Using mathematical modeling of time series for forecasting taxi service orders amount

N A Andriyanov[1,2] and V A Sonin[3]

[1]Ulyanovsk Institute of Civil Aviation, 432071 ul. Mozhaiskogo, h. 8/8, Ulyanovsk, Russia,
[2]Ulyanovsk State Technical University, 432027 ul. Severniy Venets, h. 32, Ulyanovsk, Russia,
[3]Gett (Representing in Ulyanovsk), 432071 pr-t Narimanova, h. 1, b. 3, Ulyanovsk, Russia

Abstract. One of the main characteristics of modern world is the introduction of various automated systems that greatly facilitate the work of man. In particular, in the territory of the Ulyanovsk region, the technologies of automated production are successfully applied by Nemak Rus, Bridgestone Tayer Manufacturing CIS, Gett order taxi service, etc. However, there are many tasks related to continuous monitoring of incoming information. This is due to the development of technologies and the ability to organize automated data collection, increasing the amount of information stored. It is necessary to ensure the operation of such systems in the mode of 24 hours per day, 7 days per week. It is clear that this task can not be solved only by human resources. The solutions require effective algorithms for analyzing accumulated data. The paper is aimed at the development of data analysis algorithms in similar systems. Particular attention is paid to methods of forecasting the number of orders on the available data using mathematical models of random fields. The approach of the two-dimensional representation of the time sequence is presented. Such approach makes it possible to improve the efficiency of the forecast.

1. Introduction

Currently, there is no complete and universal effective solution to the task of predicting the operation of automated technical systems that monitor all processes in the taxi order service. Indeed, at present there are no relevant data analysis and processing centers that are comparable to the accumulated data volume.

In this case, the data accumulated by such systems can logically 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 operation of the taxi order service [1-2], it is clear that, for example, the distribution of orders is likely to be seasonal, and therefore the description of such sequences is not possible by known homogeneous models (wave, autoregressive).

Thus, our research is aimed, firstly, at solving an actual scientific problem related to the statistical analysis of random processes generated by autoregressive models, including multiple roots of the characteristic equations [3], inhomogeneous doubly stochastic models [4,5], and also that is of the greatest interest with particular cases of mixed models of random sequences proposed for the description of a narrow class of sequences of real data, obtained in databases of taxi services. Secondly, it is proposed to use a two-dimensional model to analyze the available time sequence of the taxi order service data. The results of comparing the accuracy of forecasting the number of orders for a certain period by the criterion of the minimum variance of the error are given.

2. Studies in the field of time series

2.1. Actuality

So, the task of analyzing and forecasting time series is of considerable interest to researchers. The classical exposition of the theory of time series and forecasting is presented by G. Box, G. Jenkins and their co-authors [6].

Stochastic models are usually used to forecast time series [7]. The analysis shows that there is a huge number of methods and models for forecasting time series. However, we will try to generalize a number of methods and models, which will significantly narrow their classification. First, consider the following known methods and models.

2.2. Regression forecasting models

A study of such models is presented by authors in the paper [8]. Also the classic publication of N. Draper and H. Smith [9] is worth seeing. Many modern researchers relies on this paper. The main advantage of such models is certainly their sufficient knowledge, but this approach also has drawbacks. For example, models having too little complexity may turn out to be inaccurate, and models with excessive complexity may be retrained.

2.3. Autoregression forecasting models (ARIMAX, GARCH, ARDLM)

For example, the paper [10] is devoted to these models. It is worth noting that autoregressive models represent the widest range of time series forecasting models. They are as following:

- ARIMAX (autoregression integrated moving average extended). The basis for this type of model is the book by G. Box and G. Jenkins.
- GARCH (generalized autoregressive conditional heteroskedasticity). This model includes the following specifications: FIGARCH, NGARCH, IGARCH, EGARCH, GARCH-M.
- ARDLM (autoregression distributed lag model). The model is used mainly in econometrics, but it also occurs in solving problems in other areas.

The advantage of autoregressive models is a well-developed mathematical apparatus, the presence of a complete set of algorithms for processing such models, and the possibility of rapid forecasting.

For example, for autoregression of arbitrary order m you can use the following equation [6]:

$$x_i = \rho_1 x_{i-1} + \rho_2 x_{i-2} + ... + \rho_m x_{i-m} + \beta\xi_i,\ i = 2,3,...,n, \quad (1)$$

where ξ_i, $i = 1,2,...,n$, are obtained as random variables with a normal distribution, and the expectation is zero, and the variance is equal to one; β is a scale coefficient of information random field x_i. Covariance function of a random sequence (1) can be represented by the sum of the exhibitors for characteristic equation with different roots z_ν, $\nu = 1,2,...,m$,

$$z^m - \rho_1 z^{m-1} - \rho_2 z^{m-2} - ... - \rho_m = 0 \quad (2)$$

under the condition of sustainability $|z_\nu| < 1, \nu = 1,2,...,m.$

The equation (2) having root $z = \rho$ and if it's multiplicity is m will be written as follows $(z-\rho)^m = 0$. And the process of order m is easy to rewrite in operator form [3]:

$$(1 - \rho z^{-1})^m x_i = \beta\xi_i, \quad (3)$$

where $z^{-k} x_i = x_{i-k}$.

