Multivariate Histogram Comparison

S.I. BITYUKOV, V.V. SMIRNOVA (IHEP, Protvino),
N.V. KRASNIKOV (INR RAS, Moscow)

IML LHC Machine Learning WG Meeting
May 17, 2016
CERN, Switzerland

Histogram

Suppose, there is given a set of *non-overlapping intervals*.

A histogram represents the frequency distribution of data which populates those intervals. This distribution is obtained during data processing of the sample (which is taken from the flow of *events*) with observed values of random variable). These intervals usually are called as *bins*.

The filling: two extreme cases: ***a***) one event - one histogram,

b) one event – one value is put to histogram.

The filling of histogram in the case ***b*** is a chain of *independent* measurements with gradual filling of the histogram.

We consider the multivariate approach which is proposed in the method of *Statistical Comparison of Histograms (SCH)* (arXiv:1302.2651, EPJ+ 128 (2013) 143).

Distance between histograms

Given two histograms, how do we assess whether they are similar or not? What does it means "similar"? Several standard procedures exist for this task.

Suppose, a *reference histogram* is known. Usually, the proximity of *test histogram* and reference histogram is measured via a test statistics, that provides the quantitative expression of the ``distance" between histograms. The smaller the distance the more similar they are.

There are several definitions of distance in the literature (Int. Journ. of economics and statistics, 4 (2016) 98), for example, the Kolmogorov distance, the Kullback-Leibner, the chi-square distance and so on. Usually, it is the some test statistics, distribution of which can be calculated via formulae or *constructed by Monte Carlo*.

Distinguishability or consistency

Often a goal of histograms comparison is a testing of their *consistency*. Consistency here is the statement that both histograms are produced during data processing of independent samples which are taken *from the same flow of events* (or from the same population of events).

The method of statistical comparison of histograms (SCH) allows to estimate the distinguishability of histograms and, correspondingly, the distinguishability of parent events flows (or parent samples).

We use the distribution of some test statistics (*significances of difference*) instead of single test statistics in other methods. This distribution has statistical moments (the mean, RMS, asymmetry, excess, ...), i.e. the distribution can be considered as ***multivariate test statistics*** with, for example, the mean and RMS as two coordinates and, in principle, any other univariate test statistics as additional coordinates.

Significance of the difference I

Suppose there are two histograms *hist1* and *hist2* (with *M* bins) which produced during processing of two samples:

hist1: $\hat{n}_{11}\pm\hat{\sigma}_{11}, \hat{n}_{21}\pm\hat{\sigma}_{21}, \ldots, \hat{n}_{M1} \pm \hat{\sigma}_{M1}$;

hist2: $\hat{n}_{12}\pm\hat{\sigma}_{12}, \hat{n}_{22}\pm\hat{\sigma}_{22}, \ldots, \hat{n}_{M2} \pm \hat{\sigma}_{M2}$.

Then significance of difference for bin # *i*, *i=1,M* is calculated as

$$\hat{S}_i = \frac{\hat{n}_{i1} - K\,\hat{n}_{i2}}{\sqrt{\hat{\sigma}_{i1}{}^2 + K^2\,\hat{\sigma}_{i2}{}^2}}$$, where *K* is a coefficient of normalization.

Each of these *M* test statistics $\hat{S}_i$ obeys the distribution which close to standard normal distribution *N(0,1)* if both *independent* samples are taken from the same flow of events. It means that distribution of *M* values $\hat{S}_i$ also must be close to *N(0,1).*

"God made man, but Samuel Colt made them equal"

This famous slogan can be rephrased for propose method as

"God made bins, but significance of difference made them equal".

Significance of the difference II

As a result we can calculate statistical moments of this distribution and, in principle, we have information about distinguishability of samples under testing.

In example below we use only two moments: the mean value $\bar{S}$ and the *rms* of this distribution, i.e. *bidimentional test statistic*

$SRMS = (\bar{S}, rms)$.

In ideal case:

if $SRMS = (0,0)$, then histograms are identical (often it is bad);

if $SRMS \approx (0,1)$, then samples are taken from the same flow of events;

if the previous conditions are not valid, then origin flows of events have difference.

