National Research University Higher School of Economics

As manuscript

Elizaveta Prokofyeva

IMPROVING THE PROCESS OF HEALTHCARE SERVICE IN LOCAL POLYCLINICS ON THE BASIS OF ADAPTIVE MODELS OF THE PATIENT FLOW FORECASTING

PhD Dissertation Summary

for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow- 2021

The PhD Dissertation was prepared at National Research University Higher School of Economics.

Academic supervisor: Svetlana V. Maltseva, Doctor of Sciences, Professor, National Research University Higher School of Economics

Subject of the thesis

Rapidly advancing data analysis technologies are of paramount importance in healthcare. The current level of medical service automation enables the accumulation and processing of large amounts of data and the use thereof to solve optimization problems. Besides, new opportunities for advanced analytics, in particular, of big data, are emerging because of not only the volume, but also the variety of information used in healthcare. In this regard, open data of Internet resources, patient chats and feedback on the services, electronic prescriptions, pharmacy reviews, newsfeeds of social media, online consultations and many other data sources become the resources of new valuable knowledge in the field of healthcare and promote the development of the personalized service model.

As part of the national project on digitalization of the economy, the programme "Digital Economy of the Russian Federation" was developed and approved by Order of the Government of the Russian Federation dated July 28, 2017 No. 1632-r. It focuses on, inter alia, healthcare as a separate sector. Primary end-to-end digital technologies of the Programme include big data, neurotechnology and artificial intelligence. Therefore, today the problems of analysis and modelling of large arrays of city healthcare data using artificial intelligence technology are of particular importance for the development of industry solutions to ensure digitalisation of the economy, with data as a key production factor.

A critical area of application of data processing technology in healthcare is the improvement of work of medical institutions: creation of an effective work schedule for medical staff, patient flow prediction, resource planning and allocation, reducing queues, and solution of other tasks. Up-to-date analytical technologies make it possible to develop decision-making tools based on empirical data. For example, aggregated data on the actual movement of patients between hospitals and the movement of specialists within the institutions allow to plan the workload, ensure a high level of service availability, and optimise the institution's work based on the actual demand for the services. Such data are used for the active development of information systems of patient flow management.

Big data analytics enables not only the detection of the main patterns of movement within medical and preventive treatment facilities, but also the revelation of the most critical bottlenecks of the medical service system in order to meet the public demand for services with the given economic parameters. The key aspect is the ability to form predicted values to be used as the basis for the creation of an alternative organisational service model using optimal resources. One of the essential requirements for the prediction models is their adaptability, since it

ensures that the external influence on the input flows of a medical institution is considered.

International healthcare systems have showed an unprecedented growth of demand for adaptive patient service methods due to the spread of SARS-CoV-2 recognised by the WHO (World Health Organization) as a pandemic. The development of such methods will contribute to the significant improvement of the medical service efficiency and reduction of organisational costs, thus making the services available to the population.

Therefore, the development and application of data analysis technologies in the analysis and prediction of patient flows in the medical and preventive treatment facilities form the basis for solving the problems of service and resource planning improvement, which confirms the relevance of the subject of this research.

Purpose and objectives of the research

The purpose of the thesis research is to develop a method of determination of the rational resource structure of a multi-field service organisation to ensure management decision-making by developing prediction models to meet the demand for services under existing economic constraints.

To achieve the purpose of the research, the following objectives were established:

1. Analysis of existing approaches and methods for modelling and improvement of the rational structure of service organisations, considering the peculiarities of input flows of service subjects.
2. Development and implementation of the method of analysis and typing of the input flow of requests based on fuzzy clustering algorithms.
3. Development of the method of prediction of the input flow of requests, which allows to consider the peculiarities of external changes.
4. Development of the method of integration of the simulation model of service in a medical and preventive treatment facility, including methods of analysis, typing and prediction of the input flow of requests.

Scientific novelty

This paper for the first time establishes the following scientific results being defended:

1. Original method of analysis and typing of the input flow of requests based on fuzzy clustering within the theory of additive regularisation of topic models.
2. Adaptive method of prediction of input flows, which allows to consider external influence on the input flows of a medical and preventive treatment facility.

3. Advanced model of service in a medical and preventive treatment facility based on the integration of the simulation model, including the developed methods of analysis, typing and prediction of the input flow of requests.

Practical significance

The practical use of the results of this paper will add new functionality to the bundled software tools that support the work of medical institutions. Besides, the research results can be introduced into the training course "Introduction to Big Data" at the Faculty of Business and Management at the National Research University Higher School of Economics, Department of Innovation and Business in Information Technologies, programme: Business Informatics, 3rd year, 1 module, 2018-2019, 2019-2020. The research results were also used in the reports presented at the lectures and seminars of the Winter School of Business Informatics for applicants to the Master's Programme at the National Research University Higher School of Economics during the period of 2017-2020.

Methodological base of the research

Methods of data analysis, systems theory and systems analysis, graph theory, set theory, mathematical modelling, probability theory and mathematical statistics, methods of analysis, modelling and monitoring of the clinical pathways of patients, advanced programming and simulation technologies.

