STATISTICAL ESTIMATION OF CHRONOLOGICAL NEARNESS OF HISTORICAL TEXTS

V. V. Fedorov and A. T. Fomenko

1. The "principle of correlation of maxima" of the plots of the volume of historical texts has been formulated and tested for the first time in [1] for the case of a uniform distribution (see also [2]). This principle and the related method of dating of events described in historical texts (with a time scale) were found to be necessary in various chronological investigations carried out in [1-3, 10-16]. The importance of the results obtained in these papers, and especially with the use of the principle of correlation of maxima (see the corresponding formulation below or in [1, 2, 10, 16]), shows the utility of testing the stability of this principle and of the corresponding method of dating of events with respect to other procedures of statistical processing of the volume functions of texts. All the volume functions of historical texts used in this paper have been calculated in [1, 10].

2. Let us recall the principle of correlation of maxima and the method of dating of events based on it. Suppose that a historical period from the year A to the year B in the history of a region G (i.e., a state, a city, etc.) has been described in a fairly comprehensive year-by-year text (chronicles, annals, etc.), i.e., the text X is split into sections or "chapters" $X(T)$, each of which describes the events of one year T. We calculate the volume H(X, T) of each such section X(T) measured, for example, by the number of lines (or words, or symbols, or pages, etc.) (see Fig. 1). For another year text Y that describes this same time interval (A, B) and this same region G, the corresponding plot of $H(Y,T)$ will in general have a different form, since the distribution of the text volume is considerably affected by the personal interests of the chroniclers (the authors of the texts). For example, the chronicle X of the history of art and the military chronicle Y will be written with an entirely different emphasis. To what extent are these differences essential, i.e., do there exist characteristics of the volume functions that are determined only by the time interval (A, B) and the region G and that unambiguously specify all (or almost all) the texts describ-

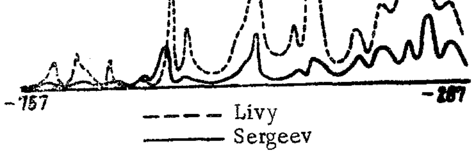

Fig. 1

Translated from Problemy Ustoichivosti Stokhasticheskikh Modelei — Trudy Seminara, pp. 101-107, 1983.

0090-4104/86/3206-0668$12.50

ing this interval of time. It is found that an important characteristic of such a plot consists of the years in which the plot has local maxima (see [1, 10]). For simplicity we shall assume that the latter are nondegenerate, i.e., reached locally at one point. The local maxima of the function $H(X, T)$ indicate the "years described in detail" in the time interval (A, B). Let $C(T)$ be the volume of all the texts written about the year T by contemporaries (i.e., persons living at that time). The plot of $C(T)$ is not known to us, since the texts are lost over the years, and the information vanishes. The following model of loss of information is constructed in [1, 10]: For the years about which a very large number of texts have been written, the number of texts that have been preserved will also be larger than usual. In such a form it is difficult to test the model, since we do not know the plot of $C(T)$. However, it is possible to test one of the consequences of this model, i.e., in view of the fact that later chroniclers X and Y, who describe this same period (A, B), are no longer contemporaries of these ancient events, they must rely on more or less the same collection of texts passed over to them, so that they must ("on the average") describe in more detail the years for which more texts have been preserved, and in less detail the years about which little is known (a small number of texts are available). The principle of correlation of maxima of the volume functions of texts X and Y has been formulated in [1, 2, 10] as follows: The plots of the volume of "chapters" for correlated texts X and Y (i.e., which describe the same period of time (A, B) and the same region G) must reach simultaneously local maxima in the interval (A, B), i.e., the years described in detail in X and the years described in detail in Y must be either close to one another, or they must coincide. In contrast, if the texts X and Y are independent, i.e., they describe either quite different historical periods (A, B) and (C, D) of the same length, or different regions, then the plots of the volume functions $H(X, T)$ and $H(Y, T)$ will reach local maxima at different points [provided that we let the segments (A, B) and (C, D) overlap]. This principle of correlation of maxima can be substantiated if for the majority of pairs of actual correlated historical texts X and Y, i.e., which describe practically the same events, the volume functions of the "chapters" for X and Y reach their maxima in roughly the same years. Here the value of the maxima must differ considerably. In contrast, for actual independent texts there must be no correlation whatsoever of the point of the maxima. In actual fact, in [1, 2, 10] the comparison was carried out not just with two texts, but with two groups of texts, and the averaged plot of the volume was calculated for each group.

