Adv Syst Sci Appl 2020; **04**; 113-124
Published online at https://ijassa.ipu.ru.

Development of Unified Approaches to Building Neural Network and Mathematical Models Based on Digital Data

Nailia Gabdrakhmanova[1*], Maria Pilgun[2]

[1)] *Peoples Friendship University of Russia (RUDN University), Moscow, Russia*

E-mail: gabd-nelli@yandex.ru

[2)] *Institute of Linguistics Russian Academy of Sciences, Moscow, Russia*

E-mail: mpilgun@iling-ran.ru

Abstract: The paper considers the problem of developing approaches to building mathematical models based on digital data of real objects. The data are in text format and contains information about the behavior of the dynamic system. The information selected from the text data enables building of neural network and mathematical models of the dynamic system. The adequacy of the models is evaluated by analytical and numerical methods. The results are meaningfully interpreted. As a result of the study, it was confirmed that the algorithms and approaches for building mathematical models to solve the considered range of problems using digital data can be unified. The analysis of the obtained solutions showed that the conclusions drawn on the basis of the built mathematical models and the conclusions drawn with the semantic neural network analysis of texts are consistent with each other. Therefore, one can talk about the positive results of the models developed. The models developed can be used in solving managerial tasks, planning and situation prediction.

Keywords: neural network, time series, differential equations, dynamic system, stability, text analysis

1. INTRODUCTION

The purpose of this study is to develop approaches for building mathematical models based on extracting information from large data arrays of various formats. It is assumed that the built neural network and mathematical models will help solve the problems of choosing a managerial decision. For example, the city administration faces such tasks, when choosing a project to be implemented in the city. In such tasks, it becomes necessary to obtain feedback, some evaluation of the attitude of the population to the projects being implemented. The building of mathematical models from digital data, in contrast to mathematical models of technical systems, poses a number of difficulties: the data are in text format; it is impossible to conduct experiments; the problem belongs to the "hard to formalize" class. At present, the progress of neural network technologies allows solving many problems. As shown in this paper, machine learning make it possible to extract hidden information from data of various formats and build mathematical models on their basis.

1.1. Data

The material for the study was the data of social media, microblogs, blogs, instant messengers, videos, forums and reviews on the construction of the South East Chord (SEC)

* Corresponding author: gabd-nelli@yandex.ru

and Nord West Chord (NWC) in Moscow. The data were collected between 00:00 on July 7, 2019 and 11:59 PM on December 31, 2019 (see Table 1.1.).

Table 1.1. Data characteristics

Parameter	Messages	Authors	Loyalty	Involvement	Audience
Nord West Chord		433	0,9	7 858	6 822 650
South East Chord	12 456	7 459	0,1	144 028	17 688 221

1.2. Method

The study involved several parallel models. With the help of the semantic neural network model, a text analysis was performed; the topic structure and the semantic network were analyzed, as well as the sentiment and the level of aggression; digital conflictogenic zones were identified; and the indices of social stress and well-being were calculated. In parallel with the neural network model, a statistical analysis of experimental data was performed and dynamic models of processes were built. The following models were used for data analysis: a model based on the neural network paradigm of using neural-like elements with time summation of signals [5] and that with differential equations [1]. Analysis of solutions for all models showed the consistency of solutions.

The chosen model enables automatic semantic ranking of the text data base using several algorithms: an algorithm for forming a homogeneous frequency network of text using an artificial neural network based on neural-like elements with time summation of signals, and an iterative Hopfield-like algorithm for ranking network vertices on a scale of "0 -100%" values. In addition, the n-gram network representation is formed by iterative re-weighting at a given number of steps, or based on the ranking process convergence criterion. Thus, lexemes are analyzed in the context of syntagmas of a given (n) length on a semantic network formed on the basis of text analysis. The frequency network of the text is built as a set of pairs of words that occur in the sentences of the text. The network vertices are weighted by their frequency of occurrence in the text. The weight of a pair of vertices in the network corresponds to the frequency of occurrence of word pairs in sentences of the text. The neural network approach makes it possible to make a correct analysis of sentiment and aggression and derive indices of social stress and well-being, which are necessary for solving the problems of this study [5; 6].

2. RESULTS AND DISCUSSION

2.1. Semantic neural network model

The dynamics of users' activity (Figures 2.2; 2.4) and generated content (Figures 2.1; 2.3) regarding the implementation of the SEC and NWC projects, enable determination of the communication features and identification of potential digital conflictogenic zones.

