THE MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE RUSSIAN FEDERATION

ST. PETERSBURG STATE POLYTECHNICAL UNIVERSITY JOURNAL

Physics and Mathematics

VOLUME 11, No. 3
2018

Polytechnical University Publishing House
Saint Petersburg
2018

ST. PETERSBURG STATE POLYTECHNICAL UNIVERSITY JOURNAL. PHYSICS AND MATHEMATICS

The journal is included in the List of leading peer-reviewed scientific journals and other editions to publish major findings of theses for the research degrees of Doctor of Sciences and Candidate of Sciences.

The publications are presented in the VINITI RAS Abstract Journal and Ulrich's Periodical Directory International Database.

The journal is published since 2008 as part of the periodical edition 'Nauchno-tekhnicheskie vedomosti SPb-GPU'.

The journal is registered with the Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications (ROSKOMNADZOR). Certificate ПИ № ФС77-52144 issued December 11, 2012.

The journal is distributed through the CIS countries catalogue, the «Press of Russia» joint catalogue and the «Press by subscription» Internet catalogue. The subscription index is **71823**.

The journal is in the Russian Science Citation Index (RSCI) database.

Address: 195251 Politekhnicheskaya St. 29, St. Petersburg, Russia.
Phone: (812) 294-22-85.
http://ntv.spbstu.ru/physics

Contents

Astrophysics

THE WOOD IN THE INHOMOGENEOUS TEMPERATURE FIELD: ESTIMATION OF CELLULOSE STRUCTURE PARAMETER FLUCTUATIONS

N.N. Matveev[1], Hoai Thuong Nguyen[2], N.S. Kamalova[1], N.Yu. Evsikova[1], A.S. Chernykh[1]

[1]Voronezh State University of Forestry and Technologies named after G.F. Morozov, Voronezh, Russian Federation;

[2] Industrial University of Ho Chi Minh City, Vietnam

In the paper, the cellulose as a fiber-forming component of wood (natural composite) has been studied. The authors put forward a technique for estimating fluctuations of cellulose microstructure in the wood through monitoring the potential difference of the thermal polarization that arises in the samples placed into an inhomogeneous temperature field with a constant temperature gradient. Formalized simulation was used for an analysis of experimental results. The proposed technique made it possible to establish that the percent of the large-sized cellulose crystallites in the wood grew with increasing smoothly temperature gradient. Similar dynamics is not typical of linear crystalline polymers whose polarization decreases with growing temperature. The obtained effect can be assigned to the fact that natural wood exhibits heterogeneous structure.

Key words: microstructure, crystallite, composite, cellulose macromolecule, synthesized material

Citation: N.N. Matveev, H.T. Nguyen, N.S. Kamalova, N.Yu. Evsikova, A.S. Chernykh, The wood in the inhomogeneous temperature field: Estimation of cellulose structure parameter fluctuations, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 5–10. DOI: 10.18721/JPM.11301

Introduction

Creating synthesized bioplastics with highly resistant physical properties such as strength, surface hardness and permissible hydrophobicity is one of the most urgent tasks in technology [1 – 8]. The arboform exemplifies these materials; it can be obtained through the synthesis from natural cellulose and "sulphate soap" released during the paper production process. The physical characteristics of these synthesized materials are determined by the orderliness of their fiber-forming component microstructure, as cellulose microstructure in our case. In this regard, the development of nondestructive methods for estimating microstructure fluctuations in the fiber-forming component of composite materials always attracts attention of the scientific community.

It is well known that wood is naturally occurring composite material, and its main components are partially crystalline cellulose and lignin. Cellulose is a stereoregular syndiotactic polymer [9 – 12]. The macromolecules of the fiber-forming wood component (cellulose) are schematically arranged in the form of a coiled tape with a cross-section of 0.39 × 0.83 nm. Molecular chains of cellulose are packed in a mean length of 15 – 17 nm with a «loosening» section of 2.5 – 3.0 nm in length following. In addition, hollows of 0.5 – 1.0 nm are always located inside amorphous regions [13]. Thus, the packaging process of cellulose macromolecules is characterized by the alternation of crystalline and amorphous phases and the presence of pores in microfibrils in the wood. The peculiarities of this structure allow us to assume that the response of biocomposite such as the wood substance to the change in external factors

depends on the concentration of crystallites in the fiber-forming cellulose and their physical properties.

In the present work, a formalized model is proposed for estimating the fluctuations in the microstructure of cellulose in the wood on exposure to external nonuniform temperature fields. For this purpose, the concentration of cellulose crystallites has been chosen as the fluctuation parameter.

Experimental results

The temperature-scanning method was used for experimental investigation as described in detail in Refs. [14, 15]. In this method, an inhomogeneous temperature field providing a constant temperature gradient ∇T was applied to a thin-layer composite sample, and a thermal-origin electric field evolving as a result. The origin of this electric field in the wood can be bound up with the structural difference between cellulose and lignin and with pyroelectric and piezoelectric properties of fiber-cellulose crystallites as well [15]. The potential difference (PD) across this field depends on the degree of crystallinity of cellulose and is measured with controlled accuracy using electrical measuring instruments.

To determine the response of cellulose in the wood to the applied inhomogeneous temperature field, studies in fluctuations of the PD in the samples were carried out. The samples were prepared from birch wood containing up to 40 % moisture. The sample thickness l_0 was about 100 μm. A special measuring cell was used to change the temperature gradient in the wood layer as given in Ref. [15]. Thin sections of the wood were placed between massive brass rodes with the lower one heated. Therefore, the temperature gradient in the wood layer was controlled by the heating rate of the heated lower electrode. The PD was initially removed from the electrodes.

Fig. 1 shows the dynamics of the temperature gradient in a thin layer of wood during the tests. Fig. 2 shows the experimental data for the measurement of the corresponding PD presented in the form of circles. Comparing the two figures, we can affirm that the PD correlates with the changes in the temperature gradient in the layer as established in various studies [14 – 16]. Thus, the temperature-scanning method makes it possible to control the value of the temperature gradient in the layer using electrical measuring instruments. In this regard, we propose to estimate the average size of cellulose crystallites by analyzing the obtained data on the basis of a formalized model [17–20].

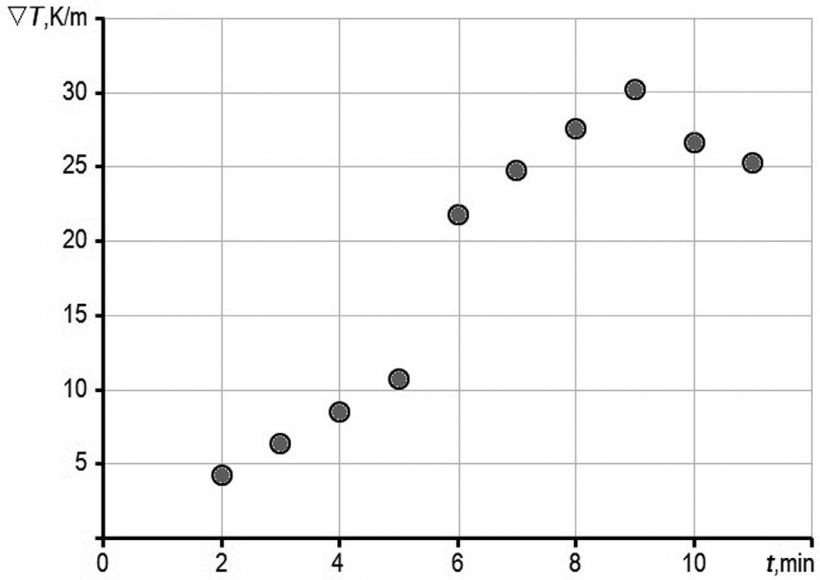

Fig. 1. The thermal-gradient dynamics for a wood thin layer during the test process

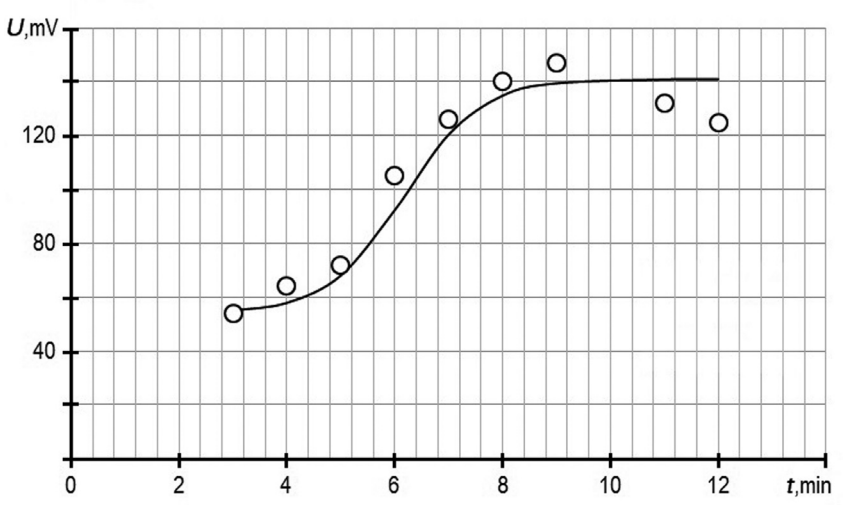

Fig. 2. The experimental (curcles) and simulated (the solid line) PD-time relation for a wood thin layer

Justification of the formalized model

It is known that the relative change in the concentration n of cellulose crystallites depends on the relative rate of a crystal growth under smoothly changing external conditions and characterized by the rate G [16]:

$$\frac{dn}{n} = Gdt. \qquad (1)$$

However, the crystal growth causes the diffusion of non-crystallizing fragments. This process is characterized by the coefficient k_D [16]:

$$\frac{dn}{dt} = -k_D \frac{dn}{dx}. \qquad (2)$$

These two processes (1) and (2) balance each other in a stationary state. Therefore, the equations can be transformed to the form:

$$\frac{dn}{dx} = (-G/k_D)n. \qquad (3)$$

The exponential function

$$n = n_0 \exp(-\Delta x/x_k)$$

is a solution of Eq. (3), where n_0 is the concentration of crystallites near crystallization centers, $x_k = k_D/G$ is the average size of the fiber-forming component crystallite.

This average size is determined when the concentration of crystallites located at a distance equal to crystallite size decreases by e times as compared to n_0.

It should be noted that these concepts of crystallites growth do not take into account the peculiarity of experimental conditions. The constant temperature gradient creates inhomogeneous growth conditions along the thickness of a sample. According to the obtained experimental results we can assume that the average size of the cellulose crystallite x_k depends on the increment of the crystallite concentration as follows:

$$x_k(n) = x_{k0}(1 + \chi\Delta n), \qquad (4)$$

where χ is a coefficient that characterizes the crystallinity degree of the cellulose in a sample, x_{k0} is the initial value of x_k.

The solution of differential Eq. (3) taking Eq. (4) into account is transformed to the following form:

$$\frac{\Delta n}{n_0} = \frac{\exp(-\Delta x/x_k)}{(1 + \chi \exp(-\Delta x/x_k))n_0}, \qquad (5)$$

where $\Delta x = \alpha l_0^2 \nabla T(t)$ is the value of the total compression of cellulose crystallites in a sample with the thickness l_0 during the expansion of lignin, α is the coefficient of thermal expansion of lignin.

According to Ref. [21] the ratio $\Delta n / n_0$ equals the relative change of crystallinity degree of cellulose in the wood. As reported in

Ref. [22], the PD appeared in the wood on exposure to an inhomogeneous temperature field is directly proportional to crystallinity degree of cellulose. Thus, the relative change in PD in the sample within the framework of this approach is simulated by the following relationship:

$$\frac{U - U_0}{U_0} = k_U \frac{\Delta n}{n_0} = \frac{k_U \exp(-\Delta x/x_k)}{(1 + \chi \exp(-\Delta x/x_k))n_0}, \quad (6)$$

where k_U is a parameter that depends on the percolation features of thermal polarization processes occurred in the composite, U_0 is the PD initial value.

Finally, we obtain the relation for estimating the PD:

$$U = U_0 \left\{ 1 + \frac{k_U \exp(-\alpha l_0^2 \nabla T/x_k)}{[1 + \chi \exp(-\alpha l_0^2 \nabla T/x_k)]n_0} \right\}. \quad (7)$$

Eq. (7) connects the PD in the sample on exposure to an inhomogeneous temperature field with the fluctuations in external conditions such as the changes in $\nabla T(t)$ and the features of fiber-forming microstructure (x_k, χ) and filler (α).

Furthermore, Eq. (7) is the basic axiom of the formalized model for the method of estimating the response of natural component-containing microstructure to fluctuations of external conditions in general and temperature in particular. The model experiment was implemented by the linear regression method using Excel spreadsheets. The results are presented by the solid line in Fig. 2. Comparing the results of the real and simulated experiments (see Fig. 2) we can conclude that it is possible to estimate the values of χ, x_k and k_U parameters from the results of physical and simulated experiments with controlled accuracy.

Summary

Thus, it has been shown that the temperature-scanning method using elements of formalized simulation makes it possible to estimate the fluctuations of supramolecular structure of the fiber-forming component in a composite when changing the external conditions. Consequently, it can also be used to study the microstructure of arboforms and synthesized plastics.

Furthermore, analysis of the PD dynamics with a smoothly increasing temperature gradient suggests that the fraction of cellulose crystallites with a large size in the wood grows with increasing the temperature gradient value. It should be noted that similar dynamics do not characterize linear crystallizing polymers, in which polarization decreases with increasing temperature. Perhaps, the considered effect is due to interaction between wood components and the cellulose characterized by the complexity of supramolecular structure.

The work was supported by the grant «Development of Innovative Ideas "Growth Points – 2017"» of the Federal State Budgetary Educational Institution of Higher Education «Voronezh State University of Forestry and Technologies named after G.F. Morozov».

REFERENCES

[1] **A.M. Kamalov, M.E. Borisova,** The influence of moisture on charge relaxation in modified polyimide films, St. Petersburg Polytechnical University Journal. Physics and Mathematics. (2) (2016) 188–192.

[2] **A. Abdulkhani, J. Hosseinzadeh, S. Dadashi, M. Mousavi,** A study of morphological, thermal, mechanical and barrier properties of PLA based biocomposites prepared with micro and nano sized cellulosic fibers, Cellulose Chemistry and Technology. 49 (7–8) (2015) 597–605.

[3] **Yu.M. Spivak, V.A. Moshnikov,** Features of photosensitive polycrystalline PbCdSe layers with a network-like structure, Journal of Surface Investigation. X-Ray, Synchrotron and Neutron Techniques. 4(1) (2010) 71–76.

[4] **E.V. Maraeva, V.A. Moshnikov, Yu.M. Tairov,** Models of the formation of oxide phases in nanostructured materials based on lead chalcogenides subjected to treatment in oxygen and iodine vapors, Semiconductors. 47(10) (2013) 1422–1425.

[5] **D. Lingam, A.R. Parikh, J. Huang, et al.,** Nano/microscale pyroelectric energy harvesting: challenges and opportunities, International Journal of Smart and Nano Materials. 4(4) (2013) 229–245.

[6] **S.K.T. Ravindran, T. Huesgen, M. Kroener, P. Woias,** A self-sustaining micro thermomechanic-pyroelectric generator, Applied Physics Letters. 99 (10) (2011) 104102.

[7] **T.C. Harman, P.J. Taylor, M.P. Walsh, B.E.**

LaForge, Quantum dot superlattice thermoelectric materials and devices, Science. 5590 (297) (2002) 2229–2232.

[8] **A.J. Boukai, Yu. Bunimovich, J. Tahir-Kheli, et al.,** Silicon nanowires as efficient thermoelectric materials, Nature. 451(10) (2008) 168–171.

[9] **H.T. Nguyen, A.S. Sidorkin, S.D. Milovidova, O.V. Rogazinskaya,** Investigation of dielectric relaxation in ferroelectric composite nanocrystalline cellulose-triglycine sulfate, Ferroelectrics. 498 (1) (2016) 27–35.

[10] **H.T. Nguyen, S.D. Milovidova, A.S. Sidorkin, O.V. Rogazinskaya,** Dielectric properties of composites based on nanocrystalline cellulose with triglycinesulfate, Physics of the Solid State. 57(3) (2015) 503–506.

[11] **A.Yu. Milinskiy,** Dielectric properties of nanocrystalline cellulose-potassium iodide composites, St. Petersburg Polytechnical University Journal. Physics and Mathematics. 3 (2017) 57–62.

[12] **B. Lindner, L. Petridis, J.C. Smith, P. Langan,** Determination of cellulose crystallinity from powder diffraction diagrams, Biopolymers. 103(2) (2015) 67–73.

[13] **L. Mandelkurn,** Crystallization of polymers, 2nd edition, Vol. 1, Cambridge University Press, Cambridge, 2004.

[14] **N.N. Matveev, N.S. Kamalova, N.Yu. Evsikova, O. Farberovich,** Influence of structural inhomogeneities on the formation of the pyroelectric phase in polymers, Physics of the Solid State. 57(6) (2015) 1148–1150.

[15] **N.N. Matveev, N.S. Kamalova, N.Yu. Evsikova, et al.**, Emergence of differences in potential in wood as a result of natural changes in temperature, Bulletin of the Russian Academy of Sciences: Physics. 80(9) (2016) 1158–1160.

[16] **B. Wunderlich,** Macromolecular Physics, Vol. 3: Crystal Melting, Academic Press, New York, 1980.

[17] **V.I. Arnold,** «Zhestkie» i «myagkie» matematicheskie modeli [«Hard» and «soft» mathematical models], MCNMO, Moscow, 2004.

[18] **K. Chattopadhyay, A.K. Tiwari, D. Singh, et al.,** A systematic analytical study on lignocelluloses originated inhibitors in hydrolyzed biomass, Cellulose Chemistry and Technology. 49(1) (2015), 81–85.

[19] **D.A. Kozhanov,** The features of finite-element modeling of a structural element of flexible woven composites, St. Petersburg Polytechnical University Journal. Physics and Mathematics. 2 (2016) 1–6.

[20] **K.C. Cunha, V.H. Rusu, I.F.T. Viana, et al.,** Assessing protein conformational sampling and structural stability via de novo design and molecular dynamics simulations, Biopolymers. 103(6) (2015) 351–361.

[21] **V.V. Postnikov, N.Yu. Evsikova, N.S. Kamalova, N.N. Matveev,** Stepen kristallichnosti tsellyulozy i termicheskoye skanirovaniye [The degree of cellulose crystallinity and the thermal scanning], INTERMATIC-2009: Materials of the International Scientific and Technical Conference, The Russian Academy of Sciences, Energoatomizdat, Moscow, 1 (2009) 197–199.

[22] **N.Y. Evsikova, N.S. Kamalova, N.N. Matveev, V.V. Postnikov,** A new approach to determining the degree of cellulose crystallinity in wood, Bulletin of the Russian Academy of Sciences: Physics. 74(9) (2010) 1317–1318.

Received 28.12.2017, accepted 20.06.2018.

THE AUTHORS

MATVEEV Nikolay N.
Voronezh State University of Forestry and Technologies named after G.F. Morozov
8 Timiryazev St., Voronezh, 394087, Russian Federation
nmtv@vglta.vrn.ru

NGUYEN Hoai Thuong
Industrial University of Ho Chi Minh City
12 Nguyen Van Bao, Ward 4, Go Vap, Ho Chi Minh, Vietnam
nguyenthuongfee@iuh.edu.vn

KAMALOVA Nina S.
Voronezh State University of Forestry and Technologies named after G.F. Morozov
8 Timiryazev St., Voronezh, 394087, Russian Federation
rc@icmail.ru

EVSIKOVA Nataliya Yu.
Voronezh State University of Forestry and Technologies named after G.F. Morozov
8 Timiryazev St., Voronezh, 394087, Russian Federation
natalyaevsikova@mail.ru

CHERNYKH Alexander S.
Voronezh State University of Forestry and Technologies named after G.F. Morozov
8 Timiryazev St., Voronezh, 394087, Russian Federation
edu-ltu@vglta.vrn.ru

THE PHASE SEPARATION PHENOMENOLOGICAL MODEL: MANGANITE AS AN EXAMPLE

T.S. Shaposhnikova, R.F. Mamin

Zavoisky Physical-Technical Institute, FRC KazanSC of RAS, Kazan, Russian Federation

In the paper, an effect of a second order phase transition has been considered in the context of the phenomenological model for a 2D charged system (2DCS) frustrated by the Coulomb interaction. The relationship between the order parameter and the charge was treated as a local temperature in the 2DCS. The existence of phase-separated states was shown to be a possibility in such a system. Various types of those states (strips, rings, etc.) were found by numerical calculations, and their parameters were determined. As the temperature is lowered, the 2DCS passes several phase transitions successively. Using the $La(1-x)Sr(x)MnO_3$ manganite as an example it was shown that such a phenomenological model could be used to describe the phase separation close to a magnetic phase transition from a ferromagnetic state to a paramagnetic one when $0.10 < x < 0.15$ and at the temperatures of $100 < T < 200$ K.

Key words: second order phase transition, phase separation, manganite, Coulomb interaction, doping level

Citation: T.S. Shaposhnikova, R.F. Mamin, The phase separation phenomenological model: Manganite as an example, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 11–18. DOI: 10.18721/JPM.11302

Introduction

The problem of phase separation has attracted much attention from researchers [1 – 8]. There are two classes of materials where phase transitions (PT) are observed with various types of structural, magnetic, charge, and orbital ordering.

The first class is manganites with colossal negative magnetoresistance of the $R_{1-x}A_xMnO_3$ type (R = La, Pr, Sm, etc., A = Ca, Sr, etc.), whose physical properties are greatly affected by the concentration x of the divalent element A varying from zero to unity [1 – 5, 8].

The second class are high-temperature cuprate superconductors, where a pseudo-gap state and charge-density waves are observed [6, 7].

Phase separation of substances is often accompanied by charge inhomogeneities. Such inhomogeneities were observed by scanning tunneling microscopy [9], angle-resolved photoemission spectroscopy (ARPES) [6], X-ray and neutron diffraction [7]. For the above-mentioned compounds, there is a certain range of temperatures and doping levels for which the ground energy state corresponds to phase coexistence. The spatial size of single-phase regions depends on the ratio between the Coulomb energy (important in the presence of a doping-induced overcharge) and the energy gain due to the presence of a more ordered phase [10, 11]. States with charge inhomogeneity have been the focus of many theoretical studies (see, for example, Refs. [12 – 15]), usually considering a first-order phase transition frustrated by the Coulomb interaction. The scalar order parameter for this type of PT is either linearly coupled with charge density or proportional to it [13, 14]. These studies have established that such models are unstable with respect to phase separation. The phase-separated state is a group of charged regions of different phases with different values of the order parameter. Notably, this type of coupling of the order parameter to charge density is forbidden in case of a second-order phase transition and the order parameter is not a scalar quantity.

In this study, we discuss a second-order PT frustrated by the Coulomb interaction. We have considered the coupling between the charge density and the squared order parameter. We have established within the framework of this model that a phase-separated state with charge inhomogeneities can exist near the phase transition temperature T_c, where the high-temperature phase matrix with an order parameter equal to η_1 contains inclusions of a

low-temperature phase with an order parameter $\eta_2 > \eta_1$. Several successive first and second-order PTs can be observed with a change in the temperature.

We have applied the phenomenological approximation, based on the Ginzburg – Landau theory, to describe static phase separation in a two-dimensional system in the vicinity of a second-order phase transition. In this case, the presence of the Coulomb interaction associated with doping-induced overcharging is taken into account. Since the above-mentioned materials are quasi-two-dimensional (CuO planes in cuprates and MnO planes in manganites), the two-dimensional description adopted is a reasonable approximation. We have defined a set of parameters, related to temperature and doping, for which phase separation is energetically favorable. We have also found the region of the phase diagram where the inhomogeneous phases coexist.

We have calculated the model parameters suitable for describing phase separation near the magnetic PT of the second order in $La_{1-x}Sr_xMnO_3$ with $0.10 < x < 0.15$.

Theoretical model

Let us consider a two-dimensional system near a second-order phase transition. A study by Nobel Prize winner Pierre-Gilles de Gennes [16] investigated the effect of double exchange in mixed-valence compounds such as manganites $(La_{1-x}Ca_x)(Mn_{1-x}^{3+}Mn_x^{4+})O_3$. It was established that introducing extra holes or extra electrons into the antiferromagnet lowered the energy of the system. Additionally, the Curie temperature was found to depend on the doping level x. Following de Gennes's study, we start with a Hamiltonian, adding to it a term with the Coulomb interaction. The Hamiltonian for a "layer" antiferromagnet can be written in the following form:

$$H = -\sum_{ij} J_{ij}\mathbf{S}_i \cdot \mathbf{S}_j - \sum_{ij\sigma} t_{ij} a_{i\sigma}^{+} a_{j\sigma} - \\ - J_{\mathrm{H}} \sum_i \mathbf{S}_i \cdot \mathbf{s}_i + H_{\mathrm{Coul}}. \quad (1)$$

The first term in this expression describes the exchange interaction of Mn ions; $\mathbf{S}_i$, $\mathbf{S}_j$ are the spin operators of ionic spin at sites i and j; J_{ij} is the exchange integral (connecting only neighboring magnetic sites i and j); the second and third terms of Hamiltonian (1) describe a double exchange: the second term describes the jumps of an electron with spin σ along the ij sites of the lattice; $a_{i\sigma}^{+}$ ($a_{i\sigma}$) is the creation (annihilation) operator for an electron at site i; t_{ij} is the hopping integral; the third term in Hamiltonian (1) describes Hund's coupling [17], $\mathbf{s}_i$ is the spin operator of the conduction electron (it can be expressed in terms of the creation and annihilation operators for the electron and the Pauli matrices); the last term describes the Coulomb interaction.

Following de Gennes, we have assumed that spin ordering of an unperturbed system is of the "antiferromagnetic layer" type. Each ionic spin S is ferromagnetically coupled to z' neighboring spins the same layer and antiferromagnetically to z spins in the neighboring layers. The exchange integrals are equal to t_{ij}' and t_{ij}, respectively. Zener charge carriers [18] hop within their layer (with the hopping integral t_{ij}') and from one layer to another (with the hopping integral t_{ij}).

Let the number of magnetic ions per unit volume of the sample be equal to N and the number of Zener carriers to xN. The model of double exchange is a model of exchange under the conditions of strong coupling $J_H >> zt, z't'$.

A phenomenological expression for free energy was derived in [16] in the finite temperature limit and for low values of relative sublattice magnetizations.

The density of the thermodynamic potential of the system in the strong coupling limit $J_H \to \infty$ then takes the form

$$\phi(\eta, \rho) = \phi_0 + \phi_\eta + \phi_{int} + \phi_{\mathrm{Coul}}, \quad (2)$$

while for a second-order PT, the second term should be expressed as

$$\phi_\eta = \frac{\alpha}{2}\eta^2 + \frac{\beta}{4}\eta^4 + \frac{\delta}{6}\eta^6 + \frac{\zeta}{8}\eta^8 + \frac{D}{2}(\nabla\eta)^2, \quad (3)$$

where the order parameter η describes the relative magnetization of each sublattice;

$$\alpha = \alpha'(T - T_c)$$

(T_c is the PT temperature without doping).

Expression (3) contains a second-order term with respect to η, positive terms of the

fourth, sixth and eighth orders with respect to η and a gradient term. In addition, expression (3) includes constants

$$\alpha = 2N(1,5k_{\mathrm{B}}T - S^2(zJ + z'J')), \quad (4)$$

$$\beta = 4N(0,45k_{\mathrm{B}}T + 0,034x(zt + z't')); \quad (5)$$

$$\delta = 6N(0,325k_{\mathrm{B}}T + 0,27x(zt + z't')); \quad (6)$$

$$\zeta = 8N(0,06k_{\mathrm{B}}T + 2,21x(zt + z't')), \quad (7)$$

where k_{B} is the Boltzmann constant.

The energy ϕ_{int} in function (2) describes the interaction of the order parameter η with the local charge density ρ:

$$\phi_{int} = -\frac{\sigma_1}{2}\eta^2\rho. \quad (8)$$

This formula is obtained from the temperature-averaged terms of expression (1) describing the double exchange. The interaction energy (8) is written in this case as a local temperature; σ_1 is the constant of this interaction.

The main physical properties of the system are determined by the parameter σ_1, found from the following expression:

$$\overline{\rho}\sigma_1 = \frac{4N}{5}x(zt + z't'). \quad (9)$$

The energy of the Coulomb interaction ϕ_{Coul} is expressed by the integral:

$$\phi_{\mathrm{Coul}} = \frac{\gamma}{2}\int\frac{(\rho(\mathbf{r}) - \overline{\rho})(\rho(\mathbf{r}') - \overline{\rho})}{|\mathbf{r} - \mathbf{r}'|}d\mathbf{r}', \quad (10)$$

where the constant γ is the dielectric constant of the medium.

In the absence of the ϕ_{int} and ϕ_{Coul} terms, a second-order PT is observed at $\alpha = 0$. There is an equilibrium value of the order parameter for $\alpha = 0$. For $\alpha < 0$, the equilibrium value of the order parameter $\eta = 0$, i.e., there is no order which can be determined by the parameter η.

In expressions (9) and (10), $\overline{\rho}$ is the mean 2D surface charge density:

$$\overline{\rho} = \frac{1}{S}\int_S \rho d^2\mathbf{r}, \quad (11)$$

where $\mathbf{r}$ is a two-dimensional vector.

The total free energy Φ, which is expressed as

$$\Phi = \int\phi(\eta,\rho)d^2\mathbf{r}, \quad (12)$$

should be minimized with respect to η and ρ.

Minimizing the energy Φ with respect to the local charge density ρ yields the equality

$$-\frac{\sigma_1}{2}\nabla_{3D}^2\eta^2 = 4\pi\gamma(\rho(\mathbf{r}) - \overline{\rho})\delta(z)d. \quad (13)$$

The thickness of the two-dimensional layer d is introduced to preserve the dimensionality; $\delta(z)$ is the Dirac delta function

Substituting (13) into expression (2), we obtain

$$\phi = \phi_0 + \frac{\alpha}{2}\eta^2 + \frac{\beta}{4}\eta^4 + \frac{\delta}{6}\eta^6 + \frac{\zeta}{8}\eta^8 + \frac{D}{2}(\nabla\eta)^2 - \frac{\sigma_1}{2}\eta^2\overline{\rho} - \frac{\sigma_1^2}{32\pi^2\gamma d^2}\int\frac{\nabla_{2D}\eta^2(\mathbf{r})\nabla_{2D}\eta^2(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}d\mathbf{r}', \quad (14)$$

where **r, r'** are two-dimensional vectors.

The last two terms in expression (14) are negative. The term $-(\sigma_1/2)\eta^2\overline{\rho}$ renormalizes the critical temperature of the phase transition that now becomes dependent on the mean charge density.

The coefficient before the parameter η^2 should now be changed (to $\tilde{\alpha}$ instead of α):

$$\tilde{\alpha} = \alpha - \sigma_1\overline{\rho}. \quad (15)$$

Notably, the presence of the last non-local term in expression (14) leads to instability of the homogeneous state.

Let us introduce dimensionless parameters Λ and ξ as

$$\Lambda = \eta/\eta_0;\ \xi_x = x/a;\ \xi_y = y/a,$$

where $\eta_0^4 = \beta/\zeta$, $a = [D\zeta^{1/2}/(2\beta^{3/2})]^{1/2}\chi$ (χ is a constant, the expressions for the constants β and ζ are given by formulae (5) and (7)).

We chose the value of the constant χ in the interval from 3 to 20. This allowed to vary the size of the region where the spatial distribution of the order parameter was calculated.

Expression (14) then takes the following form:

$$\phi = U_0(\tau\Lambda^2 + \frac{\Lambda^4}{2} + \tilde{\delta}\frac{\Lambda^6}{3} + \frac{\Lambda^8}{4} + \frac{2}{\chi^2}(\nabla\Lambda)^2 - \frac{A}{\chi}\int\frac{\nabla_{2D}\Lambda^2(\xi)\nabla_{2D}\Lambda^2(\xi')}{|\xi - \xi'|}d\xi'). \quad (16)$$

The parameters U_0, τ, χ, A and $\tilde{\delta}$ in this expression are defined as follows:

$$U_0 = \frac{\beta}{2}\eta_0^4 = \frac{\beta^2}{2\zeta}, \quad (17)$$

$$\tau = \frac{\tilde{\alpha}}{\beta\eta_0^2} = \sqrt{\frac{\zeta}{\beta^3}}\alpha'\left(T - T_c - \frac{\sigma_1\bar{\rho}}{\alpha'}\right), \quad (18)$$

$$\chi = a\eta_0\sqrt{\frac{2\beta}{D}}, \quad (19)$$

$$A = \frac{\sigma_1^2}{8\gamma d^2\pi^2\sqrt{2D}\sqrt[4]{\beta\zeta}}, \quad (20)$$

$$\tilde{\delta} = \frac{\delta}{\beta}\eta_0^2 = \frac{2\delta}{\sqrt{\beta\zeta}}. \quad (21)$$

Calculation results and discussion

To find the minimum of the free energy (12), we applied the conjugated gradient method (CGM). We introduced $N \times N$ ($N = 128$) discrete points on a square with a side a and applied periodic boundary conditions. Three parameters A, τ and χ were taken in the calculations.

We analyzed the dependence of the free energy on the parameters A and τ for a fixed value of the constant χ.

Fig. 1,*a* shows the spatial distribution of the order parameter $\Lambda = \Lambda(\xi_x, \xi_y)$ for the parameter values $A = 2.5$, $\tau = 0.6$, $\chi = 10$, and $N = 128$. Phase separation can be observed at these values. For a homogeneous background with an order parameter equal to zero (see the scale on the right), there is a ring with a non-zero order parameter, i.e.,

$$0 < \Lambda(\xi_x, \xi_y) \le 1.8.$$

The free energy is negative in this state ($\Phi < 0$). This means that the spatially inhomogeneous distribution of the order parameter corresponds to minimal free energy. This state is more energetically favorable than the homogeneous one whose free energy is zero ($\Phi = 0$ for $\Lambda(\xi_x, \xi_y) = 0$). Such inhomogeneous states are formed due to charge redistribution. A triple extra-charged layer exists in the region where the parameter $\Lambda(\xi_x, \xi_y)$ is distributed inhomogeneously. The total charge in this layer is equal to zero with high accuracy, $\Delta\rho > 0$ in the center of the stripe and $\Delta\rho < 0$ on each side.

For a fixed value of the parameter $A = 2.5$, the inhomogeneous distribution of the order parameter exists for the values of the parameter τ lying in the range

$$\tau_2 \le \tau \le \tau_3$$

($\tau_2 = -9$ and $\tau_3 = 1.5$).

The free energy is less than zero for the region $\tau \le \tau_1$ ($\tau_1 = 0.8$ for $A = 2.5$). In accordance with expression (18), τ is a linear function of the difference $T - T_c$ and varies depending on

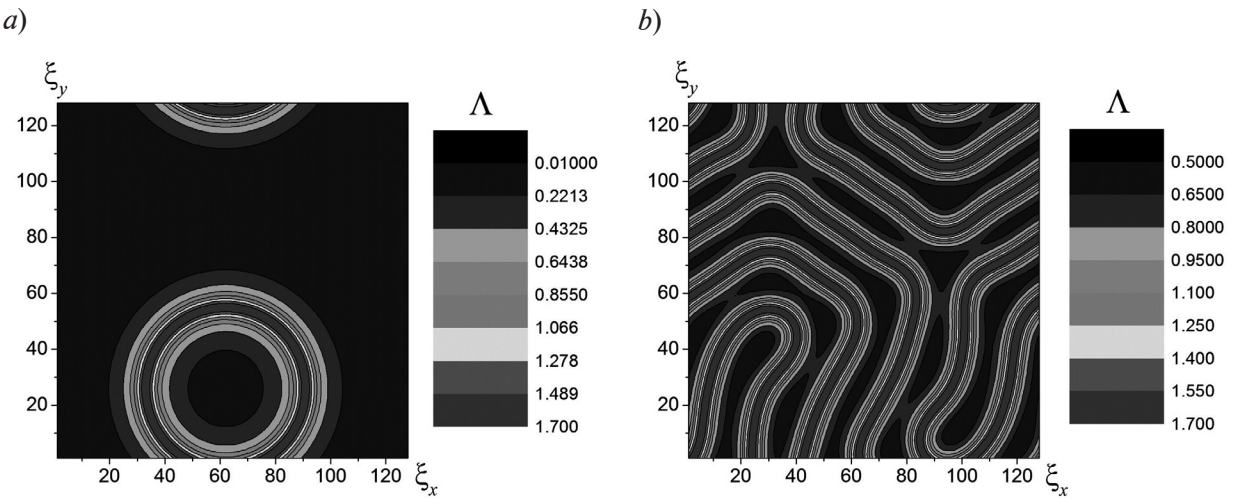

Fig. 1. Calculated distributions of the order parameter $\Lambda(\xi_x, \xi_y)$ in the inhomogeneous state for $\tau = 0.6$ (*a*) and -3.0 (*b*); $A = 2.5$, $\chi = 10$.
The number of discrete points $N^2 = 128^2 = 16384$

$\bar{\rho}$. Here, T_c is the temperature of the phase transition without interaction (i.e., for $\Phi_{int} = 0$); $\bar{\rho}$ is the value of the mean charge proportional to the doping level. The parameter A (see Eq. (20)) depends on the coupling parameter σ_1 and on the strength of the Coulomb interaction. As the latter increases, the parameter A decreases. This shortens the interval of τ values where phase separation can be observed.

Fig. 1,*b* shows the inhomogeneous distribution of the order parameter $\Lambda(\xi_x, \xi_y)$ for the parameter values $A = 2.5$, $\tau = -3, 0$, $\chi = 10$ ($N = 128$). The order parameter varies from $\Lambda_{\min} = 0.5$ to $\Lambda_{\max} = 1.7$ (see the scale on the right). An inhomogeneous distribution of the extra charge exists in the region where the order parameter Λ is distributed inhomogeneously. Calculations show that varying the parameter χ from 3 to 20 (with A = const) does not affect the interval of τ values where inhomogeneous states are formed.

Fig. 1 illustrates how the phase separation landscape changes with the changing τ. Phase separation is observed in the form of stripes or rings for $\tau > 0$ (see Fig. 1,*a*). A stripe with $\Lambda > 0$ appears on the background with a zero order parameter $\Lambda = 0$. The stripes may be straight or have a complex closed form. The number of such stripes decreases with increasing τ, and the rings are compressed. Notably, the order parameter value does not change at the center of these stripes. The shapes of the loops change as the value of τ becomes negative (see Fig. 1,*b*) and with a further decrease in τ: the loops bend more, and the order parameter value in the "background" becomes different from zero ($\Lambda_{\min} = 0.5$ in Fig. 1,*b*). Phase separation becomes shallower with a further decrease in τ (these data are not shown in Fig. 1). The difference between the Λ value inside and outside the "stripes" drops to zero at $\tau = \tau_2$, and a transition to a homogeneous state with $\Lambda = \text{const}$ is observed.

Our model includes the interaction of the order parameter with the charge ($\sigma_1 \neq 0$). In the presence of this interaction, an inhomogeneous phase-separated state with the order parameter varying from $\Lambda_{\min}$ to $\Lambda_{\max}$ (see Fig. 1,*b*) has a minimal negative free energy $\Phi_{\text{inhom}} < 0$.

Let us consider the change of phases observed with decreasing τ and at a constant value of $A = 2.5$. The inhomogeneous phase state appears abruptly (a second-order PT) at $\tau = \tau_1 = 0.8$. Stripes with $\Lambda \neq 0$ grow on the "background" with a zero order parameter $\Lambda = 0$. In these stripes, $\Lambda_{\max} = 1.8$. The number of such stripes increases as τ decreases from τ_1 to 0. Notably, the values $\Lambda_{\max} = 1.8$ and $\Lambda_{\min} = 0$ do not change in this region of τ. The phase-separated state starts to change at $\tau = 0$: $\Lambda_{\max}$ starts to decrease and $\Lambda_{\min}$ to increase. With further decrease of $\tau < 0$, the difference between $\Lambda_{\max}$ and $\Lambda_{\min}$ decreases and $\Lambda_{\max} = \Lambda_{\min} = \Lambda$ when $\tau = \tau_2 = -9$, i.e., a second-order PT from an inhomogeneous to a homogeneous state is observed. In this case, the energy of the inhomogeneous state $\Phi_{inhom} < 0$ is less than the energy of the homogeneous state Φ_{hom} (this state exists in the absence of interaction between the order parameter and the charge, i.e., with $\sigma_1 = 0$) for the entire range of values of the parameter $\tau_2 < \tau < \tau_1$.

The phase diagram of inhomogeneous states in Fig. 2 is shown in the axes $1/A - T$ (T is the temperature), for which $\Delta\Phi < 0$, i.e., the energy of the inhomogeneous

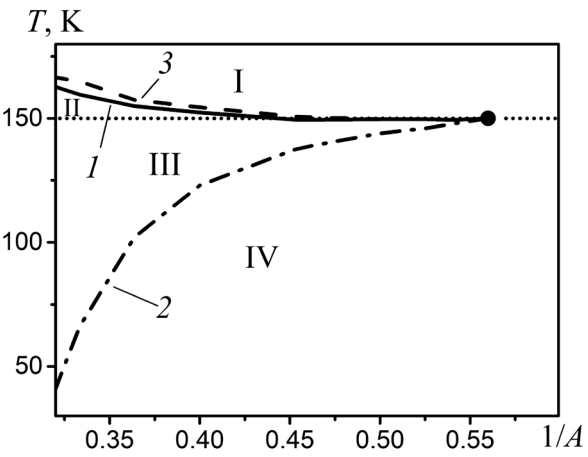

Fig. 2. Phase diagram of inhomogeneous states in $1/A - T$ axes (T is the temperature): region I corresponds to a homogeneous non-magnetic state, regions II and III correspond to a phase-separated state, region IV corresponds to a homogeneous magnetic state; $\Lambda = 0$ in region I; $\Lambda = 0$ and $\Lambda \neq 0$ in region II, $\Lambda \neq 0$ in regions III and IV. The parameters used are given in the text. The boundaries of the regions are $T = f_1(\tau_1)$ (curve *1*), $T = f_2(\tau_2)$ (*2*), $T = f_3(\tau_3)$ (*3*); $1/A = 0.555$ is the final critical point; if $1/A > 0.555$, phase separation is impossible

phase-separated state Φ_{inhom} is less than the energy of the homogeneous state Φ_{hom} at $\tau_2 < \tau < \tau_1$. The difference $\Delta\Phi = \Phi_{hom} - \Phi_{inhom}$. The following parameters were used:

$$T_c + \frac{\sigma_1\bar{\rho}}{\alpha'} = 150 \text{ K}; \quad \frac{\tau}{\alpha'}\sqrt{\frac{\beta^3}{\varsigma}} = 3 \text{ K}.$$

Regions I and IV correspond to homogeneous phases with zero and nonzero order parameters, respectively. Regions II and III correspond to inhomogeneous phases. The value of $1/A$ is directly proportional to the value of the Coulomb interaction γ and inversely proportional to the square σ_1 (see expression (20)). As the parameter A decreases ($1/A$ increases), the interval of τ values narrows, and, consequently, so does the temperature range $T(\tau)$, where the inhomogeneous distribution of the order parameter Λ is observed. Phase separation is impossible below the critical end point $A = 1.8$ ($1/A = 0.555$), which is shown in the phase diagram in Fig. 2. Indeed, with a high value of the Coulomb interaction and a low value of the double exchange energy, the Coulomb energy for charge modulation of the charge becomes so large that it is always greater than the energy gain associated with ordering.

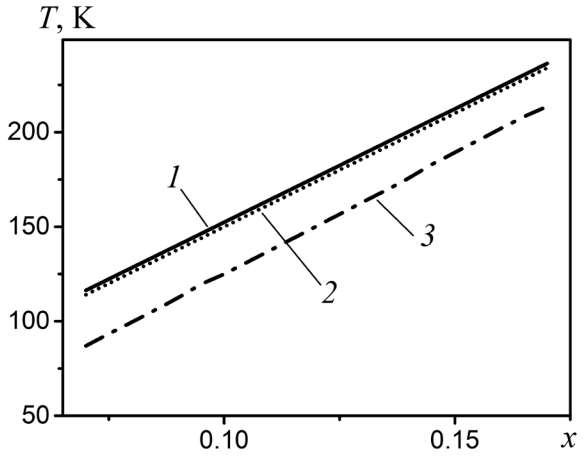

Fig. 3. Phase diagram in $x - T$ axes for different values of the parameter τ (the rest of the parameters used are given in the text). The region between lines *1* and *3* corresponds to the phase-separated state (regions II and III in Fig. 2); $\tau = 0.8$ (*1*), 0.0 (*2*), −9 (*3*).
Line *2* corresponds to the phase transition temperature in the absence of interaction between the charge and the order parameter

The line $T(\tau_3)$ in Fig. 2 indicates the boundary of the region of an inhomogeneous metastable phase. The inhomogeneous state for the interval $T(\tau_1) < T < T(\tau_3)$ corresponds to a local free energy minimum but the free energy is positive in this state ($\Phi > 0$), while the homogeneous state has an energy equal to zero. This metastable state is similar to "superheated liquid".

Fig. 3 shows the phase diagram of the inhomogeneous state in the axes $x - T$ for the values of the parameters

$$A = 2.5, \ \frac{\sigma_1\bar{\rho}}{\alpha' x} = 1200 \text{ K},$$

$$\frac{\tau}{\alpha'}\sqrt{\frac{\beta^3}{\varsigma}} = 3 \text{ K}, \ T_c = 30 \text{ K (with } x = 0).$$

Decreasing A reduces the region where phase separation is observed.

As mentioned in the Introduction, phase separation is observed in manganites and in high-temperature cuprate superconductors. In this paper we have analyzed, as an example, inhomogeneous phases in manganites where a sequence of phase transitions to inhomogeneous states is observed.

The $La_{1-x}Sr_xMnO_3$ system. Let us analyze this system. For instance, the authors of [19] suggest that the available data indicate an electronic phase-separated regime existing only in the phase diagram region $0.10 < x < 0.15$ and near the PT from the ferromagnetic to the paramagnetic state.

Let us consider a region above the temperature of the structural PT from a low-temperature pseudocubic phase to an orthorhombic phase or to a Jahn − Teller distorted orthorhombic phase at higher temperatures. Double exchange begins to play a more fundamental role in this region close to structural instability, where long-range Jahn − Teller distortions are suppressed, in the electronic phase-separated regime.

We assume that the following sequence of phase transitions can be observed in this region of the phase diagram with decreasing temperature for the strontium concentration $x = 0.125$. First, a transition from homogeneous paramagnetic to inhomogeneous state II occurs at $T = 184.5$ K. Next, as the temperature decreases, a transition to inhomogeneous

phase III is observed. Finally, only after this, at 155 K, the system undergoes a transition to a uniform ferromagnetic state. This sequence of PTs is very similar to that discussed in our paper. Additionally, such inhomogeneous states may also appear in cuprates.

Conclusion

We have considered the theory of second-order phase transitions, introducing the Coulomb interaction and charge interaction with the order parameter in addition to the standard expansion of the free energy in powers of the order parameter. We have found the distribution of this parameter and the charge distribution in a 2D plane, which corresponds to a minimum of free energy. We have carried out numerical calculations using the CGM method.

The calculations have confirmed that a region with inhomogeneous distribution of the order parameter and inhomogeneous charge distribution exists between the regions of the phase diagram characterized by constant values of the order parameter. This phase separation can exist in the form of one-dimensional stripes or two-dimensional rings or "snakes". A series of phase transitions have been found. A phase transition from a homogeneous state with a zero order parameter to a phase-separated state with two phases (with zero and nonzero order parameters) first occurred as the temperature decreased. Next, a first-order phase transition to another phase-separated state was observed, where both phases had different nonzero values of the order parameter. A transition to a homogeneous ordered state occurred only with a further decrease in temperature.

We have determined the regions in the "temperature – doping level" parameter space where the phases coexist. We have traced the changes in the type of the phase separation depending on the changes in temperature, doping level of the material, and in the coupling constant.

REFERENCES

[1] **S. Jin, T.H. Tiefel, M. McCormack, et al.,** Thousandfold change in resistivity in magnetoresistive La-Ca-Mn-O films, Science. 264 (5157) (1994) 413–415.

[2] **M.Yu. Kagan, K.I. Kugel',** Inhomogeneous charge distributions and phase separation in manganites, Phys. Usp. 44 (6) (2001) 553–570.

[3] **A.O. Sboychakov, K.I. Kugel, A.L. Rakhmanov,** Jahn-Teller distortions and phase separation in doped manganites, Phys. Rev. B. 74 (1) (2006) 014401(1–13).

[4] **E. Dagotto, T. Hotta, A. Moreo,** Colossal magnetoresistant materials: the key role of phase separation, Phys. Rep. 344 (1–3) (2001) 1–153.

[5] **E.L. Nagaev,** Lanthanum manganites and other giant-magnetoresistance magnetic conductors, Phys. Usp. 39 (8) (1996) 781–805.

[6] **K.M. Shen, F. Ronning, D.H. Lu, et al.,** Nodal quasiparticles and antinodal charge ordering in $Ca_{2-x}Na_xCuO_2Cl_2$, Science. 307 (5711) (2005) 901–904.

[7] **J.M. Tranquada, H. Woo, T.G. Perring, et al.,** Quantum magnetic excitations from stripes in copper oxide superconductors, Nature (London). 429 (6991) (2004) 534–538.

[8] **J. Deisenhofer, D. Braak, H.-A. Krug von Nidda, et al.,** Observation of a Griffith's phase in paramagnetic $La_{1-x}Sr_xMnO_3$, Phys. Rev. Lett. 95 (25) (2005) 257202(1–4).

[9] **E.H. Neto da S., P. Aynajian, A. Frano, et al.,** Ubiquitous interplay between charge ordering and high-temperature superconductivity in cuprates, Science. 343 (6169) (2014) 393–396.

[10] **V.V. Kabanov, R.F. Mamin, T.S. Shaposhnikova,** Localized charge inhomogeneities and phase separation near a second-order phase transition, Sov. Phys. JETP. 108 (2) (2009) 286–291.

[11] **V.B. Shenoy, T. Gupta, H.R. Krishnamurthy, T.V. Ramakrishnan,** Coulomb interactions and nanoscale electronic inhomogeneities in manganites, Phys. Rev. Lett. 98 (9) (2007) 097201(1–4).

[12] **J. Miranda, V.V. Kabanov,** Coulomb frustrated first order phase transition and stripes, Physica C. 468 (4) (2008) 358–361.

[13] **C. Ortix, J. Lorenzana, C. Di Castro,** Coarse grained models in Coulomb frustrated phase separation, J. Phys.: Condens. Matter. 20 (43) (2008) 434229 (1–8).

[14] **R. Jamei, S. Kivelson, B. Spivak,** Universal aspects of Coulomb-frustrated phase separation, Phys. Rev. Lett. 94 (5) (2005) 056805(1–4).

[15] **R.F. Mamin, T.S. Shaposhnikova, V.V. Kabanov,** Phase separation and second-order phase transition in the phenomenological model for a Coulomb-frustrated two-dimensional system, Phys. Rev. B. 2018. Vol. 97 (9) (2018) 094415(1–7).

[16] **P.-G. de Gennes,** Effects of double

exchange in magnetic crystals, Phys. Rev. 118 (1) (1960) 141–154.

