Scientific Visualization, 2025, volume 17, number 5, pages 38 - 51, DOI: 10.26583/sv.17.5.05

Analysis of Data Independence Using Nearest Neighbor Graphs

A.A. Kislitsyn[1]

Keldysh Institute of Applied Mathematics RAS

[1] ORCID: 0000-0003-2388-0496, kislitsyn@kiam.ru

Abstract

The article describes a new method for testing the independence of random data sets. This method uses representation of connections between data points in the form of nearest neighbor graphs and compares parameters of the resulting specific graph — such as the number of connected components and vertex degree distribution — to numerically derived critical statistics for random nearest neighbor graphs obtained by the author. The proposed method can be applied to various practical situations. It can be used to test the independence of random vectors in low-dimensional metric spaces. Such problems arise when analyzing data from physical measurements. Additionally, this approach is applicable for analyzing random points in high-dimensional spaces where direct numerical evaluations require exhaustive enumeration and are therefore impractical or impossible to obtain exactly. This problem relates to object classification characterized by many parameters. Moreover, proximity function between points may not necessarily be symmetric, which allows application of graph methods even in such cases. Alongside the task of sample testing, one could also consider comparing pseudo-random number generators by benchmarking structural graph statistics based on them. Considered probabilities of graph structure realization provide an independent set of criteria. For example, the number of fragments in some random graph might be typical for independent random variables while vertex degree distributions could differ significantly. This extends the applicability domain of statistical analysis. The paper presents a collection of model examples illustrating how the methodology works with respect to several types of practical scenarios mentioned above. A comparison of this method with other statistical approaches is provided. We emphasize that using graphs as visualization tools enables immediate identification of dependent elements within samples if there exist cluster centers represented by vertices with anomalously large degrees.

Keywords: nearest neighbor graphs, statistics, data independence, PRNG, pi-number.

1. Introduction

Analysis of large datasets and high-intensity event streams is gaining significant practical importance at the present stage, due to the computerization of virtually all areas of life. In practice, such analysis is usually carried out using methods developed within the theory of stationary random processes. At the same time, actual nonstationarity is taken into account within the framework of heuristic adaptive methods or within so-called cointegrated models of random processes. It should be noted that most machine learning methods rely on fixed, selected datasets, which therefore amount to learning on stationary distributions. In the context of analyzing high-intensity data streams, these methods tend to overestimate the accuracy of results and are therefore not particularly effective. The methodological problem lies in the fact that if data series have a stationary distribution, then it is sufficient to analyze a single sample of the required size in order to formulate a statistical criterion directly in terms of measurable quantities. If, however, the distribution is nonstationary, then, according to the definition of a nonstationary random process, one should study not the data themselves, but distribution functions constructed from non-overlapping samples, which requires substan-

tially more computational resources. For such sets of distributions, one computes distances between them — typically in the norm of summable functions or in the norm of continuous functions — after which the distribution of these distances is analyzed. In other words, homogeneity criteria for the data must be formulated not in terms of the elements of the series, but in terms of distances between sample distributions. Such analysis requires the use of methods of kinetic theory, i.e. the theory of time-varying distribution functions generated by measurable parameters of a complex system. However, the computational complexity of the corresponding procedures makes it infeasible to process large arrays of multidimensional data. Consequently, it is desirable to have a certain benchmark of statistics analogous to those for stationary series, which, however, is difficult to obtain for arbitrary nonstationary time series.

Nearest neighbor graphs (NNGs) make it possible to partially overcome this difficulty, since it has been established that for stationary distributions of distances between elements of a point set in a metric space, such statistical characteristics of the graph as its distribution by number of connected components and the degree distribution of vertices do not depend on the specific form of the distance distribution. Thus, computational experiments for determining the statistics of random NNG structures provide an opportunity to construct a new nonparametric criterion of independence for sample data. A corresponding benchmark was constructed by the author in [1, 2]. It appears important to demonstrate the methodology of data independence analysis using NNGs on a number of practical tasks, where the developed method proves to be highly effective compared to traditional approaches that employ criteria in terms of sample moments computed from observed data and similar statistics.

It should be noted that computational experiments aimed at finding percentiles of statistical criteria had been conducted before. For example, the Dickey–Fuller unit root test in autoregressive models [3] is carried out by numerical simulation of the Wiener process using the Monte Carlo method. The results of such numerical experiments are valid because they rely on proven statements about the mathematical properties of the simulated process.

The theory of random graphs, developed in the context of machine learning methods, originated in the 1950s thanks to the pioneering work of Gilbert [4], and later Erdős and Rényi [5], who laid its foundations. Since then, the theory has undergone significant development. The study of various aspects of graph theory — particularly geometric and topological — and the emergence of numerous corollaries have led to its application in a wide range of scientific disciplines, including telecommunications [6], astronomy [7], statistical physics, the social and biological sciences, and machine learning [8].