Hypotheses testing I

If the goal of the comparison of histograms is the check of their consistency, then task is reduced to hypotheses testing: *main hypothesis H0* (histograms are produced during data processing of samples taken from the same flow of events) against *alternative hypothesis H1* (histograms are produced during data processing of samples taken from different flows of events).

The determination of *critical area* allows to estimate *Type I error* (α) and *Type II error* (β) in decision about choice between *H0* and *H1*.

The *Type I error* (α) is a probability of mistake if done choice is *H1*, but *H0* is true.

The *Type II error* (β) is a probability of mistake if done choice is *H0*, but *H1* is true.

Hypotheses testing II

The selection of a *significance level* ***(α)*** allows to estimate the *power of the test* ***(1-β)***. Usually, values of significance level are *10%*, *5%*, *1%*.

If both hypotheses are equivalent, then other combinations of the ***α*** and ***β*** are used.

For example, in task about distinguishability of the flows of events works a *relative uncertainty* ***(α+β)/(2-(α+β))*** for ***α+β≤1***.

Under the *test of equal tails* the mean error ***(α+β)/2*** can be used too.

The hypotheses testing require the knowledge of the distribution of test statistics. As mentioned above the distribution of test statistics can be constructed by Monte Carlo.

Rehistogramming I

If *errors of values in bins* of at least one of histogram *are known* (for example, reference histogram), than one can construct the set of similar histograms (clones), which imitates the population of histograms which produced due to data processing of the samples taken from the same flow of events.

This set of histograms is used for construction of the distribution of test statistics for the case of *H0* hypothesis (due to comparison of the reference histogram and the produced clones - some kind of calibration).

This procedure can be named as "*rehistogramming*" in analogy with "*resampling*" in bootstrap technique.

The correctness of this procedure follows, for example, from the properties of the statistically dual distributions (Applied Mathematics, 5 (2014) 963): normal with fixed variance, Poisson and Gamma, Cauchy, Laplace, … .

Rehistogramming II

Further the set of histograms of such type is constructed for test histogram (second histogram).

New set is used for construction of the distribution of test statistics for the case of *H1* hypothesis (due to comparison of the reference histogram and the produced clones of second histogram).

The comparison of the distribution of test statistics for the case of *H0* hypothesis and the distribution of test statistics for the case of *H1* hypothesis allows to estimate the uncertainty in hypotheses testing.

Advantages of this approach

1. We have a measure of the “distance” between histograms. It is *relative uncertainty of the decision* about consistence or distinguishability of histograms.
2. We can compare multidimensional histograms likewise as unidimensional histograms.
3. We can compare two sets of several histograms simultaneously likewise as we compare a pair of histograms.
4. We can use any unidimensional test statistics (Kolmogorov-Smirnov, Anderson-Darling, …) as additional dimension in proposed multivariate test statistics.

Let us consider the example of multivariate comparison.