3. It is evident that for actual plots of volumes of correlated texts the simultaneity of their peaks will occur only approximately. For estimating the degree of simultaneity with which two volume functions reach their maxima, it is necessary to introduce a natural measure that makes it possible to estimate numerically the mismatch of the points of the maxima. Such a measure can be introduced by different methods. It is required that it should distinguish reliably between pairs of dependent (correlated) texts and pairs of independent texts. It is found that such measures exist (which is not self-evident). The first method has been proposed in [1, 2, 10]. Let us briefly describe this measure. The points at which the function $H(X, T)$ reaches its maxima are dividing the segment (A, B) into smaller parts. By measuring

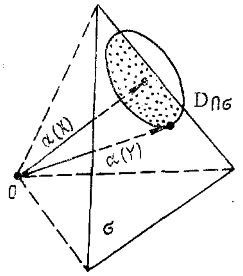

Fig. 2

their lengths in years, we obtain a sequence $(a_1, a_2, \ldots\ldots, a_p)$, of integers that specifies an integer-valued vector $a(X)$ in a Euclidean space R^p of dimension p.. For a text Y that describes a period of the same length, we obtain in general another vector $a(Y)=(b_1, b_2, \ldots, b_q)$, where the number q can differ from the number p (a different number of maxima). It can nevertheless be assumed that the number of maxima is the same. Let $p>q$; then some of the maxima of the function $H(Y, T)$ are assumed to be multiple, i.e., we assume that some maxima coalesce at a single point. This means that we are adjoining $p-q$ maxima to the plot of $H(Y, T)$. Let v be a version of introducing such multiplicities. It is evident that such a procedure is not single-valued. Thus, it can be assumed that $\sum_{i=1}^{p} a_i = \sum_{i=1}^{p} b_i = B-A$, i.e., the ends of the two vectors lie in the same simplex $\sigma=\sigma^{p-1}$ of dimension $p-1$, which is defined in the space $\boldsymbol{R}^p$ by the equation $\sum_{i=1}^{p} x_i = B-A$ (see Fig. 2). Let l be the length of the vector $a(Y)-a(X)\in\sigma$.

Let us write $p_v(X, Y)=\frac{\operatorname{vol} D\cap\sigma}{\operatorname{vol}\sigma}$, where D is a ball in $\boldsymbol{R}^p$ centered at the point X and with a radius l, and $D\cap\sigma$ is its intersection with the simplex. If C is a $(p-1)$-dimensional measurable subset in σ , then $\operatorname{vol} C$ will be either the Euclidean $(p-1)$-dimensional volume of C (the continuous case), or a number of integer points, i.e., of points with integer coordinates in C (the discrete case). Finally, we shall write $p(X, Y)=\frac{k(X, Y)+k(Y, X)}{2}$, where $k(X, Y)=\min\limits_{v} p_v(X, Y)$, i.e., the minimum is taken over all the faces of the simplex σ in the case that the original number of maxima was different. In the continuous case it is easy to verify that

$$p_v(X, Y) \leqslant \frac{\pi^{\frac{p-1}{2}} l^{p-1}(p-1)!}{G\left(\frac{p+1}{2}\right)(B-A)^{p-1}\sqrt{p}}.$$

Together with the functions $H(X, T)$ and $H(Y, T)$ also their smoothing is considered in [1, 2, 10]. The coefficient $p(X, Y)$ was calculated each time, and its minimum value was taken as the final value. By assuming that the random vector ξ is uniformly distributed on a simplex, we can interpret the number $p_v(X, Y)$ as the probability of a random event signifying that the random vector ξ is at a distance from the vector $a(X)$ that does not exceed the observed distance $l=|a(X)-a(Y)|$. If the volume functions for the texts X and Y reach their maxima simultaneously, then $p(X, Y)=0$. This procedure has been sharpened mathematically in different ways in [1, 2, 10, 16].

If the coefficient $p(X, Y)$ is "small," then the texts X and Y are dependent, but if the coefficient $p(X, Y)$ is "large," then the texts are independent, i.e., they tell different events.

