We would like to thank T. A. Azlarov, S. Rachev, and an anonymous referee for their attention to this research and useful comments.

LITERATURE CITED

1. E. J. Gumbel, "Bivariate exponential distributions," JASA, 55, No. 292, 698-707 (1960).
2. A. W. Marshall and I. Olkin, "A multivariate exponential distribution," JASA, 62, No. 317, 30-44 (1967).
3. J. E. Freund, "A bivariate extension of the exponential distribution," JASA, 56, No. 296, 971-977 (1961).
4. P. S. Puri and H. Rubin, "On a characterization of the family of distributions with constant multivariate failure rate," Ann. Prob., 2, No. 4, 738-740 (1974).
5. H. W. Block, "A characterization of a bivariate exponential distribution," Ann. Stat., 5, No. 4, 808-812 (1977).
6. L. Lee, "Multivariate distributions having Weibull properties," J. Multivar. Anal., 9, No. 2, 267-277 (1979).
7. P. Ferman, "Characterization of multivariate exponential distributions," Vestn. Mosk. Gos. Univ., Ser. Mekh. Mat., No. 4, 44-47 (1981).
8. P. Barlow and F. Proschan, Statistical Theory of Reliability an Life Testing, Holt, Reinhart and Winston, New York (1974).
9. B. Dimitrov, L. B. Klebanov, and S. Rachev, "Characterization stability of the exponential distribution," in: Stability Problems of Stochastic Models, Proceedings of a Seminar [in Russian], VNIISI, Moscow (1982), pp. 39-46.

NEW TECHNIQUES FOR COMPARING THE VOLUME FUNCTIONS OF HISTORICAL TEXTS

V. V. Kalashnikov, S. T. Rachev,
and A. T. Fomenko

We propose new techniques for estimating the degree of dependence of historical texts, such as annals, chronicles, etc. We consider texts "parametrized by time." This means that the text can be divided into a union of disjoint fragments, each describing the events of one year (or one decade, etc.). We also assume that the texts describe events over time intervals of the same length (say, a period of a few decades or centuries). Following [1], two texts X and Y are called dependent if they describe events over the same time interval and in the history of the same region, or have a common prototype. Dependent texts may have the same origin, rely on the same volume of archival data, or be versions of the same prototype. Texts are said to be independent if they describe events in essentially different time intervals (i.e., time intervals that intersect over not more than half their combined length) or describe events in different regions. It is relevant to consider techniques for estimating the degree of dependence of a pair of texts.

Consider a text X that describes events over the time interval from A to B (in some system of chronology). Let the parameter t run over the years from A to B. Represent the text X as the union of fragments X(t), where X(t) describes the events of one year t. Count the volume of the fragment X(t), e.g., in lines (or in pages, etc.). The result is a certain graph $f(X, t) = \mathrm{vol}\,X(t)$. Similarly construct the graph f(Y, t) for the text Y, which is also assumed to be given on the interval [A, B]. Identify the splash points (i.e., the points of local maxima) in the volume of the text X on the interval [A, B].

The following correlation principle of maximum points has been formulated and experimentally tested by Fomenko [1-3]:

1) If the texts X and Y are dependent, then the splashes in their volume functions occur virtually at the same time, i.e., the points of local maxima of the volume functions $\mathrm{vol}\,X(t)$ and $\mathrm{vol}\,Y(t)$ are correlated.

Translated from Problemy Ustoichivosti Stokhasticheskikh Modelei, Trudy Seminara, pp. 33-45, 1986.

 0090-4104/89/4701-2302$12.50

2) If the texts X and Y are independent, then the points of local maxima of their volume functions are uncorrelated (assuming that the time intervals of equal length described in both texts overlap).

For a discussion of the maximum correlation principle and its experimental verification, see [1-4]. Here we deal with "pointed information," i.e., we track the maximum points and ignore the magnitude of the splashes. The maximum correlation principle has been successfully applied by historians, and, together with the frequency decay principle, also formulated in [2, 3], it has been used in [5] to analyze the dependence of particular historical texts.

In this paper, we advance and test the following hypothesis: 1) for two texts X and Y that are known to be dependent, the volume functions (and not only the points of local splashes of the volume functions) should be "correlated" (assuming that the problem has been properly posed); 2) for two texts that are known to be independent, no reasonable correlation of volume functions should be observed.

This hypothesis, of course, is more complicated than the maximum correlation principle described above. It incorporates "more information" – both the location of the splash points and the magnitude of the splashes.

The original maximum correlation principle [1-3] relied on the fact that different chroniclers describing the same period in the history of the same region draw mainly upon the same "store of preserved information" (ancient texts), and as a result they tend to describe in greater detail those years for which a larger number of texts have survived and in less detail years with only few surviving texts. Now that we want to take into account also the amplitude of the volume function, we have to allow for the obvious fact that although the chronicler "makes a splash" in describing a particular year, the magnitude of this splash may depend on a variety of intractable factors, such as personal sympathies and antipathies with the events being described.