At present, the main theoretical results in random graph theory are collected in V.F. Kolchin's monograph [9], as well as in the course [10] of the School of Mathematics, University of Birmingham. Their practical applications are associated with the involvement of external scientific disciplines. One current direction that requires further development is the use of random graphs and their connection to pseudorandom number generators. Sample statistical analysis of random graphs is challenging because a "sample of graphs" does not exist as such, since a graph merely visualizes some property of the studied group of objects. Thus, the concept of sampling applies to the group of objects themselves, which must be homogeneous with respect to the studied property in a statistical sense. But since the property under investigation is revealed precisely as a result of analyzing the graph structure, it is impossible to pre-form groups of objects with the desired property. In this case, one can proceed as follows. By studying the distribution of parameters that are considered key to describing a given system model, one can replicate samples with the empirically obtained distribution law of these parameters, and then study the statistics of graphs corresponding (under the model's assumption) to such a family of systems. Since the parameter sample is finite, the corresponding empirical distribution function fluctuates from sample to sample. Therefore, one should examine the variability of the structure of the family of graphs as the parameter distribution of the system changes. It is worth noting that similar issues, albeit in a somewhat different formulation, were considered in the work of V.A. Kalyagin and co-authors [11], where graphs

were used as a tool for detecting relationships between elements of network structures with randomly scattered parameters.

The nearest-neighbor graph is a convenient tool for qualitative analysis of sample data dependence. Along with numerical estimates of vertex statistics and connectivity components, visualization immediately reveals the presence of dependence, since for independent data the sum of frequencies of vertices with degrees from zero to six accounts for more than 99% of the normalization. Therefore, the appearance of a vertex of degree, say, 13 in a graph with 10,000 vertices has a probability of less than 10^{-7}. The paper will present examples demonstrating the effectiveness of visualization for obtaining quick qualitative assessments.

2. Key properties of nearest neighbor graph structure statistics

We assume that pairwise distances between points are distinct, meaning every point has a unique nearest neighbor. Under this assumption, Nearest Neighbor Graphs possess the following main properties:

- Each vertex has exactly one outgoing edge, hence no isolated vertices appear in NNGs.
- The number of vertices equals the number of edges, stemming from the uniqueness of the nearest neighbor for any order.
- The sum of vertex degrees over incoming edges equals the total number of vertices.
- Every connected component of an NNG contains precisely one cycle of length two.

In reference [1], a series of new results concerning NNG structures have been proved. The primary statements utilized in designing algorithms for generating NNGs and collecting statistics include:

Theorem 1. Probabilities of realizing NNG structures are independent of the distribution function governing intervertex distances.

Proof relies on monotonic mapping linking uniform and arbitrary distributions, preserving the ordering of sample elements interpreted as distances rather than coordinates.

Theorem 2. Realization probabilities of NNG structures remain unaffected by whether triangle inequality holds for matrix elements treated as distance matrices.

Proof constructs an algorithm generating random matrices satisfying triangle inequality for any three elements. Furthermore, it demonstrates that no coordinate distribution exists in R^n, where the distribution of distances between points is uniform. Therefore, transforming the element distribution of the constructed matrix to a uniform distribution will inevitably violate the triangle inequality while maintaining the original ordering among elements.

Consequently, these assertions imply the possibility of constructing a benchmark for statistics of NNG structures based on uniformly distributed random matrices.

In reference [12], distributions of NNGs by numbers of connected components and vertex degrees were built for random uniform distributions of point coordinates in low-dimensional spaces $n = 1 \div 7$. It turned out that as dimensionality increases, the probability distributions of NNG structures converge towards those obtained from generating elements of random matrices irrespective of whether these elements represent actual distances or simply positive random values. Thus, we can numerically derive the asymptotic behavior of vertex degree distributions in NNGs for high-dimensional spaces, e.g., when $n > 10^4$, without actually conducting computationally intractable numerical experiments involving random coordinate generation and subsequent construction of corresponding distance matrices.

To construct a benchmark for NNG statistics, simulations of random graphs with varying numbers of vertices *N* from 100 to 1500 in steps of 50 were performed. Statistics were collected from 1 million realizations of symmetric random matrices for each size. The graph was constructed according to the following procedure: in each row of the matrix, the minimal nonzero entry was identified. The column index containing this minimum corresponded to the vertex closest to the vertex indexed by the current row. It was discovered that the distribution of first-neighbor graphs by the number of connected components closely approxi-

mates a Gaussian distribution. The mode of this distribution occurs at the point $[N/4]$. The range of the distribution is $N/2$, and the variance turned out to be approximately $N/16$, so the distribution density is approximated by the function

$$G_N(Q) = \frac{2\sqrt{2}}{\sqrt{\pi N}\,erf\left(\sqrt{N/2}\right)} \exp\left(-8N\left(\frac{Q}{N} - \frac{1}{4}\right)^2\right), \quad \frac{Q}{N} \in (0; 1/2] \tag{1}$$

In particular, for $N = 250$, the probabilities of NNG realizations with a specific ratio Q/N are presented in Table 1.

Table 1. Distribution of NNG by the number of connected components, $N = 250$

Q/N	Frequency	Q/N	Frequency	Q/N	Frequency
0,175	< 0,00001	0,225	0,03614	0,275	0,03614
0,180	0,00001	0,230	0,05669	0,280	0,02085
0,185	0,00003	0,235	0,08044	0,285	0,01089
0,190	0,00009	0,240	0,10329	0,290	0,00514
0,195	0,00030	0,245	0,12000	0,295	0,00220
0,200	0,00085	0,250	0,12616	0,300	0,00085
0,205	0,00220	0,255	0,12000	0,305	0,00030
0,210	0,00514	0,260	0,10329	0,310	0,00009
0,215	0,01089	0,265	0,08044	0,315	0,00003
0,220	0,02085	0,270	0,05669	0,320	0,00001

If, instead of a random matrix, we directly generate N points with uniformly distributed coordinates in an n-dimensional cube, the resulting distribution of the NNGs by the number of fragments will be the same as distribution (1), but shifted to the right. The mode of distribution (1), which we will call "dimensionless," is located at the point $Q/N = 0.25$. This is an exact result derived from the properties of Young diagrams for representing the number of ways a natural number can be partitioned into a sum of natural numbers. For distributions based on the number of connected components of the NNG in spaces of small dimensions, the numerically determined positions of the probability maxima are given in Table 2.

