The representativeness of the statistical data frame in the quantitative image analysis

Vladimir N. Petrushin [1,*], *Yuriy V.* Rudyak [2], and *Georgy O.* Rytikov [2]

[1] Institute of Contemporary Arts, 121309 Moscow, Russia
[2] Moscow Polytechnic University, 107023 Moscow, Russia

Abstract. In the article the concept of representativeness of the statistical data frame is discussed. Also the statistical measure of representativeness and a method for the quantitative assessment of the degree of representativeness of the general population are proposed. To demonstrate the techniques and methods of application, we consider a case of identifying the amount of statistically reliable sampling in the analysis of polymer composite surface's SEM-images.

1 Introduction

Use A4 paper size (210 x 297 mm) and adjust the margins to those shown in Table 1. The final printed area will be 172 x 252 mm.

Affiliations of authors should be typed in 9-point Times. They should be preceded by a numerical superscript corresponding to the same superscript after the name of the author concerned. Please ensure that affiliations are as full and complete as possible and include the country.

In the description and modeling of the scientific and industrial research results, the statistical approach is one of the best known and most widely used. The combination of quantitative methods of the statistical approach relies primarily on the conceptual apparatus and many techniques to apply elements of probability theory and mathematical statistics, combined with the concept of "applied statistics". The basic concepts of applied statistics are general population and data frame [1-4].

Due to the classic epistemological problem (the inability of the formal definition of basic concepts), the unified and generally accepted definitions of General population and statistical data framing, apparently, do not exist. However, many researchers of the subject area successfully understand each other intuitively in the use of the discussed concepts. The general population is usually understood as all the "target" set of objects about which some insights or predictions on the results of statistical research should be received. The data frame is the set of objects which it is possible to measure and/or calculate and on the basis of that to draw certain statistical models. With the help of these models the conclusions are drawn and the predictions are made about the general population's «behavior».

Thus it is implicitly assumed that the models successfully describing the frame will also be effective to describe the population. All the criticism of the applied statistics' methods generally and these of the parametric statistics specifically is based on the impossibility of proving or disproving this assumption. In order to remove the problem under discussion, when statistically analyzing the data in different ways, the concept of "evaluation of the data frame representativeness» is introduced. It postulates that in the process of the quantitative statistical models formation a data frame is used which is so similar to the general population that the results of the predictions for the frame and the population should be the same. Obviously, it is impossible to test this claim if the general population is not known. If the general population is known, it is not often necessary to use the statistical methods to describe it. However, predictions for the general population (the set of elements of which is not known) are not actually made, because the verification of predictions and models is always carried out only on a sample from the general population. In fact, the problem of quantitative statistical modeling and predicting in the most general form is as follows: it is required to obtain a model based on the results of some primary data frame analysis (or some aggregate thereof); the model is also expected to be sufficiently reliable from the point of view of the researcher, and efficient to describe the validation data frame (or some aggregate thereof) in case of randomness of frames and independence of measurement results from each other.

Thus, for the effective usage of applied statistics methods of data analyzing and predictions building, it is possible to formulate the concept of relative representativeness (as measured quantitatively on an interval [0;1]) of statistical data frame: primary data frame from the general population is representative, according to some criterion, relative to the validation frame, if the value of the corresponding criterion,

* Corresponding author: petrushinvn@mail.ru

calculated on the basis of statistical analysis of the primary and validation frames, takes the extreme value.

With that, the more powerful is the aggregation of validation frames for which the underlying frame is representative, the higher the quality of the statistical models generated on the basis of the primary frame is, on condition of minimizing the cost of obtaining the primary frame (minimize the amount of the primary frame and the power of the aggregation of the primary frames analyzed to determine the statistical model).

It is important to note that the above concept as presented relates mainly to the analysis of stationary random variables. In random processes' analysis (in particular, non-stationary), it requires the significant additions and refinements (for example, you may need to evaluate the results of dynamic statistical analysis of a large number of repeated experiments), considered by the authors as the problems for further research.