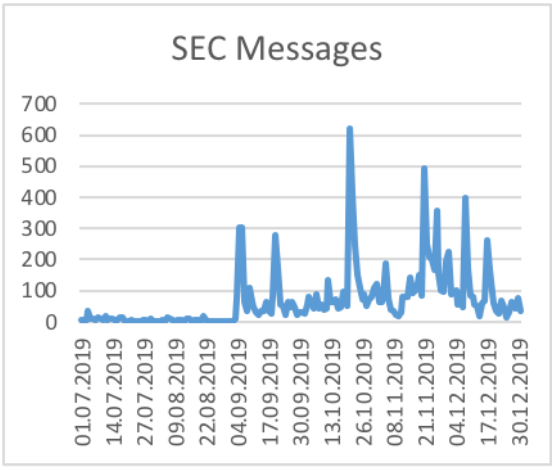

Fig. 2.1. Dynamics of messages (SEC)

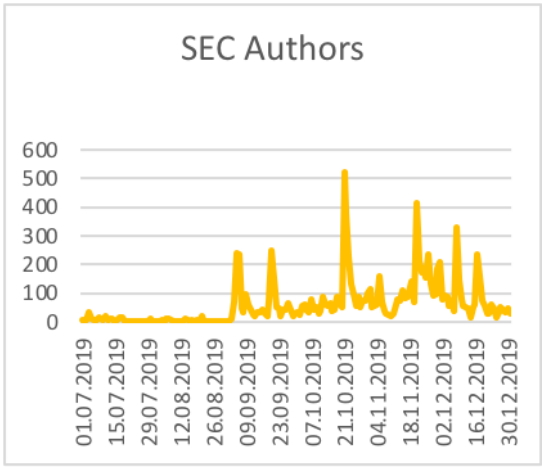

Fig. 2.2. Dynamics of users' activity (SEC)

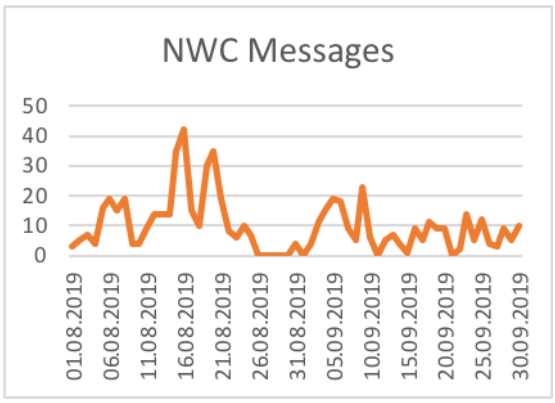

Fig. 2.3. Dynamics of messages (NWC)

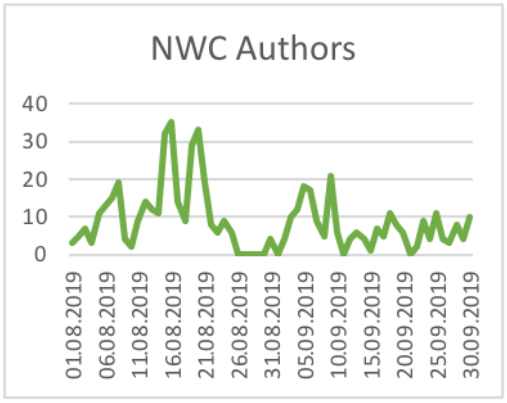

Fig. 2.4. Dynamics of users' activity (NWC)

To identify key topics, the user-generated content was analyzed with the allocation of the semantic network, as well as the topic structure, followed by the content analysis. The core of the semantic network composed of nominations with a link weight of 98-100, clearly demonstrates the problems that concern the residents and the complaints of Muscovites against these projects (Figures 2.5; 2.6.).

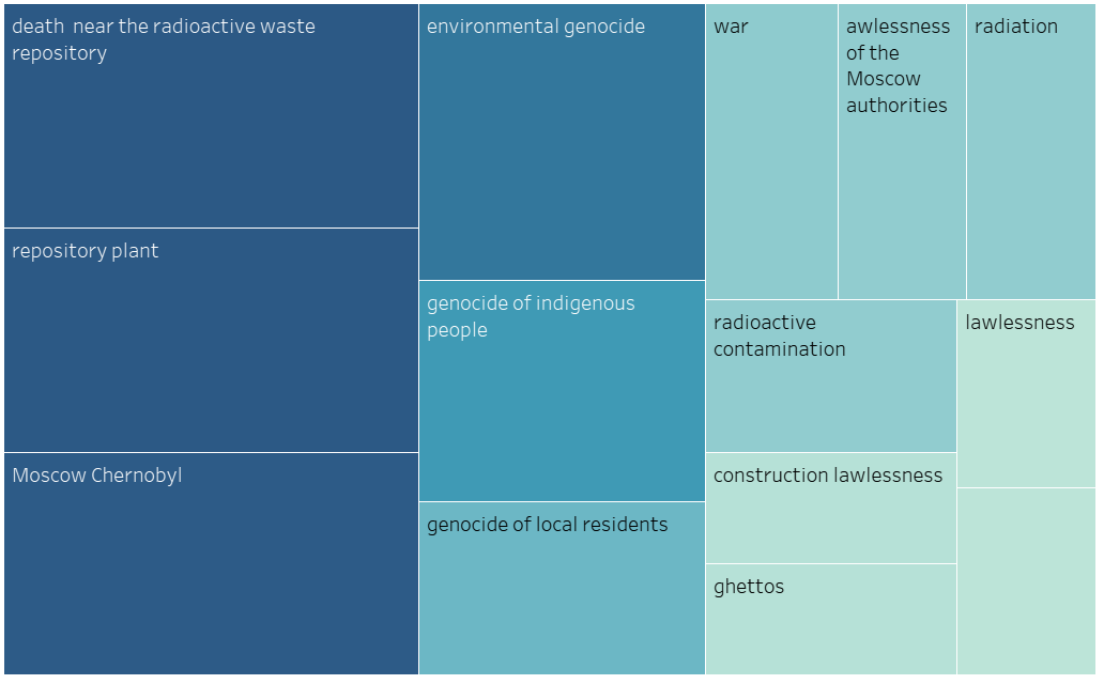

Fig. 2.5. Core of the semantic network (SEC)