Table 2. Most likely values of the number of connected components of NNG in n-dimensional space

n	1	2	3	4	5	6	7
argmax(G)	0,340	0,310	0,290	0,275	0,265	0,260	0,255

Thus, the distribution by the number of fragments for a 7-th dimensional space turns out to be very close to the "dimensionless" situation. As the dimensionality increases further, these distributions coincide with high accuracy. Therefore, in practice, for high-dimensional random vectors, the "dimensionless" benchmark, which can be obtained with relatively low computational costs, can be used.

The empirical joint distribution of the number of vertices of the NNG by the degrees $\{0, 1, ..., p\}$ of the incoming edges is formally described by a polynomial distribution, which, for large dimensions and a large number of vertices, asymptotically becomes a multivariate normal distribution:

$$f\left(n_0, n_1, ..., n_p\right) = \frac{1}{\sqrt{\det C}}\left(\frac{N}{\pi}\right)^{(p+1)/2} \exp\left(-N\sum_{i,j=0}^{p} C_{ij}^{-1}\left(n_i/N - \mu_i\right)\left(n_j/N - \mu_j\right)\right), \tag{2}$$

where $\det C$ is the determinant of the covariance matrix, the elements of which are normalized by the value $N/2$, C_{ij}^{-1} – are the elements of the inverse covariance matrix, n_j – is the number of vertices with degree j, μ_j – is the average fraction of vertices with degree j. Based on the results from 1 million graph generations, the following values of μ_j, independent of N were obtained:

$$\mu_0 = 0,367;\ \mu_1 = 0,367;\ \mu_2 = 0,184;\ \mu_3 = 0,061;$$
$$\mu_4 = 0,015;\ \mu_5 = 0,003;\ \mu_6 = 0,0005. \quad (3)$$

The first four frequencies have a relative accuracy of 0.002, for the μ_4 accuracy is 0.01, for the μ_5 it is 0.03, and for the μ_6, we have a relative accuracy of 0.1. These values (3) are the same for both symmetric matrices (distances) and arbitrary random matrices with a zero main diagonal (quasi-distances). The values presented in (3) are obtained for the first time. They are universal for any stationary distributions of distances between random points. It is noteworthy that the sum of the frequencies of the first seven degrees in (3) equals 0.9975, i.e., the probability of finding a vertex with a degree greater than 6 in a random graph is less than 0.003. Therefore, the presence of such vertices would indicate a non-random nature of the distances between the elements of the set.

Due to the different levels of significance in defining empirical frequencies (3), using a multivariate distribution to test data for independence is not convenient. To test the corresponding statistical hypothesis, it is sufficient to consider the distribution of vertices with zero degree. This distribution, at a significance level of 0.01, is approximated by a normal distribution of the form:

$$f_0(n) = \sqrt{\frac{A_0}{\pi N}} \exp\left(-\frac{A_0}{N}\left(n - \mu_0 N\right)^2 \right), \quad A_0 = 4,9,\ \mu_0 = 0,367. \quad (4)$$

An example of the benchmark probability of dependence of quantities based on the statistics of the number of vertices with zero degree for $N = 250$ is presented in Table 3. In this table, the probability of deviation of the quantity $\nu = |n/N - \mu_0|$ from zero equal $erf\left(\nu\sqrt{AN}\right)$. Value $1 - erf\left(\nu\sqrt{AN}\right)$ is an estimate of the probability that the data under consideration are statistically independent.

Table 3. Probability of dependency between variables based on the deviation ν, $N = 250$

ν	P	ν	P	ν	P
0,001	0,040	0,020	0,684	0,039	0,949
0,002	0,080	0,021	0,707	0,040	0,954
0,003	0,120	0,022	0,729	0,041	0,959
0,004	0,160	0,023	0,750	0,042	0,964
0,005	0,199	0,024	0,770	0,043	0,968
0,006	0,237	0,025	0,789	0,044	0,972
0,007	0,275	0,026	0,807	0,045	0,975
0,008	0,313	0,027	0,823	0,046	0,978
0,009	0,349	0,028	0,839	0,047	0,981
0,010	0,385	0,029	0,853	0,048	0,983
0,011	0,419	0,030	0,866	0,049	0,986
0,012	0,453	0,031	0,879	0,050	0,987
0,013	0,486	0,032	0,890	0,051	0,989
0,014	0,518	0,033	0,901	0,052	0,991
0,015	0,548	0,034	0,911	0,053	0,992
0,016	0,578	0,035	0,920	0,054	0,993
0,017	0,606	0,036	0,928	0,055	0,994
0,018	0,633	0,037	0,935	0,056	0,995
0,019	0,659	0,038	0,942	0,057	0,996

It is evident that even a very small deviation from the mean leads to the conclusion that the quantities are dependent.

