IDENTIFYING DUPLICATES IN CHRONOLOGICAL LISTS (THE METHOD OF FREQUENCY HISTOGRAMS FOR THE SPREAD OF RELATED NAMES)

G. V. Nosovskii and A. T. Fomenko

UDC 519.21

1. Introduction

The present article is devoted to empirical/statistical methods of identifying duplicates in a collection of texts (lists) containing a time scale. We give examples of the application of these methods to texts on ancient and medieval history. Duplicates are defined as texts (fragments) that are dependent in the framework of the statistical model being used. Thus we have in mind "statistical duplicates" (the substantive interpretation of them is a separate question, cf. [3]). The results obtained in this article are a continuation of studies in the construction of an "optimal statistical chronology" begun by A. T. Fomenko in a series of publications [1]–[5].

The problem of identifying duplicates was stated by Fomenko in [1] and [2], where several empirical/statistical methods of solving it were proposed. All these methods were tested on reliable material and their efficiency was confirmed completely. The results of applying them to historical texts (chronicles, annals, chronological tablets, and others) turned out to be consistent with one another and were summarized by Fomenko in the decomposition of a global chronological chart into four layers duplicating one another C_1, C_2, C_3, and C_4 (cf. [1– 6]). This decomposition, roughly speaking, means that the global chronological chart (a modern "textbook" on ancient and medieval history) can be obtained from the shorter chronicle C_1 using the following procedure of "reproduction and shuffling." The chronicle C_1 is taken in four copies C_1, C_2, C_3, C_4 (reproduction) and then copies C_2, C_3, and C_4 are shifted backwards along the time axis with respect to the original chronicle C_1 by 333, 1053, and 1778 years respectively. Next all four copies are "injected" into one another (as in shuffling a deck of cards), forming as a result the entire global chronological chart.

The procedure just described of reproduction and shuffling of the text (the initial "structure") is a simplified model of the appearance of duplicates. In reality numerous local distortions may be imposed on the global structure of systems of duplicates arising from this procedure, along with "imbedded" systems of shorter duplicates. In other words the original data may be very "noisy." Nevertheless the fact that there exists a global duplicate structure in the text (list) being studied and that this structure arises from reproduction and shuffling of a shorter list can be verified by the methods of mathematical statistics. A method of verification was proposed by the authors in [8] and [9].

In the present article we describe an empirical/statistical procedure for identification and subsequent analysis of the duplicate structure in very noisy data arising as a result of reproduction and shuffling. The ideas on which this method is based are closely connected with the principle of frequency damping stated by Fomenko in [1] and [2]. The method includes the following stages: 1) formalization of the problem and construction of a finite probabilistic scheme from the initial data, in terms of which the hypothesis of the existence of a global duplicate structure in the text being studied is stated; 2) testing this hypothesis and, when duplicates are detected, determining the size of the shifts between the basic duplicate systems.

The application of this procedure to texts of historical content gave results consistent with the decomposition found by Fomenko for the global chronological chart [1–4]. In the present article we give several examples. In doing so we take chronological lists of names (rulers) as initial data. We note that the decomposition of the global chronological chart was originally obtained by Fomenko in [1–4] on the basis of analysis of data of a completely different type—the functions that give the size of the texts and lengths of reigns. The fact that results consistent with this decomposition can also be obtained in statistical analysis of lists of names is worthy of attention.

Translated from: *Problemy Ustoichivosti Stokhasticheskikh Modelei, Trudy Seminara*, 1989, pp. 112–125.

0090-4104/92/5904-0980 $12.50

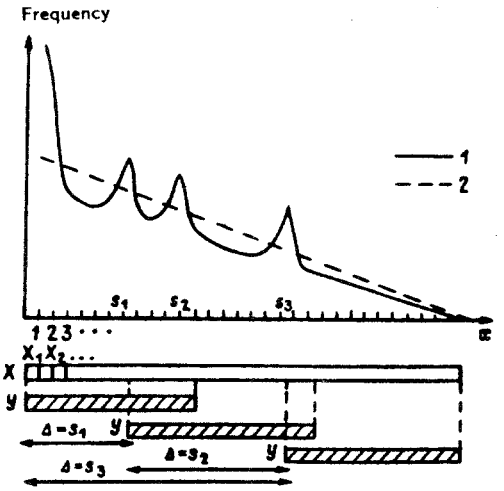

Fig. 1. Frequency histograms for the spreads:

1—frequency histogram for the spreads of related names in the list X (conditional probability distribution $\mathbf{P}(x|A_1)$); *2*—distribution of the random variable for the list X

2. A model problem and formalization of the hypothesis of the existence of duplicates in a list

1. Consider the problem of identifying duplicates, which will serve us as a typical example in constructing the method. We shall assume that we have at our disposal a certain sequence of cards X (a deck of cards), which may contain repeated cards. We shall say that the deck X *contains duplicates* if it has been obtained from several shorter decks of cards Y that are identical in content and order (possibly containing repeated cards), that were combined to form a single deck $YY\ldots Y$ and then shuffled. When this was done, before shuffling each of the original (short) decks Y may have been distorted in a random manner. Assume that these local distortions in the various parts of the original decks Y are pairwise independent. Distortions here are taken to mean exclusion, duplication, or replacement of an individual card or a sequence of successive cards (a segment of the deck). When the deck of cards X being studied contains no duplicates we shall call the order of cards in X *correct.*

We shall say that two segments Δ_1 and Δ_2 of the deck X separated from each other contain layers that duplicate each other or that they are duplicates for short, if they contain respectively cards that were originally close to each other in the initial ordering of cards in the deck Y.

The problem is to test the hypothesis H_0 that the order of cards in X is correct, i.e., that X has no duplicates, from the known sequence of cards in the deck X. If the hypothesis H_0 is rejected, one must determine which segments of the deck X duplicate one another (are duplicates).

