Development of software system for analysis and optimization of taxi services effciency by statistical modeling methods

Pavel Azanov[1], Andrey Danilov[2], Nikita Andriyanov[3]

[1]Tango Telecom, 122 Krasnaya street, 426057, Izhevsk, Republic Udmurtia, Russia
[2]Taxi Ulyanovsk, 1/3 Narimanova prospect, 432071, Ulyanovsk, Russia
[3]Ulyanovsk State Technical University, 32 Severniy Venets stret, 432027,Ulyanovsk, Russia

Abstract

The text considers using of statistical models for taxi service data analysis and forecasting. Special attention is paid to the model parameters identification and short-term forecasting. We suggest to use the mathematical models of images to account the alternating character, associated with the dependence of the taxi orders number on various parameters. In addition the possibility of improving the efffectiveness of evaluation by use of mixed random fields models is shown.

Keywords: random processes; mixed models; time series forecasting; taxi service; data analysis; image processing

1. Introduction

The following algorithm of the taxi service was quite common recently. Firstly, a dispatcher received the call, then the dispatcher communicated with a driver. During the communication the driver could accept the order or reject it. Usually all connections were provided by the radio devices. However, promising opportunities for use Internet in the taxi order service have appeared [1] due to the rapid Internet development. Now it is not difficult to order a taxi directly on the portal on the Internet or by using special applications for smartphones. In such cases, a very important source of receipt of orders from customers, that we will call customers from the phone, is not taken into account.

At the same time, it should be noted that such processing also provides a sufficient collection of statistics, the analysis of which may allow in the future to improve the quality of the taxi service. Increasing the volume of the telephone calls database warrants the possibility of analyzing the talk time, determining the most popular places in the city, etc. You will get a fairly complete statistical description of the taxi service operation adding to this statistics for orders, including the time of their execution, the waiting time of the car, the distribution by hours and other parameters.

Thus, there is an urgent task of analyzing an information collected in order to increase the efficiency of the service. So, for example, you can anticipate the number of dispatchers and drivers in advance by making precise forecasts of the calls number and tracking the orders percentage. In this case, both time series [2] and various models of random processes (RP) can be used to work with accumulated information [3,4].

2. Service architecture and statistics collection

Consider the taxi service project based on the contact center. At the same time, telephony is sent to operators through the Internet, and it requires only having a computer with a headset. To organize a dispatch taxi, you need a powerful software and hardware system. Its application allows several thousand taxi cars to work in real time.

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

For the contact center organization we need the presence of a multi-channel phone number, which will allow receiving many calls simultaneously. That's why we should use IP-telephony technologies. One of the most common telephony servers (PBX) is the Asterisk server [5], which allows to use SIP-telephony [6]. Such a telephone PBX should be set up to make calls distribution to taxi service operators. To process incoming calls we use a special program that represents the operator the form of a taxi order based on the Internet browser. To store information about calls, a database server is used, for example, MySQL or MsSQL server. Tariffs are set up using a separate module called Tarifficator. This module is programmed for its use in the web.

Thus, it is advisable to use virtualization methods to separate different servers, including a telephony server, a data base of telephony server, and a web server. In addition, an application server is needed. It provides information transfer from the contact center to the drivers. The special program for taxi service implements such transfer. And we suggest to use one more database server to store order information.

Fig. 1 presents full architecture of the considered taxi service.

The application for the Taxi program can have a version running just under java or common modern devices running by Android and iOS.

When a particular driver receives an order, the database is updated. The updates include information about the car, time of order picking, etc. These data can be used to inform the client about the assigned car.

The statistics is collected using database servers, but to present information in a convenient form it is necessary to use the Tari_cator. Tari_cator program allows you to display statistics either in a text document or in an excel format document. Fig. 2 presents the revised information on the distribution of orders, preserving the properties of the real sequence. We will make models fit according to this data.

It should be noted that the process in Fig. 2 has a heterogeneous structure, as well as some recurrent features. It is therefore necessary to select the most adequate model to more accurately describe all the peculiar distribution characteristics.