In spaces of small dimensions, the distribution of vertices by degree remains the same, only the maxima of the distributions (i.e., the mean frequencies) shift slightly. The corresponding values are presented in Table 4.

Table 4. Mean values of the frequencies of the main vertex degrees of the NNGs in n-dimensional space.

n	1	2	3	4	5	6	7
$\mu_0(n)$	0,421	0,392	0,382	0,375	0,369	0,368	0,367
$\mu_1(n)$	0,449	0,427	0,418	0,400	0,375	0,369	0,368
$\mu_2(n)$	0,130	0,143	0,148	0,166	0,178	0,181	0,183

Next, practical examples of applying the constructed benchmark to various situations related to data independence analysis will be considered.

3. Random data in a low-dimensional space

Consider an example that presents significant methodological interest. It is required to determine at what confidence level it can be assumed that the distances between the capitals of the world's countries form an independent sequence of random variables. The coordinates (latitude and longitude) of the world's capitals are known: they can be found, for example, on the website [www.flagpictures.com], which lists the latitude ϕ and longitude θ for the capitals of 251 countries. Assuming the Earth's surface is approximately a sphere with a radius of R = 6371 km (average radius), the distances between points on the sphere can be calculated using the formula:

$$d_{ij} = 2R\sqrt{\sin^2\left(\frac{\phi_i - \phi_j}{2}\right) + \cos\phi_i \cos\phi_j \sin^2\left(\frac{\theta_i - \theta_j}{2}\right)},$$

After that, the corresponding distance matrix is constructed. As a result, we obtain about 30,000 data points.

Various homogeneity tests can be applied to the obtained values. For example, by arbitrarily splitting the data into two equal parts, one can use the Kolmogorov-Smirnov test and conclude that both parts are drawn from the same distribution with a probability of 0.95. Therefore, if the data are independent, one would naturally expect typical values for the parameters of the structure of a random nearest-neighbor graph, which, in this case, is constructed for points in a two-dimensional space.

On one hand, the capitals of countries seem to be randomly distributed, meaning the geographical or other metric principles by which the positions of capitals are chosen are not clear. On the other hand, it is evident that a capital city, like any other city, is unlikely to be located at an arbitrary point on the surface, as most of this surface is covered by oceans.

The nearest neighbor graph for this set is shown in Figure 1. It was constructed using the Networkx library for Python 3.10.

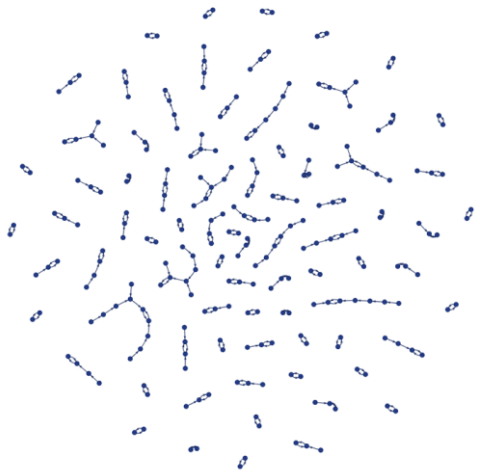

Fig. 1. Nearest neighbor graph for world capitals

Visually, major cluster centers are absent in this example. However, there are quite a few linear sequential structures, indicating a large number of vertices with degree 1.

Indeed, calculations showed that the graph consists of 80 disconnected fragments, yielding a ratio Q/N equal to 0.3187. From Table 2, for the case $n=2$, we find that the most probable fragment fraction value for random data is 0.310. The deviation of the obtained value from the optimal benchmark value is less than 0.01. Referring to Table 1, we observe that at a deviation of 0.05 from the maximum, the probability of graph realization decreases by 0.06. Therefore, in linear approximation interpolation, with confidence level approximately 0.99, the data can be considered independent according to the criterion of NNGs connected components.

On the other hand, the proportion of vertices with zero degree is 0.255. Vertices with degree 1 make up 0.506, those with degree 2 amount to 0.223, and those with degree 3 contribute 0.016. No other degrees occur in this graph. According to Table 4, the most probable fraction of vertices with zero degree in the two-dimensional case is 0.392, which exceeds the observed value by 0.137. Based on Table 3, even at half this deviation, the likelihood of dependency in the data already surpasses 0.96. Therefore, with a probability greater than 0.999, the arrangement of data points should be regarded as dependent.

Thus, applying various criteria for estimating the probability of NNGs structure allowed detecting dependency among the data. Note that this result does not contradict the fact that other criteria indicated independence of point locations. The reason lies in the observation that if a graph is not random, it may still share similar characteristics with a random graph, whereas the converse statement does not hold true.

4. Testing data for independence regardless of dimensionality

This methodology can also be used to analyze independence between data in an arbitrary sample. Such a situation typically arises when analyzing a sequence of some physical measurements. In this case, the data are arranged in the same sequence as in the experiment, in the form of the upper triangular part of a square matrix, and are interpreted as conditional distances between points, regardless of the dimension.