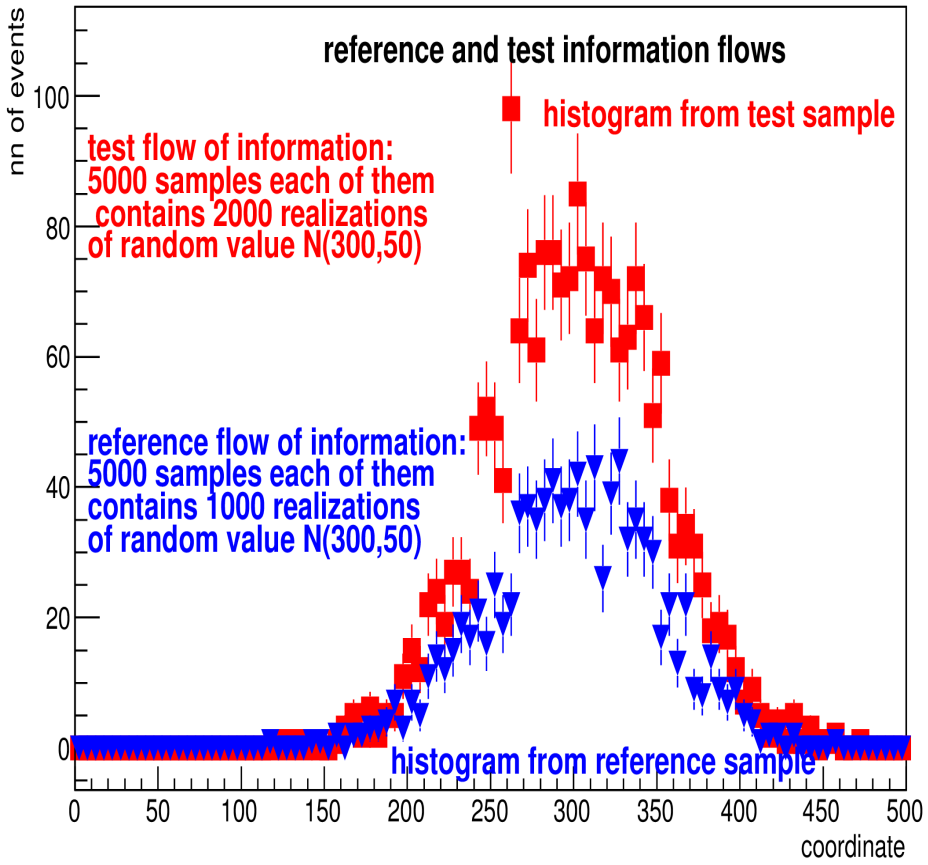
Monte Carlo experiment

Two pairs (reference pair and test pair) of independent flows of samples with realizations of random variables (each realization is ``event'') are produced to estimate the possibility of methods for distinguishing of samples from different information flows.

The volume of each flow equals 5000 samples.

First flow from each pair is a reference flow of samples with 1000 events (1000 realizations of random variable $N(300,50)$).
Second flow from first pair also is flow of samples with 2000 events (realizations of the same random variable) (Fig.A, left).
Second flow from second pair is test flow of samples with 2000 of events (realizations of random variable $N(300,44)$). The examples of distributions in samples are shown in Fig.A (right).