In our research we used several long chronicles describing the events in Russian history in the 9th through 17th centuries. Each of these texts contains a clear division by years (introduced by the original chroniclers). The chronicler states the year (reckoning from the creation of the world) and then enumerates the events which (in his opinion) occurred in that year.

a) As the first pair of dependent texts, we chose the Nikiforovaskaya chronicle (X) and the Suprasl'skaya chronicle (Y), both from Complete Collection of Russian Chronicles, Vol. 35, Moscow (1980). As the interval (A, B) described in both texts we chose the period of 406 years from 850 A.D. to 1256 A.D. The choice of this particular interval can be justified as follows. The brief introduction at the beginning of the Nikiforovskaya chronicle covers a long historical period from Adam to the Flood and then up to year 6362 from the creation of the world. This introductory part contains no detailed chronological markers (no dates) and is extremely brief (less than half a page), all of which suggested that we should omit the description of the period from deep antiquity until the year 6362 from the creation of the world. It is only starting with this year that the text is divided into "chapters" describing different years. For example: "In the summer of 6362. The beginning of the land of Rus," etc. The key events described inthe chronicle include legends about the beginning of Rus, Ryurik, the brothers Kii, Shchek, Khoriv, the baptism of the Bulgars, Oleg, Igor, the campaign against the Greeks, Greeks and Russia, Vladimir (in detail), Yaroslav, Novgorod, Suzdal, Smolensk, the invasion of Mamai, the history of Vitovt (Vytautas), the war against the Tartars, Lithuania. The text ends in 1430 A.D. However, starting with 1112 A.D. large lacunae appear in the chronicle. We therefore decided to end the sample period in 1256, where a particularly large 50-yr lacuna begins.

The second text is the Suprasl'skaya chronicle. In both chronicles, the volume of the fragments X(t) was determind by line count. The two chronicles describe roughly the same epoch in the history of Russia and some adjacent regions. Their dependence is particularly remarkable in that the two chronicles are definitely not identical, although possibly both have common sources. The chronicles substantially differ in style and in emphasis on the assessment of events. Thus, the author of the Nikiforovskaya chronicle devotes 36 lines to the year 970, while the author of the Suprasl'skaya chronicle devotes only 7 lines to this year. On the other hand, the Nikiforovskaya chronicler had nothing to report about the events of the year 977, while the Suprasl'skaya chronicler devoted 4 lines to this year. Despite all this, the correlation of the maximum points is quite pronounced [1-3]. In addition to the different distribution of fragment volumes, different events are sometimes

described in the same years. For instance, the Suprasl'skaya chronicle reports in 1233 the wedding of Aleksandr Nevskii, while the Nikiforovskaya chronicle does not mention this event. Both chroniclers thus increase (or decrease) the degree of detail of their description, sometimes by describing different events.

b) Another pair of dependent texts included the Kholmogorskaya chronicle (X) and the Tale of Bygone Years (known in English as Russian Primary Chronicle) (Y), from Complete Collection of Russian Chronicles, Vol. 33, Leningrad (1977) and the series Literary Monuments of Old Russia, Moscow (1950). Here A = 850 A.D., B = 1000 A.D. These chronicles essentially differ from each other in degree of detail. Nevertheless, the maximum correlation principle points to pronounced dependence of the two chronicles [1-3].

c) A third pair of dependent texts included the Dvinskii chronicle (sort edition) (X) and the Dvinskii chronicle (complete, extended edition) (Y), both from Complete Collection of Russian Chronicles, Vol. 33, Leningrad (1977). Here A = 1390 A.D., B = 1717 A.D.

d) The fourth pair of dependent texts included the Akademicheskaya chronicle [see Complete Collection of Russian Chronicles, Vol. 35, Moscow (1980)] and the part of the Suprasl'skaya chronicle describing the events from 1336 to 1374 A.D. (Y).

Independent pairs of texts are generated quite simply, e.g., by the following formal technique. Take some text X and as an independent text Y take the same text X "reading it backward," i.e., the sequence of years is reversed (the last year becomes the first, and so on).

It is sometimes helpful to treat the graph of the volume function of the text X as the result of observations of some stochastic process. Such a stochastic process is the sequence of events in the history of the given region (over the given time interval). Each chronicler is a "black box" processing this sequence and producing on the "output" his own chronicle, which in particular determines the volume of description of each year. In this way, different chroniclers may generate texts that will have roughly the same or substantially different degree of detail. The most stable results (in statistical terms) are obtained of course when we compare texts of "equal order" (i.e., "poor" with "poor" or "rich" with "rich"). The comparison of texts of different order (i.e., "poor" with "rich") should be approached more carefully.

Let us formulate some useful text manipulation rules: 1) Volume graphs should not be treated as "ideally exact," they should be regarded as "fuzzy" information. If two chroniclers "made splashes" close to each other (e.g., one of them "erred" by 1 year in dating a particular event), then these splashes should be treated as "approximately coincident," since an error of this kind is quite natural when describing events removed by many tens or hundreds of years into the past. 2) It is helpful to "smooth" the volume graphs and to repeat the comparison each time, taking the least value of the proximity coefficient of the graphs. 3) It is useful to focus only on the "largest" splashes, ignoring "small ripples" on the volume graph.