There are a lot of approaches and criteria [5–12], used to evaluate the representativeness of the data frame relative to the general population (we will call this theoretical representativeness «absolute» one). All the existing criteria require statistical calculation on the frame. It is obvious that the power of the criteria based on this principle, is substantially less than the power of the criteria based on the comparison of the empirical density functions and the empirical functions of probability distributions, as for all additional calculations of statistical indicators the number of degrees of freedom decreases significantly. The main advantage of the approach associated with the comparison of empirical distributions is that there is no need to justify the choice of theoretical distribution, presumably characterizing the general population.

The result of the comparison of empirical distributions is a quantitative assessment of the relative representativeness of the sample. The concept of relative representativeness application allows to combine the advantages of private and general criteria of representativeness, because no assumption is done on the nature of the general population data distribution, no values of any "alleged" criteria are calculated, but the empirical distributions of the primary and validation data frames are built and compared according to the criteria .

From our point of view, one of the "closest to reality" criteria of representativeness is the overlap area of the histograms of the base and validation data frames' empirical distributions. However, for simplicity of calculations it is traditional to apply the sum of squares of relative frequencies deviations to find the volume of a representative frame, in fact, using the method of the least squares. It is obvious that the maximum overlap of the areas of the histograms of empirical distributions corresponds to the minimum of the sum of the squares of their mutual deviations.

2 The data for the computational experiment

For a generalized description of the structure of the experimental data involved in the statistical analysis and processing, in most cases, the model of multidimensional tensor is ample, projected onto the space of variables in which the problem is stated and the results of the statistical analysis are formulated. Since visualization of more than a three-dimensional space causes some difficulties, we will consider an example of data analysis described within the traditional three-dimensional Cartesian coordinate system.

As a test data we consider the table of values of the image pixels brightness formed by a scanning electron microscope in the study of the sub-micro relief (topographic morphology) of low density polyethylene membrane surface modified by the method of gas phase sulfonation (see Fig. 1).

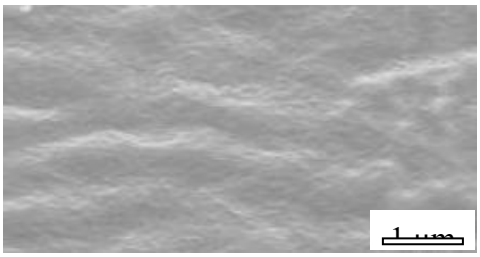

Fig. 1. The investigated fragment of a SEM-image of sample surface of LDPE modified [13] with the method of gas phase sulfonation [14].

When you visualize this data in the Cartesian coordinate system, along the axes of abscissa and ordinate the coordinates of the pixel will be plotted , and its brightness will be along the axis of the applicat. Despite the fact that when describing a real surface there is no reason to suppose that the corresponding coordinates take only integer values, the measuring procedure using the scanning electron microscope automatically discretizes the measurements. As a result, when visualizing, along the axes of the coordinate system only natural numbers (numbers of columns and rows of the data table on the axes of abscissa and ordinate, and the values of the brightness levels of the pixels range from 0..255 on the axis of the applicat) are plotted. One of the features of this example is that it is a typical situation when in order to describe continuous variables a limited number of discrete values of these variables are used (results of experimental measurements or observations).

It is also convenient that the main objective of the statistical analysis of these SEM images is the identification of submicrometer structures (regular or stochastic [15, 16]) on the investigated surfaces; these structures are obviously smaller than the entire field of view of the device, but they provide typical values of those or other physics-chemical and/or functional properties of the tested materials (What is a "molecule" according to the original definition? It's such a tiny part of matter that is characterized by the same set of chemical properties, as the substance in general. With the modern refinements, in our case it is more about something like a domain or macromolecules, as for individual molecules there are many size effects arising in the analysis of chemical reactions on monomolecular

level, but "large" (tens of thousands or more) conglomerates of molecules "behave" like the substance in general).

