Nonparametric Analysis of Tracking Data in the Context of COVID-19 Pandemic

D. A. Klyushin

Abstract Methods of statistical pattern recognition are powerful tools for analysis of statistical data on COVID-19 pandemic. In this chapter, we offer a new effective online algorithm for detection of change-points in tracking data (movement data, health rate data etc.). We developed a non-parametric test for evaluation of the statistical hypothesis that data in two adjacent time intervals have the same distribution. In the context of the mobile phone tracking this means that the coordinates of the tracked object does not deviated from the base point significantly. For estimation of the health rate it means the absence of significant deviations from the norm. The significance level for the test is less than 0.05. The test permits ties in samples. Also, we show that results of comparison of the test with well-known Kolmogorov–Smirnov test. These results show that the proposed test is more robust, sensitive and accurate than the alternative ones. In addition, the new method does not require high computational capacity.

Keywords COVID-19 · Time series · Change-Point problem · Mobile control · Health rate control · Nonparametric test · Homogeneity measure

1 Introduction

In the midst of the COVID-19 pandemic, an avalanche of information is crashing down on scientists and decision makers. Among this huge volume of data ones of the most important are geolocation data of millions of mobile phone users in different regions and health data obtained from wearable devices. This data allows tracking migration of isolated patients and prevent their uncontrolled walks. Also, it allows detection of significance changes in the health rate of a user (temperature, pulse etc.). As far these data have a random nature, quite high rate of arrival and volatility, it is

D. A. Klyushin (✉)
Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Akademika Glushkova Avenue 4D, Kiev 03680, Ukraine
e-mail: dokmed5@gmail.com

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2020
A.-E. Hassanien et al. (eds.), Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach, Studies in Big Data 78, https://doi.org/10.1007/978-3-030-55258-9_3
impossible to correctly process and analyze them manually, and therefore make the right decisions based on it without automatic analysis. In this regard, the development of artificial intelligence methods for the analysis of sequential data about isolated patient movements and analysis of health data of wearable devices users becomes an urgent task. In particular, now there were developed highly effective methods for decision making during the COVID-19 pandemic [1], diagnosis and prediction models for COVID-19 patients' response to treatment [2], forecasting of coronavirus outbreak [3, 4], and methods of analysis of computer tomography images of COVID-19 patients [5]. But, the search of current publications in the repositories of preprints arXiv and medRxiv shows that yet there are no papers on analysis of data obtained via wearing gadgets.

Storing a huge amount of information is one of the main problems of big data. Offline algorithms using complete time series become impossible in such cases. Therefore, we must use online algorithms of time series analysis for detection of change points. In the context of tracking mobile phones and wearable devices the change of time series means the deviation of data from a reference point. It is convenient to use sliding window to analyze such data. However, this approach has two drawbacks. First, the sliding window size should be small enough to prevent the problems with storing data. Second, due to the small size of the sliding window algorithms of analysis must be quite sensitive and recognize changes using small samples. In this work, we develop such an algorithm and show its advantages over an alternative (the Kolmogorov–Smirnov statistics).

In the chapter it is proposed to use the novel nonparametric test for checking the statistical hypothesis that two given samples belong to the same distribution, i.e. they are homogeneous. This test is based on the measure of samples homogeneity, which is used to reject the hypothesis that the samples are identically distributed at a given significance level. The change-point detection is equivalent to change in distribution of data in adjacent intervals. Therefore, we reformulate this problem as the problem of two-sample homogeneity and solve it using the sliding window approach.

The motivation of this chapter is to describe a new approach to change-point detection in time series and demonstrate its application in the context of COVID-19 pandemic (analysis of geolocation tracking data, detection of changes in health data of patient etc.). Now, one of the most popular online statistical methods for solving such problems is the Kolmogorov–Smirnov test [6–10]. This non-parametric test is simple and effective when samples are not overlapping or they have small overlapping. But, it is very sensitive to outliers and may produce high level of false-positive responses when the samples are highly overlapping. The novelty of the proposed approach is that it is based on the Klyushin–Petunin non-parametric test for samples’ homogeneity that allows very effective comparing not only non-overlapping samples but the samples with highly overlapping. It is the first application of the Klyushin–Petunin test for detection of change-point of time series and as it will be shown it provides more effective estimation of homogeneity of overlapping samples comparing with the Kolmogorov–Smirnov test. The contribution of this chapter is that it describes a fast, accurate and robust online algorithm to detect change-points in time series that may be an effective tool for measuring the violation of quarantine
regime, keeping the social distance between people, and remote controlling the health 
data of patients suffering from COVID-19.