Let us briefly summarize the findings of our research.

a) The proposed statistical methods confidently discriminate between pairs of texts that are known to be dependent and those that are known to be independent (allowing for the amplitudes of the volume functions!). b) The sharpness of this discrimination is different for different techniques (see below). c) Pairs of dependent texts (of equal degree of detail) are confidently discriminated (by all procedures) from pairs of independent texts. d) Pairs of dependent texts of different degree of detail (poor and rich) are still discriminated from pairs of independent texts, but (for some techniques) with lower degree of confidence.

1. TECHNIQUES TREATING THE VOLUME FUNCTION AS A PROBABILITY DISTRIBUTION

1.1. Let us consider a modification of Fomenko's comparison methods [1-4], based on Kantorovich's multidimensional theorem of displacement of masses [6]. Let f(t) = f(X, t) be the volume function of the text X on the interval [A, B]. Consider the full volume of the text X. Then we may write

$$vol\, X = \int_A^B f(t)\, dt = \sum_{t=A}^{B} vol\, X(t).$$

Construct the function

$$S(t)=S(x,t)=\frac{1}{vol\,X}\int_A^t f(u)\,du.$$

Clearly, $0\leqslant S(t)\leqslant 1$ on the interval [A, B] and S(t) is a nondecreasing function. The function $S(\cdot)$ will be called for brevity the "accumulated sum" of the text X.

Consider two texts X and Y. Let us estimate their dependence or independence by comparing the accumulated sums S(X, t) and S(Y, t) on the interval [A, A + T], where T = B − A is the length of the time period described in the texts. The accumulated sums "smooth out" small fluctuations of the volume graphs, and it is therefore natural to try and apply them for text dependence analysis. We use the functions f(X, ·), f(Y, ·) to define the probability measures $\mathbf{P}_X(\cdot)$ and $\mathbf{P}_Y(\cdot)$, treating f(X, ·) and f(Y, ·) as the distribution functions of these measures. The measures $\mathbf{P}_X$ and $\mathbf{P}_Y$ obviously have the same support − the interval D = [A, B]. The measure $\mathbf{P}_X$ is called the (normalized) mass of the text. Following the terminology of the problem of displacement of masses (see [6]), we call the comparison plan of the texts X and Y any probability measure $\mathbf{P}$ on the direct product D × D with the projections $\mathbf{P}_X(\cdot)=\mathbf{P}(\cdot\times D)$, $\mathbf{P}_Y(\cdot)=\mathbf{P}(D\times\cdot)$. For any intervals I_1 and I_2 ($I_j \subset D$, j = 1, 2), $\mathbf{P}(I_1\times I_2)$ is the fraction of the mass of the text X on the interval I_1 which is identified with the mass of the text Y from the interval I_2 under the plan $\mathbf{P}$ (see Fig. 1). We have the obvious equalities

$$\mathbf{P}(I_1\times D)=\mathbf{P}_X(I_1),\ \mathbf{P}(D\times I_2)=\mathbf{P}_Y(I_2).$$

This identification is essentially interpreted as identification of the dates of the events in the two texts: an event in the text X dated by the year t_1 is identified with the event in text Y dated by the year t_2. Such time shifts are obviously undesirable, and we assess the damage caused by this redating with the aid of a nonnegative function $c(t_1, t_2)$, $t_1, t_2\in D$, where c(t, t) = 0 for all $t\in D$. It is sometimes convenient to define the function c in the form

$$c(t_1,t_2)=H(|t_1-t_2|),\ t_1,t_2\in D, \tag{1}$$

where H is some nondecreasing convex function. We will only consider the most general case, i.e., we assume that c is a 2-antitone function, i.e., it satisfies the inequality

$$c(t_1+\Delta_1,\ t_2+\Delta_2)-c(t_1,\ t_2+\Delta_2)-c(t_1+\Delta_1,\ t_2)+c(t_1,\ t_2)\leqslant 0$$

for all $\Delta_1>0$, $\Delta_2>0$, $t_1, t_2, t_1+\Delta_1, t_2+\Delta_2\in D$. If c has the representation (1), then c is clearly a 2-antitone function.

We denote the collection of comparison plans by $\mathscr{CP}=\mathscr{CP}(X, Y)$. The total cost associated with the realization of each plan $\mathbf{P}\in\mathscr{CP}$ is naturally evaluated by the integral

$$\mathrm{Cost}(\boldsymbol{P})=\int_{D\times D} c(t_1,t_2)\,\boldsymbol{P}(dt_1,dt_2). \tag{2}$$

Therefore, the sought optimal comparison is characterized by the measure $\mathbf{P}^*\in\mathscr{CP}$ on which the minimum

$$\min\{\mathrm{Cost}(\mathbf{P}):\mathbf{P}\in\mathscr{CP}\}=\mathrm{cost}(\mathbf{P}^*) \tag{3}$$

is attained.

We denote the left-hand side of the equality (3) by $\mathscr{E}(X, Y; c)$. The number $\mathscr{E}(X, Y; c)$ is naturally considered as a measure of difference of the texts X and Y with cost function c relative to the characteristics f(X, ·) and f(Y, ·). The measure $\mathbf{P}^*$ is called the optimal comparison plan.

