On the statistics of anomalous clumps in random point images

Aleksandr L. Reznik[1], Aleksandr A. Soloviev[1] and Andrey V. Torgov[1]

[1] *Institute of Automation and Electrometry of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*

Abstract

New algorithms for calculating exact analytical formulas describing two related probabilities are proposed, substantiated and software implemented: 1) the probability of the formation of anomalously large local groups in a random point image; 2) the probability of the absence of significant local groupings in a random point image.

Keywords

Random point image, computer analysis, local groupings.

1. Introduction

In many scientific and technical disciplines, when solving applied problems related to digital image and signal processing, it becomes necessary to assess the degree of randomness of the analyzed image fragments, depending on the presence or absence of local groupings of point objects on them (as a result, the analyzed fragment significantly differs from ambient background). Such problems arise in many scientific and technical disciplines and can be both purely theoretical [1, 2] and purely applied [3]. For example, the presence of local clumps in processed aerospace images [4, 5] may indicate the presence of latent objects within the analyzed fragment that require more detailed study. In computer processing of biomedical images, one of the most important moments of the preliminary processing stage is the search for abnormal heterogeneities and thickenings, which may be evidence of various disease-causing abnormalities that require priority attention [6, 7]. In correlation rhythmography, a method is known that makes it possible to construct prognostic assessments of the possibility of restoration of sinus rhythm by a set of intervals on the cardiogram that form an autoregressive cloud (scatterogram) [8, 9].

Mathematically similar problems arise when studying the process of registering random point fields using a scanning aperture with a limited number of threshold levels. When fixing random coordinates of small-sized (ideally, point) objects that form such a field, a failure occurs at the moment when the number of signal point objects located within the scanning aperture exceeds the specified threshold level. It is shown in [10] that in cases where the analyzed image is formed by a random Poisson flux of constant intensity, the two-dimensional problem of

SDM-2021: All-Russian conference, August 24–27, 2021, Novosibirsk, Russia

reznik@iae.nsk.su (A. L. Reznik); soloviev@iae.nsk.su (A. A. Soloviev); torgov@iae.nsk.su (A. V. Torgov)

CEUR Workshop Proceedings (CEUR-WS.org)

estimating the probability that the registration process will be carried out without failures is reduced (this is achieved by means of standard factorization) to the following one-dimensional problem:

"It is required to find the probability of the event that if n points are randomly dropped on the interval $(0, 1)$, not a single grouping will be formed, located in a certain subinterval $\Omega_\varepsilon \subset (0, 1)$, having a length ε and including more than k points".

When solving applied problems related to the detection of abnormally large clumps in the analyzed images (as, for example, in the above-mentioned problems related to the processing and analysis of aerospace or biomedical images), knowledge of probability formulas $P_{n,k}(\varepsilon)$ is required in which the value of the integer parameter k is as close as possible to the value n. In such cases, it becomes necessary to know the exact analytical dependencies $P_{n,n-1}(\varepsilon)$, $P_{n,n-2}(\varepsilon)$, $P_{n,n-3}(\varepsilon)$, etc. In particular, the probability $P_{n,n-1}(\varepsilon)$ corresponds to the fact that if n points are randomly thrown over the interval $(0, 1)$, all of them will not "merge" into one ε-grouping, the probability $P_{n,n-2}(\varepsilon)$ is that no ε-grouping with a size larger than $n - 2$ will be formed, the probability $P_{n,n-3}(\varepsilon)$ — that no ε-groupings larger than $n - 3$ will be formed, etc.

On the other hand, in a number of applied problems, on the contrary, it is required to know the exact analytical relations for the probabilities when the value of the threshold parameter k is minimal, i.e. when $k = 1, 2, 3$, etc. Such formulas are needed in cases when, by the nature of the research, it is required to estimate the probability that, within the studied interval, the distribution of n random point markers-objects is such that the number of any ε-group does not exceed the threshold level $k = 1, 2, 3$, etc. The purpose of this work is to propose analytical methods and software algorithms for finding exact probabilistic formulas $P_{n,k}(\varepsilon)$ for both the maximum (that is, as close as possible to n) and the minimum values of the threshold parameter k.