As an example, consider the task of analyzing the stability of locomotion during human movement on a treadmill. The experiment was described in detail in the work [13]. A sequence of pressure sensor readings is analyzed as a person moves on a special platform at a constant speed. The pressure curve is quasi-periodic, which allows for the extraction of the pattern shape of one step, i.e., the determination of some average pressure profile for a single step. After this, the obtained average profile is subtracted from the profiles of the actual steps. This results in a series of residuals. If this series is stationary, it indicates stable movement, meaning the body does not generate mechanical disorders. If the residual series is non-stationary, it indicates a disturbance in locomotor functions. In this example, the movement of an astronaut returning from a long spaceflight was analyzed. Therefore, it is reasonable to assume that the movement would not be entirely stable. A statistical analysis of this residual series can provide a valid quantitative assessment of this instability. However, the difficulty lies in the fact that the amount of data is not particularly large — approximately 20,000 data points. This prevents the use of time-varying distribution function analysis, as there should be many such segments, each of sufficient length for a reliable estimation of the corresponding probabilities. At the same time, this data can be represented as elements of an upper triangular matrix of size $200*200$, where each row is considered as a sequence of distances between the point numbered by that row and the other points. A nearest-neighbor graph for this structure is then constructed, as shown in Figure 2.

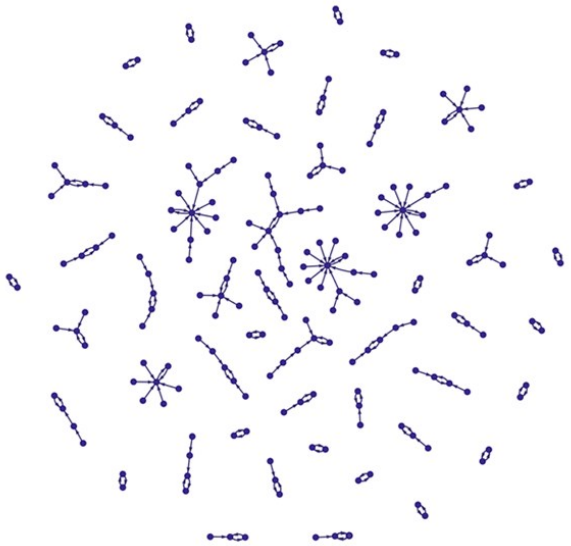

Fig. 2. Nearest neighbor graph for locomotion analysis

It is evident that the graph contains several attractor vertices with large degrees, so a high probability of dependence (non-stationarity) in the data is expected. The statistical properties of this graph are as follows. The number of structures is 50, i.e., $N/4$, which corresponds to the number of connected components for the "dimensionless" case. However, the proportion of vertices with zero degree is 0.415 instead of the reference value of 0.367. The deviation in frequencies amounted to $\nu = 0,048$. The probability of independence for such data can be determined using formula (4), which gives a value of approximately 0.02. Thus, the sequence of the residual series is non-stationary, and the astronaut's locomotor functions are disrupted.

5. Testing data for asymmetric quasi-distance matrices

The benchmark statistics of NNG structures generated by a non-symmetric random matrix can also be used in problems related to assessing the dependence of a system of original objects. In practice, such a problem arises when studying the proximity of probability density functions in a non-symmetric quasi-norm, known as relative entropy or the Kullback-Leibler distance.

Let's formulate the definition of this quasi-norm in the discrete case, which is the one considered in this example. Let $f(i), i = 1, 2, ..., n$ be some frequency distribution (empirical or theoretical), and let $g(i), i = 1, 2, ..., n$ be another distribution. We assume that both f_i and g_i are strictly positive. The Kullback-Leibler distance between the distributions f and g is defined as:

$$\rho_{fg} = \sum_{i=1}^{n} f_i \ln\left(\frac{f_i}{g_i}\right). \tag{5}$$

Consider as such distributions the empirical frequencies of letter usage in literary texts by different authors in the Russian language. In the work [14], benchmarks for over 8000 authors were constructed. The matrix of quasi-distances, calculated using formula (5), contains about 64,000 elements. By determining the minimum element in each row, we construct an adjacency matrix. The corresponding NNG is shown in Figure 3. It is evident that the graph contains quite a few clustering points — these are authors who, either consciously or unconsciously, try to imitate each other, creating texts that are statistically similar. Even without referring to the benchmark, it can be confidently said that the representative vectors of the authors are dependent with a probability close to 1.

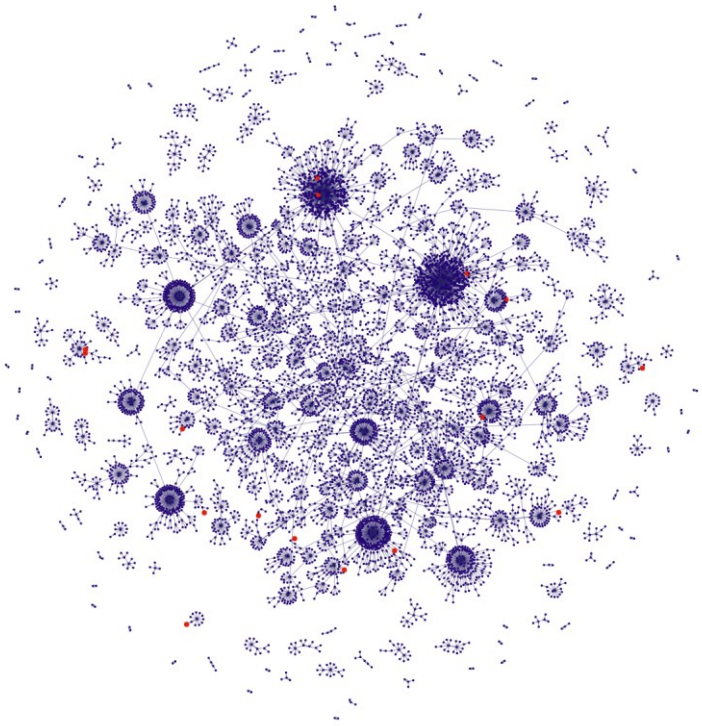

Fig. 3. Nearest-neighbor graph for author benchmark letter frequencies in the Kullback-Leibler quasi-norm.