[17] **Yu.A. Izyumov, Yu.N. Skryabin,** Double exchange model and the unique properties of the manganites, Phys. Usp. 2001. Vol. 44. No. 2. Pp. 109–134.

[18] **C. Zener,** Interaction between the *d*-shells in the transition metals. II. Ferromagnetic compounds of manganese with perovskite structure, Phys. Rev. 82 (3) (1951) 403–405.

[19] **M. Paraskevopoulos, F. Mayr, J. Hemberger, et al.,** Magnetic properties and the phase diagram of $La_{1-x}Sr_xMnO_3$ for $x \leq 0.2$, J. Phys.: Condens. Matter. 12 (17) (2000) 3993–4011.

Received 23.05.2018, accepted 24.05.2018.

THE AUTHORS

SHAPOSHNIKOVA Tatyana S.
Zavoisky Physical-Technical Institute, FRC KazanSC of RAS
10/7, Sibirsky tract, Kazan, 420029, Russian Federation
t_shap@kfti.knc.ru

MAMIN Rinat F.
Zavoisky Physical-Technical Institute, FRC KazanSC of RAS
10/7, Sibirsky tract, Kazan, 420029, Russian Federation
mamin@kfti.knc.ru

WEAK SOLUTIONS OF THE CROCCO BOUNDARY PROBLEMS

M.R. Petrichenko, D.D. Zaborova, E.V. Kotov, T.A. Musorina

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation

A procedure for designing an approximate solution of the Crocco boundary typical problem has been proposed in the paper. The procedure calls for the change of this initial problem by a nonlinear integral equation. The latter was solved by direct calculation of the integral using the mean-value theorem. . The averaging parameter was eliminated by integrating over the parameter in the (0, 1) interval. Widening the scope of the solution procedure was demonstrated and weak solutions were found. For the classical case, the weak solution was not too different from the Blasius exact one. The approximate value of the Blasius constant turned out to be 1/3 and differed from the exact one (0.33206) by 0.3 %.

Key words: Cauchy problem, integral equation, mean value theorem, group of transformations, solitary wave

Citation: M.R. Petrichenko, D.D. Zaborova, E.V. Kotov, T.A. Musorina, Weak solutions of the Crocco boundary problems, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 19–27. DOI: 10.18721/JPM.11303

Introduction

Crocco's boundary problems are primarily used for hydrodynamic applications, in particular, longitudinal viscous flow past a plate, with unsteady seepage in homogeneous and isotropic (scalar) porous media [1 – 3].

A typical Crocco boundary problem is stated as follows [1]:

$$2\varphi\frac{d^2\varphi}{du^2}+u=0,$$
$$D(\varphi)=(u:0\le u_0<u<1),\varphi\in C^{(2)}(D(\varphi)),$$
$$\left(\frac{d\varphi}{du}\right)_{u=u_0}=\varphi(1)=0, \quad (1)$$
$$\mathrm{Im}(\varphi)=(0,a),a:=\varphi(u_0),$$

and $u_0 = 0$ in the classical Blasius case.

Problem (1) in the given form is widely used in hydrodynamics, where the variable u is interpreted as the longitudinal velocity and the distribution φ as the shear stress [1].

Problem (1) is involved in seepage theory for calculating a solitary flow-rate wave, i.e., for solving the boundary problem for the Boussinesq equation [2, 3]:

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial s}\left(ku\frac{\partial u}{\partial s}\right), \quad (1a)$$

where $u = u(t,s) \le 1$ is the seepage flow depth ($t > 0$, $s > 0$); k is the hydraulic conductivity.

In a particular case,

$u(0,s) - 1 = u(t,0) = 0$.

In the general case,

$$k=k(u),\ c=-k(u)\left(\frac{\partial u}{\partial s}\right)^c,$$

where c is the seepage rate.

In the classical Boussinesq case, $k = 1$, $c = 1$. The Boussinesq equation then takes the form

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial s}\left(u\frac{\partial u}{\partial s}\right). \quad (1b)$$

Finally, the Crocco equation is used in problems on jet motion of viscous fluid (free convection in heated channels, free submerged and near-wall jets, etc.) [4, 5].

In contrast to the "natural" statement of applied boundary problems, problem (1) is

convenient because it allows finding an injective mapping of the compact set $(u_0,1)$ into a compact set

$$(0,a), \varphi : (u_0,1) \to (0,a).$$

More specifically, we suppose that each branch of the solution of boundary problem (1) is a 2-diffeomorphism $\varphi : (u_0,1) \to (0,a)$.

The solutions of boundary problem (1) are given in [6 – 32]. These studies fall into two classes:

the first class uses analytical methods, including solutions obtained in the form of power and splitting (flat) series;

studies of the second class primarily rely on numerical solutions.

Analytical papers (belonging to the first class) include those using methods of the theory of Lie transformation groups and expansions into power series and enveloping series.

For example, Boussinesq equation (1b) admits a linear transformation

$$z = \alpha t \pm \sqrt{\alpha} s,$$

and, consequently, there exists a solution of the Boussinesq equation in the form of a solitary flow-rate wave:

$$z = u + c_1 - c_2 \ln(u + c_2).$$

The case $c_2 = 0$ in this solution corresponds to a centered flow-rate wave propagating with a velocity of $\pm\alpha^{1/2}$ either upstream or downstream.

A particular group of studies [26 – 31] offer analytical tools for constructing solutions of ordinary differential equations (ODEs) of the given type near a singular point. These studies emphasize that the properties of an analytic function mainly depend on its singularities, which cannot be examined in the real interval. Moving onto the complex plane automatically means constructing a mapping onto the Riemann surface of the solution [31].

"Exact" solutions of the boundary problem for the Crocco equation are also obtained by using power series with respect to u. However, the Tauberian theorems are unknown for these solutions: namely, the series for the function $\varphi(u)$ turn out to be "poorly" convergent for $u \to 1 - 0$. For example, the divergence of the series for $\varphi(u)$ at the outer edge of the boundary layer $(u \to 1 - 0)$ and bifurcation of the solution in the outer (i.e., jet) part of the boundary layer was established in [32].

We have omitted to discuss the so-called integral methods, where, instead of the equations for the distribution density, ODEs are solved for the actual distributions (integral relations), in the above overview. These methods are ideologically closer to completely different types of methods such as direct or variational.

Thus, flat series for analytical solutions are currently the only alternative to numerical methods for solving Blasius, Chazy, and Crocco equations.

The goal of this study has been to construct an approximate solution of a typical Crocco boundary problem using the averaging procedure.

Constructing the solution of the problem

In this study, we use the method of constructing an approximate solution of boundary problem (1), based on replacing this initial boundary problem by an integral equation, and subsequently introducing an artificial parameter and averaging over this parameter. In other words, a distribution (functional) with a density coinciding with the approximate solution is used instead of an exact solution.

This technique consists in the following. Let

$$f(u) \in C^{(1)}(0,1),\ f(u) \geq 0.$$

The Crocco boundary problem for the interval (0, 1) has the form

$$2\varphi \frac{d^2\varphi}{du^2} + f(u) = 0, \varphi'(0) = \varphi(1) = 0$$

and admits a formal order reduction:

$$2\frac{d\varphi}{du} = -\int_0^u \frac{f(v)dv}{\varphi(v)}.$$

For a positive branch of the solution $\varphi := \varphi^+$, the function $1 / \varphi(u)$ is a positive monotonically decreasing distribution mapping the interval $u \in (u_0,\ 1)$ onto the interval $\varphi \in (0,a)$, where $a := \varphi(0)$.

Then, according to the Bonnet theorem,

$$\frac{d\varphi_\theta^2}{du} = -\int_{\theta u}^u f(v)dv, \qquad (\#)$$

where $0 < \theta < 1$ is the parameter (proper fraction).

Let $u = 1$ in this formula. In this case, with $u \to 1 - 0$,

$$d\varphi_\theta{}^2 / du = O(1).$$

This means that the equalities

$$\varphi_\theta(u) = O(\varepsilon^m), \quad d\varphi_\theta / du = O(\varepsilon^{-m}),$$
$$m = n,$$

where m, n are positive parameters, have to hold true.

Integrating equality (#) one more time, we obtain:

$$\varphi_\theta^2(u) = \int_u^1 dv \int_{\theta v}^{v} f(t)dt.$$

This equality is actually the approximate θ solution of the Crocco equation (indicated by the subscript θ). This solution depends continuously on the fraction (parameter) θ, and, evidently,

$$\varphi_1^2(u) = 0, \varphi_0^2(u) = \int_u^1 dv \int_0^u f(t)dt =$$
$$= \int_0^1 (1-t)f(t)dt - \int_0^u (u-t)f(t)dt > \varphi_\theta^2(u) > 0,$$
$$\forall \theta \in (0,1),$$

while, generally speaking, $\partial\varphi / \partial\theta < 0$.

The parameter θ can be eliminated, for example, by averaging the derivative with respect to it:

$$\frac{d\varphi^2}{du} := \int_0^1 \frac{d\varphi_\theta^2}{du} d\theta,$$

which leads to the expression

$$\frac{d\varphi^2}{du} = -1 / u \int_0^u vf(v)dv.$$

Finally, we can write the following approximate solution of boundary problem (1):

$$\varphi^2(u) = \int_0^1 vf(v) \ln \frac{1}{v} dv - \int_0^u vf(v) \ln \frac{u}{v} dv. \quad (2)$$

The result obtained does not depend on the order of integration with respect to the parameter θ and to the argument u. Let us call solution (2) a weak θ solution.

Properties of solutions of boundary problem (1)

We are going to list the properties of the solutions of boundary problem (1) in this section (the proofs of these properties are omitted).

1. The boundary conditions in problem (1) can be replaced by one-point (Cauchy) conditions:

$$\left(\frac{d\varphi}{du}\right)_{u=u_0} = \varphi(0) - a = 0, \quad (3)$$

with the parameter a in initial conditions (3) chosen so that $\varphi(1) = 0$. Imposing these conditions is justified by the continuous dependence of φ on the parameter a.

2. There are two branches of the solution of boundary problem (1) and, respectively, of one-point boundary problem (3):

$$\varphi^+(u) \text{ and } \varphi^-(u)$$

(Fig. 1). These branches are related as follows:

$$\varphi^+(u) + \varphi^-(u) = 0, \ 0 < u < 1,$$

and

$$0 \le \varphi^+(u) \le a, \frac{d\varphi^+}{du} < 0, \frac{d^2\varphi^+}{du^2} < 0;$$
$$a \le \varphi^-(u) \le 0, \frac{d\varphi^-}{du} > 0, \frac{d^2\varphi^-}{du^2} > 0.$$

Boundary problem (1) is typical, since, in particular, the homogeneous Crocco boundary problem is reduced to it. Let $u_0 = 0$, and then instead of representation (1) we consider the homogeneous Crocco boundary problem:

$$2\varphi \frac{d^2\varphi}{du^2} + u = 0, \quad (1c)$$
$$\varphi(0) = \varphi(1) = 0.$$

The solution of homogeneous boundary problem (1c) consists of two branches:

$$0 \le \varphi^+(u) \text{ and } \varphi^-(u) \le 0$$

(negative and positive), such that

$$\varphi^+(u) + \varphi^-(u) = 0$$

for each u value from the interval $0 < u < 1$.

There exists a value of $u = u^*$ $(0 < u^* < 1)$, such that $d\varphi^\pm / du^* = 0$ (according to Rolle's theorem). Therefore, each of the branches $\varphi^\pm(u)$ of the solution of homogeneous prob-

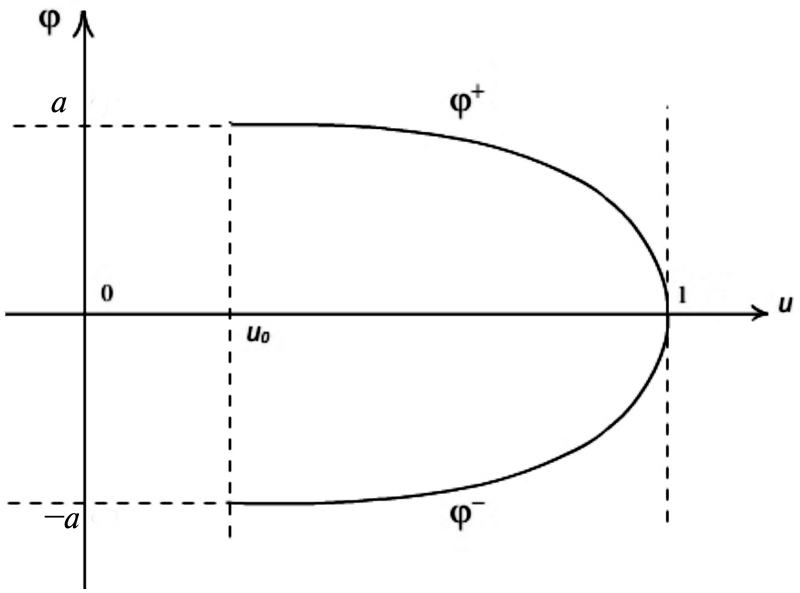

Fig. 1. Solution of a typical Crocco boundary problem: positive ($\varphi^+(u)$) and negative ($\varphi^-(u)$) monotonic branches; the vertical dashed lines indicate the boundaries of the interval

lem (1) is decomposed into two solutions of the typical Crocco boundary problem (1):

$$\varphi_l^{\pm}(u),\ D(\varphi_l^{\pm}) = (0, u^*);$$

$$\varphi_r^{\pm}(u),\ D(\varphi_r^{\pm}) = (u^*, 1).$$

In this case, the solutions $\varphi(u) = \varphi_{l,r}^{\pm}(u)$ of positive and negative typical boundary problems are mapped continuously and smoothly at the point $u = u^*$ (Fig. 2):

$$\varphi_l^{\pm}(0) = \left(\frac{d\varphi_l^{\pm}}{du}\right)_{u=u^*-0} =$$

$$= \left(\frac{d\varphi_r^{\pm}}{du}\right)_{u=u^*+0} = \varphi_r^{\pm}(1) = 0.$$

The branches of the solutions of the homogeneous boundary problem are decomposed into solutions of typical boundary problems

$$\varphi_l^{\pm}(u^* - 0) - \varphi_r^{\pm}(u^* + 0) = 0,$$
$$\left(\frac{d^2\varphi^+}{du^2}\right)_{u=u^*} < 0, \left(\frac{d^2\varphi^-}{du^2}\right)_{u=u^*} > 0. \qquad (\#\#)$$

3. Solution of boundary problem (1) – (3) satisfies the identity

$$\int_{u_0}^{1}\left(\frac{d\varphi}{du}\right)^2 du = \frac{1-u_0^2}{4}. \qquad (4)$$

4. Solution of boundary problem (1) – (3) is equivalent to the problem of the minimum positive functional (distribution):

$$F(\varphi) = (1/2)\int_{u_0}^{1}\left(\left(\frac{d\varphi}{du}\right)^2 + u\ln\frac{a}{\varphi}\right)du > 0.$$

In other words, the condition $dF \le \delta F$ is satisfied along the extremals of the functional $F(\varphi)$, where dF is the variation along the characteristic (the trajectory of the solution), and δF is the variation along the admissible (virtual) trajectory. The basis for the proof of property 4 is that the necessary minimum condition $F(\varphi)$ coincides with the Crocco equation, and the sufficiency of the condition is guaranteed by the convexity of the Lagrangian density $F(\varphi)$.

5. The property of the solution for $u_0 = 0$. In this case, boundary problem (1) is equivalent to the following nonlinear integral equation:

$$\frac{d\varphi}{du} = -\frac{1}{2}\int_0^u \frac{vdv}{\varphi(v)},$$

$$\varphi(u) = \frac{1}{2}\int_u^1 dv\int_0^v \frac{tdt}{\varphi(t)} =$$

$$= \frac{1}{2}\left(\int_0^1 \frac{(1-t)tdt}{\varphi(t)} - \int_0^u \frac{(u-t)tdt}{\varphi(t)}\right).$$

An iterative process can be used to solve this integral equation.

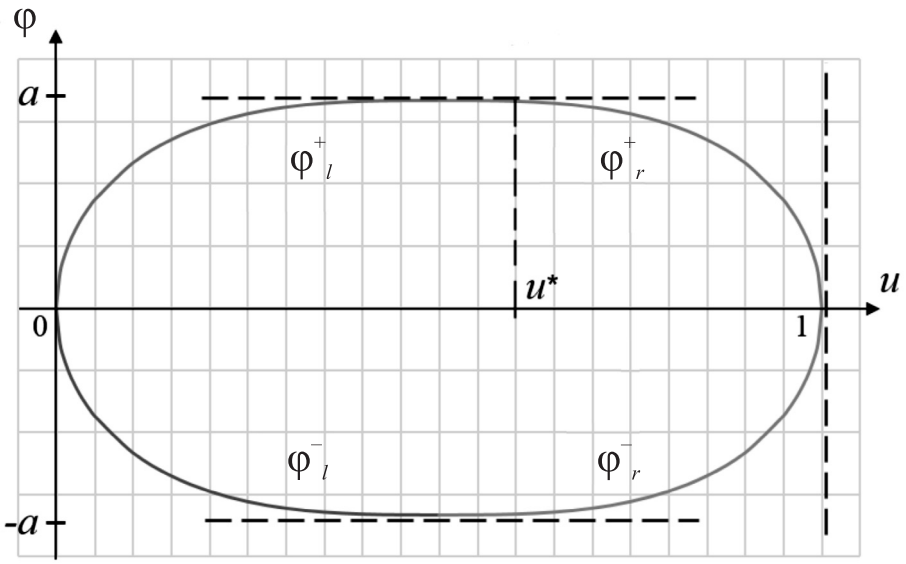

Fig. 2 . The $\varphi^{\pm}(u)$ dependences illustrating how the branches of the solution of a homogeneous boundary problem are decomposed into solutions (##) of typical boundary problems; u^* corresponds to the maxima of the function $|\varphi|$

Let the subscript denote the iteration number, and then the process of solution is expressed as

$$\frac{d\varphi_s}{du} = -\int_0^u \frac{vdv}{\varphi_{s-1}(v)},$$

$$\varphi_s(u) = \frac{1}{2}\int_u^1 dv \int_0^v \frac{tdt}{\varphi_{s-1}(t)}.$$

If $1/\varphi \in L_1(0,1)$, then $\varphi \in C^{(1)}(0,1)$. Assuming that the sequence of iterations of the function $1/\varphi_s$ forms a Cauchy sequence, $\varphi_s \to \varphi$ almost everywhere on the interval $0 < u < 1$, since the domain $L_1(0, 1)$ is complete.

6. The solution allows to formulate a corollary to the mean value theorem. Since $1/\varphi(u)$ is a monotonically increasing distribution, then, according to the Bonnet mean theorem, the equalities

$$2\varphi(u)\frac{d\varphi}{du} = -\frac{u^2}{2}(1-\theta^2),$$
$$\varphi_\theta^2(u) = \frac{1}{6}(1-u^3)(1-\theta^2), \qquad (1c)$$

where θ is a proper fraction ($0 < \theta < 1$), hold true.

The final expression for the approximate solution has the form

$$\varphi_\theta(u) = (1/\sqrt{6})\sqrt{(1-u^3)(1-\theta^2)}.$$

The mean square value of $\varphi_\theta(u)$, i.e., the θ approximation of the solution, is determined by the equality

$$\varphi^2(u) = \int_0^1 \varphi_\theta^2(u)d\theta = (1/9)(1-u^3). \qquad (5)$$

It follows from here that $\varphi(u) = (1/3)\sqrt{1-u^3}$, and this approximates the Blasius exact solution, especially for small u values. For example, the value of the Blasius constant is $a = 1/3$. Its exact value, recently calculated by Varin, is [29, 30]:

$$a = 0{,}33205733621519$$
$$629893718006201058$$
$$296654709356141267$$
$$981810047564019872$$
$$417401806440507049$$
$$0731855146368... .$$

The rational value of the constant differs from the reduced irrational one by less than 0.3 %.

The quantity $\mathfrak{D} := \int_0^1 \varphi du$ in problems with physical content is dissipation in the segment

(0, 1). In this case, its value is

$$\mathcal{D} = (1/3)\int_0^1 \sqrt{1-u^3}\,du = \frac{\sqrt{\pi}}{18}\frac{\Gamma(1/3)}{\Gamma(11/6)} \approx 0,27.$$

Apparently, the distribution $\varphi(u)$ is responsible for dissipation in the neighborhood of the point $u = 0$.

7. Let us formulate the general definition of the norm of θ approximation

$$\varphi_r(u) = \frac{1}{\sqrt{6}}(1-u^3)\cdot\left(\int_0^1 (1-\theta^2)^{r/2}d\theta\right)^{1/r} =$$

$$= \frac{1-u^3}{\sqrt{6}}\left(\frac{\sqrt{\pi}\Gamma(r/2+1)}{2\Gamma(r/2+3/2)}\right)^{1/r},$$

where $r > 0$ is any positive real number.

We are going to omit the subscript from now on, and the norm should be clear from the context. For example, $r = 2$ was adopted in the previous subsection, and then we obtain that

$$\varphi(u) := \varphi_2(u) = \frac{1-u^3}{\sqrt{6}}\sqrt{\frac{\sqrt{\pi}}{2\Gamma(5/2)}} =$$

$$= \frac{1}{\sqrt{6}}\sqrt{\frac{2}{3}}(1-u^3) = (1/3)(1-u^3).$$

Next, the Cauchy – Hölder inequality

$$\varphi_r(u) \le \varphi_{r+\alpha}(u), \forall\alpha > 0,$$

holds uniformly with respect to $0 < u < 1$, and the sequence of norms does not decrease as the index r increases from 0 to ∞, or, more precisely,

$$(1/\sqrt{6})\exp(-c/2 - 0,02) < \|a\|_r < 1/\sqrt{6},$$

where c is the Mascheroni constant.

8. The first generalized property of the solution. Let the Crocco boundary problem (1) have the following form

$$2\varphi\frac{d^2\varphi}{du^2} + u^m = 0,$$
$$\varphi'(0) = \varphi(0) = 0.$$

This representation of the Crocco equation follows from the Boussinesq equation if

$$k = k(h) = k_0(h/H)^{m-1}.$$

The θ approximation of the solution of boundary problem (1a) then takes the form

$$\varphi_\theta^2(u) = \frac{(1-\theta^{b+1})(1-u^{m+2})}{(m+1)(m+2)}, \tag{5a}$$

and, if we apply mean-square averaging over θ, the weak solution has the form:

$$\varphi(u) = \frac{\sqrt{1-u^{m+2}}}{m+2},$$
$$a = \frac{1}{m+2}. \tag{5b}$$

Identity (4) is written as follows:

$$\int_0^1 \left(\frac{d\varphi}{du}\right)^2 du = \frac{1}{2(m+1)}. \tag{4a}$$

The condition for a minimum quadratic functional practically does not change and is expressed as

$$F(\varphi) = (1/2)\int_0^1 \left(\left(\frac{d\varphi}{du}\right)^2 + u^m \ln\frac{a}{\varphi}\right)du \to \inf \ge 0.$$

In this case, dissipation follows the expression

$$\mathcal{D} = \frac{1}{m+2}\int_0^1 \sqrt{1-u^{m+2}}\,du =$$

$$= \frac{1}{(m+2)^2}\frac{\Gamma\left(\frac{1}{m+2}\right)\Gamma\left(\frac{3}{2}\right)}{\Gamma\left(\frac{3m+8}{2(m+2)}\right)} = \tag{6}$$

$$= O\left(\frac{1}{m+2}\right), m >> 1,$$

and decreases with increasing m.

By virtue of identity (4a), the following expression holds true:

$$\bar{\varphi}(u) := \frac{\varphi(u)}{a} = \sqrt{1-u^{m+2}},$$

and it can be seen from this that the degree to which the profile

$$\bar{\varphi}(u) := \frac{\varphi(u)}{a}$$

is complete is increased with increasing parameter m.

Giving a physical meaning to the solution, let us assume that $\varphi = \varphi(u)$ is the friction in the boundary layer. The friction on the surface of a plate with longitudinal viscous flow over it is then $a = \varphi(0)$, decreases monotonically and

persists from the near-wall to the jet part of the layer:

$$\text{with } m \to \infty, \varphi(u) \to \varphi(0) = 0,$$
$$0 < u < 1.$$

9. The second generalized property of the solution. Let the Boussinesq equation and the boundary conditions for it have the form

$$\frac{\partial u^a}{\partial t} = \frac{\partial}{\partial s}\left(u^b \left(\frac{\partial u}{\partial s} \right)^c \right),$$

where a, b, c are real parameters;

$$D(u) = (t > 0,\ s > 0),$$
$$u(0,\ s) - 1 = u(t,\ 0) = 0.$$

Then the corresponding Crocco transformation converts this equation to the form

$$u^b = \frac{a}{c+1}\varphi(u)\left(-\frac{d}{du}\left(u^{1-a}\frac{d\varphi}{du} \right) \right)^c, \qquad (7)$$

and the boundary conditions are imposed as follows:

$$\varphi(1) = \varphi'(0) = 0. \qquad (8)$$

In this case, the θ solution of boundary problem (7), (8) has the form

$$\varphi_\theta(u) = \frac{(c+1)c^{\frac{c}{c+1}}}{a^{\frac{1}{c+1}}(b+c)^{\frac{c}{c+1}}(b+ac)^{\frac{c}{c+1}}} \times \qquad (9)$$
$$\times \{(1-\theta^{b/c+1})(1-u^{a+b/c+1})\}^{\frac{c}{c+1}}.$$

Note 1. If $a = b = 1$ in the particular case, the expression known as the Khristianianovich seepage model for flat seepage flow follows from (9):

$$\varphi_\theta(u) = \frac{\left(\left(1-\theta^{\frac{c+1}{c}} \right)\left(1-u^{\frac{2c+1}{c}} \right) \right)^{\frac{c}{c+1}}}{(c+1)^{\frac{c}{c+1}}}. \qquad (9a)$$

Then

$$a_\theta = \left(\frac{1-\theta^{\frac{c+1}{c}}}{c+1} \right)^{\frac{c}{c+1}},$$

and, furthermore,

$$a^r = \frac{c}{(c+1)^{\frac{c(r+1)+1}{c}}} \frac{\Gamma\left(\frac{c}{c+1} \right)\cdot\Gamma\left(\frac{c(r+1)+1}{c+1} \right)}{\Gamma\left(\frac{c(r+2)+1}{c+1} \right)},$$

with $r \geq 0$.

Obviously,

$$\bar{\varphi}(u) := \frac{\varphi(u)}{a} = \left(1 - u^{\frac{2c+1}{c}} \right)^{\frac{c}{c+1}},$$

and the profile of the dimensionless distribution $\bar{\varphi}(u)$ is completed with decreasing parameter c

For example, if $c = 1/2$, then

$$\bar{\varphi}(u) = (1-u^4)^{1/3};$$

and if $c = 1/3$, then

$$\bar{\varphi}(u) = (1-u^5)^{1/4}.$$

Note 2. Let $a = 1$, b, c be the free parameters. Then, by virtue of θ solution (9), we obtain the expressions

$$\varphi_\theta(u) = \frac{(c+1)c^{\frac{c}{c+1}}}{(b+c)^{\frac{2c}{c+1}}}\{(1-\theta^{b/c+1})(1-u^{b/c+2})\}^{\frac{c}{c+1}},$$

$$\bar{\varphi}(u) = (1-u^{b/c+2})^{\frac{c}{c+1}},$$

and the degree to which the profile $\bar{\varphi}(u)$ is completed is increased with increasing parameter b.

For example, if $c = 1/2$, then

$$\bar{\varphi}(u) = (1-u^{2(b+1)})^{1/3} \underset{b\to\infty}{\to} 0,\ 0 < u < 1.$$

Conclusions

As a result of the study we have conducted, we have established the following:

A typical Crocco boundary problem admits a positive and a negative branch of the solution (φ^+ and φ^-), such that

$$\varphi^+(u) + \varphi^-(u) = 0.$$

A homogeneous Crocco boundary problem is reduced to two typical Crocco boundary problems, conjugate at the critical point $u = u^*$, such that

$$0 < u_0 < u^* < 1,$$

$$\left(\frac{d\varphi}{du}\right)_{u=u^*} = \varphi(u^* - 0) - \varphi(u^* + 0) = 0.$$

A typical Crocco boundary problem is equivalent to a nonlinear integral equation. The latter is solved by direct calculation of the integral using the second mean-value theorem. The averaging parameter is excluded by integration over the parameter in the interval (0, 1).

Extensions of the proposed solution method have been discussed in this paper. The weak θ solution is insignificantly different from the exact one for the classical case $a = b = c = 1$. The approximate value of the Blasius constant $a = \varphi(0)$ turns out to be 1/3. In this case, the exact value of $\varphi(0) = 0,33206$.

REFERENCES

[1] **L. Crocco,** Sulla strato limite laminare nei gas lungo una lamina plana, Rend. Math. Appl., Ser. 5. 21 (2) (1941) 138–152.

[2] **P.Ya. Polubarinova-Kochina,** Teoriya dvizheniya gruntovykh vod [Groundwater movement theory], 2nd ed., Nauka, Moscow, 1977.

[3] **M.R. Petrichenko,** Predelnaya zadacha Bussineska i yeye gidravlicheskoye prilozheniye [The Boussinesq limit problem and its hydraulic application], In the collection of scientific papers "The Scientific Review of Physics, Mathematics and Engineering Sciences for the 21th Century", "Prospero" publ. house, Moscow (2015) 3–7.

[4] **M.R. Petrichenko,** Predelnyye zadachi Krokko v teorii struy vyazkoy zhidkosti [The Crocco limit problems in the viscous-fluid jet theory], In the collection of scientific papers "The Scientific Review of Physics, Mathematics and Engineering Sciences for the 21th Century", "Prospero" publ. house, Moscow, (2016) 10–14.

[5] **D. Nemova, E. Reich, S. Subbotina, et al.,** Heat and mass transfer in a vertical channel under heat-gravitational convection conditions, Experimental Fluid Mechanics Proceedings of the International Conference, Prague (2015) 604–616.

[6] **M. Akdi, M.B. Sedra,** Numerical solution of the Blasius problem, The African Review of Physics 9 (0022) (2014) 165–168.

[7] **F.M. Allan, M.I. Syam,** On the analytic solutions of the nonhomogeneous Blasius problem, Journal of Computational and Applied Mathematics. 182 (2) (2005) 362–371.

[8] **H. Aminikhah,** Analytical approximation to the solution of nonlinear Blasius viscous flow equation by LTNHPM, International Scholarly Research Network, ISRN Mathematical Analysis. 2012 (2012) ID 957473 (10 p).

[9] **H. Aminikhah, S. Kazemi,** Numerical solution of the Blasius viscous flow problem by quartic B-spline method, Hindawi Publishing Corporation, International Journal of Engineering Mathematics. 4 (2016) ID 9014354 (6 p), http://dx.doi.org/10.1155/2016/9014354..

[10] **G. Andrzejczak, M. Nockowska-Rosiak, B. Przeradzk,** A note on Blasius type boundary value problems, Opuscula Math. 33 (1) (2013) 5–17.

[11] **R.C. Bataller,** Numerical comparisons of Blasius and Sakiadis flows, Matematika. 26 (2) (2010) 187–196.

[12] **R. Cortell,** Numerical solutions of the classical Blasius flat-plate problem, Applied Mathematics and Computation. 170 (1) (2005) 706–710.

[13] **C.-W. Chang, C.-S. Liu,** The Lie-group shooting method for boundary layer equations in fluid mechanics, Proceedings of the Conference of Global Chinese Scholars on Hydrodynamics, Beijing, 2006 –07, 103–108.

[14] **T. Fang, C.F. Lee,** A moving-wall boundary layer flow of as slightly rarefied gas free stream over a moving flat plate, Applied Mathematics Letters. 18 (5) (2005) 487–495.

[15] **Tiegang Fang, Wei Liang, Chia-fon F. Lee,** A new solution branch for the Blasius – A shrinking equation sheet problem, Computers and Mathematics with Applications. 56 (12) (2008) 3088–3095.

[16] **R. Fazio,** Blasius problem and Falkner – Skan model: Tцpfer's algorithm and its extension, Computers & Fluids. 73 (15 March) (2013) 202–209.

[17] **J. Goh, A.A. Majid, A.I.M. Ismail,** A quatic B-spline for second-order singular boundary value problems, Computers and Mathematics with Applications. 64 (12) (2012) 115–120.

[18] **J.-H. He,** A simple perturbation approach to Blasius equation, Applied Mathematics and Computation. 140 (2003) 217–222.

[19] **Chein-Shan Liu, Jiang-Ren Chang,** The Lie-groups shooting method for multiple-solutions of Falkner – Skan equation under suction-injection conditions, International Journal of Non-Linear Mechanics. 43 (2008) 844–851.

[20] **C.-S. Liu,** Cone of non-linear dynamical system and group preserving schemes, International Journal of Non-Linear Mechanics. 36 (7) (2001) 1047–1068.

[21] **C.-S. Liu,** The Lie-group shooting method for non-linear two-point boundary value problems exhibiting multiple solutions, Comput. Model. Eng.

Sci. 13 (2) (2006) 149–163.

[22] **Ding Xu, Xin Guo.** Fixed point analytical method for nonlinear differential equations, Journal of Computational and Nonlinear Dynamics. 8 (1) (2013) 011005 (9 p).

[23] **V.P. Varin,** Flat expansions and their applications, Moscow, KIAM Preprint (23) (2014). URL: http://keldysh.ru/papers/2014/prep2014_23.eng.pdf.

[24] **V.P. Varin,** Flat expansions and their applications, Computational Mathematics and Mathematical Physics. 55 (5) (2015) 797–810.

[25] **V.P. Varin,** Ploskiy asimptoticheskiy ryad Blaziusa [The Blasius's flat asymptotic expansion], In: Mathematics Forum Ser. "Sciences summary, The South of Russia" (2015) 34–47.

[26] **V.P. Varin,** Special solutions to Chazy equation, Moscow, KIAM Preprint (43) (2015). http://keldysh.ru/papers/2015/prep2015_43_eng.pdf.

[27] **V.P. Varin,** Special solutions to Chazy equation, Computational Mathematics and Mathematical Physics. 57 (2) (2017) 211–235.

[28] **V.P. Varin,** Integrirovaniye ODU na Rimanovykh poverkhnostyakh kak vychislitelnyy instrument, [ODE integration on the Riemann surfaces as numerical instrument], In: Teoreticheskiye osnovy konstruirovaniya chislennykh algoritmov i resheniye zadach matematicheskoy fiziki [Theoretical foundation of synthesizing the numerical algorithms and the solution of math-physical problems], Tezisy dokladov XXI Vserossiyskoy konferentsii i Molodezhnoy shkoly-konferentsii, posvyashchennoy pamyati K.I. Babenko [Sci. abstracts of the 21th All-Russian Conf. and the Youth School-Conf. Dedicated to the Memory of K.I. Babenko], KIAM (2016) 73–74.

[29] **Faiz Ahmad,** Application of Crocco – Wang equation to the Blasius problem, Electronic Journal "Technical Acoustics", http://www.ejta.org, 2007, 2.

Received 23.05.2018, accepted 04.06.2018.

THE AUTHORS

PETRICHENKO Mikhail R.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
fonpetrich@mail.ru

ZABOROVA Dariya D.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
zaborova-dasha@mail.ru

KOTOV Eugeniy V.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
ekotov.cfd@gmail.com

MUSORINA Tatiana A.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
flamingo-93@mail.ru

A STUDY OF THERMAL REGIME IN THE HIGH-POWER LED ARRAYS

A.V. Aladov[1], I.V. Belov[2], V.P. Valyukhov[3], A.L. Zakgeim[1], A.E. Chernyakov[1]

[1] Submicron Heterostructures for Microelectronics, Research & Engineering Center of RAS, St. Petersburg, Russian Federation;

[2] Jönköping University, School of Engineering, Jönköping, Sweden;

[3] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation

Thermal resistance and temperature distribution for high-power AlGaInN LED chip-on-board arrays were measured by different methods and tools. The p—n junction temperature was determined through measuring a temperature-dependent forward voltage drop on the p—n junction, at a low measuring current after applying a high heating current. Furthermore, the infrared thermal imaging technique was employed to obtain the temperature map for the test object. A steady-state 3D computational model of the experimental setup was created including temperature-dependent power dissipation in the LED chips. Simulations of the heat transfer in the LED array were performed to further investigate temperature gradients observed in the measurements. Simulations revealed possible thermal deformation of the assembly as the reason for the hot spot formation. The bending of the assembly was confirmed by surface curvature measurements.

Key words: LED, LED matrix, thermal resistance, infrared thermography, thermal interface, CFD model

Citation: A.V. Aladov, I.V. Belov, V.P. Valyukhov, A.L. Zakgeim, A.E. Chernyakov, A study of thermal regime in the high-power LED arrays, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 28—37. DOI: 10.18721/JPM.11304

Introduction

Recent research into development and applications of light-emitting diode (LED) sources for general lighting has involved increasing operating currents and packing densities of light-emitting chips in the arrays to provide ever higher output powers [1, 2]. Increased power and complex construction of LED sources entail paying more attention to thermal processes both in individual LEDs and in the array as a whole. It is insufficient to estimate the total thermal resistance in this case: the temperature field distribution over the area (thermal mapping) has to be analyzed.

Since temperature considerably affects the internal quantum efficiency (IQE), its distribution becomes a factor determining the overall output characteristics of LED arrays (optical power and efficiency). Accordingly, exploring the non-uniform temperature distribution over the array as a function of current is particularly important for modern high-power LED arrays.

The goal of this study has been detailed analysis of thermal resistance and temperature distribution in high-power white LED arrays.

Experimental procedure

We have experimentally studied the thermal processes in LED arrays manufactured by the chip-on-board technology [3] using high-power face-up AlInGaN LED chips [4]. Finding the exact temperature gradients on the surface of a LED array, associated with non-uniform heat generation and dissipation from each chip in this array, is of particular

interest in this context. To this end, we have carried out detailed simulation of thermal and current distribution in a real LED array, and determined the temperature of emitting chips located at different points of the array by direct and indirect experimental methods.

The direct method for estimating the temperature of LED chips is based on high-resolution infrared thermography using a SVIT IR thermal imaging camera [5, 6]. Thermal resistance was measured by relaxation of the temperature-dependent parameter (forward voltage) with a T3Ster thermal transient tester [7, 8].

This section describes the experimental samples and the experimental methods used for estimating the thermal characteristics, including measurements of thermal resistances, IR mapping, and measuring the thermal deformation of the LED array in operating mode.

Experimental samples. We have studied high-power arrays based on commercial ES-CABLV45P chips by EpiStar. The emitting chips have a face-up configuration [9], where an epitaxial AlInGaN heterostructure is stored on a sapphire substrate with low thermal conductivity (≈ 0.34 W/cm·K). Both contacts are located on the face, and the light is transmitted through a semitransparent *p*-contact. The emitting chips have a complex "branching" topology of the electrodes to achieve a uniform current distribution at an operating current of 400 mA [10]. LED chips 1140×1140 μm in size and 150 μm thick were mounted onto an aluminum-core printed circuit board (MCPCB) using the chip-on-board (COB) technology. The LED array was an assembly consisting of 100 chips 45×45×1.0 mm in size with a total input power of up to 100 W, which corresponds to a current of 350 mA passing through a single chip.

The LED array comprised 10 parallel-connected LED rows, each including 10 series-connected chips. The total area of the LED assembly was 20 Ч 20 mm. The chips were protected with a silicone gel containing luminophore.

The aluminum plate with the LED array was screwed to a heatsink. The diagonal distance between the heads of the screws fixing the board to the heatsink was 44 mm. The appearance of the LED array and a cross-sectional view of the structure are shown in Fig. 1.

Measurement of thermal resistances. Thermal resistances were determined using the electric analogy where heat flow is considered instead of electric current and temperature instead of voltage. Heat is transferred from the active region to the chip substrate, then to the aluminum plate through the glue, then through the thermal paste to the heatsink, which comprise elements of an equivalent thermal circuit: chip (R_{chip}), glue (R_{glue}), aluminum plate ($R_{Al\ plate}$). The Cauer model suggests that such a chain consists of a set of thermal resistors connected to a common bus through thermal capacities. The thermal capacities of different layers of the LED assembly affect only the transient characteristics, i.e., either the heating or cooling rates of the device when the current is switched on/off.

To determine the thermal resistance by the temperature-dependent parameter (the forward voltage drop time), the array was initially switched to a low test current of 50 mA so that

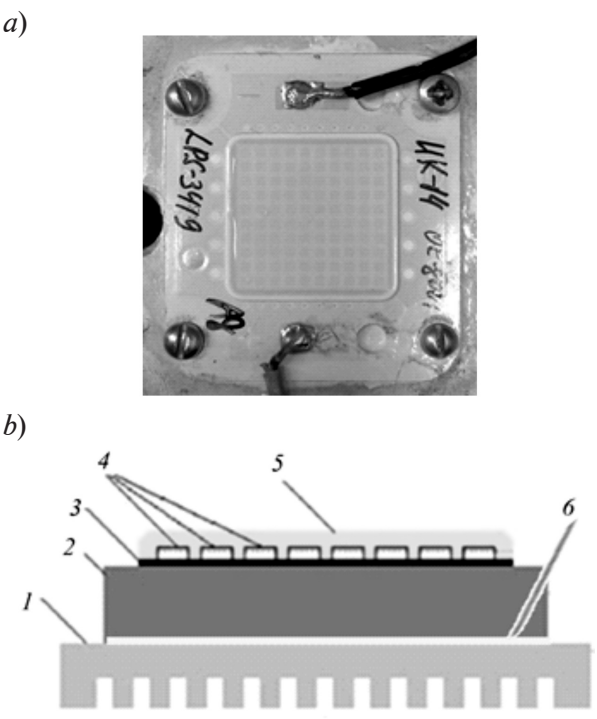

Fig. 1. LED array: appearance from above (*a*) and schematic cross-sectional view (*b*)
heatsink *1*, aluminum board *2*, glue *3*, LED chips *4*, silicone gel *5*, thermal paste *6*

the device would not self-heat, and the temperature of the p—n-junction was set by an external heater in the range of 20 – 100 °C with an accuracy of 0.5°C. Forward voltage was recorded as a function of temperature. A calibration curve for forward voltage versus temperature was obtained this way; the curve is close to a linear dependence with a coefficient of −13 mV/K. This value was subsequently used to determine the temperature of the p—n-junction in real operating mode.

The forward voltage drop (a transient characteristic) was studied with rapid switching from a low test current to a high operating one. The device was gradually heated from that moment, with heat transferred from the active region through the chip and the PC board to the heatsink and the ambient environment. The temperature evolution of the p—n junction under heating was measured by the changes in the forward voltage at the moment when short test current pulses, "cutting" the direct heating operating current, were supplied at a specific frequency. Subsequent mathematical analysis of the transient voltage characteristic in the p—n junction using the structure function approach [11] allowed to calculate the components $R_{th,I}$ and $C_{th,i}$ of the equivalent thermal circuit, the total thermal resistance ΣR_{th} and the total heat capacity ΣC_{th}. The continuous cumulative structure function was approximated by a step function, which was a direct representation of Cauer's thermal impedance model. The methods of transient characteristics and the mathematical tools involved are discussed in more detail in [12] and the references cited therein.

The T3Ster was originally intended for electronic devices, and its data processing is based on the premise that electrical power supplied to the device is completely converted to heat. However, a significant fraction of the supplied electric power in modern high-performance LEDs is converted to light and, therefore, does not contribute to heating the device. To account for this, output optical power P_{opt} was measured using the OL 770-LED High Speed LED Test and Measurement system with an integrating sphere [13]. The wall-plug efficiency for the given array amounted to 15 – 20% (depending on the input current). The corresponding part of the input power carried away by radiation was taken into account in calculations of thermal resistance.

IR thermal imaging. Temperature of the LED array surface was measured using a SVIT IR thermal imaging camera with a sensitivity range of 2.5 – 3.0 μm [10]. Measuring the temperature directly with a thermal imager allows obtaining the temperature area distribution (so-called thermal mapping).

The main methodological issues encountered in thermal mapping of AlInGaN-based structures are, firstly, that the sapphire substrate and epitaxial layers are transparent for IR wavelengths, and, secondly, that the emissivities of the materials used in LEDs (semiconductor layers, metal contacts, reflective coatings, mounting elements, etc.) are largely different [14]. For this reason, preliminary calibration is required to extract the correct temperature distributions from the IR images. This calibration was carried out, with the temperature maintained by an external heater in the range from 20 to 100 °C, by recording the IR radiation from the LED array at zero current. Using this approach, we were able to measure

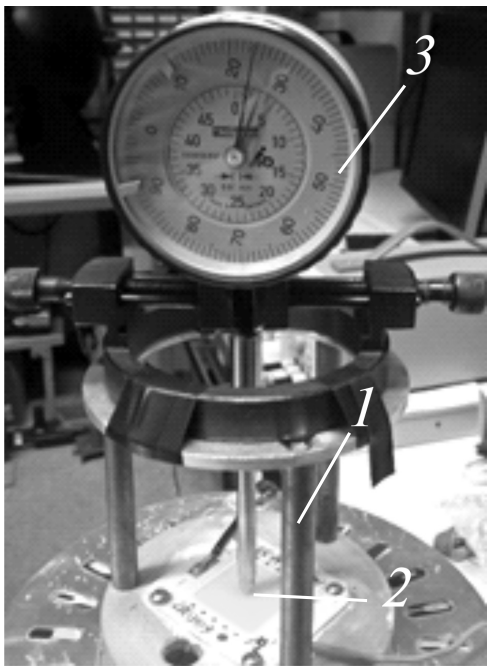

Fig. 2. Photograph of the spherometer used to measure the curvature radius of the LED array:
metal tripod 1, pointed tip 2, gauge 3

the temperature with an accuracy up to 2 K.

Measurements of the surface curvature. We used a spherometer with a dial gauge (SЬSS MicroTec, Germany) to estimate the thermal deformation of the LED array during operation, measuring the elevation of the center of the LED array in operating mode (with a current of 3.5 A) compared to the position of this center at zero current.

The spherometer consists of a metal tripod with three fixed legs of the same length (Fig. 2 [15]), a pointed tip passing along the center of the frame parallel to the legs and a standard dial gauge with a 0.01 mm graduation, showing the elevation of the tip above or below the surface on which the legs of the spherometer are resting. The position of the tip can be read with an accuracy of 5 μm.

A stationary hydrodynamic model (Computational Fluid Dynamics, CFD) was used for a flat LED array at the first stage of simulating the heat dissipation; possible deformation of the LED array (curvature of the heatsink surface) was taken into account at the next stage.

A stationary CFD model describing the thermal processes in the experimental samples was created in Flotherm 10.1 by Mentor Graphics (Fig. 3 [16]). The goal of the simulation was to reproduce the results obtained in the experiment and to study the causes of temperature gradients between the center and the periphery of the LED array. A large cylindrical aluminum heatsink was approximated by an aluminum block, and the effect of cooling fins by a high heat transfer coefficient (10,000 W/(m^2·K)) applied to the walls of the aluminum block.

The adhesive (glue) layer used to mount the chip and the thermal paste layer (15 μm thick) between the plate and the heatsink were simulated as thermal resistances at the respective interfaces. The protective silicone gel layer had a thickness of 300 μm. The thermal resistance of the glue, equal to 2.6 K/W for each chip, was obtained from the total structure function corresponding to a heating current of 1 A. This current value provided the most uniform heating of the LED array. The thermophysical properties of the materials are given in Table.

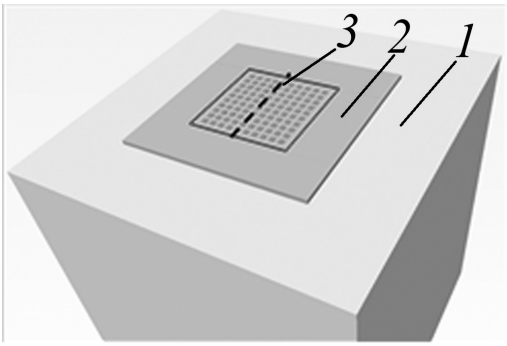

Fig. 3. CFD model of the experimental setup
aluminum block 1; LED array 2 coated with silicone gel; central part of the array (indicated by the dashed line)

Constant pressure was imposed at the boundaries of the computational domain. The ambient temperature was 20°C. The model also included thermal radiation. An algebraic turbulence model was used for the simulation. Grid-independent results were obtained by simulating a model with different cell densities. The computational domain contained 1.8 million cells with local grids for LED chip, silicone gel and aluminum board. The dependence of the power emitted by LEDs on temperature at a fixed voltage applied to the array (Fig. 4) was obtained by measurements of a single chip and included optical cooling.

Table

Thermophysical properties of the simulated system [16 −18]

System component (material)	Thermal conductivity, W/(m·K)	Density, kg/m^3
LED chip (sapphire)	36.0	3980.0
Board and heatsink (aluminum)	201.0	2710.0
Protective gel (silicone)	1.0	1000.0
Thermal paste	0.67	–

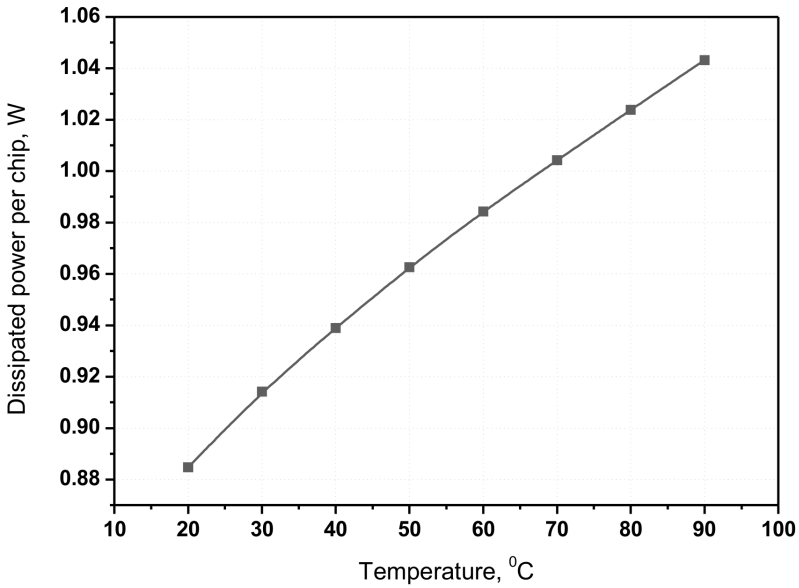

Fig. 4. Dissipated power as a function of temperature for a single LED chip

Results and discussion

The results of the measured thermal resistances for the LED array are given in Fig. 5 as cumulative structure functions. The values of thermal resistance R_{th} are plotted along the horizontal axis, and the values of specific heat C_{th} from the heat source to the ambient (shown on a logarithmic scale) are plotted along the vertical axis.

The value of the total thermal resistance

$$\Sigma R_{th} = R_{chip} + R_{glue} + R_{Al\ plate} \qquad (1)$$

was obtained for three currents: 1.0, 3.5 and 4.0 A. Evidently, quantity (1) increases from 0.3 to 0.5 K/W (by about 1.7 times) with increasing current. The inflection points on the curves indicate the thermal resistances measured along the thermal circuit. The increase in thermal resistance with increasing current can be explained by its redistribution in favor of the central arrays compared to the peripheral ones, and thus reduced sizes of the heat generation and, accordingly, the heat dissipation regions. However, the difference in the currents turned out to be insignificant (within 4 %), so heat generation can be considered nearly homogeneous in the entire LED array. This means that the increase in thermal resistance with increasing current is associated with the changes in heat dissipation rather than in in heat generation.