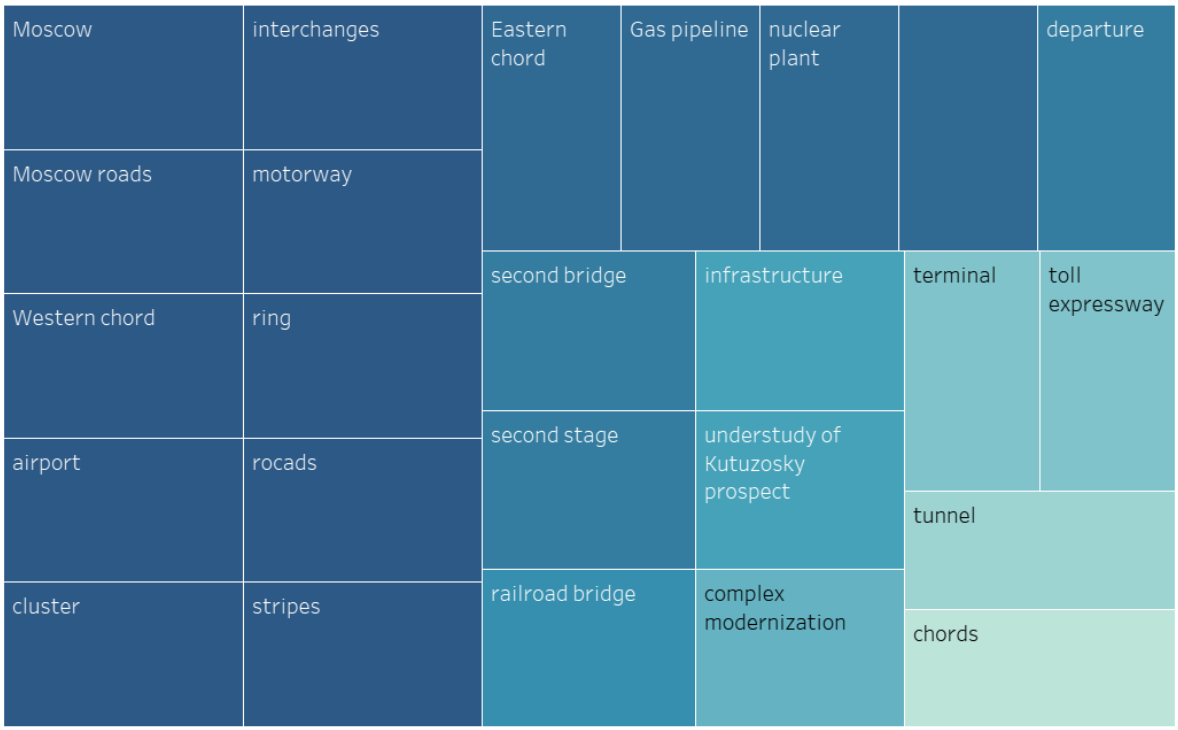

Fig. 2.6. Core of the semantic network (NWC)

The SEC-related content includes a large list of problems that concern the residents: the destruction of the radioactive burial ground of the polymer plant, the possible ecological disaster, radiation; the growth of the number of cancer diseases, the felling of trees, the all-Russian environmental protest; the residents are ready to defend their interests in protest actions both in virtual and real space; a sharp negative attitude towards the actions of representatives of the administrative apparatus; expansion of the Moscow borders, renovation, construction of other road and transport facilities; violation of work methods, lobbying for laws, deterioration of the transport situation; fake expertise papers.

Positive and neutral messages are presented in the official content containing a description of the benefits that the residents will receive from improving the transport situation and development of the city's transport system.

The North-West Chord-related content includes official messages mainly; the residents almost do not generate any content, which indicates a calm attitude towards the project. Key topics are related to coverage of different stages of the construction implementation, as well as information about similar projects; they present the North-West Chord as part of the new Moscow Chord Ring.

The sentiment analysis made it possible to clusterize the content by sentiment (Figure 2.7.).

Analysis of the level of aggression showed a high conflict potential of messages dedicated to the SEC and the absence of indicators of the conflict potential for the content dedicated to the NWC (Figure 2.8.).