Evaluation of research results. The main results of the research were presented at the following conferences and seminars:

1. XXII April International Academic Conference on Economic and Social Development, April 13-30, 2021, Moscow. Subject of the report: "Prediction of the typed input flow of patients in multi-field medical institutions".
2. The International Conference on Information Systems (ICIS). 5th Annual Workshop on Big Data Applications, 13-16 December 2020, Hyderabad, India. Subject of the report: "Forecasting Heterogeneous Patient Flow through Big Data Application in Medical Facilities for Rational Staffing".
3. 22nd IEEE International Conference on Business Informatics (IEEE CBI 2020) June 22–24, 2020, Antwerp, Belgium. Subject of the report: "Driven Approach To Patient Flow Management And Resource Utilization In Urban Medical Facilities".
4. XXI April International Academic Conference on Economic and Social Development, April 6-10, 2020, Moscow. Subject of the report: "Use of hard and fuzzy clustering methods for segmentation of the input flow of

patients in city polyclinics".

5. Winter School of Business Informatics for applicants to the Master's Programme at the National Research University Higher School of Economics, February 15-19, 2020, Solnechnaya Polyana boarding house, Volkovo, Moscow Region. Subject of the report: "Analysis and visualisation of medical data in Python for predicting patient attendance".
6. The International Conference on Information Systems (ICIS). Annual Workshop on Big Data Applications, 15-18 December 2019, Munich, Germany. Subject of the report: "Data-Driven Approaches for Efficient Patient Flow Segmentation in Polyclinics".
7. Annual Summit of Young Scientists and Engineers "Major Challenges for Society, State and Science", November 2019, "Sirius", Sochi. Subject of the report: "Analysis and modelling of clinical pathways of patients to improve the quality of service in city polyclinics".
8. 21st IEEE 21st International Conference on Business Informatics (IEEE CBI 2020), July 15-17, 2019, Moscow, Russia. Subject of the report: "Application of Modern Data Analysis Methods to Cluster the Clinical Pathways in Urban Medical Facilities".
9. XX April International Academic Conference on Economic and Social Development, April 9-12, 2019, Moscow. Subject of the report: "Application of modern data analysis methods in modelling of the clinical pathways of patients in city medical facilities".
10. Russian-French Workshop in Big Data and Applications, 2017, Moscow, Russia. Subject of the report: "Modeling Demand for Services in the Healthcare Facilities Using Big Data Techniques".

Second-tier publications:

1. E.S. Prokofyeva, S.V. Maltseva, D.A. Qiu Zheng Qing, Automation of organisation resource management based on simulation modelling by predicting input flows//Automation. Modern technologies. 2021. Vol. 75 No. 6. Pp. 272-280.
2. Prokofyeva E., Maltseva S., Tsiu-Zhen-Tsin D. Forecasting Heterogeneous Patient Flow through Big Data Application in Medical Facilities for Rational Staffing// International Conference on Information Systems 2020, Special Interest Group on Big Data Proceedings.
3. Prokofyeva E. S., Maltseva S. V., Fomichev N. Y., Kudryashov A. G. Data-Driven Approach To Patient Flow Management And Resource Utilization In Urban Medical Facilities// 22nd IEEE International Conference on Business Informatics (IEEE CBI 2020). P. 71-77.

4. E.S. Prokofyeva, R.D. Zaytsev, Analysis of clinical pathways of patients in medical facilities based on hard and fuzzy clustering methods // Business Informatics. 2020. Vol. 14. No. 1. Pp. 19–31.
5. Prokofyeva E.S., Zaytsev R.D., Maltseva S.V. Application of Modern Data Analysis Methods to Cluster the Clinical Pathways in Urban Medical Facilities// 21st IEEE International Conference on Business Informatics (IEEE CBI 2019). P. 75-83.
6. Prokofyeva E., Maltseva S. Data-Driven Approaches for Efficient Patient Flow Segmentation in Polyclinics// International Conference on Information Systems 2019, Special Interest Group on Big Data Proceedings.
7. Prokofyeva E.S., Zaytsev R.D., Maltseva S.V. The Demand for the Healthcare Services: the Opportunities of Big Data in Predicting Patient Flow//International Conference Information Systems 2017 Special Interest Group on Big Data Proceedings.

Content of the paper

The introduction validates the relevance of the research, sets the purpose and objectives, practical and theoretical significance of the research, specifies the scientific novelty, reliability and validity of the results obtained. It also describes the evaluation of the research results

The first chapter is focused on the current state of the area, features of modelling of the service system of multi-field service organisations, existing approaches and the methodological base for clinical pathway modelling as an effective healthcare management tool. The chapter establishes the main objective of the research.

Development and implementation of clinical pathways or patient flow patterns is an important healthcare management tool. In general, the clinical pathway is a patient's flow pattern when receiving a medical service in the relevant institutions.

To solve the research objective, a multichannel queuing system (QS) with the limited waiting time was considered.

Two types of input flows are studied: statistical (based on historical data) and prediction. The service request has a certain type, and the service device has performance. The request is considered served if the device performs work and renders a service to satisfy this request.

The following service characteristics were selected as the key performance criteria for this QS with the limited waiting time:

1. Average number of requests in the queue L_q;

2. Average time of the request in the queue T_q;
3. Service channel load intensity φ;
4. System operation costs f.

Determine the costs in the modelled system by the expression [1, 2]:

$$f = c_1 n + c_2 L_q(n, \rho) + c_3 \bar{k}_{\text{св}}(n) \rightarrow min \quad (1)$$

where c_1 −is the cost of operation of one additional service channel (service provider) per unit of time,

c_2 − cost equivalent of the waiting time of one request per unit of time,

c_3 − cost loss from the idle time of one channel per unit of time.