Heat transfer from the center of the array to the ambient is worse than at the periphery of the plate. This was confirmed by direct measurement of heat distribution with an IR thermal imaging camera. The observed temperature distribution along the central axis of the array is shown in Fig. 6. A noticeable temperature difference up to 13 K is observed between the central and peripheral chips of the array.

First simulations revealed a 3 – 4 K maximum temperature variation between the center and the periphery of the LED array. The CFD model initially implied that the thermal resistance between the aluminum plate and the heatsink did not vary at different points, that is, the thermal paste layer between the aluminum plate and the radiator was assumed to be homogeneous.

Upon analysis of the computational data and closer inspection of the LED array samples, we have hypothesized that the temperature variation (up to 13 K) between the central and

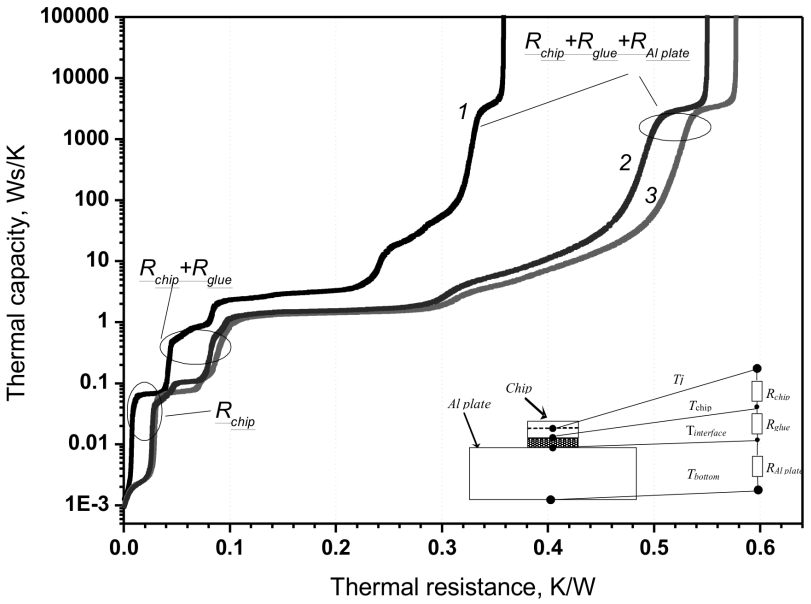

Fig. 5. Cumulative structural functions (thermal capacity versus thermal resistance) for the LED array with different currents, A: 1.0 (1) 3.5 (2) and 4.0 (3). The inset shows a simplified diagram of the thermal circuit

peripheral chips could be caused by bending of the aluminum plate. Bending also occurs due to intense heating of LED chips under rigid mechanical constraints imposed by screws in the corners of the plate

Thermal deformation of the LED array was measured by a spherometer at two points: at the center of the aluminum plate and next to one of the screws. Vertical shift in operating mode was 80 μm at the center of the LED array, and almost did not change near the screw. Therefore, the experiment confirmed

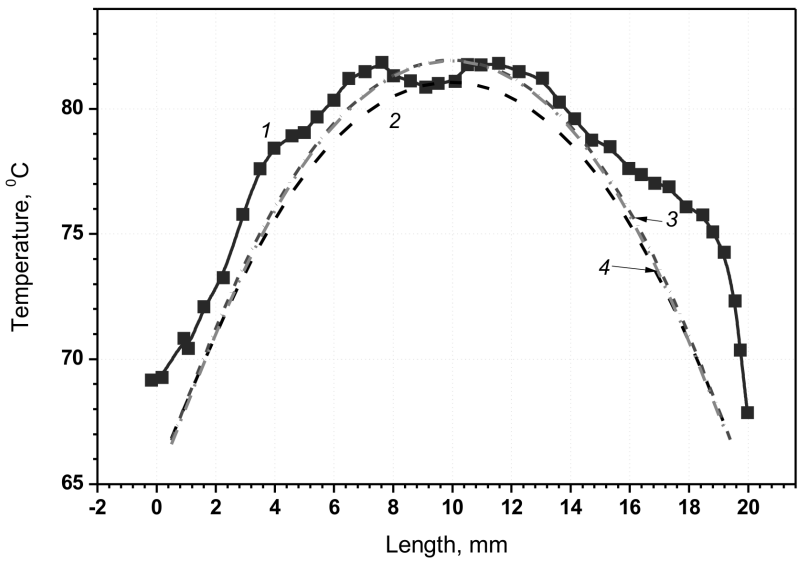

Fig. 6. Experimental (symbols 1) and calculated (lines 2 – 4) temperature distributions along the central axes of the LED arrays

that overheating at the center of the LED array is due to thermal deformation of the aluminum plate, leading to deterioration of the thermal contact between the aluminum plate and the heatsink.

The deterioration in thermal resistance at the interface between the aluminum plate and the heatsink was taken into account in the model by partitioning the thermal resistances between the heatsink and the aluminum plate into zones. Fig. 7 shows examples of the aluminum board partitioned into 4 and 6 zones of thermal resistances. The partitioning was based on the assumption that the variation (growth) in thermal resistance near the edges of the plate should be less compared to the center due to the plate's bending. The thermal resistance values assigned to individual zones were obtained by calibrating the simulated temperatures with respect to the measurements obtained in the central part of the LED array (the surface temperature of silicone is higher than 80 – 82 °C for the central chips and 65 – 68 °C for the periphery, see Fig. 6). The results of partition into thermal zones were additionally compared with the results for the resistances of a homogeneous thermal paste layer (without partitioning). Using thermal resistance zones in the model led to achieving good agreement with the experimental data in the measured temperature range. The thermal resistance values for the thermal zones are shown in Fig. 8, and the results of simulation and comparison between the measured data and the simulated temperatures are shown in Fig. 6. We have tested the results of partitions into a different number of zones. It can be seen from the data in Fig. 6 that no considerable variation could be observed in the results obtained by increasing the number of thermal resistance zones to six, compared with a coarser partition.

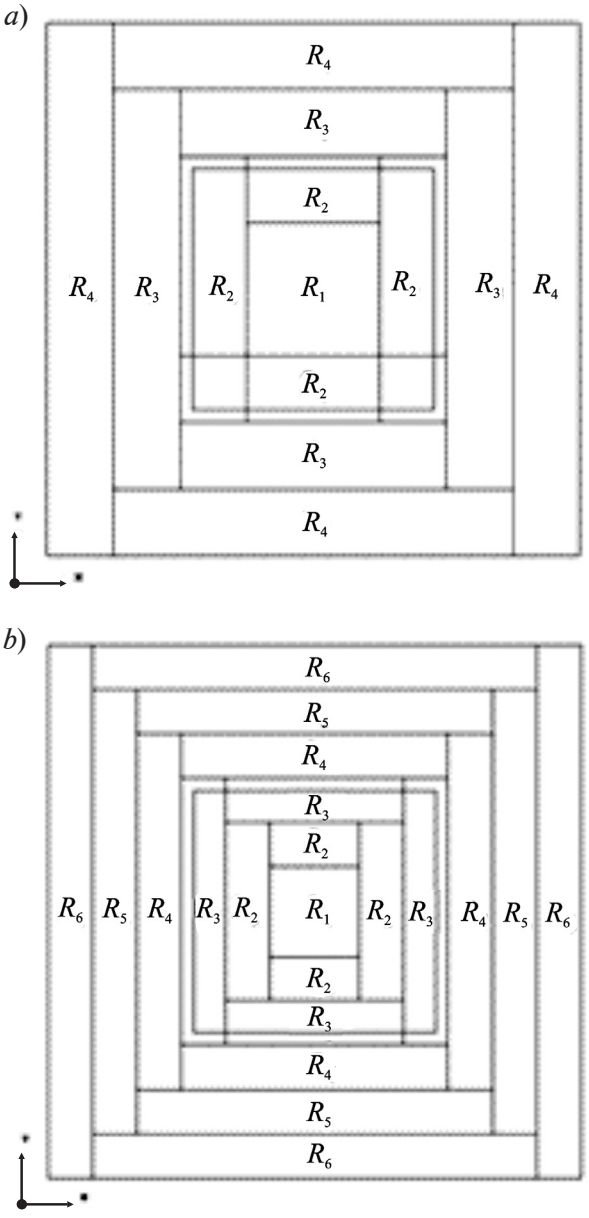

Fig. 7. Selected partitions of the aluminum board model into 4 (a) and 6 (b) zones of thermal resistance (Ri is the resistance of an ith zone)

The good agreement between the simulated and experimental temperatures obtained by partitioning the thermal resistances is shown in Fig. 6.

It is evident from the results of multizonal partitioning that thermal resistance significantly deteriorates starting from the zone next to the edge of the aluminum plate (i.e., R_3 and R_5). This is consistent with the fact that no continuous thermal paste layer could be found under the bent part of the aluminum board. Since the total measured bending height was 80 μm and the maximum thickness of the thermal paste layer was 15 μm, the area filled with thermal paste could only be located below the zone corresponding to the thermal resistance R_4 (for partition into four zones) or

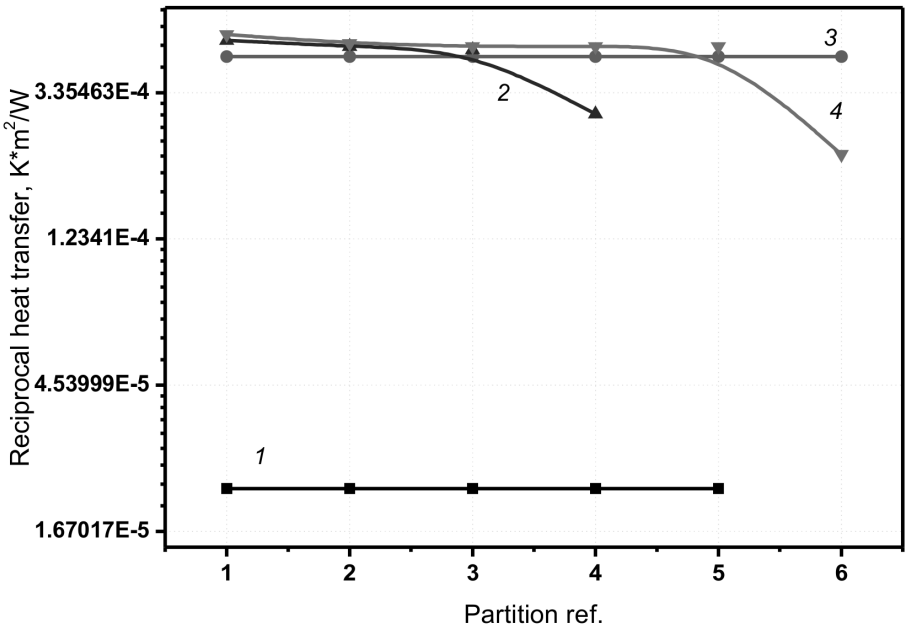

Fig. 8. Inverse thermal conductivity coefficients corresponding to different partition zones with respect to the resistance of the homogeneous thermal paste layer:
homogeneous thermal paste layer 1, four partition zones 2, without partition into zones 3, six partition zones 4

R_6 (for partition into six zones). The effective thermal resistance of these zones, 0.1 K/W, determined by calibrating the model, was approximately 10 times worse than the thermal resistance of a homogeneous thermal paste layer of 0.011 K/W under a non-deformed aluminum plate. A possible explanation for such a high value is that the aluminum plate is deformed along the edges under heating. It is also clear from the simulation results and temperature measurements that the effect of thermal deformation of the aluminum plate on the temperature distribution profile of the LED array can be represented by one effective thermal resistance, as shown in Fig. 6 and 8, i.e., without partitioning. Using this alternative is not substantiated from a physical standpoint, but leads to satisfactory agreement between the simulated and measured temperature.

The temperature distribution over the area of the LED array is shown in Fig. 9. The method of partitioning the thermal resistances allows to predict the formation of local hot spots on the array surface, and thus reproduces the effect of the measured thermal deformation of the aluminum plate.

Conclusion

We have carried out experiments and computer simulation to study the thermal properties of high-power white AlInGaN LED arrays based on emitting face-up chips mounted on an aluminum MCPCB board using the chip-on-board technology. Experimental studies

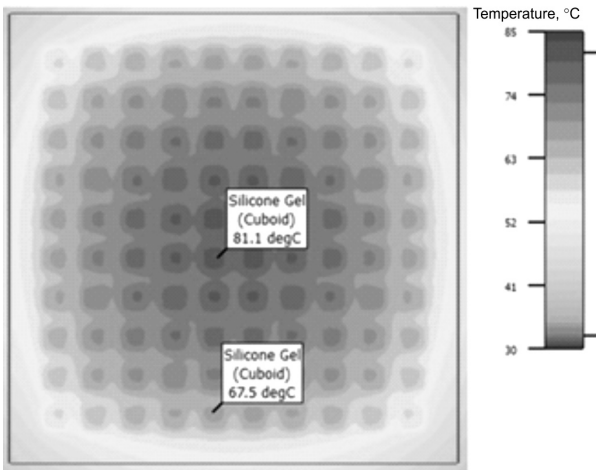

Fig. 9. Calculated temperature map for the surface of the LED array (with thermal deformation taken into account)

involved both indirect methods for determining thermal parameters from the transient temperature-dependent characteristics and a direct method for determining the temperature via IR thermal imaging.

We have established that the total thermal resistance of the LED array increased by 1.7 times as the operating current increased from 1 to 4 A, due to significant deterioration in heat removal from the chips located at the center of the array compared to those located on the periphery. This is a consequence of deformation caused by linear thermal expansion: more specifically, the central part bends as the aluminum board is fixed by the corners with screws, the gap between this board and the heatsink widens and the thermal contact deteriorates. The latter is confirmed both by the temperature distribution obtained from IR temperature mapping, and by direct measurement of the curvature of the LED array's surface in operating mode.

Mathematical and experimental simulation played a key role in understanding the observed phenomenon of thermal deformation. A small change in surface temperature, obtained as a result of CFD simulation with a homogeneous layer of thermal paste, allowed us to suggest thermal deformation of the LED array to be the reason for the actually measured temperature gradient. The simulation results were then confirmed by direct measurement of a significant bending height (up to 80 μm) of the aluminum board in operating mode.

Additionally, we have proposed a method for breaking the thermal resistance into zones, giving an example of CFD simulation of experimental samples which is in good agreement with thermal imaging results. The difference in temperature between the central and peripheral chips can reach 13 K at an input power of 100 W. Overheating of the central chips reduces the service life of the LED array. This should be taken into account when estimating the thermal resistances obtained by forward voltage drop.

Finally, the given combination of experimental methods and simulation techniques should be of help to electronics developers, facilitating analysis and solution of reliability problems caused by local hot spots evolving in the LED array as a result of thermomechanical deformation of the array's components during operation.

Measurements of LED characteristics were carried out at the Collective Use Center "Hardware components of radiophotonics and nanoelectronics: technology, diagnostics, metrology", St. Petersburg, Russian Federation.

REFERENCES

[1] **Zh. Ping, Z. Jianhua, C. Xianping C., et al.,** An experimental investigation of a 100-W high-power light-emitting diode array using vapor chamber-based plate, Advances in Mechanical Engineering. 7 (11) (2015) 1–7. https//doi.org/10.1177/1687814015620074.

[2] **F. Jiajie, X. Chaoyi, Q. Cheng, et al.,** Luminescence mechanism analysis on high power tunable color temperature Chip-on-Board white LED modules, Proc. of the 18th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE), 3–5 April 2017, Dresden, Germany (2017) 1–6.

[3] **H.L. John,** Chip on board, Technology for multichip modules, Springer, New York, 1994.

[4] **S.G. Konsowski, A.R. Helland,** Electronic packaging of high speed circuitry, McGraw Hill Professional, New York, 1997.

[5] **A.G. Shusharin, V.V. Morozov, M.P. Polovinka,** Medical infrared imaging – modern features of the method, Modern problems of science and education. (4) (2011), URL: http://science-education.ru/ru/article/view?id=4726.

[6] **S.C. Ki, C.Y. Sun, K. Jae-Young, et al.,** Precise temperature mapping of GaN-based LEDs by quantitative infrared micro-thermography, Sensors. 12 (4) (2012) 4648–4660.

[7] MicReD. T3Ster, URL: https://www.mentor.com/products/mechanical/ micred/t3ster/.

[8] Thermal management for LED applications, C.J.M. Lasance, A. Poppe (Eds), Springer, New York, 2014.

[9] ES-CABLV45P data sheet, 2017, URL: http://www.epistar.com. tw/upfiles/files_/ES-CABLV45P.pdf

[10] **A.V. Aladov, K.A. Bulashevich, A.E. Chernyakov, et al.,** Thermal resistanse and nonuniform distribution of electroluminescence and temperature in high-power AlGaInN light-emitting diodes, St. Petersburg Polytechnical University Journal: Physics and Mathematics. (2 (218)) (2015) 151–158.

[11] **D. Schweitzer, H. Pape, L. Chen, et**

al., Transient dual interface measurement – A new JEDEC standard for the measurement of the junction-to-case thermal resistance, Proc. of the 27th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM'11), 20–24 March, 2011, San Jose, USA. (2011) 222–229.

[12] **V.I. Smirnov, V.A. Sergeev, A.A. Gavrikov,** Apparatus for measurement of thermal impedance of high-power light-emitting diodes and LED assemblies, IEEE Transactions on Electron Devices. 63 (6) (2016) 2431–2435.

[13] **A.L. Zakgeim, G.L. Kuryshev, M.N. Mizerov, et al.,** A study of thermal processes in highpower InGaN/GaN FlipC-hip LEDs by IR thermal imaging microscopy, Semiconductors. 44 (3) (2010) 373–379.

[14] **R.H. Hopper, I. Haneef, S.Z. Ali, et al.,** Use of carbon micro-particles for improved infrared temperature measurement of CMOS MEMS devices, Measurement Science and Technology. 21 (4) (2010) 1–6.

[15] **R.P. Shukla, D. Udupa,** Measurement of radius of curvature of cylindrical surfaces, Journal of Optics (India). 30 (3) (2001) 131–142.

[16] Mentor Graphics Corporation. Flotherm v.10.1 User Manual (2014).

[17] Silicone Gel, 2017, URL: https://www.acc-silicones.com.

[18] Organo-silicon heat-conducting paste, Specifications, 2017, URL: http://pripoi.ru.

[19] **A.L. Zakgeim, A.E. Chernyakov,** A measuring system for obtaining spectroradiometric, photocolorimetric, and thermal characteristics of semiconductor radiators, Light and Engineering. 21 (4) (2013) 64–70.

Received 25.06.2018, accepted 09.07.2018.

THE AUTHORS

ALADOV Andrey V.
Submicron Heterostructures for Microelectronics Research and Engineering Center of the RAS
26 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
aaladov@mail.ioffe.ru

BELOV Ilia V.
Jönköping University, School of Engineering
5 Gjuterigatan St., Jönköping, Sweden
ilia.belov@ju.se.ru

VALYUKHOV Vladimir P.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
Valyukhov@yandex.ru

ZAKGEIM Alexader L.
Submicron Heterostructures for Microelectronics Research and Engineering Center of the RAS
26 Politekhnicheskaya St., St. Petersburg, 194021, Russian Federation
zakgeim@mail.ioffe.ru

CHERNYAKOV Anton E.
Submicron Heterostructures for Microelectronics Research and Engineering Center of the RAS
26 Politekhnicheskaya St., St. Petersburg, 194021, Russian Federation
chernyakov.anton@yandex.ru

GENERALIZATION OF THE PSEUDOPOTENTIAL CONCEPT FOR RADIO-FREQUENCY QUADRUPOLE FIELDS

A.S. Berdnikov[1], L.N. Gall[1], N.R. Gall[1], K.V. Solovyev[2]

[1] Institute for Analytical Instrumentation of the Russian Academy of Sciences, St. Petersburg, Russian Federation;

[2] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation

It is shown that the pseudopotential function, which describes the averaged motion of charged particles with accuracy up to quadratic terms for nonuniform radio-frequency fields, can be replaced by an infinite pseudopotential series for quadrupole radio-frequency electric fields. This replacement provides a more accurate description. It allows us to extend the parameter's range of the radio-frequency field; in this range, it makes possible to describe the motion of charged particles quantitatively and not just qualitatively. Unfortunately, even this extended concept of pseudopotential is not suitable enough for describing the motion of charged particles when approaching the region of the parametric resonance, where the motion of charged particles loses stability in the quadrupole radio-frequency fields.

Key words: high-frequency electric field, quadrupole mass filter, secular oscillation, pseudopotential

Citation: A.S. Berdnikov, L.N. Gall, N.R. Gall, K.V. Solovyev, Generalization of the pseudopotential concept for radio-frequency quadrupole fields, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 38–48. DOI: 10.18721/JPM.11305

Introduction

The pseudopotential approach is a useful tool for qualitatively describing ion motion in nonuniform radio-frequency (RF) electric fields [1 – 12]. However, the accuracy of the classical pseudopotential approach is too low for RF quadrupole mass filters [11 – 15] and (to a lesser extent) for RF quadrupole traps [16, 17], so the approach can hardly be regarded as useful for investigating the motion of charged particles in these devices. Exceptions from this include pseudopotential functions for calculating the stroboscopic values of coordinates and velocities [18 – 20], or interpretation of Floquet – Lyapunov matrices for solutions of linear differential equations with periodic coefficients in the sense of a pseudopotential model of motion [21, 22]. These pseudopotential functions are based on a fundamentally different mathematical formalism; still, these models of motion are not very convenient for practical calculations.

In this study, we have considered a reasonable compromise between classical models that are practical but not particularly accurate in terms of analyzing the motion of charged particles in RF quadrupole fields [1 – 12], and models that are mathematically accurate but not very practical [18 – 22]. The models we have proposed make it possible to significantly expand the range of parameters of the RF quadrupole electric field within the first stability zone. Using these parameters allows to fit (not only qualitatively but also quantitatively) the approximate motion trajectories to exact solutions of the corresponding differential equations.

The pseudopotential models of motion discussed in the paper follow the general ideology of the classical pseudopotential theory [1 – 12] and produce easily calculable algebraic expressions. However, these models work poorly near the boundary of the stability

region of RF quadrupoles which corresponds to parametric resonance between the driving RF field and the secular motion of charged particles violating the basic assumptions that the RF component of charged particle motion is small compared to the "slow" (averaged over RF oscillations) component of motion. Furthermore, the formulae obtained are specific for RF quadrupole electric fields and cannot be generalized to motion of charged particles in nonlinear RF electric fields.

Classical pseudopotential model for motion in a RF quadrupole field

Let us consider the motion of an ion in a RF electric field of a linear quadrupole with hyperbolic rods [11 – 17]. The electric potential $U(x, y, t)$ for such a system has the form

$$U(x,y,t) = \\ = (U_0 + V_0 \cos(\Omega t + \varphi_0))(x^2 - y^2)/r_0^2, \quad (1)$$

where U_0 is the constant component of the voltages applied to the electrodes; V_0 is the amplitude of the cosinusoidal RF component of the voltages applied to the electrodes; Ω is the circular frequency of the RF voltage, φ_0 is the phase of the RF voltage at the start of ion motion; r_0 is the shortest distance from the quadrupole axis to the hyperbolic electrodes (characterizing the interelectrode gap of the RF linear quadrupole); x and y are the Cartesian coordinates; t is the time of motion.

In dimensionless coordinates, the trajectory $x(t)$, $y(t)$ for an ion with the mass m and charge e satisfies the Mathieu-type equations [23 – 30], which are a particular case of linear differential equations with periodic coefficients:

$$\frac{d^2x}{d\xi^2} + (a + 2q\cos(2\xi + \varphi_0))x = 0, \quad (2)$$

$$\frac{d^2y}{d\xi^2} - (a + 2q\cos(2\xi + \varphi_0))y = 0, \quad (3)$$

where $\xi = \Omega t/2$ is the dimensionless time; $a = 8eU_0/m\Omega^2 r_0^2$, $q = 4eV_0/m\Omega^2 r_0^2$ are the dimensionless parameters; $f(\xi) = \cos(2\xi + \varphi_0)$ is the cosinusoidal periodic function with a dimensionless period $T' = \pi$ (the dimensionless circular frequency $\Omega' = 2$) and the initial phase φ_0.

To illustrate the principles of the classical pseudopotential approach, let us consider one-dimensional motion of an ion with mass m and charge e in an RF electric field with an electric potential of the general form

$$U(x,t) = U^0(x,t) + V(x,t)\cos(\Omega t + \varphi_0) + \\ + W(x,t)\sin(\Omega t + \varphi_0), \quad (4)$$

where $U^0(x, t)$, $V(x, t)$, $W(x, t)$ are assumed to be "slow" functions of time, in comparison with "rapidly" oscillating sinusoidal functions $\cos(\Omega t + \varphi_0)$, $\sin(\Omega t + \varphi_0)$.

Newton's equations of motion of an ion in such an electric field take the form

$$(m/e)\ddot{x} = -U_x^0(x,t) - V_x(x,t)\cos(\Omega t + \varphi_0) - \\ - W_x(x,t)\sin(\Omega t + \varphi_0), \quad (5)$$

where the subscripts denote partial derivatives, which helps subsequently avoid unnecessarily cumbersome mathematical expressions.

We assumed for the pseudopotential motion model [1 – 12] that the solution of differential equation (5) can be represented with good accuracy as a sum

$$x(t) = x_0(t) + \delta x(t),$$
$$\delta x(t) \approx \sum_{k=0}^{\infty} \frac{1}{\Omega^k}(x_k^c(t)\cos(\Omega t + \varphi_0) + \\ + x_k^s(t)\sin(\Omega t + \varphi_0) + x_k^{2c}(t)\cos 2(\Omega t + \varphi_0) + \\ + x_k^{2s}(t)\sin 2(\Omega t + \varphi_0) + \cdots), \quad (6)$$

where the "rapid" component of the trajectory $\delta x(t)$, like its time derivative, has a zero mean (calculated over the period of the RF field (4)) and is small, compared with the principal ("slow") component of the trajectory $x_0(t)$.

Let us substitute sum (6) into Eq. (5) and expand both the functions $U^0(x, t)$, $V(x, t)$ and $W(x, t)$ and their partial derivatives to truncated Taylor series with respect to the small $\delta x(t)$ increment. In this case, under certain conditions, namely:

a) assuming that the functions $x_k^c(t)$, $x_k^s(t)$, $x_k^{2c}(t)$, $x_k^{2s}(t)$, ... are "slow",

b) combining together terms that are basic trigonometric functions with the same frequencies and the same powers of Ω,

c) demanding that the corresponding coefficients (except the terms corresponding to the zero harmonic of the RF field) vanish sepa-

rately, the following approximate relations can be obtained:

$$x(t) \approx x_0(t) + \frac{e}{m\Omega^2} V_x(x_0(t),t)\cos(\Omega t + \varphi_0) + \\ + \frac{e}{m\Omega^2} W_x(x_0(t),t)\sin(\Omega t + \varphi_0) + \cdots; \quad (7)$$

$$\dot{x}(t) \approx \dot{x}_0(t) + \frac{e}{m\Omega} W_x(x_0(t),t)\cos(\Omega t + \varphi_0) - \\ - \frac{e}{m\Omega} V_x(x_0(t),t)\sin(\Omega t + \varphi_0) - \\ - \frac{e}{m\Omega^2}[V_{xt}(x_0(t),t) + \dot{x}_0(t)V_{xx}(x_0(t),t)]\times \\ \times\cos(\Omega t + \varphi_0) - \frac{e}{m\Omega^2}[W_{xt}(x_0(t),t) + \\ + \dot{x}_0(t)W_{xx}(x_0(t),t)]\sin(\Omega t + \varphi_0) + \cdots; \quad (8)$$

$$\ddot{x}_0(t) \approx -\frac{e}{m}U_x(x_0(t),t) - \frac{e}{m}\bar{U}_x^{rf}(x_0(t),t) + \cdots. \quad (9)$$

The powers of Ω up to $1/\Omega^2$ are preserved here, and the higher powers, which are small corrections due to the assumption that the RF electric field has a "high" frequency, are omitted.

It should be noted, however, that in order to obtain the correct expression for the velocity $\dot{x}(t)$ (up to the terms of the form $1/\Omega^2$), the cubic terms $1/\Omega^3$ also have to be preserved in the calculations before differentiating the function $x(t)$ with respect to time; these terms can be eliminated only after the function $\dot{x}(t)$ has been determined correctly.

The function

$$\bar{U}^{rf}(x,t) = \frac{e}{4m\Omega^2}[(V_x(x,t))^2 + (W_x(x,t))^2] \quad (10)$$

is called the pseudopotential (effective potential, RF potential, ponderomotive force potential, etc.), and Eq. (9) can be interpreted as the motion of an ion with mass m and charge e in a quasi-stationary electric field with the potential $U(x,t) + \bar{U}^{rf}(x,t)$.

Importantly, not only the pseudopotential Eq. (9) for the "slow" component of the ion trajectory, but also Eqs. (7), (8) are an integral part of the pseudopotential model of motion. The latter equations allow to explicitly express the high-frequency corrections for the trajectory and velocity of the ion, and thus find an approximate expression for the true trajectory of the ion in the RF electric field. In particular, it follows from Eqs. (7) and (8) that rapidly oscillating corrections to the "slow" component of ion trajectory are directly proportional to the amplitude of the RF component of electric field strength at the given point of the trajectory. Moreover, nonlinear algebraic Eqs. (7), (8) can be used to express the functions $x_0(t)$, $\dot{x}_0(t)$ in terms of functions $x(t)$, $\dot{x}(t)$ as a series in the powers of $1/\Omega^k$:

$$x_0(t) \approx x(t) - \frac{e}{m\Omega^2} V_x(x(t),t)\cos(\Omega t + \varphi_0) - \\ - \frac{e}{m\Omega^2} W_x(x(t),t)\sin(\Omega t + \varphi_0) + \cdots; \quad (11)$$

$$\dot{x}_0(t) \approx \dot{x}(t) - \frac{e}{m\Omega} W_x(x(t),t)\cos(\Omega t + \varphi_0) + \\ + \frac{e}{m\Omega} V_x(x(t),t)\sin(\Omega t + \varphi_0) + \\ + \frac{e}{m\Omega^2}[V_{xt}(x(t),t) + \dot{x}(t)V_{xx}(x(t),t)]\times \\ \times\cos(\Omega t + \varphi_0) + \frac{e}{m\Omega^2}[W_{xt}(x(t),t) + \\ + \dot{x}(t)W_{xx}(x(t),t)]\sin(\Omega t + \varphi_0) + \cdots, \quad (12)$$

where the terms are preserved up to $1/\Omega^2$ both for $x_0(t)$ and for $\dot{x}_0(t)$.

In particular, Eqs. (11) and (12) allow to explicitly express the initial conditions for "slow" motion (9) in terms of the initial conditions of true motion (5) in the RF field.

Notice that the discrepancy between the initial conditions for the functions $x_0(t)$, $\dot{x}_0(t)$ and $x(t)$, $\dot{x}(t)$, as well as the difference between the averaged $x_0(t)$, $\dot{x}_0(t)$ trajectories and the approximate $x(t)$, $\dot{x}(t)$ trajectories are not always taken into account in studies on assessing the accuracy of the pseudopotential model of motion, which yields estimates worse than the actual accuracy.

The normalized equation of motion (2) is obtained from Eq. (5) by the following substitution:

$$U^0(x,t) = ax^2/2, \quad V(x,t) = qx^2, \\ W(x,t) = 0,, \quad \Omega = 2, \\ e = 1, \quad m = 1, \quad t = \xi.$$

As a result, the pseudopotential model of ion motion (7) – (12) yields an approximate solution for Eq. (2), written in dimensionless form:

$$x_0''(\xi) \approx -\left(a + \frac{q^2}{2}\right) x_0(\xi) + \cdots; \qquad (13)$$

$$x_0(0) \approx x(0)\left(1 - \frac{q}{2}\cos\varphi_0\right) + \cdots;$$
$$x_0'(0) \approx x(0) q \sin\varphi_0 + x'(0)\left(1 + \frac{q}{2}\cos\varphi_0\right) + \cdots; \qquad (14)$$

$$x(\xi) \approx x_0(\xi)\left(1 + \frac{q}{2}\cos(2\xi + \varphi_0)\right) + \cdots; \qquad (15)$$

$$x'(\xi) \approx -q x_0(\xi)\sin(2\xi + \varphi_0) + x_0'(\xi)\left(1 - \frac{q}{2}\cos(2\xi + \varphi_0)\right) + \cdots. \qquad (15)$$

It is assumed here that

$$a + q^2/2 = \tilde{\beta}^2 > 0,$$

where $\tilde{\beta} = \sqrt{a + q^2/2}$ is the pseudopotential approximation for the exact value of the normalized secular frequency β [23 – 31].

The condition

$$a + q^2/2 = \tilde{\beta}^2 > 0$$

corresponds to stable ion motion in the RF qua-

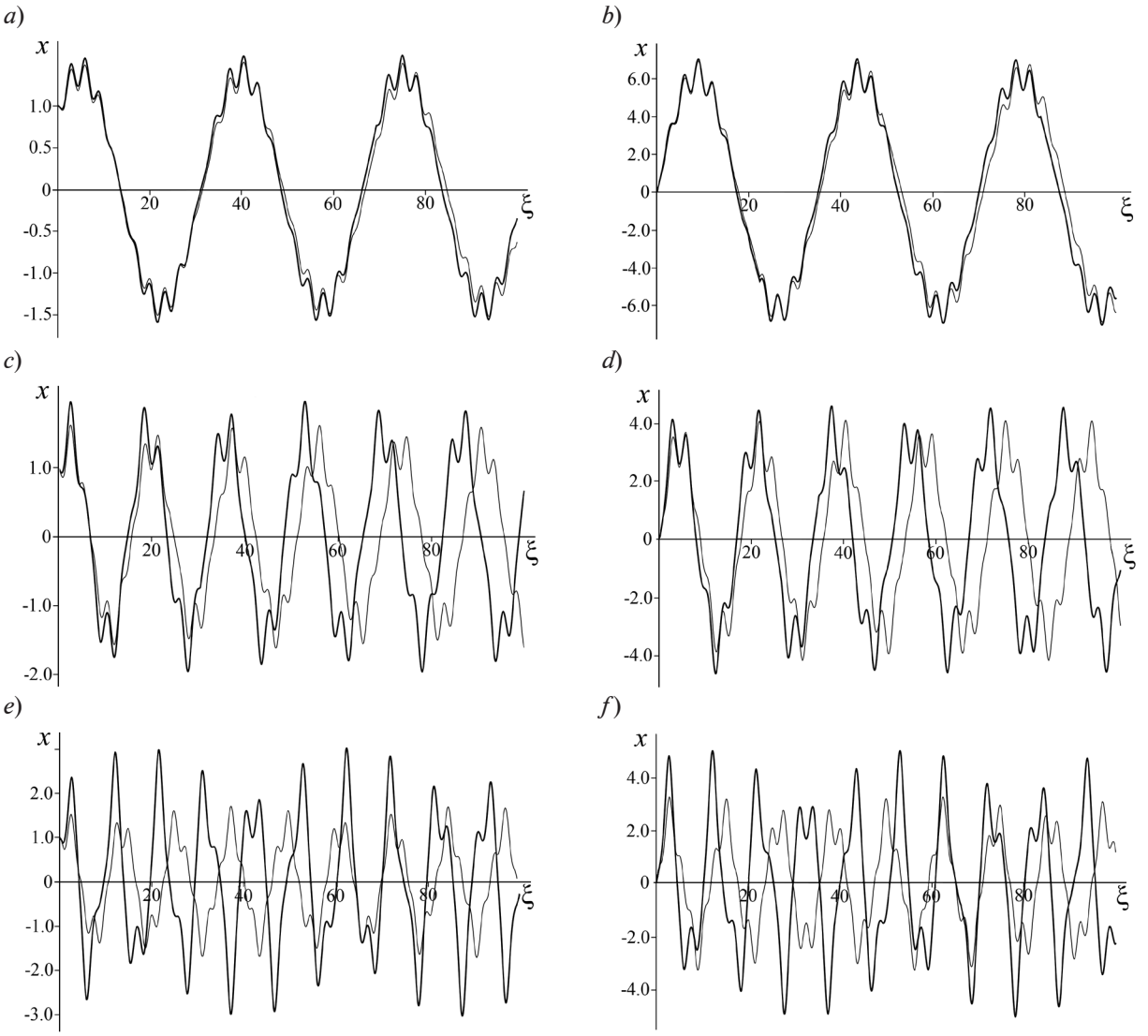

Fig. 1. Comparison of numerically obtained trajectories of Eq. (2) (thin lines) with approximate trajectories calculated by pseudopotential theory (14), (15) (solid lines).
The values of the parameters in Eq. (2) are given in Table

Table

Values of parameters in Eq. (2) for calculating its exact solutions

Fig. 1, 2	q	$x(0)$	$x'(0)$
a	0.25	1	0
b	0.25	0	1
c	0.50	1	0
d	0.50	0	1
e	0.75	1	0
f	0.75	0	1

Note. The parameter $a = 0$ for the entire given set of other parameters

drupole electric field within the pseudopotential model. Fig. 1 shows the difference between the approximate trajectories (13) – (15) and the exact (calculated) solutions of Eq. (2) with $a = 0$ for different values of the parameter q.

Pseudo-potential expansion in an infinite series

If more powers of the form $1/\Omega^k$ are preserved in expansion (6), refined equations can be obtained for "slow" motion, as well as refined coupling equations for true motion $x_0(t)$ and for slow (averaged) motion $x_0(t)$. Unfortunately, in the general case of an arbitrary RF electric field, the expressions obtained by this method turn out to be extremely complex and it is no longer possible to use the elegant and physically transparent classical pseudopotential model to interpret them (conversely, however, see Ref. [32]). Quadrupole electric fields (where the dependence of the electric potential on the coordinates is expressed by a quadratic polynomial) are an exception: high-order corrections for these fields still keep the form of an artificially constructed pseudopotential function.

As an example, let us consider one-dimensional motion in a cosinusoidal RF electric field with a quadratic electric potential:

$$\frac{dx}{dt} = v;$$
$$\frac{dv}{dt} = -\hat{U}x - \hat{V}x\cos(\Omega t + \varphi_0), \tag{16}$$

where, with respect to the linear quadrupole with electric potential (1), the quantity

$$(m/2e)\hat{U}x^2 = U_0\, x^2/r_0^2$$

is the constant component of the electric potential, and the quantity

$$(m/2e)\hat{V}x^2 = V_0\, x^2/r_0^2$$

is the amplitude of the RF component of the electric potential.

The pseudopotential expansion for the solutions of system of equations (16) can be written in the form of a specific series representing a hybrid of trigonometric Fourier series and Taylor power series:

$$x(t) = x_0(t) + \\ + x_0(t)\sum_{k=1,\infty}\cos k(\Omega t + \varphi_0)\left(\sum_{j=k,\infty}\frac{x^{(c)}_{k,2j}}{\Omega^{2j}}\right) + \\ + v_0(t)\sum_{k=1,\infty}\sin k(\Omega t + \varphi_0)\left(\sum_{j=k,\infty}\frac{x^{(s)}_{k,2j+1}}{\Omega^{2j+1}}\right);$$
$$v(t) = v_0(t) + \\ + x_0(t)\sum_{k=1,\infty}\sin k(\Omega t + \varphi_0)\left(\sum_{j=k,\infty}\frac{v^{(s)}_{k,2j-1}}{\Omega^{2j-1}}\right) + \\ + v_0(t)\sum_{k=1,\infty}\cos k(\Omega t + \varphi_0)\left(\sum_{j=k,\infty}\frac{v^{(c)}_{k,2j}}{\Omega^{2j}}\right); \tag{17}$$

$$\dot{x}_0(t) = v_0(t);$$
$$\dot{v}_0(t) = -\left(X_0 + \sum_{j=1,\infty}\frac{1}{\Omega^{2j}}X_{2j}\right)x_0(t). \tag{18}$$

In these equations, $x^{(c)}_{k,2j}$, $x^{(s)}_{k,2j+1}$, $v^{(s)}_{k,2j-1}$, $v^{(c)}_{k,2j}$, X_{2j} are unknown constants, which should be selected so that solution (17), (18) satisfies system of equations (16). Indeed, after substituting solution (17), (18) into system (16) and combining together the coefficients for trigonometric terms similar to

$$\cos k(\Omega t + \varphi_0),\ \ \sin k(\Omega t + \varphi_0)$$

and power terms $1/\Omega^j$, we can express the constants $x^{(c)}_{k,2j}$, $x^{(s)}_{k,2j+1}$, $v^{(s)}_{k,2j-1}$, $v^{(c)}_{k,2j}$, X_{2j} in a consistent manner using the recurrence relations in terms of the constants $\hat{U}$ and $\hat{V}$ entering Eqs. (16). In this case, the function

$$\hat{U}^{rf}(x_0) = \frac{1}{2}\left(X_0 + \sum_{j=1,\infty}\frac{1}{\Omega^{2j}}X_{2j}\right)x_0^2 = \frac{1}{2}\hat{\beta}^2 x_0^2, \tag{19}$$

used for writing the differential equation (18) in the form

$$\ddot{x}_0 = -d\hat{U}^{rf}(x_0)/dx_0,$$

can be interpreted as the refined quadratic pseudopotential. The latter characterizes "slow" (secular) ion motion in a quadratic RF electric field.

In particular, nonzero coefficients $x^{(c)}_{k,2j}$, $x^{(s)}_{k,2j+1}$, $v^{(s)}_{k,2j-1}$, $v^{(c)}_{k,2j}$, X_{2j}, required for calculating Eqs. (17), (18) up to terms of the form $1/\Omega^6$ are determined as

$$X_0 = \hat{U},\ X_2 = \frac{1}{2}\hat{V}^2,\ X_4 = 2\hat{U}\hat{V}^2,$$
$$X_6 = \hat{V}^2\left(8\hat{U}^2 + \frac{25}{32}\hat{V}^2\right);$$
$$v^{(s)}_{1,1} = -\hat{V},\ v^{(s)}_{1,3} = -2\hat{U}\hat{V},$$
$$v^{(s)}_{1,5} = -\hat{V}\left(8\hat{U}^2 + \frac{9}{16}\hat{V}^2\right);$$
$$v^{(c)}_{1,2} = -\hat{V},\ v^{(c)}_{1,4} = -4\hat{U}\hat{V},$$
$$v^{(c)}_{1,6} = -\hat{V}\left(16\hat{U}^2 + \frac{3}{4}\hat{V}^2\right);$$
$$x^{(c)}_{1,2} = \hat{V},\ x^{(c)}_{1,4} = 4\hat{U}\hat{V},$$
$$x^{(c)}_{1,6} = \hat{V}\left(16\hat{U}^2 + \frac{25}{16}\hat{V}^2\right);$$
$$x^{(s)}_{1,3} = -2\hat{V},\ x^{(s)}_{1,5} = -8\hat{U}\hat{V};$$
$$v^{(s)}_{2,3} = -\frac{1}{4}\hat{V}^2,\ v^{(s)}_{2,5} = -\frac{11}{8}\hat{U}\hat{V}^2,$$
$$v^{(c)}_{2,4} = -\frac{5}{8}\hat{V}^2,\ v^{(c)}_{2,6} = -\frac{23}{8}\hat{U}\hat{V}^2;$$
$$x^{(c)}_{2,4} = \frac{1}{8}\hat{V}^2,\ x^{(c)}_{2,6} = \frac{7}{8}\hat{U}\hat{V}^2,\ x^{(s)}_{2,5} = -\frac{3}{8}\hat{V}^2;$$
$$v^{(s)}_{3,5} = -\frac{1}{48}\hat{V}^3,\ v^{(c)}_{3,6} = -\frac{5}{72}\hat{V}^3,\ x^{(c)}_{3,6} = \frac{1}{144}\hat{V}^3,\ \ldots\ . \qquad (20)$$

Functions $x_0(t), v_0(t)$ can be expressed in terms of functions $x(t), v(t)$ with the help of linear equations (17), allowing, in particular, to correctly calculate the initial conditions for "slow" motion $x_0(t), v_0(t)$ in terms of the initial conditions given for the trajectory $x(t), v(t)$.

When the resulting expressions are expanded in a power series with respect to $1/\Omega^k$, we obtain expressions of the form

$$x_0(t) = x(t)\left(1 + \sum_{k=2,\infty}\frac{\tilde{x}^{(0)}_{2k}}{\Omega^{2k}}\right) + x(t)\sum_{k=1,\infty}\cos k(\Omega t + \varphi_0)\left(\sum_{j=k,\infty}\frac{\tilde{x}^{(c)}_{k,2j}}{\Omega^{2j}}\right) + v(t)\sum_{k=1,\infty}\sin k(\Omega t + \varphi_0)\left(\sum_{j=k,\infty}\frac{\tilde{x}^{(s)}_{k,2j+1}}{\Omega^{2j+1}}\right);$$
$$v_0(t) = v(t)\left(1 + \sum_{k=2,\infty}\frac{\tilde{v}^{(0)}_{2k}}{\Omega^{2k}}\right) + x(t)\sum_{k=1,\infty}\sin k(\Omega t + \varphi_0)\left(\sum_{j=k,\infty}\frac{\tilde{v}^{(s)}_{k,2j-1}}{\Omega^{2j-1}}\right) + v_0(t)\sum_{k=1,\infty}\cos k(\Omega t + \varphi_0)\left(\sum_{j=k,\infty}\frac{\tilde{v}^{(c)}_{k,2j}}{\Omega^{2j}}\right). \qquad (21)$$

In particular, substituting expressions (21) into relations (17) and combining such terms, we can express the unknown coefficients $x^{(c)}_{k,2j}$, $x^{(s)}_{k,2j+1}$, $\tilde{x}^{(0)}_{2k}$, $v^{(s)}_{k,2j-1}$, $v^{(c)}_{k,2j}$, $\tilde{v}^{(0)}_{2k}$ directly through a system of recurrent algebraic relations:

$$\tilde{v}^{(0)}_4 = \frac{3}{2}\hat{V}^2,\ \tilde{v}^{(0)}_6 = 10\hat{U}\hat{V}^2;$$
$$\tilde{x}^{(0)}_4 = \frac{3}{2}\hat{V}^2,\ \tilde{x}^{(0)}_6 = 10\hat{U}\hat{V}^2;$$
$$\tilde{v}^{(s)}_{1,1} = \hat{V},\ \tilde{v}^{(s)}_{1,3} = 2\hat{U}\hat{V},\ \tilde{v}^{(s)}_{1,5} = \hat{V}\left(8\hat{U}^2 + \frac{33}{16}\hat{V}^2\right);$$
$$\tilde{v}^{(c)}_{1,2} = \hat{V},\ \tilde{v}^{(c)}_{1,4} = 4\hat{U}\hat{V},$$
$$\tilde{v}^{(c)}_{1,6} = -\hat{V}\left(16\hat{U}^2 + \frac{49}{16}\hat{V}^2\right);$$
$$\tilde{x}^{(s)}_{1,3} = 2\hat{V},\ \tilde{x}^{(s)}_{1,5} = 8\hat{U}\hat{V};$$
$$\tilde{x}^{(c)}_{1,2} = -\hat{V},\ \tilde{x}^{(c)}_{1,4} = -4\hat{U}\hat{V},$$
$$\tilde{x}^{(c)}_{1,6} = -\hat{V}\left(16\hat{U}^2 + \frac{9}{4}\hat{V}^2\right);$$
$$\tilde{v}^{(s)}_{2,3} = \frac{1}{4}\hat{V}^2,\ \tilde{v}^{(s)}_{2,5} = \frac{11}{8}\hat{U}\hat{V}^2,\ \tilde{v}^{(c)}_{2,4} = \frac{1}{8}\hat{V}^2,$$
$$\tilde{v}^{(c)}_{2,6} = \frac{7}{8}\hat{U}\hat{V}^2;$$
$$\tilde{x}^{(s)}_{2,5} = \frac{3}{8}\hat{V}^2,\ \tilde{x}^{(c)}_{2,4} = -\frac{5}{8}\hat{V}^2,\ \tilde{x}^{(c)}_{2,6} = -\frac{23}{8}\hat{U}\hat{V}^2;$$
$$\tilde{v}^{(s)}_{3,5} = \frac{1}{48}\hat{V}^3,\ \tilde{v}^{(c)}_{3,6} = \frac{1}{144}\hat{V}^3,\ \tilde{x}^{(c)}_{3,6} = -\frac{5}{72}\hat{V}^3,\ \ldots\ . \qquad (22)$$

For transition from system of equations (16) to the dimensionless equation, we use the substitution

$$\hat{U} = a,\ \hat{V} = 2q,\ \Omega = 2.$$

Fig. 2 compares approximate solutions, constructed with the help of relations (17),

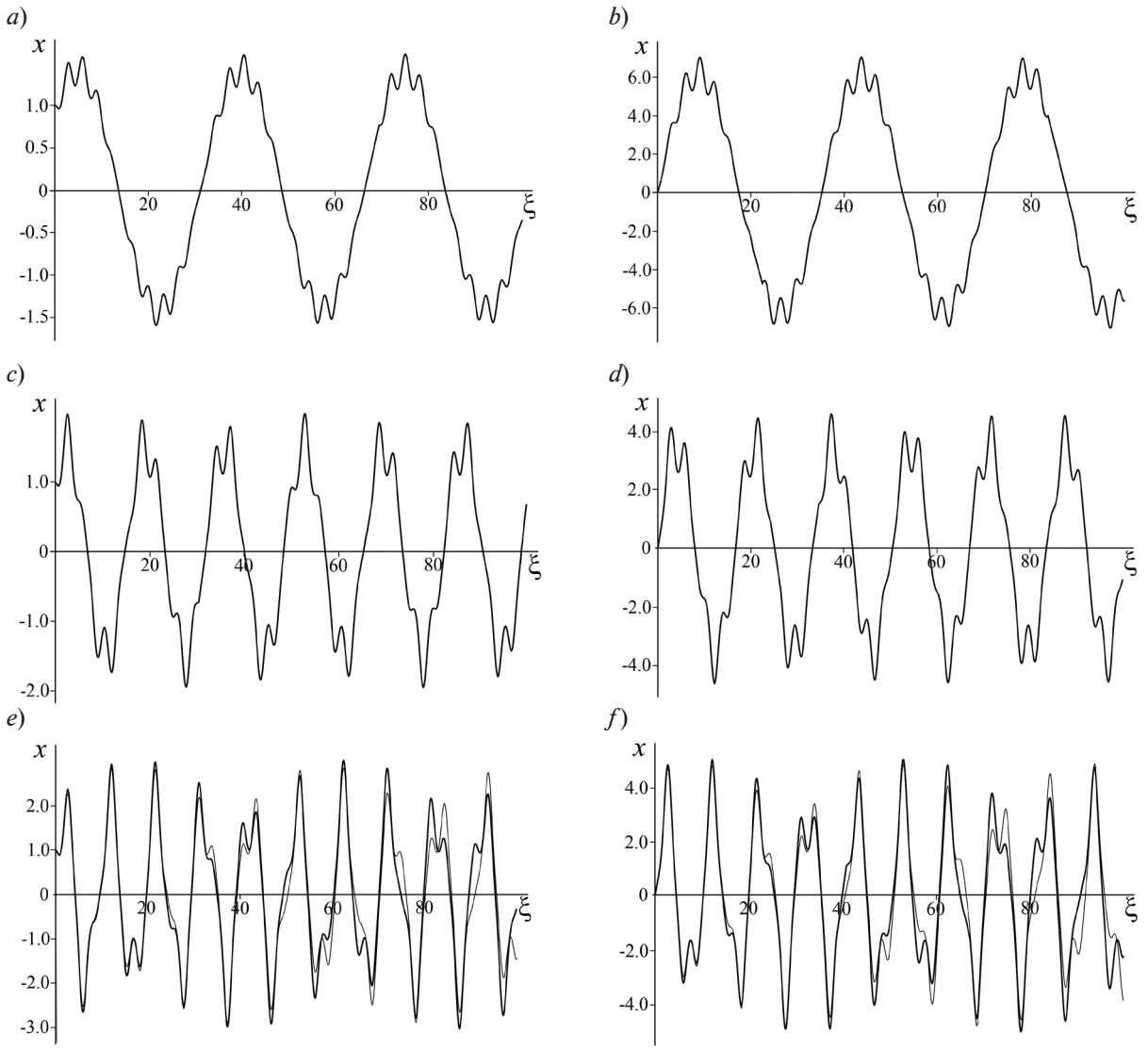

Fig. 2. Comparison of numerically obtained trajectories of Eq. (2) (thin lines) with approximate trajectories calculated by pseudopotential decomposition (17) – (22) up to terms of the form $1/\Omega^{14}$ (solid lines).
The parameter values used are given in Table. Thin and solid lines in Figs. *a* – *d* overlap, so they are visually indistinguishable (in contrast to the curves in Fig. 1)

(18), (20) and including expansion terms up to $1/\Omega^{14}$, with exact (numerical) solutions of system of equations (16). As expected, the accuracy deteriorates rapidly as the parameter q increases (i.e., when approaching the far end of the stability region), so the expressions obtained above are suitable only for moderately high q values (more precisely, only for moderately high secular frequencies $\beta \le 0,62$). Divergence is quite natural when approaching the far end of the stability region corresponding to the secular frequency $\beta = 1$, since the basic assumptions that using the representation of solutions in the form (17) is based on and derivation of the final expressions are not satisfied under parametric resonance between the proper secular oscillations of ions and the forced RF oscillations. The latter are due to external effect of the radio frequency electric field. However, ion trajectories can be calculated with sufficient accuracy using the approximate formulae obtained for the range of secular frequencies $0 \le \beta \le 0,62$.