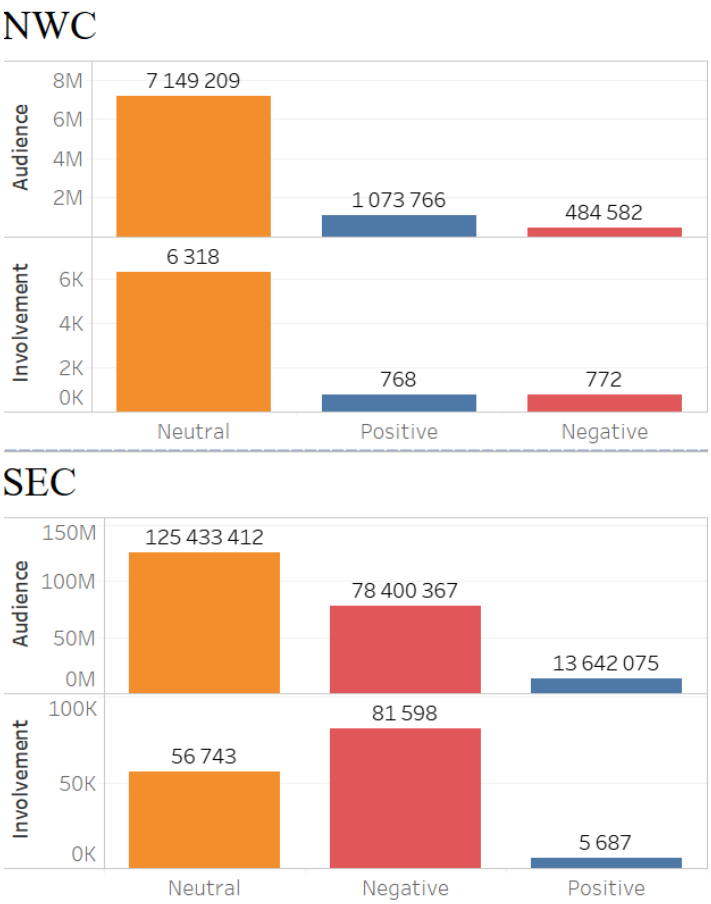

Fig. 2.7. Sentiment-analysis results, sentiment clusterization

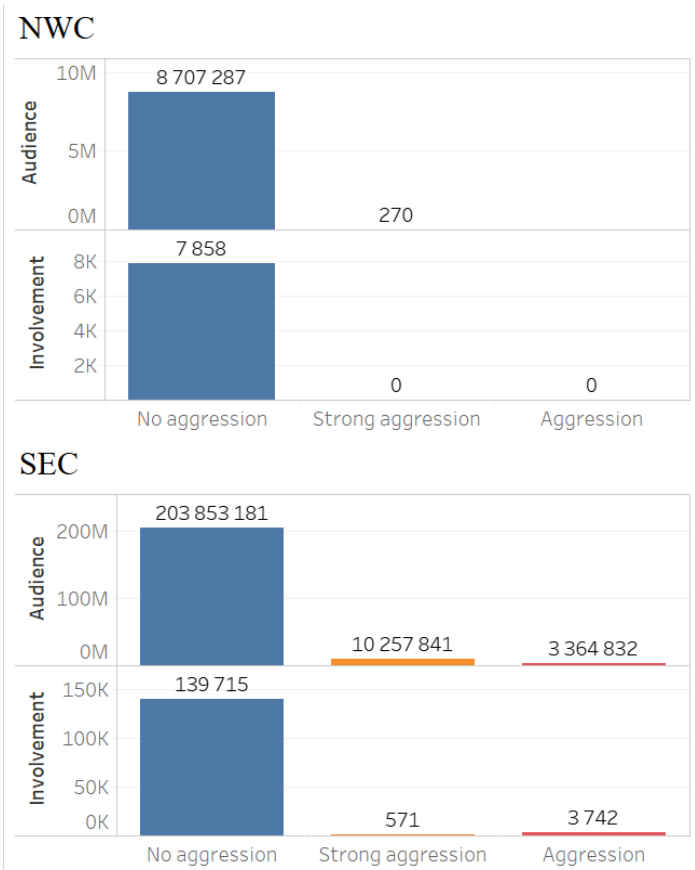

Fig. 2.8. Level of aggression in the content

Digital conflictogenic zones were identified taking into account the dynamics of the actors' activity, the user-generated content, as well as the analysis of textual data that make up the negative cluster. The most indicative criterion was the identification of the presence or absence of aggression; in particular, the presence of strong aggression indicates a high level of negative attitude of local residents towards the SEC project.

The derivation of the social stress and well-being indices made it possible to calculate the quantitative expression of the residents' attitude towards the implementation of the SEC and NWC urban development projects (Figure 2.9.).

.

