**Discrete & Continuous Models**
**& Applied Computational Science**
ISSN 2658-7149 (Online), 2658-4670 (Print)

**2024, 32 (2) 213–221**
http://journals.rudn.ru/miph

*Research article*

UDC 004.8
DOI: 10.22363/2658-4670-2024-32-2-213–221
EDN: CPUADE

# Statistical causality analysis

Alexander A. Grusho[1,2], Nikolai A. Grusho[1], Michael I. Zabezhailo[1],
Konstantin E. Samouylov[2], Elena E. Timonina[1,2]

[1] *Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44 Vavilova St, bldg. 2, Moscow, 119133, Russian Federation*
[2] *RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation*

(received: April 14, 2024; revised: April 25, 2024; accepted: April 30, 2024)

**Abstract.** The problem of identifying deterministic cause-and-effect relationships, initially hidden in accumulated empirical data, is discussed. Statistical methods were used to identify such relationships. A simple mathematical model of cause-and-effect relationships is proposed, in the framework of which several models of causal dependencies in data are described – for the simplest relationship between cause and effect, for many effects of one cause, as well as for chains of cause-and-effect relationships (so-called transitive causes). Estimates are formulated that allow using the de Moivre–Laplace theorem to determine the parameters of causal dependencies linking events in a polynomial scheme trials. The statements about the unambiguous identification of cause-and-effect dependencies that are reconstructed from accumulated data are proved. The possibilities of using such data analysis schemes in medical diagnostics and cybersecurity tasks are discussed.

**Key words and phrases:** finite classification task, cause-and-effect relationships, machine learning

**Citation:** Grusho A. A., Grusho N. A., Zabezhailo M. I., Samouylov K. E., Timonina E. E., Statistical causality analysis. *Discrete and Continuous Models and Applied Computational Science* **32** (2), 213–221. doi: 10.22363/2658-4670-2024-32-2-213–221. edn: CPUADE (2024).

## 1. Introduction

The simplest idea of a causal relationship is given by the functional dependence $y = f(x)$, in which $x$ is the cause of the appearance of the consequence $y$, if their dependence is described by the functional relationship $f$. In this simplest example, the determinism of the appearance of an effect is expressed when a cause appears and a known $f$. However, the mechanism of generating an effect from a cause is not always known, but this does not reduce the importance of causality in solving practical problems. Therefore, the task arises of searching for cause-and-effect relationships using statistical methods in some limited conditions. There are many probabilistic models and ways to restore\reconstruct cause-and-effect relationships with some confidence [1]. To a large extent, these models and methods are associated with various practical applications of cause-and-effect relationships [2]. Let's look at some of them.

Judea Pearl's causal inference model [3, 4] can be expressed in its simplest form by the ratio $y = f(x, u)$, where, in addition to the initial definition of causality, an argument $u$ of randomness is added. Then all calculations related to the causal effect are better expressed in terms of mathematical statistics. This is due to the fact that, as a rule, nothing is known about the distribution of the random component and statistical estimates are used in practice.

This model, among other things, allows us to formulate and solve problems of the influence of therapeutic effects of drugs on patients with a certain disease. Here, $u$ plays the role of individual patient characteristics that affect the effectiveness of treatment with the drug in question.

Another interpretation of this model is that the degree of lesion in a certain diagnosis depends not only on the underlying cause of the disease, but also on the individual characteristics of a particular patient. These features form a random component of the depth of the patient's lesion [5].

Let's turn to the practical procedure of making a diagnosis in a medical study. As a rule, there are a large number of parameters that the doctor considers when making a diagnosis. Each value of these parameters defines a property that can carry information about a possible diagnosis, but may not affect the diagnosis at all. Based on experience, the doctor identifies properties that have a causal relationship with a possible diagnosis. Then, in the collected analyses and other auxiliary data, we are talking about the consequences of the cause, which is determined by the property of diagnosis (disease) and manifests itself in the consequences observed by the doctor.