However, at the same time, they also have weak sides, in particular, they are characterized by spatial homogeneity, the impossibility of describing processes with a complex internal structure quite accurately without computational costs.

2.4. Models of exponential smoothing (ES)

You can get acquainted with such models in [11]. It also should be noted that these models make it possible to obtain smooth forecasting trends, but their scope is rather narrow in case of complex processes with many significant changes in values over time.

2.5. Model on the most similar pattern (MMSP)

This model was considered in [7]. The author shows that MMSP provides efficiency on a number of tasks. However, it should be noted that in many cases, for example, when applied to the FOREX series and exchanges, we get unsatisfactory results.

2.6. Algorithms on neural networks (ANN)

Recent years are characterized by the rapid development of the application of technologies related to neural networks. The positive results obtained in various areas cause such a high popularity of ANN time series forecasting models [12]. Nevertheless, the drawbacks of such models are the complexity of training and the associated computational costs.

2.7. Markov chains

The Markov chain is suitable for use in describing a mathematical model for changing random processes. The advantages of the Markov chain over the probability of the distribution of random processes are the fact that the Markov chain allows one to investigate the operating modes when changing these processes. Markov chains are well amenable to automation and allow in an automatic mode to calculate the probability of the dynamics of the process change, as well as the probability of occurrence of certain transitional regimes. In addition, Markov chains are used for forecasting of economic processes. In work [13] the model of forecasting using the Markov chain is presented as a useful tool for situations where several time series are interrelated. By estimating the transition probability matrix, the Markov chain prediction model can easily capture transitions between different states and allow the predictor to predict all interrelated time series simultaneously. Despite all the advantages, the models on Markov chains are not universal and are significantly inferior to autoregressive models in a number of problems.

2.8. Classification And Regression trees Model (CART)

Decision trees are one of the methods of automatic data analysis. The first ideas for creating decision trees go back to the works of Hoveland and Hunt in the late 50s of the 20th century. However, the fundamental work that gave impetus to the development of this direction was the book by Hunt E.B., Marin J. and Stone P.J. Analysis of the literature shows that quite a few studies have been devoted to this model, but there are a number of good papers. In particular, the paper by Yohannes Y.Y. and Webb P. [14]. Thus, CART is an algorithm for constructing a binary decision tree. It is a dichotomous classification model. Each node of a tree has only two descendants when it is partitioned. As can be seen from the name of the algorithm, it solves the problems of classification and regression. The disadvantage of classification-regression trees is the following: most of the known algorithms are "greedy algorithms". If once an attribute was selected and a partitioning into subsets was made, the algorithm can not go back and select another attribute that would give the best partition. And therefore at the stage of construction it is impossible to say whether the chosen attribute will provide the optimal partition.

2.9. Genetic Algorithm model (GA)

Genetic algorithms are actively used in robotics, computer games, learning neural networks, creating artificial life models, scheduling, optimizing queries to databases, finding optimal routes, etc. [15]. However, the effectiveness of its use is low. For example, the genetic algorithm is used to solve optimization problems (extremum search), but in some works it is used to forecast time series.

2.10. Support Vectors Method (SVM)

This method solves the classification and regression problems by constructing a nonlinear plane that separates solutions. Due to the special characteristics of the feature space nature in which the decision boundaries are constructed, the SVM has a high degree of flexibility in solving regression problems and classifying various levels of complexity. There are different types of SVM models: linear, polynomial, RBF (radial basis functions), and sigmoid [16]. SVM models are also based on neural networks, and are more likely to be used for clustering than for forecasting.

2.11. Transfer Functions (TF)

The TF algorithms belong to a class of algorithms based on neural networks. The authors of [17] present an algorithm for detecting the anomalous state of a dam based on the TF model between water level signals and pore pressure (water pressure in soil pores) in a dam. The main idea of the proposed approach is to apply methods of detecting abnormal behavior that are trained on "raw" and (or) pre-processed data. Thus, the main disadvantage of these algorithms is their computational complexity.