Examples of distributions in samples



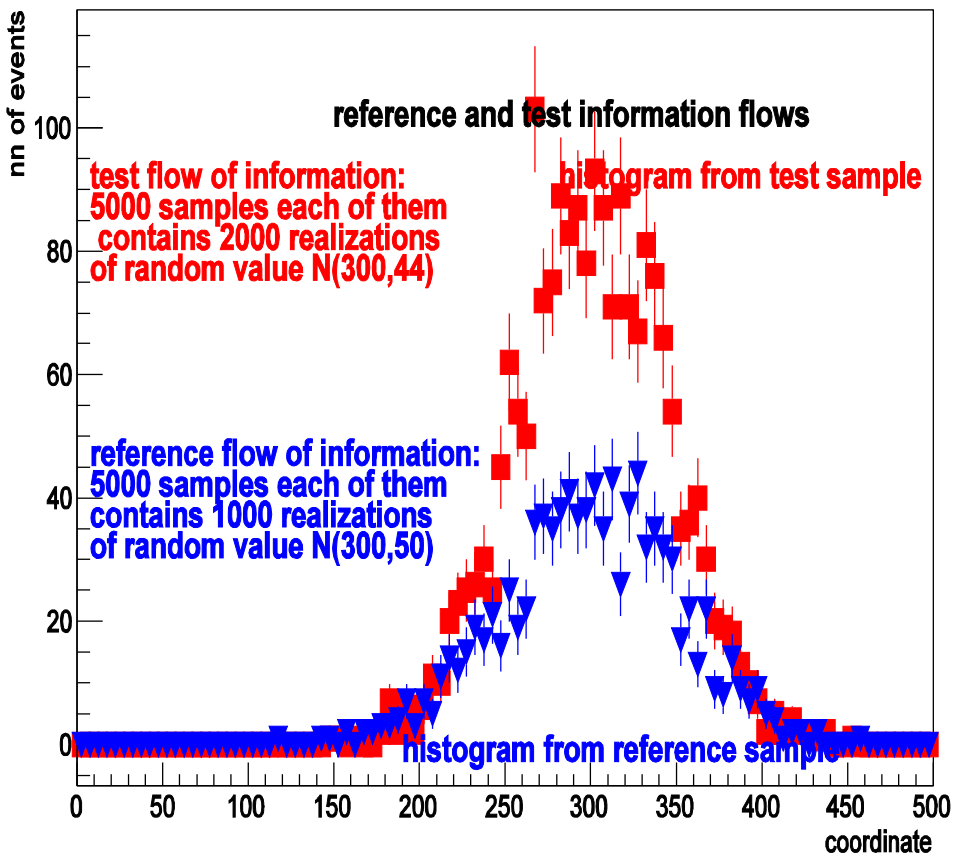


Fig.A Examples of distributions from two pairs of flows: left histogram - 1000 realizations of reference random variable *N(300,50)* and 2000 realizations of test random variable *N(300,50)*), right histogram - 1000 realizations of reference random variable *N(300,50)* and 2000 realizations of test random variable *N(300,44)*).

SCH approach

In the case of Statistical Comparison of Histograms (SCH) method for each sample is constructed the histogram. After that for each pair of samples the comparison of histograms is performed with calculation of the mean value *(<S>)* of *significances of the difference* between corresponding bins of histograms and *root mean square* (*rms*) of the distribution of these ``significances".

The distribution of bidimensional values *(<S>,rms)* is constructed for each pair of samples from both pair of flows (see, Fig.B). After that the critical line (for equal tailed test) is calculated. Correspondingly, the values α (*Type I error*) and β (*Type II error*) are determined.

The quality of the distinguishing is estimated as a *relative uncertainty*

$$(\alpha+\beta)/(2-(\alpha+\beta))\ .$$

Hypotheses testing (SCH test)