Eq. (19) yields an improved version of the approximate formula

$$\tilde{\beta}^2 \approx a + q^2/2$$

for the secular oscillation frequency, which is obtained from classical pseudopotential theory:

$$\hat{\beta}^2(a,q) \approx a + \frac{1}{2}q^2 + \frac{1}{2}aq^2 + \\ + \left(\frac{1}{2}a^2q^2 + \frac{25}{128}q^4\right) + \left(\frac{1}{2}a^3q^2 + \frac{273}{512}aq^4\right) + \\ + \left(\frac{1}{2}a^4q^2 + \frac{2049}{2048}a^2q^4 + \frac{1169}{9216}q^6\right) + \ldots . \quad (23)$$

The inequality $0 \le \hat{\beta}^2 \le 1$, or, more precisely, a pair of inequalities

$$\beta_x^2 = \beta^2(a,q) \le 1, \quad \beta_y^2 = \beta^2(-a,-q) \ge 0,$$

with $a \ge 0$, $q \ge 0$, can be used to approximately calculate the boundaries of the first stability region. Notably, while the inequality $\beta_y^2 = \beta^2(-a,-q) \ge 0$ is sufficiently accurate for describing the near end of the first stability region, then the inequality $\beta_x^2 = \beta^2(a,q) \le 1$ at best provides a qualitative description of the far end of the first stability region.

It follows from the data in Fig. 3 that series (23) diverges near the far end of the first stability region, and, therefore, when approaching the far end of the first stability region, reasonable accuracy can only be achieved by using an incredibly large number of terms in the series.

Conclusion

As a result of the study we have carried out, we have found that the concept of a pseudopotential function can be generalized in a completely constructive way for RF quadrupole fields. The purpose of such generalization is in reducing the discrepancy between the exact and analytical solutions obtained in analysis of simplified models of the object under consideration. In this case, exact solutions cannot be obtained analytically. The resulting algebraic expression in the form of a truncated pseudopotential series allows to significantly expand the range of parameters of the RF field. Motion of charged particles within the framework of the traditional pseudopotential approach, which is characterized by conceptual simplicity and physical clarity, can be described not only qualitatively but also quantitatively in this range. Notably, the approaches offered in [21, 22] lack these advantages.

Unfortunately, the concept of the pseudopotential expanded in this way is not particularly suitable for describing the motion of charged particles when approaching the region of parametric

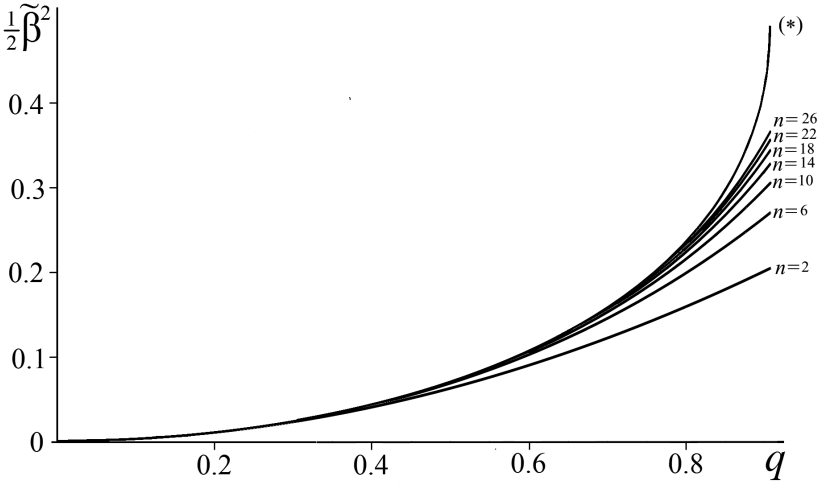

Fig. 3. Quadratic coefficient of pseudopotential (PP) function (23), calculated via PP expansions (17) – (22) with different accuracy orders $1/\Omega^n$ for $n = 2 - 26$ in the range $0 \le q \le 0,9080$ ($a = 0$), as a function of q.

The curve (*) corresponds to the function for an analytically accurate value of secular oscillation frequency (calculated in accordance with [21, 22, 31])

resonance ($\beta \approx 1$), where the motion of charged particles in RF quadrupole fields loses stability. In this case, pseudopotential models [21, 22] that are accurate yet rather cumbersome turn out to be a preferable alternative. The results prove to be acceptable for moderately high secular frequencies lying in the range $0 \le \beta \le 0,62$, while the range of permissible values of the parameter providing acceptable accuracy of calculations is far more narrow ($0 \le \beta \le 0,2$) for the classical pseudopotential theory.

It is recommended to use the exact theory of quadratic pseudopotential for RF quadrupole fields [21, 22], rather than approximate pseudopotential expansions, for large values of secular frequencies.

Acknowledgment

The authors are grateful to the developers, employees and sponsors of the Numdam digital library [37] for providing open access to a rare publication [23].

This study was carried out within the framework of state task no. 007-00229-18-00 for the Institute of Analytical Instrumentation for the Russian Academy of Sciences.

REFERENCES

[1] **L.D. Landau, E.M. Lifshitz,** Mechanics, 2nd ed., Course of theoretical physics, Vol. 1, Pergamon Press, 1969.

[2] **V.A. Gaponov, M.A. Miller,** Potential wells for charged particles in a high frequency electromagnetic field, Journal of Experimental and Theoretical Physics. 7(2) (1958) 242–243.

[3] **M.A. Miller,** Dvizheniye zaryazhennykh chastits v vysokochastotnykh elektromagnitnykh polyakh [The motion of charged particles in the high-frequency electromagnetic fields], Radiophysics and Quantum Electronics. 1 (3) (1958) 110–123.

[4] **A.G. Litvak, M.A. Miller, N.V. Sholokhov,** Utochneniye usrednennogo uravneniya dvizheniya zaryazhennykh chastits v pole stoyachey elektromagnitnoy volny [The refinement of the averaged equation of the motion of charged particles in the field of a standing electromagnetic wave], Radiophysics and Quantum Electronics. 5 (6) (1962) 1160–1174.

[5] **D.V. Sivukhin,** Dreyfovaya teoriya dvizheniya zaryazhennoy chastitsy v elektromagnitnykh polyakh, V kn.: Voprosy teorii plazmy [Drift theory of charged particle motion in the electromagnetic fields, In a book "Plazma theory problems"], Iss. 1, Gosatomizdat, Moscow, 1963, Pp. 7–97.

[6] **A.I. Morozov, L.S. Solovyev,** Dvizheniye zaryazhennoy chastitsy v elektromagnitnykh polyakh, V kn.: Voprosy teorii plazmy, [Charged particle motion in the electromagnetic fields, In a book "Plazma theory problems"], Iss. 2, Gosatomizdat, Moscow, 1963, Pp. 177–261.

[7] **V.I. Geyko, G.M. Fraiman,** Accuracy of the averaged particles in high-frequency fields, Journal of Experimental and Theoretical Physics. 107 (6) (2008) 960–964.

[8] **P.L. Kapitza,** High power electronics, Soviet Physics Uspekhi. 5 (5) (1963) 777–826.

[9] **A.G. Chirkov,** Asimptoticheskaya teoriya vzaimodeystviya zaryazhennykh chastits i kvantovykh sistem s vneshnimi elektromagnitnymi polyami [Asymptotic theory of interaction of charged particles and quantum systems with external electromagnetic fields], St. Petersburg State Polytechnic University, St. Petersburg, 2001.

[10] **R.Z. Sagdeev, D.A. Usikov, G.M. Zaslavsky,** Nonlinear physics: from the pendulum to turbulence and chaos (Ser. "Contemporary Concepts in Physics", Vol. 4), Harwood Academic Publishers, Chur, London, Paris, New York, Melbourne, 1988.

[11] **D. Gerlich,** Inhomogeneous RF fields: a versatile tool for the study of processes with slow ions, In: State-selected and state-to-state ion-molecule reaction dynamics. Part 1: Experiment, advances in chemical physics series, C.-Y. Ng, M. Baer (Eds.), Vol. LXXXII, John Wiley & Sons Inc., New York, 1992, Pp. 1–176.

[12] **M.I. Yavor,** Optics of charged particle analyzers, Academic Press, Amsterdam, 2009.

[13] **G.I. Slobodenyuk,** Kvadrupolnyye mass spektrometry [Quadrupole mass spectrometers], Atomizdat, Moscow, 1974.

[14] **P.H. Dawson,** Quadrupole mass spectrometry and its applications, American Institute of Physics, Woodbury, 1995.

[15] **R.E. March, J.F. Todd,** Quadrupole ion trap mass spectrometry, Ser. "Chemical Analysis", Vol. 165, 2nd Ed., John Wiley and Sons, Hoboken, New Jersey, 2005.

[16] **F.G. Major, V.N. Gheorghe, G. Werth,** Charged particle traps. Physics and techniques of charged particle field confinement, Springer-Verlag, Berlin, Heidelberg, New York, 2005.

[17] **G. Werth, V.N. Gheorghe, F.G. Major,**

Charged particle traps II, Applications, Springer-Verlag, Berlin, Heidelberg, 2009.

[18] **M.Yu. Sudakov, M.V. Apatskaya,** Concept of the effective potential in describing the motion of ions in a quadrupole mass filter, Journal of Experimental and Theoretical Physics. 115 (2) (2012) 194–200.

[19] **M.Y. Sudakov, E.V. Mamontov,** Analysis of the quadrupole mass filter with quadrupole excitation by the envelope equation method, Technical Physics. 2016. 61 (11) (2016) 1715–1723.

[20] **M. Sudakov**, Nonlinear equations of the ion vibration envelope in quadrupole mass filters with cylindrical rods, International Journal of Mass Spectrometry. 422 (2017) 62–73.

[21] **D.J. Douglas., A.S. Berdnikov, N.V. Konenkov,** The effective potential for ion motion in a radio frequency quadrupole field revisited, International Journal of Mass Spectrometry. 377 (2015) 345–354.

[22] **A.S. Berdnikov, D.J. Douglas, N.V. Konenkov,** The pseudopotential for quadrupole fields up to q = 0.9080, International Journal of Mass Spectrometry. 421 (2017) 204–223.

[23] **G. Floquet,** Sur les equations différentielles linéaires à coefficients périodiques, Annales scientifiques de l'École Normale Supérieure, 2e serié. 12 (1883) 47–88.

[24] **G.V. Bondarenko,** Uravneniye Khilla i yego primeneniye v oblasti tekhnicheskikh kolebaniy [The Hill equation and its application in the technical oscillation region], SA USSR, Moscow, Leningrad, 1936.

[25] **N.W. McLachlan,** Theory and application of Mathieu functions, Oxford Univ. Press, Oxford, 1947.

[26] **N.P. Erugin,** Lappo-Danilevskiy metod in the theory of differential equations, Leningrad University Press, Leningrad, 1956.

[27] **N.P. Erugin,** Linear systems of ordinary differential equations with periodic and quasi-periodic coefficients, Academic Press, New York, 1966.

[28] **F.R. Gantmacher,** The theory of matrices, Chelsea Pub. Co., USA, 1960.

[29] **B.P. Demidovich,** Lektsii po matematicheskoy teorii ustoychivosti [The course of lectures on the mathematical theory of stability], Moscow, Nauka, 1967.

[30] **V.A. Jakubovich, V.H. Starzhinskij,** Linear differential equations with periodic coefficients, Wiley, New York, 1975.

[31] **N.V. Konenkov, M. Sudakov, D.J. Douglas,** Matrix methods to calculate stability diagrams in quadrupole mass spectrometry, Journal of American Society for Mass Spectrometry. 13 (6) (2002) 597–613.

[32] **A.S. Berdnikov,** A pseudopotential description of the motion of charged particles in RF fields, Microscopy and Microanalysis. 21 (S4) (2015) 78–83.

[33] **A.L. Bulyanitsa, V.E. Kurochkin,** Studying ordering processes in open systems (on the example of pattern evolution in colonies of imperfect mycelial fungi), Nauchnoye priborostroyeniye. 10 (2) (2000) 43–49.

[34] **A.L. Bulyanitsa, V.E. Kurochkin, D.A. Burylov,** Implementation of the constant signal estimation procedure based on Tsypkin's modification of the stochastic approximation method, Journal of Communications Technology and Electronics. 47 (3) (2002) 307–309.

[35] **A.A. Evstrapov, A.L. Bulyanitsa, G.E. Rudnitskaya, et al.,** Characteristic features of digital signal filtering algorithms as applied to electrophoresis on a microchip, Nauchnoye priborostroyeniye. 13 (2) (2003) 57–63.

[36] **A.L. Bulyanitsa,** Mathematical modeling in microfluidics: basic concepts, Nauchnoye priborostroyeniye. 15 (2) (2005) 51–66.

[37] Numdam, the French digital mathematics library, URL: http://www.numdam.org/.

Received 18.07.2018, accepted 26.07.2018.

THE AUTHORS

BERDNIKOV Alexander S.
Institute for Analytical Instrumentation of the Russian Academy of Sciences
26 Rizhsky Ave., St. Petersburg, 190103, Russian Federation
asberd@yandex.ru

GALL Lidiya N.
Institute for Analytical Instrumentation of the Russian Academy of Sciences
26 Rizhsky Ave., St. Petersburg, 190103, Russian Federation
lngall@yandex.ru

GALL Nikolay R.
Institute for Analytical Instrumentation of the Russian Academy of Sciences
26 Rizhsky Ave., St. Petersburg, 190103, Russian Federation
gall@ms.ioffe.ru

SOLOVYEV Konstantin V.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
k-solovyev@mail.ru

CUMULATIVE PROTONS PRODUCTION DURING THE CARBON NUCLEUS FRAGMENTATION ON THE BERYLLIUM TARGET

D.M. Larionova, M.M. Larionova, Yu. M. Mitrankov, V.S. Borisov, V.N. Solovev, A.Ya. Berdnikov

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation

The collision of carbon nuclei with beryllium targets has been simulated in the framework of the Liège Intranuclear Cascade model at the carbon nuclei initial kinetic energies of 0.60, 0.95, 2.00 GeV / nucleon. Proton production invariant cross-sections at the nuclei collision angle of 3.5 degrees were obtained. It was shown that the dependence of experimental invariant cross-sections on the cumulative variable x in the range $0.9 < x < 2.4$ could be interpret on the basis of taking into account the Fermi motion of nucleons in nuclei, multiple scattering processes, and the formation of delta resonance. The calculation results were compared with experimental data and findings of investigation where data was analyzed in the context of the quark cluster model.

Key words: cumulative particle, delta resonance, Liège Intranuclear Cascade model, beryllium target.

Citation: D.M. Larionova, M.M. Larionova, Yu.M. Mitrankov, V.S. Borisov, V.N. Solovev, A.Ya. Berdnikov, Cumulative protons production during the carbon nucleus fragmentation on the beryllium target, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 48–55. DOI: 10.18721/JPM.11306

Introduction

Cumulative production of particles means that particles are generated in nucleus-nucleus collisions in the region kinematically forbidden for free nucleon-nucleon collisions [1, 2].

A dimensionless quantity x, also called the dimensionless variable, is introduced to characterize the cumulative particles [3].

There are different definitions for this quantity [1, 2]; in this paper, the cumulative variable x is used as a ratio of the momentum p of the detected proton to the momentum p_0 of the nucleon in the carbon nucleus [4], in the laboratory frame of reference (rest frame of the ^{9}Be target):

$$x = p / p_0.$$

Currently, there are two fundamentally different models describing the production of cumulative particles.

The first model takes into account the Fermi motion, multiple scattering within the nucleus as the hadron projectile or its fragmentation products experience several successive rescatterings [1], and the processes associated with the formation of resonances. As a result, a particle can be produced in the last intranuclear collision in a region that is kinematically forbidden in scattering by a single free nucleon.

The second model is based on the processes occurring at distances that are much smaller than the characteristic nuclear distances [5]. The most common models describing such processes are the fluctuon model [5] and the nmodel of nucleon correlations at short distances [6, 7].

The flucton models are divided into two classes: “cold” and “hot”. The former assume that fluctuons always exist in the initial nucleus [1, 5, 8, 9], while according to the latter, fluctons are formed during the collision [10].

The FRAGM experiment at the TWAC-ITEP heavy ion acceleration-accumulation facility at the Institute for Theoretical and Experimental Physics (Moscow, Russia) [11] measured proton yields at an angle of 3.5°

with the fragmentation of carbon ions with energies of 0.60, 0.95, and 2.00 GeV/nucleon on a beryllium target. The obtained data are presented as invariant cross-sections of the proton yield versus the cumulative variable x in the range $0.9 < x < 2.4$.

The experimental data from Ref. [11] were analyzed in [4] based on the quark cluster model [8]. According to this model, the nucleus contains clusters consisting of $3k$ (k = 1, 2, 3) valence quarks. The value $k = 1$ corresponds to ordinary nucleons.

However, Ref. [4] did not take into account the contribution of the processes that are not related to formation of quark clusters, namely, the Fermi motion of nucleons in the nucleus, multiple scattering, and the formation of resonances.

The goal of this study was to calculate the cross-sections for the production of cumulative protons in an inclusive reaction

$$^{12}\mathrm{C} + {}^{9}\mathrm{Be} = {}^{1}p + X, \qquad (1)$$

where ^{1}p is the proton, X is the remaining products of the reaction.

Initial kinetic energies of carbon ions were taken to be 0.60, 0.95, and 2.00 GeV/nucleon. The computational model took into account the Fermi motion of nucleons, multiple rescattering, and the formation of resonances. The hypothesis of quark clusters was not included in the model.

Simulation procedure

We used the Extension of the Liиge Intranuclear Cascade Model [12] to assess the contribution from Fermi motion, multiple scattering and the formation.

According to the Intranuclear Cascade Model, the collision of two nucleons leads either to elastic or to inelastic scattering.

Total cross-sections $\sigma_{tot,pp}$ of nucleon-nucleon scattering in millibarns (mb) were calculated using the following formulae [13, 14]:

$$\sigma^{\mathrm{I}}_{tot,pp} = 34\left(\frac{p_{lab}}{0,4}\right)^{-2,104} \quad \text{with} \quad p_{lab} < 0,44;$$

$$\sigma^{\mathrm{II}}_{tot,pp} = 23,5 + 1000(p_{lab} - 0,7)^4$$

$$\text{with} \quad 0,44 < p_{lab} < 0,80; \qquad (3)$$

$$\sigma^{\mathrm{III}}_{tot,pp} = 23,5 + \frac{24,6}{1 + \exp\left(-\frac{p_{lab} - 1,2}{0,1}\right)}$$

$$\text{with} \quad 0,8 < p_{lab} < 1,5; \qquad (4)$$

$$\sigma^{\mathrm{IV}}_{tot,pp} = 41 + 60(p_{lab} - 0,9)\exp(-1,2p_{lab})$$

$$\text{with} \quad 1,5 < p_{lab} < 3,0; \qquad (5)$$

$$\sigma^{\mathrm{V}}_{tot,pp} = 45,6 - 219p_{lab}^{-4,23} + 0,41\log^2(p_{lab}) - 3,41\log(p_{lab})$$

$$\text{with} \quad p_{lab} > 3,0; \qquad (6)$$

$$\sigma^{\mathrm{I}}_{tot,np} = 6,3555\exp[-3,2481\log(p_{lab}) - 0,377\log^2(p_{lab})]$$

$$\text{with} \quad p_{lab} < 0,446; \qquad (7)$$

$$\sigma^{\mathrm{II}}_{tot,np} = 33 + 196\left|p_{lab} - 0,95\right|^{2,5}$$

$$\text{with} \quad 0,446 < p_{lab} < 1,000; \qquad (8)$$

$$\sigma^{\mathrm{III}}_{tot,np} = 24,2 + 8,9p_{lab}$$

$$\text{with} \quad 1 < p_{lab} < 1,924; \qquad (9)$$

$$\sigma^{\mathrm{IV}}_{tot,np} = 48,9 - 33,7p_{lab}^{-3,08} + 0,619\log^2(p_{lab}) - 5,12\log(p_{lab})$$

$$\text{with} \quad 1,924 < p_{lab}, \qquad (10)$$

where p_{lab}, GeV/c is the momentum in the laboratory frame of reference.

The cross-sections $\sigma_{el,pp}$ for nucleon-nucleon elastic scattering are calculated in the extended model using the following formulae:

$$\sigma^{\mathrm{I}}_{el,pp} = \sigma_{tot,pp} \quad \text{with} \quad p_{lab} < 0,8; \qquad (11)$$

$$\sigma^{\mathrm{II}}_{el,pp} = \frac{1250}{p_{lab} + 50} - 4(p_{lab} - 1,3)^2$$

$$\text{with} \quad 0,8 < p_{lab} < 2,0; \qquad (12)$$

$$\sigma^{\mathrm{III}}_{el,pp} = \frac{77}{p_{lab} + 1,5}$$

$$\text{with} \quad 2,000 < p_{lab} < 3,096; \qquad (13)$$

$$\sigma^{\mathrm{IV}}_{el,pp} = 11,2 - 22,5p_{lab}^{-1,12} + 0,151\log^2(p_{lab}) - 1,62\log(p_{lab})$$

$$\text{with} \quad 2,096 < p_{lab}; \qquad (14)$$

$$\sigma^{I}_{el,np} = \sigma_{tot,np} \text{ with } p_{lab} < 0,85; \quad (15)$$

$$\sigma^{II}_{el,np} = \frac{31}{\sqrt{p_{lab}}} \text{ with } 0,85 < p_{lab} < 2,00; \quad (16)$$

$$\sigma^{III}_{el,np} = \frac{77}{p_{lab} + 1,5} \text{ with } 2,00 < p_{lab}. \quad (17)$$

The cross-sections for the formation of inelastic processes can be calculated as the difference between the total cross-section of nucleon-nucleon scattering and the cross-section of elastic scattering.

Computational study

The results for the simulation of intranuclear cascade are shown in Fig. 1 as the dependence of invariant cross-section (σ_{inv}) for proton production in the given reaction versus the cumulative variable x. The invariant cross-section for proton production was calculated by the formula:

$$\sigma_{inv} = \frac{E}{p_0} \frac{d^2\sigma}{dxd(p_t)^2},$$

where σ is the total cross-section of the reaction; p_0 is the per nucleon incident momentum; E and p_t are the total energy and the transverse momentum of the proton in the laboratory frame of reference [4].

Figs. 1 – 3 show the simulation results without (i.e., exclusively due to Fermi motion and multiple scattering) and with the formation of delta resonances Δ(1232) taken into account. Figs. 1 – 3 also give a comparison of the data for the simulation of intranuclear cascade with the experimental data and the results obtained based on the quark cluster model.

Discussion

It follows from the data shown in Figs. 1 – 3 that simultaneously taking into account Fermi motion, multiple scattering and delta resonance formation leads to production of cumulative particles in the range $x > 1$.

Fig. 4 shows examples of the processes leading to the production of cumulative particles.

Example 1 (Fig. 4,*a*). Let us consider the production of a cumulative particle due to Fermi motion of nucleons in the incident nucleus and

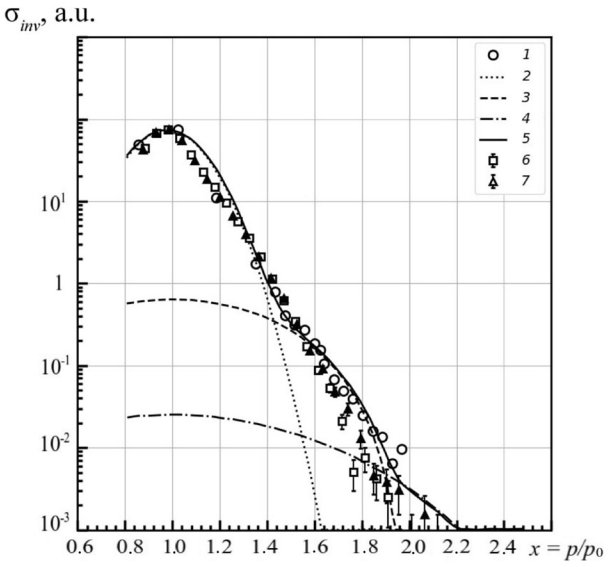

Fig. 1. Experimental (symbols *1*) [11] and simulated (*2* – *7*) curves for the cross-section of proton production in reaction (1) (angle 3.5°) versus the cumulative variable, at an initial energy of 0.60 GeV/nucleon.

The data were processed based on: the Extension of the Liиge Intranuclear Cascade Model [12], without (*6*) and with (*7*) delta resonance formation taken into account; the quark cluster model (*2* – *5*) used to assess the contributions of one- (*2*), two- (*3*) and three- (*4*) nucleon clusters and the total contribution (*5*) of quark clusters

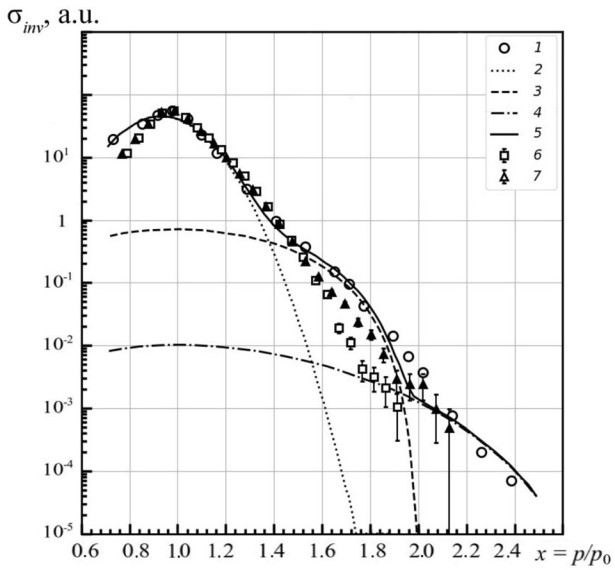

Fig. 2. The data shown are similar to those in Fig. 1 but were obtained with the initial energy of 0.95 GeV/nucleon

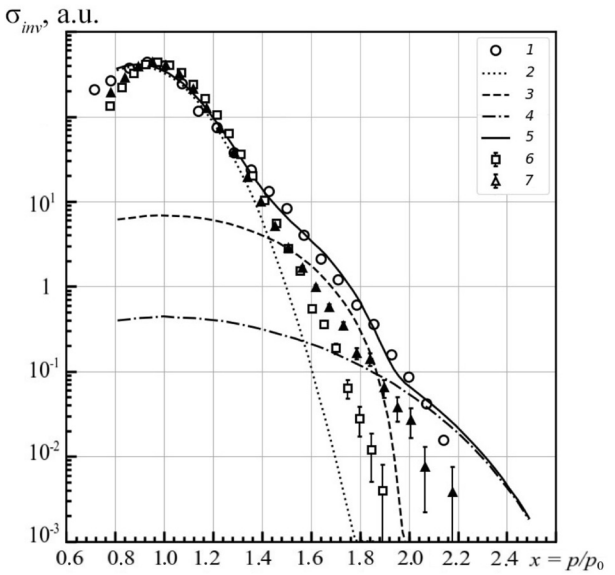

Fig. 3. The data shown are similar to those in Figs. 1 and 2 but were obtained with the initial energy of 2.00 GeV/nucleon

to multiple scattering. A cumulative proton with $x = 1.58$ was detected in this particular example of an event.

The per nucleon incident momentum can exceed the p_0 value due to Fermi motion in the given nucleus. According to the Liège intranuclear cascade model, the momenta of the nucleons in the nucleus obey the Gaussian distribution, for which the root-mean-square (RMS) value of the quantity is expressed as

$$\mathrm{RMS} = \sqrt{\frac{3}{5}} p_{\mathrm{F}},$$

where $p_{\mathrm{F}} = 270$ MeV/c is the Fermi momentum [15].

In the first stage (I) of the given event, proton 0, whose momentum is 1603 MeV/c, elastically collides with neutron 1; as a result,

the momentum of proton 0 decreases to 1337 MeV/c (the proton loses energy due to elastic collision).The second stage (*2*) is the elastic collision of proton 3, whose momentum is 1421 MeV/c, with proton 0, whose momentum is 1337 MeV/c. Due to this collision, the momentum of proton 0 increases to 1925 MeV/c. This proton is the one actually detected in this event as cumulative, and the momentum of 1925 MeV/c corresponds to the value of the cumulative variable $x = 1.58$.

Example 2 (Fig. 4,*b*). Let us consider an event with the formation of a delta resonance. In the first stage (*1*) of such an event, proton 1, whose momentum is 1346 MeV/c, collides with proton 0; as a result, delta resonance 0 with a momentum of 1098 MeV/c is produced. The second stage (2) is the collision of proton 2, whose momentum is 1108 MeV/c, with a delta resonance. After this, proton 0, whose momentum is 1872 MeV/c, is produced, and it is the one detected in this event as cumulative. The momentum of 1872 MeV/c corresponds to the value of the cumulative variable $x = 1.53$.

It follows from expressions (4) and (5) that the cross-section of inelastic processes, including the formation of delta resonances, is zero for per nucleon momenta of carbon nuclei smaller than 0.8 GeV/(c · nucleon).

Thus, the cross-sections obtained by simu-

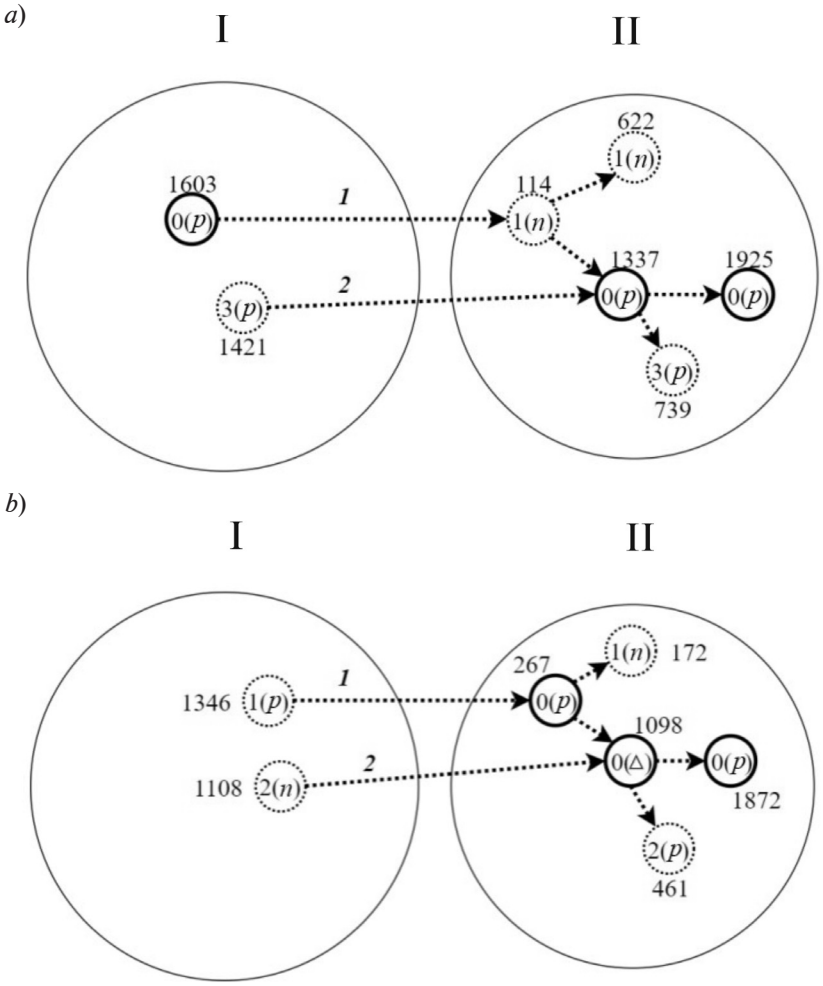

Fig. 4. Examples of events occurring in the collision of C (I) and Be (II) nuclei, without (*a*) and with (*b*) the formation of delta resonance 0(Δ); *1* and *2* are the stages of the processes. Small dashed circles indicate intranuclear nucleons, with their momenta in MeV/c shown beside them; (n) and (p) are the neutron and the proton; 0, 1, 2, 3 are their indices. The small solid circles indicate the cumulative particles formed in the processes

lation, both with and without the formation of delta resonances taken into account, coincide in the region $x < 0.8 / 0.6 \approx 1.3$.

If the formation of delta resonances in the region $x > 1.6$ is taken into account, the invariant cross-section increases and becomes closer to the experimental values.

Let us compare the results of the simulation carried out in our study with the predictions of the hypothesis based on the existence of quark clusters in nuclei (see Figs. 1 – 3).

Evidently, the processes of multiple scattering and formation of delta resonances in the region $x < 1.4$ are as adequate for describing the experimental data as the quark clusters approach, but yield lower values of invariant cross-sections in the region $x > 1.4$. The obtained results start to considerably deviate from the experimental data with increasing initial kinetic energies of carbon ions.

Conclusion

We have obtained the cumulative variable distributions of invariant cross-sections taking into account the processes of multiple scattering and resonance formation, without using the hypothesis of quark clusters in inclusive reaction (1)

$$^{12}C + {}^{9}Be = {}^{1}p + X$$

with initial kinetic energies of carbon ions of 0.60, 0.95, 2.00 GeV/nucleon.

We have established that the processes of multiple scattering and delta resonance formation lead to production of cumulative particles and make a significant contribution to the cross-section for the production of cumulative particles. The obtained results are in agreement with the experimental data for the initial kinetic energy of carbon ions of 0.60 GeV/nucleon.The obtained values are lower than the experimental values with increasing energy in the region x > 1.4, which indicates potential new mechanisms for the production of cumulative particles, for example, taking into account other nucleon resonances.

The study was carried out with the financial support of the Ministry of Education and Science of the Russian Federation, state task 3.1498.2017/4.6.

REFERENCES

[1] **A.V. Efremov,** Quark-parton picture of the cumulative production, Physics of Elementary Particles and Atomic Nuclei. 13(3) (1982) 613–634.

[2] **V.S. Stavinsky,** Limiting fragmentation of nuclei – the cumulative effect, Physics of Elementary Particles and Atomic Nuclei. 10(5) (1979) 949–995.

[3] **V.S. Stavinskij,** Unique algorithm for calculation of inclusive cross sections of particle production with big transverse momenta and of cumulative Type Hadrons, JINR Rapid Communications. 5 (18) (1986) 5–17.

[4] **B.M. Abramov, P.N. Alekseev, Yu.A. Borodin, et al.,** Manifestation of quark clusters in the emission of cumulative protons in the experiment of the fragmentation of carbon ions, JETP Letters. 97(8) (2013) 439 –443.

[5] **D.I. Blokhintsev,** On the fluctuations of nuclear matter, JETP.6(5) (1958) 995–999.

[6] **A. Tang, J.W. Watson, J. Aclander, et al.,** *n–p* Short-range correlations from (*p*, 2*p* + *n*) measurements, Physical Review Letters. 90 (4) (2003) 042301.

[7] **A. Tang, J. Alster, G. Asryan, et al**., *n–p* Short-range correlations from (*p*, 2*p* + *n*) measurements, AIP Conf. Proc. 549 (1) (2000) 451–454.

[8] **A.V. Efremov, A.B. Kaidalov, V.T. Kim, et al.,** Cumulative hadron production in quark models of flucton fragmentation, Sov. J. Nucl. Phys. 47 (5) (1988) 1364–1374.

[9] **M.A. Braun, V.V. Vechernin,** Cumulative phenomena in the QCD approach, Nuclear Physics, B, Proc. Suppl. 92 (1–3) (2001) 156–161.

[10] **A. Motornenko, M.I. Gorenstein,** Cumulative production of pions by heavy baryonic resonances in proton-nucleus collisions, arXiv:1604.04308v1[hep-ph] 14 Apr. 2016. Pp. 1–21.

[11] **B.M. Abramov, Yu.A. Borodin, S.A. Bulychev, et al.,** Cumulative protons in the ^{9}Be (^{12}C, *p*) reaction at 0.2 – 3.2 Gev/nucleon, Bull. Russ. Acad. Sci.: Physics. 75 (4) (2011) 500–504.

[12] **S. Pedoux,** Extension of the Liège intranuclear cascade model to the 2 –15 GeV incident energy range, PhD thesis, University of Liège, Liège (2011–2012).

[13] **Th. Aoust,** Amelioration du modèle de cascade intranucleaire de Liège en vue de l'ètude de cibles de spallation pour les systèmes hybrids, PhD thesis, University of Liège, Liège (2006–2007).

[14] **A. Baldini, V. Flaminio, W.G. Moorhead, D.R.O. Morrison,** Numerical data and functional relationships in science and technology, New

series Group 1, No. 12, Springer-Verlag, Berlin, Heidelberg, 1988.

[15] **D. Mancusi, A. Boudard, J. Cugnon, et al.,** Extension of the Liège intranuclear-cascade model to reactions induced by light nuclei, Phys. Rev. C. 90 (5) (2014) 054602.

Received 05.06.2018, accepted 07.06.2018.

THE AUTHORS

LARIONOVA Dariya M.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
dlar@bk.ru

LARIONOVA Mariya M.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
mlario@bk.ru

MITRANKOV Yuriy M.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
mitrankovy@gmail.com

BORISOV Vladislav S.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
vl-borisof@yandex.ru

SOLOVEV Vladimir N.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
hydraca39@gmail.com

BERDNIKOV Aleksandr Ya.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
alexber@phmf.spbstu.ru

NOTES ON USING THE PRINCIPAL COMPONENTS IN THE MATHEMATICAL SIMULATION

Yu.A. Pichugin

Saint-Petersburg State University of Aerospace Instrumentation,
St. Petersburg, Russian Federation

The paper discusses the issues related to the use of principal components analysis (PCA) in mathematical simulation. The paper significantly expands the range of the solved problems using PCA. In particular, the solutions of the following three tasks are given: (*i*) structural similarity and homogeneity estimation for random Gaussian vectors; (*ii*) recovery of missing data; (*iii*) the forecast of non-stationary time series based on the caterpillar method, which is a generalization of PCA for non-stationary time series. To solve the problems, to restore missing data and to predict the data, the author offers an unbiased estimation of the variance of the error of the regression on the PCs base for the cases of large and small samples. All the main statements are formulated in the form of theorems proved by the author.

Key words: principal component, variance of the regression error, small sample volume

Citation: Yu.A. Pichugin, Notes on using the principal components in the mathematical simulation, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 56–69. DOI: 10.18721/JPM.11307

Introduction

Principal component analysis (PCA) is a well-known procedure of mathematical statistics. This method was proposed by Pearson in 1901 [1] and consists in the following.

If the coordinate system of an *n*-dimensional space is properly rotated (by orthogonal transformation) so that the coordinate axes coincide with the principal axes of the concentration ellipsoid, then the components of the normally distributed *n*-dimensional centered vector are uncorrelated and, by virtue of the normal distribution law, independent.

In the algebraic sense, this actually means that a covariance matrix is converted to diagonal form by orthogonal transformation, and a quadratic form in the exponent of the density function of the multidimensional normal distribution is converted to canonical form. The well-known Karhunen-Loиve transformation [2, 3] is, in fact, precisely this coordinate transformation. Transition to independent PC variables typically allows to significantly reduce the dimension of the given problem with minimal loss of information

In view of this, PCs are often derived as a solution to the optimization problem in the literature, although all of their optimal properties are quite evident from the very spectrum of the covariance matrix (the spectrum shows the fraction of the variance that is discarded, see below).

PCA is known as Singular Spectrum Analysis (SSA) in time series analysis, where it is used to solve the problem of redundancy in classical spectral analysis [4 – 6]. SSA is peculiar in that the dimension of the vector is equal to N in this case, and the dimension of the mutual covariance matrix is N Ч N (N is the length of the given time series). The elements of the covariance matrix are calculated by a special technique, with the divisor equal to N regardless of the magnitude of the shift, or, respectively, of the number of terms. Even though such estimates are obviously biased, by themselves they do not produce a skewed

(towards overestimation, as Jenkins and Watts note in [7]) wavelength in spectral analysis. It was explained in [8] that the problem of high dimension of the covariance matrix in SSA is easily solved by applying the von Mises iterations, since all rows of the covariance matrix of the time series can be obtained from the first row by shifting, duplicating and rearranging the elements. The eigenvalues and eigenvectors of the covariance matrix are obtained by applying a sequence of simple iterations without rotation to a matrix of dimension N Ч N.

Alternatives to SSA are the so-called caterpillar method [9], and the method proposed later in this paper (see the sections "Problem of a relatively small sample" and "Forecast of non-stationary time series"), where the forecast scheme is constructed based on the caterpillar method.

Many methods are close to PCA, for example, the components in Independent Component Analysis (ICA) may obey Student's, Cauchy and Dirichlet distribution besides the Gaussian. Notice that the ICA method is also known as analysis of these components.

The method of principal curves and manifolds is a generalization of PCA. Recently, PCA has been widely used for visualization and graphical representation of multidimensional data (a projection of the sample on the plane of the first two principal axes was considered [10, 11]). The initial data do not have to obey the normal distribution in this case.

Here, there is a wide array of essentially similar methods, such as multidimensional scaling, nonlinear mapping, projection pursuit, as well as methods of neural network problems, such as the bottleneck method, Kohonen self-organizing maps, etc. Notably, graphical representation of multidimensional data by projection onto the plane of the first two principal axes of PCs allows to obtain a fairly good initial approximation of the sampling distribution in the solution of the classification problem in [12].

The goal of this study has been to expand the range of problems solved using principal component analysis.

In view of this goal, the study considers the problems of structural similarity analysis, missing data recovery, and forecast of non-stationary series. We have refined the details of the method of principal components directly related to problems of recovering the missing data and forecasting. The issues of reduced dimension and visualization of multidimensional data are treated as secondary in this study.

Brief overview of the techniques used for PCA

It is assumed that the vector $\mathbf{y}$ has the dimension m (dim$\mathbf{y}$ = m) and obeys the multidimensional normal distribution, i.e., $\mathbf{y} \sim N(\boldsymbol{\theta}_y, \mathbf{V}_y)$.

Let $\mathbf{P}$ be an orthogonal matrix such that

$$\mathbf{P}^T\mathbf{V}_y\mathbf{P} = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$$

$$\text{and } \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m,$$

where T is the transposition symbol (operator).

Recall that the columns of the matrix $\mathbf{P}$ are the eigenvectors of the matrix $\mathbf{V}_y$, and the set of eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ is called the spectrum of this matrix. The columns of the matrix $\mathbf{P}$ are called the basis of principal components in mathematical statistics, and the components of the vector $\mathbf{P}^T(\mathbf{y} - \boldsymbol{\theta}_y)$. are called the principal components. The parameters of the distribution $N(\boldsymbol{\theta}_y, \mathbf{V}_y)$, are typically unknown in applications. If there is a sample $\mathbf{y}_j$, $j = 1, 2, \ldots, n$, we can calculate unbiased estimates of unknown parameters:

$$\hat{\boldsymbol{\theta}}_y = \frac{1}{n}\sum_{j=1}^{n}\mathbf{y}_j,$$
$$\hat{\mathbf{V}}_y = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{y}_j - \hat{\boldsymbol{\theta}}_y)(\mathbf{y}_j - \hat{\boldsymbol{\theta}}_y)^T. \quad (1)$$

In this case, we take the orthogonal matrix $\hat{\mathbf{P}}$, as P, which reduces the estimate $\hat{\mathbf{V}}_y$ to a diagonal form, i.e.,

$$\hat{\mathbf{P}}^T\hat{\mathbf{V}}_y\hat{\mathbf{P}} = \hat{\boldsymbol{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m)$$

$$\text{and } \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_m. \quad (2)$$

Obviously, the matrices $\mathbf{V}_y$ and $\hat{\mathbf{V}}_y$ are not equal, and therefore, the matrices $\mathbf{P}$ and $\hat{\mathbf{P}}$ are not equal as well.

The problem of reducing the dimension, like the other problems considered in this paper, is directly related to testing the following hypotheses:

$$\mathrm{H}: \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_m. \quad (3)$$

Adopting this hypothesis allows to consider a vector of lower dimension

$$\widehat{\mathbf{P}}^T_{(k)}(\mathbf{y} - \widehat{\boldsymbol{\theta}}_y),$$

where the matrix $\widehat{\mathbf{P}}_{(k)}$ contains only the first k columns of matrix $\mathbf{P}$.

The first known procedure for testing the hypothesis H (see formula (3)) was Bartlett's test [13, 14] (see, for example, monograph [15] or handbook [16]). However, while Bartlett's test is described absolutely correctly in [15], two inaccuracies (a sign and a multiplier) are immediately found in [16].

Indeed, the χ^2 statistics for testing hypothesis H in [15] is expressed as

$$\gamma_\eta = n' \left\{ (m-k) \ln \left(\frac{1}{m-k} \sum_{i=k+1}^{m} \widehat{\lambda}_i \right) - \ln \left(\prod_{i=k+1}^{m} \widehat{\lambda}_i \right) \right\},$$

where

$$n' = n - k - \frac{1}{6}\left[2(m-k) + 1 + \frac{2}{(m-k)} \right],$$

while in [16] this statistic follows the expression

$$\gamma_\eta = (n-1) \left\{ (m-k) \ln \left(\frac{1}{m-k} \sum_{i=k+1}^{m} \widehat{\lambda}_i \right) + \ln \left(\prod_{i=k+1}^{m} \widehat{\lambda}_i \right) \right\}.$$

Both publications cite the same number of degrees of freedom η.

In Ref. [15]:

$$\eta = \frac{1}{2}(m-k+2)(m-k-1);$$

In Ref. [16]:

$$\eta = \frac{1}{2}(m-k+1)(m-k) - 1.$$

Nowadays, the so-called Broken Stick Model has become widely known (see, for example, Ref. [17]). According to this test, the maximum value of k, for which the inequality

$$\frac{\widehat{\lambda}_k}{\mathrm{tr}\widehat{\mathbf{V}}_y} > \frac{1/k + 1/(k+1) + \ldots + 1/m}{m},$$

where tr is the matrix trace, holds true should be chosen.

The simplest means of finding k is visual analysis of the graphical representation of the covariance matrix spectrum in descending order. In this case, the value that precedes the change of the relatively fast decline in eigenvalues to a relatively slow (smooth) one is taken as k, which essentially repeats the Broken Stick method (at an intuitive, non-formalized level). Many handbooks and textbooks propose to determine k from the ratio

$$\left(\sum_{i=1}^{k} \widehat{\lambda}_i \Big/ \sum_{i=1}^{m} \widehat{\lambda}_i \right) 100\% \approx K\%,$$

where K is the predetermined percentage of the total variance.

The choice of the method for determining k ultimately depends on the nature of the problem being solved. Let us outline the results.

Obviously, an error occurs in reducing the dimension, and the dispersion of this error must be somehow related to the rejected part of the sample spectrum. The following two sections of the paper are dedicated to the solution of this problem.

A priori variance estimate for errors of regression on PCs

The regression of the vector on PCs (on the components of the vector $\mathbf{z}$) is a relation expressed by an equation of the form

$$\mathbf{y} = \boldsymbol{\theta}_y + \mathbf{P}_{(k)}\mathbf{z} + \boldsymbol{\varepsilon}. \quad (4)$$

According to the general principles of classical regression analysis, it is assumed that

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \quad (5)$$

where $\mathbf{0}$ is a zero vector and $\mathbf{I}$ is the unit matrix of the corresponding dimension.

Assumption (5) means that

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2 \quad (i = 1, 2, \ldots, m)$$

$$\text{and} \quad \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \; (i \neq j),$$

where E, var and covare the the expectation, variance, and covariance operators, respectively.

Suppose that some implementation of the vector $\mathbf{y}$ is considered in the above-described conditions. This may be one of

the implementations used for calculating the estimates of the parameters $(\hat{\boldsymbol{\theta}}_y, \hat{\mathbf{V}}_y)$, i.e. $\mathbf{y}_j$, $(j = 1, 2, \ldots, n)$, or one of the subsequent $\mathbf{y}_{n+l}$, $(l \geq 1)$, which is of no consequence. Therefore, the subscript can be omitted for now. In practical simulation problems we have, instead of Eq. (4), a regression model of the form

$$\mathbf{y} = \hat{\boldsymbol{\theta}}_y + \hat{\mathbf{P}}_{(k)}\mathbf{z} + \boldsymbol{\varepsilon}. \qquad (6)$$

First of all, we should note that, according to classical regression analysis, Eq. (6) should be called regression on the PC basis, since the PCs themselves (the components of vector $\mathbf{z}$) are the parameters defined at the stage when the constructed model (6) is applied. Regression on PCs appears in the proof of the following theorem.