2.12. Fuzzy Logic (FL)

Fuzzy logic is used to solve a wide range of tasks, often not related to forecasting time series. Nevertheless, this model deserves consideration and is effective in solving complex problems. Particular attention is paid to this model in the works of the scientific school under the direction of N.G. Yarushkina [18, 19]. However, we note a number of shortcomings of fuzzy systems. They are as following:

- lack of a standard methodology for designing fuzzy systems;
- impossibility of mathematical analysis of fuzzy systems by existing methods;
- the application of the fuzzy approach in comparison with the probabilistic approach does not lead to an increase in the accuracy of the calculations.

2.13. Doubly Stochastic Models (DSM).

For the description of non-stationary in time processes and fields inhomogeneous in space, doubly stochastic models have been proposed. Analysis of the literature shows that they have found wide application in solving problems of image representation and processing [20,21].

Let us consider the one-dimensional case in more detail. Let the random process be formed using the model

$$x_i = (m_\rho + \rho_{i-1})x_{i-1} + \xi_i, \quad (4)$$

where $\{\xi_i\}$ are independent gaussian random variables with $m_\xi = 0$, $\sigma_{\xi i} = \sigma_x\sqrt{1-(m_\rho + \rho_{i-1})^2}$; m_ρ is constant value that has the meaning of the average value of the correlation coefficient, $i = 1,2,...,M$.

Variations of the correlation coefficient are determined by the following autoregressive model

$$\rho_i = r\rho_{i-1} + \zeta_i, \quad (5)$$

where $\{\zeta_i\}$ are independent gaussian random variables with $m_\zeta = 0$, $\sigma_\zeta = \sigma_\rho\sqrt{1-r^2}$; r is correlation coefficient of the process of changing parameters in the doubly stochastic process x_i.

It should be noted, that the following restrictions are imposed: $-1 < (\rho_i \pm m_\rho) < 1$.

Thus, changing the parameters of the models allows you to take into account various drops and surges, for example, in the taxi service associated with public holidays.

2.14. Summary

Thus, the conducted analysis confirms the interest of researchers in the tasks of modeling and forecasting time series. Obviously, even such a short review of the literature allows us to conclude that the development of time series modeling algorithms successfully finds its niche in various private applied problems. When forecasting the distribution of taxi service orders, we made a choice in favor

of autoregressive models, since there is sufficient groundwork in this direction, and the models themselves have proved themselves in statistical radio engineering applications, including those related to the processing of one-dimensional and multidimensional signals. Furthermore, we use doubly stochastic models because they allow to imitate abrupt changes in properties inside the process.

3. Brief description of the taxi order service

In the taxi order service the relevance of investigated subject is also undoubted. First, it is an increase in the efficiency of the taxi order service. Indeed, modern information technologies have radically changed the market of taxi services. He will never be the same again. But even such as today, the situation can not remain. The spread of smartphones running iOS and Android has already led to the spread of applications related to the taxi order service. The taxi market is so competitive that the emergence of a monopoly on it is an impossible fact. At the same time, it is possible to achieve lower prices and win customers, at the expense of almost 100% automation of service execution, perfect software, outsourcing by outsiders, etc. It should be noted that the fully automation requires a lot of tasks. For example, it is necessary to monitor in real time the tracking of drivers, their distribution by areas. Timely analysis of this information allows you to quickly vary prices. Analysis of incoming information about orders will allow to predict the required number of taxi drivers and the number of operators receiving calls if you use call center.

Consider the project based on the call center. At the same time, telephony is sent to operators directly through the Internet, which requires only the presence of a computer with a headset. For the dispatching taxi organization, a powerful software and hardware complex is needed. Its application allows several thousand machines to work in real time.

Obviously, the use of this technology allows you to effectively manage resources, increase the speed of order processing, always have exact customer numbers, reduce the time for applications.

So, the contact center requires the presence of a multi-channel phone number, which will allow receiving many calls simultaneously. To do this, you can use the technology of IP-telephony. One of the most common telephony servers is the Asterisk server, which allows you to work with SIP-telephony. Such a telephone exchange should be set up to distribute calls to taxi service operators. Call processing is carried out using a special program that represents the operator a taxi order form based on the Internet browser. To store information about calls, a database server is used, for example, MsSQL. Tariffs are set up using a separate module - "Tarificator", which is programmed to use it on the web.

Figure 1 shows the complete architecture of the considered service.

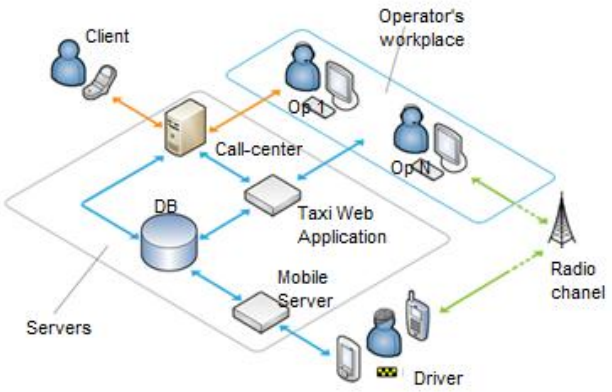

Figure. 1. Block diagram of the taxi order service.