The chapter is organized in the following way. Section 2 describes the homogeneity 
measure for samples without ties. Section 3 describes the homogeneity 
measure for samples with ties. Section 4 contains the results of the experiments with 
samples from different distributions (normal, lognormal, uniform and gamma distri-
butions) with different degree of overlapping (from completely overlapped samples 
to disjoined samples) with ties and without ties. Section 5 summarizes the chapter.

2 Homogeneity Measure for Samples Without Ties

Consider samples \( u = (u_1, u_2, \ldots, u_n) \in G_1 \) and \( v = (v_1, v_2, \ldots, v_n) \in G_2 \) from 
populations \( G_1 \) and \( G_2 \) with distribution functions \( F_1 \) and \( F_2 \) that are absolutely 
continuous. The null hypothesis states that \( F_1 = F_2 \) and the alternative hypoth-
esis states that \( F_1 \neq F_2 \). There are several categories of criteria for testing such 
hypotheses: permutation criteria, rank criteria, randomization criteria, and distance 
criteria. In addition, these tests are divided into universal tests that are valid against 
any pair of alternatives (for example, Kolmogorov–Smirnov criterion [11, 12], and 
criteria that are correct against pairs of different alternatives of a particular class 
(Dickson [13], Wald and Wolfowitz [14], Mathisen [15], Wilcoxon [16], Mann–
Whitney [17], Wilks [18] etc.). Also, they could be divided into two large groups: 
nonparametric and conditionally nonparametric ones. Nonparametric criteria are 
criteria for testing the hypothesis of the homogeneity of general populations regard-
less of their distribution assumptions [11–18]. Conditionally non-parametric criteria 
(Pitman [19], Lehmann [20], Rosenblatt [21], Dwass [22], Fisz [23], Barnard [24], 
Birnbaum [25], Jockel [26], Allen [27], Efron and Tibshirani [28], Dufour and Farhat 
[29]) use some assumptions on distributions.

According to Hill’s assumption \( A_{(n)} \) [30], if random values \( u_1, u_2, \ldots, u_n \in G \) 
are exchangeable and belong to absolutely continuous distribution then

\[
P(u \in (u_{(i)}, u_{(j)})) = \frac{j - i}{n + 1},
\]

where \( j > i, u \in G \) is a sample value, and \( (u_{(i)}, u_{(j)}) \) is an interval formed by the 
\( i \)-th and \( j \)-th order statistics. This assumption was proved in papers of Yu.I. Petunin 
et al. for independent identically distributed random values [31] and for exchangeable 
identically distributed random values [32]. Also, it was developed a nonparametric 
test for detecting the homogeneity of samples without ties [33]. Later, this test was 
extended on the samples with ties [34].

These tests estimate homogeneity of samples \( u_1, u_2, \ldots, u_n \) and \( v_1, v_2, \ldots, v_n \) 
under the strong random experiment and does not depends on their distribution. 
Suppose that \( F_1 = F_2 \) and construct the variational series \( u_{(1)}, u_{(2)}, \ldots, u_{(n)} \). Denote 
as \( A_{(n)}^{(k)} \) an event \( \{ v_k \in (u_{(i)}, u_{(j)}) \} \). According to the Hill’s assumption, for \( j > i \)
where \( h_{ij}^{(n,k)} \) is the relative frequency of the occurrence of the event \( A_{ij}^{(k)} \) in \( n \) trials. Then, construct the confidence interval \( I_{ij}^{(n)} = \left( p_{ij}^{(1)} , p_{ij}^{(2)} \right) \) with a significance level defined by the parameter \( g \). Note, that if \( g \) is equal to 3 than the significance level of \( I_{ij}^{(n)} \) is less than 0.05 [33]. Also, the value of \( p \)-statistics very weakly depends on the choice of the confidence intervals for a binomial proportion [35]. Let \( B \) be an event \( \left\{ p_{ij} = \frac{j-i}{n+1} \in I_{ij}^{(n)} \right\} \). Put \( N = (n-1)n/2 \) and find \( L = \#B \). Then, \( h = L/N \) is a homogeneity measure of samples \( x \) and \( y \) which we shall call \( p \)-statistics.

