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

P.T. Azanov[a], A.N. Danilov[b], N.A. Andriyanov[c]

[a] *"Tango Telecom", 426057 Krasnaya street 122, Izhevsk, Republic Udmurtia, Russia*
[b] *"Taxi Ulyanovsk", 432071, pr-t Narimanova, 1/3, Ulyanovsk, Russia*
[c] *Ulyanovsk State Technical University, 432027, 32 Severniy Venets street, Ulyanovsk, Russia*

Abstract

In the present work it is suggested to use statistical models for taxi service data analysis and forecasting. Special attention is paid to the application of methods of model parameters identification and short-term forecasts making. We suggest to use the mathematical models of images to account the alternating character, associated with the dependence of the number of taxi orders from various parameters. In addition the possibility of improving the effectiveness of evaluation through the use of mixed models of random fields is shown.

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

1. Introduction

In recent years, the following algorithm of processing taxi data has been enough distributed. Firstly, dispatcher received a call, then operator had to communicate with the driver and send the latter order information, e.g. by radio. However, with the development of the Internet provided promising opportunities to use this network in taxi service [1]. It is easy now to book a taxi directly to the portal on the Internet or by using special applications for Smartphone. In such cases, we do not consider a very important source of orders from customers, namely "customers from phone or cell phone".

In doing so, it may be noted that this treatment also provides sufficient statistics, analysis of which may allow, in the future, improve the quality of taxi service. In the case of inserting a new data into DB of phone calls we have a possibility to analyze talk time, the definition of "most popular" establishments in the city, etc. Adding to such information the statistics on orders, including execution time, standby time, car distribution by hours and other parameters, you can have a fairly complete statistical description of the work of the such service.

Thus, there is a pressing task of analysis of the collected information to improve the service effectiveness. So, e.g., performing accurate forecasts for the number of calls and tracking the progress of the orders, you can anticipate the required number of dispatchers and drivers. In doing so, to work with the stored information it can be used as a time-series [2], and different models of stochastic processes (RP) [3, 4].

2. Service Architecture and Statistics Collection

Let's consider the project based on contact center. Such way means a using of IP-telephony for operators and requires only a computer with a headset. To organize a taxi dispatching office it is necessary to have a powerful hardware-software complex. Its application allows multiple thousands of machines to work in real time mode.

It is obvious that the use of this technology allows you to effectively manage resources, increase the speed of processing orders, always have the exact phone numbers of the customers and reduce the time of receiving applications.

Call-center operating requires a multi-channel phone number that will allow you to make lots of calls simultaneously. To do this, you can use IP telephony technology. One of the most common telephony servers is Asterisk Server [5] that enables you to work with SIP telephony [6]. Such PBX must be configured to call distribution by taxi service operators. Call processing takes place using a special program representing to the operator the form of taxi order based on the Internet browser. To store information about calls we can use the database server MySQL. Configuring tariffs is done using a separate module-"Tarifficator", which is programmed for use on the web..

Thus, it is advisable to use virtualization techniques for separation of various servers, including the telephony server, database server and web server. In addition, there was also a need for an application server through which you transfer information from the call-center to the drivers. This is ensured by the special program for the taxi. And here we recommend you to use another database server for the storage of information about orders.

Fig. 1 presents full considered service architecture.

Application for the "Taxi" program can have a version running just under java or modern-oriented devices running Android and iOS.

With the reception of the order by certain driver database is updated, e.g., program prescribed information on car, taking order time, etc. This can be used to inform the client about the designated car.

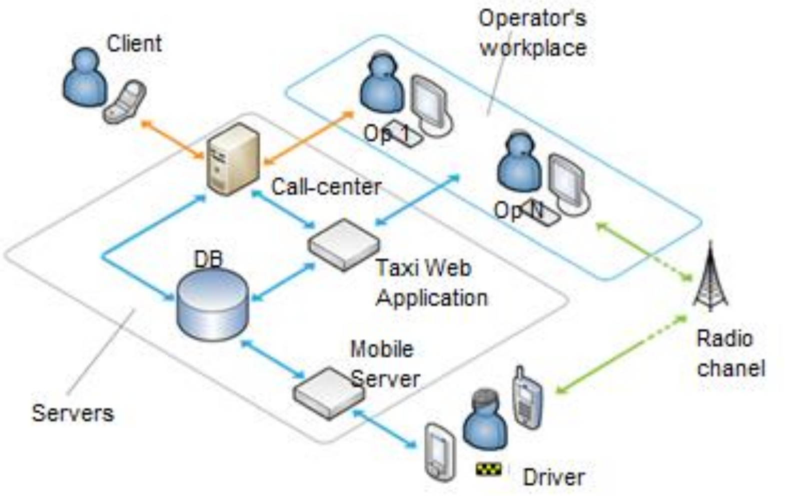

Fig. 1. Block diagram of the taxi service.

