Journal of
open research software

pyHomogeneity: A Python Package for Homogeneity Test of Time Series Data

SOFTWARE METAPAPER

MD. MANJURUL HUSSAIN

ISHTIAK MAHMUD

SHEIKH HEFZUL BARI

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

The pyHomogeneity package is a pure python package for performing homogeneity tests. The homogeneity test is applied to detect one (or more) change-point/breakpoint in the time series or sequential data. It is one of the essential tests for time series data (e.g. financial time series, hydrology, climate study, etc.). However, till now, there is no Python package available for the homogeneity test. This is where the pyHomogeneity package comes in. Currently, Pettitt's Test, four variants of Buishand's Test, and SNHT Test can be performed using this package. The package is freely available for public use.

CORRESPONDING AUTHOR:

Md. Manjurul Hussain

Institute of Water and Flood Management, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

mmhs013@gmail.com

KEYWORDS:

Homogeneity test; Pettitt's Test; Buishand's test; SNHT test; Change point detection; Breakpoint detection

TO CITE THIS ARTICLE:

Hussain MM, Mahmud I, Bari SH 2023 pyHomogeneity: A Python Package for Homogeneity Test of Time Series Data. *Journal of Open Research Software,* 11: 4. DOI: https://doi.org/10.5334/jors.427

(1) OVERVIEW

INTRODUCTION

The homogeneity test is a statistical test method that checks if two (or more) datasets came from the same distribution or not. In the time series, the homogeneity test is applied to detect one (or more) change-point/breakpoint in the series. This breakpoint occurs when the dataset changes its distribution. This detection of distribution change makes the homogeneity test an essential test in statistical analysis. In fact, homogeneity is one of the most important assumptions in time series analysis. For example, predicting financial time series, analysis or forecasting of historical meteorological data, etc. If non-homogeneity is undetected, selection of the best model could be influenced by the selected sample. There are several tests available to check homogeneity viz. Pettitt Test, Standard Normal Homogeneity Test (SNHT), Buishand's Test, etc. These tests can be performed using different commercial software packages like XLSTAT [1]. Different programming languages like R, Matlab, etc. have different packages or scripts for those tests [2, 3]. However, a python based single package to perform the most widely used homogeneity tests will save time and bring diversity into the analysis. Because Python is one of the most widely used tool used by data scientists. A large number of data analysis and research tools are also developed using Python. However, till now, there is no Python package available for the homogeneity test. pyHomogeneity package will fill this gap.

IMPLEMENTATION AND ARCHITECTURE

pyHomogeneity is a pure Python implementation for the homogeneity test. Two core python packages named NumPy [4] and SciPy [5] are used to build pyHomogeneity. The vectorization approach is used instead of the traditional for loop to improve calculation speed. This package can perform six different homogeneity tests (three unique tests with four variants of Buishand's test), which are widely used in time series analysis. Available tests in the pyHomogeneity package are briefly discussed below:

Pettitt Test

In 1979, A. N. Pettitt proposed a change-point detection test based on Mann-Whitney two-sample test [6]. For a continuous dataset, Pettitt statistic U(k) can be calculated using the equation below:

$$U(k) = 2\sum_{i=1}^{k} r_i - k(n+1)$$

Where, $r_1, r_2, r_3,\dots\dots\dots, r_k$ are the ranks of the k observations $x_1, x_2, x_3,\dots\dots, x_k$ in the complete sample of n observations and U(k) is calculated for every k = 1,2,3,……,n. The maximum of absolute values of $U(k)$ refers to the probable change point at k-th data. The approximate probability for a two-sided test is given by

$$p = 2exp\left(\frac{-6 * \left(max\left(\left|U(k)\right|\right)\right)^2}{n^3 + n^2}\right)$$

Where, the approximate probability is good for p ≤ 0.5 [6]. The probability or critical values for the test statistic also can be estimated using Monte Carlo simulation.