Let us put \( h_{ij}^{(n)} = h, n = N \) and \( g = 3 \), and construct the Wilson (for definiteness) confidence interval \( I_{ij} = (p_1, p_2) \) for \( p(B) \). The confidence intervals \( I_{ij}^{(n)} = \left( p_{ij}^{(1)} , p_{ij}^{(2)} \right) \) and \( I = (p_1, p_2) \) are called intervals based on the 3 s-rule. The scheme of trials, where the events \( A_{ij}^{(k)} \) can arise when the hypothesis that distributions are identical holds is true, called a generalized Bernoulli scheme [36–38]. If the hypothesis does not hold, this scheme is called a modified Bernoulli scheme. In general case, when the null hypothesis can be either true or false, the trial scheme is called MP-scheme (Matveychuk–Petunin scheme). If \( F_1 = F_2, \lim_{n \to \infty} \frac{j-i}{n+1} \in (0, 1) \), and \( \lim_{n \to \infty} \frac{j-i}{n+1} \in (0, 1) \), then the asymptotic significance level \( \beta \) of a sequence of confidence intervals \( I_{ij}^{(n)} \) based on the 3 s-rule, is less than 0.05 [33]. Let \( B_1, B_2, \ldots \) be a sequence of events that may arise in a random experiment \( E, \lim_{n \to \infty} p(B_k) = p^* \), \( h_n(B_1), h_n(B_2), \ldots \) be a sequence of relative frequencies of the events \( B_1, B_2, \ldots \), respectively, and \( \frac{k}{n} \to 0 \) as \( k \to \infty \). We shall call an experiment \( E \) a strong random experiment if \( h_n(B_k) \to p^* \) as \( k \to \infty \). In a strong random experiment the asymptotical significance level of the Wilson confidence interval \( I_n \) as \( g = 3 \) is less than 0.05. The test for the null hypothesis \( F_1 = F_2 \) with a significance level, which is less that 0.05, is following: if \( I_n \) contains 0.95, the null hypothesis is accepted, else the null hypothesis is rejected.
3 Homogeneity Measure for Samples with Ties

Unfortunately, the main assumption used in the test based on \( p \)-statistics is that the distributions \( F_1 \) and \( F_2 \) are absolutely continuous, i.e. the sample must not contain any repetitions. Therefore, it is necessary to extend the above ideas to samples with repetitions, i.e. ties.

Consider the sample \( u = (u_1, u_2, \ldots, u_n) \) from distribution \( F \) obtained in a strong random experiment. A tie is a sample value \( u_k \) that has duplicates. Multiplicity \( t(u_k) \) is the number of duplicates \( u_k \) in \( u \). If a joint distribution function is absolutely continuous and there are no duplicates in a sample, then the probability of the tie in the sample is equal to zero. In practice, we obtain sample values as the result of measuring a quantitative features. The measurements have a restricted accuracy, so a sample may contain ties. We propose a modification of \( p \)-statistics that could be used to estimate homogeneity of samples containing ties.

Consider a sample \( u = (u_1, u_2, \ldots, u_n) \) drawn from a population \( G \) with a distribution function that is absolutely continuous. If the sample values \( \tilde{u}_k \in \tilde{u} \) are approximations of the elements of the hypothetical sample \( u \), we shall call the sample \( \tilde{u} = (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_n) \) empirical. Population \( \tilde{G} \) we shall call an empirical population corresponding to \( G \). As a rule, real empirical samples contain ties.

Denote as \( P_\alpha \) the flooring of a value:

\[
|P(u) - u| \leq 10^{-\alpha} = \Delta
\]

Let \( u_{(1)} < u_{(2)} < \cdots < u_{(n)} \) and \( \tilde{u}_{(1)} < \tilde{u}_{(2)} < \cdots < \tilde{u}_{(n)} \) be variational series corresponding to the hypothetical and empirical samples.

For a sample value \( u^* \) that is drawn from \( G \) independently from \( u \) the following equation holds [30]:

\[
p\left(u^* \in \left[u_{(k)}, u_{(k+1)}\right]\right) = \frac{1}{n+1},
\]

where \( k = 0, 1, \ldots, n, u_{(0)} = -\infty \), and \( u_{(n+1)} = \infty \).

This formula may be extended to empirical samples [34].