4. The approach described in the previous section has been illustrated in [1, 3, 10] with the aid of a comprehensive computational experiment involving calculation of the numbers $p(X, Y)$ for different pairs X and Y of actual historical texts. Let us present here the principal result of this experiment. It was found that the coefficient $p(X, Y)$ distinguishes clearly between dependent and independent pairs of historical texts. For all the pairs of texts X, Y, studied in [10] and which describe quite different events (different epochs or different regions), i.e., for independent texts, the number $p(X, Y)$ fluctuates between 1 and 1/100 with a number of maxima ranging from 10 to 15. On the other hand, if the texts X and Y are dependent, i.e., they describe the same events (this being known by a preliminary historiographical analysis), then the coefficient $p(X, Y)$ does not exceed 10^{-8} for this same number of maxima. In Fig. 1 we plotted a typical example of two dependent texts, i.e., X which denotes V. S. Sergeev's monohraph "Topics in the History of Ancient Rome" [4], and which denotes the "history of Rome" by Titus Livius (Livy) [5]. Here (A, B) represents the years 757-287 B.C., and $p(X, Y)=2\cdot 10^{-12}$. Both texts describe the same period of Roman history. Another example of dependent texts is X which denotes the Kholmogory Annals [7], and Y which denotes the Tale of Past Years [8]. In this case (A, B) represents the years 850-1000 A.D., and $p(X, Y) =10^{-15}$. A similar example of dependent texts is X, denoting the Nikiforov Annals, and Y the Suprasl' Annals [7]. In this case (A, B) represents the years 850-1255 A.D., and $p(X, Y) =10^{-24}$. Yet another example of dependent texts which have been detected as a result of a numerical experiment carried out in [1, 3, 10] is given by X, representing part of the "History of Mediaeval Rome" by F. Gregorovius [9], which covers the history period from the year 300 to 745 A.D., whereas the second text, Y, is the "History of Rome" by Titus Livius [5], which covers the period from the 1st to the 459th year counted from the foundation of Rome, i.e., from the year 753-294 B.C. (ancient Rome). The volume functions of these two dependent texts are plotted in Fig. 3. Here $p(X, Y) =6\cdot 10^{-10}$. Let us recall that for quite independent texts the coefficient $p(X, Y)$ is not smaller than 1/100 (see [1, 2, 10]).

In analyzing quite independent texts, we calculated a lower bound for the number $p(X, Y)$ by approximating a multidimensional region by "cubic layers." In this way it was possible to compare (see [1, 2, 10]): a) Ancient texts with ancient texts, b) ancient with contemporary texts, and c) contemporary with contemporary texts. Instead of the volume functions $X(T)$ of

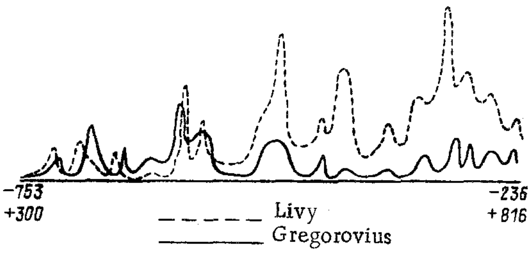

Fig. 3

"chapters," we compared also other quantitative characteristics of texts such as plots of the number of names mentioned (in each year T), plots of the number of times a certain year is mentioned in the text, plots of the frequency of referring to any other text, etc. (for details see [1, 3, 10]). It was found that all these characteristics are governed by the same statistical laws, i.e., the plots of dependent texts reach their local maxima practically simultaneously, whereas for independent texts the peaks of the plots are not at all correlated. The following procedure of dating of texts has been proposed and tested in [1, 3, 10]. Let Y be a text that describes events unknown to us whose absolute dating has been lost, with the years T being counted in the text Y on the basis of an event of local importance such as the founding of a city, or the day on which a ruler has been crowned, etc., the absolute dating of such an event having been lost. How shall we date the events described in Y ? For the text Y we shall calculate its volume function $Y(T)$ of "chapters" and compare it with the volume functions of other texts for which the absolute dating of events described in them is known to us. If among these texts we can find a text X, for which the number $p(X,Y)$ is small, i.e., it has the same order of magnitude as for pairs of dependent texts (i.e., it does not exceed 10^{-8} in the case of a number of maxima ranging from 10 to 15), then with a fairly high probability [the higher, the smaller the number $p(X, Y)$ % we can conclude that the events described in these texts either coincide, or are very near in time. In [1, 10] this procedure has been tested on mediaeval texts with a priori known dating. A typical example is Y , which represents the Dvinsk Chronicle (short edition) that describes the events taking place over a period of 327 years [6]. In going through the list of chronicles in the "Complete Collection of Russian Chronicles," we can find a text X, whose function $H(X, T)$ has maxima in practically the same years as the function $H(Y, T)$ (after letting the time intervals described in them overlap). A calculation yields $p(X, Y) = 2\cdot 10^{-25}$. It is found that X is a verbose edition of this same Dvinsk Chronicle (see [6]). Here (A, B) represents the years 1390-1717 A.D. The dating of the text Y obtained in [10] coincides with its standard dating. As another example let X denote the Academic Chronicle [7]. Following the technique described in [10], we find that the text X is part of the Suprasl' Annals [7], with (A, B) representing the years 1336-1374 A.D. In this case, $p(X, Y) = 10^{-14}$. Other examples and tables can be found in [1, 10].