Let us now describe the explicit formula for $\mathscr{E}(X, Y; c)$. Let

$$S^{-1}(u)=S^{-1}(X,u)=\max\{t: S(X,t)<u\},\ u\in[0,1] \tag{4}$$

be the inverse function of the accumulated sum S(X, ·). Now, if c is a 2-antitone function, then

$$\mathscr{E}(X,Y;c)=\int_0^1 c(S^{-1}(X,u),S^{-1}(Y,u))\,du \tag{5}$$

and the optimal comparison plan $\mathbf{P}^*$ is given by the equality

$$\mathbf{P}^*([A,\ t_1]\times[A,\ t_2])=\min\{S(X,\ t_1),\ S(Y,\ t_2)\},\ t_1,\ t_2\in D \tag{6}$$

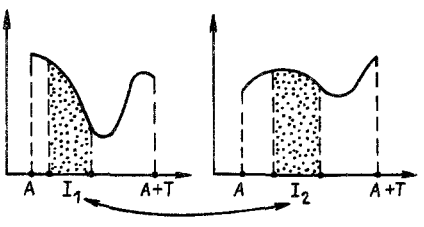

Fig. 1

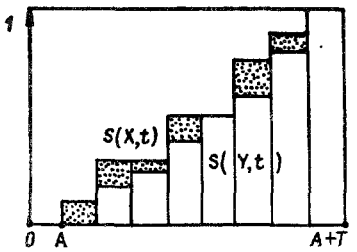

Fig. 2

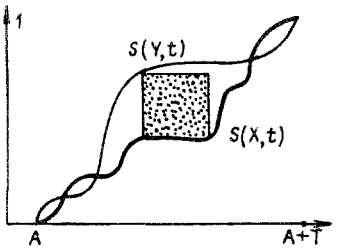

Fig. 3

(about the formulas (5) and (6), see the survey [6]). If c(t, s) = |t − s|, then the difference measure $\mathscr{E}(X, Y; c)$ of the texts X and Y is called Kantorovich metric and

$$\mathscr{E}(X, Y; c)=\varkappa(X, Y)=\int_A^B |S(X, t)-S(Y, t)|\,dt \tag{7}$$

(see Fig. 2).

The constructions described above can be extended also to the comparison of N > 2 texts (see [6]), but so far this possibility has not been tested empirically.

Optimal comparison plans are particularly important for the needs of chronology, because they ensure the best overlapping of the two texts.

Another problem is choosing the bounds a and b so that for $\mathscr{E}=\mathscr{E}(X, Y; c)<a$ we can say that the texts and dependent and for $\mathscr{E}\geqslant b$ that they are independent. These bounds can be determined empirically by analyzing a large number of texts that are known to be dependent or independent [3].

1.2. The loss function associated with the realization of some comparison plan $\mathbf{P}\in\mathscr{CP}(X, Y)$ may be constructed without resorting to the auxiliary function c. We will define the loss function using the concept of Hausdorff distance between sets (see [7]) and assuming that the comparison plan **P** is "good" (i.e., involves small time "displacements" of the texts) if $\mathbf{P}([A, t], (s, B]$ is small for "close" values of t and s. Moreover, we define the total cost associated with the realization of the plan **P** with the aid of the following Hausdorff distance:

$$\begin{aligned}\mathrm{Cost}_L(\boldsymbol{P})=\max\{&\sup_{t\in D}\inf_{s\in D}\max\{|t-s|, \boldsymbol{P}([A, t]\times(s, B])\},\\ &\sup_{t\in D}\inf_{s\in D}\max\{|t-s|, \boldsymbol{P}((s, B]\times[A, t])\}\}\end{aligned} \tag{8}$$

(see [8, 9]).

The optimal plan $\mathbf{P}^*$ among all the plans $\mathbf{P}\in\mathscr{CP}(X, Y)$ is defined by the equality

$$\mathrm{Cost}_L(\mathbf{P}^*)=\min\{\mathrm{Cost}_L(\mathbf{P}) : \mathbf{P}\in\mathscr{CP}(X, Y)\}. \tag{9}$$

The optimal comparison plan $\mathbf{P}^*$ exists and is defined by formula (6) (see [8, 9]), and the total cost associated with the realization of $\mathbf{P}^*$ is determined by Levy's metric between the accumulated sums S(X, ·), S(Y, ·), i.e.,

$$\mathrm{Cost}_L(P^*)=L(S(X, \cdot), S(Y, \cdot))=L(X, Y) \tag{10}$$

(see [8, 9]): this is the length of the largest square that can be inscribed between the completed graphs S(X, ·), S(Y, ·) (see Fig. 3).

Levy's metric L(X, Y) may be treated as a measure of deviation of the volume functions f(X, ·), f(Y, ·) in the spirit of the problem of displacement of masses.

1.3. The tests presented in Secs. 1.1-1.2 were applied to two texts that were known to be dependent – Nikiforovskaya (X) and Suprasl'skaya (Y) chronicles (850-950 A.D.). We obtained $\varkappa(X, Y)=1.5$ and L(X, Y) = 0.01. Application of the same tests to a pair of texts known to be independent (X is the Nikiforovskaya chronicle for 850-950 A.D., Y is the Nikiforovskaya chronicle for 950-1050 A.D.) produced the following values: $\varkappa(X, Y)=8.4$, $L(X, Y)=0.08$.