Statistics gathering is carried out with the help of database servers, but we get the presentation of information in a convenient form by using the "Tarifficator" that allows you to display statistics either in the text document, or in the document in excel format. Fig. 2 shows the redesigned information about the distribution of orders, which has preserved properties real sequence. We will produce fit models to this data.

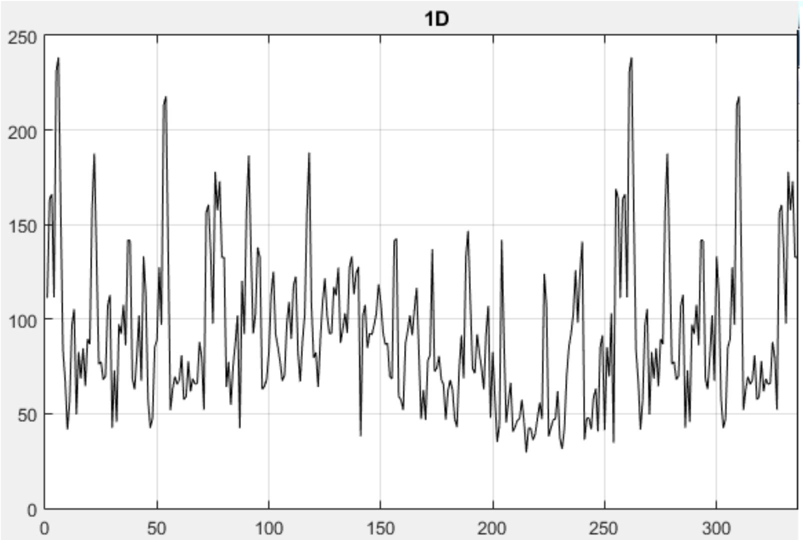

Fig. 2. Distribution of orders daily with the conversion (along the X axis is the number of orders, along the Y - the certain day).

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.

3. Mathematical models for the presentation of statistics taxi service

Let's look at a few of the options how describe the collected statistics. Let the data are collected since the beginning of the year (January) and before the end of the year, with some simplification, which will be used when we suppose to transform linear data to the images.

3.1. One-dimensional Autoregressive process

Let's imagine a sequence of data available on orders $\{O\}$ using an expression for the Autoregressive (AR) of the first order:

$$O_i = \rho O_{i-1} + \xi_i, i = 1...N, \tag{1}$$

where ρ is a coefficient of correlation throughout the sequence and can easily be evaluated on the basis of existing data; ξ_i - accidental admixture with zero mathematical expectation and variance $\sigma_\xi^2 = \sigma_O^2(1-\rho^2)$.

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

For a more precise description of the data we can use higher orders of AR process. In this case, it is advisable to use the equations of Yule-Walker [7] for determination the correlation parameters.

3.2. One-dimensional doubly stochastic model of Random Process

Description of the heterogeneity and the periodic feature of real data can be achieved by using mixed models of Random Fields (RF). One of the ways to implement mixed models is doubly stochastic model [8,9] correlation parameters of which constitute implementation of RF:

$$O_i = \rho_i O_{i-1} + \xi_i, i = 1...N, \tag{2}$$

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

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

where r - is the constant correlation coefficient; m_ρ - is the average value of the basic correlation coefficient; 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.

For models (2) and accordingly for its parameters (3) we can also use improving process orders. However, the process illustrated in Fig. 1. looks fairly "spiky" that allows the using of a first-order model.

It should be noted that the evaluation of all parameters of the model can be performed using the methods of mathematical statistics according to sample, but also satisfactory results can be obtained with a slight increase in complexity, e.g., in the assessment of all model parameters in sliding window [10] or by using non-linear Kalman filter [11], and the algorithms can be adapted to model dimensionality..

3.3. Presentation in the form of a Randpm Field

Noticed kvaziperiodic feature of the process (fig. 2.), allow to suggest that it is possible to use the RF models for presenting this kind of information. Let's consider the doubly stochastic models of images that allow you to describe mixed signals [12]. As an example, let's use the following model:

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

where O_{ij} - is modeled RF with a normal distribution having $M\{O_{ij}\} = 0$, $M\{O_{ij}^2\} = \sigma_O^2$; ξ_{ij} - is RF of independent standard Gaussian 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 characteristic equations of frequency rate (2,2) [13]; b_{ij} - is scale coefficient of simulated RF.

Random variables ρ_{xij} and ρ_{yij} have the Gaussian probability distribution function and can be described by AR equations of the first order or higher orders.