**Theorem** If the distribution \( F \) is differentiable and Lipschitz continuous, i.e.

\[
|F(x) - F(y)| \leq K|x - y|
\]

the sample value \( u^* \) is independent on \( u \), and the order statistics \( \tilde{u}_k = P_\alpha(u_{(i)}) \), \( k \leq i \) of the empirical sample \( \tilde{u} = P_\alpha(u) = (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_n) \) is a tie with multiplicity \( t(\tilde{u}_k) \), then
\[ \gamma(\tilde{u}_{(k)}) + \frac{1}{n + 1} - K\Delta \leq p(\tilde{u}^* \in [\tilde{u}_{(k)}, \tilde{u}_{(k+1)}) \leq \gamma(\tilde{u}_{(k)}) + \frac{1}{n + 1} + K\Delta, \]

where \( \tilde{u}^* = P_a(u^*), \tilde{u}_{(k)} = P_a(u_{(i)}), \gamma(\tilde{u}_{(k)}) = \frac{t(\tilde{u}_{(k)}) - 1}{n + 1}, \lambda = t(\tilde{u}_k) - 1, 1 \leq k \leq n, \) and \( \Delta = 10^{-\alpha}. \)

**Corollary** With precision to the rounding error

\[ p(\tilde{u}^* \in [\tilde{u}_{(k)}, \tilde{u}_{(k+1)}) \approx \gamma(\tilde{u}_{(k)}) + \frac{1}{n + 1}. \]  (4)

Thus, if \( p(\tilde{u}^* \in [\tilde{u}_{(i)}, \tilde{u}_{(j)}) \), \( i < j \), and \( 1 \leq i, j \leq n \), then

\[ p_{ij} = p(A_{ij}) = p(\tilde{u}^* \in [\tilde{u}_{(i)}, \tilde{u}_{(j)}) \approx \gamma_i + \gamma_{i+1} + \cdots + \gamma_{j-1} + \frac{j - i}{n + 1}, \]  (5)

where \( \gamma_i = \gamma(\tilde{u}_{(i)}), A_{ij} = \{\tilde{u}^* \in [\tilde{u}_{(i)}, \tilde{u}_{(j)}]\}. \)

**Remark** If \( \tilde{u}_{(i)}, \ i \leq l \leq j - 1 \) is not a tie, then \( \gamma_i = 0 \) and (5) transforms to (1).

Denote by \( H \) the null hypothesis on identity of absolutely continuous distribution functions \( F_1 \) and \( F_2 \) of populations \( G_1 \) and \( G_2 \) respectively. Consider \( \tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_n) \in G_1 \) and \( \tilde{u}' = (\tilde{u}'_1, \ldots, \tilde{u}'_n) \in G_2 \), and let \( \tilde{u}_{(1)} \leq \cdots \leq \tilde{u}_{(n)} \), \( \tilde{u}'_{(1)} \leq \cdots \leq \tilde{u}'_{(n)} \) be their variance series, where \( \tilde{G}_1 \) and \( \tilde{G}_2 \) are the empirical populations corresponding to \( G_1 \) and \( G_2 \). Suppose that \( F_1 = F_2 \) and denote by \( A_{ij}^{(k)} \)

\( k = 1, 2, \ldots, n \) the random event \( \left\{ \tilde{u}_k \in \left( \tilde{u}_{(i)}, \tilde{u}_{(j)} \right) \right\} \). If \( F_1 = F_2 \) then the probability of \( A_{ij}^{(k)} \) is equal to (5). Construct \( I_{ij}^{(n)} = (p_{ij}^{(1)}, p_{ij}^{(2)}) \) as a Wilson confidence interval (1) for the unknown probability of the event \( A_{ij}^{(k)} \). Denote by \( N = (n - 1)n/2 \) the number of all possible confidence intervals \( I_{ij}^{(n)} = (p_{ij}^{(1)}, p_{ij}^{(2)}) \), by \( B \) an event

\[ \left\{ p_{ij} = \frac{j - i}{n + 1} \in I_{ij}^{(n)} \right\} \] and by \( L = \#B \). Put \( h^{(n)} = \frac{L}{N} \). Since \( h^{(n)} \) is a relative frequency of \( B \), then, for \( h_{ij}^{(n)} = h^{(n)}, n = N \) and \( g = 3 \), using (1) we obtain the confidence interval \( I_{ij}^{(n)} = (p^{(1)}, p^{(2)}) \) for the probability \( p(B) = 1 - \beta \). The statistics \( h^{(n)} \) is called empirical \( p \)-statistics. It estimates the homogeneity of samples \( \tilde{u} \) and \( \tilde{u}' \).

4 Experiments and Results

Let us divide the time series in portions of size \( n \) and consider two windows of lengths \( n \).

\[ \overbrace{X_1, X_2, \ldots, X_n, X_{n+1}, \ldots, X_{2n}}^{\text{window 1}} \]

\[ \overbrace{\text{window 2}} \]
In this experiment we use samples generated with the given distribution. If the samples \( x_1, x_2, \ldots, x_n \) and \( x_{n+1}, x_2, \ldots, x_{2n} \) are heterogeneous, i.e. they have different distributions, then the point \( x_{n+1} \) is a change point, that is a tracked date deviate from the reference point significantly (Fig. 1). The windows must have equal sizes but may have different starting points. As an initial point we selected two windows with the same left end, let one window be fixed and the second window be moving with the step \( l \) (for example, \( l = 1 \)) overlapping with the first window. As a variant, we can consider adjacent non-overlapping windows also.