It is necessary to determine the rational number of service channels n with the known intensity of the input flow λ and the service flow μ,, and known constants c_1, c_2, c_3. It is also necessary to compare the system operation indicators $p_0, Q, A, L_q, T_q, L_s, T_s$ for different system constraints: queue length m, service channel load intensity φ.

Given the complexity of the control object and the need to use real big data, the problem has no exact solution. Therefore, simulation modelling methods represent one of the approaches for the study of queuing system operation processes.

This paper suggests for the first time an advanced method of typing of the input flow of the simulation model based on fuzzy clustering within the theory of additive regularisation of topic models (ARTM). The scheme of inclusion of prediction input flows into the simulation model was developed.

Input flows of the simulation model are represented by time series due to the specific structure of the initial data of medical organisations.

The second chapter starts with the description of the formal representation of medical service based on the set-theoretic model.

The structure of a medical institution based on the set-theoretical description can be represented as a certain system of patient service, which can be described by the following system in statics:

$$S_L = \{V, B, H, R\} \quad (2)$$

where $V = \{v_j\}, \mathrm{j} = 1,2, \dots, M$ is a set of service points in a medical institution, v_j is a code of the jth service point, , M −is the number of service points; the number of service points corresponds to the number of functional rooms of the medical institution. The list of service points is determined by the

Regulation [1]on the organisation of delivery of primary medical care to the adult population.

For the elements of this set, a critical parameter is the number of concurrently served patients V' and the number of concurrent service providers V''. Set in the form of a set of parameters :

$V' = \left\{v_j'\right\}, \mathrm{j} = 1,2,\dots,M$where v_j' $-$ is the number of concurrently served patients in the jth service point.

$V'' = \left\{v_j''\right\}, \mathrm{j} = 1,2,\dots,M$where v_j'' $-$is the number of concurrent service providers in the jth service point.

$H = \{h_q\}, q = 1,2,\dots,Q$ is a set of service providers, h_q is a code of the qth provider,, $Q -$ is the number of providers. Service providers mean medical specialists and nursing staff who provide medical services.

The set of connections *R* between objects of the system assigned by the sets *V,B,H,* can be assigned by two matrices – *R´ and R´´*.

Since a single provider can render several services, the connection between services and providers is described by the matrix:

$R' = \left\{r_{lq}''\right\}, l = 1,2,\dots,L;\ q = 1,2,\dots,Q$where r_{lq}' = 1, if the lth type of service can be provided by the qth specialist, 0 – otherwise.

Given the stable distribution of service providers among service points, it is also necessary to consider the connection between providers and service points:

$R'' = \left\{r_{ql}''\right\}, q = 1,2,\dots,Q; l = 1,2,\dots,L$ where r_{ql}'' = 1, if the qth provider renders services at the lth point, 0 – otherwise.

It is evident that a particular service can be rendered at different points, for example, therapists can see patients in several rooms.

The system dynamics is primarily determined by peculiarities of the patient flow, as well as by the dynamics of composition of system objects, for example, dismissal or temporary absence of physicians or medical staff, equipment failure, etc.

$P = \{p_i\}, i = 1,2,\dots,N -$ is a set of patients, p_i is the *i*th patient,
$N -$ is the number of patients. In the existing system $N \to \infty$ can be considered, but it is reasonable to consider *N* that corresponds to a certain average number of patients for a particular period, for example, taking into account the work schedule of specialists.

[1] By Order of the Ministry of Health and Social Development of the Russian Federation dated May 15, 2012 No. 543n "On the approval of the Regulation on the organisation of delivery of primary medical care to the adult population" (as amended), Appendix.

$A = \{a_k\}, k = 1,2, \dots, K$ – is a set of patient types (in accordance with diagnoses and types of clinical pathways), a_k is a code of the *k*th type of the patient,, $K -$ is the number of patient types.

The type of the patient is a parameter for p_i, which determines the clinical pathway, therefore, a set of parameters of patients can be determined using this type of classification.

$A' = \{a_i'\}, i = 1,2, \dots, N$ is a set of patient types from the set *P*, $a_i'$is the type of the *i*th patient, $a_i' \epsilon A$.

The recommended patient's clinical pathway is a diagnosis-based sequence of medical services provided to the patient. In certain cases, the sequence of service provision to patients can be changed. *The recommended clinical pathway* in this research is based on patient management algorithms.

We introduce the following designations:

$S = \{s_i\}, i = 1, \dots, N$ is a set of recommended patient's clinical pathways, s_i is a code of the clinical pathway of the *i*th patient.

We establish the services included in the clinical pathway s_i :

$A^i = \{a_n^i\}, n = 1, \dots, N^i$ where $a_n^i \in B$.

A clinical pathway s_i may include the number of the initial service, $n_0 \in [1,N]$, as well as the matrix of all allowable transitions between services that can lead to different sequences:

$U^i = \{u_{lr}^i\}, l = 1, \dots, L; r = 1, \dots, L$where u_{lr}^i = 1, if the clinical pathway s_i allows the transition from the service of the *r*th type to the service of the *l*th type, 0 – otherwise.