Theorem 1. *Given the conditions of model (6), where the elements of the model are calculated by formulae (1), (2) provided that hypothesis* H *is correct, or, in other words, accepted (see expression (3)) and with assumption (5), an a priori unbiased estimate* σ^2 *is expressed as*

$$\hat{\sigma}^2 = \frac{n-1}{(m-k)(n-k-1)} \sum_{i=k+1}^{m} \hat{\lambda}_i. \qquad (7)$$

Proof. Let us move on to covariances in regression model (4) which relates three normally distributed centered vectors $(\mathbf{y} - \boldsymbol{\theta}_y$, $\mathbf{z}$ and $\boldsymbol{\varepsilon})$. Then we have the following equality:

$$\mathbf{V}_y = \mathbf{P}_{(k)}\mathbf{V}_z\mathbf{P}_{(k)}^T + \mathbf{V}_\varepsilon, \qquad (8)$$

where

$$\mathbf{V}_\varepsilon = \sigma^2\mathbf{I}, \quad \mathbf{V}_z = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2),$$

$$\sigma_i^2 = \mathrm{var}(z_i), \ (i = 1, 2, \ldots, k).$$

Multiplying equality (8) by $\mathbf{P}^T$ on the left-hand side and by $\mathbf{P}$ on the right-hand side, we obtain:

$$\begin{aligned} &\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k, \lambda_{k+1}, \ldots, \lambda_m) = \\ &= \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2, 0, \ldots, 0) + \\ &\quad + \mathrm{diag}(\sigma^2, \sigma^2, \ldots, \sigma^2), \end{aligned} \qquad (9)$$

i.e.,

$$\lambda_i = \sigma_i^2 + \sigma^2 \ (i = 1, 2, \ldots, k),$$

$$\lambda_i = \sigma^2 \ (i = k+1, k+2, \ldots, m).$$

Notably, not just equality (9) holds true only with if hypothesis H is correct (see (3)), but assumption (5) is also possible for a large sample only if hypothesis H is correct. This is evident from equality (9), where $\lambda_i = \sigma^2$. However, this logic is violated for a small sample (see below).

Equality (9) also implies that

$$\sigma^2 = \frac{1}{(m-k)} \sum_{i=k+1}^{m} \lambda_i. \qquad (10)$$

Let us consider a classical regression model [18]:

$$\begin{aligned} x_j &= \beta_0 + \beta_1 z_{j,1} + \beta_2 z_{j,2} + \ldots \\ &\ldots + \beta_{p-1} z_{j,p-1} + \varepsilon_j \ (j = 1, 2, \ldots, n). \end{aligned} \qquad (11)$$

For an arbitrarily chosen i $(i = 1, 2, \ldots, m)$, let us substitute the initial $y_{i,j}$, values into the set of equations (11) instead of x_j and instead of $z_{j,l}$ $(l = 1, 2, \ldots, k;\ k = p-1)$ let us substitute the values of sample PCs calculated by the formula

$$(\hat{z}_{j,1}, \hat{z}_{j,2}, \ldots, \hat{z}_{j,k}) = (\mathbf{y}_j - \hat{\boldsymbol{\theta}}_\mathbf{y})^T \hat{\mathbf{P}}_{(k)},$$

where $\mathbf{y}_j$ is the j-th implementation of $\mathbf{y}$ in the initial sample.

In the latter case, we keep using our previous notations. These are PCs of the original sample, but we interchange the indices (numbers) of the component and the implementation (l and j), thus complying with the standards of regression analysis, i.e., model (11). In this context, these PCs values

$$\{\hat{z}_{j,1}, \hat{z}_{j,2}, \ldots, \hat{z}_{j,k}\}$$

(6). are assumed to be known and model (11) here is indeed a regression on PCs. Then the Eq. (11) corresponds to the ith row of matrix equality (6).

It follows from the condition

$$\bar{z}_l = \frac{1}{n} \sum_{j=1}^{n} z_{j,l} = 0 \quad (l = 1, 2, \ldots, k)$$

that

$$\hat{\beta}_0 = \bar{x},$$

where $\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$, i.e., $\hat{\beta}_0 = \frac{1}{n} \sum_{j=1}^{n} y_{i,j}$ (see above).

PCA, as well as the ordinary least squares method (OLS), ensures that the residual sum of squares (RSS) is minimal, therefore, the vector row of OLS estimates

$$(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_{p-1})$$

matches the i-th row of the matrix $\hat{\mathbf{P}}_{(k)}$, i.e.,

$$(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_{p-1}) = [\hat{\mathbf{P}}_{(k)}]_i,$$

where $[\hat{\mathbf{P}}_{(k)}]_i$ is the ith row of the matrix $\hat{\mathbf{P}}_{(k)}$.

This statement is easily proved by direct calculation. Let us define the following matrices of initial data:

$$\mathbf{Y} = (\mathbf{y}_1 - \hat{\boldsymbol{\theta}}, \mathbf{y}_2 - \hat{\boldsymbol{\theta}}, ..., \mathbf{y}_n - \hat{\boldsymbol{\theta}}),$$

$$\mathbf{Z} = \hat{\mathbf{P}}_{(k)}^T \mathbf{Y}.$$

Then the equivalent of model (6) for these matrices can be written in the form

$$\mathbf{Y} = \hat{\mathbf{P}}_{(k)} \mathbf{Z} + \mathbf{E},$$

where $\mathbf{E}$ is the $(m \times n)$-matrix of all regression residuals.

Accordingly, the equivalent of (11) is written in the form

$$\mathbf{Y}^T = \mathbf{Z}^T \hat{\mathbf{P}}_{(k)}^T + \mathbf{E}^T.$$

Next, we need to verify the equality

$$\hat{\mathbf{P}}_{(k)}^T = (\mathbf{Z}\mathbf{Z}^T)^{-1} \mathbf{Z}\mathbf{Y}^T,$$

which actually means that the matrix $\hat{\mathbf{P}}_{(k)}$, and, accordingly, all its rows are essentially OLS estimates.

Substituting the expression for the matrix $\mathbf{Z}$ (see above), multiplying by $\hat{\mathbf{P}}_{(k)} \hat{\mathbf{P}}_{(k)}^T$ on the right-hand side and placing additional brackets, we obtain an obvious identity:

$$(\hat{\mathbf{P}}_{(k)}^T \hat{\mathbf{P}}_{(k)}) \hat{\mathbf{P}}_{(k)}^T =$$
$$= (\hat{\mathbf{P}}_{(k)}^T \mathbf{Y}\mathbf{Y}^T \hat{\mathbf{P}}_{(k)})^{-1} (\hat{\mathbf{P}}_{(k)}^T \mathbf{Y}\mathbf{Y}^T \hat{\mathbf{P}}_{(k)}) \hat{\mathbf{P}}_{(k)}^T,$$

because $\hat{\mathbf{P}}_{(k)}^T \hat{\mathbf{P}}_{(k)} = \mathbf{I}$.

Multiplying by $\hat{\mathbf{P}}_{(k)} \hat{\mathbf{P}}_{(k)}^T$ is correct because the rank of the matrices does not decrease in this case:

$$\operatorname{rank}(\hat{\mathbf{P}}_{(k)} \hat{\mathbf{P}}_{(k)}^T) = \operatorname{rank} \hat{\mathbf{P}}_{(k)} = k.$$

The assumptions we made and equality (10) imply that RSS S^2 for regression model (11) obeys the expression

$$S^2 = \frac{n-1}{(m-k)} \sum_{i=k+1}^{m} \hat{\lambda}_i.$$

In accordance with the theory of linear regression [18], the estimate

$$\tilde{\sigma}^2 = \frac{S^2}{n-p} = \frac{n-1}{(m-k)(n-k-1)} \sum_{i=k+1}^{m} \hat{\lambda}_i$$

is unbiased (recall that $p = k + 1$).

Theorem 1 is proved.

We should note that estimate (7) and a brief description of the proof of Theorem 1 outlining the main idea (see formula (9)) were proposed in an earlier study [19]. It is actually the biased estimate σ^2 that is required in some problems on assessing the informativeness (see [20]). In this case, it is preferable to use the estimate

$$\tilde{\sigma}^2 = \frac{n-1}{(m-k)n} \sum_{i=k+1}^{m} \hat{\lambda}_i,$$

which follows directly from formula (10). We are going to further discuss using a priori estimates below (see the section "Recovering missing data").

The problem of a relatively small sample

The situation when the sample size n is less than the vector dimension m ($n < m$) is common for many problems. In this case,

$$\hat{\lambda}_{n+1} = \hat{\lambda}_{n+2} = ... = \hat{\lambda}_m = 0,$$

which does not at all correspond to equality (9), and, consequently, (10). However, this does not exclude the possibility of testing the hypothesis

$$\mathrm{H}_1: \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k \geq \lambda_{k+1} = \lambda_{k+2} = ... = \lambda_n \quad (12)$$

for further consideration of model (6).

Assuming the vector $\hat{\boldsymbol{\theta}}_y$ and the matrix $\hat{\mathbf{P}}_{(k)}$ to be known, for any of the subsequent implementations of the vector $\mathbf{y}$, for example the $(n + l)$th $(l \geq 1)$, the unbiased a posteriori σ^2 estimate in regression model (6) follows the expression

$$\tilde{\sigma}^2 = \frac{1}{m-k} \sum_{i=1}^{m} (y_{i,n+l} - \bar{y}_i - [\hat{\mathbf{P}}_{(k)}]_i \hat{\mathbf{z}}_{n+l})^2, \quad (13)$$

where $\bar{y}_i$ is the ith component of the a priori estimate of the vector $\boldsymbol{\theta}_y$ $(i = 1, 2, ..., m)$, and the number of estimated parameters is equal to k.

On the other hand, assuming only the matrix $\hat{\mathbf{P}}_{(k)}$ to be known, this estimate is as follows:

$$\hat{\sigma}^2 = \frac{1}{m-k-1}\sum_{i=1}^{m}(y_{i,n+l} - \bar{y}_{n+l} - [\hat{\mathbf{P}}_{(k)}]_i \hat{\mathbf{z}}_{n+l})^2, \quad (14)$$

where $\bar{y}_{n+l} = \frac{1}{m}\sum_{i=1}^{m} y_{i,n+l}$; $y_{i,n+l}$ is the ith component of the vector $\mathbf{y}_{n+l}$.

In these formulae, the same as above, in both cases, corresponding to formulae (13) and (14), $[\hat{\mathbf{P}}_{(k)}]_i$ is the ith row of the matrix $\hat{\mathbf{P}}_{(k)}$, and

$$\hat{\mathbf{z}}_{n+l} = \hat{\mathbf{P}}_{(k)}^T(\mathbf{y}_{n+l} - \hat{\boldsymbol{\theta}}_y).$$

Nevertheless, the vector $\hat{\boldsymbol{\theta}}_y$ in expression (13) is calculated by formula (1) using all the implementations of the initial sample, while in (14) we substitute all components of $\hat{\boldsymbol{\theta}}_y$with the mean value for the components of the new implementation $\mathbf{y}_{n+l}$ ($\theta_1 = \theta_2 = ...=\theta_m = \bar{y}_{n+l}$, see above). However, the vector $\hat{\boldsymbol{\theta}}_y$ is used for calculating $\hat{\mathbf{V}}_y$ and $\hat{\mathbf{P}}$, which should be taken into account in the a priori estimate σ^2. In view of this, let us consider a different method for calculating the elements of model (6).

Let us define the matrix $\mathbf{Y}$ in a different manner. Let

$$\mathbf{Y} = (\mathbf{y}_1 - \hat{\boldsymbol{\theta}}_1, \mathbf{y}_2 - \hat{\boldsymbol{\theta}}_2, ..., \mathbf{y}_n - \hat{\boldsymbol{\theta}}_n),$$

where the components of each vector $\hat{\boldsymbol{\theta}}_j$ are equal to each other and equal to the component mean of the implementation $\mathbf{y}_j$, i.e.,

$$\theta_{1j} = \theta_{2j} = ... = \theta_{mj} = \bar{y}_j = \frac{1}{m}\sum_{i=1}^{m} y_{i,j},$$
$$(j = 1, 2, ..., n).$$

Let us calculate the estimate $\mathbf{V}_y$ by formula

$$\hat{\mathbf{V}}_y = (n-1)^{-1}\mathbf{Y}\mathbf{Y}^T. \quad (15)$$

Next, let us take all of the above-describe steps: calculate $\hat{\mathbf{P}}$ and $\{\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_n\}$; check the hypothesis H_1 (see (12), $n < m$) and define the matrix $\hat{\mathbf{P}}_{(k)}$. Not only $\hat{\mathbf{V}}_y$ and $\hat{\mathbf{P}}$ but the value of k can slightly change in this case.

Theorem 2. *Given the conditions of model (6), where the elements of the model are calculated by scheme (15) provided that hypothesis* H_1 *is correct, or, in other words, accepted (see expression (12)) and with assumption (5), an a priori unbiased estimate* σ^2, *equal to*

$$\hat{\sigma}^2 = \frac{n-1}{(m-k-1)n}\sum_{i=k+1}^{n}\hat{\lambda}_i, \quad (16)$$

is unbiased.

Proof. Since the expected value operator is linear, the mean value of unbiased estimate (14), calculated from the initial sample ($j = 1, 2, ..., n$), is actually the unbiased a priori estimate sough for, equal to

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{m-k-1}(\mathbf{y}_j - \hat{\boldsymbol{\theta}}_j)^T \times$$
$$\times (\mathbf{I} - \hat{\mathbf{P}}_{(k)}\hat{\mathbf{P}}_{(k)}^T)(\mathbf{y}_j - \hat{\boldsymbol{\theta}}_j) =$$
$$= \frac{1}{(m-k-1)n}\mathrm{tr}(\mathbf{Y}^T(\mathbf{I} - \hat{\mathbf{P}}_{(k)}\hat{\mathbf{P}}_{(k)}^T)\mathbf{Y}) =$$
$$= \frac{1}{(m-k-1)n}(\mathrm{tr}\mathbf{Y}^T\mathbf{Y} - \mathrm{tr}\mathbf{Z}^T\mathbf{Z}) =$$
$$= \frac{n-1}{(m-k-1)n}\sum_{i=k+1}^{n}\hat{\lambda}_i$$

(here, the same as above, $\mathbf{Z} = \hat{\mathbf{P}}_{(k)}^T\mathbf{Y}$).

Theorem 2 is proved.

Notice that estimate (16) requires only the expected values of estimates (14) but not the expected values of mean $\hat{\boldsymbol{\theta}}_j$ estimates to coincide. This means that estimate (16) is suitable for the case of time series containing trends. In turn, this (aside from the relatively small sample) may be another reason use estimates (15) and (16). Comparing estimates (7) and (16), we can also see that in case of a small sample ($n < m$), incorrect use of (7) yields an overestimated value of σ^2. The biased equivalent of estimate (16) is the estimate

$$\tilde{\sigma}^2 = \frac{n-1}{mn}\sum_{i=k+1}^{n}\hat{\lambda}_i.$$

The index j often corresponds to a certain time count in mathematical modeling; this means that estimate (16) can be accepted without resorting to the above-described procedure for adjusting model (6), provided that the mean values of the components of vector $\mathbf{y}$, calculated by the first method (see estimates (1)), are sufficiently stationarity differ insignificantly. Errors and inaccuracies arising from this are negligible. Conversely, the above-described procedure for estimating $\hat{\mathbf{V}}_y$ should be used in the non-stationary case.

This implies recalculating $\hat{\boldsymbol{\theta}}_y$ (see above) by the implementation of $\mathbf{y}$ explicitly appearing in model (6).

The method for estimating the elements of model (6) considered in this section is most effective when the caterpillar method is applied to forecasting non-stationary time series with a pronounced trend (see the section "Forecasting of non-stationary time series").

Important remark. Before we can discuss further considerations, we should focus on a point that is very important for general understanding. In the case of a non-degenerate distribution and a large sample, we can regard the main components as regressors, and the elements of the matrix $\hat{\mathbf{P}}_{(k)}$, which is the basis of the principal components, as estimated parameters (see the proof for Theorem 1). On the other hand, in case of a small sample (the dimension is larger than the size), the elements of the matrix $\hat{\mathbf{P}}_{(k)}$ have to be regarded as regressors, and the principal components as the estimated parameters (Theorem 2). Centering should be carried out by subtracting not the mean implementation value for each component, but the mean component value for each implementation (see above). The gist of the problem is that if the sample is small, equality (9) does not have a sample equivalent, since the sample spectrum is not complete and the logic of Theorem 1 collapses. Therefore, the situation has to be considered from a different standpoint.

Finally, we should note that the case of a fundamentally degenerate distribution, when the sample spectrum is incomplete and the sample is large, brings us to the computational scheme of this section and estimate (16).

Minimal risk estimates

The formula

$$\tilde{\mathbf{y}} = \hat{\boldsymbol{\theta}}_y + \hat{\mathbf{P}}_{(k)}\hat{\mathbf{z}},$$

where $\hat{\mathbf{z}} = \hat{\mathbf{P}}_{(k)}^T(\mathbf{y} - \hat{\boldsymbol{\theta}}_y)$, is typically used if model (6) is chosen for use in applied problems.

Handbook [21] suggests the formula

$$\tilde{\mathbf{y}} = \hat{\boldsymbol{\theta}}_y + \hat{\mathbf{P}}_{(k)}\mathbf{G}\hat{\mathbf{z}}, \qquad (17)$$

where $\mathbf{G} = \mathrm{diag}(g_1, g_2, \ldots, g_k)$.

In this case the values of g_i are determined from the condition of the minimum quadratic risk

$$R^2 = \mathrm{E}(z_i - g_i\hat{z}_i)^2 \quad (i = 1, 2, \ldots, k).$$

In this case, the components of the vector $\mathbf{G}\hat{\mathbf{z}}$ are called minimum risk estimates. Since

$$z_i - g_i\hat{z}_i = z_i(1 - g_i) - g_i(\hat{z}_i - z_i),$$

we obtain the equality

$$R^2 = z_i^2(g_i - 1)^2 + g_i^2\sigma_{\tilde{z}i}^2, \qquad (18)$$

where $\mathrm{E}(\hat{z}_i - z_i)^2 = \sigma_{\tilde{z}i}^2$.

Equating the derivative R^2 with respect to g_i to zero, we obtain the equality

$$g_i(z_i^2 + \sigma_{\tilde{z}i}^2) = z_i^2, \qquad (19)$$

or $g_i = \dfrac{1}{1 + \sigma_{\tilde{z}i}^2/z_i^2}$.

Substituting the minimum risk estimate $g_i\hat{z}_i$, instead of z_i we obtain a simple quadratic equation of the form

$$g_i^2 - g_i + \delta_i^2 = 0, \quad \delta_i^2 = \sigma_{\tilde{z}i}^2/\hat{z}_i^2$$

with the following final formula g_i:

$$g_i = \frac{1}{2} + \sqrt{\frac{1}{4} - \delta_i^2}.$$

Handbook [21] suggests the following algorithm for calculating g_i :

$$\begin{cases} g_i = \dfrac{1}{2} + \sqrt{\dfrac{1}{4} - \delta_i^2}, \text{ if } \delta_i^2 \le \dfrac{1}{4}; \\ g_i = 0, \text{ if } \delta_i^2 > \dfrac{1}{4}. \end{cases} \qquad (20)$$

An alternative algorithm was proposed in [22]:

$$\begin{cases} g_i = \dfrac{1}{2} + \sqrt{\dfrac{1}{4} - \delta_i^2}, \text{ if } \delta_i^2 \le \dfrac{1}{4}; \\ g_i = \dfrac{1}{2}, \text{ if } \dfrac{1}{4} < \delta_i^2 \le 1; \\ g_i = \dfrac{1}{1 + \delta_i^2}, \text{ if } \delta_i^2 > 1. \end{cases} \qquad (21)$$

Theorem 3. *Assuming that the difference between the quantities* $\sigma_{\tilde{z}i}^2/\hat{z}_i^2$ *and* $\sigma_{\tilde{z}i}^2/z_i^2$ *is negligible, the proposed algorithm (21) provides a lower value of the quadratic risk* R^2*, compared with algorithm (20), for the case* $\delta_i^2 > \dfrac{1}{4}$.

Proof. If $\delta_i^2 > \dfrac{1}{4}$, algorithm (20) yields the

value $R^2 = z_i^2$. If $\frac{1}{4} < \delta_i^2 \le 1$, we substitute the value $g_i = \frac{1}{2}$ in (18). Since $\sigma_{\tilde{z}i}^2 < z_i^2$, we obtain that

$$R^2 = \frac{1}{4}(z_i^2 + \sigma_{\tilde{z}i}^2) < \frac{z_i^2}{2} < z_i^2.$$

On the other hand, if $\delta_i^2 > 1$, then, according to algorithm (21), the value of g_i is such that equality (19) holds true. Substituting (19) to (18), we obtain

$$R^2 = g_i^2(z_i^2 + \sigma_{\tilde{z}i}^2) - 2g_i z_i^2 + z_i^2 =$$
$$= g_i z_i^2 - 2g_i z_i^2 + z_i^2 = z_i^2(1 - g_i) < z_i^2.$$

Theorem 3 is proved.

The result obtained in this section consists in clarification of details. Using PCs in studies of multi-dimensional processes and phenomena, especially natural ones, does not necessarily have reduction of dimension as the ultimate goal, but may be aimed at analyzing the internal structure of the phenomenon, as discussed in the next section.

Structural similarity and homogeneity

Let us consider, besides the vector **y**, a vector **x** that has the same dimension:

$$\dim \mathbf{x} = \dim \mathbf{y} = m.$$

Suppose there is a sample of implementations of this vector. Samples of implementations of the vectors **x** and **y** may have different sizes. Using formulae (1) for the vector **x**, we can calculate estimates for the parameters of the distributions $\hat{\mathbf{e}}_x$ and $\hat{\mathbf{V}}_x$. Let the orthogonal matrix $\hat{\mathbf{Q}}$ be such that the following equality holds true:

$$\hat{\mathbf{Q}}^T \hat{\mathbf{V}}_x \hat{\mathbf{Q}} = \mathrm{diag}(\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_m), \qquad (22)$$

$$\hat{\mu}_1 \ge \hat{\mu}_2 \ge ... \ge \hat{\mu}_m. \qquad (23)$$

In this situation, we have two sets of statistical characteristics of PCs of both vectors:

$$\{\hat{\lambda}_i, \hat{\mathbf{p}}_i\}_{i=1}^m \text{ и } \{\hat{\mu}_i, \hat{\mathbf{q}}_i\}_{i=1}^m,$$

where $\hat{\mathbf{q}}_i$, $\hat{\mathbf{p}}_i$ are the ith columns of the matrices $\hat{\mathbf{Q}}$ and $\hat{\mathbf{P}}$, respectively, i.e.,

$$\hat{\mathbf{Q}} = (\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, ..., \hat{\mathbf{q}}_m),$$

$$\hat{\mathbf{P}} = (\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, ..., \hat{\mathbf{p}}_m).$$

The estimate of the coefficient of structural similarity between vectors **x** and **y** is defined as the quantity

$$\hat{s}_{xy} = \frac{\sum_{i=1}^m \sqrt{\hat{\mu}_i \hat{\lambda}_i}\, |\hat{\mathbf{q}}_i^T \hat{\mathbf{p}}_i|}{\sqrt{\mathrm{tr}\hat{\mathbf{V}}_x \mathrm{tr}\hat{\mathbf{V}}_y}}. \qquad (24)$$

This coefficient indicates the extent to which the structures of oscillations of the given vectors agree in relative fractions of variance. In some cases it is preferable to calculate $\hat{s}_{xy}$ by the formula

$$\hat{s}_{xy} = \max_{\varphi(i)} \frac{\sum_{i=1}^m \sqrt{\hat{\mu}_{\varphi(i)} \hat{\lambda}_i}\, |\hat{\mathbf{q}}_{\varphi(i)}^T \hat{\mathbf{p}}_i|}{\sqrt{\mathrm{tr}\hat{\mathbf{V}}_x \mathrm{tr}\hat{\mathbf{V}}_y}}, \qquad (25)$$

where $\varphi(i)$ is the permutation of the indices, i.e., the order of the statistical characteristics of PCs of one of the vectors (here it is **x**) is varied.

We may have to apply formula (25) if close eigenvalues are found in the spectrum of at least one of the vectors. If the following hypothesis is true (tested):

$$\mathrm{H}_x\colon \mu_1 \ge \mu_2 \ge ... \ge \mu_l \ge \mu_{l+1} = \mu_{l+2} = ... = \mu_m,$$

then we can consider the filtered coefficient of structural similarity in the form

$$\hat{s}_{xy}^f = \frac{\sum_{i=1}^p \sqrt{\hat{\mu}_i \hat{\lambda}_i}\, |\hat{\mathbf{q}}_i^T \hat{\mathbf{p}}_i|}{\sqrt{\sum_{i=1}^p \hat{\mu}_i \sum_{i=1}^p \hat{\lambda}_i}},$$

where $p = \min(k, l)$,

or the relative coefficient of structural similarity in the form

$$\hat{s}_{xy}^r = \frac{\sum_{i=1}^p \sqrt{\hat{\mu}_i \hat{\lambda}_i}\, |\hat{\mathbf{q}}_i^T \hat{\mathbf{p}}_i|}{\sqrt{\mathrm{tr}\hat{\mathbf{V}}_x \mathrm{tr}\hat{\mathbf{V}}_y}}.$$

If the neighboring eigenvalues differ only slightly, it is preferable to use formulae similar to formula (25) to find the estimates of the coefficients $\hat{s}_{xy}^f$ and $\hat{s}_{xy}^r$.

Different studies use the coefficient of structural similarity to compare meteorological, climatic and oceanographic fields, as well as to analyze the fields of environmental and health monitoring of different regions. In mi-

croelectronics, when several types of microelectronic devices are manufactured in each cell of a crystal plate (see, for example, [23]), the coefficient of structural similarity is convenient for comparing the manufacturing errors of different devices and determining whether this error depends on the position of a cell on a crystal plate. Similar issues may arise in the process of equipment tuning. In time series analysis, the coefficient of structural similarity is applied based on singular spectrum analysis (SSA) mentioned in the Introduction. The author's first attempts to build the coefficient of structural similarity were made for time series analysis [24].

If a sample of values of the same vector **y** is used as a sample of the vector **x**, and the second set of characteristics of PCs $\{\hat{\mu}_i, \hat{\mathbf{q}}_i\}_{i=1}^m$ is calculated from the estimate of the correlation matrix $\hat{\mathbf{R}}_y$, i.e.,

$$\hat{\mathbf{Q}}^T \hat{\mathbf{R}}_y \hat{\mathbf{Q}} = \mathrm{diag}(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_m),$$

then the coefficient of structural similarity, in all its variants, becomes the coefficient of homogeneity of the vector **y**. In time series analysis, it can be used only in combination with the caterpillar method (see above), where normalization can affect the forms of the eigenvectors, since the forms of eigenvectors of the autocovariance and autocorrelation matrices are actually different.

A natural question is whether the hypothesis that the coefficient of structural similarity equals zero, i.e., H_s: $s_{xy} = 0$, can be tested. First of all, we should note that in practice we can only obtain the estimate $\hat{s}_{xy}$, while the true value s_{xy} could only be obtained if we used the matrices $\mathbf{V}_y$, $\mathbf{V}_x$, $\mathbf{P}$ and $\mathbf{Q}$ in our calculations, which is only hypothetically possible.

Let us consider two composite vectors:

$$\mathbf{Y}_s = (\sqrt{\hat{\lambda}_1}\hat{\mathbf{p}}_1^T, \sqrt{\hat{\lambda}_2}\hat{\mathbf{p}}_2^T, \ldots, \sqrt{\hat{\lambda}_m}\hat{\mathbf{p}}_m^T)^T = (Y_1, Y_2, \ldots, Y_M)^T;$$

$$\mathbf{X}_s = (\sqrt{\hat{\mu}_1}\hat{\mathbf{q}}_1^T, \sqrt{\hat{\mu}_2}\hat{\mathbf{q}}_2^T, \ldots, \sqrt{\hat{\mu}_m}\hat{\mathbf{q}}_m^T)^T = (X_1, X_2, \ldots, X_M)^T,$$

where $M = m \times m$.

If necessary, the numbering of subvectors

$$\sqrt{\hat{\mu}_i}\hat{\mathbf{q}}_i \ (i = 1, 2, \ldots, m)$$

of the composite vector $\mathbf{X}_s$ can be set in accordance with formula (25). This can violate only condition (23), which does not matter to us. The signs of some columns of the matrix

$$\hat{\mathbf{Q}} = (\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \ldots, \hat{\mathbf{q}}_m)$$

have to be changed to opposite ones so that all products $\hat{\mathbf{q}}_i^T\hat{\mathbf{p}}_i$ or $\hat{\mathbf{q}}_{\varphi(i)}^T\hat{\mathbf{p}}_i$ are positive. This change of signs does not violate equality (22). Then, if all these conditions are satisfied, it follows from the orthogonality of the matrices $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ that

$$\hat{s}_{xy} = \frac{\mathbf{X}_s^T\mathbf{Y}_s}{\sqrt{\mathbf{X}_s^T\mathbf{X}_s\mathbf{Y}_s^T\mathbf{Y}_s}} = \frac{\sum_{i=1}^M X_i Y_i}{\sqrt{\sum_{i=1}^M X_i^2 \sum_{i=1}^M Y_i^2}}. \quad (26)$$

Let us consider two regression equations relating the components of the composite vectors $\mathbf{Y}_s$ and $\mathbf{X}_s$:

$$Y_i = \beta X_i + \varepsilon_i \ ; \quad (27)$$

$$Y_i = b_0 + b_1 X_i + e_i \ (i = 1, 2, \ldots, M),$$

where it is assumed that the residuals (errors) of each of the regressions are normally distributed and mutually independent and have the same variance.

Theorem 4. *If the hypothesis* H_ε: $\mathrm{E}(\varepsilon) = 0$ *(or an equivalent hypothesis* H_0: $b_0 = 0$*) is accepted (not rejected), then, given that the hypothesis* H_s: $s_{xy} = 0$ *is correct, the quantity*

$$\frac{\hat{s}_{xy}\sqrt{M-1}}{\sqrt{1-\hat{s}_{xy}^2}} \sim t_{M-1}, \quad (28)$$

i.e., follows Student's distribution with the number of degrees of freedom $M - 1$.

The methods for testing the hypotheses H_0: $b_0 = 0$ and H_ε: $\mathrm{E}(\varepsilon) = 0$ are well-known. Therefore, let us consider a simple proof that differs little from the well-known one for the ordinary correlation coefficient.

Proof. It follows from general theory of the least squares method and formula (26) that the OLS estimate of regression parameter (27) has the form

$$\hat{\beta} = \frac{\sum_{i=1}^M X_i Y_i}{\sum_{i=1}^M X_i^2} = \hat{s}_{xy}\frac{\sqrt{\sum_{i=1}^M Y_i^2}}{\sqrt{\sum_{i=1}^M X_i^2}}, \quad (29)$$

$$\mathrm{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{M} X_i^2},$$

where $\sigma^2 = \mathrm{var}(\varepsilon_i)$, and

$$\sum_{i=1}^{M} X_i \varepsilon_i = 0. \quad (30)$$

It follows from the assumption $\mathrm{E}(\varepsilon) = 0$ (see the hypothesis H_ε) and equality (30) that the unbiased estimate σ^2 is expressed as

$$\hat{\sigma}^2 = \frac{1}{M-1}\sum_{i=1}^{M}(Y_i - \hat{\beta}X_i)^2 = \\ = \frac{1}{M-1}\left(\sum_{i=1}^{M} Y_i^2 - \hat{\beta}^2 \sum_{i=1}^{M} X_i^2\right). \quad (31)$$

It is obvious that the equality $s_{xy} = 0$ is equivalent to $\beta = 0$ (see (29)). Therefore, taking into account expression (31), we have:

$$t_{M-1} \sim \frac{\hat{\beta}\sqrt{\sum_{i=1}^{M} X_i^2}}{\hat{\sigma}} = \frac{\hat{s}_{xy}\sqrt{\sum_{i=1}^{M} Y_i^2}\sqrt{M-1}}{\sqrt{\sum_{i=1}^{M} Y_i^2 - \hat{\beta}^2 \sum_{i=1}^{M} X_i^2}} = \\ = \frac{\hat{s}_{xy}\sqrt{M-1}}{\sqrt{1-\hat{s}_{xy}^2}}. \quad (33)$$

Theorem 4 is proved.

As noted above, changing all signs in the matrix $\tilde{\mathbf{Q}}$ does not violate formula (22) but changes the sign of $\hat{s}_{xy}$ and the sign of (28) to the opposite. This is consistent with the symmetry of Student's distribution.It is evident that the hypothesis H_s can be also tested for the filtered coefficient of structural similarity s_{xy}^f. The only difference is that in this case $M = m \times p$, and it also follows from $s_{xy}^f = 0$ that $s_{xy}^r = 0$. It is impossible to test the hypothesis H_s (the hypothesis H_ε is rejected) but it is not a fundamental obstacle to calculating $\hat{s}_{xy}$. In most applications, we are actually more interested in rejecting the hypothesis H_s, that is, in establishing structural similarity. The exception is the case when we wish to establish the inhomogeneity of a certain field or time series in the above-described manner.

Let us now discuss some practical problems directly using the results of the sections "A priori estimate of the variance of regression error on PCs", "TProblem of relatively small sample" and "Minimum risk estimates".

Recovery of missing data

Let us consider the situation when the distribution is non-degenerate and the sample size is sufficiently larger than the dimension. To reduce the number of indices and other signs, let us rewrite equation (6) in the form

$$\mathbf{y} = \bar{\mathbf{y}} + \mathbf{F}\mathbf{z} + \boldsymbol{\varepsilon}, \quad (32)$$

where $\bar{\mathbf{y}} = \hat{\boldsymbol{\theta}}_y$, $\mathbf{F} = \hat{\mathbf{P}}_{(k)}$.

Let $\mathbf{y}_{n+l}$ $(l \geq 1)$ be some implementation of the vector $\mathbf{y}$, which has u dimensions and v gaps $(m = u + v)$.

Then, with the corresponding numbering, we have the following partitions into blocks:

$$\mathbf{y}_{n+l} = (\mathbf{y}_1^T, \mathbf{y}_2^T)^T, \ \bar{\mathbf{y}} = (\bar{\mathbf{y}}_1^T, \bar{\mathbf{y}}_2^T)^T, \\ \mathbf{F} = (\mathbf{F}_1^T, \mathbf{F}_2^T)^T,$$

where the blocks $\mathbf{y}_1$, $\bar{\mathbf{y}}_1$ and $\mathbf{F}_1$ correspond to the measured components $\mathbf{y}_{n+l}$, and the blocks $\mathbf{y}_2$, $\bar{\mathbf{y}}_2$ and $\mathbf{F}_2$ to the missing data.

Let us calculate the estimate for $\mathbf{y}_2$ by the formula

$$\hat{\mathbf{y}}_2 = \bar{\mathbf{y}}_2 + \mathbf{F}_2\mathbf{G}(\mathbf{F}_1^T\mathbf{F}_1)^{-1}\mathbf{F}_1^T(\mathbf{y}_1 - \bar{\mathbf{y}}_1), \quad (33)$$

where is the matrix $\mathbf{G}$ (see formula (17)) has the dimension $k \times k$, and its components are calculated by algorithm (21).

In this case, in algorithm (21),

$$\sigma_{\hat{z}i}^2 = \hat{\sigma}^2[(\mathbf{F}_1^T\mathbf{F}_1)^{-1}]_{i,i},$$

where $[...]_{i,j}$ is the operator for taking the matrix element from the given row (i) and column (j); $\hat{z}_i$ is the ith component of the vector $\hat{\mathbf{z}}$ of OLS assessments:

$$\hat{\mathbf{z}} = (\mathbf{F}_1^T\mathbf{F}_1)^{-1}\mathbf{F}_1^T(\mathbf{y}_1 - \bar{\mathbf{y}}_1) \ \ (i = 1, 2, ..., k).$$

Due to missing rows, the columns of the matrix $\mathbf{F}_1$ are not orthogonal, the matrix $\mathbf{F}_1^T\mathbf{F}_1$ is not diagonal, and the components of the vector $\hat{\mathbf{z}}$ are dependent on each other:

$$\mathbf{V}_{\hat{z}} = \sigma^2(\mathbf{F}_1^T\mathbf{F}_1)^{-1}.$$

In this case, the a posteriori estimate has the form (see expression (13)):

$$\hat{\sigma}^2 = \frac{1}{u-k}(\mathbf{y}_1 - \bar{\mathbf{y}}_1)^T \times \\ \times (\mathbf{I} - \mathbf{F}_1(\mathbf{F}_1^T\mathbf{F}_1)^{-1}\ \mathbf{F}_1^T)(\mathbf{y}_1 - \bar{\mathbf{y}}_1).$$

However, this is the situation where it is natural to use a more reliable a priori estimate (7) computed from a substantially larger sample. This estimate is used in all of the following formulae in this section.

It should be noted here that $\hat{\mathbf{y}}_2$ is, in essence, the center of the conditional distribution $\mathbf{y}_2$. Indeed, the unconditional or a priori distribution $\mathbf{y}_2$ is $N(\bar{\mathbf{y}}_2, \hat{\mathbf{V}}_2)$ in some approximation (with a sufficiently large n), where $\hat{\mathbf{V}}_2$ is obtained from $\hat{\mathbf{V}}_y$ by eliminating rows and columns that are not related to $\mathbf{y}_2$. The conditional or a posteriori distribution $\mathbf{y}_2$ is also $N(\hat{\mathbf{y}}_2, \hat{\sigma}^2\mathbf{I})$ only in some approximation, which follows immediately from relation (5) and equations (32) and (33).

A relatively accurate Student's distribution can be given for v missing components of the vector $\mathbf{y}$:

$$\frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \mathbf{f}_i\mathbf{G}(\mathbf{F}_1^T\mathbf{F}_1)^{-1}\mathbf{G}\mathbf{f}_i^T}} \sim t_\eta, \qquad (34)$$

where the number of degrees of freedom $\eta = n - k - 1$, if estimate (7) is used and $\eta = m - k - 1$, if estimate (16) is used for some reason; $\mathbf{f}_i$ is the ith row of $\mathbf{F}$, $(i = u+1, u+2, \ldots, m;$ i.e., row $\mathbf{f}_i$ corresponds to block $\mathbf{F}_2)$.

According to formula (34) and the chosen confidence level $1 - \alpha$ (where α is the significance level) for these values, we obtain the boundaries of $(1-\alpha)\%$ Student's confidence intervals:

$$\hat{y}_i \pm t_\eta^{\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \mathbf{f}_i\mathbf{G}(\mathbf{F}_1^T\mathbf{F}_1)^{-1}\mathbf{G}\mathbf{f}_i^T}, \qquad (35)$$

where $t_\eta^{\alpha/2}$ is Student's quantile since Student's t-test is a two-tailed criterion ($\alpha/2$ is taken because Student's distribution is symmetrical).

The efficiency of using model (32) and formula (33) is the higher, the more significant the inequality

$$t_\eta\hat{\sigma}\sqrt{1 + \frac{1}{n} + \mathbf{f}_i\mathbf{G}(\mathbf{F}_1^T\mathbf{F}_1)^{-1}\mathbf{G}\mathbf{f}_i^T} < t_{n-1}\sqrt{\left(1 + \frac{1}{n}\right)[\hat{\mathbf{V}}_y]_{i,i}}, \qquad (36)$$

i.e., the narrower the confidence interval constructed by formula (35), compared with the corresponding interval constructed by the sample estimates of the distribution parameters $(\hat{\mathbf{V}}_y)$, i.e., the less the uncertainty for the data recovered.

Forecasting non-stationary time series

Model (32) and formula (33) can be used to forecast non-stationary time series [25, 26]. Here, model (32) is constructed by the caterpillar method, where $\mathbf{y}$ is a moving segment (caterpillar) of the time series $\{y_i\}_{i=1}^N$, and $\dim\mathbf{y} = m$ is the length of the caterpillar.

The sample for calculating the estimate of the covariance matrix is constructed by a stepwise shift, i.e.,

$$\mathbf{y}_1 = (y_1, y_2, \ldots, y_m)^T,$$
$$\mathbf{y}_2 = (y_2, y_3, \ldots, y_{m+1})^T,$$
$$\ldots\ldots\ldots\ldots\ldots,$$
$$\mathbf{y}_n = (y_{N-m+1}, y_{N-m+2}, \ldots, y_N)^T,$$

where $n = N - m + 1$, N is the total length of the time series.

The estimate of the covariance matrix should be calculated by the algorithm for calculating estimate (15). The final value of m for determining the caterpillar's length (preliminary calculations will be very useful in this case) should be taken no less than the period of the wave carrying the largest part of the variance, i.e., corresponding to $\hat{\lambda}_1$. Trying to fulfill this condition can lead to the situation of small sample size $(n < m)$. It was mentioned in the section "The problem of a relatively small sample" that using estimates (15) and (16) may be due to two reasons: small sample size or non-stationarity.

The authors of the caterpillar method [9] follow the classical calculation scheme for estimating the covariance matrix (see sections "Brief Description of the PCA Mathematical Apparatus" and "Problem of Relatively Small Sample") and incorrectly determine the number of degrees of freedom by the minimum size of the sample matrix of initial data (whether $\eta = n - k - 1$ or $\eta = m - k - 1$, is selected (see above) is determined by $\min(m, n)$). The rows and columns of the initial data matrix are of the same nature in the caterpillar method, which makes the duality described at the end of the

"Problem of a relatively small sample" section all the more obvious. However, the caterpillar method in [9] was used mainly for filtering time series, and the question about the number of degrees of freedom was not as crucial as in the case of forecasting by the scheme considered here.

Forecasting $\mathbf{y}_1$ (in formula (33)) includes a vector of the last u values of the time series:

$$\mathbf{y}_1 = (y_{N-u+1}, y_{N-u+2}, \ldots, y_N)^T,$$

($u < m$, because $m = u + v$, see «Recovery of the missing data»), and $\mathbf{y}_2$ is the vector of the forecast values:

$$\mathbf{y}_2 = (y_{N+1}, y_{N+2}, \ldots, y_{N+v})^T.$$

It follows from the above (see the section "Problem of a Relatively Small Sample") that estimate (16) should be used as the a priori estimate σ^2. The true meaning of forecasting is not so much in calculating the values of $\hat{\mathbf{y}}_2$, as in constructing a sufficiently narrow confidence interval (see above) for the components of $\mathbf{y}_2$. In this case formulae (35) and (36) contain $\eta = m - k - 1$, and the value $1/m$ instead of $1/n$, if the components of the vector $\bar{\mathbf{y}}$ equal to each other (see the "Problem of a relatively small sample" section) are calculated as the mean over the last m values of the time series, or $1/u$, if the mean over the last u values (over the components) is calculated, which is quite acceptable. In any case, the final choice of the parameters of the forecasting scheme is determined by inequality (36).

Conclusion

Having analyzed the existing methods for solving the problems based on the principal components and the proposed modification of these methods, we can formulate the following results.

1. We have obtained estimates for the variance of regression error on the principal components are obtained for cases of a large and a small (relative to the dimension of the problem) sample and have proved that these estimates are unbiased. The estimates obtained are an important part of the schemes for the methods for recovering missing data and forecasting non-stationary series proposed in this paper. The condition that the estimates be unbiased is necessary for constructing confidence intervals for recovered or predicted values (see below).

2. We have theoretically substantiated the previously known minimal risk estimates, also used in the above-discussed practical tasks.

3. We have introduced the coefficient of structural similarity and theoretically substantiated the statistics for testing the hypothesis that this coefficient equals zero.

4. We have proposed schemes for recovering missing data and forecasting non-stationary series. We have found the validity criteria and confidence intervals for reconstructed or predicted values.

As a last consideration we should mention that in constructing statistical models, we had to choose which elements of the model should be given the property of statistical stability and included in the model, and which ones should not. The success of applying the constructed models in practice largely depends on the adequacy of this choice, perhaps even more so than on the accuracy of the formulae used. We have assumed in this study that the basis of principal components is actually the most statistically stable part of the model.

We hope that the estimates obtained here and the solutions to the problems pesented will find practical application in a wide range of subject areas.

REFERENCES

[1] **K. Pearson,** On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Ser. 6. 2 (11) (1901) 559–572.

[2] **K. Karhunen,** Über lineare Methoden in der Wahrscheinlichkeitsrechnung, Ann. Acad. Sci. Fennicae, Ser. A, I. Math.-Phys. 1947 (37) (1947) 1–79.

[3] **M. Loève,** Probability theory, Vol. II, 4th ed., Graduate texts in mathematics, Springer-Verlag. 46 (1978).

[4] **D.S. Broomhead, G.P. King,** Extracting qualitative dynamics from experimental data, Physica. D: Nonlinear Phenomena. 20 (2–3) (1986) 217–236.

[5] **D.S. Broomhead, G.P. King,** On the

qualitative analysis of experimental dynamical systems, Nonlinear Phenomena and Chaos, S. Sarkar. (Ed.), CRC Press, Bristol (1986) 113–144.

[6] **M. Ghil, R. Vautard,** Interdecadal oscillations and the warming trend in global temperature time series, Nature. 350 (6316) (1991) 324–327.

[7] **G.M. Jenrins, D.G. Watts,** Spectral analysis and its aplications, Holden-Day, San Fracisco – Cambridge – London – Amsterdam, 1969.

[8] **Yu.A. Pichugin,** Iterative singular-spectrum analysis in estimating natural cyclicities in meteorological observation data, Meteorology and Hydrology. 10 (2001) 34–39.

[9] Glavnyye komponenty vremennykh ryadov: metod "Gusenitsa" [Principal components of time series: Caterpillar method], D.L. Danilov, A.A. Zhiglyavskiy (Eds.), SPbSU, St. Petersburg, 1997.

[10] The transform and data compression handbook, K. Rao, P. Yip (Eds.), CRC Press LLC, Boca Raton, USA, 2001.

[11] **D.D. Muresan, T.W. Parks,** Adaptive principal components and image denoising, Proceedings of IEEE International Conference on Image Processing (ICIP), 14–17 Sept. 1 (2003) I-101– I-104.

[12] **Yu.A. Pichugin,** Classification of summer weather regions in St. Petersburg, Meteorology and Hydrology. 5 (2000) 31–39.

[13] **M.S. Bartlett,** The effect of standardization on a χ^2 approximation in factor analysis, Biometrika. 38 (3–4) (1951) 337–344.

[14] **M.S. Bartlett,** A note on the multiplying factor for various 2 approximations, J. Roy. Statist. Soc. B16 (1954) 296–298.

[15] **D.N. Lawley, A.E. Maxwell,** Factor analysis as a statistical method, Butterworths, London, 1963.

[16] **S.A. Ayvazyan, V.M. Bukhshtaber, I.S. Enyukov, L.D. Meshalkin,** Prikladnaya statistika. Klassifikatsiya i snizheniye razmernosti [Applied Statistics. Classification and dimension reduction], Finansy i statistika, Moscow, 1989.

[17] **D. Jackson,** Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches, Ecology. 74 (8) (1993) 2204–2214.

[18] **G.A.F. Seber,** Linear regression analysis, John Wiley & Sons, New York, London, Sydney, Toronto (1977).

[19] **Yu.A. Pichugin**, The problem of statistical control of observation data on surface temperature at distant stations, Meteorology and Hydrology. 10 (2000) 18–24.

[20] **Yu.A. Pichugin,** Ekologicheskiy monitoring i metody mnogomernoy matematicheskoy statistiki [Environmental quality monitoring and multivariate mathematical statistics], Astrakhanskiy vestnik ekologicheskogo obrazovaniya. (2) (2012) 101–105.

[21] **S.A. Ayvazyan, V.M. Bukhshtaber, I.S. Yenyukov, L.D. Meshalkin,** Prikladnaya statistika. Issledovaniye zavisimostey [Applied Statistics. Relation studies]. Finansy i statistika, Moscow, 1985.

[22] **Yu.A. Pichugin,** Consideration of seasonal effects in problem of SAT forecasting and data control, Meteorology and Hydrology. 4 (1996) 52–64.

[23] **A.S. Mikhalchuk, Yu.A. Pichugin,** Dispersionnyy analiz pogreshnostey tekhnologicheskikh protsessov mikroelektroniki [The variance analysis of errors in the microelectronics technological processes], In collection of papers: Modelirovaniye i situatsionnoye upravleniye kachestvom slozhnykh sistem: sbornik dokladov [Simulation and quality control of complicated systems], SUAI, St. Petersburg (2017) 35–38.

[24] **Yu.A. Pichugin,** Empirical components of annual march of surface temperature, Meteorology and Hydrology. 12 (1994) 34–43.

[25] **Yu.A. Pichugin, O.A. Malafeyev,** Optimizatsiya i prognoz v dinamicheskoy modeli upravleniya portfelem tsennykh bumag [Optimization and prediction in the dynamic model of investment portfolio governance], In: Materialy sektsionnykh zasedaniy simpoziuma «Nobelevskiye laureaty po ekonomike i rossiyskiye ekonomicheskiye shkoly» [In: Proceedings of symp. "Nobel Prize winners in economics and Russian economic schools of sciences"] , SPbSU, St. Petersburg (2003) 183–185.

[28] **Yu.A. Pichugin,** Glavnyye komponenty mnogomernykh vremennykh ryadov: analiz i prognoz [Principal components of multivariate time series: Analysis and prediction], In: Collection of papers of "The 13th International Youth Scientific Conf. on Soft Computing", Vol. 1, The 1st Electrotechnical University «LETI», St. Petersburg (2010) 160–163.

Received 26.03.2018, accepted 21.06.2018.

THE AUTHOR

PICHUGIN Yury A.
Saint-Petersburg State University of Aerospace Instrumentation
61 Bolshaya Morskaya St., St. Petersburg, 190000, Russian Federation
yury-pichugin@mail.ru

THE EXPONENTIAL MODEL OF THE CELL GROWTH: A SIMULATION ERROR

V.I. Antonov[1], E.A. Blagoveshchenskaya[2], O.A. Bogomolov[3], V.V. Garbaruk[2], J.G. Yakovleva[3]

[1]Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation;

[2]Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russian Federation;

[3]Russian Research Center for Radiology and Surgical Technologies, St. Petersburg, Russian Federation

Mathematical modeling of pathological changes in the body is the means of obtaining information for making decisions about the method of treatment. Numerous studies have shown that the exponential model describes the tumor cells growth, and the time of antigen doubling determines the aggression of cancer cells growth. The present work investigates inaccuracies in determining the antigen doubling time as a function of measurement errors. The study showed that the decision on the method of treatment could be changed by taking into account errors in the prognosis of patient's condition. For patient's stratification in groups of high, medium and low risks, various threshold values corresponding to the antigen level are proposed. The results are presented in the form of a table and graphs.

Key words: mathematical modeling, pathological changes, antigen, simulation error

Citation: V.I. Antonov, E.A. Blagoveshchenskaya, O.A. Bogomolov, V.V. Garbaruk, J.G. Yakovleva, The exponential model of the cell growth: A simulation error, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 70–76. DOI: 10.18721/JPM.11308

Introduction

Cancer is one of the most common fatal diseases. Cancer incidence is on the rise. About six million new cases of malignant tumors are diagnosed every year. Cancer ranks as the third leading cause of death in the world, coming after cardiovascular and respiratory diseases.

Mathematical modeling of pathological changes in the human body is an important tool, providing data for effective decision-making in selecting treatment methods and timing. Either deterministic and stochastic models or those using methods of nonlinear dynamics are commonly chosen as basic models [1 – 11]. Most models rely on experimental data, which entails accounting for errors in setting the problem's parameters. This approach is necessary because a large number of factors affect the course of different diseases.

Prostate cancer is considered the most diagnosed cancer in men and the second (according to statistical data) cause of death from cancer [12]. The level of prostate-specific antigen p (PSA) in serum, measured in ng/ml, is one of the best-studied markers, widely used for early detection of this cancer. The kinetics of the marker may reflect the actual growth rate of the tumor.

The goal of this study has been to analyze the effect of the errors in measuring PSA in serum on the result of determining the antigen's doubling time.

Exponential model

An increase in the number of tumor cells is generally described by an exponential model, and the p level linearly depends on the number of these cells in many cases [12]. The doubling time t_d for p (measured in months in this model) determines the aggressiveness of cancer cell growth. This parameter allows to control the tumor's growth rate, choose the optimal therapeutic approach and assess treatment effectiveness. However, empirical data intrinsically contain errors; for this reason, decision-making based on the predictions of an unstable model should involve error estimation [13].