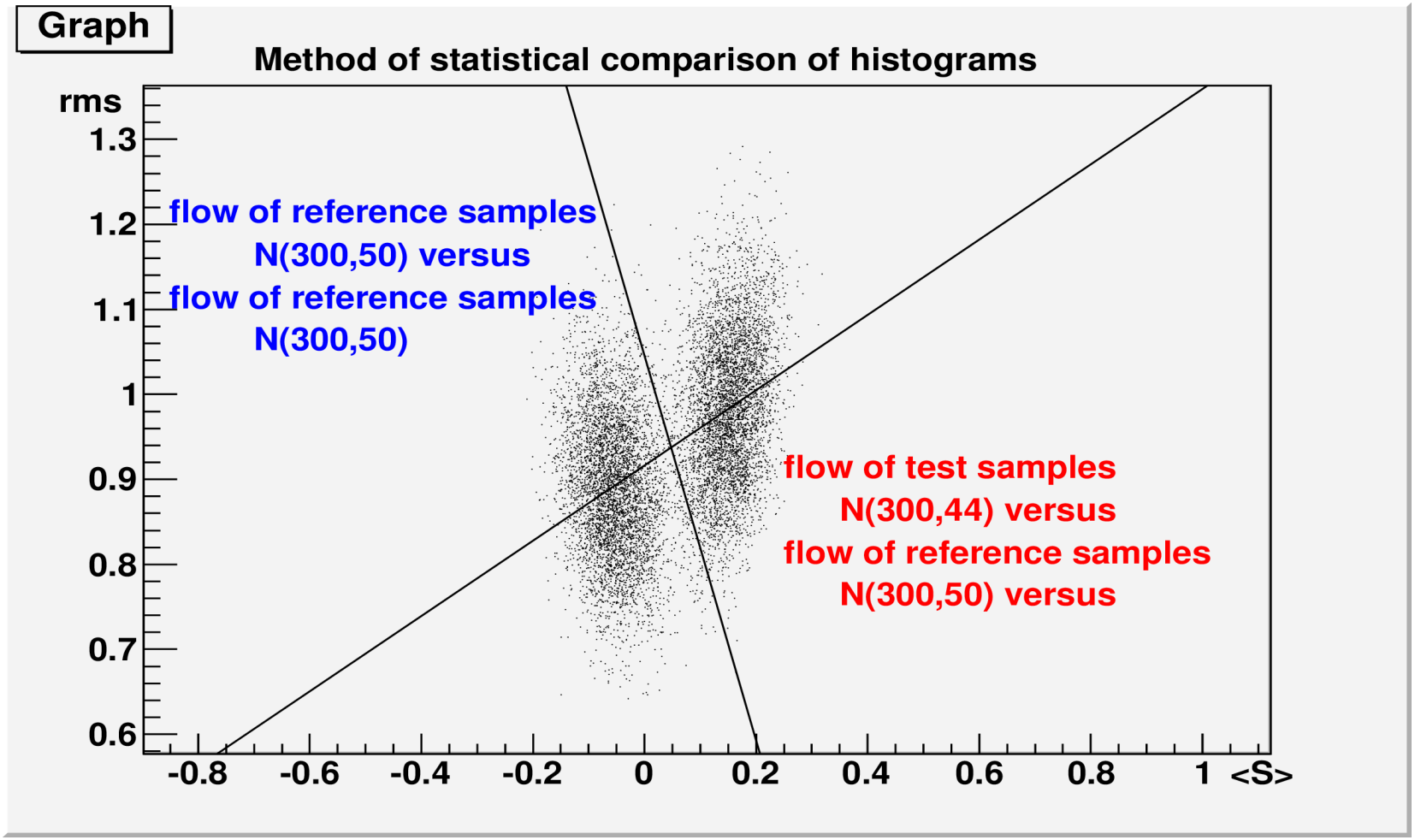


Fig.B Two distributions *(<S>,rms)* values for 5000 comparisons: the case of the comparison of samples from two similar flows (reference flow) and the case of the comparison of samples from reference flow and samples from test *N(300,44)* flow.

Conclusion

The possibility of the use of the multivariate test statistics for the comparative analysis of histograms (or dependences) in frame of the statistical comparison method is considered.

This method allows to use the multivariate test statistics for determination of the “distance” between histograms.

This method allows to compare as multidimensional histograms and sets of histograms.

In principle, the multivariate test-statistics can include any univariate test statistics as an additional coordinates.