Thus, it is advisable to use virtualization methods to separate different servers, including a telephony server, a telephony database server and a web server. In addition, an application server is needed, through which information is transferred from the call center to the drivers. This is provided

by a special program for taxis. And here it is recommended to use one more database server to store order information.

The application for the Taxi program can have versions that work just under java, or targeted at modern devices running Android and iOS.

With the receipt of an order by a particular driver, the database is updated. For example, we can store information about the car, time of order picking, etc. It can be used to inform the client about the car assigned for order.

Statistics are collected using database servers, but the presentation of information in a convenient form is obtained using the Tarificator, which allows you to display statistics either in a text document or in an excel format document. Figure 2 shows the revised information on the distribution of orders, preserving the properties of the real sequence. We will make models fit according to this data.

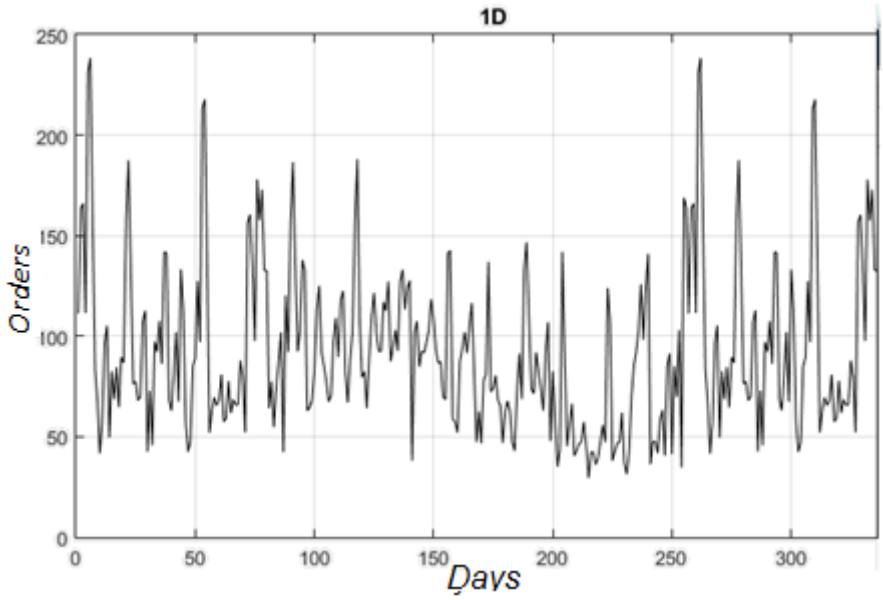

Figure 2. Distribution of orders by day with conversion (X-axis corresponds to certain day in the year, Y-axis corresponds to number of orders).

It should be noted that the process presented at Figure 2 has an inhomogeneous structure, as well as some periodic singularities. Therefore, it is necessary to select the most appropriate model to more accurately describe all the inherent characteristics of the distribution.

It is also important that Figure 2 presents the sequence which is obtained from real data collected for the year by a taxi order service operating in the territory of the Ulyanovsk region. The X-axis (days) begin on January 1st, and end closer to December 20th. In this regard, the task of forecasting is complicated, because you need to make a correct forecast for the last days of the year. When presenting data, linear transformations were used, which did not affect the nature of the time sequence, where bursts are clearly visible in the first days of the year.

4. Representation of the orders time series by mathematical models

4.1. Introduced conditions

Let's consider some variants of the description of the collected statistics on service. Let the data be collected from the beginning of the year (January) and until the end of the year (December) with some simplification, which will be used when approaching the representation in the form of an image.

4.2. One-dimensional autoregression process

Imagine an existing sequence of data about orders made as $\{O\}$. We will use expression for autoregression of the first order:

$$O_i = \rho O_{i-1} + \xi_i, i = 1...N, \quad (6)$$

where ρ is the correlation coefficient over the whole sequence and can be easily estimated from the available data; ξ_i is random additive with zero mathematical expectation and variance $\sigma_\xi^2 = \sigma_O^2(1-\rho^2)$. In this case, variance for orders is also estimated based on the sample.

Autoregression processes of higher orders can be used for a more accurate description.