In this paper, we consider a model of deterministic causation in the presence of a significant number of properties that are not related to the causal effect of some properties on others. Despite the simplified nature of the model, it manages to describe an important case of the appearance of some properties that can enhance or weaken the degree of damage when exposed to the underlying cause (confounding [6, 7]). This case is described in the article with the help of a complex cause, when several properties form a cause, as a result of which a consequence property appears.

Another class of applications of the models of the appearance of deterministic cause-and-effect relationship considered in the article against the background of a sequence with other properties that are not directly related to the desired causality is the search for traces of an insider in the observed sequences of data monitoring user actions in banking applications [8, 9]. In the considered task, several possible scenarios of insider actions were analyzed, the properties of which could manifest themselves in the monitoring data. These properties were tracked in Big Data monitoring, which made it possible to successfully solve operational information security tasks. The desired properties appear in the monitoring data as inclusions of effect properties, where the cause is not the observed actions of an insider. This scheme is transferred to interspersed anomalies arising from causes caused by failures of software and hardware systems.

The problem of trust in the results of computer data analysis in diagnostic tasks and critical calculations is important. The scheme proposed in the article can be used to increase confidence in the results of complex computer calculations. For example, we highlight a number of input and output parameters that consistently belong to input and output data clusters. If consistency is maintained in the next iteration of calculations, then this fact increases confidence in the results of calculations. If there are contradictions in the consistency of input and output data, then it is necessary to look for the reasons\causes for such a mismatch. At the same time, the remaining input and output data do not participate in such control and are considered random noise [10]. The explicability of the formed conclusions is most often realized with the help of causal relationships identified in the analyzed data [11].

## 2. A simple mathematical model of cause-and-effect relationships

Let $A = \{x_1, ..., x_n\}$ be the set of observable properties of the data. Let's consider the simplest case, when the data is a random sequence of properties. Randomness is introduced using a polynomial scheme of $M$ trials with known probabilities $p(x_k) = p_k$ where $k = 1, ..., n$. After the implementation of the polynomial scheme trials, deterministic appearances of consequences are additionally embedded in the sequence, the causes of which appeared randomly in the sequence under consideration, and

the consequences increase the total length of the sequence $M$ by an amount equal to the number of appearances of causes. The constructed model will be called the information space.

First, let's consider the following schemes of cause-and-effect relationships.

1) The scheme "$a \rightarrow b$" is determined when property $a$ arises, then property $b$ arises deterministically.
2) The scheme "$a \rightarrow b \,\&\, a \rightarrow c$" occurs when property $a$ has several consequences of $b$ and $c$.
3) The scheme "$a \rightarrow b \,\&\, b \rightarrow c$" is defined as a sequence of causal relationships: $a$ entails the appearance of $b$ and $b$ entails the appearance of $c$. This scheme is also called transitive cause $a$ for effect $c$.

In scheme 1), the effect should appear after a fixed $t$ steps from the cause. *At* $t = 1$, the effect immediately follows the cause. In scheme 2) each consequence arises with its own interval $t_1$ or $t_2$.

In the future, we will change the models in accordance with approximations to real conditions.

## 3. Statistical identification of cause-and-effect relationships in a simple model

Under the conditions of the constructed model, it is possible to statistically determine cause-and-effect relationships. For the original polynomial scheme trials

$$p_{ij} = p(x_i, x_j) = p(x_i)p(x_j).$$

If $(x_i, x_j)$ are connected by a causal relationship, then after $t$ steps after $x_i$ the next property $x_j$ is embedded in the polynomial trial sequence.