2. We now state a corollary of the hypothesis H_0 that can be verified by the methods of mathematical statistics.

Let the total number of cards in the deck X be n, of which k are distinct. Denote the set of different kinds of cards occurring in X by $t_1,\ldots,t_k$. Partition the deck X into segments of equal length: $X = X_1X_2\ldots X_N$, where N is the number of segments in the partition. Let each of the segments X_i contain p cards, where $p \ll n$.

Consider a finite probabilistic scheme of equally likely choice with replacement of two cards from the deck X (i.e., we choose one card from the deck X at random, make a note of it, return it to its place, and then choose another card in the same way). To the chosen pair we assign their *spread* ρ in the deck X, i.e., the absolute value of the difference of the indices of the segments X_{i_1} and X_{i_2} containing the two cards: $\rho = |i_1 - i_2|$. The spread is a random variable in the probabilistic scheme just introduced. We shall denote it by ξ. The distribution of ξ is easy to calculate (cf. below).

Let A be an event in the probabilistic scheme under consideration. We shall call the event A a *local event* if it is determined by the contents of the cards in just one segment X_i of the partition (which may depend on the case). The simplest example of a local event is the event A_0 that some segment of the

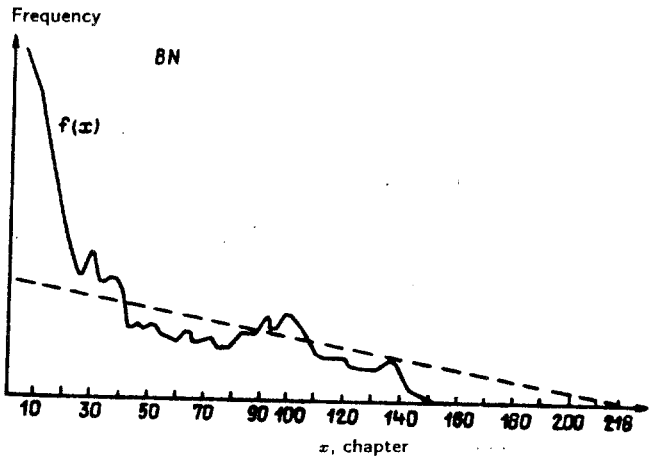

Fig. 2. Frequency histograms for the spreads in the list BN
For the notation cf. Fig. 1

partition of the deck X contains cards of the both of the kinds chosen.

Consider the conditional distribution $\mathbf{P}(\xi = x|A)$ of the random variable ξ given that the local event A has occurred.

We shall use the following intuitively obvious proposition on the statistical properties of a correct ordering of the cards in a deck. If the deck X contains no duplicates or it has been completely shuffled and the initial duplicate structure cannot be recovered, a local condition imposed on the kinds of cards chosen cannot influence the character of the global distribution of cards of these kinds in the entire deck, i.e., cannot cause any noticeable change in the distribution of the random variable ξ outside a certain neighborhood of zero, determined by the radius of damping of dependence between local distortions in different parts of the deck (cf. above). This means that if the order of cards in X is correct, then the distribution $\mathbf{P}(\xi = x|\, A,\, \xi \geq \varepsilon)$ must coincide with (be close to) the distribution $\mathbf{P}(\xi = x|\, \xi \geq \varepsilon)$ for any local event A and some $\varepsilon > 0$ determining the length of a "local segment" of the deck X. And if the hypothesis H_0 is false and the deck X contains duplicates, then for each segment X_{i_0} of the partition cards of the kinds occurring in X_{i_0} will as a rule also occur in the duplicates of the segment X_{i_0}. This leads to the conclusion that the spread of cards of these kinds will more often be near to zero or to values of the basic shifts in the duplicate system containing X_{i_0} than the spread of an arbitrarily chosen pair of cards. Consequently when X contains duplicates, there exist local events A such that the conditional distribution $\mathbf{P}(\xi = x|A)$ differs significantly from the unconditional distribution of ξ on the whole interval of possible values $0 \leq x \leq N-1$. The event A_0 defined above is an example of such a local event.

This proposition makes it possible to test the hypothesis H_0. Moreover, it turns out that by comparing these distributions among themselves for different A one can determine the typical shifts between duplicates in X.

3. The spread of related names

1. Suppose given a chronological list of names $X = \{x_1, \ldots, x_n\}$, i.e., a list consisting of names with an indication of an interval of time (lifetime, dates of reign, and the like) for each name. The names in X are arranged in chronological order and may be repeated. Let the total number of names in X be n, of which k are distinct. Assume that the list X is partitioned into segments $X_1, \ldots, X_N$ of approximately the same length, and the length of each segment X_i (the number of names occurring in it) is much smaller than the length of the whole list. The partition segments $X_1, \ldots, X_N$ will be called the *chapters* of the list. In what follows we shall give examples of real chronological lists partitioned into chapters.

Consider the problem of verifying the hypothesis H_0 that the chronology of X is "correct" (or "optimal in the statistical sense"). We shall call the chronology (the arrangement of names on the time axis) of a list X that is not the result of reproduction and shuffling any other shorter list Y *correct* (cf. above). Here, as in the model problem with a deck of cards, we admit the possibility of random distortions of copies of the list Y used in the shuffling and we assume that the local distortions in different parts of the list are