4.3. One-dimensional doubly stochastic model of a random process

The descriptions of the inhomogeneity and periodicity of real data can be achieved using doubly stochastic models of random processes whose parameters are realizations of a random process:

$$O_i = \rho_i O_{i-1} + \xi_i, i = 1...N, \quad (7)$$

where ξ_i is random additive with zero mathematical expectation and variance $\sigma_\xi^2 = \sigma_O^2(1-\rho_i^2)$, ρ_i is sequence of correlation parameters, defined as

$$\rho_i = \tilde{\rho}_i + m_\rho \text{ и } \tilde{\rho}_i = r\tilde{\rho}_{i-1} + \sqrt{\sigma_\rho^2(1-r^2)}\varsigma_i, \quad (8)$$

where r is constant correlation coefficient; m_ρ is mean value of the main correlation coefficient; σ_ρ^2 is variance in the process of changing correlation parameters; $\{\varsigma_i\}$ is the generating field of Gaussian random variables with zero mathematical expectation and unit variance.

For model (7) and its parameters (8), an increase in the order of the process can also be used. However, the process shown in Figure 2 looks quite "prickly", which allows the use of first-order models.

It should be noted that the estimation of all parameters of the model can be performed by mathematical statistics using the available sample or by the Kalman nonlinear filter, and the algorithms themselves can be adjusted to different dimensionalities of the models.

4.4. Representation in the form of a random field

The observed quasi-periodicity of the process shown in Figure 2, allows us to conclude that it is possible to use models of random fields to represent information of this kind. Consider, for example, doubly stochastic models of images that allow describing inhomogeneous signals [4,5]. As an example, we will use the following model:

$$\begin{aligned} O_{ij} &= 2\rho_{xij}O_{i-1,j} + 2\rho_{yij}O_{i,j-1} - 4\rho_{xij}\rho_{yij}O_{i-1,j-1} - \rho_{xij}^2 O_{i-2,j} - \rho_{yij}^2 O_{i,j-2} + \\ &+ 2\rho_{xij}^2\rho_{yij}O_{i-2,j-1} + 2\rho_{yij}^2\rho_{xij}O_{i-1,j-2} - \rho_{xij}^2\rho_{yij}^2 O_{i-2,j-2} + b_{ij}\xi_{ij} \end{aligned} \quad (9)$$

where O_{ij} is simulated random field with normal distribution $M\{O_{ij}\} = 0$, $M\{O_{ij}^2\} = \sigma_O^2$; ξ_{ij} is the random field of independent standard Gaussian random variables with $M\{\xi_{ij}\} = 0$, $M\{\xi_{ij}^2\} = \sigma_\xi^2 = 1$; ρ_{xij} and ρ_{yij} are correlation coefficients of the model with multiple roots of the characteristic equations of (2;2) multiplicity; b_{ij} is scale factor of the modeled random field.

Random variables ρ_{xij} and ρ_{yij} with a Gaussian probability distribution density can be described by first-order autoregressive equations or higher-order equations.

It is easy to see that the model (9) is a transformation of the ordinary two-dimensional autoregressive model of the first order. This model of random process can also be used to describe a two-dimensional array of data and has the form:

$$\begin{aligned} O_{ij} &= 2\rho_x O_{i-1,j} + 2\rho_y O_{i,j-1} - 4\rho_x\rho_y O_{i-1,j-1} - \rho_x^2 O_{i-2,j} - \rho_y^2 O_{i,j-2} + \\ &+ 2\rho_x^2\rho_y O_{i-2,j-1} + 2\rho_y^2\rho_x O_{i-1,j-2} - \rho_x^2\rho_y^2 O_{i-2,j-2} + b_{ij}\xi_{ij} \end{aligned} \quad (10)$$

Note that the model (9), unlike the model with constant parameters (10), imitates the random fields that are heterogeneous in its structure, so it can fairly well reflect sharp surges in the number of orders on weekends and holidays. In order to estimate the parameters of such an image, one can use a vector

nonlinear Kalman filter. To do this, we combine the elements of the image line into a vector $\bar{x}_i = (x_{i1}, x_{i2}, \text{K}, x_{iN})^T$. Then, the model of a single image frame can be written as:

$$\bar{x}_i = diag(\bar{\rho}_{xi})\bar{x}_{i-1} + \upsilon(\bar{\rho}_{xi}, \bar{\rho}_{yi})\bar{\xi}_i, \quad \bar{\rho}_{xi} = r_{1x}\bar{\rho}_{x(i-1)} + \upsilon_{\rho x}\bar{\xi}_{xi}, \quad \bar{\rho}_{yi} = r_{1y}\bar{\rho}_{y(i-1)} + \upsilon_{\rho y}\bar{\xi}_{yi},$$

where $diag(\bar{\rho}_{xi})$ is diagonal matrix with elements $\bar{\rho}_{xi}$ on the main diagonal; lower-triangular matrix υ is the matrix, which is determined by the decomposition of the covariance matrix: $V_x = \upsilon\upsilon^T$.