The clinical pathway of the $i -$th patient is determined based on the patient type a_i', and diagnosis, a_i''if the patient is assigned to the type…

Along with the recommended patient's clinical pathway, it is possible to determine *the implemented clinical pathway* based on the actual examinations of the patient and visits to specialists.

It is critical that *the implemented* clinical pathway may differ from *the recommended one* due to the limited resources of the institution (for example, in case of queues), the patient's specific behavioural pattern, the provider's solution, and any other factors.

Therefore, it is possible to set *the implemented clinical pathway* in the same way as the recommended one:

$S_p = \{s_{pi}\}, i = 1, \dots, N$ is a set of implemented patient's clinical pathways, s_{pi} is a code of the implemented clinical pathway of the *i*th patient.

We establish the services included in the clinical pathway s_{pi} :

$A_p^i = \{a_{pn}^i\}, n = 1, \dots, N_p^i$ where $a_{pn}^i \in B$.

The clinical pathway s_{pi} may include the sequential numbers of services, codes of specialists who rendered services, the service waiting time, the service provision time, the cost of the service in the following arrays:

In terms of patient service, the system $S_L = \{V,B,H,R\}$ shall minimise the service waiting time and the cost of the service.

The system shall also minimise system operation costs, which implies, in particular, the compliance of the physicians' load with the standard load indicators, as well as the provision of services under all requests of the patient's clinical pathways, which requires the involvement of a wide range of various specialists.

The dynamics of work of a medical institution is also determined by uneven flows of patients at different time of the day, as well as by various emergencies, for example, epidemics, when the demand for medical services goes up with a jump.

The input flow typing is reduced to the solution of the object clustering problem. To construct a cluster model, two methods were compared – *k*-medoids (a more outlier-resistant version of the *k*-means algorithm) and a hierarchical agglomerative algorithm with the Ward connection type.

After discarding abnormally long pathways, the upper limit of the pathway length was set equal to 26 events: $Q50 + 3 \cdot (Q75 - Q50)$, where $Q50$ is the median, and $Q75$ corresponds to 75% of the quantile. R programming language was chosen for data analysis. The language is well suited for research tasks, since it contains a large library of packages for various scenarios. The use of R language also enables data visualisation to understand the overall picture of the subject area. The Stringdist package was used to construct the distance matrix with the help of the OCA method (Damerau–Levenshtein distance metric).

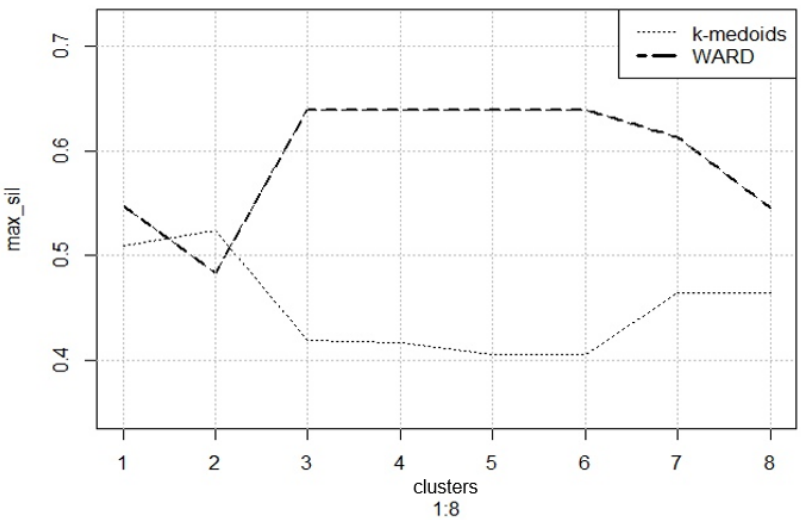

Figure 1. Dependence of the maximum silhouette coefficient in the model on the number of clusters for *k*-medoids and Ward methods

To find a model containing the cluster with the maximum silhouette coefficient, the Ward method and *k*-medoids were compared (Figure 1). Clusters are located on the X axis, and silhouette values are located on the Y axis.

Following the analysis of the coefficient values, the Ward method was chosen, and the trends of the resulting groups were revealed (Figure 2). Clusters with a low value were excluded (Figure 3). As a result of the experiments, clusters 5 and 6 with the highest silhouette coefficient were chosen (Table 1).

Table 1.

Silhouette coefficient values for clusters

Cluster number	Number of objects	Silhouette coefficient value
1	118	0.02
2	239	–0.02
3	193	0.002
4	228	–0.05
5	79	0.29
6	118	0.64