Suppose that $M$ is large enough so that the relative frequencies of properties in a polynomial scheme trials with a sufficient degree of confidence uniquely identify the various unequal probabilities of this scheme. Let's denote $\nu(i)$ the frequency of occurrence $x_i$ in a polynomial sequence of outcomes of length $M$. If $(x_i, x_j)$ are causally related, then the length of the sequence $M^*$ increases from $M$ by $\nu(i)$, and the frequency of $x_j$ occurrence increases and will be equal to $\nu^*(j) = \nu(i) + \nu(j)$. Then

$$\frac{\nu^*(j)}{M} = \frac{\nu(i)}{M} + \frac{\nu(j)}{M} > p(x_j).$$

It follows first of all that the property $x_j$ is a consequence of some cause. If there is only one cause $x_i$ in the observed sequence, then

$$\frac{M^* - M}{M} = \frac{\nu(i)}{M},$$

which uniquely determines the cause $x_i$ in the event that the probability value $p(x_i)$ differs from any other probabilities of the polynomial scheme trials. If there are several such probabilities, then each occurrence of a cause in a random sequence deterministically entails the appearance of $x_j$ exactly after $t$ steps. The random repetition of this distance from other properties to $x_j$, having a probability equal to $p(x_j)$, decreases exponentially, which makes such a possibility unlikely with a sufficiently large occurrence of $\nu(i)$. In particular, the requirement of a priori knowledge of $t$ is used to simplify the scheme. It is sufficient to assume the existence of such a $t$ because, when potential variants of the cause are identified, the statistics of the distances between $(x_i, x_j)$ will dominate and uniquely determine $t$. Thus, the following theorem is proved.

**Theorem 1.** *With a well-known polynomial scheme trials of length M and assuming the existence of one causal relationship that generates an effect always through the same number of sequence steps, such a causal relationship is uniquely identified.*

It is possible that with known probabilities of a polynomial scheme trials, its true length $M$ is unknown. Let's consider this case. Let's calculate the frequency of occurrence of all outcomes of the observed sequence of properties $\{\nu(k), k = 1, ..., n\}$. According to the de Moivre–Laplace theorem [12] for each $\{\nu(k), k = 1, ..., n\}$ for large $M^*$ we have the following estimates

$$M_k = \frac{\nu(k)}{p(x_k)} = M + o(\ln M\sqrt{M}).$$

Then

$$\frac{\nu(i) + \nu(j)}{p(x_j)} = M + \frac{Mp(x_i)}{p(x_j)} + o(\ln M\sqrt{M}) > M_k,$$

which is the most of all $M_k$ for $k \neq j$. Thus, in this case, the consequence is identified. Using a fixed distance between $(x_i, x_j)$, as shown above, we establish the cause $x_i$.

Let's consider the simplest case of Scheme 2). In this case $x_i$, the cause has two effects $x_j$ and $x_k$, which appear at distances $t_1$ and $t_2$ from the cause. If all the probabilities of the polynomial scheme trials are known and the length of the polynomial sequence is $M$, then the length of the sequence with included consequences is $M^* = M + 2\nu(i)$. From here we find $\nu(i)$ and for sufficiently large $M$ we find $p(x_i)$. If there are no other similar probabilities, then we find the cause. If there are more probabilities $p(x_i)$, then the frequencies $\nu^*(j) = \nu(i) + \nu(j)$ and $\nu^*(k) = \nu(i) + \nu(k)$ allow us to reconstruct the effects and by the distances $t_1$ and $t_2$ we find the true cause. Here it is also sufficient to allow constant distances between cause and effect $x_i$. The values $t_1$ and $t_2$ are easily calculated from the repeatability of each cause-effect pair.

If $M$ is unknown, then the method proposed for Scheme 1) can be used to identify each consequence.

Scheme 3) is equivalent to the two simplest Schemes 1). In this case, there are two causes $x_i$ and $x_v$, each of which has a corresponding effect $x_j$ or $x_k$ with fixed distances $t_1$ and $t_2$ to the causes. For a known $M$, the real length $M^* = M + \nu(i) + \nu(v)$. For an increased frequency of $\nu^*(j) = \nu(i) + \nu(j)$ and $\nu^*(k) = \nu(v) + \nu(k)$ we are recovered effects in the investigation. We find the causes by the distances. Here it is also possible to abandon a priori knowledge of the values $t_1$ and $t_2$, but only assume the constancy of these values. In the case of a transitive cause differs in that if $(x_i, x_j, x_k)$ are connected by a transitive causal relationship, then $\nu^*(k) = \nu(i) + \nu(j) + \nu(k)$.