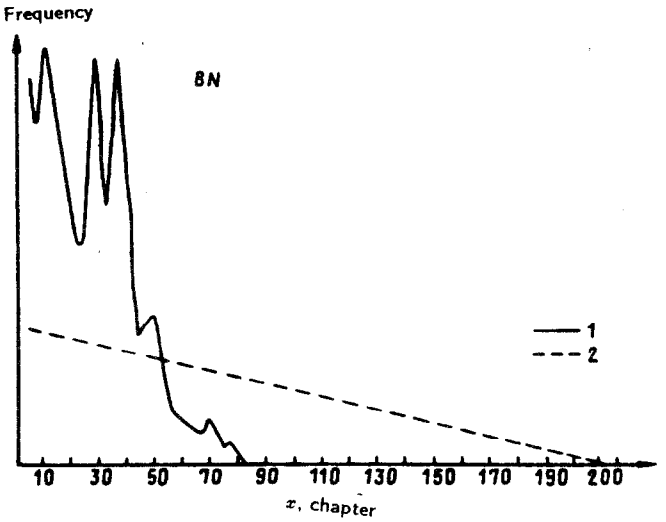

Fig. 3. Frequency histogram for spreads of names related in a subset of chapters of the list BN: *1*—The histogram $f^D(x)$, where $D = \{X_{101}, \ldots, X_{213}\}$ is a subset of the chapters of the list; *2*—for the notational conventions cf. Fig. 1

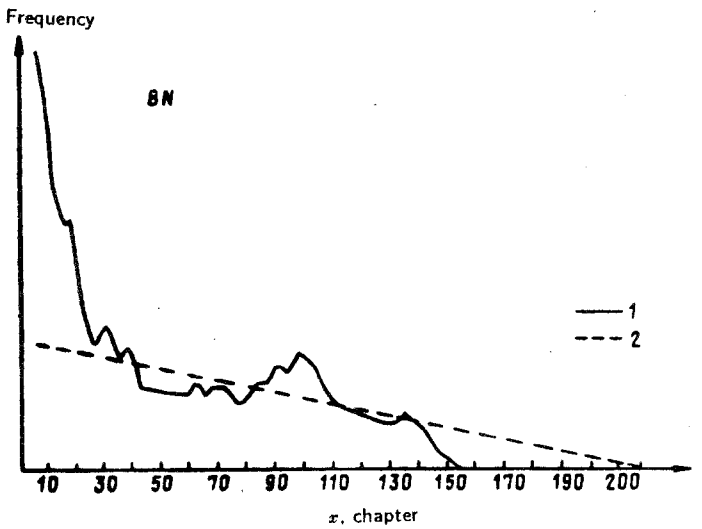

Fig. 4. Frequency histogram for spreads of names related in a subset of chapters of the list BN: *1*—The histogram $f^D(x)$, where $D = \{X_1, \ldots, X_{100}\}$ is the first hundred chapters of the list

independent (cf. Sec. 2).

Following the method of Sec. 2 we consider a probabilistic scheme of a random equally likely choice of two names from the list X with replacement and define the random variable ξ in this scheme to be the spread of the names chosen:

$$\xi = |i_1 - i_2|,$$

where i_1 and i_2 are the indices of the chapters containing the names chosen. The random variable ξ assumes nonnegative integer values from 0 to $N - 1$. We shall need the following simple proposition.

2. Lemma. *When the number of names in all the chapters of the list X is the same, the probability distribution of the random variable ξ is given by the formula*

$$\mathbf{P}(\xi = 1) = \begin{cases} \frac{1}{N}, & x = 0, \\ \frac{2(N-x)}{N^2}, & 1 \le x \le N, \end{cases} \tag{1}$$

i.e., it is a linearly decreasing function for $1 \le x \le N$.

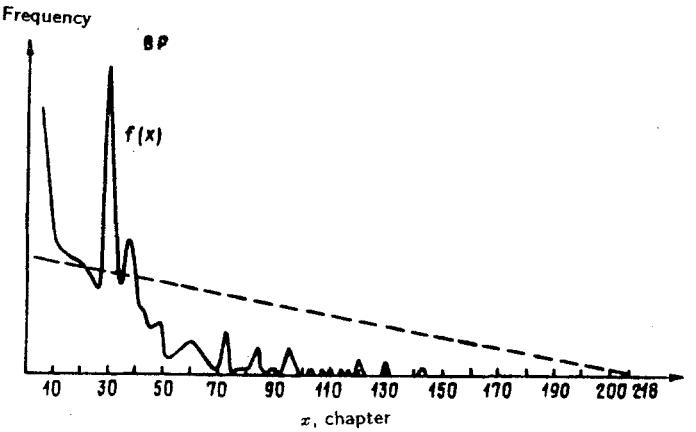

Fig. 5. Frequency histograms for spreads in the list BP
For the notational conventions see Fig. 1

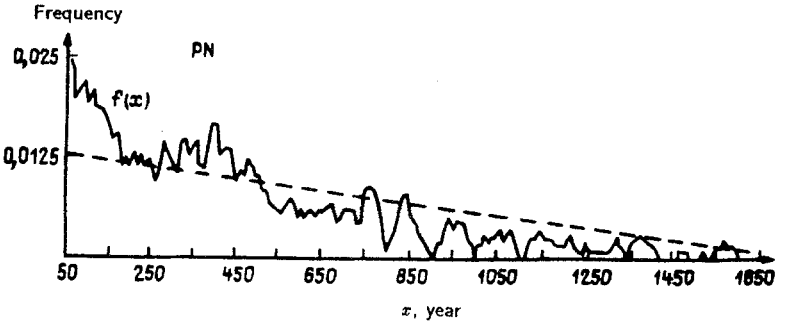

Fig. 6. Frequency histogram for the spreads in the list PN
For the notational conventions see Fig. 1

Proof. Since the random variable ξ is determined by the number of chapters containing the names chosen, we may assume that it is not the names themselves but the chapters that have been chosen. Since the size of the chapters is constant by hypothesis, the choice of each fixed chapter at the first step occurs with the same probability $\frac{1}{N}$. The same is true for the second step of the choice. Let $1 \leq x \leq N$. Then there exist exactly $N - x$ possibilities for fixing the chapter with smaller index in a pair of chapters with spread x (i.e., whose indices differ by x). The second chapter in the pair has an index that is larger by x and is thereby determined uniquely. Taking account of the fact that the chapter with smaller index may occur either as the first or the second step in the choice, we find that the number of ways of choosing a pair of chapters with spread x is $2(N - x)$. The probability of choosing a fixed pair of chapters taking account of the order of choice is $\frac{1}{N^2}$, so that by the formula for total probability $\mathbf{P}(\xi = x) = \frac{2(N-x)}{N^2}$. If $x = 0$, the same chapter occurs at both steps in the choice. There are N chapters in all and each of them can be chosen twice in succession with probability $\frac{1}{N^2}$. Consequently $\mathbf{P}(\xi = 0) = \frac{1}{N}$. The lemma is now proved.