The given element p is proportional to its increment Δp, leading to an exponential model. In this case, the equality

$$dp = kpdt, \quad (1)$$

holds true, and, consequently,

$$p = \tilde{C}e^{kt}. \quad (2)$$

The law of exponential growth is valid at a certain stage for cell populations in tissue, including tumor cells [1]. The exponential model should be used bearing in mind that the solution of differential equation (2) is Lyapunov-unstable for $k > 0$ [14], i.e., small variations in the initial conditions correspond to significant errors in the final calculations. The exponential model is widespread and appears valid to use provided that its parameters can be adjusted by the observation results or by qualitative study of the system's behavior.

With the known values of p, for example, p_1 and p_2, measured at different times t_1 and t_2, the coefficients of the solution of differential equation (1), written as

$$\ln p = C + kt,$$

have the form

$$C = \frac{t_2 \ln p_1 - t_1 \ln p_2}{t_2 - t_1}; \qquad k = \frac{\ln p_2 - \ln p_1}{t_2 - t_1}. \quad (3)$$

Notably, the coefficient C is a dimensionless quantity, while the coefficient k is measured in (months) $^{-1}$.

The time t_d, elapsed from the time t_2, that it takes for p_2 to double is predicted by the solution of the equation

$$2p_2 = p_2 e^{k \cdot t_d};$$

it follows from here that the following equality should hold true:

$$t_d = \ln 2 \cdot \frac{t_2 - t_1}{\ln p_2 - \ln p_1}. \quad (4)$$

We are going to assume from now on that an absolute measurement error Δp_i $(i = 1, 2)$ can be made in the value of p, with $|\Delta p_i| \le \varepsilon \cdot p_i$. Then the value of p is estimated as

$$p_i \pm \Delta p_i = p_i(1 \pm \varepsilon_i) = q_i \cdot p_i.$$

Here $q_i \cdot 100\%$ is the percentage of relative measurement error for p_i.

In finding the p_1 and p_2 levels with the respective errors q_1 and q_2, the doubling time t_d^{er} for p, taking into account errors, and the relative error δt_d of doubling time prediction are calculated by formulae

$$t_d^{er} = \ln 2 \cdot \frac{t_2 - t_1}{\ln \frac{q_2 p_2}{q_1 p_1}}, \quad (5)$$

$$\delta t_d = \left| \frac{t_d^{er} - t_d}{t_d^{er}} \right| = \left| \frac{\ln q_2 - \ln q_1}{\ln p_2 - \ln p_1} \right|. \quad (6)$$

The relative measurement error for p typically varies from 2 to 20 % [15]. Errors in measuring p lead to large errors in determining t_d. Notably, the projected doubling time is calculated without error even with large but identical relative errors in determining the p levels, which means that it is preferable to measure the p level at the same laboratory with the same equipment.

The denominator in formulae (4), (5) is close to zero for a small time interval $(t_2 - t_1)$ between measurements of p, which significantly increases the error in predicting t_d. To provide the given relative error Q for calculating the doubling time, the time interval between two measurements of p should satisfy the inequality

$$t_2 - t_1 \ge \frac{\left| \ln \frac{q_2}{q_1} \right|}{Q \cdot \ln 2}.$$

For example, with a 5 % error in determining the level of p, the ratio q_2/q_1 can vary from $(100 - 5) / (100 + 5)$ to $(100 + 5)/(100 - 5)$, i.e., from about 0.9 to 1.1, and from 0.82 to 1.22 with a 10% error.

Calculation results and discussion

The data in Table 1 can be used to estimate, for example, the margin of the possible error in predicting t_d^{er} with $p_2/p_1 = 1.51$ and the difference $(t_2 - t_1) = 12$ months. Instead of $t_d = 20$ months, t_d^{er} values range from 17 to 27 months, i.e., include the values below the critical. This means that more intensive treatment should be started at $t_d^{er} = 27$ months taking into

Table

Predicted values for the doubling times t_d^{er} for the cancer marker p depending on the errors q in measuring the marker with different parameters

$q_2/\ q_1$	t_d^{er}, months		
	$p_2 = 1{,}51$ ng/ml, $t_d = 20$ months	$p_2 = 1{,}46$ ng/ml, $t_d = 22$ months	$p_2 = 1{,}56$ ng/ml, $t_d = 19$ months
0.90	27	30	25
0.92	25	28	23
0.94	24	26	22
0.96	22	25	21
0.98	21	23	**20**
1.00	**20**	22	19
1.02	19	21	18
1.04	18	**20**	17
1.06	18	19	17
1.08	17	18	16
1.10	17	18	15

Notations: q_1 and q_2, %, are the errors of measured values of the markers p_1 and p_2, obtained at times t_1 and t_2; t_d is the predicted doubling time without measurement errors.

Notes. 1. t_d^{er} should be calculated by formula (5), assuming that the initial value of the marker p_1 is the same and is 1 ng/ml; the difference $t_2 - t_1 = 12$ months 2. The values of $t_d^{er} = 20$ months are highlighted in bold as critical: the growth rate of cancer cells is regarded as threatening below these values.

account the model's error.

It follows from formulae (4), (5) and Table 1 that the absolute and relative errors of determining t_d increase with smaller values of the p_2/p_1 ratio. Small t_d^{er} values correspond to a large p_2/p_1ratio, while the error in determining the doubling time decreases.

Different threshold values of p were proposed for stratification of patients by groups of high, medium and low risks in accordance with their PSA t_d levels [12]. Let us denote the values corresponding to these risks as p_{top} and p_{low} for further calculations. Patients with $p < p_{low}$ undergo preventive health screenings. Radical treatment is started if $p > p_{top}$. The $[p_{low}; p_{top}]$ interval is commonly referred to as the gray zone [15], as different treatment plans can be chosen for the p values lying in this range. Predicting whether the given p value falls in the gray or critical zone makes it possible to calculate the recommended time for the next measurement of p. If the model for the variation of p corresponds to the exponential one with parameters (3), then the value of p equal to p_b is reached at time t_b, for which one of the following equalities holds true, either

$$t_b = \frac{\ln\left(p_b^{(t_2-t_1)} \cdot \frac{p_2^{t_1}}{p_1^{t_2}}\right)}{\ln\left(\frac{p_2}{p_1}\right)},$$

or

$$t_b - t_2 = (t_2 - t_1) \cdot \frac{\ln \frac{p_b}{p_2}}{\ln \frac{p_2}{p_1}} = \frac{\ln \frac{p_b}{p_2}}{\ln 2} \cdot t_d. \qquad (7)$$

To calculate the t_b prediction taking into account the error in measuring p, q_1p_1 and q_2p_2 should be substituted into formula (7) instead of p_1 and p_2:

$$(t_b^{er} - t_2) = (t_2 - t_1) \cdot \frac{\ln(p_b / q_2p_2)}{\ln(q_2p_2 / q_2p_1)}.$$

Fig. 1,*a* shows how quickly the p value in

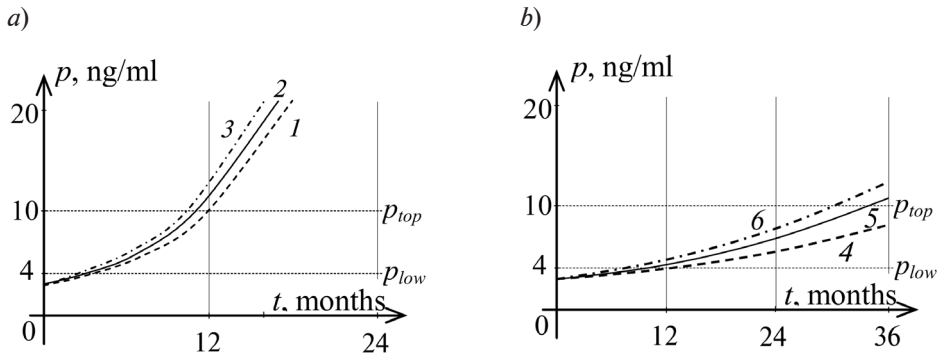

Fig. 1. Growth kinetics for the values of cancer marker p for different values of the t_d parameter, months: 5.61 (*1*), 6.00 (*2*), 6.49 (*3*) (*a*) and 17 (*4*), 20 (*5*) and 27 (*6*) (*b*); p_{top} and p_{low} are the boundaries of the gray zone; $p > p_{top}$ corresponds to the critical zone; p_2= 3 ng/ml; $(t_2 - t_1)$ = month

the gray area is reached and a transition into the critical zone is made with a high PSA growth rate (t_d = 6 months, p_{low} = 4 ng/ml, p_{top} = 10 ng/ml and $(t_2 - t_1)$ = 6 months with p_2= 3 ng/ml). In this case,

$$t_b - t_2 = 6 \cdot \frac{\ln(10 / 3)}{\ln 2} \approx 10,4.$$

This means that the next p measurement should be scheduled in about 10 months, since the level of p is going to fall into the critical zone after 12 months. Taking into account the error in determining p can change this interval by a month. The value of p might fall into the gray zone in 2.5 months; this should be kept in mind when scheduling the p measurement date.

Fig. 1,*b* shows when the gray zone is reached with the same value of p_2 and t_d = 20 months. In this case, the possibility that the lower boundary of the gray zone might be reached should be taken into account and the next p measurement should be scheduled in 8 months. This interval can be varied from 7 to 11 months when accounting for the error in measuring p.

Measuring p at a third time t_3 allows to adjust the values of coefficients (3) provided that the exponential model agrees with the experimental data obtained. The adequacy of the model can be tested in several ways.

If

$$\frac{p_3 - p_2}{t_3 - t_2} \approx \frac{p_2 - p_1}{t_2 - t_1}$$

(or $p_3 + p_1 \approx 2p_2$, provided that the measurements were carried out at equal intervals of time), then p increases linearly and the exponential model should be abandoned. This means that the increase in p is not caused by the growth of the tumor, but by other factors. The date for measuring p was selected with respect to the time when the boundary value p_b might be reached. If the obtained value of p_3 differs little from the predicted one, then the exponential model is chosen correctly. Then, with constant parameters of the model and no error in measuring p, the doubling time is constant, and the results of t_d calculations should be the same for choosing any two measurements taken at different times. The exponential model is adequate given the approximate equality of the value

$$t_d = \ln 2 \cdot \frac{t_2 - t_1}{\ln p_2 - \ln p_1}$$

and the quantities

$$t_{d32} = \ln 2 \cdot \frac{t_3 - t_2}{\ln p_3 - \ln p_2},$$

$$t_{d31} = \ln 2 \cdot \frac{t_3 - t_1}{\ln p_3 - \ln p_1},$$

i.e., with

$$\frac{\ln p_3 - \ln p_2}{t_3 - t_2} \approx \frac{\ln p_2 - \ln p_1}{t_2 - t_1}$$

(or $p_3 \cdot p_1 \approx p_2^2$, if the measurements were carried out at regular time intervals).

The coefficients of the exponent that de-

viates the least from the given three points $(t_1; p_1)$, $(t_2; p_2)$, $(t_3; p_3)$ can be then tailored to estimate the values of the residuals from the experimental points.

In this case, we have an inconsistent system of three equations with two unknowns:

$$\begin{cases} C + kt_1 = \ln p_1; \\ C + kt_2 = \ln p_2; \\ C + kt_3 = \ln p_3. \end{cases} \quad (8)$$

The coefficients C and k, approximately satisfying all the equations of the system, can be found by the least squares method:

$$a = \sum_{i=1}^{3} t_i^2, \quad b = \sum_{i=1}^{3} t_i,$$

$$u = \sum_{i=1}^{3} \ln p_i, \quad v = \sum_{i=1}^{3} t_i \ln p_i,$$

$$\begin{cases} C = \dfrac{a \cdot u - b \cdot v}{3a - b^2}; \\ k = \dfrac{3 \cdot v - b \cdot u}{3a - b^2}. \end{cases} \quad (9)$$

If an exponential model with coefficients (9) is adopted, then the adjusted doubling time for p is calculated by the formula

$$t_d = \frac{(\ln 4)(\tau_{12}\tau_{13} + \tau_{21}\tau_{23} + \tau_{32}\tau_{31})}{\ln\left(\left(\dfrac{p_2}{p_1}\right)^{\tau_{21}} \cdot \left(\dfrac{p_3}{p_2}\right)^{\tau_{32}} \cdot \left(\dfrac{p_3}{p_1}\right)^{\tau_{31}}\right)}, \quad (10)$$

$$\tau_{ij} = t_i - t_j.$$

The error of calculating t_d is found by the formula

$$\delta t_d = \left| \frac{\ln\left(\left(\dfrac{q_2}{q_1}\right)^{\tau_{21}} \left(\dfrac{q_3}{q_2}\right)^{\tau_{32}} \left(\dfrac{q_3}{q_1}\right)^{\tau_{31}}\right)}{\ln\left(\left(\dfrac{p_2}{p_1}\right)^{\tau_{21}} \left(\dfrac{p_3}{p_2}\right)^{\tau_{32}} \left(\dfrac{p_3}{p_1}\right)^{\tau_{31}}\right)} \right|. \quad (11)$$

Formulae (10) and (11) coincide with formulae (4) and (5), provided that the measurements were carried at equal time intervals of time, i.e., with $(t_3 - t_2) = (t_2 - t_1)$:

$$t_{d31} = \ln 2 \frac{t_3 - t_1}{\ln \dfrac{p_3}{p_1}}; \; \delta t_{d31} = \left| \frac{\ln \dfrac{q_3}{q_1}}{\ln \dfrac{p_3}{p_1}} \right|.$$

The error does not depend on the mean measurement error in this case.

Conclusion

Analysis of the growth kinetics of cancer cells [16, 17], established based on an exponential model, is a key step in assessing the effect of the method chosen for patient treatment. Prognosis of the disease outcome in a patient should take into account the total errors of the model, which, as we have established in this study, exceed the error in measuring the characteristics of the patient's condition.

We have obtained the formulae for calculating the relative error of the model, and found potential methods for reducing the effect of this error on the predictive capabilities of the exponential model.

We have confirmed that the decision on choosing a method for treating a patient may change upon taking into account possible errors in predicting the patient's condition.

We have proposed a method for calculating the time interval between patient assessments, necessary for adjusting the parameters of the model describing the patient's condition.

Additional data available on the patient's condition allows to assess the adequacy of the model by the several methods we have described.

Our findings can have beneficial applications not only in medicine, since the the exponential model is effective at some stages of growth rate analysis of consumption, capital, population, etc. [18].

REFERENCES

[1] **G.Yu. Reznichenko, A.B. Rubin,** Matematicheskoye modelirovaniye biologicheskikh produktsionnykh protsessov [Simulation of the biological production processes], MSU, Moscow, 1993.

[2] **V. Antonov, A. Zagainov, A. Kovalenko,** Fractal analysis of biological signals in a real time mode, Global and Stochastic Analysis. 3 (2) (2016) 75–84.

[3] **V. Antonov, A. Zagaynov,** Software package for calculating the fractal and cross spectral parameters of cerebral hemodynamic in a real time mode, New Trends in Stochastic Modeling and

Data Analysis, Ch. 7. Demography and Related Applications, ISAST (440) (2015) 339–345.

[4] **G.I. Marchuk,** Matematicheskiye modeli v immunologii [Mathematical models in immunology], Nauka, Moscow, 1985.

[5] **I.V. Ashmetov, A.Ya. Bunicheva, S.I. Mukhin S.I., et al.,** Matematicheskoye modelirovaniye gemodinamiki v mozge i v bolshom kruge krovoobrashcheniya [Simulation of hemodynamics in the brain and greater circulation], In: Computer and Brain. New technologies, Nauka, Moscow, 2005.

[6] **S.A. Astanin, A.V. Kolobov, A.I. Lobanov,** Vliyaniye prostranstvennoy geterogennoy sredy na rost i invaziyu opukholi. Analiz metodami matematicheskogo modelirovaniya [The influence of the spatial heterogeneous medium on the tumor growth and invasion, An analysis by mathematical modeling], In: Health Care in the Mirror of Informatics, Nauka, Moscow (2006) 163–194.

[7] **R. Molina-Pena, M.M. Alvarez,** A simple mathematical model based on the cancer stem cell hypothesis suggests kinetic commonalities in solid tumor growth, PLOS. One. 7(2): e26233, doi: 10.1371/journal.pone.0026233 (2012).

[8] **A.V. Kolobov, A.A. Polezhaev, G.I. Solyanik,** The role of cell motility in metastatic cell dominance phenomenon: analysis by a mathematical model, Journal of Theoretical Medicine. 3 (1) (2001) 63–77.

[9] **M.J. Williams, B. Werner, C.P. Barnes, et al.,** Identification of neutral tumor evolution across cancer types, Nature Genetics (48) (2016) 238–244. doi:10.1038/ng.3489.

[10] **N.A. Babushkina, L.A. Ostrovskaya, V.A. Rykova, et al.,** Modelirovaniye effektivnosti deystviya protivoopukholevykh preparatov v sverkhmalykh dozakh dlya optimizatsii rezhimov ikh vvedeniya [Simulation of the curative efficacy of the anticancer drug at a very-low-dose for dose-schedule optimization], Control Problems. (4) (2005) 47–54.

[11] **S. Benzekry, C. Lamont, A. Beheshti, et al.,** Classical mathematical models for description and prediction of experimental tumor growth, PLOS Comput. Biol. 10(8) 2014; e1003800. doi: 10.1371/journal.pcbi.1003800.

[12] **G.M. Zharinov, O.A. Bogomolov,** The pretreatment prostate-specific antigen-doubling time: clinical and prognostic values in patients with prostate cancer, Cancer Urology. (1) (2014) 44–48.

[13] **J.R. Taylor.** Introduction to the theory of errors. Moscow: Mir, 1985. 272 p.

[14] **Z.S. Galanova, V.V. Garbaruk,** Issledovaniye ustoychivosti avtonomnykh system [Studies in stability of independent systems], Petersburg State Transport University, St. Petersburg, 2005.

[15] **A.N. Kurzanov, E.A. Strygina, V.L. Medvedev,** Diagnostic and prognostic markers in prostate cancer, Modern Problems of Science and Education. (2) (2016) URL: http://www.science-education.ru/ru/article/view?id=24439

[16] **M.L. Ramirez, E.C. Nelson, R.W. deVere White, et al.,** Current applications for prostate-specific antigen doubling time, European Urology. 54 (2) (2008) 291–300.

[17] **M. Grosh, A. Dagher, F. El-Karar,** Prostate-specific antigen doubling time and response to cabazitaxel in a hormone-resistant metastatic prostate cancer patient, Journal of Biomedical Research. 29 (5) (2015) 420–422.

[18] **D. Meadows, J. Randers, D. Meadows,** The limits to growth. The 30-year update, Chelsea Green Publishing Company, Vermont, 1972.

Received 06.06.2018, accepted 05.09.2018.

THE AUTHORS

ANTONOV Valeriy I.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
antonovvi@mail.ru

BLAGOVESHCHENSKAYA Ekaterina A.
Emperor Alexander I St. Petersburg State Transport University
9 Moskovsky Ave., St. Petersburg, 190031, Russian Federation
kblag2002@yahoo.com

BOGOMOLOV Oleg A.
Russian Research Center for Radiology and Surgical Technologies
70 Leningradskaya St., St. Petersburg, Pesochniy Settl., 197758, Russian Federation
urologbogomolov@gmail.com

GARBARUK Victor V.
Emperor Alexander I St. Petersburg State Transport University
9 Moskovsky Ave., St. Petersburg, 190031, Russian Federation
vigarb@mail.ru

YAKOVLEVA Julia G.
Russian Research Center for Radiology and Surgical Technologies
70 Leningradskaya St., St. Petersburg, Pesochniy Settl., 197758, Russian Federation
vmkaf@pgups.ru

SHARP V-NOTCH FRACTURE CRITERIA UNDER ANTIPLANE DEFORMATION

V.V. Tikhomirov

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation

The criteria for brittle fracture of a sharp V-notch when it is loaded with antiplane concentrated forces have been considered: a criterion for the maximum average stress, a criterion for the average energy density of deformation, and an approach based on the joint use of the force and energy criteria. Failure loads estimates on the basis of the exact solutions and using asymptotics of stresses near the V-notch tip were found. A comparative analysis of the failure loads obtained through those criteria was carried out. For the asymmetric loading, the initial angle of the crack propagation from the V-notch tip was determined. In the calculation of this angle, the application of the stress asymptotics was shown to result in significant errors and to require the consideration of regular terms in the stress representations.

Key words: antiplane deformation, sharp V-notch, fracture criterion, average stress, deformation energy

Citation: V.V. Tikhomirov, Sharp V-notch fracture criteria under antiplane deformation, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 77–83. DOI: 10.18721/JPM.11309

Introduction

The tIps of sharp v-notches in elastic bodies are singular points of stress fields. Cracks may develop from these singularities under certain conditions, leading to fracture of elastic structures. For this reason, study of the stress-strain state of bodies with notches, as well as developing fracture criteria and confirming them experimentally are issues of particular interest.

Plane cracks with known fracture criteria (first developed by Griffith and Irwin) are a particular case of notches to which these criteria are not directly applicable. In view of this, several other criteria for fracture of structural elements with sharp notches have been proposed:

force [1 – 5];

energy [6 – 10];

formulated within the framework of so-called finite fracture mechanics and based on combined application of force and energy conditions [11, 12].

The majority of studies applied these criteria within the framework of a plane problem for structures with finite or semi-infinite notches. The criteria were in fact used based on asymptotic representations of stresses near the stress concentrations. It was established that the critical values of fracture parameters, such as ultimate loads, can be expressed in terms of macroscopic characteristics of materials, such as ultimate tensile strength and fracture toughness.

Some recent studies estimated the effect that including non-singular terms in stress expansions in the vicinity of the notch tip has on parameters such as the generalized stress intensity factor [13] and the crack initiation angle [14]. The results of these studies indicate that including a non-singular first term in the Williams expansion significantly affects these fracture parameters.

Antiplane problems related to our subject matter were discussed in very few studies [8, 15, 16], with no comparative analysis of fracture criteria for wedge-shaped structures.

The main goal of this study is to extend the fracture criteria developed for plane problems to the case of antiplane deformation of notched bodies and comparative analysis of these criteria in determining ultimate loads.

Since an elastic solution for a uniform wedge-shaped region can be obtained in closed form as explicit representations for stresses and displacements in case of antiplane loading, it is possible to estimate the accuracy of calculating the fracture load using stress asymptotes at the tip of the notch.

Green's functions for a sharp notch

Let us consider antiplane deformation of a homogeneous isotropic wedge-shaped region with a vertex angle 2α $(\pi/2 < \alpha \le \pi)$. The notch then is defined by the angle $\beta \in [0, \pi)$. A concentrated force $2T$ directed from the plane is applied to the face of the wedge $\theta = \alpha$ at a distance r_0 from the tip. The general problem on finding the stress-strain state of a plane with a notch is linear, so it can be represented as a superposition of two problems:

1) with symmetric loading of the notch faces, when

$$\tau_{\theta z}(r, \alpha) = \tau_{\theta z}(r, -\alpha) = T\delta(r - r_0), \qquad (1)$$

2) with antisymmetric loading, when

$$\tau_{\theta z}(r, \alpha) = -\tau_{\theta z}(r, -\alpha) = T\delta(r - r_0). \qquad (2)$$

($\delta(r - r_0)$ is the Dirac delta function).

Next, let us apply a Mellin integral transform to the harmonic equilibrium equation, satisfying boundary conditions (1) and (2); in addition, let us use the residue theorem. As a result, we obtain the following representations for stresses:

in problem 1:

$$\tau_{\theta z}(r, \theta) = \frac{K_3^N}{\sqrt{2\pi} r^{1-\lambda}} \frac{1 + \rho^2}{1 + 2\rho^2 \cos 2\lambda\theta + \rho^4} \cos\lambda\theta,$$
$$\tau_{rz}(r, \theta) = \frac{K_3^N}{\sqrt{2\pi} r^{1-\lambda}} \frac{1 - \rho^2}{1 + 2\rho^2 \cos 2\lambda\theta + \rho^4} \sin\lambda\theta; \qquad (3)$$

in problem 2:

$$\tau_{\theta z}(r, \theta) = \frac{K_3^N}{\sqrt{2\pi} r^{1-\lambda}} \frac{\rho}{1 + 2\rho^2 \cos 2\lambda\theta + \rho^4} \sin 2\lambda\theta;$$
$$\tau_{rz}(r, \theta) = -\frac{K_3^N}{\sqrt{2\pi} r^{1-\lambda}} \frac{\rho(\rho^2 + \cos 2\lambda\theta)}{1 + 2\rho^2 \cos 2\lambda\theta + \rho^4}. \qquad (4)$$

$$\rho = (r/r_0)^{\lambda}.$$

Here, $\lambda = \pi/(2\alpha)$ and $\lambda = 1$ in the case of a half-plane $(\alpha = \pi/2)$ and $\lambda = 1/2$ in the case of a semi-infinite crack in an unbounded plane $(\alpha = \pi)$.

The quantity K_3^N in relations (3) is the generalized stress intensity factor (GSIF), defined by the formula

$$K_3^N = \lim_{r \to 0} \sqrt{2\pi} r^{1-\lambda} \tau_{\theta z}(r, 0) = \frac{\sqrt{2\pi} T}{\alpha r_0^{\lambda}}. \qquad (5)$$

With $\alpha = \pi$, the GSIF coincides with the stress intensity factor (SIF) at the tip of a semi-infinite crack:

$$K_3^N(\pi) = K_3 = T\sqrt{\frac{2}{\pi r_0}}. \qquad (6)$$

When concentrated forces take critical values equal to T_c, formulae (5) and (6) define the critical intensity factors.

$$K_{3c}^N = \frac{\sqrt{2\pi} T_c}{\alpha r_0^{\lambda}}, \quad K_{3c} = T_c\sqrt{\frac{2}{\pi r_0}}. \qquad (7)$$

It should be emphasized that, in contrast with the fracture toughness constant, the critical stress intensity factor K_{3c}^N at the tip of the notch is not a constant of the material, since it depends on the angle α. We should also note that stresses (3) at the tip of the notch in problem 1 have a power singularity, while stresses (4) in problem 2 do not. Stress asymptotes (3) with $r \to 0$ are determined by the formulae

$$\tau_{\theta z}(r, \theta) = \frac{K_3^N}{\sqrt{2\pi}} r^{1-\lambda} \cos\lambda\theta,$$
$$\tau_{rz}(r, \theta) = \frac{K_3^N}{\sqrt{2\pi}} r^{1-\lambda} \sin\lambda\theta. \qquad (8)$$

Notably, formulae (3) for stresses are consistent with the results given in [17].

Summing solutions (3) and (4), we obtain the stresses in the problem on the action of a concentrated force $2T$ at the face of the notch $\theta = \alpha$:

$$\tau_{\theta z}(r, \theta) = \frac{K_3^N}{\sqrt{2\pi}} r^{1-\lambda} \frac{\cos\lambda\theta}{1 - 2\rho \sin \lambda\theta + \rho^2},$$
$$\tau_{rz}(r, \theta) = \frac{K_3^N}{\sqrt{2\pi}} r^{1-\lambda} \frac{\sin\lambda\theta - \rho}{1 - 2\rho \sin \lambda\theta + \rho^2}. \qquad (9)$$

Evidently, stresses (9) have asymptotes (8) if $r \to 0$.

Criteria for sharp notch fracture

Let us consider application of the fracture criteria with the example of a notch with symmetrically loaded faces. In this case, by virtue of symmetry, the stress $\tau_{\theta z}$ reaches the maximum value on the ray $\theta = 0$ and, therefore, the crack is going to propagate from the tip of the notch along this ray.

Force criterion. Similarly to the assumptions adopted in [1, 2], we assume that fracture of the notch starts when the maximum mean stress calculated at a certain distance d from its tip reaches a critical value equal to the shear strength τ_c of the material:

$$\overline{\tau} = \frac{1}{d}\int_0^d \max_{-\alpha<\theta<\alpha} \tau_{\theta z}(r,\theta)dr = \tau_c. \qquad (10)$$

Substituting the stress $\tau_{\theta z}$ found by formula (3) with $\theta = 0$ to expression (10), we obtain the following equality:

$$\frac{K_{3c}^N(\alpha) r_0^{\lambda}}{\sqrt{2\pi}\lambda d}\operatorname{arctg}\left(\frac{d}{r_0}\right)^{\lambda} = \tau_c. \qquad (11)$$

It is valid for any value of the angle $\alpha \in (\pi/2, \pi]$.. Let us determine the parameter d for the angle $\alpha = \pi$, i.e., for the case of a crack with $\lambda = 1/2$ and $K_{3c}^N(\pi) = K_{3c}$, in other words, mode III fracture toughness. Then we obtain the following equation for determining the relative distance $x = d/r_0$ from condition (11):

$$x = \gamma \operatorname{arctg}\sqrt{x}. \qquad (12)$$

Here we have introduced a dimensionless parameter

$$\gamma = \sqrt{\frac{2}{\pi r_0}}\frac{K_{3c}}{\tau_c}, \qquad (13)$$

whose equivalent was used for plane problem in [12], where a different linear dimension, notch depth, was used instead of the distance r_0 to the load application point. This parameter was called the brittleness parameter, or the brittleness number in [12].

Let us estimate the value of γ using the example of a brittle material such as graphite. The fracture toughness of graphite in mode III is, according to [18], $K_{3c} = 0.415$ MPa·m$^{1/2}$. Since $\tau_c = \sqrt{3}\sigma_c$ (σ_c is the ultimate tensile strength, which takes the value of 20 MPa for graphite [19]), we obtain, according to formula (13), $\gamma = 0,0287/\sqrt{r_0}$. Then, for example, if $r_0 = 0.01$ m, we obtain the value $\gamma = 0,287$.

Using criterial relation (11) and representation (7), we obtain the estimate for the ratio of critical forces for the case of a notch and a crack:

$$\frac{T_c^N}{T_c} = \frac{x}{\gamma \operatorname{arctg}(x^{\lambda})}. \qquad (14)$$

With $\gamma << 1$ the root of equation (12) can be represented as

$$x = \gamma^2 + O(\gamma^3)$$

and, therefore, the asymptote of the relative critical load (14) is determined by the formula

$$\frac{T_c^N}{T_c} = \gamma^{1-2\lambda}. \qquad (15)$$

Since λ lies in the range $1/2 \le \lambda < 1$ for any value of the angle α from $\pi/2 < \alpha \le \pi$, the inequalities $-1 < 1 - 2\lambda \le 0$ hold true. Then it follows from formula (15) that larger forces have to be applied for fracture of a sharp notch at small values of the parameter γ, compared to the forces required for crack propagation. In other words, a crack, considered as a limiting case of a notch with $\alpha \to \pi$, can be regarded as the most dangerous notch. This conclusion agrees qualitatively with the result obtained for uniaxial tension of the notch in [12].

Notably, if only the singular terms of stress expansion (3), that is, asymptotes (8), are used in fracture criterion (10), then we also obtain equality (15) for the ultimate load. Thus, the estimates of the fracture load, constructed from exact and asymptotic solutions, coincide if the distances $\mathbf{r}_0$ from the tip of the notch to the force application points are large enough.

Energy criterion. Fracture of the notch by a forming crack starts when the mean deformation energy density, calculated in a finite volume of radius R with the center at the tip of the notch, reaches a critical value Π_c [6]:

$$\frac{1}{2\mu\alpha R^2}\int_0^R\int_{-\alpha}^{\alpha}(\tau_{rz}^2 + \tau_{\theta z}^2)rdrd\theta = \Pi_c, \qquad (16)$$

where μ is the shear modulus of the material.

The radius of the control volume R depends on the properties of the material.

The critical value of the mean deformation energy density assuming that it does not depend on the vertex angle of the notch can be expressed in terms of the shear strength of the material τ_c :

$$\Pi_c = \tau_c^2/(2\mu).$$

Then, using the representations for stresses (3) for the critical state of the material and calculating the integrals in criterion (16), we obtain the equality

$$\frac{(K_{3c}^N)^2 r_0^{2\lambda}}{8\lambda^2 \alpha R^2} \ln \frac{1+(R/r_0)^{2\lambda}}{1-(R/r_0)^{2\lambda}} = \tau_c^2. \qquad (17)$$

In the limiting case, when the notch degenerates into a crack, i.e., with $\alpha = \pi$ and, consequently, with $K_{3c}^N(\pi) = K_{3c}$, equality (17) yields the equation for determining the radius of the control volume:

$$y = \frac{\gamma}{2}\sqrt{\ln \frac{1+y}{1-y}}, \quad y = \frac{R}{r_0}. \qquad (18)$$

In view of equality (7) and equation (18), condition (17) leads to the following estimate of the fracture load for the notch:

$$\frac{T_c^N}{T_c} = \sqrt{\frac{2}{\lambda}}\frac{y}{\gamma}\Big/\sqrt{\ln \frac{1+y^{2\lambda}}{1-y^{2\lambda}}}. \qquad (19)$$

We obtain from formula (19) with $y = R/r_0 << 1$, that

$$\frac{T_c^N}{T_c} = \frac{1}{f(\lambda)}\gamma^{1-2\lambda}, \qquad (20)$$

and the function $f(\lambda) = 2^{1-\lambda}\sqrt{\lambda} \geq 1$ for any $\lambda \in [0{,}5;\ 1, 0]$.

Then, comparing estimates (15) and (20), we conclude that the ultimate load obtained from the force criterion exceeds the ultimate load found using the mean energy density criterion at any angle $\alpha \in (\pi/2, \pi]$.

Notably, using energy criterion (16) with only asymptotic representations (8) also leads to an estimate of the form (20).

Criterion based on finite fracture mechanics [12]. Criterion based on finite fracture mechanics [12]. In this case, it is assumed that two conditions must be simultaneously satisfied for finite propagation Δ of a crack from the top of the notch: the force and the energy conditions (for stresses and energy balance):

$$\int_0^\Delta \tau_{\theta z}(r, 0)dr \geq \tau_c \Delta,$$
$$\int_0^\Delta K_3^2(\varepsilon)d\varepsilon \geq K_{3c}^2 \Delta, \qquad (21)$$

where $K_3(\varepsilon)$ is the stress intensity factor (SIF) at the tip of the crack of length ε.

Thus, in order to use this criterion, we need to obtain, in addition to stress field (3), the solution of the problem on a crack of finite length ε, propagating from the tip of the notch (see Fig. 1).

Let us now apply the Mellin integral transform to the harmonic equilibrium equation, to condition (1) on the face $\theta = \alpha$ and to the following mixed conditions on the beam $\theta = 0$:

$$\tau_{\theta z}(r, +0) = 0 \quad (0 \leq r \leq \varepsilon),$$
$$w(r, +0) = 0 \quad (\varepsilon \leq r < \infty).$$

As a result, we obtain the Wiener – Hopf equation:

$$\operatorname{ctg}(p\alpha)\ T_-(p) + \frac{\mu}{\varepsilon}U_+(p) = \frac{Tr_0^p}{\varepsilon^{p+1}\sin(p\alpha)} \qquad (22)$$
$$(p \in L).$$

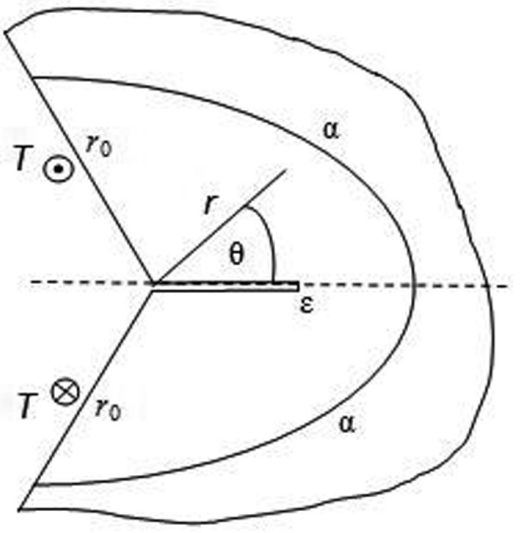

Fig. 1. A sharp notch with a symmetric crack of length ε propagating from its tip; 2α is the vertex angle of the wedge-shaped region; r_0 is the distance from the tip to the application point of concentrated forces T, directed from the plane; r, θ are the coordinates

Here p is the Mellin transform parameter. The stress transforms $T_-(p)$ and displacement transforms $U_+(p)$ along the beam are analytical functions in the left and right (relative to the contour L) half-planes.

Using the technique developed in [20], we obtain the exact solution of equation (22), which allows to express the SIF at the crack tip as

$$K_3 = K_3^N \psi(\lambda)\varepsilon^{\lambda-1/2}, \tag{23}$$

where $\psi(\lambda) = \{2\lambda[1 + (\varepsilon/r_0)^{2\lambda}]\}^{-1/2}$.

Substituting stresses (3) and SIF (23) into criterion (21) (at the critical state of the notch), we obtain the equalities

$$\begin{gathered} K_{3c}^N \frac{r_0^\lambda}{\lambda\sqrt{2\pi}} \operatorname{arctg}(\Delta/r_0)^\lambda = \tau_c\Delta, \\ (K_{3c}^N)^2 \frac{r_0^{2\lambda}}{4\lambda^2} \ln[1 + (\Delta/r_0)^{2\lambda}] = K_{3c}^2\Delta. \end{gathered} \tag{24}$$

From here we obtain the equation determining relative propagation of a crack with $\varsigma = \Delta/r_0$:

$$\varsigma = \gamma^2 \frac{\operatorname{arctg}^2\varsigma^\lambda}{\ln(1+\varsigma^{2\lambda})}. \tag{25}$$

Using equalities (7), we find from the first equation in (24) the relative fracture load in the form

$$\frac{T_c^N}{T_c} = \frac{\varsigma}{\gamma \operatorname{arctg}\varsigma^\lambda}. \tag{26}$$

Notice that equation (25) has a root $\varsigma \approx \gamma^2$ with $\varsigma << 1$. In this case, equality (26) leads to the following estimate of the fracture load:

$$\frac{T_c^N}{T_c} = \gamma^{1-2\lambda},$$

which coincides with formula (15) for using the force criterion of fracture.

The angle of initial propagation of the crack under asymmetric loading of the notch

To determine the initial angle of crack propagation from the tip of the notch under asymmetric loading, let us use, for example, the force criterion proposed within the framework of the plane problem [5]. Crack initialization occurs along the beam $\theta = \theta_*$, where the mean shear stress takes the maximum value:

$$\overline{\tau}_{\theta z}(\theta) = \frac{1}{d}\int_0^d \max_{-\alpha<\theta<\alpha} \tau_{\theta z}(r,\theta)dr, \tag{27}$$

$$\left.\frac{\partial\overline{\tau}_{\theta z}(\theta)}{\partial\theta}\right|_{\theta=\theta_*} = 0. \tag{28}$$

Substituting expression (9) into formula (27), we obtain the following representation for the mean tangential stress:

$$\begin{gathered} \overline{\tau}_{\theta z}(\theta) = \frac{K_3^N r_0^\lambda}{\sqrt{2\pi}\lambda d} \times \\ \times\left[\operatorname{arctg}\frac{(d/r_0)^\lambda - \sin\lambda\theta}{\cos\lambda\theta} + \lambda\theta\right]. \end{gathered} \tag{29}$$

After using condition (28), we find from here the angle θ_* describing the direction of initial crack growth:

$$\theta_* = \frac{1}{\lambda}\arcsin(d/r_0)^\lambda. \tag{30}$$

We should note that the follow estimate of the fracture load follows from fracture criterion (10) and formulae (29) and (30):

$$\frac{T_c^N}{T_c} = \frac{x}{\gamma \arcsin(x^\lambda)},$$

where $x = d/r_0$ is the root of the equation

$$x = \gamma \arcsin\sqrt{x}.$$

Numerical results and discussion

Based on the three given criteria, we have calculated fracture loads under symmetric loading of the notch depending on the parameter γ and different angles α. Comparative analysis of the results based on the exact solution of problem (3) shows that all the criteria yield similar results and the maximum discrepancy does not exceed 3 % for small values of the parameter ($\gamma < 0.1$). In this case, according to formulae (15) and (20), the fracture load has an asymptotic estimate $T_c^N/T_c = O(\gamma^{1-2\lambda})$.

With increasing parameter γ, the relative ultimate load decreases, and its values, determined using criteria (10), (16) and (21), diverge. The criterion based on finite fracture mechanics yields the greatest value for this load, and the criterion of mean deformation energy density provides a lower-bound estimate of the load. For example, with $\gamma = 0.8$ and a

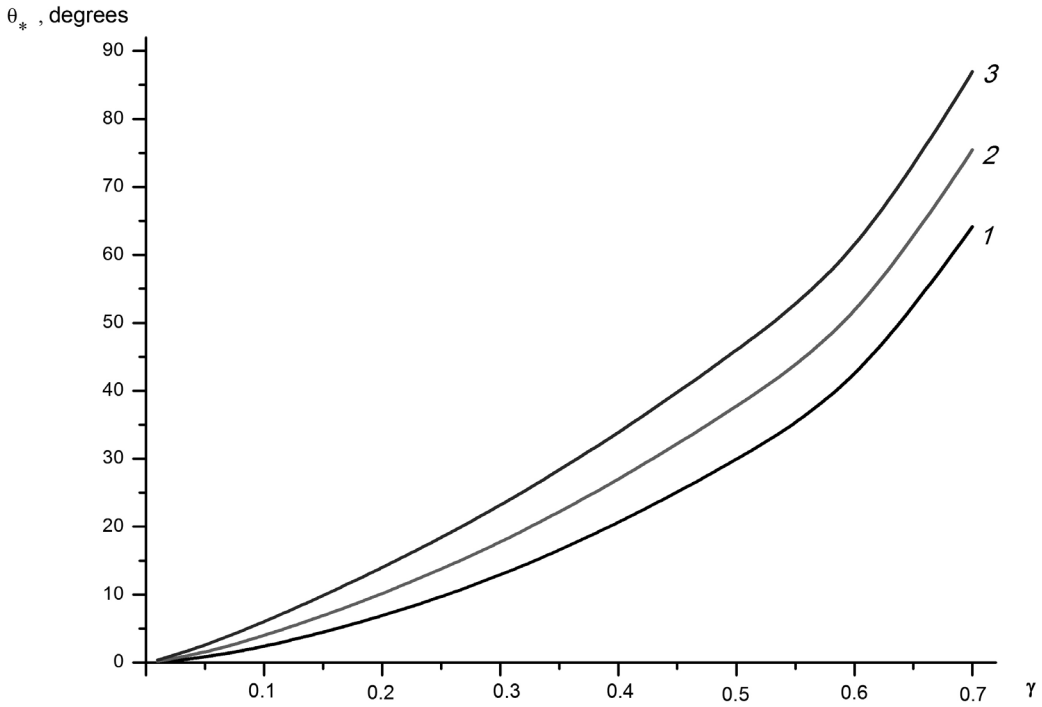

Fig. 2. Dependences of the initial angle of crack propagation from the tip of the notch on the parameter γ with different notch vertex angles α, deg: 120 (*1*) 135 (*2*), 150 (*3*)

notch with an angle of 90°, the difference in estimates for T_c^N / T_c based on these criteria is about 13 %.

The values of fracture loads for the notch found using stress asymptotes (8) almost coincide, up to $\gamma = 0.5$, with the values calculated from exact solution (3).

Thus, using the asymptotes of the stress field near the tip of the notch to estimate the fracture load within the framework of the antiplane problem is fairly acceptable.

Under asymmetric loading of the notch faces, the initial angle of crack propagation depends significantly on the regular terms in stress representation (9). Using only stress field asymptotes (9) in the form (8) with criterion (27), (28) determines the initial angle $\theta_*^{as} = 0$. However, this angle, calculated from exact solution (9) using formula (30), may considerably differ from the value of θ_*^{as} (Fig. 2). It follows then that non-singular terms have to be included in the formulae for stresses if $r \to 0$ for finding the direction of the initial growth of a crack from the tip of the notch.

Conclusion

The paper considers the criteria for brittle fracture of a sharp notch under antiplane loading with concentrated forces: a) maximum mean stress, b) mean deformation energy density, c) an approach based on combining force and energy criteria.

We have established that the fracture loads resulting from application of different criterial relationships are expressed in terms of a single dimensionless parameter depending on the material constants (shear strength and fracture toughness in mode III). Apparently, the ultimate loads found using different approaches are quite close.

However, the angle of initial propagation of a crack from the tip of the notch considerably depends on the accuracy of calculating the stresses near this tip, i.e, calculating this angle based on the stress asymptotes leads to significant errors.

REFERENCES

[1] **V.V. Novozhilov,** On the necessary and sufficient criterion for brittle strength, J. Appl. Math. Mech. 33 (2) (1969) 212 −222.

[2] **Z. Knesl,** A criterion of V-notch stability, Int. J. Fract. 48 (4) (1991) R79−R83.

[3] **A. Sewerin,** Brittle fracture criterion for structures with sharp notches, Eng. Fract. Mech. 47 (5) (1994) 673−681.

[4] **M.L. Dunn, W. Suwito, S.J. Cunningham,** Fracture initiation at sharp notches: Correlation using critical stress intensities, Int. J. Solids Struct. 34 (29) (1997) 3873−3883.

[5] **J. Klusak, T. Profant, M. Kotoul,** Determination of the threshold values of orthotropic bi-material notches, Proc. Engineering. 2 (1) (2010) 1635−1642.

[6] **P. Lazzarin, R. Zambardi,** A finite-volume-energy based approach to predict the static and fatigue behavior of components with sharp V-shaped notches, Int. J. Fract. 112 (3) (2001) 275−298.

[7] **Z. Yosibash, A. Bussiba, I. Giland,** Failure criteria for brittle elastic materials, Int. J. Fract. 125 (2) (2004) 307−333.

[8] **M. Treifi, O. Oyadiji,** Strain energy approach to compute stress intensity factors for isotropic homogeneous and bi-material V-notches, Int. J. Solids Struct. 50 (14−15) (2013) 2196− 2212.

[9] **P. Lazzarin, A. Campagnolo, F. Berto,** A comparison among some recent energy- and stress-based criteria for the fracture assessment of sharp V-notched components under Mode I loading, Theor. Appl. Fract. Mech. 71 (1) (2014) 21−30.

[10] **A. Campagnolo, F. Berto, D. Leguillon,** Fracture assessment of sharp V-notched components under Mode II loading: a comparison among some recent criteria, Theor. Appl. Fract. Mech. 85 B (2016) 217−226.

[11] **I.G. Garsia, D. Leguillon,** Mixed-mode crack initiation at a V-notch in presence of an adhesive joint, Int. J. Solids Struct. 49 (15−16) (2012) 2138− 2149.

[12] **A. Carpinteri, P. Cornetti, N. Pugno, A. Sapora,** On the most dangerous V-notch, Int. J. Solids Struct. 47 (7−8) (2010) 887−893.

[13] **M.R. Ayatollahi, M. Dehghany, M. Nejati,** Fracture analysis of V-notched components − effects of first non-singular stress term, Int. J. Solids Struct. 48 (10) (2011) 1579−1589.

[14] **M.M. Mirsayar, M.R.M. Aliha, A.T. Samaei,** On fracture initiation angle near bi-material notches − effects of first non-singular stress term, Eng. Fract. Mech. 119 (1) (214) 124−131.

[15] **M. Zappalorto, P. Lazzarin,** A unified approach to the analysis of nonlinear stress and strain fields ahead of mode III-loaded notches and cracks, Int. J. Solids Struct. 47 (6) (2010) 851− 864.

[16] **W. Shi,** Path-independent integral for the sharp V-notch in longitudinal shear problem, Int. J. Solids Struct. 48 (3-4) (2011) 567−572.

[17] **C-H. Chuo, W-B. Wei, T.J-C. Liu,** The antiplane electro-mechanical field of a piezoelectric wedge under a pair of concentrated forces and free charges, J. Chin. Inst. Eng. 26 (5) (2003) 575−583.

[18] **M.R.M. Aliha, A. Bahmani, S. Akhondi,** Determination of mode III fracture toughness for different materials using a new designed test configuration, Mat. Design. 86 (2015) 863−871.

[19] **C.N. Morrison, A.P. Jivkov, Ye. Vertyagina, T.J. Marrow,** Multi-scale modelling of nuclear graphite tensile strength using the site-bond lattice model, Carbon. 100 (2016) 273−282.

[20] **V.V. Tikhomirov,** Mode III crack approaching to the wedge-shaped elastic inclusion, St. Petersburg Polytechnical University Journal. Physics and Mathematics. 10 (2) (2017) 99−109.

Received 19.03.2018, accepted 20.05.2018.

THE AUTHOR

TIKHOMIROV Victor V.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
victikh@mail.ru

THE FAR FIELD OF A SUBMERGED LAMINAR JET: LINEAR HYDRODYNAMIC STABILITY

R.I. Mullyadzhanov, N.I. Yavorsky

Kutateladze Institute of Thermal Physics, Novosibirsk, Russian Federation

A linear stability problem for a submerged Landau – Squire jet has been considered. It was shown that in the space, the intrinsic perturbation amplitude varied as a power function of the spherical radius R, read from the motion source. It was established that the increment in the sinusoidal disturbance became more than that for axisymmetric one for Re $_D > 31$. The linear stability theory was applied to the value of the laminar-turbulent transition coordinate as a function of the Reynolds number. A model criterion for a laminar-turbulent transition in the far jet region was proposed. For the first time, this made it possible to obtain a good agreement between the theoretical results and experimental data for Re $_D < 2000$.

Key words: laminar jet, Landau solution, hydrodynamic stability, far field

Citation: R.I. Mullyadzhanov, N.I. Yavorsky, The far field of a submerged laminar jet: Linear hydrodynamic stability, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 84–94. DOI: 10.18721/JPM.11310

Introduction

Hydrodynamic stability theory studies the conditions under which one flow regime flow of a fluid or a gas is replaced by another [1 – 3]. Situations like that often happen in a wide range of natural phenomena and technical devices; therefore, new results in this field have numerous fundamental and practical applications. Free shear flows are one of the widest classes in hydrodynamics, with jet flows playing a central role. The classical problem on the stability of a circular flooded laminar jet issuing from a local source still has no definitive solution, stimulating further interest in this issue.

It was experimentally proved that a circular jet loses stability at relatively low flow velocities. Some of the first experiments on this problem, described in [4], were carried out by Schade. These experiments indicated that stable jet flow could be obtained at Reynolds numbers about several hundred. Further, Viilu [5] obtained in 1962 a result that somewhat contradicted Schade's data, determining the critical Reynolds number in the range of only 10.5 – 11.8. In the same year, A.J. Reynolds published the results of similar experiments [6], with a fairly detailed description of the scenarios dealing with the loss of flow stability.

The inlet conditions in such experiments are often modeled with a long tube; the velocity profile at the exit of this tube should be close to the parabolic Poiseuille profile. However, the outlet characteristics are highly dependent on the length of the tube nozzle.

More thorough studies of the outlet velocity profile were carried out in a relatively recent series of experiments [7, 8]. Measurements have shown that the length of the nozzle, which is about 200 channel diameters, is sufficient for forming a parabolic velocity profile up to Reynolds numbers of about 6700. It was also found that a non-axisymmetric mode, visualized in the cross-section, starts to develop at high flow velocities and close enough to the nozzle.