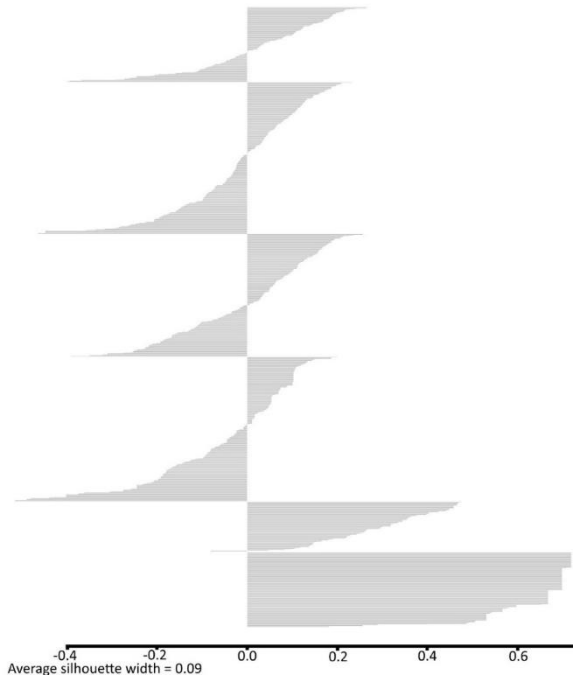

Figure 2. Silhouette coefficients of six resulting clusters

Following determination of the optimal number of clusters, process maps were separately formed for the resulting groups. For example, the process map for cluster 5 contains the main stages of the clinical pathway of patients with sepsis as

graph nodes: start, registration with the corresponding department, antibiotic intake, etc. The graph arcs show the transition of patients through the stages of treatment. The numbers on the arcs correspond to the number of persons making the transition between the nodes. More significant patient patterns are shown with wider arcs. Therefore, the process map enables the quick assessment of the most loaded routes of a medical institution. Besides, such maps can be interpreted by medical specialists by comparing with the accepted medical standards to reveal congested resource units and reorganise the service process.

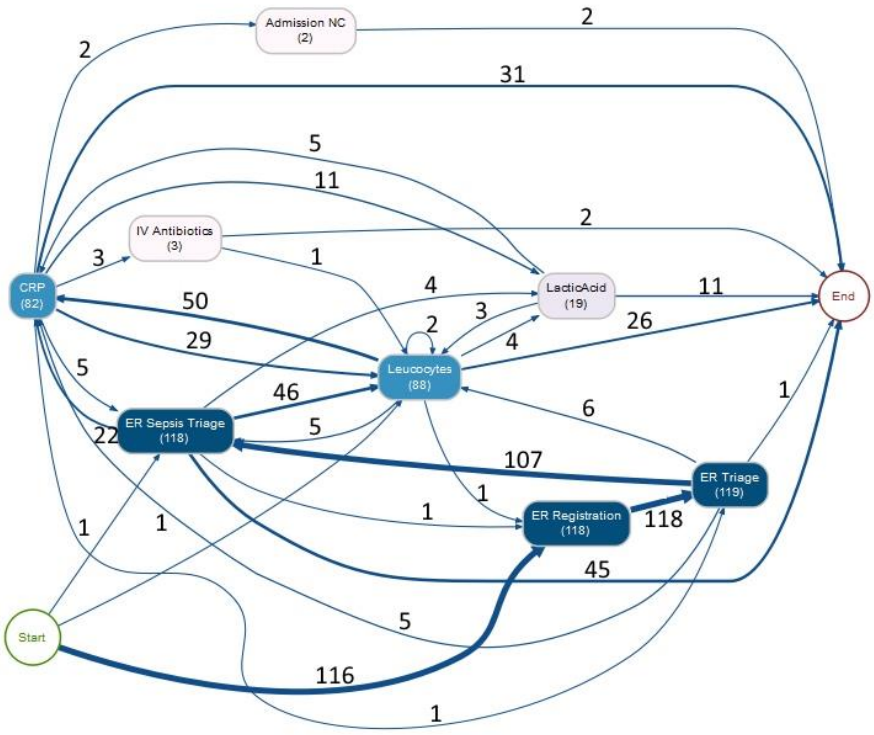

Figure 3. Process map for cluster 5

The next stage included fuzzy, or soft, clustering of the initial data using topic modelling methods, in which the patient pathway can be assigned to several patterns (topic clusters) with different probabilities.

An advanced method of typing of the object's input flow based on fuzzy clustering within the theory of additive regularisation of topic models (ARTM) was proposed.

To construct a topic model of the collection of documents D , it is necessary to:

- find a set of topics T,
- carry out the observable distribution of words by topics $p(w|t)$ for all $t \in T$,
- distribute topics $p(t|d)$ for all documents $d \in D$.

In topic modelling terms, the services provided to the patient as part of medical care are correlated with the words of the model (Table 1). The patient's route, consisting of services, is similar to a word document. Therefore, the hidden topics discovered by the algorithm are interpreted as patterns of the patient's clinical pathways.

The hypothesis of the independence of sampling elements ("bag of words") is equivalent to the assumption that the order of terms in documents is not important for topic identification, i.e., the topic of the document can be identified even after a random rearrangement of terms.

Therefore, the fuzzy clustering method allows to divide the input flow for a system with self-arrangement elements, when the sequence of visits to specialists is determined by the patient.