In terms of the definition proposed in the Introduction, the General population is the totality of all possible SEM-images of surfaces of the investigated class materials (i.e., as expected, unattainable large amount of measurements); the validation data frame is the table of values pixels' brightnesses describing one specific analyzed SEM-image, and possible primary data frime is made up of a subset of the validation frame (i.e., they are fragments of the investigated SEM-image). Identification of the primary data frame volume, characterized by the same empirical distribution of pixel brightness as the verifying data frame, allows us to estimate the characteristic sizes of the regular or stochastic structures, making a decisive contribution to the formation of those or other physics-chemical and functional properties of all the studied material. Thus, an abstract mathematical problem about finding the way to assess the degree of representativeness of data frames is correlated with a specific area of application in chemistry and materials science (an important problem is solved in the formation of structural-functional model of the material associated with the assessment of the characteristic dimensions of the structural domains of the material surface).

3 The results of the computational experiment

For the SEM-images presented in Fig.1 we built the histograms of the empirical pixels' brightnesses distributions for a series of 40 primary data frames (Fig.2 shows a typical histogram in integral and differential forms of representation) and the histogram of the empirical pixels' brightnesses distribution for the all analyzed image (see Fig.3).

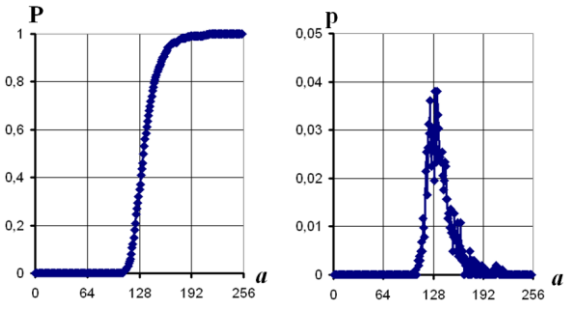

Fig. 2. The distribution function and the probability density function that characterize pixel brightness in the primary data frame of 32x32 elements.

The integral functions of the empirical distributions, in this case, seem to be preferable due to the fact that they make it possible to use the traditional parametric mathematical statistics approaches like minimizing the sum of squared deviations (or the sum of absolute deviations) in the representativeness analysis. Although it is also possible to carry out the appropriate calculations requiring the statistical proximity for the probability densities if it is necessary.

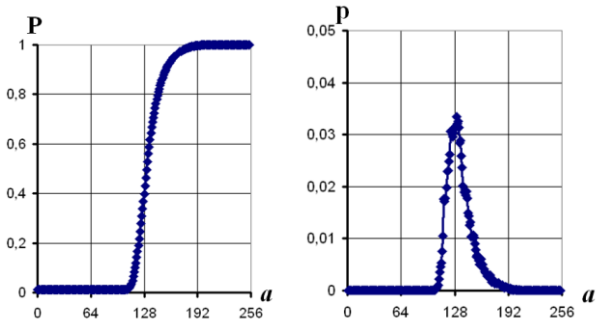

Fig. 3. The distribution function and the probability density function that characterize the pixel brightness in the validation data frame of 220×152 elements.

Let us consider, for example, the sum of the squares of mutual deviations for the empirical distributions that characterize various primary and validation data frames.

A trivial solution for the squares sum minimizing is to match the primary and the validation data frames, which guarantee the zero value for the sum. However, this decision does not answer our purpose, because we want to find such subsets of the considered set that the distributions would coincide as much as possible and the capacity of the subsets would be minimal.

As the empirical distribution for the validation data frame is known, and the empirical distributions for the primary data frame is calculated numerically, we present (Fig. 4) the fragment of a chart based on the sum of the squares of the empirical distributions deviations for the primary and the validation data frames as function of the volume of the primary data frame. (The low volume zone is clipped because there is absolutely no sense to talk about the distribution for the sample if it is less than 4×4 elements, and the high volume zone is clipped because it is clear that there will be 0 «at the end». And what we need to find is the local extrema corresponding to the smallest possible sample size).

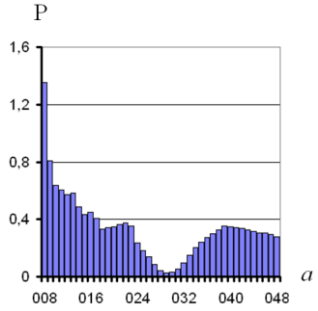

Fig. 4. The sum of squared deviations for the probability distribution functions of the primary and the validation data frames as function of the frame size (in pixels).

In our example, the primary data frame is formed in the shape of a square, as the method of the source material processing does not imply any anisotropy, which is confirmed by the preliminary experiments in

wettability and other physics - chemical properties of the surface studied.