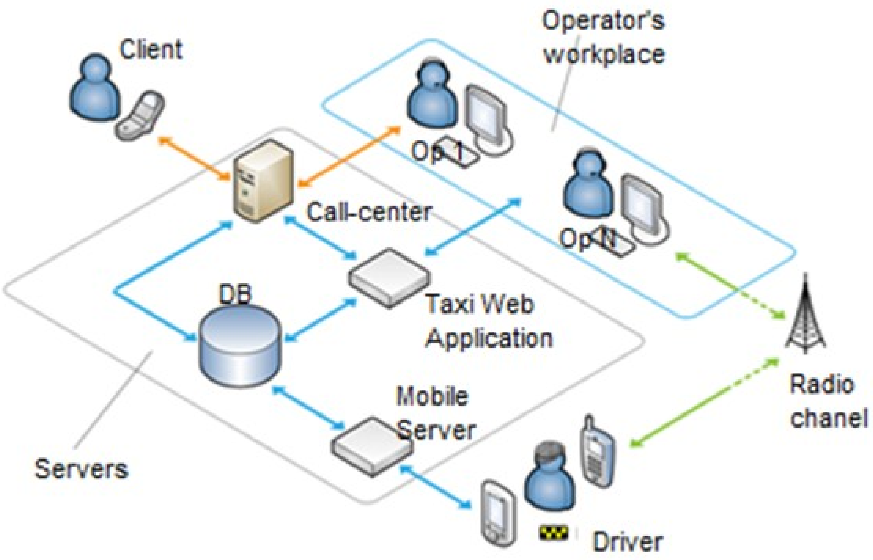

Fig. 1. Block diagram of the taxi service.

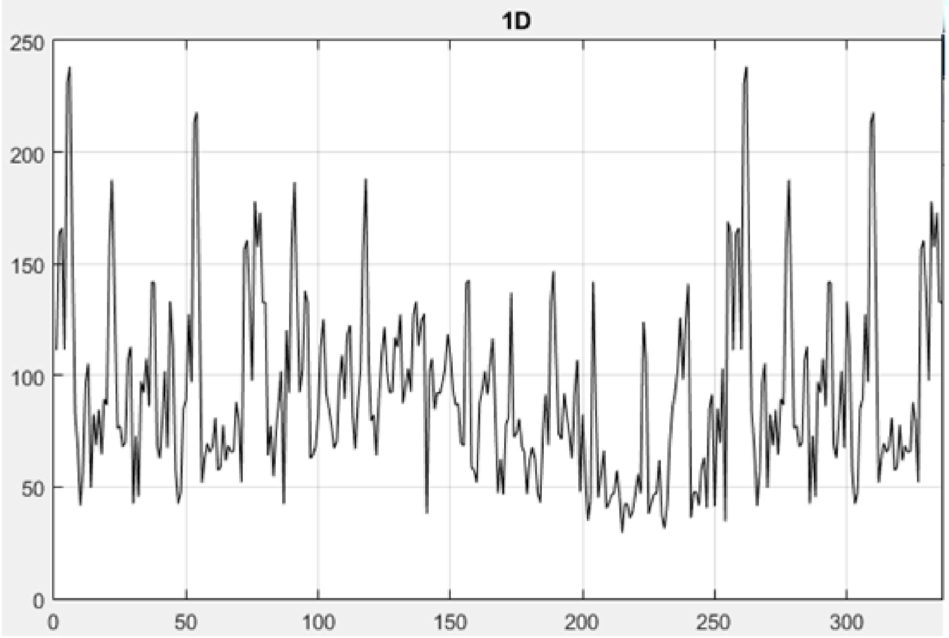

Fig. 2. Distribution of orders daily with the conversion (along the X axis is the number of orders, along the Y axis is the certain day).

3. Mathematical models for the presentation of taxi service statistics

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 (from January) and until the end of the year (to December) with some simplification, which will be used in the presentation approach in the form of an image.