Table 2. Correlation of topic modelling terms with modelling objects

Topic modelling terms	System's modelled objects
$D -$ collection of documents	$L -$ organisation's event log
$T -$ set of topics	$Z -$ set of pathway patterns
$W -$ glossary of terms	$A -$ list of services rendered
$d \in D -$ document	$s \in S -$ patient's clinical pathway
$t \in T -$ topic	$z \in Z -$ pattern (typical pathway)
$w \in W -$ word	$a \in A -$ medical service
ϕ_{wt} −distribution of words in the topic t	$\phi_{az} -$ distribution of services in each pattern z
$\theta_{td} -$ distribution of topics in the document d	$\theta_{zs} -$ probabilities of patterns in each route s

It is assumed that the appearance of services a in the route s, related to the pattern z, is described by the general distribution $p(a|z)$ for the organisation's event log L and does not depend on the route's document. In topic modelling terms this assumption is called a conditional independence hypothesis:

$$p(a|s,z) = p(a|z); \quad (3)$$

$$p(s|a,z) = p(s|z); \quad (4)$$

$$p(s,a|z) = p(s|z)p(a|z) \quad (5)$$

According to the definition of conditional probability, the total probability formula and the conditional independence hypothesis, the probabilistic topic model is as follows:

$$P(a|s) = \sum_{t \in T} p(z|s)p(a|z) = \sum_z \phi_{az}\theta_{zs}, \quad (6)$$

where $\phi_{az} = P(a|z) -$ are the probabilities of activities in each pattern. Each route is set by s conditional distribution on the set of all patterns z:

$\theta_{zs} = P(z|s) -$ probabilities of patterns in each route. Such distributions allow to calculate the probability of a particular clinical activity of the patient

The ARTM approach is based on the multicriteria regularisation: it allows to construct models that meet several constraints simultaneously. Each constraint is formalised as a regulariser, model parameter-dependent optimisation criterion $R_i(\Phi, \Theta) \to max$.

Weighted sum of all such criteria:

$$R_i(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta) \quad (7)$$

Task of maximisation together with the main likelihood criterion:

$$L(\Phi, \Theta) = \sum_{s \in S} \sum_{a \in A} n_{sa} log \sum_{z \in Z} \phi_{az}\theta_{zs} + R(\Phi, \Theta) \to \max_{\Phi, \Theta} \quad (8)$$

$$\sum_{a \in A} \phi_{az} = 1;\ \phi_{az} \geq 0; \quad (9)$$

$$\sum_{a \in A} \theta_{zs} = 1;\ \theta_{zs} \geq 0; \quad (10)$$

To solve the regularised likelihood task, an EM-algorithm with the modified M-step formulas is used:

$$\phi_{az} \propto (n_{az} + \phi_{az} \frac{\partial R}{\partial \phi_{az}})+ \quad (11)$$

$$\phi_{az} \propto (n_{az} + \phi_{az} \frac{\partial R}{\partial \phi_{az}})+ \quad (12)$$

The thesis study proposes a regularisation strategy based on the assessment of the impact of each regulariser and the combination thereof on the quality of the final model. During optimisation, regularisers are added sequentially at each stage in order to assess the contribution of each of them, in contrast to the previously considered approaches [26–29] in which regularisers have a fixed structure during optimisation.

The standard EM-algorithm for ARTM (Algorithm 2) was modified by adding an external cycle in which the best new model is selected based on the perplexity metric. The adaptation of this standard strategy requires the consideration of data specifics. In terms of regularisation, these subject areas have high dimensionality, and contain a large number of numerical and categorical indicators. The object

type is determined based on the implemented pathway, represented by a time series.

Regularisation coefficients were selected by an experiment, since automatic correction of regularisation strategies in ARTM is an unsolved scientific problem so far.

Algorithm 1. EM-algorithm for ARTM (initial)

Input: collection of documents D, number of topics $|T|$;
Output: Φ, Θ;

initialise the column vectors ϕ_t, θ_d randomly;
repeat
  set to zero n_{wt}, n_{td} for all $d \in D, w \in W, t \in T$;
    for all $d \in D, w \in d$
      $p(w|d) \coloneqq \sum_{t \in T} \phi_{wt}\, \theta_{td}$;
      for all $t \in T$
        $p(t|d,w) \coloneqq \phi_{wt}\theta_{td}/p(w|d)$;
        increase n_{wt}, n_{td} by $n_{dw}p(t|d,w)$;
  $\phi_{wt} \propto (n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}})+$ for all $w \in W, t \in T$;
  $\theta_{td} \propto (n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}})+$ for all $d \in D, t \in T$;
until Θ, Φ converge;

Algorithm 2. EM-algorithm for ARTM (modified)

Input: collection of documents D, number of topics $|T|$;
Output: Φ, Θ;

initialise the column vectors ϕ_t, θ_d randomly;
repeat
  for all $r \in R'$
    set to zero n_{wt}, n_{td} for all $d \in D, w \in W, t \in T$;
    for all $d \in D, w \in d$
      $p(w|d) \coloneqq \sum_{t \in T} \phi_{wt}\, \theta_{td}$;
      for all $t \in T$
        $p(t|d,w) \coloneqq \phi_{wt}\theta_{td}/p(w|d)$;
        increase n_{wt}, n_{td} by $n_{dw}p(t|d,w)$;
    $\phi_{wt} \propto (n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}})+$ for all $w \in W, t \in T$;
    $\theta_{td} \propto (n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}})+$ for all $d \in D, t \in T$;
  ϕ'_{wt}, θ'_{td} are the best
until Θ, Φ converge;

As a result of a series of experiments, the combination of the following regularisers was selected in the uniform grid of values with empirically supported coefficients:

- Decorrelation of service distributions in routes;
- Smoothing/sparsification of service distributions in route patterns;
- Smoothing/sparsification of pattern distributions in routes;

The application of this more flexible approach to clustering of the patient's clinical pathways has not been previously covered by studies. BigARTM open source library in Python, which is based on additive regularisation, was used. Data were converted to Vowpal Wabbit format, which accepts input data in a specific structure:

tag | A feature1: value1 | B feature2: value2.