If the parameter $M$ is unknown, the effects of various causes can be determined using the method for Scheme 1). Then, using the distances $t_1$ and $t_2$, we find both causes.

Let's consider a case where there are two causes with the same effect. Let the effect $x_j$ be a common consequence of each of the two causes $x_i$ and $x_k$. Then, using the method described above, we find the consequence $x_j$ and the frequency of its occurrence $\nu^*(j) = \nu(i) + \nu(j) + \nu(k)$. If the distances $t_1$ and $t_2$ are known, then it is easy to find the causes. From each encountered $x_j$, we select two elements at distances $t_1$ and $t_2$. The set of selected elements will also include $(x_i, x_k)$ and random elements with polynomial probabilities. Let's calculate the frequency of occurrence of each of the encountered elements. It will be shown further that the two highest frequencies corresponding to the causes and the causes themselves are uniquely distinguished from them.

We will carry out the proof for the most difficult case, when $M$ is unknown and only the constraints $t_1 \leq t$ and $t_2 \leq t$ are known. From each encountered $x_j$, we select all the elements on the left at a distance $t$. The random component of occurrence $x_j$ is equal to $\nu(j)$. By the de Moivre–Laplace theorem

$$\nu(j) = Mp(x_j) + o(\ln M\sqrt{M}),$$

where the parameter $M$ is unknown, but

$$M \geq \frac{M^*}{3}.$$

An error in determining the causes may occur due to the accidental coincidence of conditions of the same distance from $x_j$ a large number of randomly appearing properties. The largest number of identical random elements does not exceed $\nu(j)$. Such a random sequence can be located at a distance from each $x_j$ at a distance $t^*$ with a probability of $1/t$. If this is a property $x_f$, then the probability of a random occurrence of a long sequence of length

$$M^{**} = \frac{\min\limits_{i,j,k}\{p(x_i), p(x_j), p(x_k)\}M}{2},$$

is estimated by the value

$$\left(\frac{p(x_f)}{t}\right)^{M^{**}},$$

which is significantly less than the probability estimates in the de Moivre–Laplace theorem. Given that $t$ and the number of properties are limited, we obtain the following theorem.

**Theorem 2.** *With a well-known polynomial scheme trials of length $M$ and assuming the existence of two causal relationships that generate a common consequence always through the same number of steps in the resulting sequence, such a causal relationship is uniquely identified. Moreover, it can be assumed that only $M^*$ is known, which includes the length of the random part of the sequence and the number of all the consequences of the desired causes.*

## 4. Causal relationships in several information spaces

Let's consider a finite set of information spaces (IS) [13], which we will denote as $\mathrm{IS}^* = \{\mathrm{IS}_1, ..., \mathrm{IS}_k\}$. Each IS has its own set of properties $A_i$. Randomness is determined using polynomial scheme from $M$ trials and with known probabilities, but the probabilities of polynomial schemes trials for different $\mathrm{IS}_i$ are different. If $a$ belongs to $\mathrm{IS}_1$, and $b$ belongs to $\mathrm{IS}_2$ and $a$ is the cause of the deterministic appearance of consequence $b$, then if there is a connection from $\mathrm{IS}_1$ to $\mathrm{IS}_2$ that allows initiating the consequence, property $b$ appears in the sequence of properties of $\mathrm{IS}_2$ at the same moment (the sequences in all IS are synchronized), and from this moment the remaining part of the sequence in $\mathrm{IS}_2$ is shifted by 1. These shifts do not interfere with finding synchronous pairs in $\mathrm{IS}_1$ and $\mathrm{IS}_2$, since the number of shifts each time is equal to the number of occurrences in $\mathrm{IS}_1$ of the cause candidate minus one.

If cause $a$ generates consequences in several different IS, then the scheme of the appearance of consequences $a$ is such as described above.

For simplicity, let's first consider the case of two spaces $\mathrm{IS}_1$ and $\mathrm{IS}_2$. The task is to statistically identify cause-and-effect relationships. Let $A_1 = \{x_1, ..., x_n\}$ – are properties in $\mathrm{IS}_1$, $A_2 = \{y_1, ..., y_m\}$ – are properties in $\mathrm{IS}_2$.