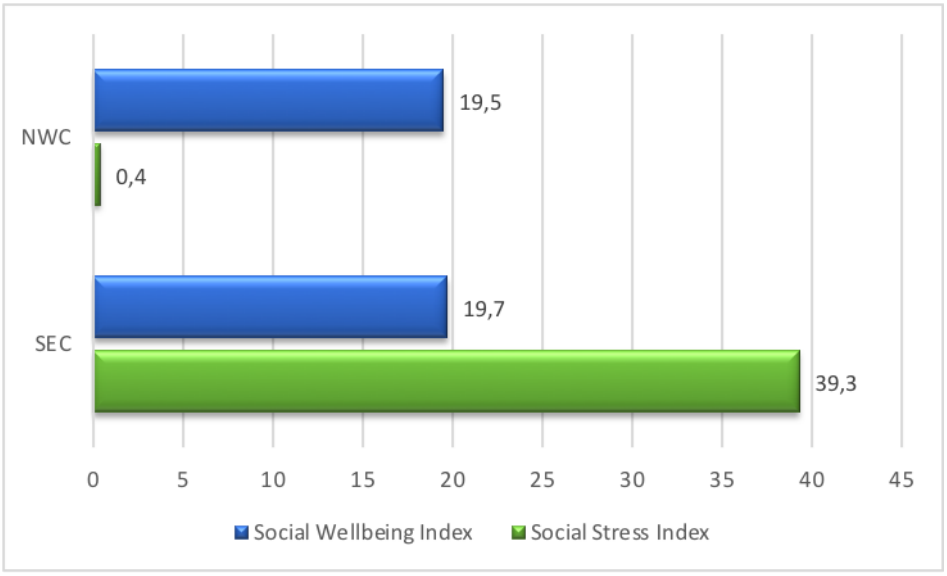

Fig. 2.9. Social Stress Index and Social Well-being Index

2.2. Building of mathematical models

This section is dedicated to the building of mathematical models with a focus on the study of the system dynamics (the dynamics of the population's attitude towards the project P under consideration). The problem was solved on the basis of research data on the attitude of the population to the construction of the South-East (SE) Chord and the North-West (NW) chord.

In the previous section, text data were analyzed after filtering using associative neural networks and visualization programs. Further, based on the data selected using the machine learning method, it is required to build mathematical models. To build mathematical models, it is necessary to formalize the problem under consideration.

2.3. Formalization of the problem of building a mathematical model

Given: set W'={w'(t)}, t is within [0,T], where w'(t) is a numerical vector corresponding to the text at time moment t after its vectorization. The elements of the set are divided into clusters according to the criterion of nearness to some feature set. In our problem, this is a relation to a certain project P. In textual analysis, several methods have been developed for solving this problem. For example, one way to determine the proximity of vectors is to calculate the cosine distance using the cosine of the angle between the vectors.

The clustering resulted in array $W = \{w(t)\}_{t=1}^{T}$, where t is time, w(t) is a numeric vector of messages after filtering by the criterion of nearness to the feature space.

At the next step, the set of elements W was clustered in two ways presented below.

First. The entire set W is divided into two clusters by the feature of positive and negative attitude towards the project. The resulting clusters with numerical vectors of texts were designated as K1 ('aggression present') and K2 ('no aggression'). "Aggression present" means a negative attitude towards project P. "No aggression" means a positive attitude towards project P.

Second. The entire set W is divided into three subclusters according to the degree of negative or positive attitude to project P. We called this feature tonality and identified three degrees of tonality. The constructed subclusters are designated X1 ("negative"), X2 ("neutral"), X3 ("positive").

2.4. Data pre-processing

The entire time interval [0; T] is divided into n intervals with the same increment. The following designations are introduced: t is the number of the interval, t=1,2,…,n.

The following terms are introduced:
$k_1(t)$ is the number of elements in cluster K1 at time t,
$k_2(t)$ is the number of elements in cluster K2 at time t,
$x_1(t)$ is the number of elements in cluster X1 at time t,
$x_2(t)$ is the number of elements in cluster X2 at time t,
$x_3(t)$ is the number of elements in cluster X3 at time t.

In the models under consideration, it is assumed that an unaccounted individual can join one of the groups or move from one group to another through information received either from the media or through interpersonal communication from a previously informed individual.

To build models, the data of all groups are standardized according to formula $\tilde{x} = \frac{x-\bar{x}}{\hat{\sigma}}$, where $\bar{x}$ is the estimation of expectation, $\hat{\sigma}$ is the estimation of the mean square deviation, and $\tilde{x}$ is the standardized value of x. Fragments of standardized data $\{x_1(t)\}$, $\{x_2(t)\}$, $\{x_3(t)\}$, t=$\overline{1,n}$ are presented in Table 2.1 and in the graphs in Fig. 2.10. and Fig. 2.11.

Table 2.1. Fragment of initial data

t	x_1	x_2	x_3
1	-1.28	-1.66	-1.13
2	-0.93	-0.57	-0.72
3	-0.34	-0,91	-0.59
4	0.23	1.51	0.28
5	0.09	0.24	0.03
6	-0.65	0.01	-0.96
7	-0.59	-0.22	-0.68
8	-0.37	-0.68	-0.46
9	-0.45	-0.28	-0.85
10	-0.56	0.35	-0.43
11	-0.40	-0.80	-0.30
12	0.22	0.76	-0.11
13	-0.64	-0.80	-0.71