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

$$\begin{aligned} O_{ij} &= 2\rho_x O_{i-1,j} + 2\rho_y O_{i,j-1} - 4\rho_x\rho_y O_{i-1,j-1} - \rho_x^2 O_{i-2,j} - \rho_y^2 O_{i,j-2} + \\ &+ 2\rho_x^2\rho_y O_{i-2,j-1} + 2\rho_y^2\rho_x O_{i-1,j-2} - \rho_x^2\rho_y^2 O_{i-2,j-2} + b_{ij}\xi_{ij} \end{aligned} \tag{5}$$

Let's note that model (4), unlike models with constant parameters (5), simulates the heterogeneous structure of the RF, so can make a good reflection of surges in the number of orders during weekends and holidays. Parameter estimation for such images, you can use the vector (row) non-linear Kalman filter. To do this, combine the elements of the image string into a vector $\bar{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iN})^T$. Then model for a single frame of the image can be written as follows:

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

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

Evaluation process is described by non-linear Kalman filter::

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

where $\bar{x}_{эpi} = \Phi\left(\bar{x}_{p(i-1)}\right)$, $\Phi_p(\bar{x}_{p(i-1)}) = \begin{pmatrix} \Phi(\rho, x) \\ r_{1x}\bar{\rho}_{x(i-1)} \\ r_{1y}\bar{\rho}_{y(i-1)} \end{pmatrix}$, $\bar{\xi}_i = \begin{pmatrix} \bar{\xi}_i \\ \bar{\xi}_{xi} \\ \bar{\xi}_{yi} \end{pmatrix}$, P_i - covariation matrix of the filtration errors.

The use of this algorithm is possible if exactly known characteristics of information RF, i.e. the 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. For this purpose can be used pseudogradient assessment procedures, as well as expressions for covariation function for doubly stochastic models. 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.

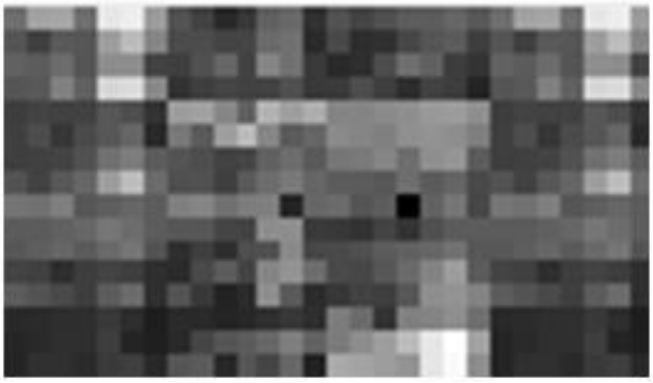

Fig. 3. View statistics of orders as an image.

Thus, we see that the resulting image, on the one hand, is not strongly correlated, and on the other hand, the image has several areas with higher values of brightness that testifies about the properties of heterogeneity. In general, it was suggested 5 models to describe the available data. But they also can have some modifications.

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). On the basis of models considered we will produce forecasting the past 21 values of a sequence. It should be noted that the image data will be structured by seasons and weeks, as presented in Table 1.

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 forecast (prediction) based on one-dimensional AR model;
2) A2 is the forecast (prediction) based on one-dimensional doubly stochastic model;
3) A2* is the forecast (prediction) based on one-dimensional mixed model with the evaluation parameters through the Kalman filter;
4) A3 is the forecast (prediction) based on two-dimensional AR model;
5) A4 is the forecast (prediction) based on two-dimensional doubly stochastic model;
6) A4* is the forecast (prediction) based on mixed model with evaluation parameters through the Kalman filter in two-dimension mode.

Table 1. Data structure when converting 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																					

Fig. 4 presents the results of statistical modelling.

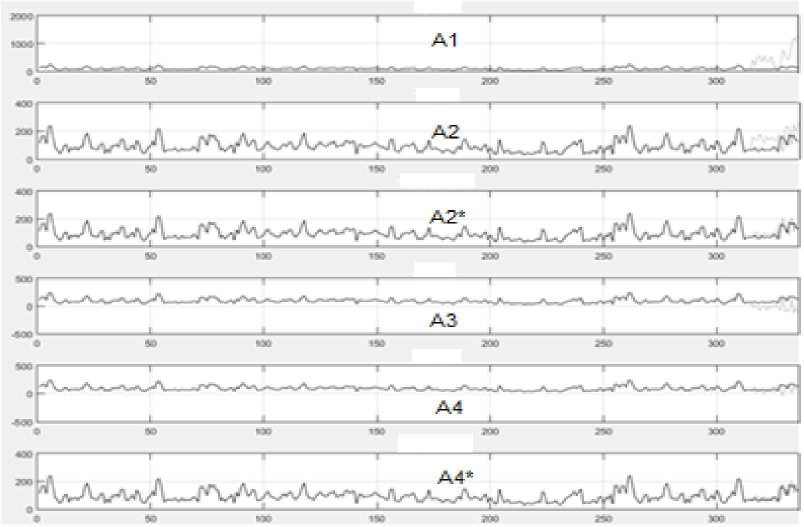

Fig. 4. Predicting the past values of the taxi service orders and real data (on X axis we have converted number of orders, on Y - the curtain day).