Hereinafter \( N(m, \nu) \) is a normal distribution, where \( m \) is the expected value and \( \nu \) is the variance, \( LN(m, \nu) \) is a lognormal distribution, where \( m \) is the expected value and \( \nu \) is the variance, \( U(a, b) \) is the uniform distribution on an interval \( (a, b) \), and \( G(a, b) \) is a two-parameter gamma distribution. For comparison we selected \( p \)-statistics and wide-used Kolmogorov–Smirnov statistics. The width of sliding window is equal to 40 (due to the requirement that the significance level of Klyushin-Petunin test must be less than 0.05). The sliding step is equal to 1. The averaging was made on 10 experiments. The results demonstrate the high effectiveness of the proposed method (see Figs. 2, 3, 4 and 5).

Here, we selected the distribution parameters such that to demonstrate the deviation of the expected values of the samples from the reference point, but the same experiments may be conducted with the same means and various standard deviation. As was shown in [33] the \( p \)-statistics has great advantage in these cases.

As we see, the curves of \( p \)-statistics and the Kolmogorov–Smirnov statistics decrease from the left upper corner (corresponding to the same samples which imply the high homogeneity measure) toward right lower corner (corresponding to the very different samples which implies the low homogeneity measure). The sensitiveness is estimated by the slope of these curves (the more steep curve means the more sensitive homogeneity measure). The monotonicity is estimated visually.
Fig. 2 P-statistics for a normal and b lognormal distribution functions
Fig. 3 P-statistics for a uniform and b gamma distribution functions
Fig. 4 $P$-value of the Kolmogorov–Smirnov test for a normal and b lognormal distribution functions
Fig. 5 $p$-value of the Kolmogorov–Smirnov test for a uniform and b gamma distribution functions

The results show that the $p$-statistics is monotonic, sensitive, and robust homogeneity measure. Since the step of sliding window is equal to 1, at every time step the first sample becomes more “contaminated” by the sample values of the second sample. Thus, we handle the variants with different degrees of overlapping between
samples: from total coinciding to disjoining. As we see, in many variants of distributions the Kolmogorov–Smirnov statistics as opposed to the $p$-statistics demonstrate non-monotonic and non-robust behavior.

There is one more method to estimate the sensitiveness of the test statistics. Remember, that the null hypothesis about identity of distribution functions in the Kolmogorov–Smirnov test is rejected if the Kolmogorov–Smirnov statistics is less than 0.05. Vice versa, in the Klyushin–Petunin test the null hypothesis is rejected when the confidence interval for the $p$-statistics does not contain 0.95. As we see on Fig. 2 and Fig. 4 (shifted normal and log-normal distributions), the $p$-statistics rejects the null hypothesis starting from the half point, while the Kolmogorov–Smirnov statistics rejects the null hypothesis starting from the one quarter point. Therefore, in the cases of shifted normal and lognormal distributions the Kolmogorov–Smirnov test is more sensitive. From the other side, Figs. 3 and 5 demonstrate the advantage of the $p$-statistics in the case when distributions are different but significantly overlapping. As previously, the $p$-statistics rejects the null hypothesis starting from the half point but in the cases of overlapped uniform and gamma distributions the Kolmogorov–Smirnov test cannot distinguish them.

We may conclude that the Kolmogorov–Smirnov statistics is more preferable when the distributions are shifted (i.e. they have different expected values) and the $p$-statistics in useful in the same cases but overcome the Kolmogorov–Smirnov test in the cases when the distributions have significant overlapping (for example, normal distributions with the same expected values and different variances or uniform distributions on overlapping intervals). It may be explained by the different nature of these statistical tests. The Kolmogorov–Smirnov test detects the difference between the empirical distributions functions and is more sensitive to shifts of distributions and outliers. Instead, the $p$-statistics uses the information on permutations and can detect both shifting and changes of the shapes of distributions remaining robust to outliers.

Next, let us consider two quasi-real experiments imitation the deviation of the tracked object from the reference point. Obviously, the real experiment requires real data but there are private and inaccessible, so we just demonstrate the capabilities of the proposed test for solving such problems on the artificial data.