3.1. One-dimensional Autoregressive process

Let's imagine a sequence of data available on orders fOg using an expression for the Autoregressive (AR) of the first order

$$O_i = \rho O_{i-1} + \xi_i, i = 1, \dots, N \qquad (1)$$

where ρ is a coefficient of correlation throughout the sequence and can easily be evaluated on the basis of existing data; ξ_i is accidental admixture with zero mathematical expectation and variance $\sigma^2_{\xi_i} = \sigma^2_O(1\text{-}\rho^2)$.

Besides the variance for orders is also estimated on the basis of the sample.

AR processes of higher orders can be used for a more accurate description. In this case, it is need to use the Yule-Walker equations [7] to determine the correlation parameters.

3.2. One-dimensional doubly stochastic model of Random Process

Descriptions of the heterogeneity and periodicity of real data can be achieved using mixed models of Random Fields (RF). One of the variants to realize mixed models is the doubly stochastic model [8,9], whose correlation parameters also represent the implementation of the RF:

$$O_i = \rho_i O_{i-1} + \xi_i, i = 1, \dots, N \qquad (2)$$

where ξ_i is the random additive value with zero mathematical expectation and variance $\sigma_{\xi_i}^2 = \sigma_O^2(1\text{-}\rho_i^2)$; ρ_i is a sequence of correlation parameters

$$\rho_i = \tilde{\rho}_i + m_\rho, \tilde{\rho}_i = r\tilde{\rho}_{i-1} + \sqrt{\sigma_\rho^2(1 - r^2)}\varsigma_i, i = 1, ..., N \tag{3}$$

where r is the constant correlation coefficient; m_p is the average value of the basic correlation coeffcient; σ_ρ^2 is the dispersion of the process describing change in the correlation parameters; $\{\varsigma_i\}$ is a field of Gaussian random variables with zero mathematical expectation and variance of unit.

An increase in the order of the process can also be used for the model (2) and its parameters (3), respectively. However, Fig. 1 shows the process which looks fairly ”prickly”. This fact allows the use of first-order models.

It is important that the estimation of all parameters of the model can be performed by mathematical statistics using the available sample, but also satisfactory results can be obtained with a slight increase in complexity, for example, in estimating all the parameters of the model in a sliding window [10] or using a nonlinear Kalman filter [11]. In addition, such algorithms can be adapted to different dimensionalities of the models.

3.3. Presentation in the form of a Random Field

The observed quasi-periodicity of the process shown in Fig. 2, allows us to conclude that it is possible to use models of random fields to represent information of this kind. Consider, for example, the doubly stochastic models of images that allow describing heterogeneous signals [12]. As an example, we will use the following model:

$$\begin{gathered} O_{i,j} = 2\rho_{xi,j}O_{i-1,j} + 2\rho_{yi,j}O_{i,j-1} - 4\rho_{xi,j}\rho_{yi,j}O_{i-1,j-1} - \rho_{xi,j}^2 O_{i-2,j} - \rho_{yi,j}^2 O_{i,j-2} + \\ +2\rho_{xi,j}^2\rho_{yi,j}O_{i-2,j-1} + 2\rho_{xi,j}\rho_{yi,j}^2 O_{i-1,j-2} - \rho_{xi,j}^2\rho_{yi,j}^2 O_{i-2,j-2} + b_{i,j}\xi_{i,j}, i = 1, ..., N_1, j = 1, ..., N_2, \end{gathered} \tag{4}$$

where $O_{i,j}$ is modeled RF with a normal distribution having $M\{O_{i,j}\} = 0, M\{O_{i,j}^2\} = \sigma_O^2$; $\{\xi_{i,j}\}$ is RF of independent standard Gaussian variables with $M\{\xi_{i,j}\} = 0, M\{xi_{i,j}^2\} = \sigma_\xi^2 = 1$; $\rho_{xi,j}$ and $\rho_{yi,j}$ are correlation coefficients of the model with multiple roots of characteristic equations of frequency rate (2,2) [13]; $b_{i,j}$ is a scale coefficient of simulated RF.