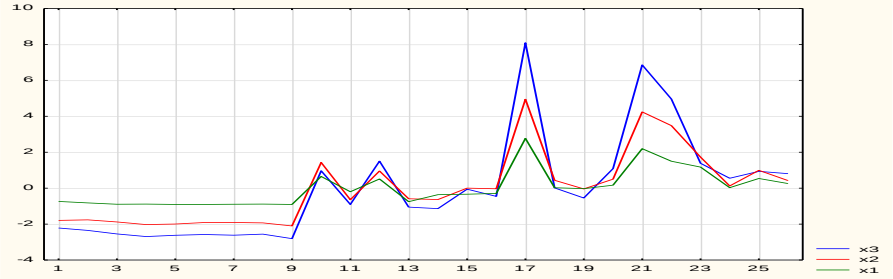

Fig. 2.10. Fragment of time series $\{x_1(t),x_2(t),x_3(t)\}$ for the SEC data

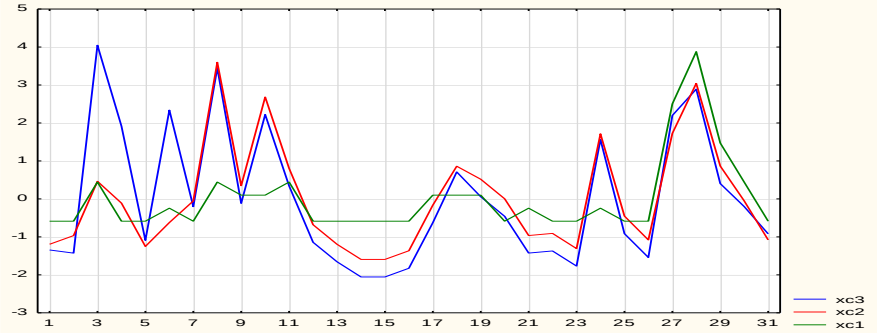

Fig. 2.11. Fragment of time series $\{x_1(t),x_2(t),x_3(t)\}$ for the NWC data

2.5. Model 1. Autoregressive time series models

Data $\{x_1(t)\},\{x_2(t)\},\{x_3(t)\}$, $t=\overline{1,n}$ are time series. Time series models can be built using neural networks [3,4] and methods of mathematical statistics. For our time series, we built autoregressive models. Autoregressive models can be used to predict the dynamics of series a few steps ahead. Coefficients in time series models indicate the direction in the behavior dynamics.

Studies of the behavior of time series have shown that the values of the series are affected by such factors as holidays, weekends or weekdays, time of day, etc.

To smooth over these factors, the following transformation was used (see Fig. 2.12.):

$F: (x_1(t),x_2(t),x_3(t)) \to (N1(t),N2(t),N3(t))$ according to the following rule:

$$N1 = \frac{x_1}{x_1+x_2+x_3}, \quad N2 = \frac{x_2}{x_1+x_2+x_3}, \quad N3 = \frac{x_3}{x_1+x_2+x_3}.$$

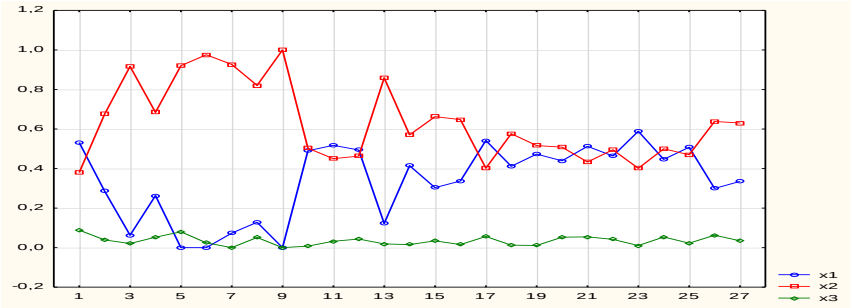

Fig. 2.12. Fragment of transformed time series $\{x_1(t),x_2(t),x_3(t)\}$ for the SEC data.

The time series of the NWC and SEC data were studied for stationarity; the coefficients of autocorrelation and partial autocorrelation were found; and estimates of the structural orders were obtained. Autoregressive models were built for each time series using the Statistica 10 software.

Model 1 results

Based on the experimental data for the period of n NW and SE facilities, autoregressive models of time series $\{x_1(t)\}$, $\{x_2(t)\}$,$\{x_3(t)\}$ were built. Fig. 2.13 shows trends of time series $\{x_1(t)\}$, $\{x_2(t)\}$,$\{x_3(t)\}$ for the SE facility of cluster K1. The graphs show that the dynamics of SE rows is positive for all rows. To build an autoregressive model, it is required to transform the series so that the series are stationary. Therefore, autoregressive models are built according to series {N1(t)},{N2(t)},{N3(t)}. To set the structure of models (p, q, r), estimates of autocorrelation and partial autocorrelation of each time series were calculated. The results of estimating the coefficients of autoregressive models are shown in Tables 3, 4. It is important to note that almost all observations for SE facilities fell into cluster K1, and for NW – into cluster K2. The analysis of the coefficient estimates shows that for the SE facilities the negative attitude towards project P is growing. In models (1), the coefficient before N1 is positive and has a value close to 1, and the coefficient before N3 is negative. The coefficients show that the negative attitude is growing strongly, and the positive attitude is decreasing. Models (2) indicate that for the NW facility, the positive attitude to project P grows faster than the negative one.