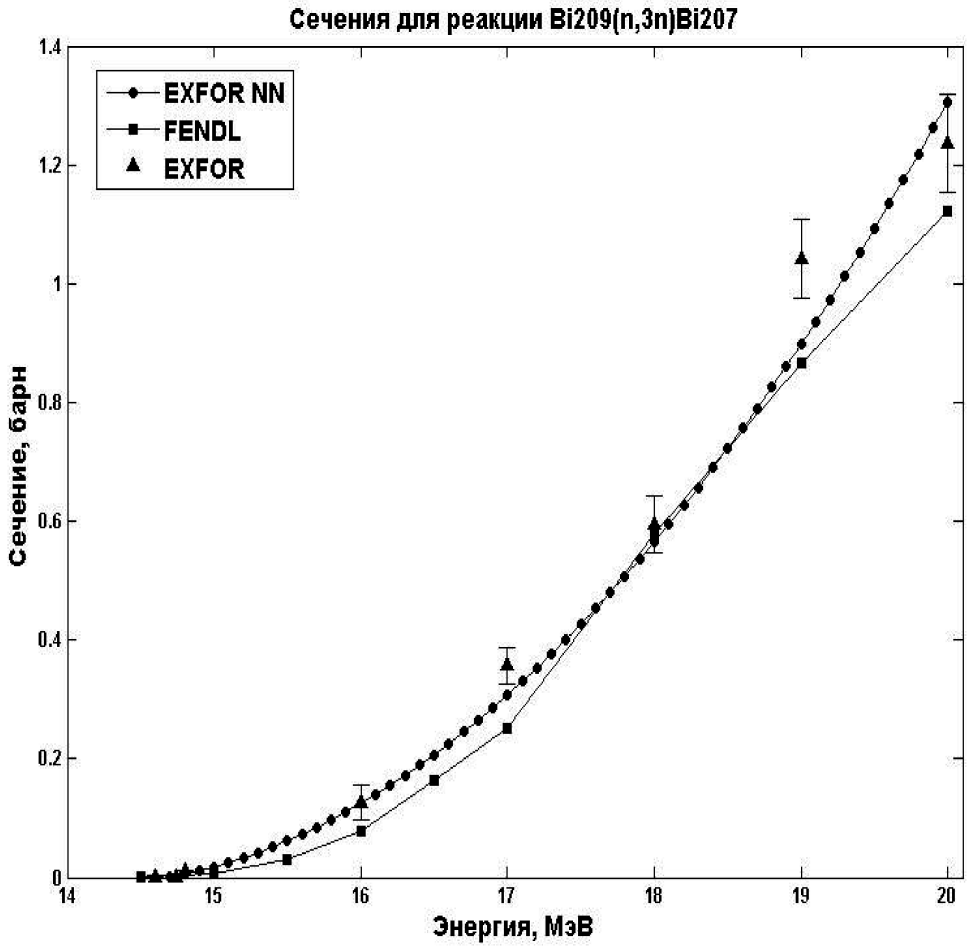
Acknowledgements

The authors are grateful to Prof. V. Kachanov, Prof. Yu. Korovin, Dr. S. Gleyzer, Dr. L. Moneta, Dr. N. Korneeva for helpful discussions.

The work was supported by the Ministry of Education and Science RF (Agreement on October, 17, 2014 N 14.610.21.0004, id. PNIER RFMEFI61014X0004).

Backup slide

Statistical comparison of data sets



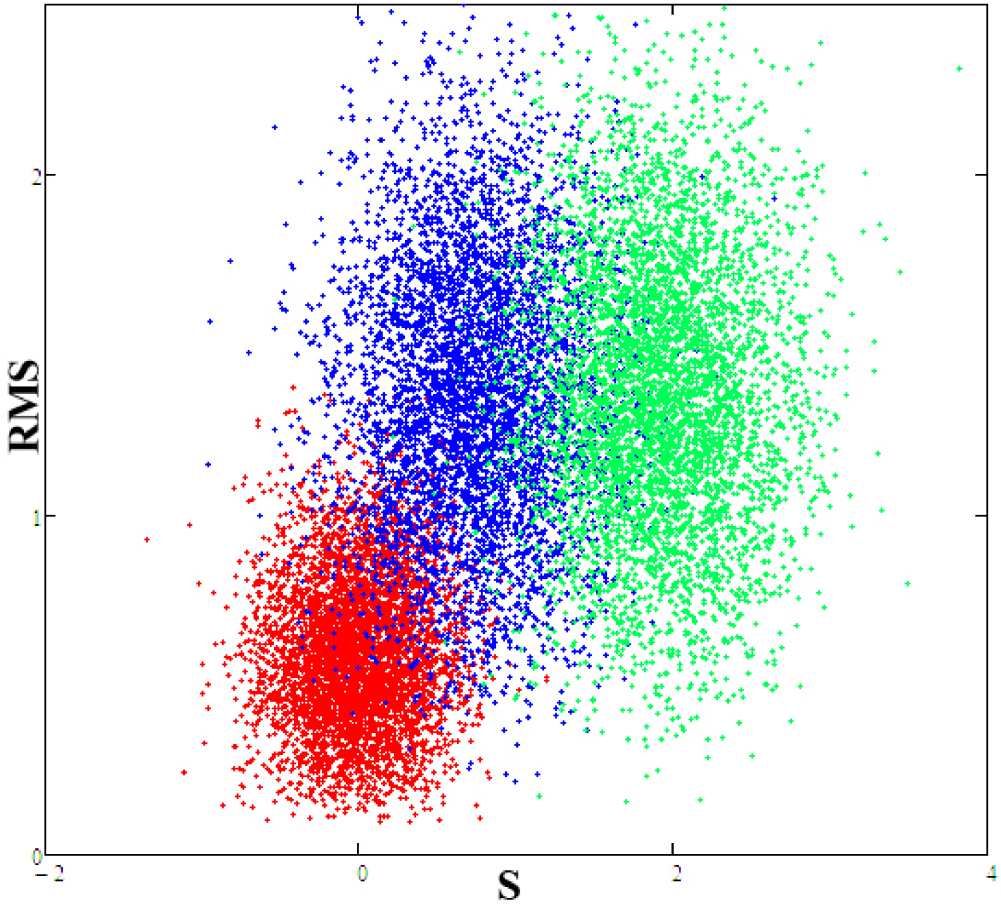


Fig.C The comparison of the experimental data (triangles with errors and red spot) with results obtained in frame of two models (blue and green spots)
[A. Maksimushkina, V. Smirnova, Method for data statistical visualization, Scientific Visualization, 7, Issue 5 (2015) 26-37].