5. The coefficient $p(X, Y)$ described above is based on the concept of spherical neighborhood of a point, and this makes it difficult to process by computer the experimental material, i.e., the plots of the text volume. For analyzing the stability of the principle of correlation of maxima described above and in [1, 10], and for utilizing a computer, it is possible to resort to the following method of statistical estimation of the chronological nearness of sequences of points of maxima of the volume functions of texts. Just as in Sec. 3, let $a(X)$ be a vector that describes the instants of time at which certain events are mentioned most in the source X . It is natural to assume that: a) This vector is closely related to the vector of actual events $a(\Pi)$ where Π is the historical period that is being described; b) the instants of occurrence of certain events that deserve to be mentioned in a certain text constitute a random point process. If we have two texts X and Y , then the simplest

relationships between them can be described by the following schemes: $a(X)\leftarrow a(\Pi)\rightarrow a(Y)$ and $a(\Pi)\rightarrow a(X)\rightarrow a(Y)$. In either case it is necessary that $a(X)$ and $a(Y)$ should be realizations of certain random processes that are close to one another. As a measure of proximity between $a(X)$ and $a(Y)$ we shall use the following quantity:

$$R(X, Y)=\sum_{i=1}^{N} \min_{j=\overline{1,M}} |a_i(X)-a_j(Y)|+\sum_{j=1}^{M} \min_{i=\overline{1,N}} |a_i(X)-a_j(Y)|,$$

where the subscripts $i=\overline{1,N}$ and $j=\overline{1,M}$ are marking the components of the vectors $a(X)$ and $a(Y)$. For brevity we shall henceforth say that $R(X, Y)$ is the distance between the texts X and Y. In other words, we fix a maximum of one text, and then find the nearest maximum of the other text. We calculate the distance between them. After that we find the sum of these distances with respect to all the maxima of the first text. Then we repeat this procedure by letting the two texts change place. As a result we obtain the above number. With such an approach the two texts that are being compared can have a different number of maxima, and we are not obliged to equate their number by introducing multiple maxima. Let us note that such a choice of the measure of proximity is mainly due to simplicity of calculation. It is certainly possible to use also other measures of proximity which are found to be (on the basis of experiment) reliable in distinguishing between pairs of dependent and pairs of independent texts. The reader can see that the following analysis in fact does not use the form of the function $R(X, Y)$.

Now let us utilize a technique which is fairly common in statistics, i.e., we calculate the distribution function $F_0(R)$ of the random variable R(X, Y) for a set of hypotheses which necessarily contains also the hypothesis that the vectors $a(X)$ and $a(Y)$ are independent. After that we find the distance $\hat{R}(X, Y)$ between the specific texts X and Y of interest to us. If the probability of occurrence of such a distance or of a smaller distance is small, then it is natural to discard the hypothesis that the texts X and Y are independent, and assume that they are correlated.

In this paper the function $F_0(R)$ has been calculated by the Monte Carlo method under the following assumptions:

$$a(U_\alpha^n)=\begin{cases}\hat{a}(X), & \alpha=1,\\ a^n(Y), & a=2,\end{cases}$$

where n is the number of the trial, and $\hat{a}(X)$ signifies that the values of the components of the vector $a(X)$ have been calculated on the basis of the text X, with $a_1^n(Y)=Y_1$, $a_i^n(Y)=a_{i-1}^n(Y)+\xi_i$, $i=\overline{2,M}$, $\xi_i \in F(A, S)$, where F is a distribution with a mean

$$A=\frac{1}{N-1}\sum_{i=1}^{N-1}[\hat{a}_{i+1}(X)-\hat{a}_i(X)]$$

and a variance

$$S=\frac{1}{N-1}\sum_{i=1}^{N-1}[\hat{a}_{i+1}(X)-\hat{a}_i(X)-A]^2.$$

TABLE 1

		1	2	3	4	5	6	7	8	9	10	11	12
1	Table of Past Years (850—1110) N=61	0 0	0.550 0.497	0.569 0.515	0.305 0.422								
2	Nikoforov Annals (850—1430) N=83	0.660 0.993	0 0	0.01 0.03	0.001 0.002								
3	Supprasl' Annals (850—1446) N=132	0.840 0.999	0.001 0.004	0 0	0.003 0.003								
4	Academic Chronicle (1336—1446) N=33	0.155 0.699	0.343 0.929	0.375 0.887	0 0								
5	Dvinsk Chronicle (complete) (1390—1717) N=52					0 0	0.015 0.012						
6	Dvinsk Chronicle (short) (1390—1717) N=47					0.013 0.012	0 0						
7	Nikiforov Annals (850—1255) N=31							0 0	0.006 0.008				
8	Suprasl' Annals (850—1255) N=30							0.006 0.005	0 0				
9	Titus Livius "History of Rome" (757-287 B.C.)N=15									0 0	0.002 0.108		
10	Gregorovious "History of Rome" (300-754 A.D.)N=15									0.003 0.130	0 0		
11	Suprasl' Annals (1336—1374) N=15											0 0	0.003 0.58
12	Academic Chronicle (1336—1374) N=15											0.001 0.111	0 0