Computations for real chronological lists have shown that the distribution of ξ has the form (1) even when the sizes of the chapters in the list are only approximately equal. In other words the form of the distribution of ξ is stable with respect to variations in the size of the chapters. However there are cases when a chronological list of names is partitioned into chapters of markedly different sizes. In this case the list must be normalized by dividing the multiplicity of occurrences of names in each chapter by the size of the chapter (to avoid considering fractional multiplicities one can multiply all multiplicities by the product of the sizes of the original chapters). After this norming the sizes become equal. Therefore without loss of generality we shall assume that *the probability distribution* $\mathbf{P}(\xi = x)$ *is a linearly decreasing function on the set* $1 \leq x \leq N$ *equal to* 0 *when* $x = N$.

3. Let A be a local event (cf. Sec. 2). According to a corollary of the hypothesis H_0, if the chronology of the list X is correct, then the probability distributions $\mathbf{P}(\xi = x | \xi \geq \varepsilon)$ and $\mathbf{P}(\xi = x | A, \xi \geq \varepsilon)$ must coincide (be close to each other). The number ε is chosen depending on the character of the data. It is determined

by the a priori radius of damping of dependence between the individual chapters of the list X under the hypothesis that the chronology of X is correct (cf. Sec. 2). In the examples given below each chapter corresponds to approximately one generation (10–20 years) and ε is given the values 5–10 (chapters).

Definition. The *frequency histogram of the spread of related names in the list* X is the conditional probability distribution $\mathbf{P}(\xi = x | A)$, where A is an arbitrary local event.

We now state the corollary of the hypothesis H_0 in the form in which we shall use it.

If the chronology of the list X is correct, then the frequency histograms of the spreads of related names in X on the set $\varepsilon \leq x \leq N$ must coincide with (be close to) a linearly decreasing function equal to 0 for $x = N$.

If the chronology of the list X is correct, then the probability distributions $\mathbf{P}(\xi = x | \xi \geq \varepsilon)$ and $\mathbf{P}(\xi = x | A, \xi \geq \varepsilon)$ must coincide (cf. above). But the first of these, according to the lemma just proved, is a linearly decreasing function on the set $1 \leq x \leq N$ equal to zero for $x = N$. Consequently $\mathbf{P}(\xi = x | A, \xi \geq \varepsilon)$ decreases linearly on the set $\varepsilon \leq x \leq N$, from which it follows that the frequency histogram for the spreads of related names $\mathbf{P}(\xi = x | A)$ is also linearly decreasing on this set.

4. In the repeated choice scheme we are considering we denote the first name chosen from the list X by a_1 and the second name by a_2. We define the event $A_1 = \{(a_1, a_2) :$ there exists a chapter X_i such that the first occurrences of the names a_1 and a_2 in the list X are in chapter $X_i\}$.

In the case when the event A_1 happens, the names a_1 and a_2 will be called *contemporaries* (with respect to the chronology of the list X). The event A_1 is a local event, since it is determined for each elementary outcome from the contents of just one chapter. Consequently in the case of a correct chronology of the list X the frequency histogram for the spreads of related names (names of contemporaries)

$$f(x) = \mathbf{P}(\xi = x | A_1) \tag{2}$$

with $\varepsilon \leq x \leq N$ must coincide with (be close to) a linearly decreasing function equal to zero for $x = N$. If the list X contains duplicates, names that first appeared in the list in the same chapter will as a rule be repeated in the duplicates of this chapter. Thus the variables giving the spreads of such names will assume values close to the typical shifts between duplicates in the list X with a higher frequency. On the frequency histogram for the spreads of related names (2) this leads to the appearance of spikes at the values of the fundamental shifts in the duplicate systems. This is illustrated in Fig. 1 using the example of a list X that is the result of shuffling three copies of a shorter list Y shifted by s_2, s_3, and $s_1 = s_3 - s_2$ chapters with respect to one another. The list X (cf. Fig. 1) is partitioned into chapters $X_1, X_2, \ldots, X_N$ of equal size.

5. In the case when the list X contains several duplicate systems, the shifts between duplicates in different systems do not coincide. Then the frequency histogram (2) will contain mixtures of spikes at the values of all the shifts in all the duplicate systems. However in this situation the spikes corresponding to systems that contain only a small number of duplicates will "belong to the histogram" with small weight and may not be noticeable in it. To distinguish the shifts in different duplicate systems we consider other local events and the frequency histograms for related names determined by them.

Let $D = \{X_{i_1}, \ldots, X_{i_m}\}$ be a certain set of chapters of the list X (a subset of all the chapters). To exhibit the shifts between duplicates of the chapters in the set D we define the event $A_1^D = \{(a_1, a_2) :$ there exists a chapter X_i in the set D such that the first occurrences of the names a_1 and a_2 in the list X both belong to chapter $X_i\}$.