This format is adapted to the division into categories or modalities during model training. The model was created and trained for the initial number of topics $T = 300$. Based on the calculated perplexity parameters 63.97 and sparsity coefficients Θ=0.44 и Φ=0.42, the optimal number of clusters (9) was chosen.

The probabilistic assessment of assignment to a specific cluster was carried out for each unique patient. Below is an example of distribution by patterns of clinical pathways for one of the patients in the sample: P_1=0.017998157, P_2=0.059349068, P_4=0.5676379, P_6=0.35303143. Accordingly, the next step of the patient (57% probability) will correspond to the behavioural pattern of cluster 4.

Fuzzy clustering allows to add a hierarchical view of patient routes and display the resources of medical institutions. The selected clusters will be used for the improvement of the assigned patient flow prediction, as well as for the formation of recommendations on the resource equipment of hospitals upon development of services.

The third chapter describes the simulation model. Input flows are included in the simulation model successively: first, a series of experiments is conducted based on the statistical indicators of the model medical institution. Then the input flow of requests is represented by the attendance prediction values. For each type of patient, a separate future attendance prediction is prepared based on autoregressive and neural network prediction models of time series or sequential values of attendance at a certain time.

One of the popular attendance prediction tools is SARIMA model (seasonal autoregressive integrated moving average) [30], which was chosen due to its high accuracy demonstrated in the studies aimed at the solution of similar problems [31, 32].

The Holt-Winters model [33] was also constructed and assessed, taking into account the exponential trend and additive seasonality. GRU controlled recurrent unit [34] and LSTM model [35] were chosen as deep learning models.

Creation of a meta-algorithm may improve the presented algorithms (Figure 3). We consider the simplest meta-algorithm using linear regression. In mathematical terms, linear regression for this case is as follows:

$$y_t = w_1 y_t^{EXP} + w_2 y_t^{SARIMA} + w_3 y_t^{LSTM} + w_4 y_t^{GRU} \qquad (14)$$

where y_t^{EXP} is prediction for the tth period of HWES model; y_t^{SARIMA} is prediction for the tth period of SARIMA model; y_t^{LSTM} is prediction for the $t$the period of LSTM model; y_t^{GRU} is prediction for the tth period of GRU model; w_1, w_2, w_3 и w_4 are model weights for prediction.

Another example of the meta-algorithm is a fully connected neural network optimised using the hyperopt library (http://hyperopt.github.io/hyperopt/). Predictions of models $y_t^{EXP}, y_t^{SARIMA}, y_t^{LSTM}$, y_t^{GRU} with n neurons on the first layer are inputs of such a neural network, and the final prediction y_t. is an output. The neural network is trained using back propagation of error.

To assess the prediction accuracy, RMSE (Root Mean Squared Error) was chosen [36]. This minimising metric (2) is often used to study prediction models for patient flows in medical institutions [37, 38]. Peculiarities of use of RMSE for the modelling quality analysis are described in detail in [38].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_i - O_i)^2} \qquad (15)$$

(2)

where $T_i -$ are values of the time series; $O_i -$ are prediction values obtained.

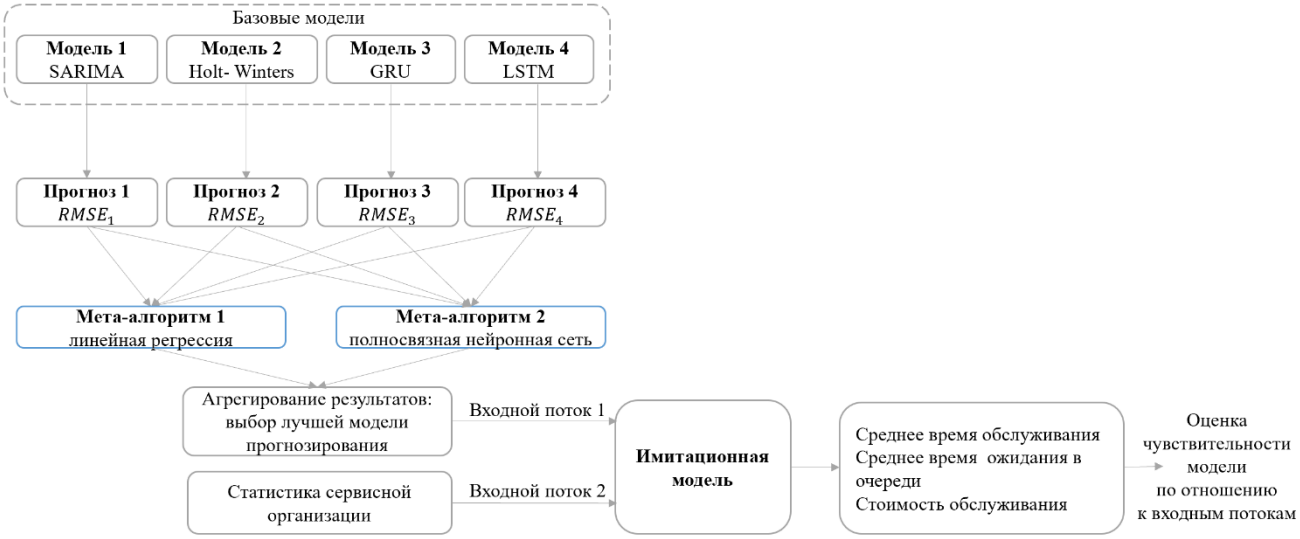

Figure 4. Integration of the predicted input flow into the simulation model

After the assessment of prediction accuracy, the results of the meta-algorithm with the best RMSE metric are integrated into the simulation model. Based on a series of experiments, the sensitivity of the model to the predicted and actual input flows is assessed.