The following results were obtained for the vectors under consideration. The proportion of vertices with zero degree was 0.503, with degree one was 0.253, with degree two was 0.122, with degree three was 0.056, and with degree four was 0.032. Comparing these results with the data from (3), we see that they differ significantly. Using the approximation (4), we find that the hypothesis of independence for the system of author benchmarks is rejected with a probability of one, as was expected from the qualitative assessment of the pattern in Figure 3.

6. Application to a long sequence analysis

Let us also present another important example of using the benchmark, when analyzing a sufficiently long sequence of random numbers. Such a sequence can be considered as the result of a pseudorandom number generator (PRNG), the quality of which is compared to the benchmark built using reliable generators. These include, for example, Xoshiro 512 [15] and Mersenne Twister [16]. If we use the existing sequence of numbers to generate an alternative benchmark for NNGs structures, the deviation of the resulting distributions from the base benchmark would indicate that the data in such a long sample are dependent.

In particular, due to the need for generating sufficiently long sequences of random independent variables, the statistical analysis of the sequence of digits in the decimal expansion of the number π is of great practical interest. On the one hand, since the number π can be calculated using mathematical formulas, it is not random. However, the question is not about that, but rather whether, if we do not know that we are dealing with the number π, and analyze the sequence of its decimal digits, it is possible to determine that this sequence is not random? The author is not aware of any statistical proof of such dependence. On the contrary, classical statistical tests show that the digits of π form a random, uniformly distributed sequence, and the correlation between samples is close to zero. Of course, we do not know the number itself, but for the analysis, a computed sequence of 10 trillion decimal digits of π was used [17]. Thus, the results of further analysis pertain not to the theory of the number π itself, but only to the given finite sequence of its digits.

A computational experiment was conducted to test how independent the digits of the number π are, using the constructed benchmark for NNGs structure statistics. The experi-

ment was as follows: a window of 20 decimal digits was considered, and it was associated with a number from the interval $[0;1]$, in which these 20 digits appeared after the decimal point. The window was moved along the sequence of digits of π in steps of 20. The generated series of numbers was treated as the result of generating a random sequence in $[0;1]$. Using this sequence, NNGs structure statistic was built for symmetric matrices of size $N=100$ based on 1 million generated graphs.

Figures 4-6 show the NNGs distributions by fragments for two variants of random uniform symmetric generation and for the "pi-generator" for comparison.

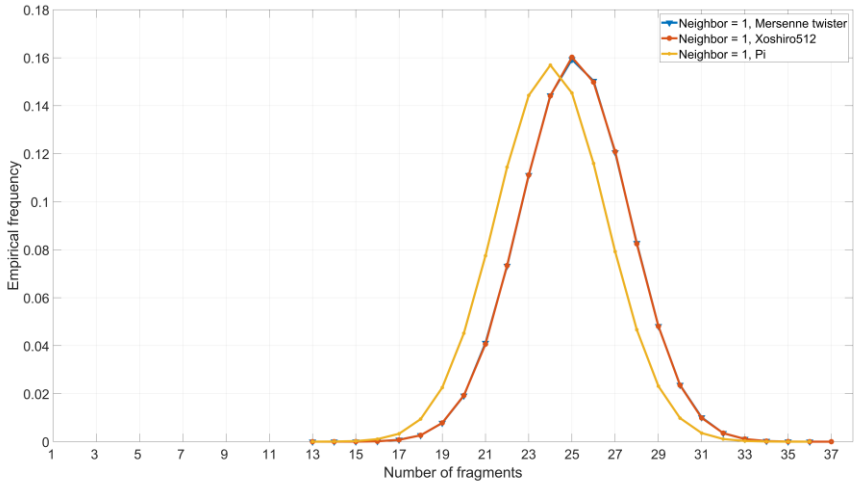

Fig. 4. NNGs distributions by the number of fragments for N = 100 for the first neighbor in regular symmetric and “pi-generation”

From these graphs, an important conclusion can be drawn: the NNG distributions for both standard, but structurally different, PRNGs coincide. This test showed that these generators are equivalent in terms of creating structures where there are no hidden connections between the elements that are not detected by standard criteria, i.e., they are sufficiently reliable. Moreover, the distributions for both generators match both in the number of connected components and in the degrees of the vertices.