In this case it is convenient to enumerate the sample size with the number of pixels forming the side of the corresponding square.

It is clearly seen that the selected criterion of representativeness reaches a local minimum when the amount of the primary data frame is 29(±1)×29(±1), i.e. the 29×29-frame is representative for minimizing the differences in distribution functions relative to the 220×152-frame, which estimates the structural domain dimension of the studied surface. The presence of the local minimum can be interpreted as a sign of similarity of the empirical distributions of the primary and the validation samples. And we can consider it to be prerequisite for the fractal surface structures observation.

The observation of a local minimum at approximately 4-fold increasing of the primary data frame volume (56(±1)×56(±1)) and the observation of the corresponding local maxima when assessing the representativeness according to the Kolmogorov-Smirnov criteria (similar to [17]) (see Fig. 5.) are the indirect confirmation of the need for special methods for fractal structures detecting. These are also the direct evidence that the evaluation of the characteristic size of the structural domain is not random.

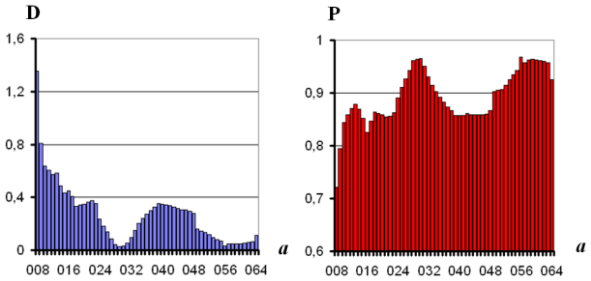

Fig. 5. The illustration of testing the hypothesis about the possibility of the fractal surface structures observation. The left figure illustrates the function of the sum of distributions' squared deviations for the primary and validation data frames. The right figure shows the primary data frame reliability (determined on the basis of the Kolmogorov-Smirnov test) relative to the validation one, as function of the frame volume parameter.

The numerical experiment results analysis (based on the application of Kolmogorov-Smirnov criteria) shows that the accuracy of the primary data frame relative to the validation one varies from 0.73 to 0.96 (when changing the data frame's size parameter in range of [8;64]) and reaches local maxima when the size is 30(±1)×30(±1), 56(±1)×56(±1), which indicates the coincidence, within the statistical error, of the interval values of the representative data frame size calculated with two different ways.

The incongruence of the representative data frame double volume value and the 4-fold increased relative data frame is due to the errors in digitizing the images and the applied statistical criteria peculiar properties. The corresponding function and distribution density of pixels' brightnesses are shown in Fig.6.

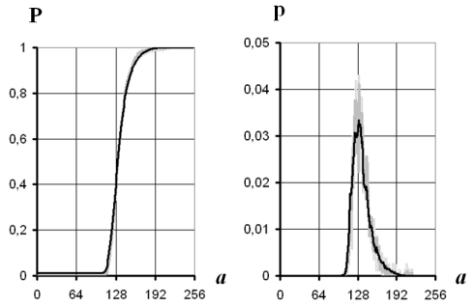

Fig. 6. The comparison of the pixels brightness distributions for the representative ("light grey line" - 29×29, "dark grey line" - "56×56"), and validation ("black line" - 220×152) samples.

To remove the problem under discussion, apparently, it is required to reconsider the structural domain on SEM-images with a larger zoom.

Generally speaking, there are several ways of primary data frame shaping. We have considered the square "from the upper left corner"; it is possible to choose the rectangles, the squares in various "locations" for a validation data frame; it is also possible to introduce the average distribution in the ensemble of samples of the same size, etc.

In the general case, the "shape" of the studied subarea of the image must be determined according to its physical aspects (for example, based on the consideration of symmetries, or on the basis of a priori information about any anisotropy) and (in multi-dimensional tasks) after the sham elimination of anisotropy with the help of scaling it can be considered as a hypercube. But this issue is not the subject of this article and will be investigated in future works.

It is important that we perform a tool for the quantitative evaluation and selection of the relative representativeness of data frames, and that a side effect is the estimation of the characteristic size of the structural domains in the analyzed images.