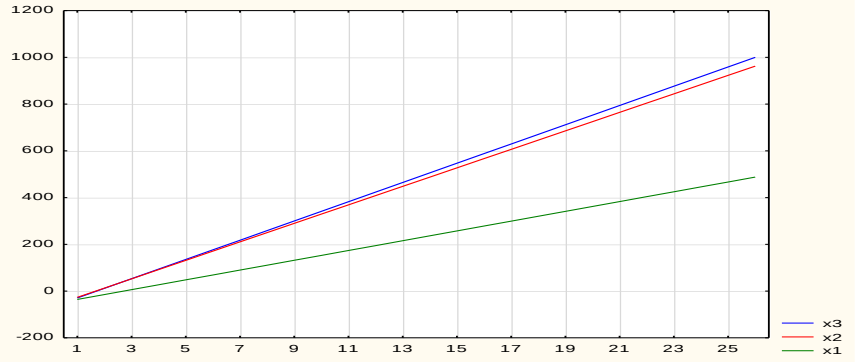

Fig. 2.13. Trends of time series $\{x_1(t)\}$, $\{x_2(t)\}$,$\{x_3(t)\}$ for the SE data.

Autoregressive models for the SE data are given below:

$$\begin{aligned} N1(t) &= 0.4 + 0.74N1(t-1) + 0.32\varepsilon\ ; \\ N2(t) &= 0{,}6 + 0.74N2(t-1) + 0.32\varepsilon\ ; \\ N3(t) &= -0.06N3(t-1) + 0.032\varepsilon\ , \end{aligned} \qquad (1)$$

where $\varepsilon \epsilon N(0,1)$.

Autoregressive models for the NW data are as follows:

$$\begin{aligned} N1(t) &= 0.48N1(t-1) - 0.32\varepsilon\ ; \\ N2(t) &= 0.77 - 0.14N2(t-1) - 0.7\varepsilon\ ; \\ N3(t) &= 0.07 + 0.7N3(t-1) + 0.4\varepsilon\ , \end{aligned} \qquad (2)$$

where $\varepsilon \epsilon N(0,1)$.

2.6. Model 2. Modeling with differential equations

The dynamics of the process under consideration can be described using the type-epidemiological mathematical model [1,2,7]. In this case, we study the dynamics of the process as some type of epidemiological contact process. There are various kinds of models. The type of models is selected based on the analysis of the interaction of the selected groups in the model. Taking into account that there is no contradiction in the problem under consideration, the type of model was chosen using systems of linear differential equations:

$$\begin{cases} \frac{dx_1}{dt} = \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3 + f_1(t) \\ \frac{dx_2}{dt} = \beta_{21}x_1 + \beta_{22}x_2 + \beta_{23}x_3 + f_2(t) \\ \frac{dx_3}{dt} = \beta_{31}x_1 + \beta_{32}x_2 + \beta_{33}x_3 + f_3(t) \end{cases} \quad (3)$$

The solution to the system of differential equations (3) is vector X(t)=(x_1 (t), x_2(t), x_3 (t)), then (1) in vector form is written as follows:

$$\frac{dX}{dt} = AX + F.$$

The parameters of our system (3) are unknown. The task of building parametric models arises when it is required to identify the value of a parameter from measurement data. Various methods have been developed to estimate the parameters of model (3) from experimental points, for example, the Nelder-Mead method.

To solve the parameterization problem, a neural network approach was used. It has been proven that the neural network approach allows obtaining sustainable solutions. This is accounted for as follows. Experimental points always have measurement errors. Neural network functions are robust against errors in the original data and can be efficiently implemented in software and hardware. Neural networks of multi-layer perceptron (MLP) and radial basis (RBF) architectures are suitable for solving the problem.

Let us explain the essence of the neural network approach using the example of the simplest problem of solving a differential equation:

$$A(u) = g,\ u = u(x),\ x \in \Omega \subset R^p\ ,\ (*)$$

here $A(u)$ is a differential operator, i.e. an algebraic expression containing derivatives of an unknown function *u*.

Let us search for an approximate solution (*) in the form of an artificial neural network of a given architecture:

$$u(x,w) = \sum_{i=1}^{n} c_i v(x, a_i)$$

Here $w=(w_1,w_2,\ldots,w_n)$ is the vector of weights $w_i=(c_i,a_i)$. The basic neuroelement *v* is set by the choice of the activation function – function *f* of one real variable.

The weights of the neural network – linear input parameters c_i and non-linear input parameters a_i – are involved in the process of step-by-step training of the network that, in the general case, is built on minimization of some error functional *J(w):*

$$J = \sum_{j=1}^{M} (A(u(\xi_j)) - g(\xi_j))^2$$

The point to be emphasized is that what is sought is not the minimum *J(w)*, but point w_η in the space of weights $J(w_\eta) < \eta$ specifying solution $u_\eta = u(x, w_\eta)$. The number $\eta > 0$ is chosen so that the built model is considered sufficiently accurate.