Standard Normal Homogeneity Test (SNHT)

Standard Normal Homogeneity Test (SNHT) is based on the Ratio Test Method [7]. This method is best suitable to detect non-homogeneity near the beginning and end of the series [8]. The T(k) is calculated by comparing the mean of the first k data of the record with the last n-k data as follows:

$$T(k) = k\overline{z_1}^2 + (n-k)\overline{z_2}^2$$

Where,

$$\overline{z_1} = \frac{1}{k}\sum_{i=1}^{k}\frac{x_i - \overline{x}}{S}$$

$$\overline{z_2} = \frac{1}{n-k}\sum_{i=k+1}^{n}\frac{x_i - \overline{x}}{S}$$

$$\overline{x} = mean = \frac{\sum_{i=1}^{n} x_n}{n}$$

$$S = sample\,standard\,deviation = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$n = number\,of\,sample\,data$$

The T(k) reaches its maximum value when a breakpoint is detected at the data point K. The test statistic T_0 is defined as:

$$T_0 = \max(T(k))$$

The null hypothesis will be rejected if T_0 is above a certain level, which is estimated using Monte Carlo simulation.

Buishand's Test

In 1982, Buishand proposed a homogeneity test method based on adjusted partial sums [9]. The test statistic is given below:

$$S(k) = \sum_{i=1}^{k}\frac{x_i - \overline{x}}{\sigma}$$

Where,

$$\sigma = standard\,deviation = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2}$$

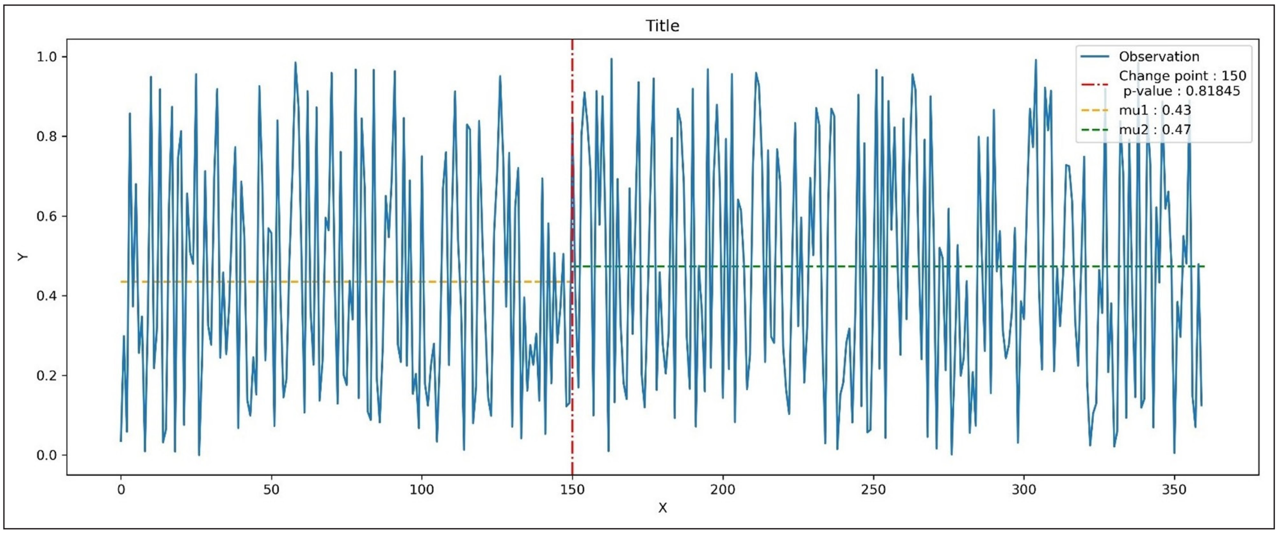

Figure 1 Homogeneity result plot.

The maximum of the absolute values of $S(k)$ is referred to as the probable change point at k-th data.

Buishand proposed four ways to check the sensitivity of this homogeneity test. These are:

Q test

In this method, Q is calculated using the equation given below and critical values for the test statistic are obtained from the table by Buishand [9] or using Monte Carlo simulation.

$$Q = \frac{\max\left(S(k)\right)}{\sqrt{n}}$$

Range test

In this method, Range R is calculated using the equation below. Critical values for the test statistic can be derived from the table by Buishand [9] or using Monte Carlo simulation.

$$R = \frac{\max\left(S(k)\right) - \min\left(S(k)\right)}{\sqrt{n}}$$

Likelihood Ratio test

The test statistic V(k) is calculated from the equation below [10]. Critical values for the test statistic are derived from the Monte Carlo simulation.