The process of line evaluation is described by the Kalman quasilinear filter:

$$\hat{\bar{x}}_{pi} = \hat{\bar{x}}_{эpi} + P_i \frac{\partial \Phi^T}{\partial \bar{x}_{pi}} V_n^{-1}(\bar{z}_i - \hat{\bar{x}}_{эpi}), \quad \bar{x}_{pi} = \begin{pmatrix} \bar{x}_i \\ \bar{\rho}_{xi} \\ \bar{\rho}_{yi} \end{pmatrix} = \Phi(\bar{\rho}_{x(i-1)}\bar{x}_{i-1}) + \upsilon(\bar{\rho}_{x(i-1)}, \bar{\rho}_{y(i-1)})\bar{\xi}_i, \tag{11}$$

where $\bar{x}_{эpi} = \Phi(\bar{x}_{p(i-1)})$, $\Phi_p(\bar{x}_{p(i-1)}) = \begin{pmatrix} \Phi(\rho, x) \\ r_{1x}\bar{\rho}_{x(i-1)} \\ r_{1y}\bar{\rho}_{y(i-1)} \end{pmatrix}$, $\bar{\xi}_i = \begin{pmatrix} \bar{\xi}_i \\ \bar{\xi}_{xi} \\ \bar{\xi}_{yi} \end{pmatrix}$.

The use of this algorithm is possible under the condition of precisely known characteristics of the information random field, i.e. when we know parameters r_{1x}, r_{2x}, r_{1y}, r_{2y}, as well as average values of correlation in row and column, variance of correlation parameters and information signal. Otherwise, a preliminary evaluation of these parameters is necessary. For this, pseudo-gradient estimation procedures can be used. The resulting sequence of parameters in the future can be further analyzed and replaced by a model. It is also permissible to use estimation in a sliding window. However, we will not give a similar description of these methods.

Figure 3 shows the transformation of the original process into an image.

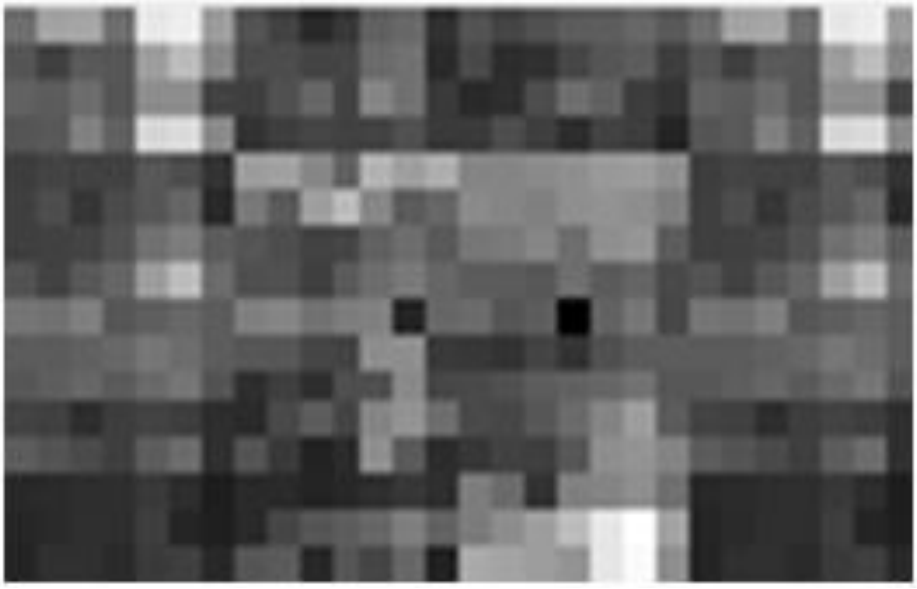

Figure 3. View order statistics as an image.

Thus, we see that the resulting image, on the one hand, is not strongly correlated, and on the other hand, there are several regions with higher brightness values on the image, which indicates the properties of the inhomogeneity.

5. Analysis of forecasting accuracy

We perform the necessary estimation of the parameters of the models (6), (7), (9), and (10). Based on the models considered, we will forecast the last 21 values of the sequence. In the case of an image, the data will be structured according to the seasons and weeks, as shown in Table 1.

Table 1. Data structure when converting to an image.