2. Obtaining particular solutions to a problem using computer analytics programs

The simplicity of the problem posed in the introduction is illusory, and its analytical solution is known only for the simplest case $k = 1$ [11, 12]:

$$P_{n,l}(\varepsilon) = (1 - (n-1)\varepsilon)^n, \quad 0 \le \varepsilon \le \frac{1}{n-1}. \tag{1}$$

Formula (1) describes the probability of an event that if n points are randomly dropped onto the interval $(0, 1)$, not a single ε-group will be formed containing at least 2 points, that all the ejected points will be located between themselves at a distance exceeding ε. The classical way to obtain solution (1) is to represent the desired probability in the form of an easily integrable iterated integral [11]:

$$P_{n,1}(\varepsilon) = n! \int\limits_{(n-1)\varepsilon}^{1} dx_n \left\{ \int\limits_{(n-2)\varepsilon}^{x_n-\varepsilon} dx_{n-1} \ldots \left\{ \int\limits_{2\varepsilon}^{x_4-\varepsilon} dx_3 \left\{ \int\limits_{\varepsilon}^{x_3-\varepsilon} dx_2 \left\{ \int\limits_{0}^{x_2-\varepsilon} dx_1 \right\} \right\} \right\} \right\}.$$

Solution (1) can be obtained in a different way. For example, in [10], a simple probabilistic-geometric method was proposed that allows one to calculate the probability (1) without resorting to the procedure of multidimensional integration. Thus, it is not difficult to find a solution to the main problem when $k = 1$. But for $k > 1$ the problem becomes much more complicated. Here, our efforts have led to the results that will be given below.

First, note that for arbitrary fixed values of n and k, the desired solution can be represented in the form of an n-fold integral:

$$P_{n,k}(\varepsilon) = n! \underset{D_{n,k}(\varepsilon)}{\int \cdots \int} dx_1 \ldots dx_n, \tag{2}$$

where the domain of integration $D_{n,k}(\varepsilon)$ is given by the system of linear inequalities

$$\begin{cases} 0 < x_1 < x_2 < \cdots < x_{n-1} < x_n < 1, \\ x_{k+1} - x_1 > \varepsilon, \\ x_{k+2} - x_2 > \varepsilon, \\ \vdots \\ x_n - x_{n-k} > \varepsilon. \end{cases}$$

To calculate integral (2), we developed a method of successive dimensionality reduction based on the step-by-step replacement of the initial n-fold integral with a set of structurally similar, but having dimension one less than the iterated integrals. Further, formalizing this method and applying cyclic recursion, two systems for analytical calculation of probabilities were designed and implemented as a computer software in order to calculate the desired piecewise-polynomial dependence in the form of functions of the continuous parameter ε. One system calculates the limits of integration for each of the iterated integrals into which the original n-fold integral (2) decomposes; the second software system is based on multiple differentiation of the integral (2) with respect to the parameter ε. In addition to the two mentioned software systems, a third algorithmic scheme was also developed and implemented as software, using a discrete-combinatorial model to calculate probabilistic formulas $P_{n,k}(\varepsilon)$.

Analytical calculations performed using the listed software systems made it possible to find a complete set of partial formulas $P_{n,k}(\varepsilon)$ in all ranges of variation of the continuous parameter ε for all values of integer parameters n and k up to $n = 14$. Note that their calculation is associated with a large amount of routine operations on setting the limits of integration, checking intermediate systems of inequalities for consistency, and performing direct integration in n-dimensional space, which is almost impossible to do "manually" even for $n = 4$. Therefore, all the necessary software calculations were carried out using the parallel computing algorithms the use of high-performance computing clusters [13].

2.1. Algorithms for software and analytical calculation of probabilistic formulas $P_{n,k}(\varepsilon)$

At the next stage, we tried, using the analysis of the obtained partial results, to establish and, if possible, reveal the general laws governing the formation of probability formulas for the case $k > 1$. And several of these analytic patterns were indeed discovered and subsequently rigorously proved. First, for $k = n - 1$, a simple dependence was traced and later prove. First, for $k = n - 1$, a simple dependence was traced and later proved

$$P_{n,n-1}(\varepsilon) = 1 - n\varepsilon^{n-1} + (n-1)\varepsilon^{n}. \tag{3}$$

(It should be recalled that formula (3) describes the probability that if n points are randomly dropped onto the interval $(0, 1)$, they will not all "merge" into one compact ε-grouping.)