The constraints used upon selection of the service organisation structure in the proposed modelling scheme include the budget, the standard staff load factor and the waiting time in the queue. The selected service organisation structure shall minimise the service waiting time and the cost of service. Besides, it is necessary to ensure that the system operation costs are minimised. Based on the results of simulation modelling, the organisation can create a reasonable development strategy and establish further stages of performance improvement.

Prediction models were trained in Google Colaboratory cloud platform and optimised using the hyperopt library. Table 1 sets out the results of the RMSE metric for each model and each group.

LSTM model has the best results by groups g_2, g_4, g_5, g_6 (highlighted in the table), while SARIMA и GRU models have the best results by groups g_1 andg_3, respectively. The best results of LSTM model are promoted by the specific architecture of this type of recurrent neural networks, which is capable of input data filtering using three types of structures called gates. The LSTM model's structure ensures long-term data storage and use thereof in predicting of attendance of patients in the considered groups.

Table 3.

Results of prediction of attendance in a medical organisation

Model	Result					
SARIMA	4.6180	7.6787	*4.7832*	4.5814	5.0447	4.6760
HWES	4.6507	7.6230	4.9137	4.5863	5.0551	4.6828
GRU	*4.5376*	7.7650	4.8439	4.4483	5.0295	4.6324
LSTM	4.5380	*7.6721*	4.8566	*4.4444*	*4.9595*	*4.6205*

Table 4.

Assessment of meta-models for predicting attendance in a medical organisation based on the RMSE metric

Model	Result					
Linear regression	4.5864	8.0750	4.7243	4.2856	4.8511	4.7925
FCNN	*4.5370*	7.3105	4.6550	4.3562	4.7629	4.7325

Based on the results of the computational experiment (see Table 2), the three-layer neural network shows a significant improvement in the prediction quality, especially for groups g_1, g_2, g_3, g_4 и g_5.

Therefore, the article describes the methods of prediction of a typed input flow of a medical organisation and proposals for the integration of the predicted input flow of requests into a simulation model.

A series of experiments has shown that the creation of a meta-algorithm based on the results of basic SARIMA, Holt-Winters models, GRU controlled recurrent unit and LSTM significantly improves the prediction quality in almost all patient groups.

Conclusion

This thesis work studies the problems of determination of the rational resource structure of a service organisation to ensure management decision-making by developing prediction models to meet the demand for services under existing economic constraints. The paper contributes greatly to the improvement of the efficiency of service in medical institutions:

1. The conducted analysis of existing approaches and methods for modelling and improvement of the rational structure of service organisations, considering the peculiarities of input flows of service subjects, showed the need to develop an alternative method based on data of organisations.

2. An original method of analysis and typing of the input flow of requests based on fuzzy clustering algorithms within the theory of additive regularisation of topic models was proposed and developed.

3. An adaptive method of prediction of the input flow, which allows to consider external influence on the input flows of a medical and preventive treatment facility, was proposed.

4. The advanced model of service in a medical and preventive treatment facility based on the integration of the simulation model, including the developed methods of analysis, typing and prediction of the input flow of requests was proposed.

The study results promoted to the significant improvement of the accuracy of modelling for the selection of a rational organisation's structure. The paper shows the possible implementation of the proposed solutions in city medical institutions.

Bibliography

1. G.S. Osipov. Study of queuing systems with the waiting time in Anylogic/ G.S. Osipov // Bulletin of Science and Practice. – 2016. – No. 10(11). – Pp. 139–151. – DOI 10.5281/zenodo.161072.
2. I.N. Pirogova. Queueing Theory/I.N. Pirogova, P.P. Skachkov, E.G. Filippova. – 2017.
3. E.S. Prokofyeva, S.V. Maltseva, D.A. Qiu Zheng Qing, Automation of organisation resource management based on simulation modelling by predicting input flows//Automation. Modern technologies. 2021. Vol. 75 No. 6. Pp. 272–280.
4. E.S. Prokofyeva, R.D. Zaytsev, Analysis of clinical pathways of patients in medical facilities based on hard and fuzzy clustering methods // Business Informatics. 2020. Vol. 14. No. 1. Pp. 19–31.
5. K.V. Vorontsov, A.A. Potapenko. Regularisation of probabilistic topic models to enhance interpretability and determine the number of topics // Computational Linguistics and Intellectual Technologies: Papers from the Annual conference “Dialogue” (Bekasovo, June 4-8, 2014). Issue 13 (20) – Moscow: RGGU, 2014. Pp. 676–687