Month	**January**							**February**							**March**						
Day	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***
Week 1																					
Week 2				Data							Data							Data			
Week 3																					
Week 4																					
Month	**April**							**May**							**June**						
Day	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***
Week 1																					
Week 2				Data							Data							Data			
Week 3																					
Week 4																					
Month	**July**							**August**							**September**						
Day	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***
Week 1																					
Week 2				Data							Data							Data			
Week 3																					
Week 4																					
Month	**October**							**November**							**December**						
Day	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***	***Mon***	***Tue***	***Wed***	***Thu***	***Fri***	***Sat***	***Sun***
Week 1																					
Week 2				Data							Data							Data			
Week 3																					
Week 4																					

The latter values will form a rectangular area in the lower right corner of the image, which is also convenient for predicting and comparing prediction results based on different models. Let's designate the forecasting methods as follows:

1) A1 corresponds to forecast based on a one-dimensional autoregressive model of a random sequence;

2) A2 corresponds to forecast based on a one-dimensional doubly stochastic random sequence model;

3) A2* corresponds to forecast based on a one-dimensional doubly stochastic model with parameter estimation by using the nonlinear Kalman filter;

4) A3 corresponds to forecast based on a two-dimensional autoregressive model of a random field;

5) A4 corresponds to forecast based on a doubly stochastic random field model;

6) A4* corresponds to forecast based on doubly stochastic random field model with parameter estimation by using the nonlinear Kalman filter.

The relative variance of the forecast errors of the last twenty-one values, respectively, is as following:

1) Error Variance is 5.898 for algorithm A1;
2) Error Variance is 0.256 for algorithm A2;
3) Error Variance is 0.064 for algorithm A2*;
4) Error Variance is 0.894 for algorithm A3;
5) Error Variance is 0.181 for algorithm A4;
6) Error Variance is 0.041 for algorithm A4*.

Figure 4 shows the results of the statistical simulation. Particular attention is paid to the predicted part of the sequence. At the same time, the difference in the visual perception of the graphs of Figure 2

and Figure 4 is due to the fact that because of the forecasts, the intervals along the Y axis have changed. However the curves marked with a solid line describe the same data, and the dotted line shows the forecast.

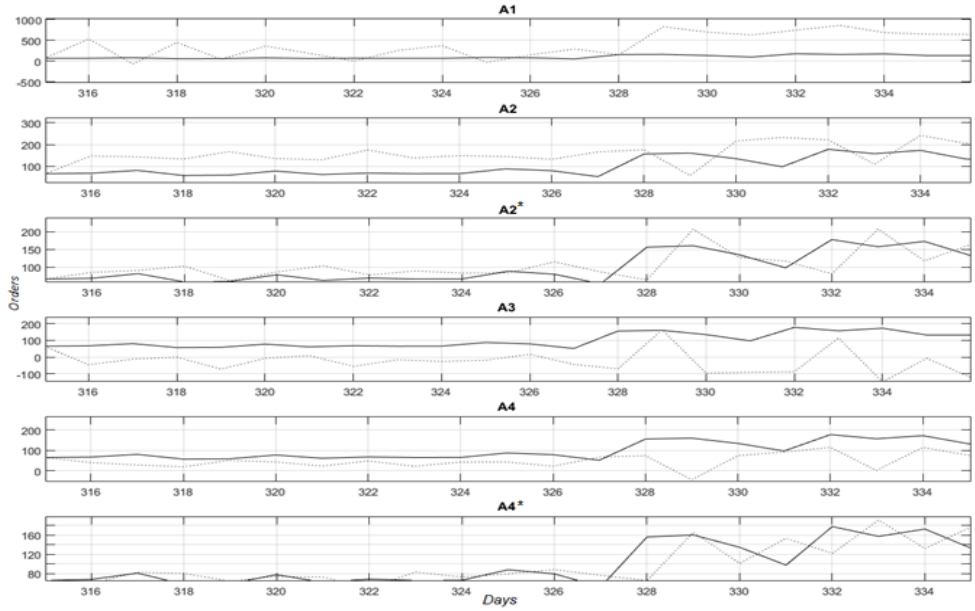

Figure 4. Forecasting the last values of taxi service orders and real data

Thus, the results of forecasting allow us to conclude that the use of autoregression models leads to unsatisfactory results in forecasting complex data. Increasing the effectiveness of the forecast can be achieved through the use of image models. However, such an assessment will also be insufficiently effective. The best indicators are provided by doubly stochastic models that take into account inhomogeneities inherent in real data. In this case, the transition to the multidimensional case leads to an improvement in the forecast, which is related to the properties of the data set being examined. In addition, the highest of the algorithms considered, the accuracy of the forecast is provided by doubly stochastic models of images, which are evaluated using the nonlinear Kalman filtering.

6. Conclusion

The problem of analysis and optimization of the taxi order service efficiency is considered. It is proposed to use doubly stochastic models of images to account for the heterogeneity of the data. A comparative analysis of forecasting based on different models is carried out. In this case, the gain in comparison with autoregressive ones can reach several orders, and by applying the nonlinear Kalman filtering it is possible to increase the forecast efficiency by another 4-5 times.