$$V = max\left(\frac{\left|S(k)\right|}{\sqrt{k(n-k)}}\right)$$

U Test

According to Buishand, U statistic is a robust test and good for detecting change point in the middle of a series [10]. The U statistic is calculated using the equation below. At the same time, critical values for the test statistic can be found in the table given by Buishand [9] or using Monte Carlo simulation.

$$U = \frac{1}{n(n+1)} \sum_{k=1}^{n-1} S(k)^2$$

EXAMPLE

A quick example of pyHomogeneity usage is given below.

```
import numpy as np
import pyhomogeneity as hg

# Data generation for analysis
data = np.random.rand(360,1)

result = hg.pettitt_test(data)
print(result)
```

Output is like this:

```
Pettitt_Test(h=False, cp=89, p=0.1428, U=3811.0,
avg=mean(mu1=0.5487521427805625,
mu2=0.46884198890609463))
```

Whereas, the output is a named tuple, so user can call by name for specific result:

```
print(result.cp)
print(result.avg.mu1)
```

or, user can directly unpack the results like this:

```
h, cp, p, U, mu = hg.pettitt_test(x, 0.05)
```

Users can plot results by following (Figure 1):

```
mn = 0
mx = len(data)
loc = result.cp
mu1 = result.avg.mu1
mu2 = result.avg.mu2

plt.figure(figsize=(16,6))
plt.plot(data, label=”Observation”)
plt.hlines(mu1, xmin=mn, xmax=loc, linestyles='--',
colors='orange',lw=1.5, label='mu1 : ' +
str(round(mu1,2)))
plt.hlines(mu2, xmin=loc, xmax=mx,
linestyles='--', colors='g', lw=1.5, label='mu2 : ' +
str(round(mu2,2)))
plt.axvline(x=loc, linestyle='-.' , color='red', lw=1.5,
label='Change point : '+ str(loc) + '\n p-value : ' +
str(result.p))

plt.title('Title')
plt.xlabel('X')
plt.ylabel('Y')
plt.legend(loc='upper right')
plt.savefig("F:/homogeneiry_results_plot.jpg",
dpi=600)
```

Users can find more examples in pyHomogeneity's Github repository's example section.

QUALITY CONTROL

Tests for pyHomogeneity package are performed using some fixed random data, where the results of those data are known. So, the performance of the functions is easily determined by comparing the output of the functions with the known results. Anyone can perform the unittest locally by using the below command in the root of the local copy of pyHomogeneity:

```
pytest –v
```

In addition, pyHomongeneity uses the continuous integration (CI) platform Travis CI for automatic testing after each change of the code base uploaded to the source repository [11]. The tests on Travis CI have been performed using different Python versions (2.7, 3.4., 3.5, 3.6, 3.7, 3.8) on Linux system. Users may raise issues on GitHub for additional support on using the package. Anyone can also contribute to this package. The contributor guideline can be found in the Contribution section on Github.

(2) AVAILABILITY

OPERATING SYSTEM

This package is platform independent, so it can be run on any operating system (GNU/Linux, Mac OSX, Windows) where python can be run.

PROGRAMMING LANGUAGE

Python 2.7 and 3.4+

ADDITIONAL SYSTEM REQUIREMENTS

None

DEPENDENCIES

pyHomogeneity is written using core python packages. Only Numpy and Scipy are required to use it.

SOFTWARE LOCATION

Archive

Name: Zenodo
Persistent identifier: http://doi.org/10.5281/zenodo.3785287
Licence: MIT
Publisher: Md. Manjurul Hussain Shourov
Version published: 1.1 and earlier versions. The DOI above always resolves to the latest version, previous versions can be identified with separate DOIs (see versions sections on the Zenodo repository page).
Date published: 04/05/2020

Code repository

Name: Github
Identifier: https://github.com/mmhs013/pyHomogeneity
Licence: MIT
Date published: 04/05/2020

LANGUAGE

English

(3) REUSE POTENTIAL

pyHomogeneity is a Python package, a widely used and freely available programming language. Because the package is for Python, it is platform-independent and therefore can be used by the majority of individuals in the data science community. It is a statistical analysis tool that performs different types of homogeneity tests for time series data. So, it can be used for data quality tests for study or academic research purposes. Many researchers have already started to use pyHomogeneity package in their research [12, 13, 14, 15].