Random variables$\rho_{xi,j}$; j and $\rho_{yi,j}$ have the Gaussian probability distribution function and can be described by AR equations of the first order or higher orders.

It is easy to see that the model (4) is a transformation of the usual two-dimensional autoregressive model of the first order. This model of RF can also be used to describe a two-dimensional array of data and has the form:

$$\begin{gathered} O_{i,j} = 2\rho_x O_{i-1,j} + 2\rho_y O_{i,j-1} - 4\rho_x\rho_y O_{i-1,j-1} - \rho_x^2 O_{i-2,j} - \rho_y^2 O_{i,j-2} + \\ +2\rho_x^2\rho_y O_{i-2,j-1} + 2\rho_x\rho_y^2 O_{i-1,j-2} - \rho_x^2\rho_y^2 O_{i-2,j-2} + b_{i,j}\xi_{i,j}, i = 1, ..., N_1, j = 1, ..., N_2. \end{gathered} \tag{5}$$

Note that the model (4), unlike the model with constant parameters (5), imitates heterogeneous in the structure of the RF, so it can fairly well reflect sharp surges on the number of orders on weekends and holidays. In order to estimate the parameters of such an image, we can use a vector (row-by-row) nonlinear Kalman filter. It requires to combine the elements of the image string into a vector $\vec{x}_i = (x_{i1}, x_{i2}, \dots, x_{iN})$. Then the model for a single frame of the image can be written as following equation:

$$\vec{x}_i = diag(\vec{\rho}_{xi})\vec{x}_{i-1} + \vartheta(\rho_{xi}, \rho_{yi})\vec{\xi}_i, \vec{\rho}_{xi} = r_{1x}\vec{\rho}_{x(i-1)} + \vartheta_{\rho_x}\vec{\xi}_{xi}, \vec{\rho}_{yi} = r_{1y}\vec{\rho}_{y(i-1)} + \vartheta_{\rho_y}\vec{\xi}_{yi},$$

where $diag(\vec{\rho}_{xi})$ is the diagonal matrix with elements $\vec{\rho}_{xi}$on the main diagonal; ϑ is down triangle matrix determined by the decomposition of covariance matrix: $V_x = \vartheta\vartheta^T$.

The evaluation process is described by the Kalman nonlinear filter:

$$\hat{\vec{x}}_{pi} = \hat{\vec{x}}_{epi} + P_i \frac{\partial\Phi^T}{\partial\vec{x}_{pi}} V_n^{-1}\left(\vec{z}_i - \hat{\vec{x}}_{epi}\right),$$

$$\vec{x}_{pi} = \begin{pmatrix} \vec{x}_i \\ \vec{\rho}_{xi} \\ \vec{\rho}_{yi} \end{pmatrix} = \Phi\left(\vec{\rho}_{x(i-1)}, \vec{x}_{i-1}\right) + \vartheta\left(\vec{\rho}_{x(i-1)}, \vec{\rho}_{y(i-1)}\right)\vec{\xi}_i,$$

where

$$\vec{x}_{epi} = \Phi\left(\vec{x}_{p(i-1)}\right), \Phi_p\left(\vec{x}_{p(i-1)}\right) = \begin{pmatrix} \Phi(\rho, x) \\ r_{1x}\vec{\rho}_{x(i-1)} \\ r_{1y}\vec{\rho}_{y(i-1)} \end{pmatrix}, \vec{\xi}_i = \begin{pmatrix} \xi_i \\ \xi_{xi} \\ \xi_{yi} \end{pmatrix}.$$

The use of this algorithm is possible if characteristics of information RF is exactly known, i.e. when we know the correlation coefficients $r_{1x}, r_{2x}, r_{1y}, r_{2y}$, as well as average values by row and column correlation, variance of correlation parameters and variance of information signal. Otherwise, a preliminary assessment of these parameters is required. Pseudogradient assessment procedures, as well as expressions for covariation function for doubly stochastic models can be used for this purpose. Produced at the output sequence of parameters can then be further parsed and replaced with any model. Also you can use and evaluation in the sliding window.