For $k = n - 2$, the relationship $P_{n,k}(\varepsilon)$ is more complex:

$$P_{n,n-2}(\varepsilon) = \begin{cases} 1 - 2C_n^2\varepsilon^{n-2}(1-\varepsilon)^2 - 2\varepsilon^n, & 0 \le \varepsilon \le \dfrac{1}{2}; \\ 1 - 2\varepsilon^n + (2\varepsilon - 1)^n - 2C_n^2\varepsilon^{n-2}(1-\varepsilon)^2, & \dfrac{1}{2} \le \varepsilon \le 1. \end{cases} \tag{4}$$

For $k = n - 3$, the dependence $P_{n,k}(\varepsilon)$ becomes so complicated that its reconstruction by analyzing particular software solutions is a completely independent and difficult task:

$$\underset{(n>6)}{P_{n,\,n-3}(\varepsilon)} = \begin{cases} 1 - 2\varepsilon^n + C_n^1(6\varepsilon^n - 4\varepsilon^{n-1}) + C_n^2(-3\varepsilon^n + \varepsilon^{n-2}) + \\ +C_n^3(9\varepsilon^n - 18\varepsilon^{n-1} + 12\varepsilon^{n-2} - 3\varepsilon^{n-3}), & 0 \le \varepsilon \le \dfrac{1}{2}; \\ 1 - 2\varepsilon^n + (2\varepsilon-1)^n + C_n^1(1-\varepsilon)(-2\varepsilon^{n-1} + 2(2\varepsilon-1)^{n-1}) + \\ +C_n^2(1-\varepsilon)^2(\varepsilon^{n-2} + (2\varepsilon-1)^{n-2}) - 3C_n^3\varepsilon^{n-3}(1-\varepsilon)^3, & \dfrac{1}{2} \le \varepsilon \le 1. \end{cases} \tag{5}$$

Formulas (3)–(5) are confirmed both by software calculations and by direct analytical integration.

2.2. Formulas $P_{n,k}(\varepsilon)$ found using software, analytical and discrete-combinatorial algorithms

The purpose of developing discrete-combinatorial methods for calculating formulas $P_{n,k}(\varepsilon)$ is that they can be used to try to find a general solution for $k = 2$ by analogy with formula (1), which is valid for $k = 1$. Unfortunately, this task turned out to be much more difficult than it seemed before the start of the research. This is primarily due to the fact that, in contrast to the case $k = 1$, the probability $P_{n,k}(\varepsilon)$ consists of several piecewise-homogeneous fragments, continuously joined at the points of "connection". Secondly, the formula itself changes depending on the parity of n. Thirdly, finding patterns in each of the parameter ε ranges variation requires the creation of an individual scheme for transferring each continuous problem corresponding to a given specific range to its own very complex discrete-probabilistic problem.

In our proposed reduction scheme, generalized Catalan numbers appear in all subproblems (i.e., in all ranges of variation of the parameter ε). Knowing their explicit form is required when ordering interdependent random number sequences. Most of these probabilistic-combinatorial problems turned out to be more convenient to fomulate and solve in a "word-linguistic" form. In a number of cases, it was possible to use the technique of finding monotonic paths in Weyl chambers [14].

In the case $k \ge 2$ for the probabilities $P_{n,k}(\varepsilon)$, we could not find a general compact analytical relation similar to formula (1) for $k = 1$. However, using all the above computer and discrete-combinatorial tools, including software-analytical calculations and generalized Catalan numbers, we have established and then proved a number of particular previously unknown analytical dependencies. In particular, for $\varepsilon \to 0$, an asymptotic formula, common for arbitrary n, was established:

$$\begin{aligned} P_{n,2}(\varepsilon) = {} & C_n^0 + C_n^2(-n+2)\varepsilon^2 + C_n^3(4n-10)\varepsilon^3 + C_n^4(3n^2-37n+86)\varepsilon^4 + \\ & + C_n^5(-40n^2+394n-922)\varepsilon^5 + C_n^6(-15n^3+625n^2-5171n-12086)\varepsilon^6 + \\ & + C_n^7(420n^3-10724n^2+79996n-187002)\varepsilon^7 + \\ & + C_n^8(105n^4-10570n^3+205499n^2-1426841n+3336406)\varepsilon^8 + \\ & + C_n^9(5040n^4-155708n^3+2267664n^2-17317506n+52315558)\varepsilon^9 + \\ & + C_n^{10}(-945n^5+189000n^4-15794625n^3+389687181n^2-3798029823n + \\ & + 12998966646)\varepsilon^{10} + o(\varepsilon^{10}). \end{aligned} \tag{6}$$

For even values of $n = 2m$ on the segment $1/m < \varepsilon < 1/(m-1)$, the previously stated hypothesis formula is rigorously proved

$$P_{2m,2}(\varepsilon) = \frac{1}{m+1} C_{2m}^m (1-(m-1)\varepsilon)^{2m}.$$

For even values of $n = 2m$ on the segment $1/(m+1) < \varepsilon < 1/m$, the formula is established

$$\begin{aligned} P_{2m,2}(\varepsilon) = {} & C_{2m}^m (1-(m-1)\varepsilon)^{2m} - C_{2m}^{m-1}(1-(m-1)\varepsilon)^{2m} - \\ & - C_{2m}^{m-2}(1-m\varepsilon)^{m+2}(1-(m-2)\varepsilon)^{m-2} + \\ & + 2C_{2m}^{m-3}(1-m\varepsilon)^{m+3}(1-(m-2)\varepsilon)^{m-3} - \\ & - C_{2m}^{m-4}(1-m\varepsilon)^{m+4}(1-(m-2)\varepsilon)^{m-4}. \end{aligned}$$