We see that the proposed tests distinguish between dependent and independent texts. Further computational work is needed in order to justify the applicability of the proposed tests.

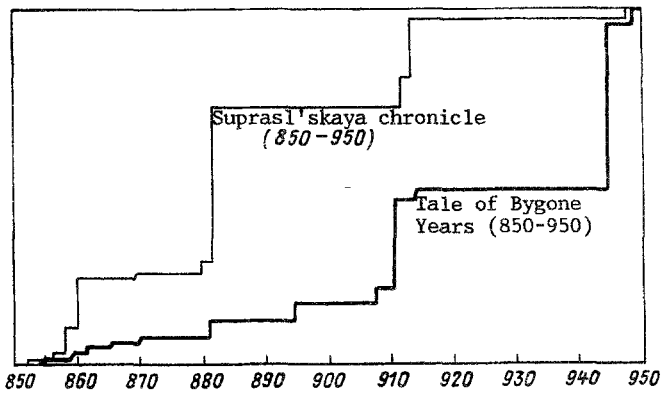

Fig. 4

1.4. The procedures described in Secs. 1.1, 1.2 are effective when dealing with texts of roughly the same volume. If we compare, say, a "poor" and a "rich" text (see Fig. 4), then the two volume curves are substantially different in the metrics $\varkappa$ and L. This suggests that we need a special cost function c that will bring rich and poor texts closer to each other. This problem has not been solved so far, and therefore for comparison of texts of different order ("poor"–"rich") we propose the "sum of jumps" tests. This test can be described as follows. Take a sufficiently small number $\varepsilon > 0$ (the approximation level). Let $\Delta S(X, t)$ be the magnitude of the jump of the graph $S(X, t)$ [resp., $\Delta S(Y, t)$] in year t.

Compute the sum of the jumps $\Delta S(X, t')$ by the following rule:

$$\sum_{t'} \lambda(t') \Delta S(X, t'),$$

where t' are the years when $\Delta S(X, t') > \varepsilon$; $\lambda(t') = 1$ if the δ-neighborhood of the year t' does not contain a year t" such that the function $S(Y, t")$ makes a jump greater than ε and $\lambda(t') = 0$ otherwise. Here $\delta > 0$ is a fixed number, characterizing the admissible time uncertainty of the dating in the text. Set $\Delta(X \to Y) = \sum_t \lambda(t') \Delta S(X, t')$. We similarly evaluate $\Delta(Y \to X)$ by interchanging the texts X and Y. As the resultant measure of distance between the texts X and Y take $\Delta(X, Y) = (1/2)(\Delta(X \to Y) + \Delta(Y \to X))$.

Thus, by computing the number $\Delta(X \to Y)$, say, we compute, roughly speaking, the sum of those maxima of the function vol(X, t) that occur "against the background" of approximate constancy of the function vol(Y, t). In other words, we compute a measure of distance between the two accumulated sums S(X, t), S(y, t).

Here we again use "pointed information" which, however, also follows for the magnitude of the jumps. Let us demonstrate the application of this procedure for $\varepsilon = 0$, 1, $\delta = 2$ (i.e., jumps in the graphs S not exceeding 0.1 were "filtered" and time "inaccuracies" of up to 4 years were allowed). For the Nikiforovskaya (X) and the Suprasl'skaya (Y) chronicles, 850-950, we obtained $\Delta(X, Y) = 0$ – ideal dependence. For the two parts fo the Nikiforovskaya chronicle (X is the part for 850-950, Y the part for 950-1050), we obtained $\Delta(X, Y) = 488$. Comparison of the Nikiforovskaya chronicle (X) with the Tale of Bygone Years (Y) on the interval 850-950 gave $\Delta(X, Y) = 11$. This test detects dependence and independence with fair degree of confidence.

1.5. Yet another test can be devised by observing the dynamics of convergence of the graphs $S(X, \cdot)$ as we gradually "deplete" the texts. For example, assume that the texts X and Y are compared by the metric $\varkappa(X, Y)$ [see (7)].

Let $\varkappa_i(X, Y; \tau_1, \tau_2, \ldots, \tau_i)$, $i \leqslant 1$ be the metric (7) between the texts X and Y after deletion of the chapters relating to the years $\tau_1, \tau_2, \ldots, \tau_i$ (these are naturally different years). Let

$$\varkappa_i(X, Y) = \min_{\tau_1, \ldots, \tau_2} \varkappa_i(X, Y; \tau_1, \ldots, \tau_i).$$

For dependent texts, we naturally expect to observe not only smaller changes in $\varkappa_i$ ($i \geqslant 1$) than for independent texts, but also relatively faster reduction of $\varkappa_i$ (for independent texts, the reduction of $\varkappa_i$ will be slower). Figure 5 illustrates this situation.