The event A_1^D is a local event. It differs from the event A_1 (defined above) in the additional requirement that the chapter X_i containing the first occurrences of a_1 and a_2 in the list X must belong to the set D. The frequency histogram for the spreads of related names

$$f^D(x) = \mathbf{P}(\xi = x | A_1^D) \tag{3}$$

behaves in a manner similar to that of the frequency histogram (2), the only difference being that when there are duplicates in X the spikes on the frequency histogram (3) arise only at values of the shifts between

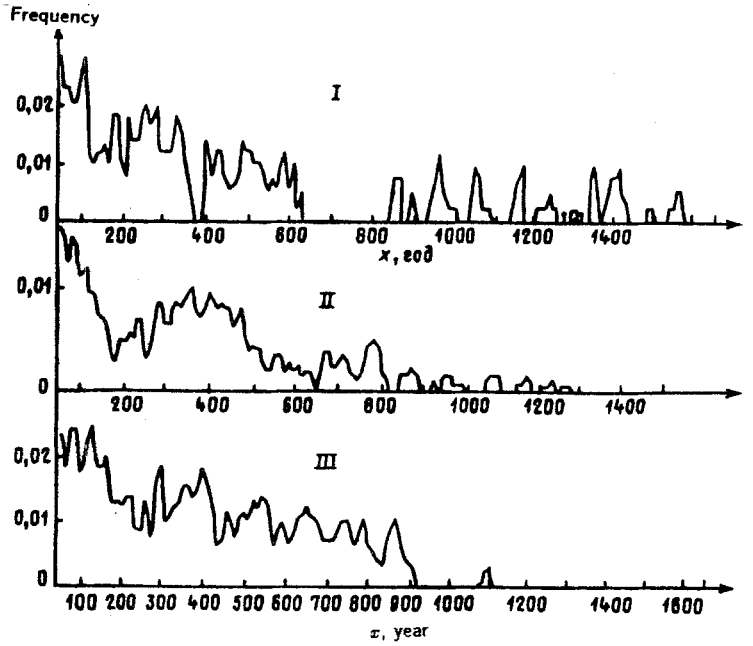

Fig. 7. Frequency histograms for the spreads of related names in different subsets of the chapters of the list PN

I—$f^D(x)$, $D = \{X_1, \ldots, X_{20}\}$ is the set of chapters of the list PN covering the time interval 50–250 C. E.
II—$f^D(x)$, $D = \{X_{31}, \ldots, X_{50}\}$ is the set of chapters of the list covering the time interval 250–550 C. E.
III—$f^D(x)$, $D = \{X_{51}, \ldots, X_{80}\}$ is the set of chapters of the list covering the time interval 550–850 C. E.

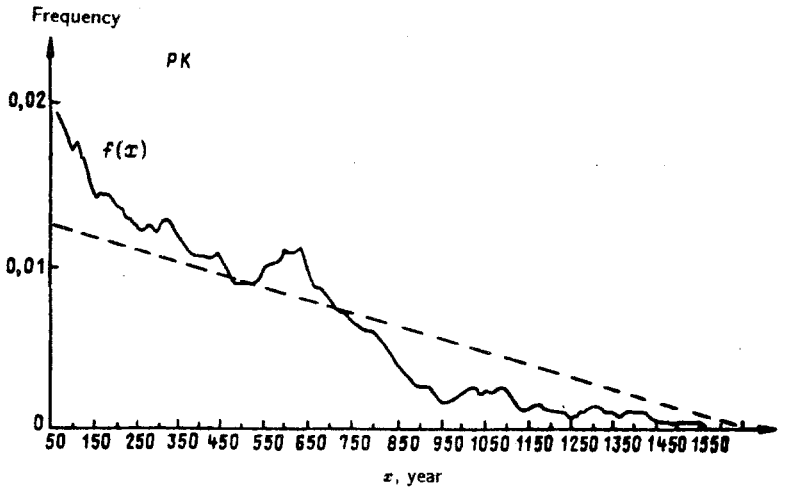

Fig. 8. Frequency histograms for the spreads in the list PK

For notational conventions cf. Fig. 1

the duplicates of chapters of the set D (which themselves, of course, may fail to belong to D). Comparing the frequency histograms (3) constructed for different subsets of chapters of the list X makes it possible to answer the question whether X contains only one duplicate system (if any duplicates are detected in X at all) or whether there are several of these systems and if so, whether they possess different fundamental shifts.

6. We now turn to the analysis of specific chronological lists of names. As the first illustrative example we consider the list BN of names mentioned in the Bible, endowed with a partition into "chapter/generations," i.e., into text fragments describing the events of approximately one generation (cf. [3]). A complete list of names of the Bible, counting multiplicity and with a partition into chapter/generations was compiled by V. P. Fomenko and T. G. Fomenko. The list BN contains tens of thousands of names, of which about 2000 are distinct. There are 218 chapters in the list. Multiple occurrences of all the names in the chapters were

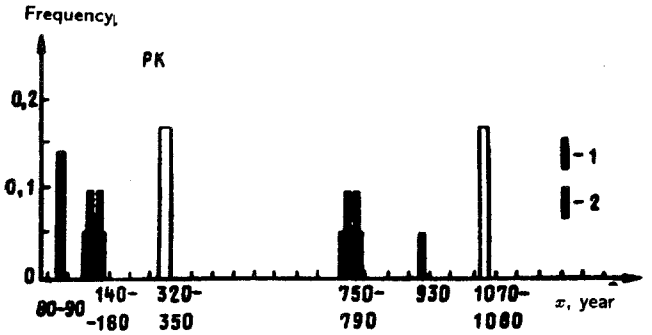

Fig. 9. Frequency histograms for spreads of related names in different chapters of the list PK
1—$f^D(x)$, $D = \{X_{21}, \ldots, X_{50}\}$ is the set of chapters in the list covering the time interval 250–550 C. E.
2—$f^D(x)$, $D = \{X_{51}, \ldots, X_{80}\}$ is the set of chapters in the list covering the time interval 550–850 C. E.