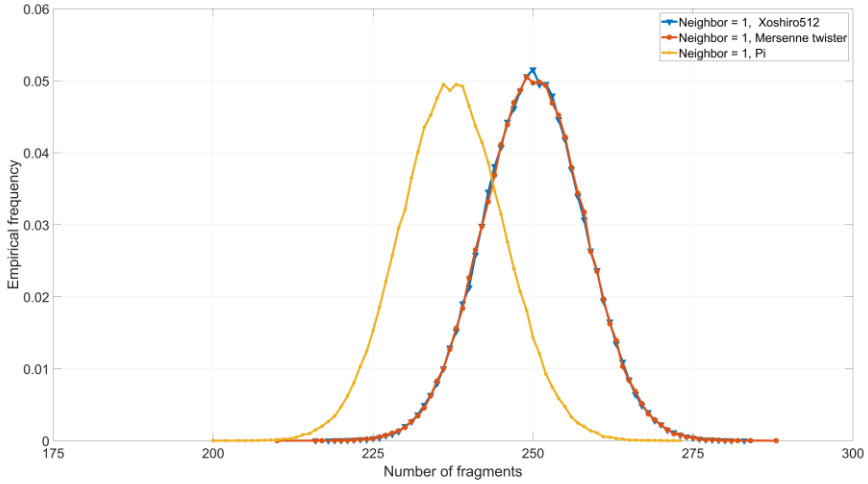

Fig. 5. NNGs distributions by the number of fragments for N = 1000 for the first neighbor in regular symmetric and “pi-generation”

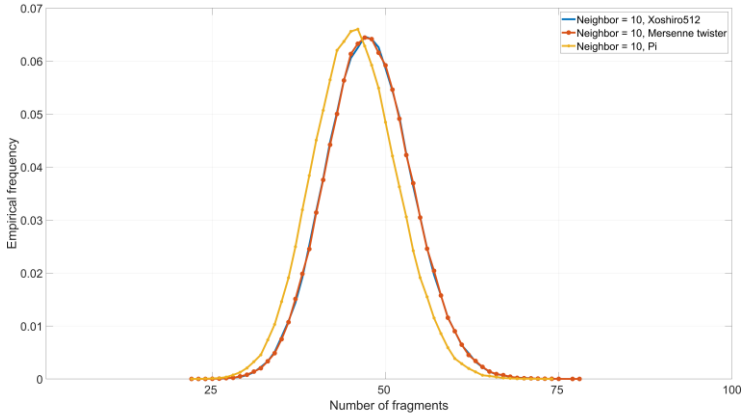

Fig. 6. NNGs distributions by the number of fragments for N = 100 for the 10-th neighbor in regular symmetric and “pi-generation”

Note that the maximum of the first neighbors' distribution for the "pi-generator" is shifted to the left of the point $N/4$. This is quite an interesting point, as for distance distributions in a finite-dimensional space, a similar graph is shifted to the right, as indicated by the data in Table 2. Therefore, the interpretation of the pi-generator as a random distribution in a small finite-dimensional space does not hold, and we are forced to admit that the pi-generator generates dependent systems of variables.

How significant are the differences in the distributions? The shift of the maximum to the left is fundamental. The difference between the two distribution functions in the C-norm was 0.174, which is a large value for samples of length around 25 million (this is the full number of data points for the number of NNGs fragments). According to the estimates for the agreed level of stationarity [18], such a distance is characteristic of samples of length 100 from the same distribution. For samples of length over 20 million, the probability of this distance, assuming the independence of the sample elements, is less than 0.01. Therefore, the distributions by fragments are different with a probability practically equal to one, and the digits of the number π cannot be considered statistically independent.

Also, for the "pi-generator," the following distribution of vertices by degrees was obtained:

$$\mu_0 = 0,383;\ \ \mu_1 = 0,353;\ \ \mu_2 = 0,176;\ \ \mu_3 = 0,063;\ \ \mu_4 = 0,018;\ \ \mu_5 = 0,005\,. \qquad (6)$$

These values differ significantly from the values presented in (3). With a probability practically equal to one, the digits of the number π form a statistically dependent sequence.

The distribution of the number of vertices with degree zero in the nearest neighbor graphs with 100 vertices, compared to standard symmetric generation, is shown in Figure 7.

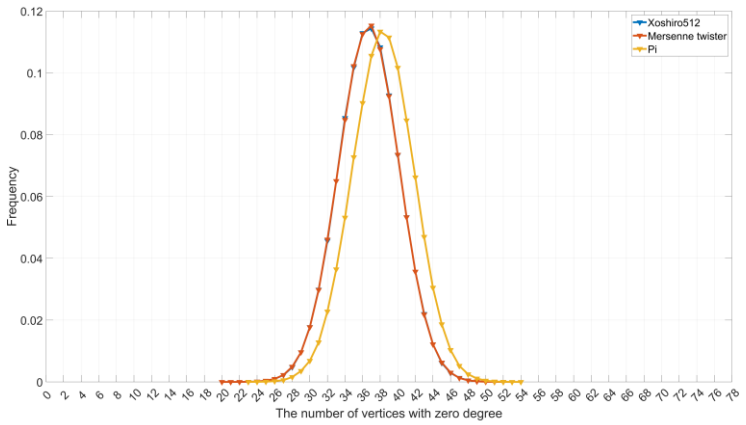

Fig. 7. Comparison of distributions of the number of vertices with zero degree for regular and “pi-generation”

The distance between the pi-distribution and both PRNGs in the C-norm was 0.209, which indicates that these are two distinct distributions. For the graphs generated by the number π, the number of vertices that are not nearest neighbors to any other vertices is slightly higher than for a regular random sample of numbers.