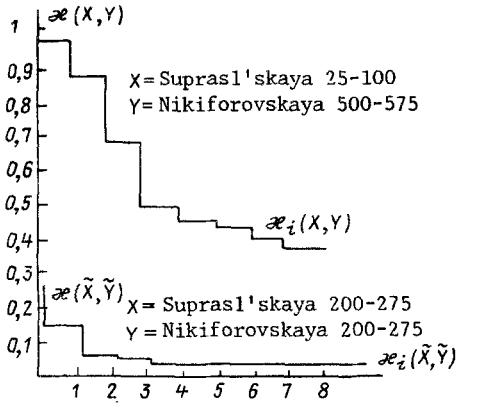

Fig. 5

2. A COMPARISON TECHNIQUE TREATING THE TEXT VOLUMES AS RANDOM SAMPLES

2.1. We start with a review of some well-known results of probability theory [10]. Let ξ and η be two nonnegative random variables (r.v.s) with joint distribution function (d.f.) $H(x, y)=\boldsymbol{P}(\xi\leqslant x, \eta\leqslant y)$ and marginal distribution functions $F(x)=\boldsymbol{P}(\xi\leqslant x)=\lim_{y\to\infty} H(x, y)$ and $G(y)=\boldsymbol{P}(\eta\leqslant y)=\lim_{x\to\infty} H(x, y)$. We measure the deviation of the r.v.s ξ and η by $\boldsymbol{E}|\xi-\eta|$. We have the equality

$$\min \boldsymbol{E}|\xi-\eta|=\int_0^\infty |F(x)-G(x)|\,dx=l(F, G), \tag{1}$$

where min is over all possible distributions H with given marginal distributions F and G. We know that the equality in (1) is attained for the function $H(x, y)=\min(F(x), G(y))$. This form of H corresponds to the case of "strong dependence" between the r.v.s ξ and η. If the r.v.s ξ and η are independent, then

$$\boldsymbol{E}|\xi-\eta|=\int_0^\infty (F(x)+G(x)-2F(x)G(x))\,dx=m(F, G). \tag{2}$$

Therefore judging by the degree of proximity of $\mathbf{E}|\xi-\eta|$ to $l(F, G)$ or to $m(F, G)$, we say that the r.v.s ξ and η are strongly dependent or "almost independent." Incidentally note that $l(F, G)$ is the minimal possible value for $\mathbf{E}|\xi-\eta|$ [see (11)]. At the same time, $m(F, G)$ is not the maximal possible value for $\mathbf{E}|\xi-\eta|$. It corresponds to the case of independent ξ and η. Even higher values of $\mathbf{E}|\xi-\eta|$ are obtained for negatively correlated ξ and η.

2.2. We treat the sequence of volumes $f(X, t)$, $A\leqslant t\leqslant B$, as a sample of independent r.v.s with d.f. $F_X(x)=\mathbf{P}\{f(X, t)\leqslant x\}$. This approach is justified by the unpredictability of real historical events, nondeterminism of the personal traits of the chronicler, and also the effect of purely conjunctural and personal factors. We similarly treat the sequence $f(Y, t)$, $A\leqslant t\leqslant B$ as a sample of independent r.v.s with d.f. $F_Y(x)=\mathbf{P}\{f(Y, t)\leqslant x\}$. Since we have no other information apart from the texts X and Y, we naturally take F_X and F_Y as the empirical distribution functions,

$$F_X(x)=\#\{t: f(X,t)\leqslant x\}. \tag{3}$$

$$F_Y(x)=\#\{t: f(Y, t)\leqslant x\}. \tag{4}$$

An analogue of the mean $\mathbf{E}|\xi-\eta|$ in this case is the empirical average

$$M(X, Y)=\frac{1}{B-A+1}\sum_{t=A}^{B} |\operatorname{vol}(X, t)-\operatorname{vol}(Y, t)|. \tag{5}$$

This interpretation of the texts X and Y suggests dependence if $M(X, Y)$ is close to $l(F_X, F_Y)$ and independence if $M(X, Y)$ is close to $m(F_X, F_Y)$.

As a more sophisticated test of dependence of the texts X and Y we can suggest comparison of the sample d.f. $H_{XY}(x, y)=\#\{t: \operatorname{vol}(X, t)\leqslant x, \operatorname{vol}(Y, t)\leqslant y\}$ with the functions $\min(F_X(x), F_Y(y))$. This test, however, is computationally costly and has not been tried so far.