The probabilities of random synchronous pairs are equal to $p(x_i, y_j) = p(x_i)p(y_j)$ If $x_i$ is the cause of occurrence $y_j$, then the probability of occurrence of such a pair is equal to $p(x_i)$.

Let's denote the frequencies of the properties in $\mathrm{IS}_1$ by $\nu$, and the frequencies in $\mathrm{IS}_2$ by $\mu$. If $x_i$ is the cause of occurrence $y_j$, then the resulting frequency of the property

$$\mu^*(y_j) = \nu(x_i) + \mu(y_j).$$

Hence, as before, we conclude that $y_j$ is a consequence of the cause of $\mathrm{IS}_1$. The restoration of the cause is reduced to calculating the statistics of properties from $\mathrm{IS}_1$, which are located in $\mathrm{IS}_1$ at the moments when properties $y_j$ appear in the sequence of $\mathrm{IS}_2$.

By the de Moivre–Laplace theorem

$$\mu(y_j) = Mp(y_j) + o(\ln M\sqrt{M}),$$
$$\nu(x_i) = Mp(x_i) + o(\ln M\sqrt{M}).$$

Let $\hat{\nu}(x_k)$ be the frequency of occurrence of properties $x_k$ from $\mathrm{IS}_1$, which are located in $\mathrm{IS}_1$ at the moments when properties $y_j$ appear in the sequence of $\mathrm{IS}_2$. If the inequality

$$p(x_k)p(y_j) < p(x_i),$$

holds for any $k \neq i$, then the estimate

$$\hat{\nu}(x_k) = Mp(y_j)p(x_k) + o(\ln M\sqrt{M}) < \nu(i),$$

is valid for any $k \neq i$ and completely determines the cause for sufficiently large $M$.

**Theorem 3.** *With known polynomial scheme trials of length M in* $\mathrm{IS}_1$ *and* $\mathrm{IS}_2$ *and under the assumption*

$$p(x_k)p(y_j) < p(x_i), \quad k \neq i,$$

*if there is one connection by which a cause belonging to* $\mathrm{IS}_1$*, through this connection generates an effect in* $\mathrm{IS}_2$ *always simultaneously with the appearance of the cause, such a causal relationship is uniquely identified for sufficiently large M.*

If a given property in $\mathrm{IS}_1$ has several consequences located in different IS, then for large $M$ these cause-and-effect relationships are determined in the same way as it is done for two IS.

Similarly, an algorithm is built to identify a transitive causal relationship in Scheme 3), when causes and effects are located in different IS.

## 5. Complex causes

In practice, there is often a situation where the cause is complex. Let $(x_i, y_j, z_k)$ be properties that form causal relationships as follows. A property $x_i$ from $\mathrm{IS}_1$ is not the cause of an effect $z_k$ from $\mathrm{IS}_3$, and a property $y_j$ from $\mathrm{IS}_2$ is not the cause of an effect $z_k$ from $\mathrm{IS}_3$. However, if, with the simultaneous appearance of $x_i$,$y_j$ in $\mathrm{IS}_1$ and $\mathrm{IS}_2$ and the presence of their connection with $\mathrm{IS}_3$, an effect $z_k$ appears in $\mathrm{IS}_3$, then the cause constructed in this way is called complex.

The task is to identify the only complex causal relationship in $\mathrm{IS}^*$.

If $\mathrm{IS}_1 \times \mathrm{IS}_2$ is known, then the number of occurrences of the cause for the effect $z_k$ is distributed according to a polynomial law with probabilities $p(x_i)p(y_j)$.

If we denote the frequency of occurrence of a pair $(x_i, y_j)$ in the probability space $\mathrm{IS}_1 \times \mathrm{IS}_2$ as $\nu(x_i, y_j)$, then as a result of the occurrence of the cause for the effect $z_k$, the statistics $\nu(z_k)$ will increase by $\nu(x_i, y_j)$. From which it follows that $z_k$ is a consequence of any cause, it is established as before. If $\mathrm{IS}_1 \times \mathrm{IS}_2$ is known, then the cause is determined in the same way as for one information space. If $\mathrm{IS}_1 \times \mathrm{IS}_2$ is unknown, then it is necessary to identify the spaces in which the parts of the cause are located.