Fig. 3 shows the transformation of the original process to the 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 heterogeneity. We propose 6 variants of the models to describe the available data. Let’s compare them in detail.

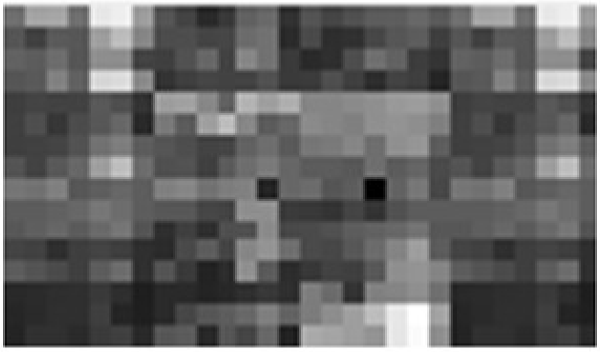

Fig. 3. Representation of orders statistics as an image.

4. Comparative analysis of efficiency of prediction based on different models

We will perform the necessary parameter estimation for models (1), (2), (4) and (5). So we produce forecasting the past 21 values of a sequence on the basis of models which was considered. It should be noted that the image data will be structured by seasons and weeks, as presented in Table 1.

Table 1. Data structure when converting it to 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	Data							Data							Data						
Week 2																					
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	Data							Data							Data						
Week 2																					
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	Data							Data							Data						
Week 2																					
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	Data							Data							Data						
Week 2																					
Week 3																					
Week 4																					

The latter values will form a rectangular area in the lower right corner of the image, which is also useful for predicting and comparing the results of prediction based on various models. Denote the forecasting methods as follows:

1) A1 is the prediction based on one-dimensional AR model;

2) A2 is the prediction based on one-dimensional doubly stochastic model;

3) A2* is the prediction based on one-dimensional mixed model with the evaluation parameters through the Kalman filter;

4) A3 is the prediction based on two-dimensional AR model;

5) A4 is the prediction based on two-dimensional doubly stochastic model;

6) A4* is the prediction based on mixed model with evaluation parameters through the Kalman filter in two-dimension mode.

Fig. 4 presents the results of statistical modeling.

Relative variance of the prediction error of the last twenty one value, respectively, are as following:

1) It equals 10.88 for one-dimensional (1D) AR model;

2) It equals 0.254 for one-dimensional (1D) doubly stochastic model;

3) It equals 0.067 for one-dimensional (1D) doubly stochastic model with Kalman filter evaluation;

4) It equals 0.870 for two-dimensional (2D) AR model;

5) It equals 0.174 for two-dimensional (2D) doubly stochastic model;

6) It equals 0.049 for two-dimensional (2D)doubly stochastic model with Kalman filter evaluation.

Thus, analysis of the diffierent models predicting results allows to say that using AR model leads to unsatisfactory results when predicting of complex data. Improving the effectiveness of predicting by the statistical models can be get using models of images. But such assessment will also not effective enough. So doubly stochastic models provide the best indicators because such models take into account the heterogeneity inherent in real data. Moving to the multivariate case leads to better forecast because of the characteristics of the analyzed data set. In addition, the highest accuracy of prediction algorithms which were considered is provided by doubly stochastic models of the images. For such models estimation of parameters is performed using the Kalman filter.

