$

In addition to stochastic models, we also make a forecast based on fuzzy models of Mamdani and Sugeno.

Let carry out the necessary assessment of the parameters of the models presented in Table 1. Based on the considered models, we will forecast the last 21 values of the sequence. The parameters can be estimated using the Yule – Walker equations or by the pseudogradient method [22].

5 Results and discussion

The relative variances of forecast errors of the last twenty-one values, respectively, are following:

1) Error using a one-dimensional autoregressive random sequence model is 0.718;
2) Error using a one-dimensional doubly stochastic random sequence model is 0.381.

Figure 2 presents the forecasts themselves and real data.

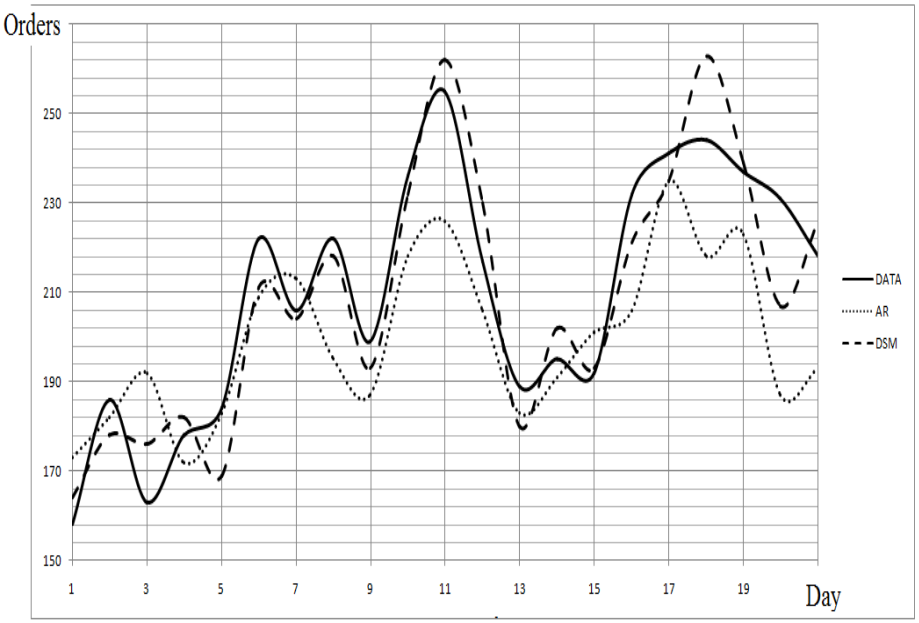

Fig. 2 Forecasting the number of taxi service orders by mathematical models of random processes.

Obviously, the use of a doubly stochastic model provides a more accurate forecast. Similarly, we compare the forecast data using the fuzzy logic of Mamdani and Sugeno. Figure 3 shows the results of forecasting based on fuzzy logic.

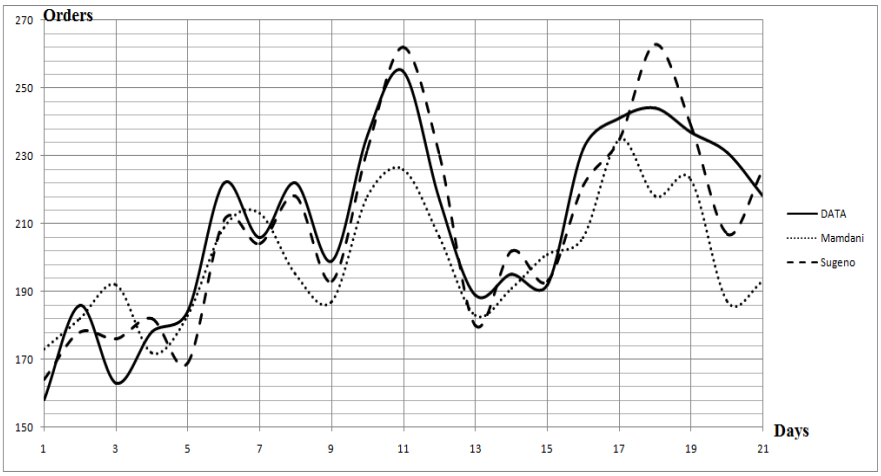

Fig. 3 Forecasting the number of taxi service orders based on fuzzy logic.

For the forecasts presented in Figure 3, the relative variance of the forecast errors of the last twenty-one values, respectively, are following:

1) Error using Mamdani fuzzy logic is 0.304;
2) Error using Sugeni fuzzy logic is 0.206.

First, models based on fuzzy logic do not differ as significantly as models based on autoregressive processes. Secondly, both models provide a lesser variance in forecasting error.

6 Conclusion

The task of analyzing and optimizing the effectiveness of a taxi order service is considered. It is proposed to use doubly stochastic models to take into account data heterogeneity. A comparative analysis of forecasting based on 2 different models is presented. In this case, the gain of the doubly stochastic model in comparison with the autoregressive model can reach 2-5 times. The use of fuzzy logic knowledge bases allows to increase the forecasting efficiency by another 5-10%.

Acknowledgements

This work is supported by RFBR, Project №18-31-00056 mol_a

References

1. A. Mecke, I. Lee, J.R. Baker jr., M.M. Banaszak Holl, B.G. Orr, Eur. Phys. J. E **14**, 7 (2004)
2. M. Ben Rabha, M.F. Boujmil, M. Saadoun, B. Bessaïs, Eur. Phys. J. Appl. Phys. (to be published)
3. F. De Lillo, F. Cecconi, G. Lacorata, A. Vulpiani, EPL, **84** (2008)
4. L. T. De Luca, *Propulsion physics* (EDP Sciences, Les Ulis, 2009)