After solving the parameterization problem (3), it is necessary to find the solution for (3) and analyze the obtained solutions. Then, on the basis of the obtained solutions, the choice of the management decision is made. Below is a brief description of the obtained solutions for parametrization problems and solving differential equations using real data of two facilities.

Outstanding mathematician A.N. Kolmogorov [3,7] proposed to abandon the explicit form of functions in the right-hand side of Equation (3) in models such as epidemiological contact processes, considering only certain restrictions on the form of these functions. If we follow the idea proposed by A.N. Kolmogorov, then neural network mapping models are well suited here. In this case, for a system of differential equations written in general form as

$\frac{dX}{dt} = F(X,t)$, the neural network mapping by experimental points will result in *F* in the form of an expansion in terms of the given basic functions of the neural network (exponential, hyperbolic tangent, identity function). The results of solving this problem are provided in the section below.

Model 2 results

When building models described using a system of differential equations in the form of (3) based on observational data {N1(t)},{N2(t)},{N3(t)}, the parameterization problem was solved. Taking into account the stationarity of the transformed series, F(t) was excluded from the right-hand side of (1). Numerous numerical experiments have been performed. As a result of experiments on the SE data, the following models were built:

$$\begin{cases} \frac{dN1}{dt} = 0{,}42N1 - 0{,}4N2 \\ \frac{dN3}{dt} = 0{,}13N1 + 0{,}4N3 \end{cases}$$

The locations of the trajectories in the vicinity of the stationary point were studied. The roots of the standard equation were calculated: $k_1=0{,}4-0{,}3i$, $k_2=0{,}4+0{,}3i$

The general solution of the system is obtained:

$$\begin{cases} N1(t) = e^{0.4t}(c_{11}cos0.3t + c_{12}sin0.3t) \\ N3(t) = e^{0.4t}(c_{21}cos0.3t + c_{22}sin0.3t) \end{cases}$$

The Lyapunov stability parameter has shown that the stationary point for the given system is unstable.

Conclusion. All models built using differential equations, mathematical statistics and neural networks have their own specifics and their significance in the analysis of system dynamics. We believe that when solving this type of problem, it is necessary to build models using a combination of neural network models together with methods of analytical description of processes. Neural network models are not very convenient for analyzing dynamic processes. Models built using differential equations (model 2) and regression models (model 1) are more suitable for analyzing the dynamics of processes. However, neural networks are more stable for approximations with distorted data. For example, in problems of parametrization of differential equations and in problems of finding approximate solutions for differential equations, neural networks have shown good results.

3. CONCLUSION

Algorithms and approaches for building mathematical models to solve the considered range of problems using digital data can be unified. The analysis of the solutions obtained showed that the conclusions drawn on the basis of the built mathematical models and the conclusions drawn using the semantic neural network analysis of texts are consistent with each other. Therefore, one can talk about the positive results of the models developed. The models developed can be used in solving managerial tasks, planning and situation prediction.

The main conclusions of the solved problem are as follows:

- the largest conflictogenic zone was identified in the SEC-related data set. The neutral and positive content include official messages mainly, which cause strong rejection of the residents. The user-generated content is characterized by negative evaluativeness and aggression, which makes it possible to predict further escalation of the conflict;
- the NWC-related content has no signs of conflictogenity.

ACKNOWLEDGEMENTS

The work was supported by RFBR grant 19-08-00261.

REFERENCES

1. Arnold, V.I. (2012) *Geometric methods in the theory of ordinary differential equations*. Moscow: ICMMO.

2. Castillo-Ch´avez, C., Song, B. (2003). Models for the transmission dynamics of fanatic behaviors. In: H.T. Banks, C. Castillo-Ch´avez (Eds.), Bioterrorism: Mathematical Modeling Applications in Homeland Security, in: *SIAM Frontiers in Applied Mathematics*, vol. 28, SIAM, Philadelphia, 2003, 155–172.

3. Gabdrakhmanova N. (2015). Neural network models for solving management problems on the main oil pipeline. *XVII all-Russian scientific and technical conference "Neuroinformatics-2015": Collection of scientific papers. Part 1*. M.: NIAMI MEPhI, 2015. -244, 172-181.

4. Gabdrakhmanova N. (2018). Forecasting time series using topological data analysis. *In ITISE 2018 International Conference on Time Series and Forecasting Proceedings of Papers* 19-21 September 2018 Granada (Spain), 1367-1374.

5. Kharlamov A,A. & Pilgun M. (Eds.) *Neuroinformatics and Semantic Representations. Theory and Applications.* Newcastle upon Tyne**:** Cambridge Scholars Publishing.

6. Kharlamov, A.A. & Pilgun, M. Analysis of the situation connotation on the example of assessing the reaction of society: social media data. *International journal of future generation communication and networking*. (IJFGCN), ISSN: 2233-7857(PRINT); 2207-9645(Online), Nadia, (2020), Vol. 13, No. 3, 37-44.
DOI 10.33832/IJFGCN.2020.13.3.04.

7. Kolmogorov, A. N. (1972). Qualitative research of population dynamics models. *Problems of Cybernetics*. Issue 25. 100-106.