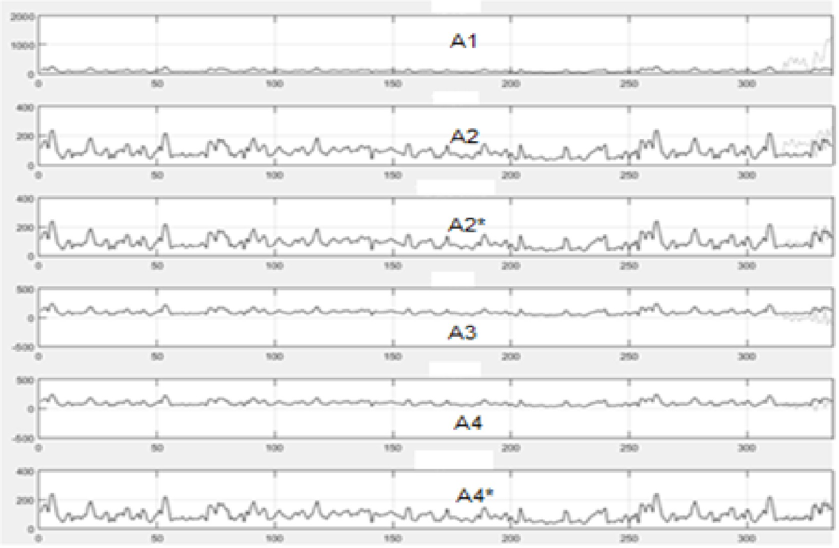

Fig. 4. Predicting the past values of the taxi service orders and real data (on X axis we have converted number of orders, on Y axis we have the certain day of the year).

5. Software package for statistical analysis of data on taxi service

As mentioned earlier the following algorithm of the taxi service is widely known. A dispatcher receives a call from the client and then communicates with a driver on the radio and transmits the order details to the driver. However, there is an alternative variant to this scheme of work at present time. The rapid growth of Internet traffic and the possibilities of IP-telephony (SIP), the availability of smartphones running under iOS or Android in each family allow you to abandon the use of radios when organizing a taxi order service.

At the same time, the task of optimizing the work of the taxi service is quite relevant, because this type of service is still in demand even during a finance crisis. Moreover, the opportunity to improve the efficiency of the service and save money by switching to automated mode is a very promising task.

Thus, the solution of this problem implies research at the junction of information and telecommunications systems. Indeed, it is necessary to realize not only communication networks that allow to exchange the information between operators, taxi drivers and customers, but also have software implementation of algorithms for handling calls and orders. At the same time, an important study is the statistical analysis of orders data.

We have solved the number of tasks during implemention of the software. First of all, the structure of the database has been developed. All the connections were thought out in the database, all the necessary information was collected. Secondly, it was suggested to use mixed or doubly stochastic autoregressive models to solute the orders forecasting problem. Third, we suggested a number of procedures based on the forecast data. We described how to calculate call traffic, determine the required number of operators. Fourthly, a Web-based interface has been developed that allows you to quickly change the settings on the telephony server.

The organization of the taxi service is performed on the basis of integration with the contact center (Telephony Server). Furthermore, the service includes:

- the database server;
- the Web server;
- the application server running the special Taxi program.

Fig. 5 shows the work of the Contact Center in more detail when the operator processes the order form. After such processing the database is updated and the order for taxi drivers is distributed.

Using the programming languages PHP and JavaScript, we developed the web-based interface for analyzing order data. As we mentioned earlier the interface can be conditionally called Tarififcator and allows you to obtain various statistical characteristics, as well as implement database modifications that are aimed at changing prices. In addition, you can view statistics on orders in real time using Tarifficator.

Another application, implemented by PHP and JavaScript, is the calculator of complex routes. The program allows you to calculate the cost of an order in the case when the driver passes several points in sequence. For example, cabbie drives first from point A to point B, and then from point B to point C.

The module for data analysis has been improved for convenience of the operating with different statistics in the languages PHP and JavaScript. This module (Fig. 6) allows to draw various statistical graphs using the library flot.js, and it also allows to make changes in the database related to setting prices. In addition, it is very important that in the Tarifficator module all necessary statistics is collected in real time mode.