N is the number of maxima, the first number in the corresponding row is the probability in the case of a normal distribution, and the second number is the probability in the case of a Poisson distribution.

The simulation was performed for the cases that $F(A, S)$ is a truncated normal distribution ($\xi_i \geqslant 0$) or an exponential distribution. The simulation results are listed in Table 1. As the basic vectors used for specifying A and S, we used, in turn, the vectors $\hat{a}(X)$ and $\hat{a}(Y)$.

It is easy to see that our approach can be satisfactorily utilized in the case of vectors $\hat{a}(X)$ and $\hat{a}(Y)$ of roughly the same length. It follows from Table 1 that the procedures used in this section and in Sec. 4 above yield basically the same qualitative results, so that we can hope that our original assumption concerning the representativeness of information about the peaks of the volume functions of historical texts is correct.

It is of interest to study other measures that make it possible to distinguish between pairs of dependent texts and pairs of independent texts. This would enable us to compare the results obtained by using different techniques, and to reach meaningful chronological conclusions. The authors express their gratitude to I. S. Shiganov for his assistance in the calculations.

LITERATURE CITED

1. A. T. Fomenko, "Some statistical regularities in the distribution of the density of information in texts with a scale," Semiotika Inf., No. 15, 99-124, VINITI Press, Moscow (1980).
2. A. T. Fomenko, "Informative functions and corresponding statistical regularities," Abstracts of Reports of the Third International Vilnius Conference on Probability Theory and Mathematical Statistics, Vol. 2, 211-212, Institute of Mathematics and Cybernetics of the Academy of Sciences of the Lith. SSR, Vilnius (1981).
3. A. T. Fomenko, "A technique of recognition of duplicates and some applications," Dokl. Akad. Nauk SSSR, 258, No. 6, 1326-1330 (1981).
4. V. S. Sergeev, Topics in the History of Ancient Rome [in Russian], Vols. 1-2, Moscow (1938).
5. Titus Livius, History of Rome [Russian translation], Vols. 1-6, Moscow (1897-1899).
6. Complete Collection of Russian Chronicles [in Russian], Vol. 33, Leningrad (1977).
7. Complete Collection of Russian Chronicles [in Russian], Vol. 35, Moscow (1980).
8. Tale of Past Years. Literary Monuments of Ancient Russia [in Russian], Khud. Lit. Press, Moscow (1978).
9. F. Gregorovius, History of Mediaeval Rome, SPB (Collection of Works) [in Russian], Vols. 1-5 (1902-1912).
10. A. T. Fomenko, "New statistical experimental techniques of dating of ancient events and applications to the global chronology of the ancient and mediaeval world," Preprint No. B07201, Nov. 9, 1981, State Committee for Television and Broadcasting, Moscow (1981).
11. A. T. Fomenko, "A new empirical statistical technique of ordering of texts with applications to dating problems," Dokl. Akad. Nauk SSSR, 268, No. 6, 1322-1327 (1983).
12. A. T. Fomenko, "Calculation of the second derivative of the Moon's elongation and statistical regularities in the distribution of certain astronomical data," Operations Research and Control Systems [in Russian], Vyshcha Shkola, No. 20, Kiev (1982), pp. 98-113.
13. A. T. Fomenko, "The jump of the second derivative of the moon's elongation," Celestial Mech., 29, 33-40 (1981).
14. A. T. Fomenko, "On the properties of the second derivative of the moon's elongation and related statistical regularities," Problems of Computational and Applied Mathematics, No. 63, 136-150, Tashkent (1981).
15. A. T. Fomenko, "The author's invariant of Russian literary texts," Methods of Quantitative Analysis of Texts of Narrative Sources, 86-109, Inst. of History of the USSR, AN SSSR, Moscow (1983).
16. A. T. Fomenko, "On the geometry of distribution of integer points in hyperregions," Proceedings Seminar on Vector and Tensor Analysis, No. 21, Moscow State Univ. (1983), pp. 106-152.