7. References

[1] P. Azanov, A. Danilov, N. Andriyanov Development of software system for analysis and optimization of taxi services efficiency by statistical modeling methods // 2017 International Conference Information Technology and Nanotechnology. Session Mathematical Modeling, IPGTIS-ITNT 2017; Samara; Russian Federation; 24 April - 27 April 2017. CEUR Workshop Proceedings, Volume 1904, 2017, pp 232-238

[2] Danilov A.N., Andriyanov N.A., Azanov P.T. Ensuring the effectiveness of the taxi order service by mathematical modeling of its work // Information technology and nanotechnology. Proceedings of ITNT-2018. - 2018. pp 1781-1789.

[3] Vasiliev K.K., Andriyanov N.A. Analysis of autoregressions with multiple roots of characteristic equations // Radiotekhnika. 2017. № 6. pp 13-17.

[4] Vasiliev K. K., Andriyanov N. A. Synthesis and analysis of doubly stochastic models of images// 2nd International Workshop on Radio Electronics and Information Technologies, REIT 2017; Yekaterinburg; Russian Federation; 15 November 2017. CEUR Workshop Proceedings, Volume 2005, 2017, pp 145-154
[5] Vasil'ev K. K., Dement'ev V. E., Andriyanov N. A. Doubly stochastic models of images // Pattern Recognition and Image Analysis. January 2015. V. 25(1). - pp 105-110.
[6] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel Time Series Analysis - Fourth Edition Copyright © 2013 by John Wiley & Sons, Inc.
[7] Chuchueva I.A. Model of forecasting time series on the most similar pattern / / Phd Thesis, MSTU. N.E. Bauman, Moscow. 2012, p 155.
[8] Krasheninnikov V.R., Kuvaiskova Yu.E., Klyachkin V.N., Shunina Yu.S. Updating the models for forecasting the state of objects in the form of time series systems and multidimensional classifiers // Vestnik of Computer and Information Technologies. 2017. No. 6 (156). pp 11-16.
[9] Draper N R & Smith H. Applied regression analysis. New York: Wiley, 1966. p 407.
[10] Zhihua Wang, Yongbo Zhang, and Huimin Fu Autoregressive Prediction with Rolling Mechanism for Time Series Forecasting with Small Sample Size // Mathematical Problems in Engineering Volume 2014 (2014), Article ID 572173, p 9.
[11] Prajakta S.K. Time series Forecasting using Holt-Winters Exponential Smoothing // Kanwal Rekhi School of Information Technology Journal. 2004. p 13. Access on: http://www.it.iitb.ac.in/~praj/acads/seminar/04329008_ExponentialSmoothing.pdf
[12] Zulifqar Ali et al Forecasting Drought Using Multilayer Perceptron Artificial Neural Network Model // Advances in Meteorology, Volume 2017 (2017), Article ID 5681308, p 9.
[13] Gongwei Chen Markov Chain Model Forecast for Interrelated Time Series Data Using SAS/IML // Access on: https://www.lexjansen.com/wuss/2014/36_Final_Paper_PDF.pdf
[14] Yohannes, Yisehac, Webb, Patrick Classification and Regression Trees, CART: A User Manual For Identifying Indicators of Vulnerability to Famine And Chronic Food Insecurity. 1999 - p 59.
[15] Pehlivanoglu I.V., Atik I. Time series forecasting via genetic algorithm for turkish air transport market // Journal of aeronautics and space technologies, July 2016, Volume 9, Number 2 - pp 23-33.
[16] Wenzel F., Galy-Fajou T., Deutsch M., Kloft M. Bayesian Nonlinear Support Vector Machines for Big Data // (PDF). Machine Learning and Knowledge Discovery in Databases (ECML PKDD).
[17] Kozionova A.P., Piaita A.L., Mokhova I.I., Ivanov Yu. P. Algorithm based on the transfer function model and the one-class classification for detecting the anomalous state of dams // Information-control systems, №6. 2015 - pp 10-18
[18] Timina I., Egov E., Yarushkina N., Kiselev S. Identification anomalies the time series of metrics of project based on entropy measures // INTERACTIVE SYSTEMS: Problems of Human-Computer Interaction collection of scientific papers. 2017. pp 246-254.
[19] Yarushkina N.G. Fundamentals of the theory of fuzzy and hybrid systems: a tutorial. - Moscow: Finance and Statistics. 2004 – p 320.
[20] K. Vasiliev, V. Dementiev and N. Andriyanov Representation and processing of multispectral satellite images and sequences // Procedia Computer Science 126 (2018), pp 49-58
[21] Andriyanov N. A., Gavrilina Yu. N. Image Models and Segmentation Algorithms Based on Discrete Doubly Stochastic Autoregressions with Multiple Roots of Characteristic Equations// 3rd International Workshop on Radio Electronics and Information Technologies, REIT 2018; Yekaterinburg; Russian Federation; 14 March 2018. CEUR Workshop Proceedings, Volume 2076, 2018, pp 19-29

Acknowledgment
The study was supported by RFBR, project № 18-31-00056.