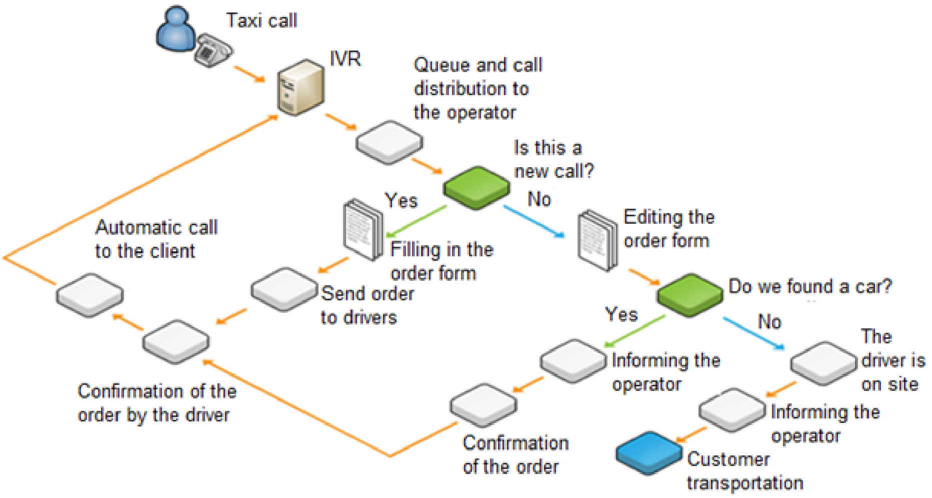

Fig. 5. Call processing algorithm.

You also can use module that allows the fitting of real data for operating the statistics module. So you can use statistical models of random sequences. Parameter identification may be implemented for the distribution of orders daily, calculated by the common AR model. Using these data the module will give forecast for the following days. Doubly stochastic models allow to consider the non-stationary in the distribution of data (bursts on weekends). For the such models you can use parameter identification algorithms based on a combination of algorithms of pseudogradient search and nonlinear Kalman filter [14].

The developed program complex allows you to accurately forecasting based on the doubly stochastic models of the images. Thus, improving the efficiency of taxi services is possible through the right choice of the necessary number of drivers in different time intervals. Similarly, it is possible to calculate, for example, the required number of call-center staff for different time periods.

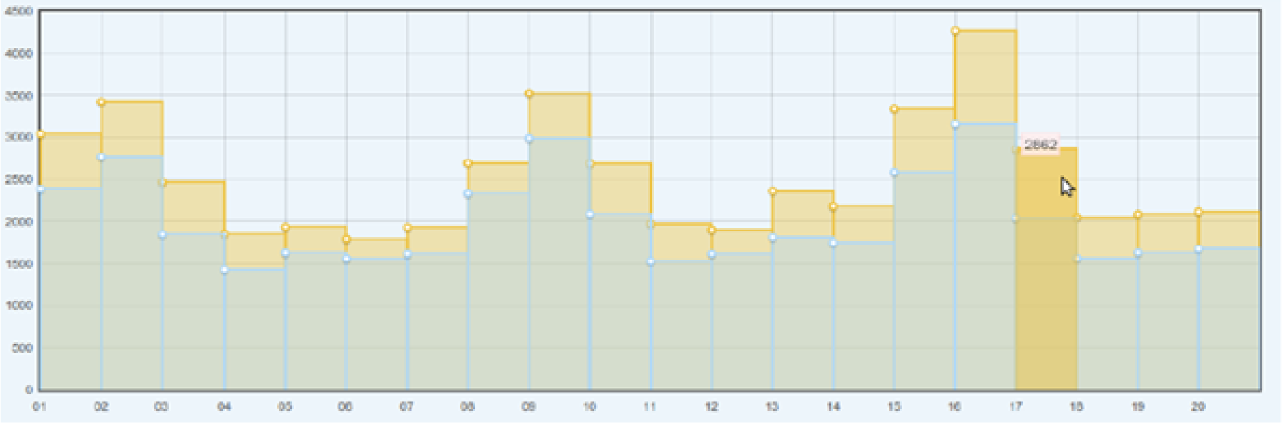

Fig. 6. Example of presenting statistics in the Tarifficator module.