4 Conclusions

In this article the concept of relative representativeness of data frames is presented, as well as the method for its quantitative estimation. As the criterion of representativeness it is proposed to use the overlap areas of the histograms of probability functions empirical distributions of the primary and validation data frames, which will make it possible to assess the degree of representativeness of the base frame relative to the validation one by our proposed numerical criterion.

The example of application and demonstration of the physical meaning of the presented concept is the analysis of the gas-sulfonated membrane of low density polyethylene (LDPE) surface's SEM- images obtained by scanning electron microscopy. It is shown that the subset of 29(±1)×29(±1) pixels (with the value of the assessment of representativeness 0.96±0,02) is the relatively representative data frame of the analyzed

image, which allows us to estimate the characteristic sizes of the structural domain of the surface.

In the considered example, the maximum overlap area of the empirical distributions histograms for the primary (29×29 and 56×56) and the validation (220×152) data frames (in both cases) was 0.96(±0,02), indicating a high degree of relative representativeness of these primary data frames. The minimum overlap area of the histograms for the primary (8×8) and validation (220×152) samples amounted to 0.73, which is interpreted as a relative lack of representativeness of this (8×8) sample.

Evaluation of the characteristic dimensions of the sub-regions of images featuring the empirical distribution of pixels' brightnesses close to the pixel brightness distribution of all the considered images, suggests the possibility to observe the fractal structures, and points to the need for specialized, more sensitive and resource-intensive methods of identifying such structures in further research.

The work was performed with financial support from the RFBR grant 16-03-00540. The authors express their deep appreciation to the referees, the interpreter (M.V. Pluzhnik) and the editor for helpful and significant suggestions and tips.

References

1. M.J. Kendall, A. Stuart, *The theory of distributions* (Nauka, 1966)
2. M.J .Kendall, A. Stuart, *The statistical inference and communications* (Nauka, 1973)
3. M.J. Kendall, A. Stuart. *The multivariate statistical analysis and time series* (Nauka, 1976)
4. V.N. Petrushin, M.V. Ulyanov, *Information sensitivity of computer algorithms* (PhisMatLit, 2010).
5. A.I. Urazbahtin, I.G. Urazbakhtin, Information and communication technologies, **4**(3), 10 (2006)
6. V.A. Banakh, I.N. Smalikho, E.L. Pichugina, A. Brewer, Optics of atmosphere and ocean, **22**(10) 966 (2009)
7. B.F. Melnikov, S.V. Pivneva, O.A. Rogova, Stochastic optimization in informatics, **6**, 74-82 (2010)
8. F.N. Ilyasov, Sociological studies, **3**, 112-116 (2011)
9. O.A. Rogova, N.V. Sofonova, Heuristic algorithms and distributed computing, **1**(6), 58-75 (2014)
10. G.O. Rytikov, M.V. Ulianov, V.N. Petrushin, Double smoothing in time series formalization, *In 2014 International conference on computer technologies in physical and engineering applications (ICCTPEA)*, pp. 150-151 (IEEE, 2014)
11. T.P. Svetlova, Proceedings of the Main geophysical Observatory named by A.I. Voeikov, **579**, 115-128 (2015)
12. E.Yu. Ivanov, M.S. Kosyakov, Software products and systems, **4**(112), 198-202 (2015)
13. V.G. Nazarov, *Surface modification of polymers* (Moscow state University of Printing Arts, 2008)
14. V.G. Nazarov, V.P. Stolyarov, S.P. Molchanov, G.A. Jurassic, M.N. Artemenko, Polimer Science. Series A, **55** 1343 (2013)
15. E.S. Kopachev, S.A. Nozdrachev, V.N. Petrushin, Yu.V. Rudyak, G.O. Rytikov, V.G. Nazarov, Physical Mesomechanics, **18** 98 (2015)
16. V.N. Petrushin, Y.V. Rudyak, G.O. Rytikov, The holistic method of the surface structure characterization, *In 14th International Baltic Conference on Atomic Layer Deposition, BALD 2016 – Proceedings,* pp. 15-19 (2016)
17. V.N. Petrushin, S.A. Drozdov, G.O. Rytikov, Cloud of Science, **2** 247-264 (2015)