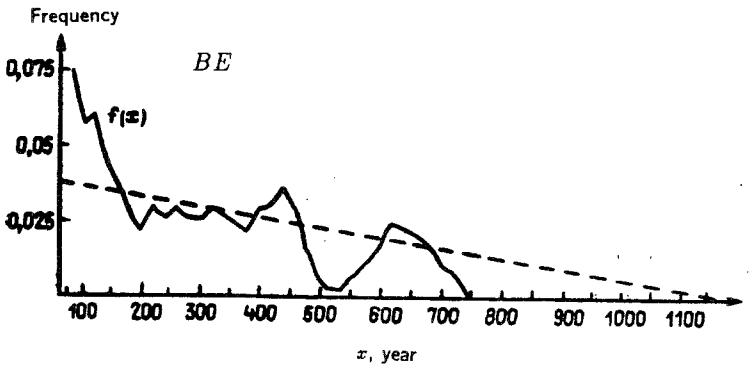

Fig. 10. Frequency histograms for spreads in the list BE
For notational conventions cf. Fig. 1

normalized (cf. above). We note that the partition of the list BN into chapter/generations we are using does not coincide with the canonical division of the Bible into chapters and verses.

It is known that there are duplicates in the Bible: chapter/generations 98–137 (I Samuel, II Samuel, I Kings, II Kings) and 138–167 (I Chronicles, II Chronicles) describe the same events. Consequently there should be a spike in the frequency histograms for the spreads of related names in the list BN at the values of 30–40 chapters ($138 - 98 = 40$, $167 - 137 = 30$). This is indeed the case. Figure 2 shows the frequency histogram of the spreads of related names (2) for the list BN with $5 \leq x \leq 218$ (in the interval $0 \leq x \leq 5$ the graph grows rapidly with approach to zero). The graph (Fig. 2) shows two extensive spikes in intervals of spread 30–40 and 90–110 chapters. Spikes are observed at spreads of 10 and 140 chapters (approximately).

We now give also the frequency histograms (3) for the spreads of related names in the list BN in various subsets of chapters of the list. Figure 3 shows the frequency histogram (3) for the set of chapters $\{X_{101}, \ldots, X_{218}\}$, which constitutes the second half of the list BN. The known series of duplicates in BN pointed out above is almost entirely contained in the chapters of this subset. The spike corresponding to this series in the spread interval of 30–40 chapters shows up much more prominently in Figure 3. It seems to consist of two spikes very close to each other at values of 30 and 36 chapters. Moreover three other sharp spikes show up in Fig. 3 (at spread values of 10, 47–54, and 72–73), showing that there are duplicates in the list BN separated by these values. Figure 4 gives the frequency histogram (3) for the set D consisting of the first 100 chapters of the list BN. This histogram shows that some of the chapters of the first half of the list also have duplicates. Both Fig. 4 and Fig. 3 have a double spike in the spread interval 30–40. This means that the duplicates of chapters 98–167 are also contained in the first half of the list BN. The facts just enumerated about the duplicates in the list BN were discovered earlier by A. T. Fomenko using other methods [1–3]. We note that if parallel passages (the list BP) are taken instead of names, the location of the set of spikes in the frequency histograms for the spreads of related names (2) and (3) is hardly altered. Figure 5 provides for purposes of comparison the frequency histogram (2) for the list BP. Comparison of

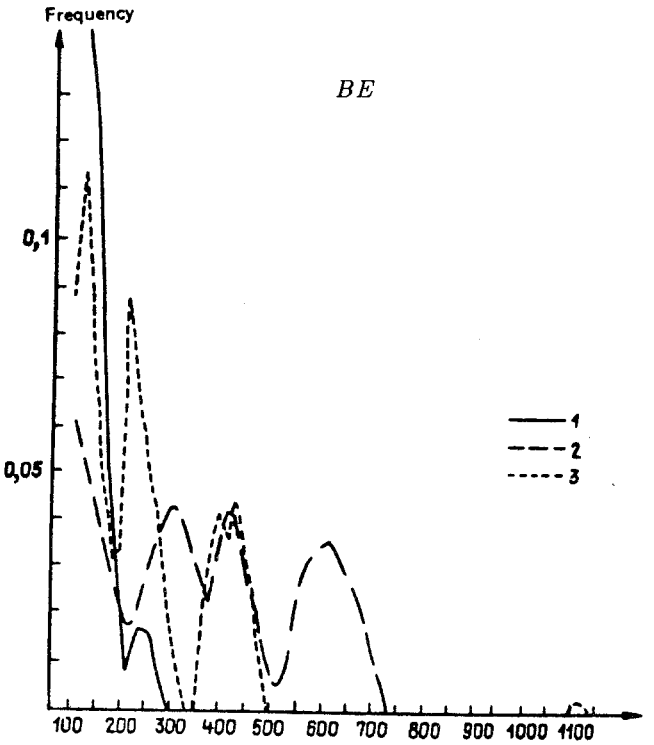

Fig. 11. Frequency histograms for the spreads of related names in different subsets of chapters of the list BE

1—$f^D(x)$, where $D = \{X_{40}, \ldots, X_{57}\}$ are the chapters in the list covering the time interval 1100–1453 C. E.
2—$f^D(x)$, where $D = \{X_1, \ldots, X_{19}\}$ are the chapters in the list covering the time interval 300–700 C. E.
3—$f^D(x)$, where $D = \{X_{20}, \ldots, X_{39}\}$ are the chapters in the list covering the time interval 700–1100 C. E.