Thus, you can get rid of radio communication and go to a software complex that handles data through the Internet. At the same time, telephony in the contact center means not the operators attached to the handset, but the people who process the data directly on the computer. In our project the dispatching taxi is organized with the help of the powerful software and hardware complex. Furthermore, thousands of cars and more can work simultaneously. So it is possible to completely abandon the use of a radio for a taxi service. Obviously, the use of such technology allows you to effectively manage resources, increase the speed of processing orders, always have exact customer numbers, reduce the time for applications running. And in order that the work of the taxi dispatcher was possible and necessary condition is the presence of a standard computer and a headset with a microphone.

6. Conclusion

The problem of analysis and optimization of the taxi order service efficiency is considered. It is suggested to use the doubly stochastic models of images to account for the heterogeneity of the data. A comparative analysis of forecasting based on 6 different models is carried out. In this case, the gain in comparison with autoregressive ones can reach several orders, and by applying the Kalman vector nonlinear filter it is possible to increase the forecast efficiency by another 4-5 times. A powerful software and hardware complex was developed. It will be used in the work of taxi order services and provide a solution to the task of real-time forecasting.

Acknowledgements

This work was supported by the OOO ”EIS-PFO” (Ulyanosk, Russia). We express our special gratitude for provided information used in the research.

References

[1] Andriyanov NA, Danilov AN. Taxi service with forecasting statistics based on complex mathematical models Advances of modern science 2016; 2(10): 114–116. (in Russian)

[2] Yarushkina NG, Afanasyeva TV, Perfilieva IG. Time series mining. Students book. Ulyanovsk: UlGTU, 2010; 320 p. (in Russian)

[3] Prokis J. Digital communications. Translated from eng. Edited by Klovskiy DD. Moscow: Radio and communications, 2000; 800 p.

[4] Borovkov AA. Probability Theory. Springer Science and Business Media; 536 p.

[5] Meggelen J, Madsen L, Smith J. Asterisk: future of the telephony. 2-nd edition, translated from eng. SPb: Symbol-Plus, 2009; 656 p.

[6] Goldstein BS, Zarubin AA, Samorezov VV. Session Initiation Protocol (SIP): Reference book. Series: Telecommunication protocols of Russia, 2005; 456 p. (in Russian)

[7] Andriyanov NA, Dementyev VE. The application of the system of equations of the Yule-Walker to simulate isotropic random fields. Modern trends of technical sciences. IV International Scientific Conference materials. Kazan, Russia, 2015: 2–6. (in Russian)

[8] Vasil'ev KK, Dement'ev VE, Andriyanov NA. Doubly stochastic models of images. Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications) 2015; 25(1): 105–110. DOI: 10.1134/S1054661815010204.

[9] Andriyanov NA. Doubly stochastic models based on the correlation interval changes. Mathematical methods and models: theory, application and role in education 2014; 3: 6–8. (in Russian)

[10] Andriyanov NA. Method of fitting images based on random field model with changing parameters. Advances of modern science 2016; 5(9): 98–100. (in Russian)

[11] Vasil'ev KK, Dement'ev VE, Andriyanov NA. Application of mixed models for solving the problem on restoring and estimating image parameters. Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications) 2016; 26(1): 240–247. DOI: 10.1134/S1054661816010284.

[12] Dementyev VE, Andriyanov NA. The using of doubly stochastic models of random processes and fields to describe complex heterogeneous signals. Actual problems of physical and functional electronics. Materials of 19-th all-Russian youth scientific schoolseminar. Ulyanovsk: UlGTU, 2016; 98–99. (in Russian)

[13] Vasiliev KK, Krasheninnikov VR. Statistical image analysis. Ulyanovsk: UlGTU, 2014; 214 p. (in Russian)

[14] Vasiliev KK, Dementyev VE, Andriyanov NA. Parameter estimation of doubly stochastic random fields. Radio 2014; 7: 103–106. (in Russian)