Every function has docstrings to ensure clarity about what each function does and available options that the user can declare. The user documentation of pyHomogeneity is hosted on GitHub repository. The documentation contains some sample examples that can be easily modified for different user scenarios. pyHomogeneity is released under the MIT license and welcomes any contributions. We encourage users to submit feedback using GitHub issue tracker, or by emailing the authors.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Md. Manjurul Hussain orcid.org/0000-0002-5361-0633
Institute of Water and Flood Management, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

Ishtiak Mahmud orcid.org/0000-0002-4753-5403
Shahjalal University of Science and Technology, Sylhet, Bangladesh

Sheikh Hefzul Bari orcid.org/0000-0003-2635-2146
Graduate School of Symbiotic Systems Science and Technology, Fukushima University, Fukushima, Japan

REFERENCES

1. XLSTAT: The Leading Data analysis and statistical solution for Microsoft Excel. [Online]. Available: https://www.xlstat.com/en/.
2. **Pohlert T.** Trend: Non-parametric trend tests and change-point detection; 2016. Available: https://cran.r-project.org/web/packages/trend/index.html.
3. **Dey P.** Pettitt Change point test for univariate time series data. *MATLAB Central File Exchange;* 2022. Available: https://www.mathworks.com/matlabcentral/fileexchange/60973-pettitt-change-point-test-for-univariate-time-series-data.
4. NumPy, fundamental package for scientific computing with Python. [Online]. Available: http://www.numpy.org/.
5. SciPy, Python-based ecosystem of open-source software. [Online]. Available: https://www.scipy.org/.
6. **Pettitt A.** A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 1979; 28(2): 126–135. DOI: https://doi.org/10.2307/2346729
7. **Alexandersson H.** A homogeneity test applied to precipitation data. *Journal of Climatology*. 1986; 6(6): 661–675. DOI: https://doi.org/10.1002/joc.3370060607
8. **Mahmud I, Bari SH, Hussain MM, Rahman MT.** Homogeneity of rainfall and temparature series in bangladesh. *International Conference on Climate Change and Water Security*; 2015. DOI: https://doi.org/10.13140/RG.2.1.4431.3688
9. **Buishand TA.** Some methods for testing the homogeneity of rainfall records. *Journal of Hydrology*. 1982; 58(1–2): 11–27. DOI: https://doi.org/10.1016/0022-1694(82)90066-X
10. **Buishand TA.** Tests for detecting a shift in the mean of hydrological time series. *Journal of Hydrology*. 1984; 73(1–2): 51–69. DOI: https://doi.org/10.1016/0022-1694(84)90032-5
11. Travis CI, continuous integration service used to build and test software projects hosted on GitHub and Bitbucket. [Online] Available: https://www.travis-ci.com/.
12. **Talukder MR.** Analyzing trend and prediction of precipitation in Germany using non-parametrical tests and machine learning algorithms. *ScienceOpen Preprints*; Nov 21 2021. DOI: https://doi.org/10.14293/S2199-1006.1.SOR-.PPCPS00.v1
13. **Parajuli B, Zhang X, Deuja S, Liu Y.** Regional and Seasonal Precipitation and Drought Trends in Ganga–Brahmaputra Basin. *Water*. Jan 2021; 13(16): 2218. DOI: https://doi.org/10.3390/w13162218
14. **Li ZH, Wang JX, Lu M, Zhang T, Wang XC, Li WW, Yu HQ.** Hospital sewage treatment facilities witness the fighting against the COVID-19 pandemic. *Journal of environmental management*. May 1 2022; 309: 114728. DOI: https://doi.org/10.1016/j.jenvman.2022.114728
15. **Dibike Y, Hartmann J, de Rham L, Beltaos S, Peters DL, Bonsal B.** Exploratory Data Analysis of the Canadian River Ice Database Variables and their Correlations with Seasonal Temperature. *21st Workshop on the Hydraulics of Ice Covered Rivers Saskatoon*, Saskatchewan, Canada; 2021.

TO CITE THIS ARTICLE:
Hussain MM, Mahmud I, Bari SH 2023 pyHomogeneity: A Python Package for Homogeneity Test of Time Series Data. *Journal of Open Research Software,* 11: 4. DOI: https://doi.org/10.5334/jors.427

Submitted: 17 April 2022 **Accepted:** 25 January 2023 **Published:** 14 February 2023

Journal of Open Research Software is a peer-reviewed open access journal published by Ubiquity Press.