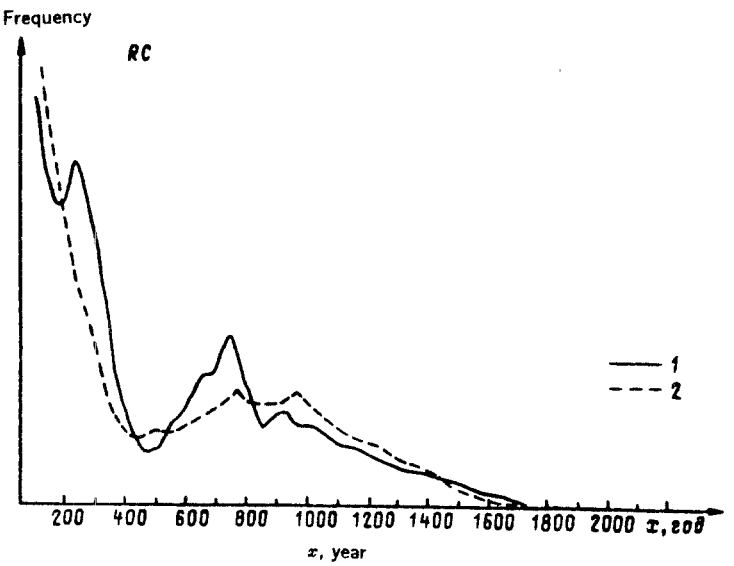

Fig. 12. Frequency histograms for the spreads in the list RC, which contains gaps:
For notational conventions cf. Fig. 1

Figs. 2 and 5 shows also that the amplitudes of the spikes, in contrast to their abscissas, depend strongly on which elements of the text (names, parallel passages, and the like) are used to compile the list.

7. We now give the results of computing the frequency histograms of related names for the lists PN and PK of names and nationalities of Popes (called Pontifici Maximi until the 11th century). These lists were

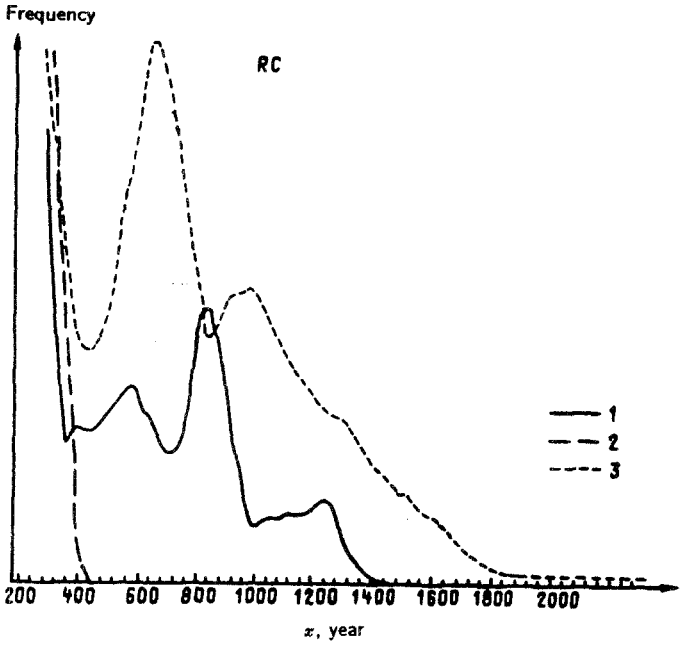

Fig. 13. Frequency histograms for the spreads of related names in different subsets of the chapters of the list RC :

1—$f^D(x)$, where $D = \{X_{41}, \ldots, X_{82}\}$ are the chapters of the list covering the time interval 50–850 C. E.
2—$f^D(x)$, where $D = \{X_{83}, \ldots, X_{123}\}$ are the chapters of the list covering the time interval 850–1700 C. E.
3—$f^D(x)$, where $D = \{X_1, \ldots, X_{40}\}$ are the chapters of the list covering the time interval 753 B. C. E.—50 C. E.

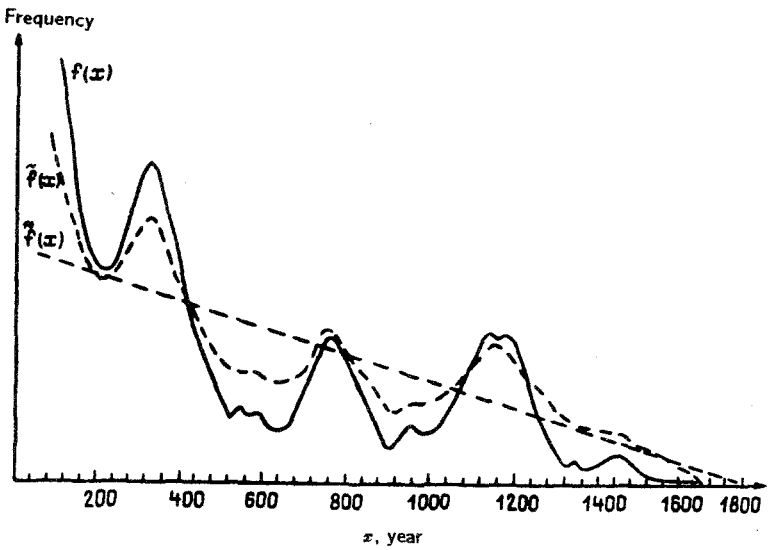

Fig. 14. Destruction of duplicate structures and the corresponding alteration of frequency histograms for the spreads of related names under a random perturbation of the list

$f(x)$ is the initial histogram for the spreads of related names in the list; $\tilde{f}(x)$ is the analogous histogram for the same list after the addition of 30 copies of the same name; $\tilde{\tilde{f}}(x)$ is the frequency histogram for the spreads of related names of the same list after a random permutation of 20% of its names (the duplicate structure is totally destroyed)

compiled from data known today on all the Popes and Antipopes in the history of the Roman Church from the year 50 C. E. until 1950 [11]. The lists PN and PK were divided into chapters of 10 years. The length

of each list is 293, and there are 190 chapters. The list PN contains 87 different names. The lists PN and PK were not normalized, since the probability distribution of ξ for them hardly differs from (1). Figure 6 shows the frequency histogram of the spreads of related names (2) for the list PN.

Figure 7 gives three frequency histograms of the form (3) for the list PN with different choices of the subset D of chapters of the list. On the time axis these subsets of chapters are situated as follows: $\{X_1, \ldots, X_{20}\} \approx$ (50–250 C. E.), $\{X_{21}, \ldots X_{50}\} \approx$ (250–550 C. E.), and $\{X_{51}, \ldots, X_{80}\} \approx$ (550–850 C. E.). First occurrences of names are very rare in chapters with larger indices, and it makes no sense to study the frequency histograms $f^D(x)$ for these chapters in view of the smallness of the sample.