At the same time, the claim of dependence between the digits of the number π in this case has a high degree of reliability (greater than 0.99). Meanwhile, the Kolmogorov-Smirnov test applied to the sample of the considered set of random numbers (matrix elements) confirms the hypothesis of their uniform distribution with a probability greater than 0.99. Therefore, statistical criteria based on probability distributions of NNGs structures are, in certain cases, a more subtle tool for detecting hidden dependencies.

7. Conclusion

The construction of the benchmark for the probabilities of NNG structure realizations for random matrices from independent variables allowed for the development of two new approaches to analyzing random data in terms of testing their stationarity.

First, when the data is relatively small — on the order of a few thousand — conducting a full analysis of the stationarity of their sample distributions depending on the sample size is not feasible, as the number of samples for analyzing evolution is small. In this case, the data can be arranged in the upper triangular part of a matrix of appropriate size, symmetrically extended to the lower part, and the NNG generated by this specific matrix can be studied. If the resulting graph has a number of disconnected fragments and a distribution of vertex degrees close to the maxima of the corresponding benchmark distributions, the sample can be considered to consist of independent data, and the significance level at which this decision is made can be specified.

Second, when there is a large amount of data, the issue of their homogeneity arises, i.e., whether they belong to the same distribution. Before proceeding with the technically complex task of identifying possible clusters and relationships between them, it is first necessary to determine whether such an analysis is even meaningful. To do this, a method of rapid graph structure generation is used, assuming that the analyzed data represent some distances generated by a PRNG. A second, test benchmark is then created and compared with the original reference benchmark. If the differences in the distributions are significant, the data are considered dependent, after which filtering procedures can be applied using various methods, depending on the specific task.

The novelty of the proposed approach to analyzing samples from a non-stationary series, compared to classical stationary analogs of this task, lies in the fact that the indicator is not the deviation of current sample distributions from one another or from some pattern, but the deviation in the probability distributions of the realizations of structures, which are indirectly related to these distributions. Examples of applying this approach to a number of practical problems described in this work demonstrated its effectiveness compared to classical stationary methods.

Acknowledgements

This work is supported by RSF, project № 23-71-10055

References

1. Kislitsyn A.A. Investigation of nearest neighbor graph statistics. (in Russian) Mathematical Modeling, 2022, Vol. 34, № 8, pp. 110-126.

2. Kislitsyn A.A. Modeling of nearest neighbor graphs to assess the probability of independence of sample data. (in Russian) Mathematical Modeling, 2023, Vol. 35, № 7, pp. 63-82.

3. Verbik M.A., Guide to Modern Econometrics. (in Russian) Moscow: Nauchnaya kniga, 2008, 615 p.

4. Gilbert E.N., Random plane networks. J. Soc. Ind. Appl. Math. 9, 1961, pp. 533–553.
5. Erdős P., Rényi A. «On random graphs», Publ. Math. Debr., 1959, Vol. 6, pp. 290–297.
6. Haenggi M., Andrews J.G., Baccelli F., Dousse O., Franceschetti M., Stochastic geometry and random graphs for the analysis and design of wireless networks. IEEE Journal on Selected Areas in Communications, 2009, Vol. 27, № 7, pp. 1029-1046.
7. Pavlou O., Michos I., Papadopoulou Lesta V., Papadopoulos M., Papaefthymiou E.S., Efstathiou A., Graph Theoretical Analysis of local ultraluminous infrared galaxies and quasars, Astronomy and Computing, 2023, 45p.
8. Waikhom L., Patgiri R. A survey of graph neural networks in various learning paradigms: methods, applications, and challenges. Artif. Intell. Rev., 2023, Vol. 56, pp. 6295–6364.
9. Kolchin V.F. Random graphs. (in Russian) Moscow: Fizmatlit, 2004. 256 p.
10. Krivelevich M., Panagiotou K., Penrose M., Mcdiarmid C. Random Graphs, Geometry and Asymptotic Structure. Fountoulakis N, Hefetz D, eds. Cambridge University Press; 2016.
11. Kalyagin V.A., Koldanov A.P., Koldanov P., Pardalos P.M. Statistical Analysis of Graph Structures in Random Variable Networks, Springer, 2020, 101 p.
12. Kislitsyn A.A., Goguev M.V. Computational expansion into nearest neighbor graphs: statistics and dimensions of space. (in Russian) Keldysh Institute Preprints, 2022, № 88, pp. 1-32.
13. Fomina E.V., Ivchenko A.Yu., Lysova N.Yu., Zhedyaev R. Yu., Orlov Yu.N. Indicators of cosmonaut locomotor functions stability: a new method for ground-reaction forces analysis. Acta Astronautica, 2021, Vol. 189, pp. 679-686.
14. Kislitsyna M.Yu., Orlov Yu.N. Statistical analysis of the complete corpus of fiction in Russian and recognition of the author. (in Russian) Keldysh Institute Preprints, 2024, № 17, pp. 1-24.
15. Blackman D., Vigna S. Scrambled Linear Pseudorandom Number Generators. arXiv: 1805.01407v3 28 Mar 2022.
16. Makoto M., Takuji N. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. ACM Trans. Model. Comput. Simul, 1998, Vol. 8, № 1, pp. 3–30.
17. Trüb P. Digit Statistics of the First 22.4 Trillion Decimal Digits of Pi. arXiv, 2016, https://doi.org/10.48550/arXiv.1612.00489.
18. Orlov Yu.N. Kinetic methods of nonstationary time series. (in Russian) Moscow: MIPT, 2014, 276 p.