Case 1 is illustrated by Fig. 1. We simulate the time series consisting of two slightly overlapping samples: the first sample consists of 50 random values from the uniform distribution $U(0,1)$ and the second sample consists of 50 random values from $U(0,1) + 0.7$. We compare two consecutive sliding windows. The width of sliding window is equal to 10. The sliding step is equal to 10. The largest difference is expected between 5th and 6th segments (5th row of Table 1). Indeed, in this case we obtain the minimal value of $p$-statistics (0.8444) and according to the Klyushin–Petunin test [33] the null hypothesis on homogeneity of the sample is rejected because the confidence interval for the $p$-statistics does not contain 0.95). The Kolmogorov–Smirnov test demonstrates the over-conservativeness. According to its $p$-value the null hypothesis is rejected in the first, second, fourth, sixth and eights cases besides that by assumption first five segment and the second five segments were drawn from
the same distributions respectively. It may be explained by that the Kolmogorov–Smirnov test based in the difference between the empirical distributions functions and is very sensitive to outliers.

In order to demonstrate the difference between empirical $p$-statistics and original $p$-statistics let us consider the case for samples rounded up to two decimal digits. As we see, the original $p$-statistics overestimate the homogeneity of the rounded samples (see Table 2). The Kolmogorov–Smirnov statistics demonstrate the same properties as in Case 1. But these results may be explained by the small size of samples. Thus, we must consider the case with more large samples.

In Case 2 we simulate the time series consisting of ten alternating samples: the odd samples consist of 50 random values from the uniform distribution $U(0,1)$ and the even samples consist of 50 random values from $U(0,1)$ plus 0.7. Now the width of sliding window is equal to 50 and the sliding step is equal to 50.

Table 1  Empirical $p$-statistics (h) and its left (hl) and rights (hu) confidence bounds, $p$-statistics without ties (ro) and its left (ro_l) and right (ro_u) confidence bounds, the Kolmogorov–Smirnov statistics (ro_ks) and its $p$-value in Case 1

| h       | h_l   | h_u   | ro_p  | ro_l  | ro_u  | ro_ks | $p$-value |
|---------|-------|-------|-------|-------|-------|-------|-----------|
| 0.9778  | 0.7983| 0.998 | 0.9778| 0.7983| 0.998 | 0      | 0         |
| 0.9111  | 0.7077| 0.9775| 0.9111| 0.7077| 0.9775| 0.1    | 2e-005    |
| 0.8889  | 0.6803| 0.9678| 0.8889| 0.6803| 0.9678| 0.5    | 0.112     |
| 0.9333  | 0.7363| 0.986 | 0.9333| 0.7363| 0.986 | 0.3    | 0.00556   |
| 0.8444  | 0.6283| 0.9457| 0.8444| 0.6283| 0.9457| 0      | 0         |
| 1       | 0.8333| 1     | 1     | 0.8333| 1     | 0.4    | 0.03008   |
| 0.9778  | 0.7983| 0.998 | 0.9778| 0.7983| 0.998 | 0.5    | 0.112     |
| 0.9556  | 0.7663| 0.993 | 0.9556| 0.7663| 0.993 | 0.7    | 0.664     |
| 1       | 0.8333| 1     | 1     | 0.8333| 1     | 0.4    | 0.03008   |

Table 2  Empirical $p$-statistics (h) and its left (hl) and rights (hu) confidence bounds, $p$-statistics without ties (ro) and its left (ro_l) and right (ro_u) confidence bounds, the Kolmogorov–Smirnov statistics (ro_ks) and its $p$-value for rounded Case 1

| h       | h_l   | h_u   | ro_p  | ro_l  | ro_u  | ro_ks | $p$-value |
|---------|-------|-------|-------|-------|-------|-------|-----------|
| 1       | 0.8333| 1     | 0.9778| 0.7983| 0.998 | 0.2   | 0.00064   |
| 1       | 0.8333| 1     | 1     | 0.8333| 1     | 0.2   | 0.00064   |
| 0.9778  | 0.7983| 0.998 | 0.9556| 0.7663| 0.993 | 0.6   | 0.3088    |
| 1       | 0.8333| 1     | 1     | 0.8333| 1     | 0.1   | 2e−005    |
| 0.9111  | 0.7077| 0.9775| 0.9111| 0.7077| 0.9775| 0.3   | 0.00556   |
| 1       | 0.8333| 1     | 1     | 0.8333| 1     | 0.3   | 0.00556   |
| 0.7778  | 0.5556| 0.9074| 1     | 0.8333| 1     | 0.6   | 0.3088    |
| 1       | 0.8333| 1     | 1     | 0.8333| 1     | 0.7   | 0.664     |
| 1       | 0.8333| 1     | 1     | 0.8333| 1     | 0.3   | 0.00556   |
As we see in Table 3, all statistics demonstrate the stable results. All pairs of segments are recognized as heterogeneous. Therefore, as a practical recommendation we may state that the sample size must exceed 40.