For odd values $n = 2m+1$ on the segment $1/(m+1) < \varepsilon < 1/m$, the formula is established

$$\begin{aligned} P_{2m+1,2}(\varepsilon) = {} & C_{2m+1}^{m+1}(1-m\varepsilon)^{m+1}(1-(m-1)\varepsilon)^m - \\ & - 2C_{2m+1}^{m+2}(1-m\varepsilon)^{m+2}(1-(m-1)\varepsilon)^{m-1} + \\ & + C_{2m+1}^{m+3}(1-m\varepsilon)^{m+3}(1-(m-1)\varepsilon)^{m-2}. \end{aligned}$$

3. Conclusion

The results presented in this paper were obtained with the help of specially created instruments of machine analytics, as well as with the use of generalized Catalan numbers, which made it possible to transfer the inherently continuous problem of finding probabilistic formulas to the category of discrete-combinatorial ones. The efficiency of the proposed discrete-combinatorial methods allows us to hope for further progress in solving the described “continuous” problem, up to finding a general analytical formula for arbitrary values of the integer parameters n and k in all variation ranges of the continuous parameter ε. The presence of such a generalized analytical solution would provide researchers with an additional tool for assessing whether the analyzed point image is random or regular.

Acknowledgments

This work was supported in part by the Russian Foundation for Basic Research (project No. 19-01-00128), and Ministry of Science and Higher Education of the Russian Federation (project No. 121022000116-0).

References

[1] Shannon C.E. A mathematical theory of communication // The Bell System Technical Journal. 1948. Vol. 27. Is. 3. P. 379–423.

[2] Gnedenko B.V., Belyayev Y.K., Solovyev A.D. Mathematical methods of reliability theory. New York: Academic press, 1969. 518 p.

[3] Birger I.A. Technical diagnostic. Moscow: Mashinostroenie, 1978. 240 p. (In Russ.)

[4] Gromilin G.I., Kosykh V.P., Popov S.A., Streltsov V.A. Suppression of the background with drastic brightness jumps in a sequence of images of dynamic small-size objects // Optoelectronics, Instrumentation and Data Processing. 2019. Vol. 55. No. 3. P. 213–221.

[5] Reznik A.L., Tuzikov A.V., Soloviev A.A., Torgov A.V., Kovalev V.A. Time-optimal algorithms focused on the search for random pulsed-point sources // Computer Optics. 2019. Vol. 43. No. 4. P. 605–610.

[6] Ablameiko S.V., Anischenko V.V., Lapitsky V.A., Tuzikov A.V. Medical information technologies and systems. Minsk: OIPI NAS Belarus, 2007. 176 p.

[7] Wójcik W., Pavlov S., Kalimoldayev M. Information technology in medical diagnostics. London: CRC Press, 2019. 336 p.

[8] Poggio T., Girosi F. Networks for approximation and learning // Proceedings of the IEEE. 1990. Vol. 78. P. 1481–1497.

[9] Stinton P., Tinker I., Vickery I.C., Yahe S.P. The scatterogram. A new method for continuous electrocardiographic monitoring // Cardiovascular Research. 1972. Vol. 6. P. 598–604.

[10] Reznik A.L., Efimov V.M., Solov'ev A.A., Torgov A.V. Reliability of readout of random point fields with a limited number of threshold levels of the scanning aperture // Optoelectronics, Instrumentation and Data Processing. 2014. Vol. 50. No. 6. P. 582–588.

[11] Parzen E. Modern probability theory and its applications. New York; London: John Wiley & Sons Inc., 1960. 464 p.

[12] Wilks S. Mathematical statistics. Princeton: Princeton University Press, 1944. 284 p.

[13] Reznik A.L., Tuzikov A.V., Soloview A.A., Torgov A.V. Intelligent software support for analysis of random digital images // Computational Technologies. 2018. Vol. 23. No. 5. P. 70–81.

[14] Gessel I.M., Zeilberger D. Random walk in a Weyl chamber // Proceedings of the AMS. 1992. Vol. 115. No. 1. P. 27–31.