Lemanov et al. [9] studied submerged jets issuing from a nozzle with a length of 100D (D is the diameter of the tube). In addition, the flow was visualized and it was established that the region of steady laminar flow decreases with increasing Reynolds number. It was found (in agreement with the results of the previous authors) that sinusoidal perturbations start to evolve in the region located before the final turbulent transition of the jet. We are going to use the experimental data in this study for qualitative and quantitative comparison with the theory presented below.

Analytical study of this problem started with a paper by Batchelor and Gill [5], who found that only sinusoidal perturbation is an unstable mode in the far field in the inviscid case. How-

ever, the study also indicated that including the expansion of the jet downstream could slightly change the conclusions obtained using a plane-parallel approximation.

Tatsumi and Kakutani [10] note that stability analysis of non-parallel flows is not sufficiently developed in hydrodynamic stability theory that regards even such flows as jets and wakes as quasi-parallel. Ling and W.C. Reynolds [11] developed an approach taking into account flow expansion within the framework of perturbation theory. Garg [12] used a more general approach, applied only to Bickley's (two-dimensional) jet [13]. In contrast to the two-dimensional case, where the perturbation characteristics vary with the axial coordinate in a non-self-similar manner and some approximations have to be applied [14, 15], the general form of the perturbations in the three-dimensional case can be written based on self-similarity considerations. This analysis was first performed by Likhachev [16] for the Schlichting jet. Aside from unstable perturbations with $m = 1$ (m is the azimuthal wave number), unstable axisymmetric modes with $m = 0$ were detected. Even though axisymmetric perturbations turned out to be the most unstable only in a narrow range of rather small Reynolds numbers, this allowed a qualitatively explain the experimentally observed axisymmetric pulsations, described by W.C. Reynolds. Recall that the perturbations with $m = 1$ are the most dangerous for relatively large Re numbers. This analysis used the Schlichting solution, which is an equivalent of the exact Landau solution in the boundary layer approximation.

Shtern and Hussain [17] carried out a similar analysis for a Landau jet. Unlike previous studies, where the dependences of the perturbation v on the axial coordinate had the form $v \propto e^{ik(x)x}$ (x is the coordinate along the direction of the jet propagation, $k(x)$ is the axial wavenumber), where the maximum value of v decreases downstream, in this case, the perturbations were considered in the form $v \propto e^{ik(R)\ln R}$ (R is the spherical radius), based on a previous study for the two-dimensional setting [14, 15]. Thus, the authors discussed perturbations as a power function of R and obtained results similar to those presented in [16]. However, only neutral solutions were considered (the imaginary part of $k = 0$).

In addition to a non-standard dependence on the spatial coordinate, perturbations also do not have a purely exponential dependence on time. Thus, stability analysis is not modal, which follows from the fact that the characteristic time in the jet problem increases as $(R/|\mathbf{u}|) \propto R^2$ downstream, where $|\mathbf{u}|$ is the local velocity on the jet axis. The perturbations, whose wavelength and characteristic pulsation time also increase with increasing R, evolve together with the main flow [12]. Based on the conclusions of [16], we can assume that if we consider the spatial evolution of a small perturbation with a fixed frequency ω_0, the neutral curve $\omega_0(\mathrm{Re})$ and the scaling $\omega_0 \propto R^{-2}$ determine the variation range of R, where this perturbation grows, for a given value of Re.

This statement was confirmed by three-dimensional calculations of the stability problem [18]. Additionally, an important remark was made in [19]: calculations of the stability problem in unbounded domains are greatly complicated by numerical difficulties from the boundary conditions at the outlet; the latter can considerably distort the results.

It follows from this brief overview that using a self-similar form of disturbances allows to avoid the above-mentioned numerical difficulties. This statement is an additional argument in favor of the self-similar approach in this problem.

Problem statement

We study the evolution of perturbations $\mathbf{v}$ of a certain laminar velocity field $\mathbf{U}$; the total velocity field is represented as $\mathbf{u} = \mathbf{U} + \mathbf{v}$. Let us substitute this representation into Navier – Stokes equations and perform linearization assuming that the velocity perturbation amplitude is small compared with the main flow. We then obtain the following equation:

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{U} \cdot \nabla)\boldsymbol{v} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{U} = -\frac{1}{\rho}\nabla\chi + \nu\Delta\boldsymbol{v}, \quad (1)$$

where χ is the pressure field perturbation, ν is the kinematic viscosity, ρ is the fluid density.

The velocity field of the main flow is described by the exact solution of Navier – Stokes equations that can be represented in spherical coordinates (R, θ, φ):

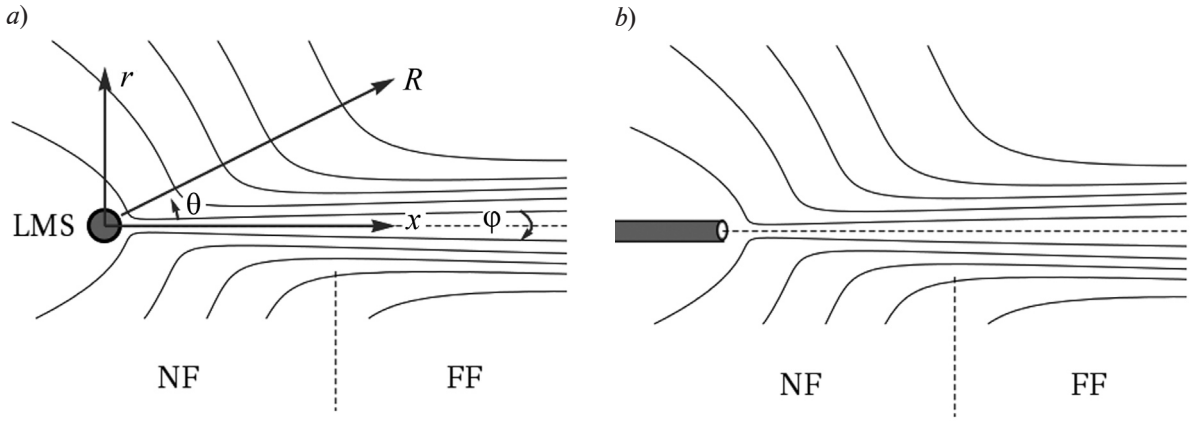

Fig. 1. Graphical representation of solution (3):
a is the idealized axisymmetric jet flow (streamlines are shown) caused by a local motion source (LMS), with coordinate systems (spherical and cylindrical); *b* is an implementation of such flow that is a fluid jet issuing from a long tube into a submerged region; NF, FF are the near and far fields, respectively

$$U_R = -\frac{\nu y'(\psi)}{R}, U_\theta = -\frac{\nu y(\psi)}{R\sqrt{1-\psi^2}},$$
$$U_\varphi = 0, y(\psi) = 2\frac{1-\psi^2}{A-\psi}, \quad (2)$$

where $\psi = \cos\theta$.

The parameter A is related as follows to the "momentum" P_x of the jet:

$$P_x = 16\pi\rho\nu^2 A\left[1 + \frac{4}{3(A^2-1)} - \frac{A}{2}\ln\frac{A+1}{A-1}\right], \quad (3)$$

This solution was obtained by Slyozkin [20], Landau [21] and Squire [22] and corresponds to jet flow caused by a point momentum source.

Fig. 1 shows a graphical representation of the solution obtained. This solution is used as the main flow in our study, since its direct comparison with experimental data yielded good agreement in the far field of the jet [23 – 25].

Since the problem statement does not include the characteristic dimension of length, for reasons of dimension, we are going to search for perturbations in the following class:

$$v_R = \frac{\nu}{R} f(\psi,\eta)e^{im\varphi},$$
$$v_\theta = -\frac{\nu}{R\sqrt{1-\psi^2}} g(\psi,\eta)e^{im\varphi}, \quad (4)$$
$$v_\varphi = \frac{\nu}{R} h(\psi,\eta)e^{im\varphi}, \chi = \frac{\rho\nu^2}{R^2} q(\psi,\eta)e^{im\varphi},$$

where the variable $\eta = \sqrt{(R/\nu t)}$.

Notably, the variables ψ and η were also used in analysis of two-dimensional [14, 15, 26, 27] and three-dimensional [28 – 30] conical flows. Using the method of variable separation, we can establish for $y = 0$ ($A \to \infty$) that the solution is expressed analytically in terms of Legendre polynomials with respect to the variable ψ and in terms of hypergeometric functions with respect to the variable η [31]. This actually means that the solution has a power dependence on η, which is not surprising as the representation of the velocity field of the main flow, constructed based on considerations of dimension, has a power dependence R^{-1}. Next, we transform the power dependence with a certain exponent n as follows:

$$\eta^n = (R/R_0)^n(\nu t/R_0^2)^{-n/2} = \exp[n\ln(R/R_0) - (n/2)\ln(\nu t/R_0^2)], \quad (5)$$

where R_0 – is some constant of the length dimension (radius of the inlet nozzle).

Evidently, if $y \neq 0$, it seems expedient to consider the problem on stability against perturbations in the form of waves in terms of new variables:

$$v = (\nu/R)v_0(\psi)\exp(ik\xi - i\omega\ln\tau + im\varphi),$$
$$\xi = \ln(R/R_0), \tau = \nu t/R_0^2, \quad (6)$$

where v_0 is a dimensionless vector depending only on the angle ψ; k and m are the radial and azimuthal dimensionless wavenumbers; ω

is the dimensionless frequency, τ is the dimensionless time.

Then the components of perturbations of the velocity and pressure fields have the form:

$$v_R = (\nu / R) f(\psi) \exp(ik\xi - i\omega \ln \tau + im\varphi),$$

$$v_\theta = \frac{\nu}{R\sqrt{1-\psi^2}} g(\psi) \exp(ik\xi - i\omega \ln \tau + im\varphi), \quad (7)$$

$$v_\varphi = \frac{\nu}{R\sqrt{1-\psi^2}} ih(\psi) \exp(ik\xi - i\omega \ln \tau + im\varphi),$$

$$\chi = \frac{\rho\nu^2}{R^2} q(\psi) \exp(ik\xi - i\omega \ln \tau + im\varphi),$$

where f, g, h, q are the dimensionless functions of only the angular variable ψ.

Substituting representation (7) into Eqs. (1), and making some transformations, we obtain a system of ordinary differential equations:

$$i\Omega f + \frac{2mh}{1-\psi^2} + (2-ik)q - 2g' + y''g + \frac{2yg}{1-\psi^2} - \left(2 + ik + k^2 + \frac{m^2}{1-\psi^2}\right) f - (2-ik)y'f - yf' - 2\psi f' + (1-\psi^2) f'' = 0;$$

$$i\Omega g + mh' + (1-\psi^2)(2f - (1+ik) f') - \left(ik + k^2 - \frac{m^2}{1-\psi^2}\right) g - (1-ik)y'g - \frac{2\psi yg}{1-\psi^2} - yg' - (1-\psi^2)q' = 0;$$

$$i\Omega h - mq + 2mf - \frac{2m\psi g}{1-\psi^2} - \left(ik + k + \frac{m^2}{1-\psi^2}\right) h + iky'h - yh' + (1-\psi^2)h'' = 0;$$

$$(1+ik) f + g' - \frac{mh}{1-\psi^2} = 0, \quad (8)$$

where $\Omega = \omega R^2 / (\nu t)$ is some constant parameter acting as the generalized frequency; it includes the dependence on the radius and time (proportional to the variable η^2).

The order of the derivative of the function g is reduced from the second to the first in the second equation of system (8) using the continuity equation. We should note that Eqs. (8) are identical to the equations obtained by Shtern and Hussain (who actually considered the exponential dependence of perturbation on time, or, more precisely, on $1/\eta^2$, used the far-field approximation ($\eta \to \infty$, which is equivalent to $\tau \to 0$) to derive the equations, and discarded some terms with high powers of τ). No approximations have been used in our study to derive these equations, except that Ω is assumed to be a constant parameter.

For complete statement of the problem, system of equations (8) should be supplemented with suitable boundary conditions. The following conditions imposed on the velocity field follow from representation (7):

$$g(\pm 1) = 0, h(\pm 1) = 0, \quad (9)$$

meeting the requirements that functions g and h be bounded.

Procedure of numerical solution

The procedure for numerical solution of the resulting system of equations is shown schematically in Fig. 2. Since the points $\psi = \pm 1.0$ are singular, we need to find the asymptotic expansion of the functions of the problem in their neighborhood of these points and shift the start of numerical integration. Asymptotic expansions of a certain test function Ψ in the neighborhood of singular points $\psi = \pm 1.0$ are used in the ranges $\psi \in [-1.0; \psi_c]$ and $[\psi_p; 1.0]$ (see expansion (10)). Next, two solutions of Eqs. (8) are constructed by numerical integration from ψ_c to ψ_m and from ψ_p to ψ_m. The values of the function Ψ and its derivatives should be kept continuous at the point ψ_m, in accordance with the order of the system of differential equations (see conditions (11)).

It can be shown for Legendre-type equations [32] that the functions of the problem are proportional to the factor $(1 - x^2)^{m/2}$ and a certain analytical (in the neighborhood of $\psi = \pm 1.0$) function, which, in turn, can be represented as a Taylor series.

Thus, some test function Ψ (f, g, h or q) in the neighborhood of the point $\psi = 1.0$ can be represented in the following form:

$$\Psi = (1-\psi^2)^{m/2} (\Psi_0 + \Psi_1(1-\psi) + \Psi_2(1-\psi^2) + \Psi_3(1-\psi^3) + ...), \quad (10)$$

where the complex-valued expansion coefficients $\Psi_0, \Psi_1, \Psi_2, \Psi_3$ are determined by substituting function (10) into system of equations

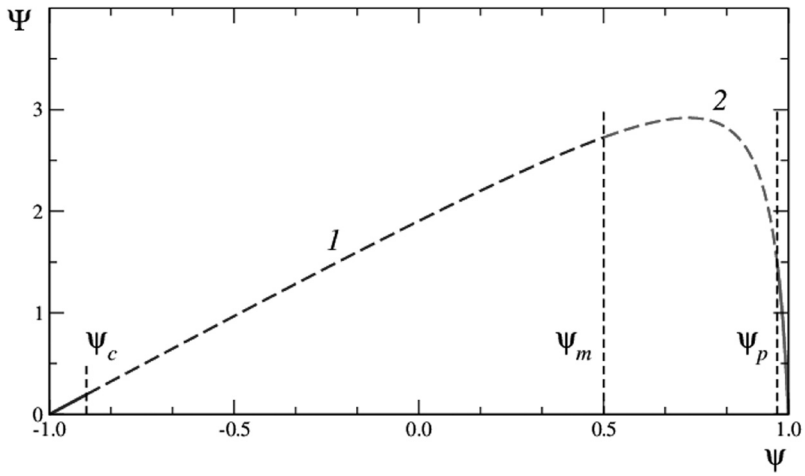

Fig. 2. Scheme of the numerical algorithm used:
Asymptotic expansions of a certain test function Ψ in the neighborhood of singular points $\psi = \pm 1.0$ are used in the ranges $\psi \in [-1{,}0;\ \psi_c]$ and $[\psi_p;\ 1{,}0]$ dashed curves *1*, *2* are the domains of further numerical integration; the condition that the values of the function ψ and its derivatives be continuous is imposed at ψ_m

(8). Some parameters remain undefined (free); these should be found by actually solving the spectral problem.

A decomposition similar to expression (10) can also be written in the neighborhood of the point $\psi = -1, 0$. Next, two numerical solutions need to be constructed for selected values of A (in the function y), Ω and a set of free parameters, and integration of Eqs. (8) starts from the points $\psi_c = -1.0 + \varepsilon_c$ and $\psi_p = 1.0 - \varepsilon_p$, where ε_c and ε_p are small parameters (in the range $10^{-5} - 10^{-3}$). The continuity conditions for the functions of the problem and their derivatives should be satisfied, in accordance with the order of the system of ordinary differential equations, at some point ψ_m ($\psi_m = 0.9$ for the solutions found below); the choice of this point does not affect the result. Namely, the following conditions should be fulfilled:

$$\begin{gathered} f_-(\psi_m) = f_+(\psi_m),\ f_-'(\psi_m) = f_+'(\psi_m), \\ g_-(\psi_m) = g_+(\psi_m), \\ h_-(\psi_m) = h_+(\psi_m),\ h_-'(\psi_m) = h_+'(\psi_m), \\ q_-(\psi_m) = q_+(\psi_m), \end{gathered} \qquad (11)$$

where plus and minus correspond to the solutions obtained by integrating the system of equations from the points ψ_p and ψ_c, respectively.

Conditions (11) are achieved by varying the free parameters and the wavenumber $k = k_{re} + ik_{im}$ using Newton's method. We used a similar calculation scheme in [33].

Results and discussion

Increasing perturbations (at $-k_{im} > 0$) were detected only for azimuthal wavenumbers $m = 0$ and $m = 1$, the same as in [17], which, however, discussed only neutral perturbations ($k_{im} = 0$). Thus, the k_{im}(Re) dependence was not analyzed in [17], which actually makes it possible for us to carry out a comprehensive comparison with experimental data, as will be shown below. It is convenient to use the Reynolds number, constructed from the velocity on the axis and the distance from the origin, in this problem:

$$\mathrm{Re} = \frac{U_R R}{\nu}\Big|_{\psi=1} = -y'(1) = -\frac{4}{A-1}, \qquad (12)$$

according to exact solution (2).

Fig. 3 shows the dispersion curves $-k_{im}(\Omega)$ for different Reynolds numbers Re and $m = 0$. As the Reynolds number increases above the critical value $\mathrm{Re}_{crit}^{m=0} = 26.20$, a range of Ω values appears, for which solutions exist, with $-k_{im} > 0$. Notably, $\mathrm{Re}_{crit}^{m=0} = 28.1$ was found in [17]. The small difference can be explained by the insufficiently accurate algorithm for calculating the spectral problem used in [17], where asymptotic expansions of the functions of the problem in the neighborhood of the points $\psi = \pm 1.0$ were not used.

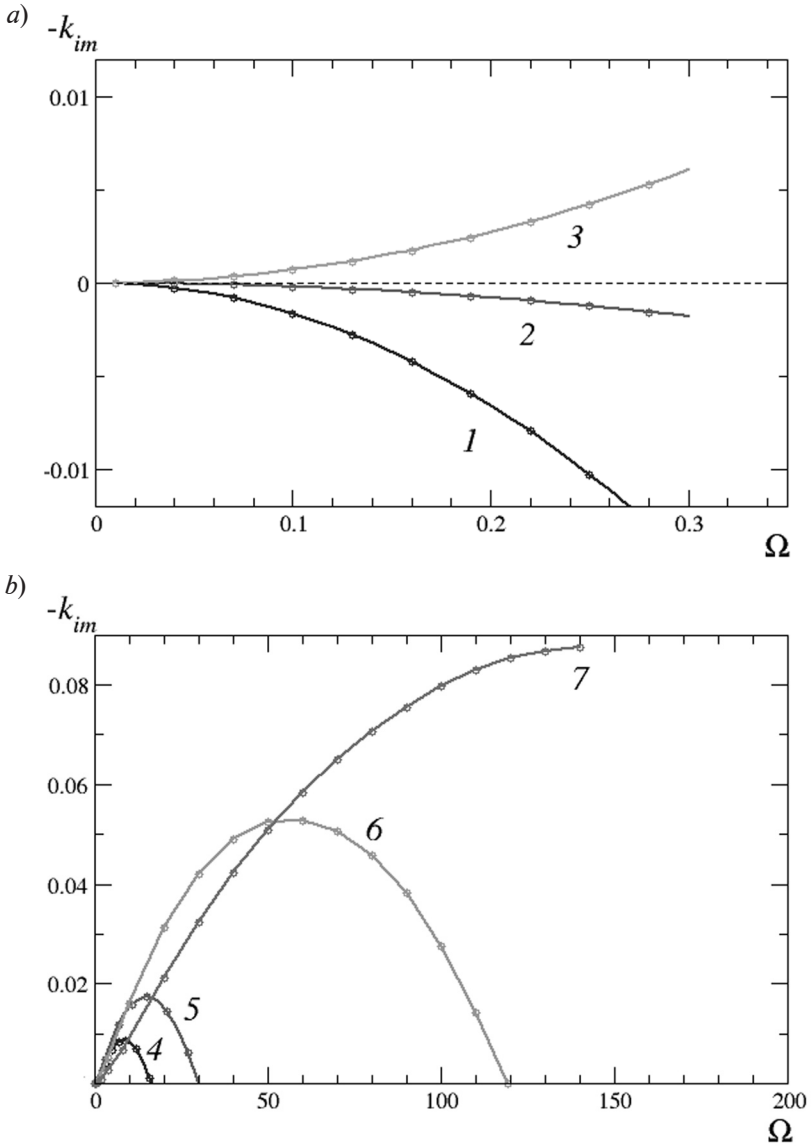

Fig. 3. Dispersion curves $-k_{im}(\Omega)$ in the ranges of the parameter Ω equal to (0 − 0.35) (*a*) and (0 − 200) (*b*), for the most unstable solution with $m = 0$, with different Reynolds numbers Re:
20 (*1*), 25 (*2*), 33,33 (*3*), 40 (*4*), 50 (*5*), 100 (*6*) и 200 (*7*)

The current statement of the problem allows studying the evolution of perturbations in the entire space, thanks to self-similarity of the main flow and the given perturbations, and is thus global. The ratio of the amplitude of perturbation velocity on the axis to the velocity of the main flow obeys the following relationship:

$$v_R / U_R = \left|[(\nu / R) f(1) e^{-k_{im}(\mathrm{Re})\xi}] \times \right.$$
$$\left. \times [(-\nu / R) y'(1)]^{-1} \right| \infty (R / R_0)^{-k_{im}(\mathrm{Re})}. \quad (13)$$

The perturbation amplitude algebraically grows or decays downstream relative to the main flow, depending on the distance measured from the origin. The growth rate is determined by the imaginary component of the wavenumber and depends on the Reynolds number. The absolute value of $-k_{im}(\mathrm{Re})$ turns out to be critical in this case.

Fig. 4 shows the dependence of the maximum value of $-k_{im}(\Omega)$, obtained for each dispersion curve, with different Re numbers. For

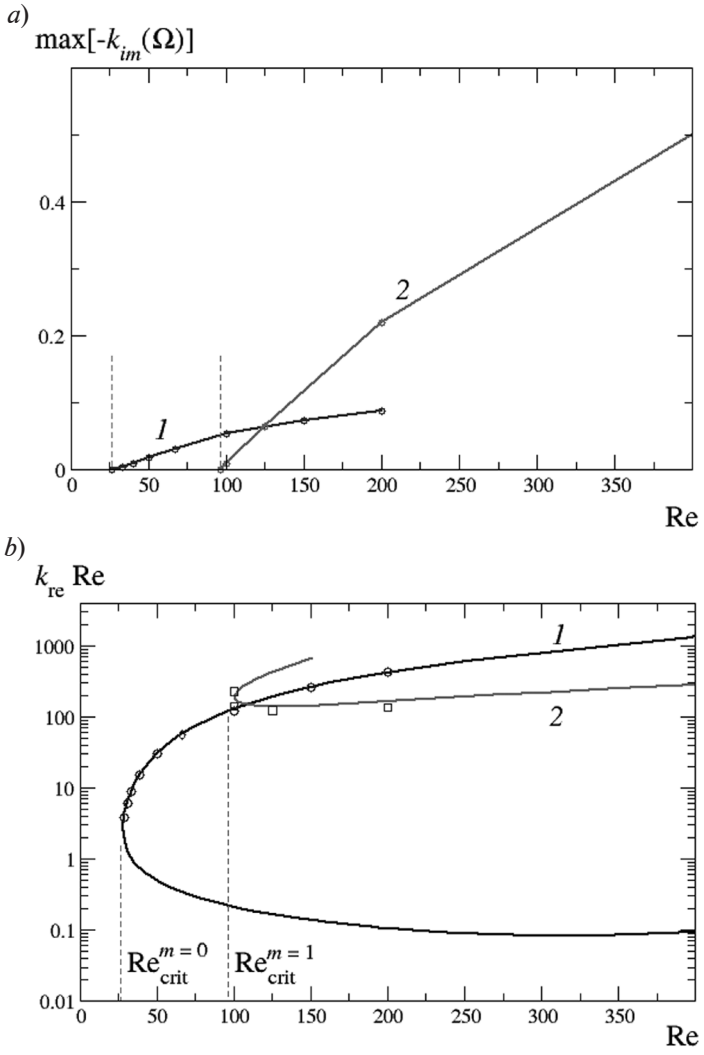

Fig. 4. Dependences of the maximum value of the imaginary component $-k_{im}$ (*a*) and the value of the real component k_{re}Re (*b*) of the wavenumber k on the Reynolds number for the most unstable solutions with $m = 0$ (*1*) and $m = 1$ (*2*). Fig. 4,*b* shows a comparison of the data in our study (symbols) with those in [17] (solid lines). The values $\mathrm{Re}_{crit}^{m=0} = 26.20$ and $\mathrm{Re}_{crit}^{m=1} = 96.29$ are marked with vertical dashes

instance, it is evident that even though positive $-k_{im}$ values exist for $\mathrm{Re} \le 40$, these values do not exceed 0.01. This suggests that the ratio of the perturbation amplitude to the main flow velocity on the axis increases by only 7 % (approximately) at a distance $R/R_0 = 10^3$, compared with this ratio at a distance $R/R_0 = 1$. With Re = 200, the peak $-k_{im}(\Omega)$ value on the dispersion curve is reached for $-k_{im} = 0.087$. The perturbation increases by 82 % at a distance $-k_{im}$ for these parameters.

Thus, we can conclude that, despite a mechanism of growing axisymmetric perturbations present in the given flow, the rate of such growth turns out to be extremely low. For this reason, axisymmetric perturbations can be characterized as neutrally stable in the first approximation. It is probably due to the weakly pronounced perturbation effect at $m = 0$ that only stable solutions are valid in the plane-parallel approximation. Fig. 4,*b* shows a comparison of the dependence of k_{re}Re on the Reynolds number Re, obtained in this study and in [17]. The curve representing the data from [17] demonstrably lacks a lower branch.

It can also be seen from Fig. 4,*b* that an

unstable solution for $m = 1$ appears if Re is increased to a value of the order of 100. The obtained value of the critical Reynolds number is $\mathrm{Re}_{crit}^{m=1} = 96.29$, which is slightly less than the corresponding value in [17] ($\mathrm{Re}_{crit}^{m=1} = 101$).

Comparing the maximum $-k_{im}(\Omega)$ values as functions of the Reynolds number Re with $m = 0$ and $m = 1$ indicates that the perturbation growth rate for $m = 1$ substantially exceeds that for $m = 0$ as the Reynolds number increases above a certain value. In this case, the maximum $-k_{im}$ values are approximately the same for $\mathrm{Re} \approx 120 - 130$.

The next stage of our study consisted in comparing the results of the above-described linear stability analysis with the experimental data given in the literature. We are going to find the relationship between the Reynolds number with the value of this number, used in experiments and numerical calculations, expressed by formula (12). The number constructed by the diameter of the exit nozzle $D = 2R_0$ and the mean flow rate U_b has the form

$$\mathrm{Re}_D = U_b D / \nu. \qquad (14)$$

Let us consider the parabolic velocity profile formed in the inlet nozzle. In cylindrical coordinates (x, r, φ) with the center in the middle of the exit section ($x = 0$), this profile has the following form:

$$U(R) = 2U_b(1 - R^2 / R_0^2), \qquad (15)$$

where R_0 is, the same as above, the radius of the tube.

The total momentum flux through the exit section is determined by the following relationship:

$$P_x = \oint \rho U^2(R) dS = 2\pi\rho \int_0^{R_0} U^2(R) R dR. \qquad (16)$$

Substituting formula (15) into relation (16), we obtain that

$$P_x = \frac{1}{3}\pi\rho\nu^2 \mathrm{Re}_D^2. \qquad (17)$$

Thus, we arrive at the following relationship:

$$\mathrm{Re}_D = \sqrt{\frac{3P_x}{\pi\rho\nu^2}}. \qquad (18)$$

Therefore, there is an explicit relationship between Re_D and Re (or between A and Re: $\mathrm{Re} = -4/(A - 1)$). The following asymptote can be written for large values of the Reynolds number:

$$\mathrm{Re}_D = \sqrt{8\mathrm{Re}} + \\ +\sqrt{2}(8 + \ln 8 - 3\ln \mathrm{Re})\mathrm{Re}^{-1/2} + \ldots, \mathrm{Re} \to \infty, \qquad (19)$$

with the first term often used in the literature ($\mathrm{Re}_D = \sqrt{8\mathrm{Re}}$).

The results of the analysis results for $m = 1$, obtained in this study, are compared in Table with the results of other authors. Notice that the

Table

Comparison of results obtained by different authors for analysis of linear stability of the Landau jet with the azimuthal wavenumber $m = 1$

Author	Re_{crit}	$\mathrm{Re}_{D,crit}$	$k_{re,crit}$	Ω_{crit}
V. Shtern, F. Hussain [17]	101.0	27.77	1.85	84.00
P.J. Morris [34]	177.1	37.64	2.12	86.66
O.A. Likhachev [16]	94.46	27.49	1.55	59.72
Our study	96.29	27.10	1.78	76.93

Notations: Re is the Reynolds number determined by formula (12), Re_D is the Reynolds number constructed from the diameter D of the exit nozzle; $k_{re,}$ is the real component of the wavenumber k; Ω is the parameter acting as the generalized frequency; the subscript "*crit*" indicates the critical value.

Notes. 1. The stability of the velocity profile was studied in [34] in a plane-parallel approximation using the Schlichting solution. 2. The same approach as in our study was used in [17].

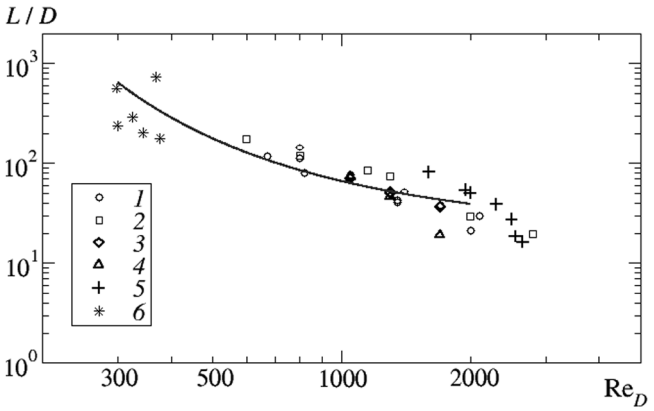

Fig. 5. Theoretical (line) and experimental (symbols) dependences of the distance at which the jet becomes turbulent versus the Reynolds number plotted for D.
The experimental data from [6, 9] were used, the theoretical curve was obtained in the present study. The diameter of the nozzle in [6] was $D = 0.32$ mm (symbols *6*). The experimental conditions in [9]: $D = 0.5$ mm (symbols *1*, *2*); 1.0 mm (*3*, *4*); 3.5 mm (*5*); velocity fluctuations were measured with a hot-wire anemometer (*2*, *4*) and visually (*1*, *3*, *5*)

critical Reynolds number Re_{crit} is significantly lower if the expansion of the jet is taken into account; however, the values of the Reynolds number $\mathrm{Re}_{D,crit}$ differ less in this case. The values of the real part of the wavenumber and the generalized frequency are also slightly lower. Nevertheless, the data obtained by Shtern and Hussain, as well as by Likhachev, are in good agreement with the results of our calculations.

Next, we estimated the distance L from the source of the jet, at which the perturbation amplitude takes some critical value, because the flow becomes turbulent. It is assumed in the calculations that perturbation grows by formula (13), in accordance with the given linear mechanism. Obviously, it is important to determine the criterion of laminar-turbulent transition in this case.

We assumed that the laminar-turbulent transition occurs when the perturbation amplitude significantly exceeds the local velocity at some point. By measuring this distance and using the dependences we found for $-k_{im}(\mathrm{Re})$ and formula (13), we obtained the dependence of L on Re.

Fig. 5 shows a comparison of the experimental data obtained by A.J. Reynolds [6] and by Lemanov et al. [9] with the theoretical dependence we have found (shown by a solid line).

The resulting expression has the form

$$L / D = 2,0 \cdot 10^{\alpha},$$
$$\alpha = 1 + 1 / (0,0081\mathrm{Re}_D{}^{0,8} - 0,11);$$

it was found by extrapolating the function $-k_{im}(\mathrm{Re})$ to higher values of the Reynolds number.

Comparison of theoretical and experimental results yields a good quantitative agreement, even though the turbulent processes, including the stage of nonlinear perturbation growth, are far more complex in reality, compared to the model. Importantly, turbulent fluctuations were observed in the flow from the tube even with $\mathrm{Re}_D > 2000$ (according to data obtained by Lemanov), which limits the range for comparison of theoretical and experimental data to $\mathrm{Re}_D < 2000$.

Conclusion

We have considered the linear stability problem for a submerged Landau – Squire jet. We have established that the amplitude of intrinsic perturbations spatially varies as a power function of the spherical radius R, read from the source of motion.

We have obtained a problem on the eigenvalues, which is solved numerically. Unstable perturbations were found for the first two azimuthal wavenumbers ($m = 0$ and 1); at the same time, the corresponding critical values of the Reynolds number, constructed

from the mean-flow velocity with a parabolic distribution inside the nozzle and its diameter, were

$\mathrm{Re}_D^{m=0} = 13.98$;

$\mathrm{Re}_D^{m=1} = 27.10$,

respectively.

We have confirmed that the increment for the growth of sinusoidal perturbations increases with values of $\mathrm{Re}_D > 31$, i.e., greater than that of axisymmetric perturbations.

We have proposed a model criterion for the laminar-turbulent transition in the far field of the jet, based on the fact that the ratio of the amplitude of the perturbation velocity to the main flow velocity spatially varies as a power function of R; the growth increment is known from the solution of the formulated spectral problem.

We have obtained for the first time a good agreement between the results of linear stability theory and the experimental data with $\mathrm{Re}_D < 2000$ for the coordinate of the laminar-turbulent transition as a function of the Reynolds number.

This study was financially supported by the grant of the Russian Science Foundation No. 14-19-01685.

REFERENCES

[1] **D.D. Joseph,** Stability of fluid motions, Vol. I. Springer-Verlag, Berlin-Heidelberg-New York, 1976.

[2] **P. Drazin, W. Reid,** Hydrodynamic stability. Cambridge University Press, Cambridge, 1981.

[3] **P.J. Schmid, D.S. Henningson,** Stability and transition in shear flows, Applied Mathematical Sciences, Springer, Berlin, 142 (2001).

[4] **G.K. Batchelor, A.E. Gill,** Analysis of the stability of axisymmetric jets, Journal of Fluid Mechanics. 14 (4) (1962) 529–551.

[5] **A. Viilu,** An experimental determination of the minimum Reynolds number for instability in a free jet, Journal of Applied Mechanics. 29 (3) (1962) 506–508.

[6] **A.J. Reynolds,** Observations of a liquid-into-liquid jet, Journal of Fluid Mechanics. 14 (4) (1962) 552–556.

[7] **G.V. Kozlov, G.R. Grek, A.M. Sorokin, Yu.A. Litvinenko,** Influence of initial conditions at the nozzle exit on the structure of round jet, Thermophysics and Aeromechanics. 15 (1) (2008) 55–68.

[8] **V.V. Kozlov, G.R. Grek, G.V. Kozlov, M.V. Litvinenko,** Subsonic round and plane jets in the transversal acoustic field, Vestnik NGU. Ser. Physics. 5(2) (2010) 28–42.

[9] **V.V. Lemanov, V.I. Terekhov, K.A. Sharov, A.A. Shumeyko,** An experimental study of submerged jets at low Reynolds numbers, Technical Physics Letters. 39 (5) (2013) 421–423.

[10] **T. Tatsumi, T. Kakutani,** The stability of a two-dimensional laminar jet, Journal of Fluid Mechanics. 4 (3) (1958) 261–275.

[11] **Chi-Hai Ling, W.C. Reynolds,** Non-parallel flow corrections for the stability of shear flows, Journal of Fluid Mechanics. 59 (3) (1973) 571–591.

[12] **V.K. Garg,** Spatial stability of the non-parallel Bickley jet, Journal of Fluid Mechanics. 102 (1981) 127–140.

[13] **W.G. Bickley,** The plane jet, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 23 (156) (1937) 727–731.

[14] **K.K. Tam,** Linear stability of the non-parallel Bickley jet, Canadian Applied Mathematics Quarterly. 3 (1) (1995) 99–110.

[15] **A. McAlpine, P.G. Drazin,** On the spatio-temporal development of small perturbations of Jeffery–Hamel flows, Fluid Dynamics Research. 22 (4) (1998) 123–138.

[16] **O.A. Likhachev,** Analysis of stability in a self-similar circular jet with consideration of the nonparallel effect, Journal of Applied Mechanics and Technical Physics. 31 (4) (1990) 621–626.

[17] **V. Shtern, F. Hussain,** Effect of deceleration on jet instability, Journal of Fluid Mechanics. 480 (2003) 283–309.

[18] **X. Garnaud, L. Lesshafft, P.J. Schmid, P. Huerre,** Modal and transient dynamics of jet flows, Physics of Fluids. 25 (4) (2013) 044103.

[19] **L. Lesshafft,** Artificial eigenmodes in truncated flow domains, Theoretical and Computational Fluid Dynamics. 32 (3) (2018) 245–262.

[20] **N.A. Slyozkin,** Ob odnom sluchaye integriruyemosti polnykh differentsialnykh uravneniy dvizheniya vyazkoy zhidkosti [On a case of integrability of the complete differential equations of the motion of viscous fluid], Uchenyye zapiski MGU. (2) (1934) 89–90.

[21] **L.D. Landau,** Ob odnom novom tochnom reshenii uravneniy Navye –Stoksa [On an innovative exact solution to the Navier – Stokes equations], Doklady USSR AS. 43 (7) (1944) 299–301.

[22] **H.B. Squire,** The round laminar jet, The Quarterly Journal of Mechanics and Applied Mathematics. 4 (3) (1951) 321–329.

[23] **E.N. Andrade da C., L.C. Tsien,** The velocity-distribution in a liquid-into-liquid jet , Proceedings of the Physical Society. 49 (4) (1937) 381–391.

[24] **G.W. Rankin, K. Sridhar, M. Arulraja, K.R. Kumar,** An experimental investigation of laminar axisymmetric submerged jets, Journal of Fluid Mechanics. 133 (1983) 217–231.

[25] **B.J. Boersma, G. Brethouwer, F.T.M. Nieuwstadt,** A numerical investigation on the effect of the inflow conditions on the self-similar region of a round jet, Physics of Fluids. 10 (4) (1998) 899–909.

[26] **A.V. Shapeev,** Unsteady self-similar flow of a viscous incompressible fluid in a plane divergent channel, Fluid Dynamics. 39 (1) (2004) 36 –41.

[27] **A.V. Shapeev,** Issledovaniye smeshannoy spektralno-raznostnoy approksimatsii na primere zadachi o vyazkom techenii v diffuzore [Studies in the mixed spectral-difference approximation illustrated by the example of the problem on the viscous flow in a diffuser], Numerical Analysis and Applications. 8 (2) (2005) 149–162.

[28] **C. Sozou, W.M. Pickering,** The round laminar jet: the development of the flow field, Journal of Fluid Mechanics. 80 (4) (1977) 673–683.

[29] **C. Sozou,** Development of the flow field of a point force in an infinite fluid, Journal of Fluid Mechanics. 91 (3) (1979) 541–546.

[30] **B.J. Cantwell,** Transition in the axisymmetric jet, Journal of Fluid Mechanics. 104 (1981) 369–386.

[31] **R.I. Mullyadzhanov,** Zatoplennyye struynyye MGD techeniya [Submerged MHD jet flows], Thesis for Ph.D., Kutateladze Institute of Thermal Physics, SO RAS, Russian Federation, 2012.

[32] **V. Shtern, F. Hussain,** Instabilities of conical flows causing steady bifurcations, Journal of Fluid Mechanics. 366 (1998) 33–85.

[33] **R.I. Mullyadzhanov, N.I. Yavorsky,** On the self-similar exact MHD jet solution, Journal of Fluid Mechanics. 746 (2014) 5–30.

[34] **P.J. Morris,** The spatial viscous instability of axisymmetric jets, Journal of Fluid Mechanics. 77 (3) (1976) 511–529.

Received 23.03.2018, accepted 23.05.2018.

THE AUTHORS

MULLYADZHANOV Rustam I.
Kutateladze Institute of Thermal Physics
1 Acad. Lavrentiev Ave., Novosibirsk, 630090, Russian Federation
rustammul@gmail.com

YAVORSKY Nikolay I.
Kutateladze Institute of Thermal Physics
1 Acad. Lavrentiev Ave., Novosibirsk, 630090, Russian Federation
nick@itp.nsc.ru

THREE-FLUID FORMULATION AND A NUMERICAL METHOD FOR SOLVING THE STATIONARY PROBLEM OF THERMAL HYDRAULICS OF A TWO-PHASE ANNULAR DISPERSED FLOW

E.E. Avdeev, A.A. Pletnev, S.V. Bulovich

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation

The article presents a one-dimensional three-fluid model for solving the stationary flow problem on a two-phase steam-water annular dispersed stream in a vertical heated channel, based on nine balance equations with an equal phase pressure assumed. The validation of the marching algorithm of the described stationary problem has been carried out by comparison with the published experimental data relating to a two-phase flow in a circular pipe under adiabatic conditions for pressures of 3 – 9 MPa, total flow rates of 500 – 3000 kg /(m^2·s) and internal diameters of 10 and 20 mm. The total pressure differences in the channel were calculated. Good qualitative agreement with experimental data was obtained. Small quantitative disagreements were found. They were shown to be reduced by the refinement of the closing relations.

Key words: two-phase steam-water flow, three-fluid model, numerical solution

Citation: E.E. Avdeev, A.A. Pletnev, S.V. Bulovich, Three-fluid formulation and a numerical method for solving the stationary problem of thermal hydraulics of a two-phase annular dispersed flow, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 95–103. DOI: 10.18721/JPM.11311

Introduction

The most common approach for describing multiphase flows in a thermal hydraulic approximation is based on the model of interpenetrating continua. The existing simulation codes for thermal hydraulics (KORSAR, TRAC, RELAP) are based on the so-called two-fluid approximation describing the case for multiphase flow represented by two "fluids": vapor and fluid. The corresponding system consists of six differential equations of mass, momentum and energy balance (for each of the fluids). The system is closed by the phase balance equation, the thermodynamic equation of state (for each of the fluids) and empirical or semi-empirical relations (for each of the fluids). These relations, called closing, describe mass, heat and momentum transfer between the individual phases and between the phases and the channel wall.

The structure of the two-phase medium and, accordingly, the closing relations in such a model are determined by the chosen flow regime map.

The two-fluid approximation was the first one used in simulation codes developed for thermal hydraulics. It is still found in many codes and is often discussed in literature. The two-fluid approach is fairly up-to-date, providing adequate descriptions for all flow regimes where it is more correct to divide two-phase flow into any two components. This includes, for example, bubble, slug, dispersed and stratified flows. The two-fluid approach yields acceptable results for all of these flow regimes; the descriptions differ only in the definitions of "fluids" in each case and, accordingly, in the closing relations.

On the other hand, the medium in the annular dispersed flow regime is divided into three fluids: vapor, droplets, and liquid film; accordingly, each of the fluids should have its own velocity and temperature, which is impossible in the two-fluid approximation. There are approximate methods for describing the specifics of the annular dispersed-regime for these cases, staying within the framework of the two-fluid model. For example, in the KORSAR simulation code, the emerging droplets are taken into account by an additional term in the balance equations but still do not have their own velocity and temperature.

The issue of inaccurate description of the annular dispersed regime is solved by replacing the two-fluid approach with the three-fluid one

where each of the fluids (i.e., vapor, droplets, and liquid film) is described by its own equations for the balance of mass, momentum, and energy. The three-fluid model is well-known: it has been considered in some Russian and foreign studies [1 – 7]. While using the approach with three fluids instead of two is nothing new from an algorithmic standpoint, it allows to construct a more complete physical model that is non-equilibrial with respect to the velocities and temperatures of the given fluids.

The three-fluid model is more accurate in describing the annular dispersed flow regime, which means that more accurate values can be obtained for the characteristics of heat transfer, friction loss, volume fractions of fluids, and, ultimately, the position of the dryout point can be determined more reliably. Using this model means that the problem of closing relations has to be considered again. Most of the correlations tested were formulated for the two-fluid approximation. However, formulation of the three-fluid model makes it possible to partially use the experience in accumulated solving two-phase problems. For example, closing relations describing the exchange processes between vapor and droplets in dispersed flow or exchange with the channel wall in single-phase flow can be used for describing exchange processes at the corresponding interfaces and in the three-fluid approximation. In this case, the exchange processes directly between the droplets and the film are characteristic for the three-fluid model distinguishing it from the two-fluid one, allowing to abandon the equilibrium model of generation and deposition move on to non-equilibrium models by processing the experimental data.

Our study presents a marching algorithm for the numerical solution of the stationary problem of annular dispersed flow in a one-dimensional approximation using the three-fluid approach. Systems considered in a two- and three-fluid approximation in a non-stationary formulation for the case of equilibrium pressure (mathematical models with total pressure in fluids) are known to lose their evolutionary properties with certain values of operating parameters [8]. The correctness of the Cauchy problem can be restored by different measures but all of them "perturb" the original system of equations. Therefore, if a stationary flow regime is possible for a given problem, then the solution obtained by the marching method of integration is a reference for evolutionary problems, allowing to quantitatively estimate the distortions introduced into the solutions by the techniques applied for regularizing the problem.

Mathematical model

The developed stationary one-dimensional three-fluid model describes annular dispersed flow with masses of vapor, droplets and liquid film moving in two-phase vapor-liquid flow. The model takes into account phase transition, exchange processes with the channel wall, entrainment and deposition of droplets on the film surface. Accordingly, we are going to solve a system of equations consisting of nine differential equations of mass, momentum and energy balance for each of the three liquids.

The differential equations solved are written in the following form.

1) *Mass balance equations.*

For vapor:

$$\frac{\partial}{\partial x}(A\alpha_v\rho_v u_v) = m_{dv}\Pi_{id} + m_{fv}\Pi_{if}, \qquad (1)$$

where x, m, is the axial coordinate; A, m^2, is the cross-sectional area of the channel; α_v is the volume fraction of vapor; ρ_v, kg/m^3, is its density; u_v, m/s, is its velocity; m_{dv}, m_{fv}, kg/(m^{2}s), are the mass sources of vapor generation from the droplets and the film, respectively (positive for condensation and negative for evaporation); Π_{id}, Π_{if}, m, are the perimeters of the surface of heat transfer with vapor for the droplets and the film respectively; the subscripts v, d, f refer to vapor, droplets, and liquid film, respectively. Below we are going to use another subscript w, referring to the channel wall.

For droplets:

$$\frac{\partial}{\partial x}(A\alpha_d\rho_d u_d) = -m_{dv}\Pi_{id} - \Pi_{if}(S_d - S_e), \quad (2)$$

where α_d, ρ_d are the volume fraction and density of the droplets, respectively; u_d, m/s, is their velocity; S_e, S_d, kg/(m^2·s), are, respectively, the entrainment and deposition rates for droplets on the surface of the liquid film.

For liquid film:

$$\frac{\partial}{\partial x}(A\alpha_f \rho_f u_f) = -m_{fv}\Pi_{id} + \Pi_{if}(S_d - S_e), \quad (3)$$

where α_f and ρ_f are, respectively, the volume fraction and the density of the liquid film; uf, m/s, is its flow velocity.

2) *Momentum balance equations.*

For vapor:

$$\frac{\partial}{\partial x}(A\alpha_v \rho_v u_v^2) + \bar{\alpha}_v \bar{A}\frac{\partial P}{\partial x} = m_{dv}\Pi_{id}(u_{di} - u_v) + \\ + m_{fv}\Pi_{if}(u_{fi} - u_v) - \Pi_{if}\tau_{vf} - \Pi_{id}\tau_{vd} + A\alpha_v \rho_v g_x, \quad (4)$$

where P, Pa, is the pressure; u_{di}, u_{fi}, m/s, are the interfacial velocities (for the vapor – droplets and vapor – film interfaces, respectively). We used the interfacial velocity u_{di} between the droplets and the vapor, assumed to be equal to the droplet velocity u_d. A dependence from the CARHARE code [9] was used for the velocity of the interface between the liquid film and the vapor:

$$u_{fi} = \frac{\alpha_v}{\alpha_v + \alpha_f}u_f + \frac{\alpha_f}{\alpha_v + \alpha_f}u_v;$$

τ_{vf}, τ_{vd}, kg/(s^{2}m) are the shear stresses (vapor – liquid film and vapor – droplets, respectively); g_x, m/s^2, is the projection of the gravity vector on the x axis; the overbar indicates the averaged value of a quantity.

For droplets:

$$\frac{\partial}{\partial x}(A\alpha_d \rho_d u_d^2) + \bar{\alpha}_d \bar{A}\frac{\partial P}{\partial x} = -m_{dv}\Pi_{id}(u_{di} - u_v) + \\ + \Pi_{id}\tau_{vd} + A\alpha_d \rho_d g_x - \Pi_{if}(S_d u_d - S_e u_f). \quad (5)$$

For liquid film:

$$\frac{\partial}{\partial x}(A\alpha_f \rho_f u_f^2) + \bar{\alpha}_f \bar{A}\frac{\partial P}{\partial x} = \\ = -m_{fv}\Pi_{if}(u_{fi} - u_v) - \Pi_{wf}\tau_{wf} + \Pi_{if}\tau_{vf} + \\ + A\alpha_f \rho_f g_x + \Pi_{if}(S_d u_d - S_e u_f), \quad (6)$$

where Π_{wf}, m, is the perimeter of the surface of heat transfer between the channel wall and the film; τ_{wf}, kg/(s^2·m), is the shear stress between the liquid film and the channel wall.

3) *Energy balance equations.*

For vapor:

$$\frac{\partial}{\partial x}(A\alpha_v \rho_v u_v H_v) = (\mathrm{HTC})_{vd}\Pi_{id}(T_{sat} - T_v) + \\ + \left(h_v + \frac{u_{di}^2}{2}\right)m_{dv}\Pi_{id} + (\mathrm{HTC})_{vf}\Pi_{if}(T_{sat} - T_v) + \\ + \left(h_v + \frac{u_{fi}^2}{2}\right)m_{fv}\Pi_{if}, \quad (7)$$

where H_v, J/kg, is the total specific enthalpy of vapor, $H_v = h_v + 0.5u^2$ (h_v is the specific enthalpy); $(\mathrm{HTC})_{vd}$, $(\mathrm{HTC})_{vf}$, W/(m^2·K), are, respectively, the coefficients of heat transfer from the vapor to the interface with the droplets and from the vapor to the interface with the liquid film; T_{sat}, T_v, K, are the saturation and vapor temperatures, respectively.

For droplets:

$$\frac{\partial}{\partial x}(A\alpha_d \rho_d u_d H_d) = (\mathrm{HTC})_{dv}\Pi_{id}(T_{sat} - T_d) - \\ - \left(h_d + \frac{u_{di}^2}{2}\right)m_{dv}\Pi_{id} - \Pi_{if}(S_d H_d - S_e H_f), \quad (8)$$

where H_d, H_f, J/kg, are, respectively, the total specific enthalpies of the droplets and the liquid film, $H_{d,f} = h_{d,f} + 0.5u^2$ ($h_{d,f}$ is the specific enthalpy); $(\mathrm{HTC})_{dv}$, W/(m^2·K), is the coefficient of heat transfer from the droplets to the interface with the vapor; T_d, K, is the temperature of the droplets.