TABLE 1

Chronicles compared	$M(x, Y)$	$l(F_X, F_Y)$	$m(F_X, F_Y)$	$\frac{M-l}{m-l}$
Suprasl'skaya (X) (850-1256 A.D.) Nikiforovskaya (Y)	0,88	0,62	2,96	0,11
Suprasl'skaya (X) (850-1256 A.D.) Inverted Nikiforovskaya (Y)	2,67	0,62	2,96	0,87
Dvinskaya complete (X) (1390-1717 A.D.) Dvinskaya short (Y)	2,89	2,55	5,63	0,11
Rachinskii's (X) (1400-1550 A.D.) Evreinovskaya (Y)	2,32	0,92	5,51	0,3
Inverted Rachinskii's (X) (1400-1550 A.D.) Evreinovskaya (V)	5,97	0,92	5,51	1,1
Vladimirskaya (X) (830-1241 A.D.) Volynskaya (Y)	0,79	0,42	0,83	0,9
Inverted Vladimirskaya (X) (830-1241 A.D.) Volynskaya (Y)	0,80	0,42	0,83	0,92
Suprasl'skaya (X) (850-1110 A.D.) Tale of Bygone Years (Y)	19,68	18,81	20,14	0,65
Suprasl'skaya (X) (850-1110 A.D.) Inverted Tale of Bygone Years (V)	20,13	18,81	20,14	0,99
Nikiforovskaya (X) (850-1110 A.D.) Tale of Bygone Years (Y)	19,91	18,95	20,08	0,85
Nikiforovskaya (X) (850-1110 A.D.) Inverted Tale of Bygone Years (Y)	20,04	18,95	20,08	0,96

Our assumptions of independent and identically distributed vol(X, t), $A \leqslant t \leqslant B$, are not fulfilled in practice, strictly speaking. Nevertheless, it seems that these assumptions are not decisive. This can be demonstrated by the following idealized example. Let the quantity ω_t $(A \leqslant t \leqslant B)$ characterize the volume of the actual events in hear t. Assume that the chronicler X processes these events as follows:

$$\mathrm{vol}(X, t) = f_X(\omega_t), \tag{6}$$

where $f_X(x)$ is a positive monotone increasing function. In other words, the chronicler writes more in years that are richer in events. Similarly, the chronicler Y processes the events with the corresponding function

$$\mathrm{vol}(Y, t) = f_Y(\omega_t).$$

It is easy to see that in this case the following equality holds without any randomness assumptions:

$$M(X, Y) = l(F_X, F_Y). \tag{7}$$

It is useful to note the following fact. If the texts X and Y are such that vol(X, t) $\leqslant$ vol(Y, t) for all $A \leqslant t \leqslant B$, then the equality (7) also holds.

Table 1 lists the values of M(X, Y), $\ell(F_X, F_Y)$, and $m(F_X, F_Y)$ for different pairs of texts X and Y. The results presented in Table 1 clearly show that the following pairs of texts are dependent:

a) Suprasl'skaya and Nikiforovskaya chronicles;

b) complete and short versions of the Dvinskaya chronicle;

c) Evreinovskaya chronicle and Rachinskii's chronicle.

When one of the texts is inverted, the pair is clearly independent. The Volynskaya chronicle and the Chronicle of the Great Prince Vladimir of Kiev (Vladimirskaya chronicle) are also independent.

The case is more complex when comparing "poor" and "rich" texts. As a rich text we use the Tale of Bygone Years. We see that its comparison with the Suprasl'skaya chronicle does not give a conclusive result. On the other hand, comparison with the Nikiforovskaya

TABLE 2

Chronicles compared (850-1000 A.D.)		$M(X, Y)$	$l(F_x, F_y)$	$m(F_x, F_y)$	$\frac{M-l}{m-l}$	Total shift years
Tale of Bygone Years (X) Kholmogorskaya (Y)	Without time shift	13,07	10,97	20,42	0,22	0
	With time shift	11,05	10,98	20,60	0,0072	8
Tale of Bygone Years (X) Suprasl'skaya (Y)	Without time shift	16,73	15,28	17,58	0,63	0
	With time shift	15,6	15,28	17,56	0,12	12
Tale of Byfone Years (X) Nikiforovskaya (Y)	Without time shift	17,16	15,46	17,45	0,85	0
	With time shift	15,64	15,46	17,48	0,09	10

chronicle (which, in its turn, is strongly dependent with the Suprasl'skaya chronicle) suggests that the two texts are independent. We will return to the case of texts of different dize in Sec. 2.4.

2.3. When the sequences of values f(X, t), f(Y, t) are treated as random samples, a natural measure of dependence is provided by the coefficient of correlation

$$r = r(X, Y) = \frac{\sum_{t=A}^{B} (f(X, t) - EX)(f(Y, t) - EY)}{\left[\sum_{t=A}^{B} (f(X, t) - EX)^2 \sum_{t=A}^{B} (f(Y, t) - EY)^2\right]^{1/2}}.$$

The possible values of r are contained in the interval [−1, 1]. Closeness of r to zero suggests that the texts X and Y are independent and its closeness to 1 suggests that they show (positive) dependence. Calculations using this technique produced the following results. Comparison of the texts of Sergeev (X) [12] and Levy (Y) [15], both dealing with ancient Rome, gave a correlation coefficient r(X, Y) = 0.48, i.e., detected noticeable dependence of these texts, which is not surprising, because both are based on the same events. Comparison of the texts of Bemont and Monod (X) [14] and Kohlrausch (Y) [13], both describing medieval Rome, gave r(X, Y) = 0.77, which also points to dependence. Comparison of the texts of Levy (X) [15] and Gregorovius (Y) [16], describing ancient Rome and medieval Rome, respectively, gave r(X, Y) = 0.528. On the other hand, comparison of the text of Sergeev [12] with the same text read in backward order gave r = −0.046, which indicates independence. Application of the proposed procedure to texts of different size (X − Tale of Bygone Years, Y − Suprasl'skaya chronicle) gave r = 0.125. In this case, the proposed technique fails to detect dependence of the texts.