### 5 Conclusions and Future Work

The new online algorithm implementing the Klyushin–Petunin test on streaming geolocation data is effective, sensitive, and robust. It does not depend on assumptions about distributions of samples and equally sensitive to difference between expected values (location hypothesis) and shapes of the distributions (scale hypothesis). Its significance level is less that 0.05. It does not require the special conditions for saving the data. The algorithm effectively solve the change-point and is more sensitive than alternative Kolmogorov–Smirnov when the size sample is small (less that 40).

The experiments with samples from normal, lognormal, uniform, and gamma distribution with parameters describing different degrees of overlapping shows that \( p \)-statistics is more stable, sensitive and monotonic than Kolmogorov–Smirnov statistics. The Kolmogorov–Smirnov statistics demonstrates non-monotonic behavior due to the sensitiveness to outliers.

The Kolmogorov–Smirnov statistics allow to detect the change-point in data stream, but the \( p \)-statistics has better accuracy and robustness, because it is very effective both for shifted samples drawn from the distributions with different means and for samples with large overlapping. The main feature of the \( p \)-statistics is the monotonicity depending on the degree of overlapping between the samples (the more overlapping the more \( p \)-statistics, and vice versa) and a sharp jump at the change-point.
The future scope of the work is to apply the test to real data obtained during COVID-19 pandemic, decrease its computational complexity, and extend it to multivariate time series.