Figure 8 gives the frequency histogram for the spreads of related names (2) for the list PK and Fig. 9 gives the frequency histogram (3) for this list with $D = \{X_{20}, \ldots, X_{50}\} \approx$ (250–550 C. E.) and $D = \{X_{50}, \ldots, X_{80}\} \approx$ (550–850 C. E.).

Figures 6–9 show the spread values in years (1 chapter = 10 years). We remark that there is a 300-year shift in both lists.

8. We now give the results for the list BE of names of Byzantine emperors (from 307 to 1453 C. E.). The list was broken up into 57 chapters of 20 years each. The list BE contains 151 names, of which 63 are distinct. It was not normalized, since for it the distribution of ξ is well approximated by formula (1).

Figure 10 gives the frequency histogram (2) for the list BE. The list BE contains duplicates between which the fundamental shifts are $\approx$ 440 and 620 years. More detailed information on the shifts in the duplicate systems of the list BE can be obtained by analyzing the frequency histograms of the form (3). Figure 11 gives these histograms for the subsets of chapters $D = \{X_1, \ldots, X_{19}\} \approx$ (300–700 C. E.), $D = \{X_{20}, \ldots, X_{39}\} \approx$ (700–1100 C. E.), and $D = \{X_{40}, \ldots, X_{57}\} \approx$ (1100–1453 C. E.). It appears from Fig. 11 that this method detects no duplicates to the right of the year 1100 in the list BE. The fundamental shift of 330 years, which does not show up in Fig. 10, nevertheless exists in the list BE as a shift between the duplicate chapters 1–19 (cf. Fig. 11).

9. Consider the list RC of Roman Caesars (from 753 B. C. E to 1700 C. E.). This list also includes the names of persons who did not formally bear the title of Caesar, but had de facto power. The list RC contains 555 names of which 193 are distinct. It was broken into chapters of 20 years each. The total number of chapters was thus 123, but some of them are empty, since during the periods of the two Roman republics there are no names of emperors. The presence of empty chapters complicates the normalizing of the list. The list RC therefore was not normalized, and Fig. 12 gives both the frequency histogram of the random variable ξ and the frequency histogram for the spreads of related names (2). The fundamental shift in the list RC determined from Fig. 12 is a shift of $\approx$ 780 years. From frequency histograms of the form (3) for $D = \{X_1, \ldots, X_{41}\} \approx$ (753 B. C. E.–70 C. E.), $D = \{X_{42}, \ldots, X_{83}\} \approx$ (70–870 C. E.) and $D = \{X_{84}, \ldots, X_{123}\} \approx$ (850–1700 C. E.), given in Fig. 13, other shifts show up, in particular fundamental shifts of 300–320, 780, and 1050 years in the history of Rome [1–6].

10. In conclusion we give graphs that show how the duplicate structure of a list is destroyed by a random perturbation. Fig. 14 gives three graphs. One of them shows the frequency histogram for the spreads of related names (2) for a list containing duplicates. The second shows the same histogram for the perturbed list, into which one and the same name has been inserted into some of the chapters (30 of the 180). Finally the dashed line shows the same histogram for a list 20% of whose names have been randomly permuted. This line coincides with the frequency histogram of ξ for the list under consideration.

These graphs show that the duplicate structure in a list X is rapidly destroyed by shuffling and the frequency histograms for the spreads of related names approach a straight line. Consequently if the frequency histograms for the spreads of related names nevertheless differ significantly from a straight line for a list with a uniform partition into chapters, this means that the list contains duplicates (in the statistical sense).

Literature Cited

1. A. T. Fomenko, "Some statistical regularities in the distribution of the density of information in texts with a scale," *Semiotika i Informatika*, No. 15, 99–124 (1980).

2. A. T. Fomenko, "Informational functions and the statistical regularities connected with them," in: *Proc. III Intern. Conf. Th. Prob. Math. Stat.* [in Russian], Inst. Mat. Kib. Akad. Nauk Lit. SSR, Vol. 2 (1981), pp. 211–212.
3. A. T. Fomenko, "New experimental/statistical methods for dating ancient events and applications to the global chronology of the ancient and medieval world," Preprint: Gosteleradio, Moscow (1981).
4. A. T. Fomenko, "A method of identifying duplicates and some applications," *Dokl. Akad. Nauk SSSR*, **258**, No. 6, 1326–1330.
5. A. T. Fomenko, "A new empirical/statistical method of ordering texts and applications to dating problems," *Dokl. Akad. Nauk SSSR*, **268**, 1322–1327 (1983).
6. A. T. Fomenko, "A new empirical/statistical method of detecting parallelisms and dating duplicates," *Probl. Ustoich. Stokh. Mod.*, 154–177 (1984).
7. V. V. Kalashnikov, S. T. Rachev, and A. T. Fomenko, "New methods of comparing size functions of historical texts," *Probl. Ustoich. Stokh. Mod.*, 33–45 (1986).
8. G. V. Nosovskii and A. T. Fomenko, "On the determination of the original structure in mixed sequences," *Tr. Sem. Vekt. Tenz. Anal.* (MGU), No. 22, 105–119 (1985).
9. A. T. Fomenko, "Duplicates in mixed sequences and the principle of frequency damping," in: *Proc. IV Intern. Conf. Theory Prob. Math. Stat.* [in Russian], Inst. Mat. Kib. Akad. Nauk Lit. SSR, Vol. 3 (1985), pp. 246–248.
10. J. Blair, *The Chronology and History of the World*, London (1779).
11. S. G. Lozinskii, *History of the Papacy* [in Russian], Moscow TsS SVB SSSR (1934).
12. E. J. Bickerman, *Chronology of the Ancient World*, Cornell University Press, Ithaca (1968).