Relative variance of the prediction error of the last twenty one value, respectively, are as following:
1) for one-dimensional (1D) AR model - 10.88;;
2) for one-dimensional (1D) doubly stochastic model -0.254;
3) for one-dimensional (1D) doubly stochastic model with Kalman filter evaluation - 0.067;
4) for two-dimensional (2D) AR model - 0.870;
5) for two-dimensional (2D) doubly stochastic model - 0.174;
6) for two-dimensional (2D)doubly stochastic model with Kalman filter evaluation - 0.049.

Thus, analysis of the forecasting results allows to say that using AR model leads to unsatisfactory results when forecasting of complex data. Improving the effectiveness of forecasting can be done through models images. This assessment will also not effective enough. The best indicators provide doubly stochastic models, which 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.

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

Using the programming languages PHP and JavaScript we has developed web interface for data analysis on orders. This interface can be conditionally named "Tarifficator" and allows to obtain various statistical characteristics, as well as to make modifications of the database, oriented for prices. In addition, here you can view statistics on orders in real time.

For work with statistics module you also can use module that allows the fitting of real data, using statistical models of random sequences. Parameter identification, e.g., is implemented for the distribution of orders daily, calculated the common AR model and the module 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 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, e.g. the required number of call-center staff for different time periods.

6. Conclusion

The problem of analyzing and optimizing the efficiency of taxi service was considered. It was suggested to use doubly stochastic models of images to account for the heterogeneity of the data. The text presented the comparative analysis of prediction based on 6 different models. While winning compared to the AR models can reach several orders of magnitude, and by applying vector Kalman filter we can increase efficiency prediction back in 4-5 times.

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, N.A. Taxi service with forecasting statistics based on complex mathematical models / N.A. Andriyanov, A.N. Danilov //Advances of modern science. - 2016. - Vol. 2. No. 10. - P. 114-116 - (in Russian)

[2] Yarushkina N.G. Time series mining/ N.G. Yarushkina, T.V. Afanasyeva, I.G. Perfilieva // Students book. Ulyanovsk: UlGTU. -2010. - 320 p. - (in Russian)

[3] Prokis, J. Digital communications / Translated from Eng., Edited by. Klovskiy D.D. - Moscow: Radio and communications. - 2000. - 800 p.

[4] Borovkov, A.A. Probability Theory. / A.A. Borovkov. - Springer Science & Business Media. 536 p. ISBN 978-1-4471-5201-9.

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

[6] Session Initiation Protocol (SIP): Reference book / edited by B.S. Goldstein, A.A. Zarubin, V.V. Samorezov. - Series: Telecommunication protocols of Russia, 2005. - 456 p. - (in Russian)

[7] Andriyanov, N.A. The application of the system of equations of the Yule-Walker to simulate isotropic random fields /N.A. Andriyanov, V.E. Dementyev // Modern trends of technical sciences. IV International Scientific Conference materials. Kazan, Russia 2015, P. 2-6. - (in Russian)

[8] Vasil'ev, K.K. Doubly stochastic models of images / K.K. Vasil’ev, V.E. Dement'ev, N.A. Andriyanov // Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications). 2015. V. 25. № 1. P. 105-110. DOI: 10.1134/S1054661815010204

[9] Andriyanov, N.A. Doubly stochastic models based on the correlation interval changes // Mathematical methods and models: theory, application and role in education. 2014. № 3. - P. 6-8. - (in Russian)

[10] Andriyanov, N.A. Method of fitting images based on random field model with changing parameters // Advances of modern science. 2016. V. 5. No. 9. P. 98-100. - (in Russian)

[11] Vasil'ev, K.K., Application of mixed models for solving the problem on restoring and estimating image parameters / K.K. Vasil’ev, V.E. Dement'ev, N.A. Andriyanov // Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications). 2016. V. 26. № 1. P. 240-247. DOI: 10.1134/S1054661816010284

[12] Dementyev, V.E. The using of doubly stochastic models of random processes and fields to describe complex heterogeneous signals / V.E. Dementyev, N.A. Andriyanov //Actual problems of physical and functional electronics. Materials of 19-th all-Russian youth scientific school-seminar, 2016. – Ulyanovsk: UlGTU, P. 98-99. - (in Russian)

[13] Vasiliev, K.K Statistical image analysis / K.K. Vasiliev, V.R. Krasheninnikov -Ulyanovsk: UlGTU, 2014. -214 p. - (in Russian)

[14] Vasiliev, K.K. Parameter estimation of doubly stochastic random fields / K.K. Vasiliev, V.E. Dementyev, N.A. Andriyanov // Radio. 2014. №7. - P. 103-106. - (in Russian)