References

1. Fong, S.J., Li, D., Dey, N., Crespo, R.G., Herrera-Viedma, E.: Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Appl. Soft Comput. 106282 (2020)
2. Elghamrawy, S.M., Hassanien, A.E.: Diagnosis and prediction model for COVID19 patients response to treatment based on convolutional neural networks and whale optimization algorithm using CT image. medRxiv 2020.04.16.20063990; (2020). https://doi.org/10.1101/2020.04.16.20063990
3. Fong, S.J., Li, D., Dey, N., Crespo, R.G., Herrera-Viedma, E.: Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak. Int. J. Interact. Multimedia Artif. Intell. 6(1), 132–140 (2020). https://doi.org/10.9781/ijimai.2020.02.002
4. Elmousalami, H.H., Hassanien, A.E.: Day level forecasting for coronavirus disease (COVID-19) spread: analysis, modeling and recommendations. arXiv preprint arXiv:2003.07778 (2020)
5. Rajinikanth, V., Dey, N., Raj, A.N.J., Hassanien, A.E., Santosh, K.C., Sri Madhava Raja, N.: Harmony-search and otsu based system for coronavirus disease (COVID-19) detection using lung CT scan images. arXiv preprint arXiv:2004.03431 (2020)
6. Shao, X., Zhang, X.: Testing for change points in time series. J. Am. Stat. Assoc. 105(491), 1228–1240 (2010). https://doi.org/10.1198/jasa.2010.tm10103
7. Tran, D., Gaber, M.M., Sattler, K.: Change detection in streaming data in the era of big data: models and issues. SIGKDD Explorations 16, 30–38 (2014)
8. Zhang, W., James, N.A., Matteson, D.S.: Pruning and nonparametric multiple change point detection. In: 2017 IEEE international conference on data mining workshops (ICDMW), New Orleans, LA, pp. 288–295 (2017)
9. Padilla, O.H.M., Athey, A., Reinhart, A., Scott, J.G.: Sequential nonparametric tests for a change in distribution: an application to detecting radiological anomalies. J. Am Statist. Assoc. 114(526), 514–528 (2019). https://doi.org/10.1080/01621459.2018.1476245
10. Padilla, O.H.M., Yu, Y., Wang, D., Rinaldo, A.: Optimal nonparametric change point detection and localization. arXiv preprint arXiv:1905.10019 (2019)
11. Smirnov, N.V.: Estimate of difference between empirical distribution curves in two independent samples. Byull Mosk Gos Univ. 2(2), 3–14 (1939)
12. Smirnov, N.V.: On the deviations of an empirical distribution curve. Mat. Sb 6(1), 3–26 (1939)
13. Dixon, W.G.: A criterion for testing the hypothesis that two samples are from the same population. Ann. Math. Stat. 11, 199–204 (1940)
14. Wald, A., Wolfowitz, J.: On a test whether two samples ate from the same population. Ann. Math. Stat. 11, 147–162 (1940)
15. Mathisen, H.C.: A method of testing the hypothesis that two samples are from the same population. Ann. Math. Stat. 14, 188–194 (1943)
16. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrika 1, 80–83 (1945)
17. Mann, H.B., Whitney, D.R.: On a test of whether one of the random variables is stochastically larger than other. Ann. Math. Stat. 18, 50–60 (1947)
18. Wilks, S.S.: A combinatorial test for the problem of two samples from continuous distributions. Proc. Fourth Berkeley Symp. Math. Stat. Prob. 1, 707–717 (1961)
19. Pitman, E.J.G.: Significance tests which may be applied to samples from any populations. J. Royal Stat. Soc. Ser. A. 4, 119–130 (1937)
20. Lehmann, E.L.: Consistency and unbiasedness of certain nonparametric tests. Ann. Math. Stat. 22, 165–179 (1947)
21. Rosenblatt, M.: Limit theorems associated with variants of the von Mises statistic. Ann. Math. Stat. 23, 617–623 (1952)
22. Dwass, M.: Modified randomization tests for nonparametric hypotheses. Ann. Math. Stat. 28, 181–187 (1957)
23. Fisz, M.: On a result be M. Rosenblatt concerning the Mises-Smirnov Test. Ann. Math. Stat. 31, 427–429 (1960)
24. Barnard, G.A.: Comment on “The spectral analysis of point processes” by M.S. Bartlett. J. Royal Stat. Soc. Ser. B 25, 294 (1963)
25. Birnbaum, Z.W.: Computers and unconventional test-statistics. In: Prochan, F., Serfling, R.J. (eds.) Reliability and Biometry, pp. 441–458. SIAM, Philadelphia, PA (1974)
26. Jockel, K.-H.: Finite sample properties and asymptotic efficiency of Monte Carlo tests. Ann. Stat. 14, 336–347 (1986)
27. Allen, D.L.: Hypothesis testing using L1-distance bootstrap. Am. Stat. 51, 145–150 (1997)
28. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Vol. 57 of Monographs on Statistics and Applied Probability. Chapman-Hall, New York (1993)
29. Dufour, J.-M., Farhat, A.: Exact nonparametric two-sample homogeneity tests for possibly discrete distributions. Center for interuniversity research in quantitative economics (CIREQ). Preprint 2001–23. California Press, pp. 707–717 (2001)
30. Hill, B.M.: Posterior distribution of percentiles: bayes’ theorem for sampling from a population. J. Am. Stat. Assoc. 63, 677–691 (1968)
31. Madreimov, I., Petunin, Yu.I: Characterization of a uniform distribution using order statistics. Teor. Ver. Mat. Statist. 27, 96–102 (1982)
32. Andrushkiw, R.I., et al.: Construction of the bulk of general population in the case of exchangeable sample values. In: Proceedings of the International Conference of Mathematics and Engineering Techniques in Medicine and Biological Science (METMBS’03), Las Vegas, Nevada, USA., pp 486–489 (2003)
33. Klyushin, D.A., Petunin, Yu.I: A nonparametric test for the equivalence of populations based on a measure of proximity of samples. Ukrainian Math. J. 55(2), 181–198 (2003)
34. Andrushkiw, R.I., Klyushin, D.A., Petunin, Yu.I.: Proximity measure between samples with repetition factor greater than one. CAMS Research Report: 0708–12, New Jersey Institute of Technology (2008)
35. Klyushin, D.A., Lyashko, S.I., Zub, S.S.: (n) Assumption in machine learning. In: Computational Linguistics and Intelligent Systems, Lviv: Lviv Politechnic Publishing House, 2019. Vol 2: Proceedings of the 3rd International Conference, COLINS 2019, Workshop, Kharkiv, Ukraine, April 18–19, pp. 32–38 (2019)
36. Matveichuk, S.A., Petunin, Yu.I: Generalization of Bernoulli schemes that arise in order statistics. I. Ukrainian Math. J. 42(4), 459–466 (1990)
37. Matveichuk, S.A., Petunin, Yu.I: Generalization of Bernoulli schemes that arise in order statistics II. Ukrainian Math. J. 43(6), 728–734 (1991)
38. Johnson, N., Kotz, S.: Some generalizations of Bernoulli and Polya-Eggenberger contagion models. Statist. Paper 32, 1–17 (1991)