2.4. The results presented in Tables 1 and 2 show that the proposed procedure fairly confidently detects dependence of texts of similar volume (e.g., Suprasl'skaya and Nikiforovskaya chronicles). At the same time, dependence of "poor" and "rich" texts (e.g., Nikiforovskaya chronicle and Tale of Bygone Years) is not detected, although the "rich" lies almost always above the "poor" text. This is attributable to the small number of splashes (4-5 cases) in the "poor" text that are located at a distance of about one year from the corresponding splashes in the "rich" text. Accuracy of one year is of course excessive for problems of this kind. Therefore, we should try to construct a dependence test combining two of the techniques described above:

1) the method based on proximity of maxima (see [1-4]),

2) the procedure that treats text volumes as r.v.s.

Under this approach, we first match the close splashes in the volume of the compared texts (and determine the required shifts) and then apply the test to compare the resulting (distorted) texts. This technique is widely used in functional analysis − see, e.g., the Skorokhod distance in the space D[0, 1] [11]. The resulting test allows for both components.

Table 2 summarizes the results obtained by applying this procedure to pairs of texts of different size. The table gives not only the value of the test statistic, but also the total number of years by which the chapters were shifted (only the "poor" texts were shifted) – each chapter was shifted by not more than one year. For example, in the Kholmogorskaya chronicle, only 8 of the 150 chapters had to be shifted by one year. The results show that these time shifts sharply reduce the value of $(M - \ell)/(m - \ell)$ for dependent texts. For independent texts, such a reduction requires a substantially greater number of shifts.

Our results are encouraging for the possibilities of detection of dependence between texts. So far, however, the question of combining the computed distances and total shifts into a single test remains open. Here, as in the other techniques, the solution of the problem will follow once more extensive computational material has been accumulated.

3. CONCLUSIONS

The techniques described in this paper are essentially experimental. So far, they have been tried on a limited volume of empirical material, and final verdict of applicability should await more detailed checks and calibration. Yet even preliminary results suggest that it is indeed possible to develop tests for classifying texts into dependent and independent with allowance for their volume.

We are grateful to N. Ya. Rives for his considerable interest in this research and for his willing assistance with computer work. His expert help has enabled us to test a number of hypotheses and to advance new ones.

LITERATURE CITED

1. A. T. Fomenko, "Some statistical regularities in the distribution of information density in texts with a scale," in: Semiotics and Informatics [in Russian], No. 15, VINITI, Moscow (1980), pp. 99-124.
2. A. T. Fomenko, "Information functions and associated statistical regularities," in: Abstracts of Papers at 3rd International Vil'nyus Conf. on Probability Theory and Mathem. Statistics [in Russian], Vol. 2, Inst. Mat. i Kibernet. AN LitSSR, Vilnius (1981), pp. 211-212.
3. A. T. Fomenko, New Empirical–Statistical Procedures for Dating of Ancient Events and Application to the Global Chronology of the Ancient and Medieval World [in Russian], Preprint, Gos. Kom. Telev. Radioveshch. order 3672 (9 Sept. 1981), No. B7201, Moscow (1981).
4. V. V. Fedorov and A. T. Fomenko, "Statistical estimation of chronological proximity of historical texts," in: Stability Problems of Stochastic Models, Proc. of a Seminar [in Russian], VNIISI, Moscow (1983), pp. 101-107.
5. L. E. Morozova, "Quantitative methods in the analysis of so-called Filaret manuscripts – a record of 'Troubled Times,'" in: Mathematical Methods and Computers in Historical Research [in Russian], Nauka, Moscow (1985), pp. 182-203.
6. S. T. Rachev, "The Monge–Kantorovich problem of displacement of masses and its application in stochastic theory," Teor. Veroyatn. Primen., 29, No. 4, 625-653 (1984).
7. F. Hausdorff, Set Theory [Russian translation], ONTO, Moscow (1937).
8. S. T. Rachev, "On minimal metrics in the space of real random variables," Dokl. AN SSSR, 257, No. 5, 1057-1070 (1981).
9. S. T. Rachev, "Minimal metrics in the real valued random variable space," Lect. Notes Math., 982, 172-180 (1983).
10. V. M. Zolotarev, "Metric distances in spaces of random variables and their distributions," Mat. Sb., 101(143), No. 3(11), 416-454 (1976).
11. P. Billingsley, Convergence of Probability Measures [Russian translation], Nauka, Moscow (1977).
12. V. S. Sergeev, Essays in the History of Ancient Rome [in Russian], Moscow State Univ. (1938).
13. Kohlrausch, German History [Russian translation], Vols. 1, 2, Moscow (1860).
14. C. Bemont and G. Monod, History of Europe in the Middle Ages [Russian translation], Petrograd (1915).
15. T. Levy, History of Rome [Russian translation], Moscow (1897-1899).
16. F. Gregorovius, The History of the City of Rome in the Middle Ages [in Russian], St. Petersburg (1902-1912).