For liquid film:

$$\frac{\partial}{\partial x}(A\alpha_f \rho_f u_f H_f) = (\mathrm{HTC})_{fv}\Pi_{if}(T_{sat} - T_f) - \\ - \left(h_f + \frac{u_{fi}^2}{2}\right)m_{fv}\Pi_{if} + \Pi_{if}(S_d H_d - S_e H_f) + \\ + q_{wf}\Pi_{fw} - q_{wfi}\Pi_{fw}, \quad (9)$$

where $(\mathrm{HTC})_{fv}$, W/(m^2·K), is the coefficient of heat transfer from the liquid film to the interface with the vapor; T_f, K, is the temperature of the liquid film; q_{wf}, W/m^2, is the heat flux from the channel wall to the liquid film; q_{wfi}, W/m^2, is the component of the heat flux from the channel wall that actually contributes to generating vapor.

The system is also complemented by *a phase balance equation*

$$\sum \alpha_k = 1 \quad (10)$$

and by *thermodynamic equations of state* taking the form

$$\rho_k = \rho_k(P,T), \quad (11)$$

$$e_k = e_k(P,T), \quad (12)$$

where e_k, J/kg, is the specific internal energy of the kth phase.

Closing relations

The system of equations (1) – (10) is closed by a set of relations describing the processes of mass, momentum and energy transfer both between the individual phases and between the phases and the channel wall.

Mass sources describing the phase transition (m_{dv} and m_{fv}) are derived by considering the heat balance at the interface. Since the interface cannot accumulate heat, we obtain:

$$m_{dv} = -[(\mathrm{HTC})_{dv}(T_{sat} - T_d) + (\mathrm{HTC})_{vd}(T_{sat} - T_v)] / (h_v - h_d). \quad (13)$$

Direct heat transfer with the channel wall only occurs for the liquid film in annular dispersed flow. Accordingly, a term $q_{wf}\Pi_{fw}$, describing the heat flux transferred to the liquid film from the channel wall is added to the right-hand side of the film's energy balance equation. A term describing the component of the heat flux from the channel wall that actually contributes to generating vapor: $q_{wf}\Pi_{fw}$, where $q_{wfi} = \psi q_{wf}; \psi = 0-1$, is also added to the energy balance equation.

Given this term, a similar mass source for the phase transition between the liquid film and the vapor has the form

$$m_{fv} = -[(\mathrm{HTC})_{fv}(T_{sat} - T_f) + (\mathrm{HTC})_{vf}(T_{sat} - T_v) + q_{wfi}] / (h_v - h_f). \quad (14)$$

The terms of the form $(\mathrm{HTC})_{kv}(T_{sat} - T_k)$ are the heat flux from the liquid phase to the interface with the vapor, and $(\mathrm{HTC})_{vk}(T_{sat} - T_v)$ the heat flux from the vapor to the interface with the liquid phase for both mass sources (13), (14).

The phase transition model used allows for heat transfer coefficients of different magnitudes on both sides of the interface, but the following assumption is used for the sake of simplicity. Heat transfer coefficients on both sides of the interface are assumed to be the same and equal to a limiting factor that is the coefficient of heat transfer from the vapor. It follows then that the coefficients of heat transfer from the vapor to the interface with the droplets and from the droplets to the interface with the vapor are equal and take the form of a correlation describing gas flow around spherical particles [10]:

$$(\mathrm{HTC})_{vd} = (\mathrm{HTC})_{dv} = \frac{\lambda_v}{D_d}(2 + 0,6\,\mathrm{Re}_d^{0,5}\,\mathrm{Pr}_v^{0,33}), \quad (15)$$

where the Reynolds number for droplets follows the expression

$$\mathrm{Re}_d = \frac{\rho_v |u_v - u_d| D_d}{\mu_v}$$

(D_d, m, is the mean droplet diameter; μ_v, Pa·s, is the coefficient of dynamic viscosity of vapor); λ_v, W/(m·K), is the thermal conductivity of vapor.

The coefficients of heat transfer from the vapor to the interface with the liquid film and from the liquid film to the interface with the vapor are calculated from the correlation for single-phase convection [11] applied to the droplet-laden vapor core:

$$(\mathrm{HTC})_{vf} = (\mathrm{HTC})_{fv} = \frac{\lambda_v}{(D - 2\delta)}(0,023\,\mathrm{Re}_v^{0,8}\,\mathrm{Pr}_v^{0,4}), \quad (16)$$

where D, m, is the internal diameter of the channel; δ, m, is the mean film thickness; Pr_v is the Prandtl number for vapor; Re_v is the Reynolds number for vapor,

$$\mathrm{Re}_v = \frac{\rho_v |u_v - u_f| (D - 2\delta)}{\mu_v}.$$

The expressions for the mass sources describing hydrodynamic entrainment and deposition on the surface of the liquid film (S_e and S_d) are borrowed from [2], where turbulent diffusion is assumed to be the dominant mechanism for deposition of droplets on the film surface:

$$S_d = 9 \cdot 10^{-3} u_v \left(\frac{C}{\rho_v}\right)^{-0,5} \mathrm{Re}_v^{-0,2}\,\mathrm{Sc}_v^{-2/3} C, \quad (17)$$

where C, kg/m^3, is the droplet concentration,

$$C = \frac{G_d}{\frac{G_v u_d}{\rho_v u_v} + \frac{G_d}{\rho_d}} = \frac{\alpha_d \rho_d}{\alpha_v + \alpha_d};$$

Sc_v is the Schmidt number for vapor (taken to equal unity for simplicity);

$$\text{here } \mathrm{Re}_v = \frac{\rho_v u_v \alpha_v D}{\mu_v}.$$

It is assumed that shearing off of roll wave crests is the main mechanism of hydrodynamic entrainment of droplets into the droplet-laden vapor core for a liquid with a low viscosity (like water):

$$S_e = 1{,}07 \frac{u_v \mu_f \tau_{fv}}{\sigma^2} \left(\frac{\rho_f}{\rho_v} \right)^{0,4} \times \times \begin{cases} k_s, \mathrm{Re}_v > 10^5; \\ k_s[2{,}136 \lg(\mathrm{Re}_v) - 9{,}68], \mathrm{Re}_v < 10^5, \end{cases} \quad (18)$$

where $k_s = 0{,}57\delta + 21{,}73 \cdot 10^3 \delta^2 - 38{,}3 \cdot 10^6 \delta^3 + + 55{,}68 \cdot 10^9 \delta^4$.

The shear stresses in the momentum balance equations (τ_{wf} between the channel wall and the liquid film; τ_{vf} between the vapor and the liquid film; τ_{vd} between the vapor and the droplets) are written in the following form:

$$\tau_{wf} = Cf_{wf} \frac{\rho_f}{2} u_f |u_f|; \quad (19)$$

$$\tau_{vf} = Cf_{vf} \frac{\rho_v}{2} (u_v - u_f) |u_v - u_f|; \quad (20)$$

$$\tau_{vd} = Cf_{vd} \frac{\rho_v}{2} (u_v - u_d) |u_v - u_d|, \quad (21)$$

where Cf_{kk} are the respective coefficients of friction (drag force).

The coefficient of friction between the vapor and the liquid film is calculated from the modified Wallis correlation [12]:

$$Cf_{vf} = \frac{0{,}079}{\mathrm{Re}_v^{0,25}} \left(1 + 300 \frac{\delta}{D} \right), \quad (22)$$

where

$$\mathrm{Re}_v = \frac{\rho_v |u_v| (D - 2\delta)}{\mu_v}.$$

The aerodynamic drag of the droplets is calculated from the dependence [13]:

$$Cf_{vd} = \left(0{,}4 + \frac{24}{\mathrm{Re}_d} + \frac{4}{\sqrt{\mathrm{Re}_d}} \right) \frac{1}{4}, \quad (23)$$

where

$$\mathrm{Re}_d = \max \left(0{,}1; \frac{\rho_v |u_v - u_d| D_d}{\mu_v} \right).$$

The coefficient of friction between the liquid film and the channel wall is expressed as follows [14]:

$$Cf_{fw} = \max \left(\frac{16}{\mathrm{Re}_{Dhf}}; \frac{0{,}079}{\mathrm{Re}_{Dhf}^{0,25}} \right), \quad (24)$$

where

$$\mathrm{Re}_{Dhf} = \frac{u_f \rho_f \alpha_f D}{\mu_f}.$$

The geometric characteristics used should also be formulated for writing the closing relationships; these include:

perimeters of interface interactions;

perimeter of the interaction between the liquid film and the channel wall;

mean thickness of the liquid film;

mean droplet diameter.

The perimeter of the interface between the vapor and the droplets follows the expression

$$\Pi_{id} = \frac{6\alpha_d}{D_d}, \quad (25)$$

The perimeter of the interface between the vapor and the liquid film:

$$\Pi_{if} = \pi(D - 2\delta), \quad (26)$$

where $\delta = 0{,}5D(1 - \sqrt{1 - \alpha_f})$ (mean thickness of the liquid film).

The perimeter of the interface between the liquid film and the channel wall:

$$\Pi_{wf} = \pi D. \quad (27)$$

The droplet diameter is calculated by the technique described in [14]:

$$D_d = \max(8{,}4 \cdot 10^{-5}; \min[D_{d1}; D_{d2}]), \quad (28)$$

$$D_{d1} = 7{,}96 \cdot 10^{-3} \frac{\sigma}{\rho_v j_v} \mathrm{Re}_v \left(\frac{\rho_d}{\rho_v} \right)^{1/3} \left(\frac{\mu_v}{\mu_d} \right)^{2/3}, \quad (29)$$

where j_v is the normalized velocity,

$$j_v = \frac{u_v \rho_v A \alpha_v}{\rho_v A} = u_v \alpha_v;$$

the Reynolds number

$$\mathrm{Re}_v = \frac{\rho_v |j_v| D}{\mu_v}.$$

$$D_{d2} = 0,254 L\left[-0,13We_v + \sqrt{16 + (0,13We_v)^2}\right], \quad (30)$$

where L is the characteristic size (taken as the internal diameter D of the channel);

$$We_v = \frac{\rho_v j_v^2 L}{\sigma}.$$

Numerical method

Moving on to the finite-difference formulation of the system, let us introduce the pseudovector notation of the following form:

$$\vec{W} = \begin{pmatrix} W_{k,m} \\ W_{k,i} \\ W_{k,e} \end{pmatrix} = \begin{pmatrix} A\alpha_k \rho_k u_k \\ A\alpha_k \rho_k u_k^2 \\ A\alpha_k \rho_k u_k H_k \end{pmatrix}; \quad (31)$$

$$Q = \begin{pmatrix} S_{k,m} \\ S_{k,i} \\ S_{k,e} \end{pmatrix}, \quad (32)$$

where **W** is the flux vector; **Q** is the right-hand side vector; the subscript k indicates the corresponding fluid (vapor, droplet, liquid film).

Using two-point approximation, let us write the finite-difference formulation of the system of equations:

$$\frac{W_{k,p}^{n+1} - W_{k,p}^{n}}{\Delta x} = S_{k,p}^{n+1}; \; p = m, e; \quad (33)$$

$$\frac{W_{k,i}^{n+1} - W_{k,i}^{n}}{\Delta x} + \overline{\alpha}\overline{A}\left(\frac{P^{n+1} - P^{n}}{\Delta x}\right) = S_{k,i}^{n+1}; \quad (34)$$

$$\sum_k (\alpha_k^{n+1}) = 1, \quad (35)$$

where the subscript p indicates the corresponding balance equation (m for mass, i for momentum, e for energy).

At the same time, the balance equations for momentum and energy are written identically in pseudovector notation, so we combined them. Since the flux vector is not linear, in order to achieve convergence of the process, at the next step we divide the sought-for quantities by the sum of the known quantity $W_p^{n+1,s}$ from the previous iteration and their increments $\Delta W_p^{n+1,s+1}$ at the sought-for iteration, which transforms the given system to the form:

$$\frac{W_{k,p}^{n+1,s} + \Delta W_{k,p}^{n+1,s+1} - W_{k,p}^{n}}{\Delta x} = S_{k,p}^{n+1,s}; \; p = m, e; \quad (36)$$

$$\frac{W_{k,i}^{n+1,s} + \Delta W_{k,i}^{n+1,s+1} - W_{k,i}^{n}}{\Delta x} + \overline{\alpha}\overline{A}\left(\frac{P^{n+1,s} + \Delta P^{n+1,s+1} - P^{n}}{\Delta x}\right) = S_{k,i}^{n+1,s}; \quad (37)$$

$$\sum_k (\alpha_k^{n+1,s} + \Delta\alpha_k^{n+1,s+1}) = 1. \quad (38)$$

Leaving the increments on the left-hand side of the equation, we move the remaining terms to the right-hand side, thus reducing the solution of the system to finding the increments of the sought-for functions. In the same way, we represent the sought-for flux vector in terms of the vector of primitive variables:

$$f = (\alpha_k, T_k, u_k, P)^T.$$

As a result, the system is reduced to a matrix notation of the form

$$M^{n+1.s} \Delta f^{n+1,s+1} = \mathbf{B}^{n+1,s}, \quad (39)$$

which is solved by the Gauss method. The vector $\mathbf{B}^{n+1,s}$ is the right-hand side vector; $M^{n+1,s}$ is the matrix of transformation of the flux vector to the vector of primitive variables.

The coefficients of the matrix $M^{n+1,s}$ can be obtained by vector differentiation of the flux vector with respect to the vector of primitive variables:

$$\Delta W^{n+1,s+1} = \left(\frac{\partial \mathbf{W}}{\partial \mathbf{f}}\right)^{n+1,s} \Delta f^{n+1,s+1} = M^{n+1,s} \Delta f^{n+1,s+1}.$$

The matrix $M^{n+1,s}$ has a repeating block structure with a block size of 3 × 3 (see Table).

The given algorithm is applicable for solving problems with codirectional liquid velocities.

Model testing

The three-fluid model described in our study was used to calculate two-phase flow of water in an adiabatic circular tube; we took as a basis the experimental data obtained by

Table

Coefficients of the transformation matrix $M^{n+1,s}$

Balance equation	$\Delta\alpha_k$	ΔT_k	Δu_k	...	ΔP
mass	$A\rho u$	$A\alpha u\dfrac{\partial\rho}{\partial T}$	$A\alpha\rho$	...	$A\alpha u\dfrac{\partial\rho}{\partial P}$
momentum	$A\rho u^2$	$A\alpha u^2\dfrac{\partial\rho}{\partial T}$	$A\alpha\rho 2u$	...	$A\alpha u^2\dfrac{\partial\rho}{\partial P}+\bar{A}\bar{\alpha}$
energy	$A\rho uH$	$A\alpha uE\dfrac{\partial\rho}{\partial T}+$ $+A\alpha\rho u\dfrac{\partial e}{\partial T}$	$A\alpha\rho(H+u^2)$	...	$A\alpha u\left(E\dfrac{\partial\rho}{\partial P}+\rho\dfrac{\partial e}{\partial P}+1\right)$
	⋮	⋮	⋮	⋮	⋮
phases	1	0	0	...	0

Würz in [15], considering two-phase water-vapor flow in adiabatic round tubes with internal diameters of 10 and 20 mm, with pressures ranging from 3 to 9 MPa and total flow rates of 500 – 3000 kg/(m²·s).

We used the given model to calculate a total of 90 experimental points. Fig. 1 shows a comparison of the calculated dependences of the total pressure drop dp/dx on the relative flow rate G_{vrel} of vapor with the measured values. G_{vrel} is the ratio of the absolute flow rate of vapor the sum of flow rates of all three "liquids".

It can be seen from the given results that the calculated dependences practically coincide with the experimental ones from a qualitative standpoint: they have the same slopes, even for dependences with an alternating-sign derivative. Quantitative analysis of the data reveals slight discrepancies, which can be clearly observed by plotting the calculated total pressure drops versus the experimentally measured ones (Fig. 2).

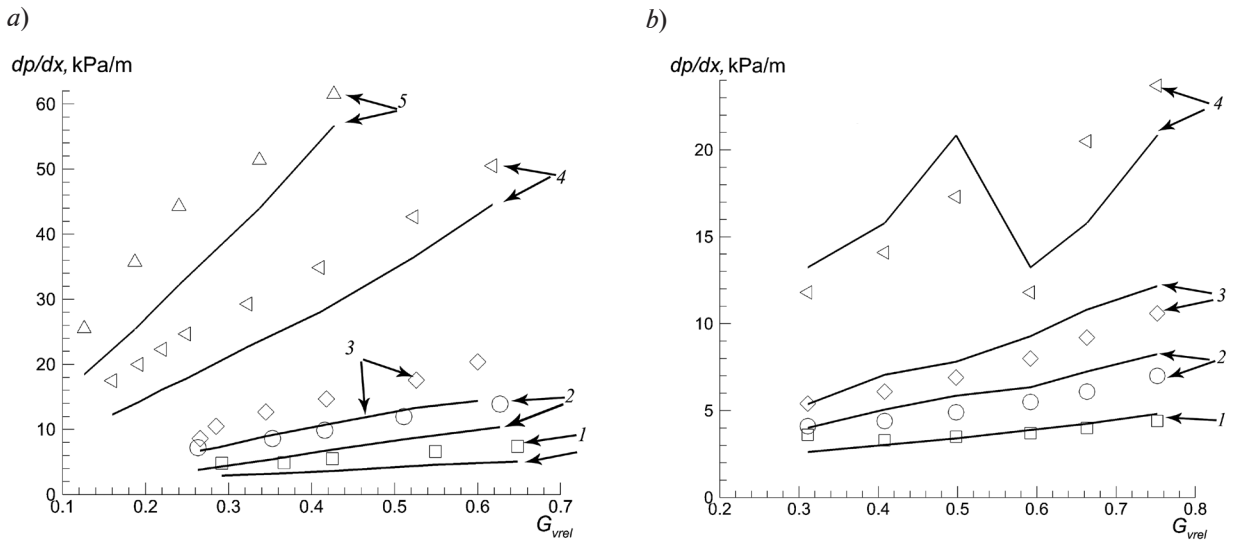

Fig. 1. Experimental (symbols) and calculated (lines) dependences of the total pressure drop in the channel versus the relative vapor flow rate for internal channel diameters D = 10 mm (*a*) and 20 mm (*b*), with different total flow rates, kg/(m²·s): 500 (*1*), 750 (*2*), 1000 (*3*), 2000 (*4*), 3000 (*5*). Pressure P = 7 Mpa

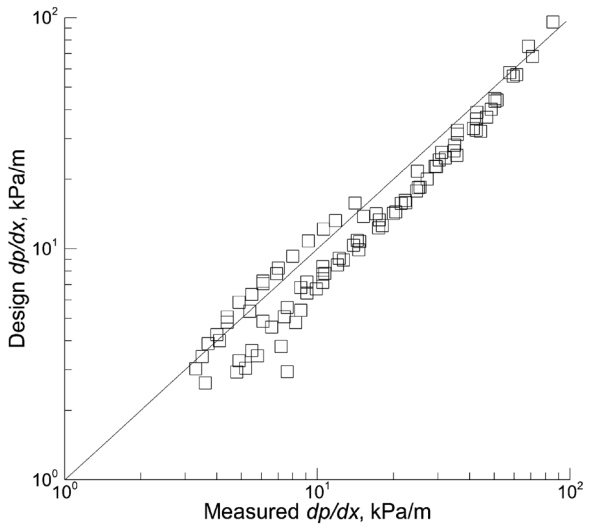

Fig. 2. Relationship between the calculated total pressure drop in the channel and the measured values for all experimental data used from [15]

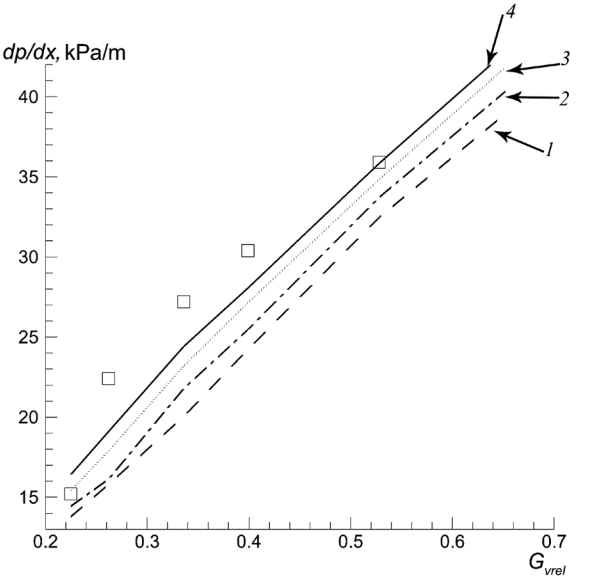

Fig. 3. Dependences similar to those shown in Fig. 1, which were obtained for different roughness coefficients: 1.00 (*1*), 1.20 (*2*), 1.50 (*3*), 1.85 (*4*); $P = 3$ MPa, $D = 10$ mm

A possible explanation for these discrepancies is that the chosen formulation of the closing relations was not optimal. It is these relations that describe the physics of the processes and, therefore, determine the calculated result in the three-fluid model. Optimizing the closing relations was not among the goals of this study. However, we can confirm that taking into account additional factors, such as tube roughness (which complicates the expression for the coefficient of friction between the film and the channel wall), makes it possible to bring the calculated results to closer agreement with the experimentally measured values. Fig. 3. shows the results obtained by adjusting the calculated results in this manner. For simplicity, the tube roughness is taken into account in the coefficient of friction as a variation of the multiplier, and the calculations were carried out only for one series of experiments (pressure of 3 MPa, total flow rate of 1000 kg/(m^2·s), internal channel diameter of 10 mm). The data in Fig. 3 clearly demonstrate that the calculated total pressure drops can be better fitted to the experimentally measured values by increasing the roughness coefficient.

Conclusion

We have developed a numerical procedure for solving the stationary problem of two-phase water-vapor flow in a one-dimensional approximation with an annular dispersed-flow regime using a three-fluid formulation. The integrated system of equations includes the heat transfer between the phases and the channel wall, the friction between the phases and the channel wall, entrainment and deposition on the surface of the liquid film, as well as interaction of the gravity field for channels with constant and variable cross-sections.

We have carried out initial testing of the developed computational model by comparing the simulation results with the experimental data reported by Würz [15]. The comparison confirmed that the three-fluid model considered in our study provides an adequate description of the given series of experiments with vapor-fluid flow and qualitatively speaking, the obtained dependences fully agree with the experimental ones. There are quantitative discrepancies but they do not exceed 20%. However, we have established that these discrepancies can be reduced by adjusting the closing relations.

REFERENCES

[1] **S.M. Sami,** An improved numerical model for annular two-phase flow with liquid entrainment, Int. Comm. Heat mass transfer. 15 (1) (1988) 281–292.

[2] **S. Sugowara,** Droplet deposition and entrainment modeling based on the three-fluid model, Nuclear Engineering and Design. 122 (1–3) (1990) 62–84.

[3] **N. Hoyer,** Calculations of dry-out and postdry-out heat transfer for tube geometry, Int. J. Multiphase Flow. 24 (2) (1997) 319–334.

[4] **V.M. Alipchenkov, L.I. Zaichik, Yu.A. Zeigarnik, et al.,** The development of a three-fluid model of two-phase flow for a dispersed-annular mode of flow in channels: size of droplets, Heat and Mass Transfer and Physical Gasdynamics. 40 (4) (2002) 641–651.

[5] **S. Jayanti, M. Valette,** Prediction of dry-out and postdry-out heat transfer at high pressures using one-dimensional three-fluid model, Int. J. Heat and Mass Transfer. 47 (22) (2004) 4895–4910.

[6] **V. Stevanovic, M. Stanojevic, D. Radic,** Three-fluid model predictions of pressure changes in condensing vertical tubes, Int. J. Heat and Mass Transfer. 51 (15–16) (2008) 3736–3744.

[7] **H. Saffari, N. Dalir,** Effect of virtual mass force on prediction of pressure changes in condensing tubes, Thermal Science. 16 (2) (2012) 613–622.

[8] **S.V. Bulovich, E.M. Smirnov**, Experience in using a numerical scheme with artificial viscosity at solving the Riemann problem for a multi-fluid model of multiphase flow, AIP Conference Proceedings. 1959 (2018) 050007-1–050007-8.

[9] **D. Bestion,** The physical closure laws in the CATHARE code, Nuclear Engineering and Design. 124 (3) (1990) 229–245.

[10] **B.I. Brounshteyn, G.A. Bishbeyn,** Gidrodinamika, masso- i teploobmen v dispersnykh sistemakh [Hydrodynamics, mass and heat exchange in the dispersion systems], Leningrad, Chemistry PH, 1977.

[11] **S.S. Kutateladze, V.M. Borishanskiy,** Spravochnik po teploperedache [Manual of heat transfer], The State Energetic PH, Moscow, 1958.

[12] **G.B. Wallis,** One-dimensional two-phase flow, McGraw-Hill, NewYork, 1969.

[13] **R.I. Nigmatulin,** Dinamika mnogofaznykh sred v 2 ch. Ch.1. [Multiphase mixture dynamics, in 2 parts, P. 1], Moscow, Nauka, 1987.

[14] **Yu.V. Yudov, S.N. Volkova, A.A. Slutskiy,** KORSAR/V1.1 teplogidravlicheskiy raschetnyy kod, metodika rascheta zamykayushchikh sootnosheniy i otdelnykh fizicheskikh yavleniy konturnoy teplogidravliki [KORSAR/V1.1 thermal and hydraulic calculation code, design procedure of constitutive relationships and separate physical phenomena of contour heathydraulics], Sosnoviy Bor, 2001.

[15] **J. Würtz,** An experimental and theoretical investigation of annular steam-water flow in tubes and annuli at 30 to 90 bar, No. 372, Technical university of Denmark, Roskilde, 1978, 141 p.

Received 13.07.2018, accepted 07.08.2018.

THE AUTHORS

AVDEEV Evgeniy E.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
avdeev-evgeni@yandex.ru

PLETNEV Alexander A.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
aapletnev@yandex.ru

BULOVICH Sergey V.
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St., St. Petersburg, 195251, Russian Federation
bulovic@yandex.ru

RADIO EMISSION OF STARS IN THE MONOCEROS CONSTELLATION

A.A. Lipovka, N.M. Lipovka

Center of Physical Studies, University of Sonora, Hermosillo, Mexico

In the present paper, the optical identifications of the bright stars from the Monoceros constellation with strong radio sources have been suggested. The Monoceros constellation is projected on the bright region of the Milky Way, where the densities of the stars and gas are rather high. 17 stars brighter than 11^m are located within the one square degree plate under investigation. All these stars were identified with radio sources from NVSS survey of NRAO observatory. Considerable radio refraction was revealed in the interstellar medium. It was found that twelve stars among seventeen ones exhibited radio emission characterized by non-thermal spectrum.

Key words: coordinates system, optical identification of radio objects, interstellar space.

Citation: A.A. Lipovka, N.M. Lipovka, Radio emission of stars in the Monoceros constellation, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 11 (3) (2018) 104–111. DOI: 10.18721/JPM.11312

Introduction

The existence of the interstellar medium was first hypothesized by Vasily Struve, who concluded in 1847, based on theoretical considerations, that the space between the stars is filled with gas. Struve's hypothesis was later confirmed by the independent observations made by Vorontsov-Velyaminov and Trumpler, who both discovered the absorption of light by the interstellar medium. Observations of outer space in the meter and centimeter wavelengths began with the advent of radio astronomy in the 1930s; independent experimental data proving that the interstellar space is filled with gas ionized by nearby stars were obtained after World War II [1]. Propagation of radio waves in outer space and in the Earth's ionosphere and atmosphere depends on the properties of the interstellar medium, which is why these properties should be taken into account in studies on celestial radio sources and identifying these sources with optical objects.

The distribution of radio emission in the Milky Way (i.e., its "radio brightness"), obtained from observations at 6.4 cm with the 12 m parabolic radio telescope [2] and the Large Pulkovo Radio Telescope (LPRT) 100 m in diameter [3], confirmed that ionized hydrogen (HII) is concentrated towards the Galactic plane, where the stellar density is high.

Analysis of all radio data was carried out in [3] for the distribution of radio brightness of the northern sky in the frequency range from 0.4 to 7,700 MHz. It was found that the Galactic corona is radio emission produced by relativistic electrons moving in the magnetic field of this corona (synchrotron radiation) and therefore has a non-thermal character, covering an area of 20 Ч 25 kiloparsecs around the center of the Galaxy. The size of the Galactic corona turned out to be two times larger than previously assumed [1]. The characteristics of the radio medium (relativistic electrons and magnetic field) in the Galactic corona [3] were calculated in accordance with the mechanism of synchrotron radio emission of relativistic electrons in a magnetic field [4].

The mismatch between the radio and optical coordinates of celestial objects was first discovered when radio astronomy emerged as an independent branch of astrophysics. The first error in matching the radio emission of celestial objects to the optical sky was made when the updated 3C Catalog [5] was introduced in 1962 as reference without listing which observed radio objects corresponded to the optical ones.

However, this was not particularly critical at the time, since the survey was carried out with poor angular resolution ($\Theta = 13.6' \times 4.6°$) at a frequency of 178 MHz, in fact, averaging all the positions of the radio sources located in a plate of one square degree. The reason for this error was that the radio telescope had a high-directivity pencil beam pattern [5].

Studies on optical identification of radio and optical sky, carried out in 1990 at the National Institute of Astrophysics, Optics and Electronics (Tonantzintla, Mexico), confirmed the mismatch between optical objects with diffuse image and radio sources [6].

A survey of the northern sky was carried out in 1993–1997 at the National Radio Astronomy Observatory (NRAO), VA, USA, via a radio telescope with high sensitivity and good resolution ($\Theta = 45''$) at 1400 MHz. The mismatch between radio and optical objects was reconfirmed by this survey [7]. It was hypothesized based on the data obtained that the sources of radio emission are mainly distant celestial objects (quasars and distant galaxies with redshifts greater than unity) [7], while the interstellar and intergalactic media are empty spaces where radio beams propagate along null geodesics and reach the most distant objects with millisecond precision.

Based on these considerations, the 2009 General Assembly of the International Astronomical Union Congress recommended the International Celestial Reference Frame (ICRF2) system [8] to match the radio and optical objects. The catalog included 3,414 reference radio sources. The identifications in the catalog were obtained by cross-correlation of the coordinates of celestial radio and optical objects, with these objects taken as reference. The majority of the reference objects were radio sources that randomly coincided with distant optical objects of a quasi-stellar structure whose density was very high. This was where the error made in ill-considered attempts to identify radio objects with optical ones originated from.

The goal of this study has been to carry out alternative optical identifications of bright stars from the Monoceros constellation, which are strong radio sources.

Substantiating why the radio object J(062153.45-041807.69) cannot be accepted as reference

In this study, we have considered a sky plate projected onto the local Galactic arm with high densities of both stars and the gas component of the interstellar medium consisting mainly of atomic and ionized hydrogen. A radio source recommended as a reference object for matching radio and optical objects by the ICRF2 catalog [8] is located at the edge of this plate.

Fig. 1 shows the image of the radio object J(062153.45-041807.69) as isophotes (lines of equal radio emission intensity) [7] superimposed on the image of the optical sky based on the data in [9]. Evidently, the given radio object has a two-component structure, and more than 15 other optical objects of the quasi-stellar structure fall into the region of this object. For this reason, it is impossible to determine which of the optical objects emits radio waves, and therefore the radio object J(062153.45-041807.69) should be excluded from the list of reference objects of the catalog [8], where it is recommended for high-precision matching of the radio and optical skies.

In addition, if we use the reference proposed in catalog [8], not a single radio source coincides with any optical object in the immediate vicinity of the reference object, which has a size of about two square degrees. Notably, the coordinates of the radio object J(062153.45-041807.69) included in the reference catalog [8] were determined in the meter wavelength range [10], where significant radio refraction is observed in the Earth's ionosphere; this phenomenon was not taken into account when radio coordinates of this object were determined.

In 1962, Komesaroff carried out a survey of the sky at a frequency of 19.7 MHz in Australia and New Zealand and discovered that radio waves experience significant radio refraction in the meter wavelength range in the Earth's ionosphere at altitudes over 350 km [11].

At present, it has been established that the coordinates of radio objects obtained in the centimeter wavelength range have been altered by radio refraction in the Earth's troposphere and, as a result, differ from the coordinates of radio objects obtained in the meter range.

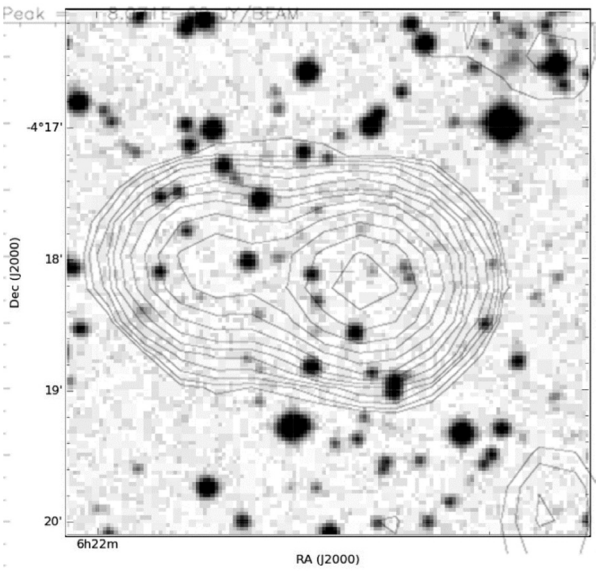

Fig. 1. The coordinates of the reference radio object J(062153.45041807.69) (according to [8]) superimposed on the optical image:
the radio object is shown as isophotes superimposed on an image of a region of the optical sky (plate). Coordinates for the epoch J2000.0 are plotted on the axes. RA is the right ascension (h, min) and DEC is the declination (deg, min)

The errors made in the matching radio and optical objects are discussed in detail in [12]. Identifications of celestial radio sources with optical object should take into account the parameters of the medium which affects the nature of radio wave propagation in interstellar and intergalactic media.

According to our procedure for matching radio and optical objects [13], identification was assumed to be correct if three or more radio sources matched the objects visible in the optical wavelength range for a given plate of one square degree. This is also necessary for taking into account the azimuth slew of the given plate, often occurring in scans of sky regions of one square degree with a radio interferometer [14].

Identification of radio objects with stars in the Monoceros constellation

We have carried out the first optical identifications at the National Institute of Astrophysics, Optics and Electronics (INAOE, Tonantzintla, Mexico) in 1985, 1990, 1993 and 1994 using a Zeiss blink comparator with the precision of

$$\sigma RA \times \sigma DEC = 1.5'' \times 1.5''$$

according to standard astrometric practices.

We have found that radio objects fall into an empty field in the optical image of the sky [6]. The general consensus on the reason for the mismatch between radio sources and optical celestial objects, which still exists today, has long been that radio emission comes from very distant compact objects (radio galaxies and quasars) located at the edge of the observable Universe.

The advent of computer technologies and the Internet has offered countless new opportunities for astrophysics and other sciences. In fact, virtually all observations of celestial objects in the radio and optical ranges can be found on the Internet, giving the perfect opportunity to reconsider the existing identifications of radio and optical objects.

In 2007, we carried out further identifications of these objects and found that radio objects were incorrectly matched with optical ones and that most bright stars emit in the RF range.

We have developed and successfully used the method for identifying radio and optical objects based on matching radio sources to bright stars (the Lipovka – Kostko – Lipovka method, or LKL) [13].

In this study, we have identified radio sources with stars in the Monoceros constellation based on the NVSS radio survey of the NRAO observatory [14]. The given sky plate (Fig. 2) is projected onto the local arm of the Galaxy with high stellar density.

Matching radio objects to stars (Fig. 2) confirmed significant radio refraction in the interstellar medium, which is rather predictable, since this region of the sky is located in the local spiral arm characterized by a high content of gas. This region also has a high stellar density, which is why 17 stars brighter than 11^m have been identified with radio objects. In addition, seven faint stars were identified with "weak" radio sources whose flux density lies below the detection threshold ($P < 2.5$ mJy); such a threshold value was taken in [14]. These stars are denoted by the letter "a" in Fig. 2 and are not considered in this paper, since they are absent in catalog [15]. However, 7 faint stars

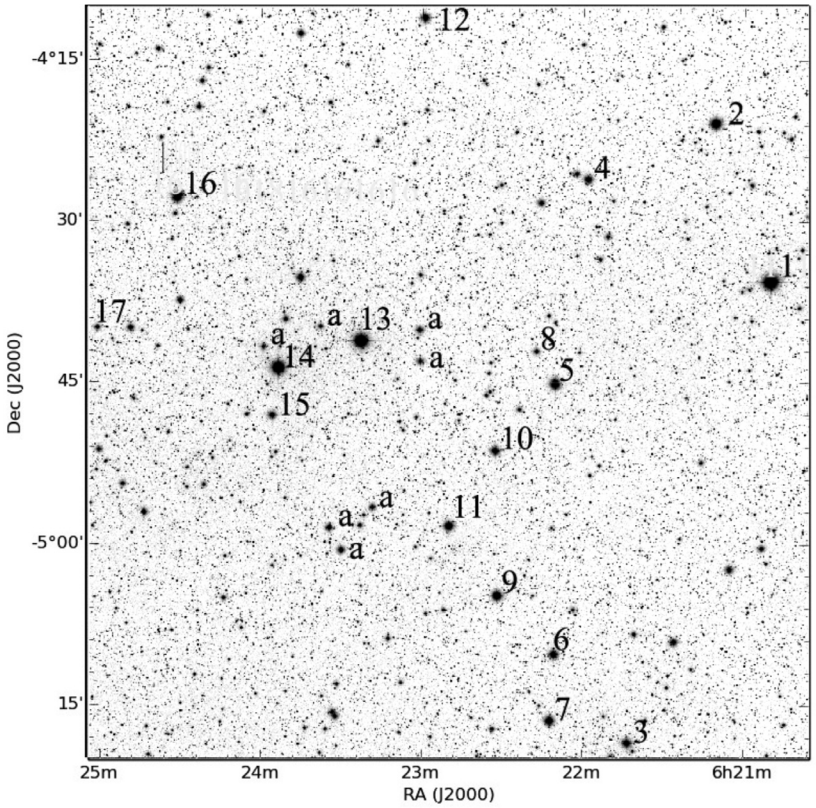

Fig. 2. Image of the sky plate with 17 stars (numbered) identified with strong radio sources according to the data of radio surveys [9, 14];
celestial objects marked with "a" have been identified with very weak radio sources and are not considered in this paper.

matching 7 weak radio sources in a plate less than 0.2 square meters in size confirms that the identifications given in Fig. 2 and in Table 2 are correct.

The method we have developed for matching the radio to the optical sky (the LKL method [13]) is based on the data from fundamental catalogs of stars [16], and radio sources are identified with stars whose density and accuracy of measured coordinates are sufficient to confidently match the coordinates of celestial radio objects to objects detected through optical observations.

The conventional names for stars [16] and their equatorial coordinates according to the UCAC3 catalog [17] for the epoch J2000.0 are given in Table 1 (columns 3 and 4). The parallaxes for the stars are given in column 5, and the stellar magnitudes in column 6.

No single radio source could be identified with an optical object in the given region of the sky based on the NVSS survey [14], while using our method for matching the radio objects [13] made it possible to identify 17 strong radio sources with bright stars (see. Fig. 2 and Table 2).

Table 2 shows the equatorial coordinates of radio sources at a frequency of 1400 MHz (columns 2 and 3) according to the data of [15], which we obtained by identifying the optically observed stars (see Table 1 and Fig. 2).

The coordinates of radio objects whose matches to stars were corrected are given in columns 6 and 7 (Table 2).

The numbering of radio sources in Table 2 corresponds to the numbering of stars in Table 1 and Fig. 2. Flux density based on the data in [15] is given in column 4 of Table 2.

Flux density measurements are available for several radio sources located in this plate (see Fig. 2), at frequencies $\nu = 150 - 1400$ MHz according to catalog [15]. The spectral index of

Table 1

Names [16] and coordinates of stars according to the UCAC3 catalog [17] for the epoch J2000.0

Star		RA(J),	DEC(J)	ε Pos	Mag
No.	Name	h m s	deg, min, s	mas	*m*
1	HD 44286	06 20 50,466	−04 35 43,70	1270	6,68
2	HD 44335	06 21 10,845	−04 21 00,18	10	7,84
3	HD 44457	06 21 43,488	−05 18 34,15	26	8,92
4	HD 294985	06 21 58,328	−04 26 14,34	34	9,02
5	HD 44546	06 22 10,445	−04 45 13,06	11	7,92
6	HD 44565	06 22 10,867	−05 10 20,73	29	8,80
7	HD 44566	06 22 12,414	−05 16 31,21	26	8,38
8	HD 294989	06 22 17,423	−04 42 10,80	21	10,74
9	HD 44620	06 22 31,967	−05 04 53,95	91	8,18
10	HD 44619	06 22 32,671	−04 51 26,77	18	9,02
11	HD 44678	06 22 49,984	−04 58 25,83	32	8,30
12	HD44702	06 22 59,160	−04 11 13,50	22	8,50
13	HR 2295	06 23 22,793	−04 41 15,20	376	6,89
14	HD 44841	06 23 53,623	−04 43 43,88	102	6,99
15	HD 44856	06 23 53,863	−04 48 09,35	32	9,29
16	HD 295031	06 24 31,450	−04 27 58,40	42	8,44
17	HD295031	06 25 01,010	−04 40 00,00	20	10,08

Note. The numbers of the stars correspond to those shown in Fig. 2.

Notations: RA(J) is the right ascension, DEC(J) is the declination, ε Pos mas is the optical parallax, Mag *m* is the stellar magnitude.

radio emission α was calculated for these objects (Table 2, column 5). The radio spectrum of these stars turned out to be non-thermal; the radio flux density is $P \sim \nu^{-\alpha}$, where ν is the frequency of observation in the radio range.

Table 3 shows corrections to the coordinates of radio sources for three groups of objects. These corrections have different values because the given objects are located at different distances from the observer and because of the apparently significant radio refraction in this direction of the interstellar medium. The numbers of radio sources in each of the three groups are given in column 1 and correspond to the numbers in Tables 1, 2 and in Fig. 2.

These corrections have different values because the given objects are located at different distances from the observer and because of the apparently significant radio refraction in this direction of the interstellar medium.

These corrections (columns 2, 4, Table 3) should be added (taking into account the sign) to the coordinates measured in the radio range (Table 2, columns 2, 3) in order to obtain the corrected coordinates of radio objects matched to optical objects (Table 2, columns 6, 7).

Conclusion

Prior to our study, the method for matching the radio to the optical sky proposed in the NVSS survey [14] proved unsuccessful for matching any radio sources to optical objects within the given region of the sky [14, 15]; the method uses the ICRF2 catalog [8] recommended for identifying radio objects with optical ones.

The method for matching the radio and optical sky that we have proposed and used (the LKL method) increased the number of radio sources identified with optical objects by tens of times. We have established that radio sources are primarily identified with stars. The corrections obtained for the radio coordinates are due to a number of factors:

Table 2

Data comparison for radio sources in the NVSS survey and the corrected matches to the stars

No.	NVSS data [14, 15]]				Corrected coordinates	
	RA(J) h m s	DEC(J) deg, min, s	*P* MJy	α	RA(J) h m s	DEC(J) deg, min, s
1	06 19 13,97	−04 35 53,2	13,5	0,60	06 20 48,8	−04 34 50,0
2	06 19 37,76	−04 22 21,0	46,7	0,70	06 21 12,7	−04 21 17,8
3	06 21 54,53	−05 27 22,6	26,7	0,86	06 21 47,6	−05 18 15,0
4	06 20 26,93	−04 27 13,1	51,7	–	06 22 01,4	−04 26 09,0
5	06 22 20,51	−04 53 57,6	319*	–	06 22 08,8	−04 44 49,3
6	06 22 17,87	−04 55 48,8	37,7	–	06 22 12,0	−05 10 03,5
7	06 22 23,71	−05 19 10,8	23,9	0,57	06 22 15,0	−05 16 25,0
8	06 22 26,70	−05 25 32,0	15,3	–	06 22 10,9	−04 41 38,8
9	06 22 47,48	−05 01 03,7	7,3	0,80	06 22 22,0	−05 05 14,4
10	06 22 28,50	−05 19 25,5	46,1	–	06 22 35,8	−04 52 00,0
11	06 22 42,36	−05 11 27,4	10,0	–	06 22 49,3	−04 57 17,0
12	06 21 19,50	−04 12 33,1	8,8	–	06 22 54,4	−04 11 29,0
13	06 23 17,87	−04 55 48,8	37,7	0,40	06 23 24,7	−04 41 38,1
14	06 23 43,57	−04 58 28,2	246,8	0,80	06 23 50,4	−04 44 18,2
15	06 22 21,56	−04 49 43,4	60,7	0,60	06 23.55,8	−04 48 49,4
16	06 24 38,40	−04 37 41,5	114,9	0,75	06 24 26,7	−04 28 34,5
17	06 24 49,73	−04 53 59,5	174,4	0,70	06 24 55,8	−04 39 49,0

Note. The numbers of the stars correspond to those shown in Fig. 2 and in Tables 1 and 2.

Notations: *P* is the flux density of radio objects, α is the spectral index of radio emission of these objects ($P \sim \nu^{-\alpha}$, ν is the frequency of observation in the radio range). The rest of the notations are given in Table 1.

Table 3

Corrections for coordinates of radio objects matched to optical data for stars

Star number	ΔRA	$\pm\sigma_1$	ΔDEC	$\pm\sigma_2$
	m s	s	min, s	s
1, 2, 4, 12, 15	1 30	2.9	−10 00	35.5
8, 9, 10, 11, 13, 14, 17	−7	1.3	−14 00	15.2
3, 5, 6, 7, 16	10	1.2	−7 00	10.6

Notes. 1. The numbers of stars correspond to the ones in Fig. 2 and in Tables 1 and 2.

2. To obtain the corrected coordinates for each star, the corrections have to be added (taking into account the sign) to the coordinates of the radio source from the NVSS survey [14, 15] (see Table 2).

Notations: σ is the absolute correction error. The rest of the notations are given in Table 1.

accuracy in matching radio and optical objects;

accuracy in measuring radio coordinates;

presence of radio refraction in the given region of space.

Our findings are conclusive proof that optical identifications for the NRAO (National Radio Astronomy Observatory, VA, USA) and DSS (Palomar Observatory, CA, USA) surveys should not be performed based on coordinate matches of radio sources and objects visible in the optical wavelength range, since these matches are incorrect. Each one-degree astronomical plate scanned in the NVSS survey should be matched to the optical sky using the LKL method, regardless of the given coordinate match.

Applying the proposed method yields correct information about the astrophysical characteristics of the identified objects in a wide wavelength range (for radio and optical objects).

Based on the results obtained, we have resolved the paradox that stars do not emit radio waves. Given the correct identification of radio and optical skies, 17 stars brighter than 11m were matched in the plate under consideration.

Identifications of radio and optical objects carried out in our study also confirmed the presence of interstellar medium in the given region of space, which is further confirmed by the image of this sky region in the infrared wavelength range.

Acknowledgments

The authors express their gratitude to the staff of the Palomar Observatory and the National Radio Astronomy Observatory for compiling a survey of the sky in the optical and radio wavelength range, to the creators of the UCAC database (USNO CCD Astrograph Catalog) and to all those involved in compiling the database for general use.

The authors express their gratitude to the Astronomical Data Center of Canada (CADC) for the opportunity to obtain all available information on celestial objects for scientific research.

The authors express their gratitude to the researchers of the Special Astrophysical Observatory of the Russian Academy of Sciences O.V. Verkhodanova, S.A. Trushkina and others for creating a database of radio catalogs.

REFERENCES

[1] **S.A. Kaplan, S.B. Pikelner,** Mezhzvezdnaya sreda [Interstellar medium], Fizmatgiz, Moscow, 1963, P. 11–118.

[2] **Y.N. Parijsky, V.Y. Golnev, N.M. Lipovka,** The preliminary results of observations of the Milky Way at the wavelength of 6.4 cm obtained at the Pulkovo Observatory, Bulletin of the Main Astronomical Observatory in Pulkovo. 24(6(182)) (1967) 199–203.

[3] **N.M. Lipovka,** Radioizlucheniye korony Galaktiki [Galactic-corona radiation], Soviet Astronomy. 54 (6) (1977) 1211–1220.

[4] **V.L. Ginzburg,** Teoreticheskaya fizika i astrofizika. Dopolnitelniye glavy [Theoretical physics and astrophysics. Supplementary chapters], Nauka, Moscow, (1976) 70–99.

[5] **A.S. Bennett,** The revised 3C catalog of radio sources, MNRAS, 125 (1962) 75–86.

[6] **E. Chavira-Navarrete, N.M. Lipovka, A.A. Lipovka,** Opticheskiye polozheniya 748 slabykh diffuznykh obyektov i galaktik v okrestnosti radioistochnikov RC kataloga [Optical positions of 748 faint diffuse objects and galactics in the vicinity of radio sources from RC catalogue], Special Astrophysical Observatory of RAS, St. Petersburg, Preprint No. 88 (1993) 1–47.

[7] **J.J. Condon, W.D. Cotton, E.W. Greisen, et al.,** The NRAO VLA Sky Survey, Astron. J. 115 (5) (1998) 1693–1716.

[8] **A.L. Fey, C. Ma, E.F. Arias, et al.,** The second extension of the International Celestial Reference Frame: ICRF-EXT. 1, Astron. J. 127 (6) (2004) 3587–3608.

[9] Digital Sky Survey System (DSS) – Canadian Astronomy Data Centre, DSS, CADC, URL: www3.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/dss/.

[10] **J.N. Douglas, F.N. Bash, F.A. Bozyan, Ch. Wolfe,** The Texas survey of radio sources covering −35.5 degrees < declination < 71.5 degrees, Astron. J. 111 (4) (1996) 1945–1963.

[11] **M.M. Komesaroff,** Ionospheric refraction in radio astronomy. I. Theory, Aust. J. Phys. 13 (2) (1960) 153–167.

[12] **A.A. Lipovka, N.M. Lipovka,** Problems of connection of radio sky to optical sky. History and perspective, Geodesy. 10 (2013) 2–7.

[13] **A.A. Lipovka, N.M. Lipovka,** Sposob privyazki koordinat nebesnykh radioistochnikov k opticheskoy astrometricheskoy sisteme koordinat.

LKL [Method of referencing of celestial radio sources coordinates to optical astrometrical coordinate system, Lipovka – Kostko – Lipovka, (LKL), Patent No. 2 010107938/28(011185), 2011.

[14] The NRAO VLA Sky Survey, URL: http://www.cv.nrao.edu/NVSS/.

[15] **O.V. Verkhodanov, S.A. Trushkin, H. Andernach, V.N. Chernenkov,** Current status of the CATS database, Bulletin SAO. 58 (2005) 118–129. (arXiv:0705.2959), URL: http://www.sao.ru/cats/.

[16] Canadian Astronomy Data Centre, www3.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/.

[17] **N. Zacharias, C. Finch, T. Girard, et al.,** The third US Naval Observatory CCD Astrograph Catalog (UCAC3), URL: http://vizier.u-strasbg.fr/viz-bin/VisieR.

Received 05.04.2018, accepted 07.05.2018.

THE AUTHORS

LIPOVKA Anton A.
Center of Physical Studies, University of Sonora, Hermosillo, Mexico
nila_lip@mail.ru

LIPOVKA Neonila M.
nila_lip@mail.ru