To identify spaces in which parts of the cause are present, you can use the procedure for disconnecting links between information spaces. At the first stage, it is necessary to create a sequence of properties in each IS according to a polynomial scheme trials of length $M$. At the second stage, for each individual IS that may have a connection with $\mathrm{IS}_3$, a procedure is used to turn on the connection between them and a frequency of $z_k$ change is considered. In the absence of a frequency change, the connection of the next IS with the $\mathrm{IS}_3$ is disabled and the same experiment is performed with the next IS.

If all IS that could have a connection with $IS_3$ did not lead to a change in the frequency of the property, then the pairs of included connections are also sorted out. In the case under consideration, the $IS_1 \times IS_2$ pair will give the desired frequency of $z_k$ change. This means that there is a complex reason for this pair. There are a couple of implementations of synchronous polynomial schemes in the $IS_1 \times IS_2$ space. Consequently, as noted above, they generate a new polynomial scheme $IS_4 = IS_1 \times IS_2$ of length $M$ with probabilities $p(x_i, y_j) = p(x_i)p(y_j)$. Using the methods described earlier, there is a complex cause.

## 6. Conclusion

The paper considers the simplest case of identifying deterministic cause-and-effect relationships in the presence of random properties that do not carry information about causes and effects.

To identify cause-and-effect relationships, information is used on the occurrence of causes and the time of occurrence of consequences after the appearance of causes. First, the consequences of some unknown causes are revealed, and then additional information allows you to identify the causes of the found consequences themselves.

This model roughly corresponds to the search for additional information in medical diagnostics and information security. The analysis of the above model allows you to find new ways to obtain additional information in search tasks.

In the future, it is supposed to investigate the model in conditions of striving for infinity of parameters $M$ and $n$. Another area of further research is to increase the size of causes and effects to sets containing several properties.

**Author Contributions:** Conceptualization, A. Grusho and M. Zabezhailo; methodology, K. Samouylov; validation, A. Grusho, M. Zabezhailo and N. Grusho; formal analysis, A. Grusho and E. Timonina; investigation, N. Grusho; data curation, A. Grusho and N. Grusho; writing—original draft preparation, E. Timonina; writing—review and editing, K. Samouylov. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Data sharing is not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Zhang, X., Hu, W. & Yang, F. Detection of Cause-Effect Relations Based on Information Granulation and Transfer Entropy. *Entropy* **24,** 212. doi:10.3390/e24020212 (2022).
2. Reimer, J., Wang, Y., Laridi, S., Urdich, J., Wilmsmeier, S. & Palmer, G. Identifying cause-and-effect relationships of manufacturing errors using sequence-to-sequence learning. *Scientific Reports* **12,** 22332. doi:10.1038/s41598-022-26534-y (2022).
3. Pearl, J. *Causal Inference* in *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008* (eds Guyon, I., Janzing, D. & Schölkopf, B.) **6** (PMLR, Whistler, Canada, Dec. 2010), 39–58.
4. Pearl, J. *The mathematics of causal inference* in *Joint Statistical Meetings Proceedings. ASA* (2013), 2515–2529.
5. Yao, L., Chu, Z., Li, S., Li, Y., Gao, J. & Zhang, A. A Survey on Causal Inference. *ACM Transactions on Knowledge Discovery from Data* **15,** 1–46. doi:10.1145/3444944 (2021).
6. Höfler, M. Causal inference based on counterfactuals. *BMC Medical Research Methodology* **5,** 1–28. doi:10.1186/1471-2288-5-28 (2005).

7. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications* **11,** 3923. doi:10.1038/s41467-020-17419-7 (2020).
8. Grusho, A. A., Grusho, N. A. & Timonina, E. E. Root Cause Anomaly Localization [Lokalizatsiya iskhodnoy prichiny anomalii]. *Information Security Problems. Computer Systems.* in Russian, 9–16 (2020).
9. Smirnov, D. V. Methodology of problem-oriented Big Data analysis in limited time mode [Metodika problemno-orientirovannogo analiza Big Data v rezhime ogranichennogo vremeni]. *International Journal of Open Information Technologies* **9.** in Russian, 88–94 (2021).
10. Grusho, A., Grusho, N., Zabezhailo, M. & Timonina, E. *Evaluation of Trust in Computer-Computed Results* in *Distributed Computer and Communication Network* (eds Vishnevskiy, V. M., Samouylov, K. E. & Kozyrev, D. V.) (Springer International Publishing, Cham, 2022), 420–432. doi:10.1007/978-3-030-97110-6_33.
11. Guyatt, G. *et al.* Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine. *JAMA* **268,** 2420–2425. doi:10.1001/jama.1992.03490170092032 (Nov. 1992).
12. Shiryaev, A. N. *Probability [Veroyatnost']* in Russian. 521 pp. (MTsNMO, Moscow, 2004).
13. Grusho, A., Grusho, N. & Timonina, E. *Method of Several Information Spaces for Identification of Anomalies* in *Intelligent Distributed Computing XIII* (eds Kotenko, I., Badica, C., Desnitsky, V., El Baz, D. & Ivanovic, M.) (Springer International Publishing, Cham, 2020), 515–520. doi:10.1007/978-3-030-32258-8_60.

## Information about the authors

**Grusho, Alexander A.**—Principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; professor of Department of Probability Theory and Cyber Security of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University), (e-mail: grusho@yandex.ru, ORCID: 0000-0003-4400-2158)

**Grusho, Nikolai A.**—Candidate of Physical and Mathematical Sciences, Senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (e-mail: info@itake.ru, ORCID: 0000-0002-5005-2744)

**Zabezhailo, Michael I.**—Professor, Doctor of Physical and Mathematical Sciences, Principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (e-mail: m.zabezhailo@yandex.ru, ORCID: 0000-0002-5067-5909)

**Samouylov, Konstantin E.**—Professor, Doctor of Technical Sciences, Head of the Department of Probability Theory and Cyber Security of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University) (e-mail: samuylov-ke@rudn.ru, ORCID: 0000-0002-6368-9680)

**Timonina, Elena E.**—Professor, Doctor of Technical Sciences, Leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; professor of Department of Probability Theory and Cyber Security of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University) (e-mail: eltimon@yandex.ru, ORCID: 0000-0002-6493-3622)

UDC 004.8
DOI: 10.22363/2658-4670-2024-32-2-213–221 EDN: CPUADE

# Статистический анализ причинно-следственных связей

А. А. Грушо[1, 2], Н. А. Грушо[1], М. И. Забежайло[1], К. Е. Самуйлов[2], Е. Е. Тимонина[1, 2]

[1] *Федеральный исследовательский центр «Информатика и управление» РАН, ул. Вавилова, д. 44, стр. 2, Москва, 119133, Российская Федерация*
[2] *Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация*

**Аннотация.** Рассмотрена проблема выявления детерминированных причинно-следственных связей, изначально скрытых в накопленных эмпирических данных. Для выявления таких связей использовались статистические методы. Предложена простая математическая модель причинно-следственных отношений, в рамках которой описано несколько моделей причинно-следственных связей в данных – для простейших отношений между причиной и следствием, для многих следствий одной причины, а также для цепей причинно-следственных связей (так называемых транзитивных причин). Сформулированы оценки, позволяющие с помощью теоремы Муавра–Лапласа определить параметры модели, связывающие события в испытаниях полиномиальной схемы. Обоснованы утверждения об однозначной идентификации причинно-следственных связей, которые восстанавливаются по накопленным данным. Обсуждаются возможности использования таких схем анализа данных в задачах медицинской диагностики и кибербезопасности.

**Ключевые слова:** задача конечной классификации, причинно-следственные связи, машинное обучение